Chapter 6

Bayesian Inference for Categorized Data

Jean-Marc Bernard

Introduction

This chapter presents Bayesian parametric inference and Bayesian predictive inference for categorized data, that is for problems involving one or several frequencies. It is as self-contained as possible and may be considered as a general introduction to the Bayesian methodology, even though the operational techniques described are specific to categorized data. One definite advantage of presenting Bayesian inference on this type of data is that the sampling model does not involve arbitrary technical assumptions (*e.g.* normality, *etc.*). Being partly freed from the burden of technical aspects, it becomes easier to focus on the concepts underlying the methods.

The Bayesian approach to inference has often been criticized as being subjective because, besides the data themselves, it requires an external element, namely the prior distribution. However, we shall see that, when one adopts what we call with Rouanet (see Chapters 1

and 2) the *data analysis methodology*, the arbitrariness involved in the choice of the prior distribution is much reduced and not larger than the one existing in the frequentist framework. The data analysis methodology leads to proposing, for each situation, a "reference prior distribution" that provides a "standard Bayesian analysis". In this chapter we shall insist on how and why this approach to inference enables one to go far beyond the traditional analyses. Our two claims are the following: *(i)* all that can be done within the frequentist framework may be reinterpreted in a more natural way within the Bayesian one; *(ii)* some problems that arise quite naturally when analyzing data, and for which the frequentist approach does not provide answers, may, on the contrary, be addressed easily by the Bayesian approach.

The other dimension that has, up to now, put a brake on the development and use of Bayesian methods was the technical difficulties involved in Bayesian computations. Nowadays the increase in power of computers and, even more importantly, the emergence of general and efficient algorithms have both contributed to fill the gap between what was theoretically conceivable and what was practically feasible.

This chapter is structured as follows. Section 6.1 deals with the inference on one frequency, that is with binary data, under either an hypergeometric or a binomial sampling model; it will enable us to introduce the key concepts involved in the Bayesian approach and to compare it to the frequentist one. From this point on, we shall focus on Bayesian inference without further attempting to provide a systematic comparison with frequentist inference. The predictive approach to inference, again on one frequency, is presented in Section 6.2. We then give, through concrete and real examples, an insight on how the Bayesian approach can be extended to situations involving several frequencies, first considering simple designs (Section 6.3), and then more complex ones (Section 6.4). The computational aspects, left aside in the first sections, are sketched in Section 6.5. Finally, Section 6.6 summarizes the major points put forward in the chapter.

6.1 From Frequentist to Bayesian Inference: An Illustration for Inference on one Frequency

We shall first consider the case of binary data sampled without replacement from a finite population (hypergeometric sampling). This situation is presumably the simplest one at the technical level, as the mathematics needed merely summarize to basic combinatorial calculus. After stating the problem of inference on one frequency (Section 6.1.1), we shall turn to its frequentist solutions (Section 6.1.2) and then to its Bayesian ones (Section 6.1.3 and Section 6.1.4). These first sections are illustrated by a miniature example (small data set and small population) for which all calculations can easily be carried out by hand. Considering a simple situation and a simple example will enable us to examine all the steps involved in the Bayesian approach at the conceptual level. In Section 6.1.5 we move to sampling from an infinite population, *i.e.* binomial sampling.

We then explore, for the two kinds of sampling models, the links between the frequentist and Bayesian approaches to inference (Section 6.1.6). A detailed analysis of a real example (Section 6.1.7) will provide a practical view on the comparison of the two approaches. Finally in Section 6.1.8, we stress the advantages of the Bayesian approach.

6.1.1 The problem of Inference on one Frequency

Let us consider the following situation: the data are a group of n binary observations, among which a are "successes" and b are "failures", with $a + b = n$. Thus the observed frequency[1] of success

1. Throughout this chapter the word "frequency" will be used for designating a "relative frequency", and the word "count" for an "absolute frequency". The inferential methods considered further will involve probabilistic statements bearing on unknown frequencies. The two concepts of "frequency" and "probability", though formally obeying identical rules, should be carefully distinguished.

is $f_{obs} = a/n$. We assume that the data constitute a sample from a larger population of finite size N, whose composition in terms of successes and failures (A, B) is unknown. The proportion $\phi = A/N$ called the *parent frequency* (also called the *population frequency* or the *true frequency*) is itself unknown and may be any value within the set $\Phi = \{0/N, 1/N, \cdots, N/N\}$.

In this situation, the problem of inference may be stated as follows: "*What may be said about the unknown frequency ϕ on the grounds of the observed data f_{obs} and n?*". The frequency ϕ is thus said to be the *parameter of inference*[2].

Random sampling frame-model. Fundamentally, the problem of inference described previously is a problem of *generalization* from a sample to a population. But in order to proceed to such a generalization, one must be assured (or must assume) that the sample is, in some sense, representative of the population. Random sampling is a privileged means to reach such a representativeness. This is the frame that we shall adopt here: we *assume* (in the sense of Chapter 2) that the group of observations is a random sample of fixed size n without replacement from the population[3].

The random sampling (with fixed size n) assumption, and the characterization of the population by a single real parameter, ϕ, constitute here by themselves the "frame-model" for this situation of inference. In the more complex situations that we shall envisage later in this chapter, there will be several categories instead of only two, so that the population will be characterized by several parent frequencies as parameters. Nevertheless, even then, no extra technical assumptions will be needed on the population.

---

2. In the following, several "objects" should be carefully distinguished: ϕ, the unknown parameter; φ, a value that this parameter may take; φ_0, a reference value of interest for the parameter; and Φ the set of all possible values for parameter ϕ.
3. As we all know, effective random sampling is actually rarely done in practice. An alternative view on this assumption is to think that the generalization goes from the data set at hand to some larger data set that might be observed and that would be composed of data items exchangeable with the available ones (see Section 6.2.3).

An illustrative example: Commitee data. To illustrate the discussion, let us once again consider the example found in Chapter 4 Section 4.1.4: a club (the population) comprises $N = 20$ members; a commitee (the sample) of $n = 5$ members is extracted from the club and is composed of $a = 4$ women and $b = 1$ man, so that the observed frequency of women is $f_{obs} = 4/5$. We call $(a = 4, b = 1)$ the *observed composition in counts* and $(f_{obs} = 0.80, 1 - f_{obs} = 0.20)$ the *observed composition in frequencies*. The corresponding *parent compositions* (A, B) and $(\phi, 1 - \phi)$ are unknown (see Table 6.1).

Table 6.1: Commitee data. Observed and unknown parent compositions in counts.

	Women	Men	Total
Sample	$a = 4$	$b = 1$	$n = 5$
Population	$A = ?$	$B = ?$	$N = 20$

We shall take $\varphi_0 = 0.30$ as a reference value of interest for ϕ, with, as the main goal of the analysis, trying to *generalize the descriptive property* $f_{obs} > 0.30$, *i.e.* to answer the question "May we say that $\phi > 0.30$?".

6.1.2 The Frequentist Solutions

Sampling distribution. We saw in Chapter 4 Section 4.1.4 that, for a given value $\varphi = A/N$ of the parameter ϕ, the proportion of samples of size n for which the statistic frequency F equals the value $f = a/n$, $a \in \{0, \cdots, n\}$, is given by the *hypergeometric distribution*[4]:

$$P(F = f \mid \phi = \varphi) = p_f^{\varphi} = \frac{\binom{A}{a}\binom{B}{b}}{\binom{N}{n}}. \tag{6.1}$$

4. The notation $\binom{A}{a}$ represents the *binomial coefficient* which may be expressed in terms of factorials as: $A!/(a!(A-a)!)$. The symbol '|' reads "if" or "conditionally on".

For example, if the parent composition is ($A = 6$, $B = 14$) *i.e.* $\varphi = 0.30$, there are $\binom{6}{4}\binom{14}{1} = 210$ samples whose composition is ($a = 4$, $b = 1$), or equivalently $f = 4/5$, out of $\binom{20}{5} = 15504$ samples of size 5 in all, so that $p_f^{\varphi} = 210/15504 = 0.0135$.

Under the random sampling frame-model, these proportions of samples convert to sampling probabilities[5]: thus p_f^{φ} represents the probability of observing the frequency value f in a sample of size n when the parent frequency equals φ. The set of such probabilities for all the possible values of the statistic F, $\{0/n, 1/n, \cdots, n/n\}$, is called the sampling distribution of F given that $\phi = \varphi$. The sampling distribution of F for the value $\varphi = 0.30$ is given in Table 6.2 and represented graphically in Figure 6.1 p. 165.

Table 6.2: Sampling distribution of F, $P(F = f \mid \phi = \varphi)$, for $\varphi = 0.30$ ($n = 5$ and $N = 20$).

f	0/5	1/5	2/5	3/5	4/5	5/5
p_f^{φ}	0.1291	0.3874	0.3522	0.1174	0.0135	0.0004

Significance test. In the *significance test* procedure, one considers a particular *hypothesis* about the true frequency ϕ, for example that ϕ equals some reference value of interest φ_0; this hypothesis is generally called a *null hypothesis* and denoted $\mathcal{H}_0$. The aim of the test is to assess whether the observed data are compatible with $\mathcal{H}_0$ or not.

This type of conclusion is reached on the sole basis of the sampling distribution of F given the hypothesis $\mathcal{H}_0 : \phi = \varphi_0$. As may be seen from Figure 6.1 p. 165, if $\phi = 0.30$, then the observable frequency on a sample of size 5 may be any value from 0 to 1 but

5. The notation "$P()$" may be read "(sampling) probability" or "proportion (of samples)" according to whether the random-sampling frame-model has been assumed or not.

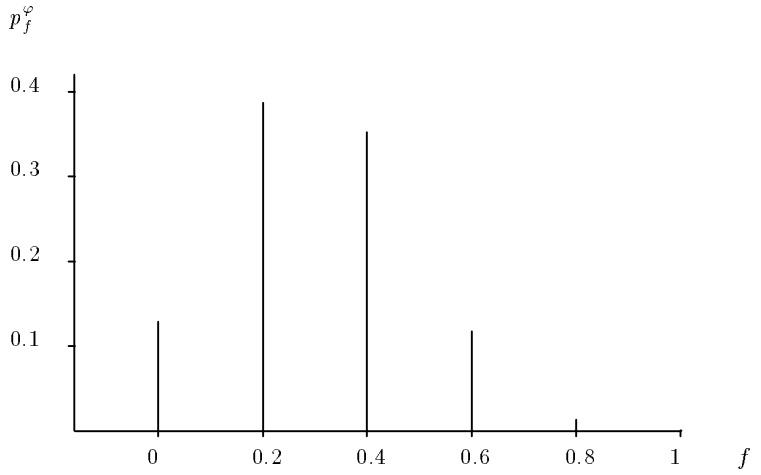

Figure 6.1: Sampling distribution of F, $P(F = f \mid \phi = \varphi)$, for $\varphi = 0.30$ ($n = 5$ and $N = 20$).

will most probably be close to 0.30: indeed, the two most probable values are 0.20 and 0.40 (0.30 cannot be obtained with $n = 5$).

The compatibility of the data with $\mathcal{H}_0$ is measured by the probability, under the hypothesis $\mathcal{H}_0$, of obtaining a frequency F *more extreme* than the observed frequency f_{obs} (including the "as extreme" case); this probability is called the *observed significance level*, or simply the *observed level*, and noted p_{obs}. In our example $f_{obs} = 0.80$ is greater than the reference value 0.30, and is thus extreme *upwise*. In this situation the observed level is the *observed upper level* p_{sup}, *i.e.* the probability of F being equal to or greater than f_{obs} (p_{sup} was denoted $\overline{p}$ in Chapter 4); according to Table 6.2 p. 164, is is thus:

$$p_{sup} = P(F \geq 0.80 \mid \phi = 0.30) \quad = \quad 0.0135 + 0.0004 = 0.0139.$$

If f_{obs} had been lower than the reference value φ_0, it would have been extreme downwise and the observed level p_{obs} would then be defined as the lower level $p_{inf} = P(F \leq 0.80 \mid \phi = 0.30)$.

If the probability p_{sup} is considered sufficiently small, say smaller than a specified one-sided reference level α_{sup}, then ϕ may be declared *significantly greater than* 0.30 at the one-sided α_{sup} level; equivalently we may say that the hypothesis $\phi = 0.30$ is *downwise incompatible* with the data. In our example, for the one-sided level $\alpha_{sup} = 0.025$, we have $p_{sup} = 0.0139 < 0.025$, so that the parent frequency ϕ can be declared significantly greater than 0.30. If we similarly tested any other reference value φ_0 less than 0.30, f_{obs} would be more extreme to the right and consequently the observed level p_{sup} would be even smaller. Thus, for all possible hypotheses such as $\phi = 0.25$, $\phi = 0.20$, *etc.*, we would again reach the conclusion of an incompatibility with the data. Hence the test performed is also a test of the extended hypothesis $\widetilde{\mathcal{H}_0} : \phi \leq 0.30$.

Remark: The observed level p_{obs} is usually defined by including the case $F = f_{obs}$; this is an *inclusive level*. But this choice is a matter of convention and the observed level could also be defined by using the *exclusive convention* $p'_{sup} = P(F > f_{obs} \mid \mathcal{H}_0)$; here we would find $p'_{sup} = 0.0004$. The inclusive convention is known to be conservative whereas the exclusive one is anti-conservative. This is why some authors have also proposed an intermediate solution between these two, the *mid-P convention*, where the observed level is defined as $(p_{sup} + p'_{sup})/2$, that is here 0.0071 (see *e.g.* Berry & Armitage, 1995). For a small sample, as in our example, the point probability $P(F = f_{obs} \mid \mathcal{H}_0)$ may not be negligible, so that the conclusion may be affected by the choice between these conventions; for larger samples, any point probability is negligible so that using one convention or another would be of no practical consequence.

Confidence limits and confidence interval for ϕ. The *confidence interval* for the parent frequency ϕ is built from the previously described test procedure by seeking the set of reference values φ_0 for which f_{obs} is not too extreme, neither upwise nor downwise[6].

6. There are actually other ways of defining confidence intervals. Restricting ourselves to this test-based construction allows us to use "*The* confidence interval" without ambiguity.

Let us consider all the possible values φ_0 for ϕ that are less than f_{obs}. For a given one-sided reference level α_{sup} (less than 0.50 of course), some of these values — the smaller ones — will be downwise incompatible with the data, while some others — the larger ones — will be compatible with them; the smallest compatible value, noted $\underline{\varphi}$, is the *lower confidence limit* for ϕ. Similarly, if we consider all possible reference values greater than f_{obs} and for each compare the observed lower level p_{inf} with the one-sided reference level α_{inf}, we get the *upper confidence limit*, $\overline{\varphi}$. Typically these two limits are defined by taking a fixed two-sided level α_{ts} and identical one-sided levels $\alpha_{sup} = \alpha_{inf} = \alpha_{ts}/2$. The resulting interval $[\underline{\varphi}; \overline{\varphi}]$ is the *confidence interval* for ϕ with *confidence level* (or *guarantee*) $\gamma_{ts} = 1 - \alpha_{ts}$; it is noted $IC_{\gamma_{ts}}$.

For the Commitee data, for $\gamma_{ts} = 0.95$, *i.e.* $\alpha_{ts} = 0.05$, we find:

$$IC_{0.95} \quad = \quad [0.35; 0.95].$$

How should this interval be interpreted? By construction, it is the set of values for ϕ that are compatible with the observed frequency f_{obs} at the confidence level γ_{ts}. From this construction follows the fundamental confidence property satisfied by this interval: *For any value φ, given that $\phi = \varphi$, the (sampling) probability for* $\mathrm{IC}_{\gamma_{\mathrm{ts}}}$ *to contain φ is at least*[7] γ_{ts}:

$$P(IC_{\gamma_{ts}} \ni \varphi \mid \phi = \varphi) \quad \doteq \quad \gamma_{ts}. \tag{6.2}$$

Frequentist probability. In the test of the hypothesis $\mathcal{H}_0 : \phi = 0.30$, each probability involved, p_f^{φ}, may be interpreted as the long-run *frequency* of the occurence of "$F = f$" if one was repeatedly collecting samples of size $n = 5$ from a population of size $N = 20$ with

7. Because the population is of finite size N, the set Φ of the possible values for ϕ is discrete. Due to this, it is not always possible to find limits having exactly some given confidence level γ. This is why we need to use the expression "at least" (and the corresponding symbol '$\doteq$'). This point is not fundamental, "is at least" becoming "is exactly" as N tends to infinity.

fixed $\phi = 0.30$. These probabilities are thus "idealized frequencies" and are said to be *frequentist probabilities*[8].

The confidence level γ_{ts} of an interval is also a frequentist probability: if one was repeatedly collecting samples of size n from the same population of size N and was, for each one, computing the confidence interval $IC_{\gamma_{ts}}$, then, in the long run, at least $\gamma_{ts} \times 100$ percent of these intervals would contain the true value of ϕ.

Quite often though, statistics users will propose the following natural interpretation: "Given the calculated interval $IC_{\gamma_{ts}}$, there is a γ_{ts} probability that the parent frequency ϕ belongs to the interval." "Natural interpretation" because the probability evoked here bears on an unknown quantity, namely the parameter ϕ. The problem is that this alternative interpretation has no meaning within the frequentist statistical framework because the frequentist approach only involves probabilities on observables given the parameter but not on the parameter itself. As we shall see, this interpretation will become possible in the Bayesian statistical framework.

To summarize, the frequentist procedures are exclusively based on the sampling distribution. In this hypothetico-deductive approach, one considers a single (for the test) or several (for the confidence interval) hypotheses concerning the parameter frequency ϕ and one concludes about the compatibility of the observed data with these hypotheses. The assessment of compatibility is done on the sole base of the sampling distribution, which provides probabilities of observable samples given a particular parent frequency considered as fixed. At no moment this approach involves any probability relative to the unknown parameter ϕ.

6.1.3 The Bayesian Approach

Towards a more natural approach of the problem of inference. From an intuitive point of view, the problem of inference

8. The word "frequentist" has nothing to do with the fact that our inference bears on an unknown "frequency"; it only expresses the status of the probabilities involved.

would appear to come up in the opposite way from the one adopted in the frequentist approach: after having observed the data, f_{obs} and n, what may be said about the unknown parent frequency ϕ? Instead of having probabilities relative to all the possible samples from a population with specified parameter $\phi = \varphi$, we would like to have probabilities relative to the unknown ϕ for the unique sample that was actually observed. In other words, the frequentist probabilities go from the unknown ($\phi = \varphi_0$) to the known (the statistic F), whereas natural probabilities would go from what is known (the data characterized by $F = f_{obs}$) to what is not (ϕ). In brief, what we have obtained, up to now, are the p_f^{φ}, *i.e.* the probabilities of all f given some value φ, and what we would like to obtain are some p_{φ}^{f}, *i.e.* some probabilities of all φ given a value f (and particularly the observed one, f_{obs}).

How to obtain inverse probabilities: Bayes' theorem. For obtaining such inverse probabilities, the Bayesian approach consists in posing *prior probabilities* on each value φ that ϕ may take; these prior probabilities are denoted p_{φ}. The idea is that in order to begin calculating the probabilities of the various possible "causes" of what was observed, one must start somewhere with some probabilities of these causes independently of what was observed. The set of the prior probabilities, p_{φ} for all $\varphi \in \Phi$, is called the *prior distribution*, or simply the *prior*, on ϕ.

From the sampling distribution (equation (6.1)) and the prior distribution, one derives a *posterior distribution*, or simply a *posterior*, *i.e.* a set of posterior probabilities p_{φ}^{f}, through the use of *Bayes' theorem*:

$$p_{\varphi}^{f} \quad = \quad \frac{p_{\varphi} p_f^{\varphi}}{\sum_{\varphi \in \Phi} p_{\varphi} p_f^{\varphi}}. \tag{6.3}$$

Each prior probability p_{φ} is updated according to the *likelihood* of f given φ, *i.e.* p_f^{φ}; the denominator on the right-hand side of the above equation is a normalizing constant for making the posterior probabilities add up to one.

The probabilities bearing on the unknown parameter ϕ, whether prior or posterior, will also be denoted with the symbol "*Prob*()". We purposely use a different symbol, *Prob*, for these Bayesian (epistemic) probabilities in order to distinguish them from the frequentist ones, generically noted P. Hence, p_φ and p_φ^f may also appear as $Prob(\phi = \varphi)$ and $Prob(\phi = \varphi \mid F = f)$ respectively.

Bayes' theorem provides probabilities on ϕ for any given value f of F. If we use it for the observed value f_{obs}, we thus have a distribution on ϕ given the observed data (and, of course, the prior distribution).

Uniform prior distribution. The Bayesian approach requires the choice of a prior distribution on ϕ. One simple choice is to consider that all possible values of ϕ have the same prior probability: as there are $(N+1)$ such values, this choice leads to $p_\varphi = 1/(N+1)$. Using the definition of p_f^φ given in (6.1) and applying Bayes' theorem (6.3), we find the posterior probabilities:

$$p_\varphi^f \quad = \quad \frac{\binom{A}{a}\binom{B}{b}}{\binom{N+1}{n+1}}. \tag{6.4}$$

This posterior distribution for the Commitee data is represented in Figure 6.2 p. 171. For the first four lower values of ϕ (0, 0.05, 0.10 and 0.15) the posterior probability is exactly 0; indeed, having observed $a = 4$ successes in the sample implies that the number of successes A in the population is necessarily at least 4, so that we know for sure (logical induction) that $\phi \geq 4/20 = 0.20$. From the value $\varphi = 0.20$, the probabilities are strictly positive and increase to reach a maximum for 0.80, value which is the same as the observed frequency, and then decrease to finally reach 0 again for the last value $\varphi = 1$ which again is logically impossible because the data comprised $b = 1$ failure.

Generalisation to Beta-binomial prior distributions. The uniform distribution is only one particular possibility for the prior distribution. For reasons that will soon become apparent, it is important to investigate the other possible choices. For this purpose,

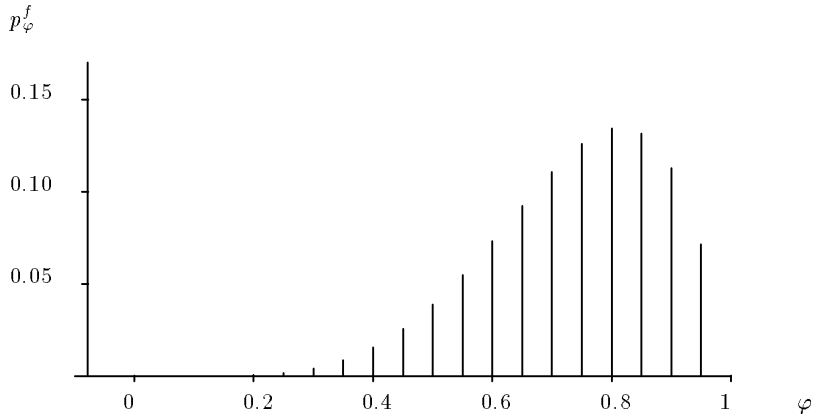

Figure 6.2: Commitee data. Posterior distribution on ϕ derived from a uniform prior ($f_{obs} = 0.80$, $n = 5$, $N = 20$).

we shall now introduce a more general class of distributions, the *Beta-binomial* distributions, containing the uniform distribution as a particular case[9]. A Beta-binomial distribution depends on two positive real "hyperparameters" α and β[10], whose sum is denoted $\nu = \alpha + \beta$; such a distribution will be noted $BeBi(\alpha, \beta; N)$. Its general expression is[11]:

$$p_\varphi = \frac{\binom{A+\alpha-1}{A}\binom{B+\beta-1}{B}}{\binom{N+\nu-1}{N}} \quad \text{with } \varphi = \frac{A}{N} = 1 - \frac{B}{N}. \qquad (6.5)$$

If the preceding distribution $BeBi(\alpha, \beta; N)$ is taken as a prior on (A, B) with fixed $A + B = N$, then the posterior distribution,

9. This distribution appears in Mosimann (1962) as the *compound binomial* and was later referred to as the *Beta-binomial* by Hoadley (1969) and most recent authors. In Bernard (1983) they are called *discrete-Beta*.
10. These hyperparameters should not be mistaken for the parameter of inference ϕ, nor for significance levels that were denoted with subscripts α_{ts}, α_{sup} or α_{inf}.
11. Our notation follows the usual extension of binomial coefficients to non-integers through the use of the Gamma function: $\binom{A+\alpha-1}{A} = \Gamma(A+\alpha)/(\Gamma(\alpha)A!)$.

after observing the counts (a, b) with fixed $a + b = n$, is still of the same type, but with transformed characteristics[12]: it is a $BeBi(a + \alpha, b + \beta; N - n)$. A first change expresses the knowledge brought by the data: the two prior hyperparameters (α, β) are incremented by the observed counts (a, b) to become $(\alpha' = a+\alpha, \beta' = b+\beta)$ with $\nu' = n+\nu$. On the other side, what remains to be known about the population reduces so that the unknown composition (A, B) is now replaced by $(A' = A-a, B' = B-b)$ with $N' = N-n$. If, in the prior equation (6.5), we substitute A, B, N, α, β and ν with their respective "primed" value, *i.e.* A', B', *etc.*, we get the posterior distribution given below:

$$p_\varphi^f = \frac{\binom{A'+\alpha'-1}{A'}\binom{B'+\beta'-1}{B'}}{\binom{N'+\nu'-1}{N'}} = \frac{\binom{A+\alpha-1}{A-a}\binom{B+\beta-1}{B-b}}{\binom{N+\nu-1}{N-n}}. \qquad (6.6)$$

Actually, the two hyperparameters α and β may be considered as *prior strengths* put on each of the two categories, that are combined additively with the observed strengths a and b provided by the data. Figure 6.3 p. 173 summarizes this updating process.

The uniform prior distribution for which we have first illustrated Bayes' theorem is a Beta-binomial $BeBi(\alpha = 1, \beta = 1; N)$. Figure 6.4 p. 174 shows the application of Bayes' theorem to the Commitee data with a non-uniform though symmetrical prior $BeBi(2, 2; 20)$. As may be seen from the comparison of the prior and the posterior, the data have intervened in two ways:

(i) As already noticed, the observed composition $(a = 4, b = 1)$ logically excludes some of the initially possible values for ϕ: ϕ is necessarily within the interval $[4/20; (20-1)/20]$. So only values in this interval have a non-null posterior probability.

12. This property is why Beta-binomial distributions are privileged priors for hypergeometric sampling: the Beta-binomial family is said to be a *conjugate family* for this sampling scheme.

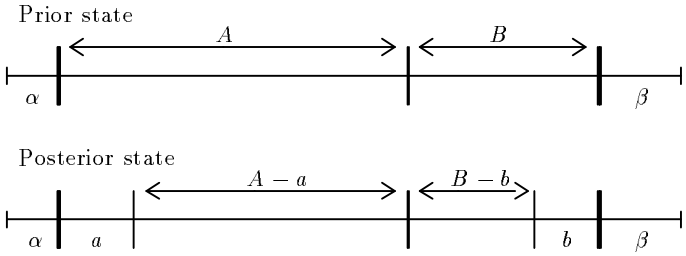

Figure 6.3: Prior and posterior states of knowledge on ϕ expressed by a Beta-binomial distribution: the quantities appearing above belong to the "unknown" (A and B in the prior state, $A - a$ and $B - b$ in the posterior state), those appearing below belong to the "known" (prior knowledge, α and β, plus data, a and b, in the posterior state).

(ii) Moreover, the probabilities of the remaining possible values for ϕ have been updated in the direction of a favouring of values that are close to f_{obs}; for example, $Prob(\phi = 0.70)$ increases from 0.059 to 0.129 whereas $Prob(\phi = 0.20)$ decreases from 0.048 to less than 0.001.

Bayes' theorem as a learning model. Fundamentally Bayes' theorem is a *probabilistic learning model* that may be schematized by the diagram below. What is known initially, before any observation, is expressed by the prior distribution; this prior state of knowledge is updated by the data into a posterior state, itself expressed by the posterior distribution[13].

Prior on ϕ	+	Data f_{obs} and n	$\longrightarrow$	Posterior on ϕ given the data

From the statistical point of view that we only focus on here, this learning model interpretation also clearly explains why the

13. This view of Bayes' theorem explains why it is now often proposed as a model for knowledge representation and updating in the fields of cognitive psychology and artificial intelligence (see *e.g.* Anderson, 1991; Walley, 1996b).

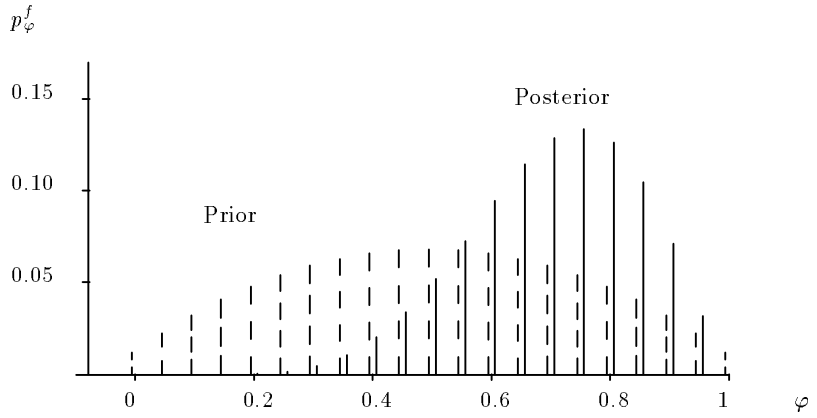

Figure 6.4: Commitee data. Prior (dashes) and posterior (solid) distributions; the prior is the non-uniform distribution $BeBi(\alpha = 2, \beta = 2; N = 20)$.

probabilities are of an *epistemic* nature. The probabilities involved in the Bayesian approach are *relative to some particular state of knowledge*; they are not the expression of some intrinsic property of some object or device of the outside world, but they describe one's uncertainty about reality. And of course, this uncertainty varies with the information available[14]. In this line, Bayesian inference should also be seen as fundamentally recursive: after some data have been observed the posterior state of knowledge becomes the new prior state to be used with some future data, and so on.

Respective roles of the prior and the data upon the posterior distribution. With the preceding view, the posterior distribution appears as the combination of two components, namely the prior and the data. Let us look into how these two components respectively affect the posterior distribution. Schematically two situations may occur:

14. This is why we use the term "probability *on* ϕ" rather than "probability *of* ϕ".

(i) The data are in *agreement* with the prior distribution in the sense that the observed frequency f_{obs} corresponds to a region with high prior probabilities. The peak of the posterior distribution then corresponds to a value close to the peek of the prior (and close to f_{obs} as well). The change from prior to posterior will then be a decrease in variance. Using the "learning model" viewpoint, the new information provided by the data goes in the same direction as the initial state of knowledge; thus this state of knowledge is reinforced, uncertainty diminishes and the resulting final state is more precise. Of course, the larger the data size n is, the more important is the gained precision, *i.e.* the decrease in variance. This first case is illustrated in Figure 6.5. [In Figures 6.5 and 6.6, we assume a population of size $N = 100$.]

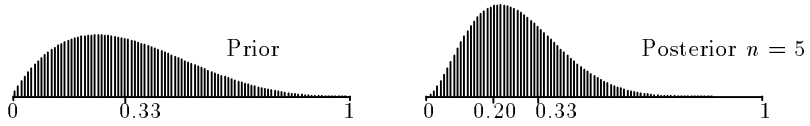

Figure 6.5: Posterior distribution (right) for ϕ when the data of size $n = 5$ are in agreement ($f_{obs} = 0.20$) with the prior $BeBi(\alpha = 2, \beta = 4; N = 100)$ (left).

(ii) The data are in *disagreement* with the prior in the sense that f_{obs} falls in a low prior probability region. In this case the updating process comprises a translation of the center of the distribution towards the observed frequency f_{obs}. If the data are not too numerous, only this center shifting occurs; but with larger data size, the decrease in variance previously described will also occur. This is illustrated in Figure 6.6 p. 176.

This attraction power of the data, whether it involves a center shifting or a variance decrease, depends on the ratio of the data's strength, *i.e.* their size n, to the *total prior strength*, $\nu = \alpha + \beta$. If ν is small, which represents little prior knowledge, then the data's

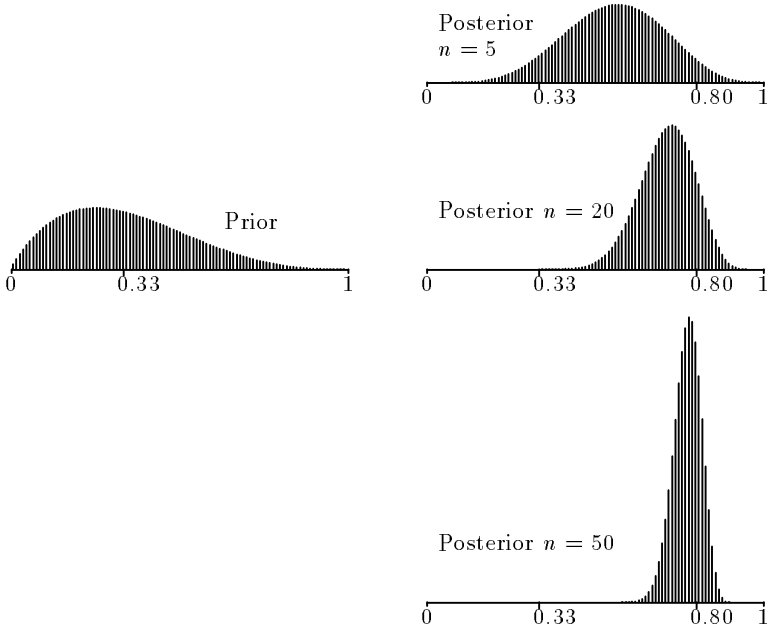

Figure 6.6: Posterior distributions (right) for ϕ when data of size $n = 5$, $n = 20$ and $n = 50$ are in disagreement ($f_{obs} = 0.80$) with the prior $BeBi(\alpha = 2, \beta = 4; N = 100)$ (left).

attractive power will be quite perceptible even for small data size n. On the other hand, if ν is large, which represents substantial prior knowledge, n must be large enough for the posterior to be noticeably affected. Intuitively speaking, it is more difficult to get someone to change his mind if his ideas are rather firm to start with. This interpretation in terms of "strengths ratio" is a guide for the question of the prior's choice that we discuss next.

Two approaches for choosing the prior distribution. According to which criteria should the prior distribution be chosen? The answer to this question actually depends on the goal which is assigned to the analysis of the data; very roughly we may categorize goals into one of the two following ones.

If the analysis must lead to decision making, one should take into account all other available relevant pieces of information (previous data, expert knowledge) besides the data themselves. These would then be incorporated into the prior distribution and would then partly affect the analysis. With such a goal, the prior strength ν may not be small relative to the data strength n[15]. In addition, such a *decisionist approach* typically involves the taking into account of the respective costs (or utilities) of each possible decision. Clearly this approach may involve several elements of subjectivity (expert knowledge, costs) and is mostly advocated by what Rouanet calls "radical Bayesians" (see Chapter 1, Appendix 1). For a recent account of this approach, see Bernardo & Smith (1994).

The other approach that we shall adopt in the sequel of this Chapter is what we call the "*data analysis methodology*". It corresponds to a situation in which no prior information is available, or in which, if some is, one does not want to take it into account in the analysis. Here the idea is to choose a prior which expresses a *prior state of ignorance* on the parameter. The resulting posterior distribution will then express "what the data have to say" about the unknown parameter, independently of any external knowledge. In the modern era of statistical inference, this "moderate Bayesian" viewpoint is largely associated with the name of Jeffreys (1938/1961)[16], in particular with the use of Jeffreys's (1946) rule for choosing prior probabilities[17].

15. For expressing substantive prior knowledge, a broader class of prior distributions might be needed. A simple and general solution is to use a weighted mixture of Beta-binomial distributions instead of a single one; the posterior obtained then is also a weighted mixture of Beta-binomial distributions.
16. Though less often referenced in Bayesian literature, Jaynes has also been a strong defender of this view (*e.g.* Jaynes, 1968).
17. Bernardo & Smith (1994, p. 68) write that "... the problem of reporting inferences is essentially a special case of a decision problem"; these authors nevertheless dedicate a rather long chapter to "Inference" including a section on "Reference analysis" where they propose "reference posterior distributions" (which often agree with Jeffreys' proposals) to fulfill the need for an inductive data analysis methodology.

Formalizing ignorance: towards standard distributions. Of course the goal of formalizing ignorance is to provide, if not an objective Bayesian method, at least a *reference method for public usage*. Much work and debate have been motivated by such a purpose, because formalizing ignorance has not proved to be as a simple matter as it might sound (for a recent review, see *e.g.* Kass & Wasserman, 1996).

For the inference on a frequency in the case of a finite population, the most common solution — initially proposed by Bayes himself and later by Laplace (1825/1986 p. 45) who justified it by the so-called *principle of insufficient reason* — consists of choosing a uniform prior $BeBi(\alpha = 1, \beta = 1; N)$: knowing nothing about ϕ is operationalized by assigning equal prior probabilities to each possible value for ϕ. Clearly the uniform solution satisfies two intuitive principles that an ignorance prior should obey: *(i)* there must be some kind of symmetry between the two categories, so that the prior strengths α and β must be of the same order of magnitude, *(ii)* the total prior strength must be small for the strength ratio to be in favour of the data. However, coherence with what was later proposed for the case of an infinite population suggests the possible use of some other priors than the uniform obeying these principles. We shall leave the discussion of that point to Section 6.1.5. Here, we shall only give the two resulting ideas.

The first idea is that, when considering all proposed priors, it is possible to define an *ignorance zone*, corresponding to prior strengths constrained by: $\alpha \geq 0$, $\beta \geq 0$ and $\nu = 1$. The difference between the two extreme points of this ignorance zone, namely $(\alpha = 1, \beta = 0)$ and $(\alpha = 0, \beta = 1)$, actually reduces to a change from one success to one failure. Thus, as soon as the sample size n is large enough, the various solutions provided by the entire ignorance zone will lead to very close results.

The second idea is that, for practical purposes, we have been led to suggest one particular prior within the ignorance zone, the *standard prior* obtained for $\alpha = \beta = 1/2$. We call the resulting posterior the *standard posterior distribution* or simply the *standard*

distribution. When using this standard prior, we shall refer to *Bayesian standard methods*.

From these two points, the overall strategy that we shall adopt for subsequent analyses in this chapter is the following. First, we shall always provide standard probabilistic statements, *i.e.* ones that are derived from the standard posterior distribution; such statements will be of the form "$Prob^{\star}(property) = \gamma$". Secondly, varying the prior within the ignorance zone provides the means to determine the sensitivity of the Bayesian results to the choice of the prior. When this is done, Bayesian probabilistic statements end up in a probability interval rather than in a single probability value; these complementary statements will be of the form "$Prob^{\star}(property) = [\gamma_1, \gamma_2]$". Often we shall summarize these two kinds of statements by omitting the upper guarantee γ_2 (less essential than γ_1 for the purpose of generalizing a property of interest), so as to obtain the more compact statement,

$$Prob^{\star}(property) \quad = \quad \gamma \ (\geq \gamma_1) \tag{6.7}$$

which should be understood as: "The probability of *property* is γ for a standard prior, and in any case greater than γ_1 for whichever prior taken in the ignorance zone"[18]. Notice that the '$\star$' superscript in "$Prob^{\star}$" indicates two things at once: first that the probabilistic statement is derived from the posterior distribution (thus the conditioning on the observed data becomes implicit) and second, that we are using an ignorance prior (either the standard prior or the ignorance zone).

As will become apparent on examples considered further, for large data sets, the several probabilities provided by the entire ignorance zone will all be very close to one another so that the single standard probability is enough to summarize them all. For small samples though, the suggested sensitivity analysis might lead

18. If one wishes to produce statements involving only one guarantee value, the most cautious (conservative) solution consists in stating: $Prob^{\star}(property) \geq \gamma_1$.

to wide probability intervals. If the inferential conclusion appears to be too easily affected by the prior's choice, a sensible conclusion is that "we do not know much more about the parameter after taking the data into account than before doing so".

6.1.4 Bayesian Answers to the Inference Problem

We have now defined a general Bayesian framework for inductive data analysis: from a prior distribution expressing an initial state of ignorance about the parameter, and the observed data, one derives a posterior distribution on the parameter which expresses probabilistically what the data have to say about the parameter.

This construction will be completed by stating that, *in the Bayesian framework, the posterior distribution is the exhaustive summary of the inductive analysis from which answers to any question pertaining to the parameter will be drawn*. Each possible question can be translated into a *property of interest* that the parameter may have, typically stating that the parameter is within some restricted region among the set of its possible values; the Bayesian answer will then be the posterior probability of the parameter belonging to that region.

Of course, the Bayesian answers that we examine next are not bound to the use of a particular prior, but, from now on, we shall restrict our attention to answers obtained with ignorance priors; thus speaking of "the" posterior or "the" probability will refer to unambiguous statements, as far as we allow for the restricted undetermination induced by the ignorance zone.

Bayesian test of an extended hypothesis. Let us turn back to our initial question of comparing ϕ to the reference value $\varphi_0 = 0.30$. Within the Bayesian framework, testing the extended hypothesis $\widetilde{\mathcal{H}_0} : \phi \leq \varphi_0$ just amounts to calculating the posterior probability of the hypothesis, which is here found to be:

$$Prob^{\star}(\phi \leq 0.30) \quad = \quad 0.0049.$$

This probability is the Bayesian counterpart of the frequentist observed level p_{sup} for $\widetilde{\mathcal{H}_0}$ (in Section 6.1.2, we found 0.0139, 0.0004 and 0.0071 for p_{sup} depending on the choice between the inclusive, the exclusive or the mid-P conventions). As it is sufficiently small here, say less than the one-sided level $\alpha_{sup} = 0.025$, we are in a position to reject $\widetilde{\mathcal{H}_0}$ at level α_{sup}. The conclusion reached here sounds similar to a conclusion obtained from a frequentist test, but the "rejection" involved here is of a quite different nature. In the frequentist framework the line of reasoning was: "If the hypothesis were true, then the observed data would be highly unprobable; hence the hypothesis must be false". Whereas here it goes: "given the observed data, the hypothesis is highly unprobable".

Instead of considering an hypothesis counter to what we observed in the data (as commonly done in frequentist methods), we may envisage the problem more directly and consider the hypothesis $\phi > 0.30$ which generalizes the observed property $f_{obs} > 0.30$. This is straightforward in the Bayesian framework: we previously found $Prob^\star(\phi \leq 0.30) = 0.0049$, so that, by considering the complementary property, we get

$$Prob^\star(\phi > 0.30) \quad = \quad 1 - 0.0049 = 0.9951.$$

This probability is close to 1, say greater than the one-sided guarantee $\gamma_{sup} = 0.975$, and we may thus conclude that the data are in favour of the hypothesis $\phi > 0.30$ at the guarantee γ_{sup}.

We have just illustrated here a quite general feature of the Bayesian approach, the fact that it enables one to reason directly about the hypothesis which generalizes the descriptive conclusion. Because of this, it is more common within the Bayesian framework to provide statements with a high guarantee (with respect to a reference guarantee) rather than reverse statements having a low level (with respect to a reference level).

If we now consider the ignorance zone as a whole and not the standard prior only, we find $Prob^\star(\phi \leq 0.30) = [0.0016, 0.0139]$ and $Prob^\star(\phi > 0.30) = [0.9861, 0.9984]$. Even though the sample size is particularly small here, $n = 5$, both intervals of probabilities are

seen to be rather narrow. In any case we may conclude that the probability of ϕ being greater than 0.30 is higher than 0.9861.

Credibility limits and credibility interval. There are also Bayesian counterparts of the confidence limits and of the confidence interval, respectively called *credibility limits* and *credibility interval*.

For a given upper reference level α_{sup}, or equivalently a given upper guarantee $\gamma_{sup} = (1 - \alpha_{sup})$, the *lower credibility limit* for ϕ, noted $\underline{\varphi}$, is defined as the largest value for ϕ such that $Prob^{\star}(\phi \geq \underline{\varphi}) \doteq \gamma_{sup}$. In the Commitee data example, for $\gamma_{sup} = 0.975$, we have $Prob^{\star}(\phi \geq 0.45) = 0.9756$ and $Prob^{\star}(\phi \geq 0.50) = 0.9552$, so that we find $\underline{\varphi} = 0.45$. The *upper credibility limit* at the lower guarantee γ_{inf} is defined in a similar way as the smallest value for ϕ such that $Prob^{\star}(\phi \leq \overline{\varphi}) \doteq \gamma_{inf}$; here we find $\overline{\varphi} = 0.95$.

For identical upper and lower guarantees $\gamma_{inf} = \gamma_{sup} = (1 + \gamma_{ts})/2$, the two values $\underline{\varphi}$ and $\overline{\varphi}$ define the *credibility interval*[19] for ϕ at the two-sided guarantee $\gamma_{ts} = 0.95$:

$$ICR_{0.95} \quad = \quad [0.45; 0.95].$$

By construction the credibility interval $ICR_{\gamma_{ts}}$ has the *fundamental credibility property*, *i.e.*: *The probability for* ϕ *belonging to* $\mathrm{ICR}_{\gamma_{\mathrm{ts}}}$ *is at least* γ_{ts}*:* $Prob^{\star}(\phi \in [\underline{\varphi}; \overline{\varphi}]) \doteq \gamma_{ts}$.

It should be noted that both the test of an extended hypothesis and the credibility interval procedure end up in statements of the same nature, $Prob^{\star}(property) = guarantee$. The only difference is that, in the former, the property is given and the guarantee to be computed, whereas, in the latter, the guarantee is given and the property has to be found (within a restricted class of properties). As we announced, the answer to any question pertaining to the parameter (or parameters in more complex problems) will be based upon statements of this kind, where the *property* will correspond to the appropriate region of the parameter's space.

19. More acurately this is *the* symmetrical credibility interval. Other intervals could be defined with differing lower and upper reference levels.

Remark: Within the framework just presented, it is important to notice that "testing a point null hypothesis", if taken as meaning "assessing the probability of the hypothesis", is void of interest. Indeed, the posterior probability $Prob^{\star}(\phi = \varphi_0)$ of *any* point hypothesis is very small provided that N is large (it tends to 0 as N increases), so that only statements involving a set of values for ϕ may possibly lead to a sufficiently high guarantee.

6.1.5 Binomial Sampling Model (N Infinite)

Up to here, we have focused on inference about a frequency from a sample from a finite population. This simple case has given us the opportunity to set up the various components of the Bayesian approach. But quite often, the population size is not specified and may be considered very large relative to the sample size. This situation corresponds to the case of an *arbitrarily large* population, *i.e.* technically an infinite N. This case has received more attention and results relative to it are more "classical" than the preceding ones (see *e.g.* Lindley, 1965; Lindley & Phillips, 1976) even though, in our opinion, they still remain too rarely applied. However, both cases are closely related, since most results for the infinite case may be obtained as limiting ones from the finite case, with N tending towards infinity.

When N is infinite, the population is only characterized by its composition in frequencies, *i.e.* by the unknown parent frequency ϕ. The sampling distribution, the p_f^{φ}, now becomes a *binomial distribution* (limiting form of the hypergeometric when N tends towards infinity). The prior and posterior distributions are now members of the family of *Beta distributions* (limiting form of the Beta-binomial family when N tends towards infinity). The major change is that these distributions are *continuous*. To each possible value φ of ϕ is now associated a *probability density*: any single value of ϕ has a null probability and only intervals for ϕ have non-null probabilities.

All of the methods previously described, whether frequentist or Bayesian, can be extended to the infinite case. However, we shall not attempt to give a detailed parallel presentation, but rather focus only on the Bayesian approach.

Prior distribution: Ignorance zone and standard prior. As in the finite case, everything in the Bayesian approach can be expressed in terms of prior and posterior strengths. A Beta prior is characterized by two strengths α and β respectively attached to the "success" and "failure" categories; we note such a prior $Beta(\alpha, \beta)$.

The choice $\alpha = \beta = 1$ leads to a distribution with a uniform density. Several other proposals can be found in the litterature depending on which specific criteria are used to formalize ignorance: Haldane (1948) proposed $\alpha = \beta = 0$ (this should read very close to 0 since the Beta distribution is not defined for null strengths); Jeffreys (1946, 1938/1961) and Perks (1947) suggested $\alpha = \beta = 1/2$.

It appears that all suggested solutions correspond to prior strengths both between 0 and 1. Moreover it can be shown (see *e.g.* Bernard, 1996) that, for any prior in this set, the probability of any one-sided hypothesis of the type $\phi > \varphi_0$ is always in-between the probabilities obtained with the two following extreme priors: $(\alpha = 0, \beta = 1)$ and $(\alpha = 1, \beta = 0)$. This result led us to suggest the idea of an *ignorance zone* defined as:

$$\alpha \geq 0, \quad \beta \geq 0, \qquad \nu = \alpha + \beta = 1. \tag{6.8}$$

A similar suggestion is also made by Walley (1991, 1996a) under the name of the "imprecise Beta model". Formalizing prior ignorance by this ignorance zone leads to a probability interval for each property of interest relative to the parameter. The lower and upper probabilities of this interval are interpreted by Walley as acceptable betting rates for and against the property.

Within the ignorance zone, it is convenient to have one single standard reference prior; the mid-point of the ignorance zone ($\alpha = \beta = 1/2$), which coincides with Jeffreys' and Perks' priors, appears as a good compromise between all proposed priors. We shall soon see

(Section 6.1.6) that there are strong connections between Bayesian methods using this ignorance zone and frequentist methods, even as far as the degree of undetermination is concerned.

Posterior distribution on ϕ. If the prior on ϕ is $Beta(\alpha, \beta)$ and the data's composition in counts is (a, b), the posterior distribution is $Beta(a+\alpha, b+\beta)$. The mean and variance of the posterior distribution are given below:

$$Mean(\phi) = \frac{a+\alpha}{n+\nu}, \tag{6.9}$$

$$\mathrm{Var}(\phi) = \frac{(a+\alpha)(b+\beta)}{(n+\nu)^2(n+\nu+1)}. \tag{6.10}$$

Notice that this mean is the ratio of the posterior "success" strength $(a+\alpha)$ to the posterior total strength $(n+\nu)$; in particular the standard distribution $(\alpha = \beta = 1/2)$ is centered on $\frac{a+1/2}{n+1}$. Furthermore, using any prior within the ignorance zone, the posterior mean is always comprised between $\frac{a}{n+1}$ and $\frac{a+1}{n+1}$; it can be seen that, when n is not too small, both values will be very close to the observed frequency $f_{obs} = a/n$.

6.1.6 Bayesian Reinterpretation of Frequentist Procedures

For the problem of inference with which we started, "What can be said about ϕ from the data f_{obs} and n?", it appears that we now have two sets of answers: the first obtained by using frequentist methods (significance test and confidence interval), and the second by using the Bayesian framework (Bayesian test and credibility interval). However, these answers are different on two levels: first, and most of all, they respectively refer to two different statistical frameworks in which the probabilities — on which the conclusions are based — are of a different nature; second, they lead to different, though close, numerical results.

We shall now see that both frameworks are closely related and, more precisely, that it is possible to reinterpret Bayesianly the frequentist procedures by an appropriate choice of the prior strengths.

Tests of an extended hypothesis. Let us again consider the frequentist test of the extended hypothesis $\phi \leq \varphi_0$ in the case $f_{obs} > \varphi_0$. As we remarked in Section 6.1.2, the observed level can be defined in (at least) two ways, inclusive or exclusive, according to whether the probability of the observed value is included or not. These two observed levels were respectively defined as:

$$p_{obs} = p_{sup} = P(F \geq f_{obs} \mid \phi = \varphi_0), \tag{6.11}$$

$$p'_{obs} = p'_{sup} = P(F > f_{obs} \mid \phi = \varphi_0). \tag{6.12}$$

For the Commitee data and $\varphi_0 = 0.30$ we found: $p_{obs} = 0.0139$ and $p'_{obs} = 0.004$.

Within the Bayesian framework we also introduced a Bayesian test of the hypothesis $\phi \leq \varphi_0$, based on the posterior distribution, that is a distribution conditionnal on the data and the two prior strengths α and β. Here also, we may consider two variants of this test whether the probability of the value φ_0 is included or not; their respective levels are:

$$Prob(\phi \leq \varphi_0 \mid F = f_{obs};\ \alpha, \beta), \tag{6.13}$$

$$Prob(\phi < \varphi_0 \mid F = f_{obs};\ \alpha, \beta). \tag{6.14}$$

The symmetry with the frequentist approach is only apparent as the two probabilities (6.13) and (6.14) are equal when N is considered infinite (because then $Prob(\phi = \varphi_0) = 0$); on the other hand the observed levels (6.11) and (6.12) are never equal (this would require an infinite sample size n).

With a standard prior, $\alpha = \beta = 1/2$, and the inclusive variant, we found a Bayesian level of 0.0049 for the Commitee data, close to the two frequentist levels. Due to what we call "Guilbaud's magical

hypergeometric identity"[20], it is actually possible to choose α and β so that both the frequentist levels and the Bayesian levels are in perfect agreement:

$$P(F \geq f_{obs} \mid \phi = \varphi_0) = Prob(\phi \leq \varphi_0 \mid F = f_{obs}; 0, 1) \quad (6.15)$$

$$P(F > f_{obs} \mid \phi = \varphi_0) = Prob(\phi < \varphi_0 \mid F = f_{obs}; 1, 0) \quad (6.16)$$

As previously noted, the difference between the two Bayesian variants (6.13) and (6.14) tends to vanish as N becomes large. The remaining and more important differences between equations (6.15) and (6.16) are that *the frequentist inclusive level corresponds to prior strengths* ($\alpha = 0, \beta = 1$) and *the exclusive frequentist level to prior strengths* ($\alpha = 1, \beta = 0$).

It can be seen that the two priors that provide this frequentist-Bayesian mutual reinterpretation are the two extreme priors of the ignorance zone. From a Bayesian point of view, each of the two frequentist tests appears slightly biased, as the inclusive one is favouring the "failure" category whereas the exclusive one is favouring the "success" category. In this regard, the "mid-P" convention (taking $(p_{obs} + p'_{obs})/2$ as the observed level) appears to be more balanced; this is quite similar to the argument that led us, on the Bayesian side, to the standard prior ($\alpha = \beta = 1/2$). Indeed these two kinds of "compromises" will generally lead to almost the same numerical value.

One major consequence of this link between the two approaches is that, if one feels puzzled about the undetermination in the ignorance zone in the Bayesian framework, one should feel so too about the choice between the inclusive and exclusive conventions in the frequentist framework. There is no more arbitrariness in the Bayesian data analysis approach than in the frequentist approach. Furthermore, the common undetermination in both approaches is

20. This identity may be found for example in Lieberman & Owen (1961, p. 19); its importance for Bayesian inference has been pointed out by Guilbaud during several seminars around 1980 (Guilbaud, 1983; see also Rouanet, Bernard, Le Roux, 1990, p. 227).

small and will be of no practical consequence if the sample size n is not too small.

This link was explored in more detail in Bernard (1996); we also showed that the ignorance zone provides bounds for frequentist levels obtained from several other sampling models[21].

Confidence and credibility. From what precedes follows a similar link between confidence intervals and crebility intervals. If $IC_\gamma = [\underline{\varphi}, \overline{\varphi}]$ is a γ confidence interval for ϕ, then it is also approximately a γ standard credibility interval, that is:

$$Prob^\star(\phi \in IC_\gamma) \quad \approx \quad \gamma. \tag{6.17}$$

Of course this approximate identity holds for any procedure used for defining confidence intervals, whether they are based upon the inclusive, mid-P, or exclusive test, but the numerical closeness will be much greater when comparing the mid-P level based confidence interval with the standard credibility interval.

There are similar approximate identities for a variety of other *elementary* cases. But, as Rouanet pointed out in Chapter 2, such identities do not hold for many other cases, either because confidence intervals are not available or because they present undesirable properties. Investigating in detail several common interval estimation problems, Jaynes (1976) summarizes his paper by saying: "... the Bayesian method is easier to apply and yields the same or better results. Indeed, the orthodox results are satisfactory only when they agree closely (or exactly) with the Bayesian results."

21. One conceptual difficulty of the frequentist approach is the dependence of p_{obs} on the sample space, and particularly on the stopping rule, since the observed data are compared to all other data sets that *might have occurred*. Hence the same data will be analyzed differently whether it is considered that the sampling process stopped because size n was reached (binomial sampling) or because a successes (or b failures) were reached (negative-binomial sampling). On the contrary, the ignorance zone idea is free from any sampling scheme assumption. This last point is made very clear in Walley (1996a).

Fiducial approach. Bayesian inference with ignorance priors is very close in spirit to Fisher's fiducial approach: these both aim at providing a distribution on the parameter which only expresses the information brought by the data. However, the fiducial idea consists in trying to reach this goal without the recourse to a prior distribution. Unfortunately, due to the discreteness of the distributions encountered in the case of categorized data, one can only derive an asymptotic fiducial distribution for ϕ, which, not so suprisingly, coincides with our suggested standard posterior distribution (Fisher, 1956/1959, pp. 60–65).

Thus, whereas in Chapter 5 the expressions "fiducial", "Bayes-fiducial" or "standard Bayesian" could be used indifferently, here, within the context of categorized data, we shall only use "standard Bayesian", since the word "fiducial" may only refer to the motivation but not to the technique.

General comments. Let us conclude this sub-section with two remarks. First, when looked at from a Bayesian viewpoint, the frequentist methods correspond to an ignorance prior state. This is the expression of a more general result: in elementary problems there is a privileged link between frequentist inference and data analysis minded Bayesian inference, of which we already had an example when considering inference on means in Chapter 5. Second, equations (6.15), (6.16) and (6.17) each provide a single numerical result with two possible statistical interpretations: in particular we think that the Bayesian interpretations of the observed significance level and of the confidence interval are both more natural than their frequentist counterparts, since they provide probabilistic statements about the unknown parameter ϕ.

6.1.7 An Example: Mendel's Peas (Shape)

We shall consider one of Mendel's experiments on the genetic transmission of pea characteristics, already mentioned in Chapter 2 Section 2.1.1. Here we only consider the "shape" attribute of the

peas, but we shall turn back to the full data, involving "shape" and "colour", in Section 6.3.4.

The prediction of the Mendelian model is that, by crossing two pure breeds of peas, respectively round and wrinkled, the second generation will provide on average 3 round peas for 1 wrinkled one, *i.e.* 3 out of 4. The composition in counts of 556 observed peas is $(a = 423, b = 133)$, so that the observed frequency of "round" is $f_{obs} = 423/556 = 0.761$, a value quite close to the reference value predicted by the Mendelian model $\varphi_0 = 3/4 = 0.75$.

First of all we may perform the usual significance test of the null hypothesis $\mathcal{H}_0$: $\phi = 0.75$. Using the binomial inclusive test, we find $p_{obs} = 0.297$; with the usual Chi-square approximate test corrected for continuity, we find $p_{obs} \approx 0.295$, a value which agrees quite closely with the exact result. Using the mid-P convention, we would find respectively 0.280 (exact) and 0.278 (approximate). Whichever way we perform the test, we reach the conclusion that the departure of the data from the Mendelian model is not significant at any usual reference level ($p_{obs} > 0.10$).

Figure 6.7 p. 191 gives the standard posterior distribution of ϕ, which corresponds to the posterior strengths $a + \alpha = 423 + 1/2$ and $b + \beta = 133 + 1/2$. The mean of this distribution is 0.760, value very close to f_{obs}; the relatively low dispersion of this distribution results from the large sample size ($n = 556$).

If, using this distribution, we proceed to the Bayesian test of the extended hypothesis $\phi \leq 0.75$, we find:

$$Prob^{\star}(\phi \leq 0.75) \quad = \quad 0.280.$$

Here we see that the mid-P test and the standard Bayesian test both lead to the same numerical result up to the third decimal place. Using the ignorance zone, we find $Prob^{\star}(\phi \leq 0.75) = [0.264, 0.297]$, the lower probability corresponding to the frequentist exclusive test (from equation (6.16)) and the upper one to the inclusive one (from equation (6.15)).

Besides the numerical equivalence between frequentist and Bayesian tests, the Bayesian interpretation throws a light on the

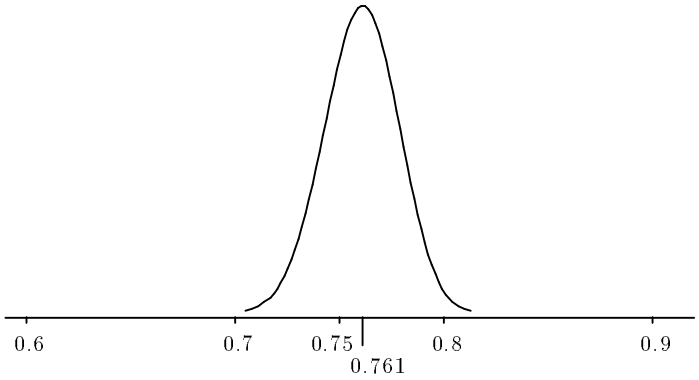

Figure 6.7: Mendel's data. Standard distribution on ϕ ($f_{obs} = 0.761$, $n = 556$).

true meaning of a "non-significant" (NS) result. The departure of $f_{obs} = 0.761$ from $\varphi_0 = 0.75$ has been qualified "NS" because the observed level (say the mid-P one) $p_{obs} = 0.280$ was not small enough. From the Bayesian viewpoint, this is equivalent to saying that $Prob^\star(\phi \leq \varphi_0)$ is not small enough, but also, as a consequence, that $Prob^\star(\phi > \varphi_0) = 1 - 0.280 = 0.720$ is not large enough. In other words, concluding "NS" implies that none of the two statements about ϕ have a sufficiently high guarantee: by itself a "NS" conclusion is just a report of an insufficient knowledge about ϕ, and may in no way mean that the Mendelian model $\phi = \varphi_0$ has been proved.

Clearly, providing a confidence or a credibility interval carries much more information about ϕ. For example, the standard Bayesian statement

$$Prob^\star(0.724 < \phi < 0.794) \quad = \quad 0.95$$

tells us a lot more about how close the unknown parameter ϕ comes to the Mendelian model.

The previous Bayesian statements are reinterpretations of the usual frequentist procedures. But the Bayesian approach may

provide a probability for any property of interest relative to the parameter. For example, we might prefer to consider a property stating that ϕ is *close* to the predicted reference value 0.75, *i.e.* for example that the departure of ϕ from 0.75 in any direction is not more than ϵ considered as a *negligible* amount. For $\epsilon = 0.05$, we get:

$$Prob^{\star}(0.75 - 0.05 < \phi < 0.75 + 0.05) \quad = \quad 0.988\,(\geq\ 0.986).$$

With the Bayesian approach introduced here, it is not possible to prove any sharp model since $Prob^{\star}(\phi = 0.75)$ is always 0 however large the sample size may be; but it is possible to conclude that, with a good guarantee, the departure from the model is negligible.

Finally, some researchers might prefer to formulate the conclusions, either descriptive or inductive, in terms of the odds ratio. The observed odds ratio is $r_{obs} = f_{obs}/(1 - f_{obs}) = 3.18$ to be compared to the reference value $\rho_0 = 3$. From the standard distribution on the parent odds ratio $\rho = \phi/(1 - \phi)$, we may for example get the statement[22],

$$Prob^{\star}(3 - 0.5 < \rho < 3 + 0.5) \quad = \quad 0.826\,(\geq 0.814),$$

in which the guarantee is clearly to small for the observed property "$3 - 0.5 < r_{obs} < 3 + 0.5$" to be generalized to ρ.

What we illustrate here is that in order to say that the departure from the model is small, one must first choose a particular scale to measure the departure (frequency, odds ratio or any other felt relevant). It must be emphasized that having this choice is allowed by the flexibility of the Bayesian approach to inference.

6.1.8 Bayesian vs Frequentist Inference

Significant or Non-significant vs Large or Small. When studying the preceding example, we argued that the "S vs NS" frequentist dichotomy provides a rather poor range of conclusions about

22. If ϕ follows a Beta distribution, the derived parameter ρ actually follows a scaled Fisher/Snedecor F distribution (the unscaled version is well known to those familiar with ANOVA). The standard posterior on ρ is precisely $\frac{2a+1}{2b+1} F(2a+1, 2b+1)$.

the model of interest. We suggest that a better way to summarize the information provided by the data about the validity of a model is to think in terms of a *large* or *small* departure from the model. This methodological point of view is not, by itself, bound to the Bayesian approach, but using that approach makes it particularly easy to adopt (for a detailed account of this viewpoint within the context of ANOVA, see Rouanet, 1996). The two figures that we examine next will help us to make the distinction between the two approaches clearer (see Figure 6.8).

Let us consider, as for Mendel's data, the model $\phi = 0.75$; to fix ideas, let us also take $d_{obs} = (f_{obs} - 0.75)$ as an index of the data's departure from the model, and define $|d_{obs}| \leq \epsilon$ as a small departure and $d_{obs} > \epsilon$ as a large positive one, with $\epsilon > 0$, say for example $\epsilon = 0.10$.

Figure 6.8a p. 195 shows how the values of d_{obs} and of n jointly determine the significance of the frequentist test (for a fixed reference level $\alpha_{sup} = 0.05$). The "NS" conclusion may be reached, either for a small d_{obs}, or for a large one and a small n. At the same time, a "S" conclusion can occur when d_{obs} is large, but also when d_{obs} is small and n is large (note that these remarks actually apply for *any* value for ϵ). With a huge n, almost any value of d_{obs} leads to an "S" conclusion. Obviously, the distinction "S" vs "NS" mixes the observed size of the departure and the size of the sample.

On the other hand, if we now adopt the Bayesian framework and take $\gamma_{sup} = 1 - \alpha_{sup} = 0.95$ as a reference guarantee, we will conclude "Large" when $Prob^{\star}(\delta > \epsilon) > 0.95$, "Small" when $Prob^{\star}(|\delta| \leq \epsilon) < 0.95$, and "Don't know" otherwise (with $\delta = (\phi - 0.75)$). Figure 6.8b p. 195 is analogous to Figure 6.8a and gives the regions where each of these conclusions are reached depending on d_{obs} and n. Now it is clear that the conclusions "Large" or "Small" cannot be reached if d_{obs} is not itself large or small. Thus, this approach can only provide inductive conclusions that generalize descriptive ones. The "Don't know" region mostly corresponds to the case of a small n (there are not enough data for being able to generalize anything), or otherwise to situations where d_{obs} is too close to ϵ to reach a

conclusion (relatively to ϵ). If we superimpose the two figures, we are now in a position to decide whether a "NS" result actually means "The model is approximately true" or "There are not enough data", two conclusions which, undoubtedly, do not sound quite the same. On the other hand, "S" is seen to be a necessary condition for "Large", but not a sufficient one[23].

Simple change of words or more? It could be argued that what we propose in this section is nothing more than replacing "significance test" by "Bayesian test" and "confidence" by "credibility". This is actually a way of thinking of the frequentist-Bayesian mutual interpretations, but only *at the numerical level*. More fundamental, we think, is that the Bayesian approach enables the researcher to formulate his/her conclusions in natural terms, that is in terms of probability statements relative to the unknown parameter.

But it should already appear that there are some decisive advantages to the Bayesian approach, such as the freedom to choose the parameter of interest (*e.g.* frequency or odds ratio) in order to match the researcher's question precisely. This will become even more obvious in the next sections where we examine questions that are either unanswerable or awkwardly answered in the frequentist framework: *e.g.* predictive inference, inference in the presence of nuisance parameters. Unlike what we did in the beginning of this section, we shall now concentrate on the Bayesian approach and shall not attempt to present the frequentist alternatives (when they exist). What will, hopefully, become clear is that the Bayesian approach to inference provides a general, unrestricted, framework for inferring from any property of the data, however complex the data or the property may be, to the corresponding property in the population.

23. Figures like 6.8a and 6.8b (p. 195) could actually be constructed for a variety of other situations; see for example Bernard (1994, p. 79) for inferring on a contrast on several means.

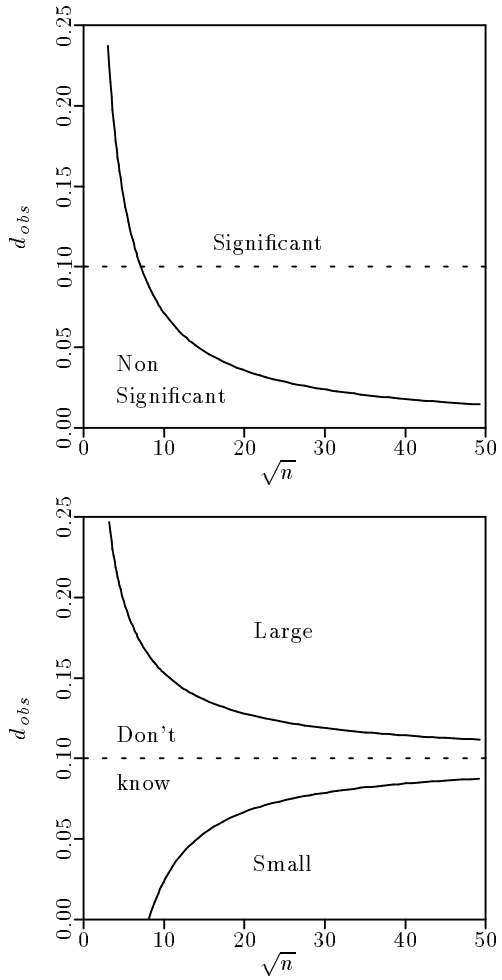

Figure 6.8: Frequentist test vs Bayesian assessment of importance. *(a)* "Significant" or "Non-significant" for $\widetilde{\mathcal{H}_0} : \delta \geq \varphi_0$ at level $\alpha_{sup} = 0.05$ (above); *(b)* "Large", "Small" or "Don't know" for the departure of δ from φ_0 (relatively to $\epsilon = 0.10$) at guarantee $\gamma_{sup} = 0.95$ (below); conclusions in both diagrams are expressed as functions of n (scale in $\sqrt{n}$) and $d_{obs} = f_{obs} - \varphi_0$, for the reference value $\varphi_0 = 0.75$.

6.2 Predictive Inference on one Frequency

6.2.1 The Predictive Approach to Inference

Another way of considering the problem of inference is to adopt a *predictive point of view* by trying to answer the following question: after having observed the composition (a, b), what may be predicted about the composition (a', b') of a future sample of size $n' = a' + b'$, or equivalently about its frequency $f' = a'/n'$?

For Mendel's data (Section 6.1.7), we found the descriptive property $0.70 < f_{obs} < 0.80$. In the predictive formulation of the problem of inference, one is interested in the probability of finding this property again in a future experiment with n' other peas. As Guttman (1983) pointed out, this simple question cannot be answered within the traditional (frequentist) framework.

6.2.2 Predictive Distribution

The predictive distribution required to answer this type of question can also be derived from Bayes' theorem[24]. Let us first consider the case of a finite population of size N. If the prior on ϕ is Beta-binomial $BeBi(\alpha, \beta; N)$, then the predictive distribution of $f' = a'/n'$ given $f = a/n$ is also a member of the same family, as it is a $BeBi(a + \alpha, b + \beta; n')$, whose expression is:

$$p^f_{f'} = \frac{\binom{a'+a+\alpha-1}{a'}\binom{b'+b+\beta-1}{b'}}{\binom{n'+n+\nu-1}{n'}}. \tag{6.18}$$

As we shall see next, this distribution summarizes all the Bayesian distributions given until now in this chapter.

24. In equation (6.3), the expression $\sum_{\varphi\in\Phi} p_\varphi p^\varphi_f$ may also be written p_f and represents *prior predictive probabilities*. After having observed (a, b) on a first sample, the resulting state of knowledge about ϕ is given by the posterior probabilities p^f_φ. From this state of knowledge taken as a new prior one, it is possible to apply Bayes' theorem again in order to integrate a second sample (a', b'); then the previous expression will provide *posterior predictive probabilities*, or, in short, *predictive probabilities*.

Predictive vs posterior distribution. Inferring on the population composition (A, B) from the observed (a, b) is nothing other than predicting what remains unknown in the population, that is the composition $(A', B') = (A - a, B - b)$ of the $N' = N - n$ remaining elements. Thus the posterior distribution can be seen as a particular predictive distribution where the prediction bears on the whole N' remaining elements, whereas in the predictive distribution, the prediction only bears on the n' next elements.

Hence, the only difference between the predictive distribution given in (6.18) and the posterior distribution of (6.6) is the extent of the prediction realized. This is obvious from their respective equations: replacing (a', b') by $(A' = A - a, B' = B - b)$ in equation (6.18) leads us back to equation (6.6).

Predictive distribution for infinite N. As we have said, the hypergeometric and binomial sampling models only differ by the status of N, finite or infinite. But it can be seen that, in the predictive distribution, the size N of the population no longer appears. Consequently the predictive distribution is the same whatever sampling model is considered. The only difference between these two models is that the finiteness of N induces a limit for n', $n' \leq (N - n)$, whereas, when N is infinite, n' may be any value. In each model, setting n' to its maximum value, either $n' = N - n$ or $n' = \infty$, will give back the posterior distribution.

Characteristics of the predictive distribution. As the posterior, the predictive distribution is centered on the relative posterior strength of the "success" category; on the other hand, its variance is larger than the posterior's:

$$Mean(f') = \frac{a + \alpha}{n + \nu}, \tag{6.19}$$

$$\mathrm{Var}(f') = \frac{(a + \alpha)(b + \beta)}{(n + \nu)(n + \nu + 1)} \left(\frac{1}{n + \nu} + \frac{1}{n'} \right). \tag{6.20}$$

In the predictive distribution, there are two sources of variance or uncertainty: the first one concerns the parameter ϕ which, as in

the posterior distribution, is only known through the n available observations; the second one comes from the fact that, for a given ϕ, there is still some uncertainty about the composition of the n' observations on which the prediction bears. These two sources of variance combine additively to give the above predictive variance. This is even more obvious in the following approximation of the variance, valid for small prior strengths and large sizes n and n':

$$\mathrm{Var}(f') \approx f(1-f)(\frac{1}{n}+\frac{1}{n'}). \tag{6.21}$$

A common way for frequentists to bypass the difficulties of the frequentist approach for answering such predictive questions is to estimate ϕ from the data, *e.g.* $\widehat{\phi} = f_{obs}$, and to predict f' assuming that $\widehat{\phi}$ is the true value of ϕ. This approach is obviously biased since it does not take into account one source of uncertainty, namely the one about ϕ, and thus leads to overprecise predictions. Raftery, Madigan & Volinsky (1996) show that using a Bayesian predictive analysis instead improves predictive performance[25].

Prediction about the next observation. The special case $n' = 1$, *i.e.* predicting the next observation, has certainly been one of the most discussed in the early era of statistical inference[26]. For the next observation there are only two possible outcomes, either $(a' = 1, b' = 0)$ so that $f' = 1$, or $(a' = 0, b' = 1)$ so that $f' = 0$. The posterior distribution is then determined by the single probability of a success in the next trial,

$$Prob(f' = 1) = \frac{a+\alpha}{n+\nu}, \tag{6.22}$$

25. This difference between the frequentist and Bayesian approaches to inference is true in more general contexts than prediction: the Bayesian approach enables one to take into account *all* sources of uncertainty when inferring on a parameter of interest (see Chapter 5, Section 5.2.3).

26. The most famous example is the Laplace (1825/1986 p. 45) "probability that the sun will rise again tomorrow" example, which, as Bernard Bru points out (id., postface, pp. 263–264), was not amongst the least arguable applications of his method.

which has the same expression as the posterior mean in the general case[27].

As may be seen, this expression differs slightly from the first intuitive answer that might be proposed: "if one observed a successes out of n, then the probability of a subsequent success is $\frac{a}{n}$". We just used the expression "first intuitive answer" because considering cases $a = 0$ or $a = n$, especially when n is small, leads one to thinking that this latter formula is over-confident. The prior strengths α and β involved in (6.22), if they are chosen strictly positive, may thus be viewed as safeguards against a too "data-glued" inferential statement, since they allow for the possibility of an event that has not yet been observed.

Adopting the ignorance zone formalization of prior ignorance, we find $Prob^\star(f' = 1) = [\frac{a}{n+1}, \frac{a+1}{n+1}]$. This suggests an alternative "predictively-minded" interpretation of the ignorance zone. After having observed a successes out of n, one may wish to give a probability statement that will remain acceptable after the next observation, *whatever it may be*, will have been observed. The first intuitive answer, $\frac{a}{n}$, will be modified into $\frac{a+1}{n+1}$ if this hypothetical next observation is a success, and into $\frac{a}{n+1}$ if it is a failure: these two values are precisely the two bounds of our ignorance-zone-based predictive probability.

Example of a predictive distribution. Figure 6.9 p. 200 gives the standard predictive distribution (prior strengths $\alpha = \beta = 1/2$) for a hypothetical *replication* of Mendel's experiment ($n' = n = 556$). From this distribution, we get the statement $Prob^\star(0.70 < f' < 0.80) = 0.929$. For several other values of n', this probability would be: 0.651 ($n' = 100$), 0.957 ($n' = 1000$) and 0.985 ($n' = 10000$). This last value is already extremely close to the result found in Section 6.1.7 from the standard posterior (0.988).

27. For a uniform prior ($\alpha = 1$ and $\nu = 2$) equation (6.22) is the famous Laplace's *rule of succession*.

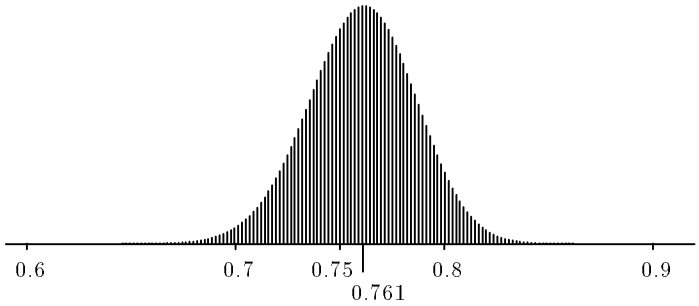

Figure 6.9: Mendel's data. Standard predictive distribution on f' for a future sample of size $n' = 556$ ($f_{obs} = 0.761$, $n = 556$).

6.2.3 Remarks on the Predictive Approach

Predictive approach and exchangeability. In this section, we derived the predictive distribution from a parametric sampling model characterized by the unknown parameter ϕ. There is actually a more direct and intuitive way to obtain the predictive distribution. The idea, due to de Finetti (1974-1975, 1981), and that we only sketch here, is to consider a frame model where observables, either actually observed or future ones, are considered *exchangeable*, and where the prior is defined on observables only. In this framework, the concept of a parameter becomes a secondary one, and is derived from considering the limiting case of a large or infinite sequence of exchangeable data items (see also Geisser, 1993, pp. 1–5).

Generality of the predictive approach. The predictive approach provides a unified framework for the two sampling schemes that were considered in Section 6.1, and this on two counts: *(i)* it includes the inference on the parameter as a particular case, and *(ii)* it does not require wondering whether the population has a finite or infinite size; the only condition is to assume that the population is large enough to contain the n' observations on which one wants to make a prediction. One implication of this is that, using the predictive viewpoint, the notion of a parent parameter may be envisaged

in a much more intuitive way: the parent frequency can be seen as the frequency in a future sample of extremely large size. This intuitive interpretation will surely be found helpful in subsequent sections dealing with more complex data structures involving several unknown parameters.

Because the predictive probabilities are obtained through Bayes' theorem, they are, like the posterior ones, of an epistemic nature: they go from what is known (the n observed data) to what is not (the n' future ones). But predictive probabilities are, we think, even more intuitive than posterior ones because they only relate *observables* between each other. As a matter of fact, Marie-Paule Lecoutre experiments (see Chapter 3) indeed showed that the predictive formulation of the problem of inference appears very natural to researchers.

Quite surprisingly — and even though "... *prediction was the earliest and most prevalent form of statistical inference*" as Geisser (1993, preface) points out —, research in predictive inference is now quite underdeveloped compared with research in parametric inference. Very few books, even amongst those adopting the Bayesian framework, actually attempt to correct this oversight; noticeable exceptions are Aitchison & Dunsmore (1975) and Geisser (1993).

6.3 Bayesian Inference on Several Frequencies (structure $S \rightarrow U_K$)

What we have dealt with up to here in this chapter are dichotomous data. This simple case has enabled us to set up the Bayesian framework. We shall now generalize some of the preceding results to the case of polytomous data, *i.e.* involving $K > 2$ categories. On top of its intrinsic interest, this situation also constitutes the key to the analysis of structured data, *i.e.* data whose design involves several factors. The stress here will not be put on technical matters but rather on the substance of the results and their interpretation.

More details about technical matters can be found in Bernard (1983, 1986).

We shall first restate the problem of inference when several parameters are involved with the aid of a simple example with $K = 3$ (Section 6.3.1) before moving to general considerations on Bayesian inference in this context (Section 6.3.2) and finally exploring in more detail two more complex examples (Section 6.3.4).

6.3.1 The Problem of Inferring on Several Frequencies

Let us consider an arbitrarily large population (we assume N infinite) whose composition in frequencies according to a ternary variable is $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$ with $\phi_1 + \phi_2 + \phi_3 = 1$. The parent composition $\boldsymbol{\phi}$ is unknown, but a sample of size n randomly extracted from the population is available and has led to the observed composition in counts $\boldsymbol{a} = (a_1, a_2, a_3)$ with $a_1 + a_2 + a_3 = n$.

For a given parent composition $\boldsymbol{\phi}$, the sampling distribution of $\boldsymbol{a}$ is given by the *multinomial distribution*, which generalizes the binomial distribution to K categories. Under the random sampling frame model, this distribution provides probabilities of $\boldsymbol{a}$ given $\boldsymbol{\phi}$.

An illustrative example: "Ordered data". For the purpose of illustration, let us consider the following data (hereafter called "Ordered data") where each observation falls into one of three categories: the observed counts are $\boldsymbol{a} = (10, 8, 6)$, with $n = 24$ and hence the observed frequencies $\boldsymbol{f} = (0.417, 0.333, 0.250)$. These have the descriptive property $f_1 > f_2 > f_3$ and a question of interest is whether this property can be extended to the population: Can we say that $\phi_1 > \phi_2 > \phi_3$?

Parameters and questions. When considering binary data, the frame model was characterized by a single parameter, the unknown frequency ϕ of one of the two categories. But now we have two free parameters (not three, because the parent frequencies ϕ_k must add up to 1). For a given question, there will usually be a corresponding *parameter of interest* which will be the subject of inference. This parameter of interest may be one of the initial parameters of the

sampling model, *e.g.* ϕ_1, or a *derived parameter*, *e.g.* $(\phi_1 - \phi_2)$ or ϕ_1/ϕ_2, *etc.*.

But some other questions may be more complex and may require statements that are relative to several parameters at the same time. This is the case in our example, as the question of interest may be stated: "Can we conclude that *both* $(\phi_1 - \phi_2)$ and $(\phi_2 - \phi_3)$ are positive?"

Shortcomings of the frequentist methods. As the sampling model now involves several parameters, some difficulties arise as far as frequentist methods are concerned. Two major difficulties are examined below.

Nuisance parameters. Suppose there is a unidimensional parameter of interest θ (*e.g.* $\phi_1 - \phi_2$) and that we want to test some hypothesis about θ, say $\theta = \theta_0$. Because the sampling model involves two parameters, this hypothesis is by itself generally insufficient to fully determine the sampling distribution: this distribution still depends on some *nuisance parameters*. In our example there are two free parameters, so that when specifying one through the hypothesis there will still be one remaining nuisance parameter (*e.g.* $\phi_1 + \phi_2$). In some simple situations, it is possible to overcome this difficulty by conditioning the model on some statistic, calculated from the data, that gives little or no information about the parameter of interest. This approach leads to *conditional tests* sometimes referred to as "exact tests". But, quite often for more complex derived parameters, this first approach cannot be used and the only remaining solution is to resort to asymptotic considerations that provide approximate tests. The difficulty is that the resulting methods are typically not valid for small samples, which explains the ritual warnings such as: "the Chi-square test is only valid if the expected absolute frequencies are all greater than 5".

Questions without answers. The problem of nuisance parameters transfers to the confidence interval procedure. Moreover, in some cases, the problem may become dramatic as it may be possible to devise a test *for some* reference values of the parameter of interest

but *not for all* of them. An example is the absence of a satisfactory confidence interval for the ratio of two frequencies from independent samples (Aitchison & Bacon-Shone, 1981).

But, in our view, the most critical shortcoming of the frequentist approach is its inadequacy to deal with complex properties of interest such as the one in our example. It is indeed hard to think of any single $\mathcal{H}_0$ whose rejection would enable us to conclude that $\phi_1 > \phi_2 > \phi_3$.

6.3.2 Bayesian Inference

On the contrary, the Bayesian approach is free from these limitations. Again it involves a prior distribution on the unknown $\boldsymbol{\phi}$ which, when combined with the data $\boldsymbol{a}$, leads to a posterior distribution on $\boldsymbol{\phi}$. Both these distributions bear simultaneously on *all* the parameters of the sampling model, so that the answer to any question is, at the theoretical level, quite simple[28]. If there is one single parameter of interest, one derives, from the overall posterior distribution, the corresponding *marginal* distribution on this parameter; if one is interested in a complex property, the answer is the posterior probability of the appropriate region of the parameters' space. There is no theoretical need to resort to asymptotic results so that the Bayesian approach can be applied to any sample, large or small.

The K-category generalization of the Beta distribution that we introduced for binary data is the *Dirichlet distribution* (for its main properties, see *e.g.* Wilks, 1962, pp. 177–182; Fang, Kotz & Ng, 1990, pp. 16–24). When combined with data sampled according to a multinomial model, a Dirichlet prior leads to a Dirichlet posterior. For a K categories problem, the prior distribution is characterized by K prior strengths, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ which are updated into posterior strengths $\boldsymbol{a} + \boldsymbol{\alpha} = (a_1 + \alpha_1, \cdots, a_K + \alpha_K)$. Again we denote ν the total prior strength: $\nu = \sum_k \alpha_k$.

28. We remind the reader that the computational issues are discussed in Section 6.5.

Ignorance/standard prior distributions. The uniform prior is obtained by taking the prior strengths α_k all equal to 1: $\alpha_k = 1$. Haldane's (1948) prior corresponds to null prior strengths, *i.e.* $\alpha_k = 0$; with Jeffreys' rule (1946, 1938/1961) all prior strengths are $\alpha_k = 1/2$, a solution which now differs from Perks' (1947) who suggested $\alpha_k = 1/K$.

The definition of an ignorance zone is more delicate in the general case of K categories[29]. Following the same line as in Bernard (1996), we suggest defining it as the region $\sum_{k=1}^{K} \alpha_k = 1$, a proposal that is in agreement with Walley's (1996a) suggestion of an "imprecise Dirichlet model" with $\nu = 1$. Within this region, we take Perks' (1947) prior, $\alpha_k = 1/K$, as the standard prior[30].

Standard posterior distribution. Figure 6.10 p. 206 shows the standard posterior distribution $Di(10 + \frac{1}{3}, 8 + \frac{1}{3}, 6 + \frac{1}{3})$ for the Ordered data. The support of this distribution is a *simplex* (all points within a triangle for $K = 3$), each vertex of which corresponds to one of the most extreme possible compositions: $(1, 0, 0)$ (left), $(0, 1, 0)$ (bottom) and $(0, 0, 1)$ (right). The relative posterior strengths determine the center of the distribution, while its dispersion mostly reflects the small overall posterior strength $n + \nu = 25$.

Figure 6.11 p. 207 shows the simplex of all possible compositions with the two necessary ingredients for answering our inductive question: Figure 6.11a gives another view of the posterior distribution where the probability density is now expressed by means of "isodensity" contours; Figure 6.11b indicates the region of the simplex for which the property $\phi_1 > \phi_2 > \phi_3$ holds. Answering our initial question simply amounts to "merging" these two

29. One difficulty is that for some of the proposed priors (the uniform and Jeffreys') the total prior strength depends on K. This may be seen as undesirable, because the number K of categories into which data are classified may sometimes be quite arbitrary. Neither our proposed standard prior nor the ignorance zone idea present this difficulty.

30. This symmetrical standard prior should be used when the K categories do not have any particular underlying structure. If some tree-structure underlies the set of categories, the standard prior needs to be adapted in order to take the tree-structure into account (Bernard, 1997).

Insert Figure 6.10 about here

Figure 6.10: Ordered data. Standard distribution on $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$ for an observed composition in counts $\boldsymbol{a} = (10, 8, 6)$.

views of the simplex by computing the posterior probability, according to Figure 6.11a, of the region given in Figure 6.11b: we find $Prob(\phi_1 > \phi_2 > \phi_3) = 0.423$ (≥ 0.375). The degree of generalizability of the descriptive property is too small; the property cannot be extended to the population[31].

6.3.3 Some Derived Parameters and Their Posterior Distributions

The posterior on $\boldsymbol{\phi}$ fully determines the posterior on *any* derived parameter. There are two properties of the Dirichlet that allow making specific inferences, namely the *pooling* and the *restriction* properties (Bernard, 1997): *(i)* when pooling two categories k and k',

31. It is only for a certain type of properties (*e.g.* those corresponding to a region of the simplex defined by a linear inequation) that the probability of the property is assured to be greater than 0.50 when the property is true for the sample. Here, there is a larger probability for the population property not to hold (0.577) than there is for it to hold (0.423). This fact would be even more striking for properties defining a smaller region of the simplex.

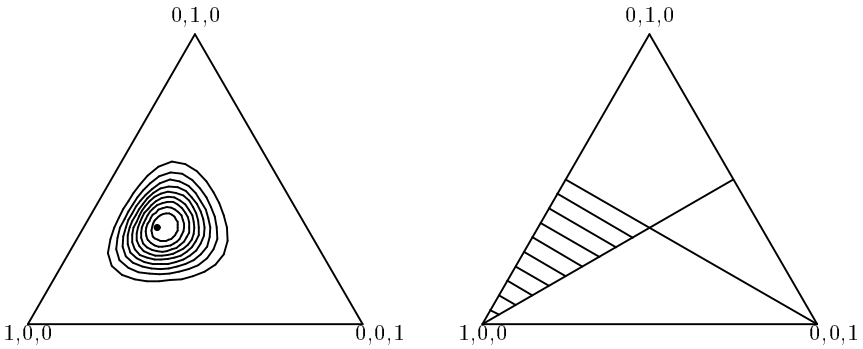

Figure 6.11: Ordered data. *(a)* Simplex with isodensity countours with respect to the standard distribution on $\boldsymbol{\phi}$ (left); *(b)* Region of the simplex for which the property $\phi_1 > \phi_2 > \phi_3$ holds (right).

the posterior is still a Dirichlet with all necessary vectors (frequencies and strengths) transformed by summing their k and k' components; *(ii)* if inference is restricted to a subset K' of K and thus bears on the associated conditional (on K') frequencies, the posterior is still a Dirichlet, but a reduced one involving only the K' categories and their respective strengths.

For the privileged problem of the *partial comparison of two frequencies*, say ϕ_j and ϕ_k with respective posterior strengths α'_j and α'_k, two derived parameters are generally considered: $\delta = \phi_j - \phi_k$ and $\rho = \phi_j / \phi_k$. The posterior distribution on δ is not standard but may be easily obtained by an appropriate software; on the other hand, the exact posterior of ρ is a scaled Fisher/Snedecor distribution: $\rho \sim (\alpha'_j/\alpha'_k) F(2\alpha'_j, 2\alpha'_k)$. We already saw a particular case of that property for $K = 2$ (see Footnote 22, p. 192).

6.3.4 Two Examples of Typical Problems

Validating a model: Mendel's data (shape and colour). Let us reconsider Mendel's full data involving the shape and colour attributes of peas, already discussed in Chapter 2 Section 2.1.1 and

that were partly analyzed in Section 6.1.7: the observed data on 556 peas and the expected frequencies according to Mendel's theory are given in Table 6.3.

Table 6.3: Mendel's data (shape and colour). Observed compositions in counts and frequencies, and theoretical frequencies.

	Round		Wrinkled		
	Yellow	Green	Yellow	Green	
Observed counts a_k	315	108	101	32	556
Observed frequencies f_k	0.567	0.194	0.182	0.058	1
Theoretical frequencies φ_{0k}	0.5625	0.1875	0.1875	0.0625	1

As we already noted in Chapter 2, the data could hardly be more in accordance with Mendel's theory: the goodness of fit indicator Phi-square was found to be 0.0008. However, we also stressed that the non-significant result obtained from the standard goodness of fit Chi-square test ($\chi^2 = 0.47$, $p_{obs} = 0.93$) is insufficient to reach the conclusion that "the model is true". The only allowed conclusion from this test is that "the data are compatible with the model".

Now, within the framework adopted in this chapter, how can we try to really validate Mendel's model? The idea is quite simple and proceeds as follows. We must first choose a relevant goodness of fit descriptive indicator. Second we must define some criteria, relative to this indicator, that can be considered as indicating a "good-enough" fit. This criteria provides a property of interest that a frequency composition (on four categories) may or may not have. Then we will just proceed as we did before: *(i)* check whether the property is true for the observed data, and, if it is, *(ii)* calculate the probability for it being true in the underlying population.

Let us first take the Phi-square, *Phi2*, as a relevant descriptive indicator: a value of 0 indicates a perfect fit and the greater the value the worse the fit is. Following Corroyer & Rouanet (1994) we can take $Phi2 < 0.20^2 = 0.04$ as a criteria of a small departure from a

perfect fit. Descriptively this property is true as the observed value of this indicator is $Phi2_{obs} = 0.0008 < 0.04$. To the observed Phi-square, $Phi2_{obs}$, computed from the observed frequencies $\boldsymbol{f}$, corresponds the unknown parent $Phi2_{par}$, a derived parameter computed in a similar way from the parent frequencies $\boldsymbol{\phi}$[32]. From the overall standard posterior distribution,

$$\boldsymbol{\phi} \quad \sim \quad Di(315 + \tfrac{1}{4}, 108 + \tfrac{1}{4}, 101 + \tfrac{1}{4}, 32 + \tfrac{1}{4})$$

we can derive the marginal distribution of $Phi2_{par}$ from which we get:

$$Prob(Phi2_{par} < 0.04) \quad = \quad 1.000\ (\geq 1.000).$$

The inductive conclusion is straightforward: the departure of $\boldsymbol{\phi}$ can be assessed negligible (defined operationaly as $Phi2_{par} < 0.04$) with a guarantee of at least 1.000 (values are rounded to three decimal places, so that ≥ 1.000 actually means ≥ 0.9995).

Of course it is not necessary to specify the 0.04 limit in advance. We could instead specify some guarantee, say $\gamma = 0.95$, and find the corresponding limit for $Phi2_{par}$. Doing so, we get:

$$Prob(Phi2_{par} < 0.015) \quad = \quad 0.95.$$

It cannot be overemphasized that the line of reasoning just presented is quite general and could be applied to *any* relevant indicator and/or criteria. For example we might prefer to consider the maximum (over the K categories) of the absolute value of the relative deviation between f_k and φ_{0k}: $MaxDev_{obs} = Max_k|\frac{f_k - \varphi_{0k}}{\varphi_{0k}}|$. For this other indicator we find descriptively $MaxDev_{obs} = 0.079$ and inductively $Prob^{\star}(MaxDev_{par} < 0.333) = 0.95$[33]. With a 0.95 guarantee, it may be said that *none* of the four true frequencies ϕ_k departs from the

32. It is common to note observables quantities with a latin letter and the corresponding unknown parameter with a related greek letter (*e.g.* f and ϕ). But of course the alphabet quickly turns out a bit short and the alternative "obs" vs "par" subscripts notation becomes more handy.

33. We challenge frequentists to define sensible confidence limits for this complex indicator.

corresponding reference frequency φ_{0k} by more than 33.3% in terms of relative deviation. This conclusion seems to be much less in favour of the Mendelian model, but it must be realized that this indicator is also a much more severe one than the Phi-square as it requires a condition simulaneously on each of the four frequencies.
Remark: The move from the observed statement $Phi2_{obs} = 0.008$ to the inductive one $Prob^{\star}(Phi2_{par} < 0.015) = 0.95$ can be thought of as "paying a tax" whose amount is the difference $0.015 - 0.008$. This "tax" is the price to pay for having a statement on the unknown parent parameters rather than one on the data set at hand only. Of course there is a trade-off between the strength of the inductive statement and its guarantee: the lower the guarantee, the stronger the statement, *i.e.* the lower the tax. In addition the tax can be reduced with more data, because then the "generalizability potential" of the data is larger.

Assessing a "quasi-implication": Fractions data. Does success to task A imply, approximately, success to task B? This type of question is quite common in developmental Psychology, since such a fact, if it was established, may point to the hypothesis that the acquisition of B-type ability is necessary for the acquisition of A-type ones. For example, in an experiment about number construction in children, Charron (1996) asked 165 school pupils to do several tests in which the task consisted of calculating some quantity through the use of a fraction which could express either a Part-Part or a Part-Whole relationship. Table 6.4 gives the observed counts for two of these tests A and B (success is denoted a and b, failure is denoted a' and b').

Amongst subjects succeeding at task A, 92.3% also succeed at task B; there is a *quasi-implication* from a to b. On the other hand only 50.0% of subjects succeeding at B also succeed at A, so that the reciprocal quasi-implication is much less supported by the data.

For Table 6.4, the observed Phi-square is $Phi2_{obs} = 0.298$ which indicates a large descriptive departure from independence; the Chi-square test for independence (corrected for continuity), $\chi^2 = 46.63$

Table 6.4: Fractions data. Observed counts for two tests A (Part-Part relationship) and B (Part-Whole relationship) from Charron (1996).

		Part-Part b	Part-Part b'
Part-Whole	a	36	3
	a'	36	90

($p < 10^{-6}$), is highly significant and thus clearly points to the existence of a departure from independence. However, none of these results tell us that this departure occurs in the *specific direction* of an implication $a \Longrightarrow b$, as both statistics would be unchanged by permuting A and B.

Hildebrand, Laing & Rosenthal (1977) proposed a general descriptive index "Del" for measuring the departure of the frequencies of an $A \times B$ cross-classification table from a specific logical model (a model which specifies that one or several cells should be empty). For the case of a simple implicative model in a 2×2 table (one empty cell), the Del index is equivalent to Loevinger's (1948) "homogeneity index" and to Rouanet, Le Roux & Bert's (1987, pp. 156–160) "association rate index". For Table 6.4 and the model $a \Longrightarrow b$ (cell ab' empty), the Del index is defined as

$$d_{a \Longrightarrow b} = 1 - \frac{f_{ab'}}{f_a f_{b'}}, \tag{6.23}$$

where $f_{ab'}$ denotes the ab'-cell frequency and f_a and $f_{b'}$ the corresponding marginal frequencies. This index equals 0 in case of independence and 1 if the logical model is descriptively true (the index may be negative if the association between A and B is negative); a high value thus indicates a high degree of quasi-implication.

One problem is that the corresponding existing inferential frequentist methods, because they are based upon asymptotic consid-

erations, do not provide valid statements, neither for small data sets, nor, paradoxically, for cases where the implicative model is almost descriptively verified (Hildebrand *et al.*, 1977, pp. 206–208).

In Bernard & Charron (1996a) we proposed a Bayesian approach for the study of oriented dependencies in 2×2 tables, called "Bayesian Implicative Analysis", which is free from these difficulties. On the descriptive side, the method is based on the idea that the implicative analysis must take into account all d indexes (one for each possibly empty cell of the table) and particularly the two positive ones; depending on their values, the descriptive conclusion may be that of a "quasi-implication", a "quasi-equivalence" or a "quasi-independence". For the Fractions data, we get:

$$\begin{aligned} d_{a \Longrightarrow b} &= 0.864 \\ d_{b \Longrightarrow a} &= 0.345 \end{aligned}$$

These values reflect the asymmetry in Table 6.4: the degree of the quasi-implication $a \longrightarrow b$ is high (close enough to 1) whereas the degree of its reciprocal is not.

Going from the descriptive side to the inductive one is again straightforward in the Bayesian framework. The first step leads to the standard Bayesian distribution on the vector of the parent frequencies $\boldsymbol{\phi} = (\phi_{ab}, \phi_{ab'}, \phi_{a'b}, \phi_{a'b'})$:

$$\boldsymbol{\phi} \sim Di(36 + \tfrac{1}{4}, 3 + \tfrac{1}{4}, 36 + \tfrac{1}{4}, 90 + \tfrac{1}{4})$$

From this overall distribution, we next derive the joint marginal distribution on the two parent Del indexes, $\delta_{a \Longrightarrow b}$ and $\delta_{b \Longrightarrow a}$, which is summarized in Figure 6.12[34].

From this joint distribution we may compute the probability of any statement relative to $\delta_{a \Longrightarrow b}$, *e.g.* $Prob^{\star}(\delta_{a \Longrightarrow b} > 0.70) = 0.964$, to $\delta_{b \Longrightarrow a}$, or to both simultaneously. For example we may define

34. The entire figure actually also contains, as is done in Bernard & Charron (1996a, p. 28), the region in which the two δ indices are negative. This region has been omitted here because it has a very low probability ($< 10^{-6}$).

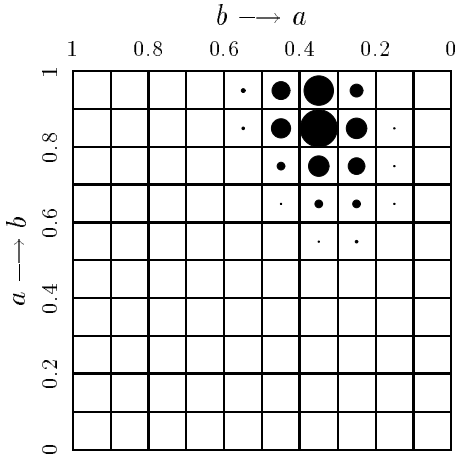

Figure 6.12: Fractions data. Joint standard Bayesian distribution on $\delta_{a \Longrightarrow b}$ and $\delta_{b \Longrightarrow a}$ with disks' surface proportional to probability; the vertical axis indicates the degree of the quasi-implication $a \longrightarrow b$, from 0 (independence) to 1 (implication $a \Longrightarrow b$); the horizontal axis reads similarly for the reverse quasi-implication $b \longrightarrow a$; the largest probability is obtained for $\delta_{a \Longrightarrow b} \in [0.8, 0.9]$ and $\delta_{a \Longrightarrow b} \in [0.3, 0.4]$ which corresponds to the observed values of the d's: 0.864 and 0.345.

operationaly the notion of a quasi-implication from a to b (meaning that $a \longrightarrow b$ at a high degree but that the reciprocal is not true) by a joint statement on $d_{a \Longrightarrow b}$ and $d_{b \Longrightarrow a}$ such as,

$$d_{a \Longrightarrow b} > 0.70 \quad \textit{and} \quad d_{b \Longrightarrow a} < 0.50.$$

This property is descriptively verified for the Fractions data; at the inductive level, we get,

$$Prob^{\star}(\delta_{a \Longrightarrow b} > 0.70 \ \ \& \ \ \delta_{b \Longrightarrow a} < 0.50) \quad = \quad 0.959 \ (\geq 0.920)$$

so that the conclusion of a quasi-implication from a to b (relatively to the threshold values 0.70 and 0.50) can be generalized to the population with a guarantee of at least 0.920.

As we announced, the method can also be used for extreme data. Suppose, for example, that the ab' count in Table 6.4 were 0 (non structural) instead of 3. Then, we would find descriptively $d_{a \Longrightarrow b} = 1$ and $d_{b \Longrightarrow a} = 0.357$, and inductively $Prob^{\star}(\delta_{a \Longrightarrow b} > 0.70 \ \& \ \delta_{b \Longrightarrow a} < 0.50) = 0.992 \ (\geq 0.992)$.

In Bernard & Charron (1996a) we extend this "Bayesian Implicative Analysis" to the study of several binary variables, the method providing, in the end, descriptive and inductive implicative graphs. In Bernard & Charron (1996b), we give extensions to more complex logical models in $A \times B$ tables.

6.4 Examples of More Complex Designs

The previous section was concerned with data whose design structure may be formally written as "$S \rightarrow U_K$" ("Subjects categorized in a set U_K of K categories")[35]. The study of such designs is actually the key to the Bayesian analysis of more complex designs, since the observation space U_K may be itself complex for example because some particular tree-structure underlies it. Indeed, one of our directions of research has been to consider the case of tree-structured categories, with applications to the analysis of sequential data from Ethology (Bernard, Blancheteau, Rouanet, 1985; Bernard, Blancheteau, 1987; Bernard, 1997).

Another direction has been to show how the Bayesian approach could be extended to "quasi-complete" designs, *i.e.* formally $S<G>\times T \rightarrow U_2$ ("Subjects nested within Groups and crossed with Trials providing binary observations"), on the theoretical level (Bernard, 1986) as well as on the practical one with the development of the specific software `IBFGT2` (Poitevineau, Bernard, 1986).

In this section, we shall only illustrate through examples the results of Bayesian inference for two paradigmatic experimental designs and questions, namely the comparison of two independent

35. In the following we shall use several such formulas which belong to the "`LID`" language used by the `EyeLID` software (see *e.g.* Bernard, 1994).

samples ($S < G_2 >\!\!\rightarrow U_2$) and of two matched samples ($S \times T_2 \rightarrow U_2$) with binary observations[36].

6.4.1 The "Aspirine Data" (Two Independent Samples)

The following data are extracted from a study on 22071 American doctors started in 1982 and whose purpose was to investigate the effect of regularly taking aspirin on several health indicators (Steering Commitee of the Physicians' Health Study Research Group, 1988); we only consider here the data relative to myocardial infarction (MI) occurence. The doctors were randomly allocated to two experimental groups: 11037 took a $325mg$ dose of aspirin every other day (group g_1), and 11034 took a placebo instead (group g_2). After 57 weeks of treatment, 104 MI occured within group g_1 and 189 within group g_2.

A frequentist analysis of these data is presented in Rouanet, Bernard, Le Roux (1990, p. 210). We shall only present here its Bayesian analysis. Descriptively, the frequencies of MI in each group are: $f_1 = 104/11037 = 0.0094$ and $f_2 = 189/11034 = 0.0171$. Both frequencies being very small, it is more natural to compare them considering their ratio than considering their difference; we have $r_{obs} = f_2/f_1 = 1.82$, so that descriptively we may conclude that taking aspirin did reduce the risk of MI by a coefficient of at least $3/2$: $r_{obs} > 3/2$.

If we want to generalize this result (to a larger population from which the set of 22071 doctors may reasonably be considered as a random sample), we need to infer on the derived parameter $\rho = \phi_1/\phi_2$ corresponding to the observed r_{obs}. From the overall

36. To give a hint to the key link between these designs and the "$S \rightarrow U_K$" design, let us only say that there are some strong equivalences (in terms of the Bayesian distributions involved) between designs "$S < G_2 >\!\!\rightarrow U_2$" and "$S \rightarrow G_2 \times U_2$" on the one hand, and between designs "$S \times T_2 \rightarrow U_2$" and "$S \rightarrow {U_2}^{T_2}$" on the other. More details, in particular for the specification of ignorance/standard priors, can be found in Bernard (1983).

standard distribution on (ϕ_1, ϕ_2)[37], we first derive the standard distribution on ρ which then leads to the following statements:

$$\begin{aligned} Prob^{\star}(\rho > 3/2) &= 0.945\ (\geq 0.937) \\ Prob^{\star}(\rho > 4/3) &= 0.995\ (\geq 0.994) \end{aligned}$$

The observed property, $r_{obs} > 3/2$, can be generalized into an inductive one, $\rho > 3/2$, with a standard guarantee of 0.945. If we consider a weaker property, $\rho > 4/3$, we get a better guarantee 0.995. This illustrates again the trade-off between the strength of the property and the guarantee in the probabilistic statement.

The unconstrained choice of the parameter of interest for realizing a comparison is a strong advantage of the Bayesian framework. Any relevant derived parameter can be used so that conclusions can be expressed either in terms of the frequencies' difference or of the frequencies' ratio, or any other relevant indicator. In the present case, it is clear that statements about ρ are more directly interpretable than statements about the difference δ between the frequencies, such as for example $Prob^{\star}(\delta > 0.0052) = 0.95$.

6.4.2 "Conflicting Data" (Two Matched Samples)

The other paradigmatic example is that of the comparison of two matched samples (design $S \times T_2 \rightarrow U_2$). It is used here as a base for stressing once again the psychological difficulties encountered in the interpretation of traditional frequentist procedures, and for showing how the Bayesian interpretation framework clears up apparently conflicting results.

Data and frequentist procedures. The data (hereafter called "Conflicting data") are fictitious data borrowed from Marie-Paule Lecoutre (see Chapter 3) concerning the interpretation of significance tests in situations where an initial experiment and a replicate of it lead to apparently conflicting conclusions; in one of M.-P. Lecoutre's

37. The standard distribution is here defined by: $\phi_1 \sim Beta(189.25, 10845.25)$, $\phi_2 \sim Beta(104.25, 10933.25)$ and $\phi_1 \perp\!\!\!\perp \phi_2$.

experiments, researchers in Psychology were presented with the following experimental data and results.

A first experiment involved 50 rats that were tested on two successive trials, $t1$ and $t2$, of a labyrinth run; between the trials, the rats received an injection of some particular drug. For each run (one rat and one trial), the experimenter measured the number of errors, and from it derived a binary measure: "+" for less than two errors, "−" for more. The raw results of experiment 1 are given in Table 6.5a.

Table 6.5: Conflicting data. Observed counts *(a)* for experiment 1 (left) and *(b)* for experiment 2 (right).

		$t2$	
		+	−
$t1$	+	12	15
	−	5	18

		$t2$	
		+	−
$t1$	+	8	15
	−	10	17

The success frequency is higher for trial $t1$ ($f_1 = (12+25)/50 = 0.54$) than for trial $t2$ ($f_2 = (12 + 5)/50 = 0.34$). If we proceed to a McNemar test for comparing the two trials with regard to the frequency of success, we get $\chi^2_{obs} = 5$ for 1 df, and hence $p_{obs} = p_{sup} = 0.013$, so that the difference is significant at the one-sided level $\alpha_{sup} = 0.025$.

In a replicate of the experiment with 50 rats again, another researcher gets the data given in Table 6.5b. Now the frequencies of success are respectively $f_1 = 0.46$ and $f_2 = 0.36$. Here, the McNemar test leads to $\chi^2_{obs} = 1$ so that the difference is not significant ($p_{obs} = p_{sup} = 0.159$).

Experiment 1 reveals inductively a negative effect of the drug on rats' performance, whereas experiment 2 does not seem to confirm this result. The conflict becomes even stronger when researchers are told that, if the two experiments were pooled together, the same

kind of analysis would provide a significant result again: $\chi^2_{obs} = 5$, $p_{sup} = 0.013 < 0.025$.

We are going to see that the perceived conflict is only apparent and that, when analyzing these data in the Bayesian framework (and in particular using the Bayesian reinterpretation of the tests), the feeling of a paradox completly vanishes.

Bayesian procedures. A preliminary remark is in order concerning the data's structure. The design of each experiment is $S \times T_2 \rightarrow U_2$: each subject is tested on each of two trials, each pair subject/trial providing a binary outcome in $U_2 = \{+, -\}$. Equivalently, we can describe the data' structure by saying that for each subject we observe a success/failure profile amongst the four following ones: "++" for two successes, "+−" for a success followed by a failure, "−+" for a failure followed by a success, and finally "−−" for two failures. If we denote V_4 the set of these four profiles, the design can be rewritten $S \rightarrow V_4$, and so we are taken back to a design envisaged in Section 6.3. With this rewriting, the sampling model involves four parameters, *i.e.* the four parent profile frequencies: ϕ_{++}, ϕ_{+-}, ϕ_{-+} and ϕ_{--}; the two frequencies that we want to compare are derived parameters, namely $\phi_{+.} = \phi_{++} + \phi_{+-}$ and $\phi_{.+} = \phi_{++} + \phi_{-+}$; finally if we want to compare them through their difference, we need to consider $\delta = \phi_{+.} - \phi_{.+}$ as the parameter of interest. Notice that δ can even be writen more simply as a partial contrast between the profile frequencies[38]: $\delta = \phi_{+-} - \phi_{-+}$.

Let us first proceed to a quick descriptive analysis of these data: For experiment 1, the observed difference in success between trial t_1 and trial t_2 is $d = 0.20$; for experiment 2 the trial effect is still positive (which corresponds to a negative effect of the drug) but smaller, $d = 0.10$. In the pooled data, we have $d = 0.15$.

38. More generally, any design of the type $S \times T \rightarrow U_K$, where T is a simple or compound factor with several modalities, can be rewriten as $S \rightarrow {U_K}^T$; contrasts on T are then transformed into contrasts on ${U_K}^T$. This property is used in the `IBFGT2` software (Bernard, 1986; Poitevineau, Bernard, 1986).

Figure 6.13 gives the standard distribution on the parent parameter δ for experiment 1. This distribution is roughly centered on the observed difference $d = 0.20$; its dispersion expresses the experimental precision. From this distribution, we get the statements:

$$\begin{aligned} Prob^{\star}(\delta > 0) &= 0.990\ (\geq 0.979) \\ Prob^{\star}(\delta > 0.05) &= 0.959\ (\geq 0.930) \end{aligned}$$

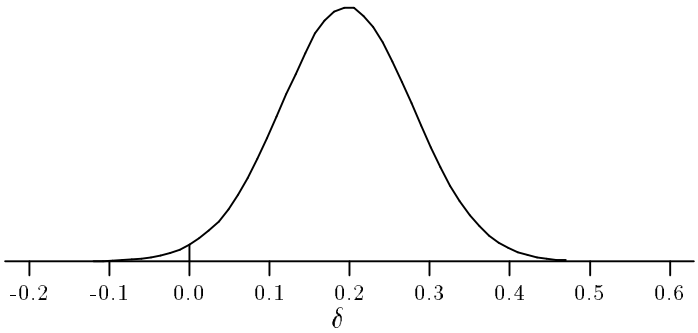

Figure 6.13: Conflicting data (experiment 1). Standard distribution on δ.

The existence of an effect is well established: there is a high guarantee, 0.990, for the true effect δ to be in the same direction as the observed one. Symmetrically, the probability of a negative effect, *i.e.* a true effect in the opposite direction to the observed one, is small: $Prob^{\star}(\delta < 0) = 1 - 0.990 = 0.010$. This last numerical result is the Bayesian counterpart (approximately) of the one-sided level of the frequentist Chi-square test: $p_{sup} = 0.013$.

Figure 6.14 p. 220 shows the standard distributions on δ for each of the two experiments as well as for the pooled data. Experiment 2 descriptively goes in the same direction as experiment 1 ($d = 0.10$), but experimental precision is not large enough to reach the conclusion of the existence of a true effect with a sufficient guarantee: $Prob^{\star}(\delta > 0) = 0.843$. The non-significant result given by the Chi-square test ($p_{sup} = 0.159$) approximately corresponds to the

Bayesian statement $Prob^{\star}(\delta < 0) = 1 - 0.843 = 0.157$. But it should be obvious from the distribution of δ in figure 6.14 that this statement in no way constitutes a proof of the absence of an effect; the distribution is not particularly concentrated around the value 0 as the following statement indicates: $Prob^{\star}(|\delta| < 0.257) = 0.95$.

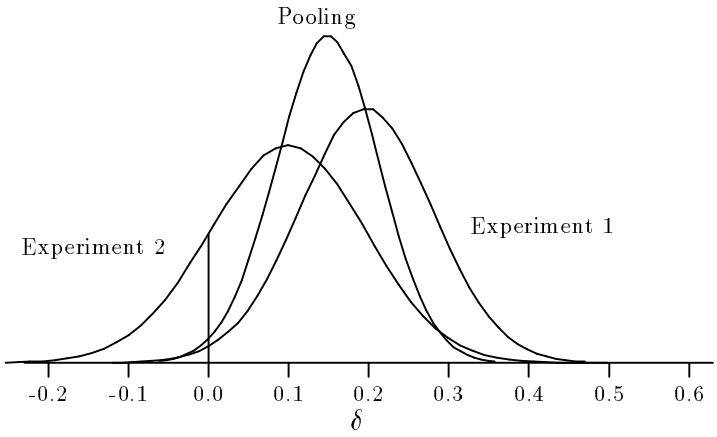

Figure 6.14: Conflicting data. Standard distribution on δ for experiments 1 and 2 and for the pooling of the two experiments.

Again (see Section 6.1.8), the Bayesian reinterpretation of the frequentist observed level sheds light on the limits of the conclusions that a test might lead to: a significant (S) result clearly allows one to conclude that the effect exists (the effect has been proved to be greater than 0), but a non-significant (NS) one does not allow one to conclude that it does not exist. The seeming paradox is only the consequence of a false identification between the dichotomies "S vs NS" and "Effect vs No effect". On the contrary, if we compare the two standard Bayesian distributions for experiments 1 and 2, it is clear that the information that they each provide on δ is not contradictory. The two data sets basically point in the same direction. When pooling the data, the experimental precision becomes sufficient again to conclude that the effect exists, $Prob^{\star}(\delta > 0) = 0.988$, despite the

fact that the overall observed difference $d = 0.15$ is smaller than in experiment 1 alone.

This example reminds us that the first step of the analysis should always be based on description: if the observed d is not small, no inductive analysis, either frequentist or Bayesian, should be able to prove that the corresponding δ is small, and still less that δ is 0. One great advantage of the Bayesian approach is that both the descriptive step and the inductive step are contained in the standard distribution on δ: the distribution is typically approximately centered on the observed d and its dispersion reflects the attained strength of our state of knowledge about the unknown δ.

Last, another point emerging from the preceding comparison between the two approaches is the extreme poorness of the "S vs NS" alternative. "NS", as we have just said, does not tell us very much, and "S" only tells us that the existence of an effect is well established without saying anything about the size of the effect. This criticism, though often pointed out, has not, until now, led many researchers (nor many referees) to enrich their statistical toolbox with more powerful devices.

Our claim, following Rouanet (1996), is that the Bayesian approach is perfectly suited to go beyond significance testing and to provide answers to the crucial question of the importance of effects (see also Bernard, 1994, in particular Chapter 6, pp. 70–80). From a standard Bayesian distribution, several much more informative statements can be derived; in somes cases, statements of the type "$Prob^{\star}(\delta > \delta_0) = \gamma$" may lead to the conclusion of a *large* effect, in some others, statements of the type "$Prob^{\star}(|\delta| < \epsilon) = \gamma$" will allow one to conclude that there is a *small* (or *negligible*) one.

6.5 Computational Aspects

On a practical level, how can we do all that was described in this chapter? Until very recently, implementing the Bayesian approach to inference posed serious technical difficulties for all but the most

elementary cases. These difficulties arose from the fact that the posterior distributions needed cannot generally be characterized analytically, and thus have to be evaluated by numerical means. Because of this, the application of Bayesian methods has long been restricted to cases for which the required computations were feasible in a reasonable amount of time on available computers. Fortunately these limitations have been pushed back, first by the tremendous increase in the power of micro-computers, but even more by the recent emergence of very efficient numerical approximation techniques.

In 1991, for the first French version of this book, we developed software that treated each case separately by some specific means. Most of the inferences presented in the present chapter were then handled by one of the `IBF2XK` and `IBFGT2` programs (Bernard, 1986; Poitevineau, Bernard, 1986)[39]. But for complex cases, today we prefer another method which we implemented in the `BayCat` software and which is based on the principle of *random sampling from the posterior distribution*. The advantage of this latter "Monte-Carlo" approach is not its numerical accuracy — though it has proved to perform quite nicely on this level —, but its generality and ease of implementation.

Elementary problems. Let us first summarize the computational needs in the elementary cases for which the Monte-Carlo approach is not necessary. All Bayesian inferences that were presented in the first sections of the chapter only involve known unidimensional distributions: *Beta-binomial* distributions for parametric inference from

39. `IBF2XK` provides inferences on one frequency ϕ and on the difference δ or ratio ρ between two frequencies for the following designs: $S \rightarrow U_2$ (one group of binary observations), $S \rightarrow U_K$ (one group of K-categorized observations), $S < G_2 > \rightarrow U_2$ (two groups of binary observations), $S < G_2 > \rightarrow U_K$ (two groups of K-categorized observations). `IBFGT2` provides inferences on user-defined linear combination between frequencies for a design of the type $S < G > \times T \rightarrow U_2$ (G groups of binary observations with T repeated measurements for each "subject" in set S). Each of these programs resorts to approximations by other distributions (scaled-*Beta* or scaled-F) when necessary.

binary data sampled without replacement from a finite population (Sections 6.1.3 and 6.1.4), *Beta* or F distributions for parametric inference from binary data sampled from an infinite population (Sections 6.1.5 and 6.1.7), and *Beta-binomial* distributions again for predictive inference from binary data (Section 6.2).

These distributions are all implemented for example in the basic Bayesian software `FirstBayes`[40]. They may also be found in most standard statistical or mathematical packages.

More complex problems. The Bayesian inferences discussed in the last sections of this chapter all involve parameters that follow *Dirichlet* distributions or parameters derived from them. Let us consider that the overall posterior distribution is given by $\boldsymbol{\phi} \sim Di(\boldsymbol{\alpha}')$, where $\boldsymbol{\alpha}'$ are the posterior strengths, and that we are interested in some derived parameter $g(\boldsymbol{\phi})$ where $g()$ may either be a numerical function or a logical function of the ϕ_k's. In this general setup, the following simple Monte-Carlo algorithm can be used:

Step 0. Set $i = 1$.

Step 1. Draw a random sample $\tilde{\boldsymbol{\varphi}}$ from the posterior distribution $Di(\boldsymbol{\alpha}')$. This can easily be done using the characterizations of the Dirichlet either in terms of independent Gamma distributions (see *e.g.* Fang, Kotz & Ng, 1990, p. 17) or in terms of independent Beta distributions (see *e.g.* Bernard, 1997).

Step 2. Calculate $\tilde{g}[i] = g(\tilde{\boldsymbol{\varphi}})$.

Step 3. Increment i. Repeat steps 1–2 as long as $i \leq I$.

Step 4. The vector $\tilde{g}[i], i = 1, \cdots, I$ provides an approximation to the distribution of $g(\boldsymbol{\phi})$, which can be summarized by any appropriate means: histogram, mean and variance, or quantiles.

40. `FirstBayes` was written by Anthony O'Hagan. Version 1.0, dated May 1994, and relevant documentation are available on the Internet at the "`http://www.maths.nott.ac.uk/personal/aoh/1b.html`" site.

It must be stressed that, on top of its simplicity, the above algorithm may be used for any function $g()$ so that it enables one to draw inferences on any derived parameter or property of interest, however complex it may be, without requiring sophisticated mathematics. The numerical accuracy of summaries of the distribution of $g(\boldsymbol{\phi})$ is controlled by the number of iterations I: the larger I, the more accurate the approximation[41].

Computation of ignorance-zone-based probability intervals. Of course, when prior ignorance is formalized by the ignorance zone, one needs to perform the above computation not for a single Dirichlet posterior distribution but for several ones; the set of distributions to consider is all $Di(\boldsymbol{\alpha}')$, with $\boldsymbol{\alpha}' = \boldsymbol{a} + \boldsymbol{\alpha}$, such that all α_k are > 0 and such that their sum ν is equal to 1. In principle, what is then required is to minimize/maximize the probability of the property of interest on $g(\boldsymbol{\phi})$, obtained from each $Di(\boldsymbol{\alpha}')$, with respect to $\boldsymbol{\alpha}$. In practice, this optimizing problem may be solved in an approximate way by only considering the K extreme vectors of prior strengths $\boldsymbol{\alpha}$, *i.e.* $\alpha_k = 1$ for some k and $\alpha_{k'} = 0$ for $k' \neq k$. This approximate solution was used for all examples in this chapter[42].

All figures of distributions and probability statements relative to the examples "Ordered data" (Section 6.3.1), "Mendel's data" and "Fractions data" (Section 6.3.4), "Aspirine data" (Section 6.4.1) and "Conflicting data" (Section 6.4.2) have been obtained by the above method with $I = 10^6$ iterations[43]. A remarkable fact is that all Bayesian guarantees and credibility limits given for these examples agree perfectly (up to the third decimal place) with results obtained in the 1991 French edition through more specific routines.

41. With this algorithm, the standard error on any probability value can easily be shown to be at most $1/(2\sqrt{I})$, *e.g.* 0.0005 for $I = 10^6$.
42. For a numerical derived parameter $g(\boldsymbol{\phi})$ and a property of the type $g(\boldsymbol{\phi}) > g_0$, this optimizing problem may be better approximated by first minimizing/maximizing $g(\frac{\boldsymbol{\alpha}'}{\nu'})$ (with $\nu' = n + \nu$) with respect to $\boldsymbol{\alpha}$, and then use each found solution for $\boldsymbol{\alpha}$ as a prior strengths vector.
43. We generally recommend to set I to at least 10^4. On a 80486-DX2/66-based computer, the computing time per thousand iterations is (very roughly) $0.1 \times K$ *seconds* where K denotes the number of categories.

The `BayCat` software implements the above Monte-Carlo algorithm for all the types of inferences described in relation to these examples[44]. This program, as well as the previous French-language `IBF2XK` and `IBFGT2`, are available from the author of this chapter.

6.6 Conclusions

We think that a quite general scientific methodology, as far as data analysis is concerned, is to *(i)* characterize the observed data by one or more properties of interest, *(ii)* attempt to generalize these properties to some future data of size n'. What we have tried to show in this chapter is that the Bayesian approach is particularly suited for fulfilling such a purpose, including the Bayesian predictive approach (generalization for small n') and the Bayesian parametric approach (generalization for large or infinite n'). Within this framework, point *(ii)* is answered by standard probability statements expressing the information brought by the data on the specific question of interest.

The two difficulties of the Bayesian approach to inference (its supposed subjectivity and its computational impracticability) are, in our opinion, perfectly dealt with, the former by the recourse to reference Bayesian distributions obtained from ignorance priors, the latter by efficient and quite general approximate algorithms. On the other hand, we have stressed several shortcomings of the frequentist approach which restrict its use to particular data structures or questions.

It is clear that several common questions of interest that may arise from the analysis of categorical data have not been mentioned at all in this chapter. Nevertheless, we think that the framework we have drawn here is quite general and may be applied to a variety of other

44. Strangely enough, though there are clear mentions of the use of the above algorithm (see *e.g.* Gelman, Carlin, Stern & Rubin, 1995, pp. 76–77 and pp. 481–482), we have not been able to find any widely available software for the inference on various derived parameters from a Dirichlet distribution. This is why we decided to develop our own software, `BayCat`.

questions. Our view is that the inductive step is straightforward within such a framework, so that the major issue in the analysis is to *ask the right question(s)*, *i.e.* to carefully design the relevant descriptive properties (based on relevant descriptive indices) that the inferential step will attempt to generalize.

To conclude, the Bayesian approach does not only appeal to us because it provides natural probabilistic statements, but also, on a very practical level, because it opens up a free and wide road to getting the data to answer *all* the questions that a researcher may need to ask.

References

Aitchison, J., Bacon-Shone, J. (1981), "Bayesian risk ratio analysis", *The American Statistician*, 35 n^o 4, pp. 254–257.

Aitchison, J., Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, Cambridge: Cambridge University Press.

Anderson, J. R. (1991), "The Adaptive Nature of Human Categorization", *Psychological Review*, 98 n^o 3, pp. 409–429.

Bernard, J.-M. (1983), *Inférence Bayésienne sur des Fréquences dans le Cas de Données Structurées: Méthodes Exactes et Approchées*, Thèse de doctorat de troisième cycle (unpublished), Université René Descartes, Paris.

Bernard, J.-M. (1986), "Méthodes d'Inférence Bayésienne sur des Fréquences", *Informatique et Sciences Humaines*, 68 pp. 89–133.

Bernard, J.-M. (1994), "Analyse Descriptive des Données planifiées", special issue of *Mathématiques, Informatique et Sciences Humaines*, 126, pp. 7–98.

Bernard, J.-M. (1996), "Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process", *The American Statistician*, 50 n^o 1, pp. 7–13.

Bernard, J.-M. (1997), "Bayesian Analysis of Tree-Structured Categorized Data", *Revue Internationale de Systémique*, 11 n^o 1, pp. 11–29.

Bernard, J.-M., Blancheteau, M. (1987), "Le Comportement Prédateur chez un Forficule, Euborellia Moesta (Géné), III. Analyse Séquentielle des Tentatives Infructueuses de Capture", *Biology of Behaviour*, 12, pp. 117–126.

Bernard, J.-M., Blancheteau, M., Rouanet H. (1985), "Le Comportement Prédateur chez un Forficule, Euborellia Moesta (Géné), II. Analyse Séquentielle au Moyen de Méthodes d'Inférence Bayésienne", *Biology of Behaviour*, 10, pp. 1–22.

Bernard, J.-M., Charron, C. (1996a), "L'Analyse Implicative Bayésienne: une méthode pour l'étude des dépendances orientées. II: Données binaires", *Mathématiques, Informatique et Sciences Humaines*, 134, pp. 5–38.

Bernard, J.-M., Charron, C. (1996b), "L'Analyse Implicative Bayésienne: une méthode pour l'étude des dépendances orientées. II: Modèle logique sur un tableau de contingence", *Mathématiques, Informatique et Sciences Humaines*, 135, pp. 5–18.

Bernardo, J. M., Smith, A. F. M. (1994), *Bayesian Theory*, Chichester: John Wiley & sons.

Berry, G., Armitage, P. (1995), "Mid-P confidence intervals: a brief review", *The Statistician*, 44 n^o 4, pp. 417-423.

Charron, C. (1996), "Categorization of Problems and Conceptualization of Fractions in Adolescents", *European Journal of Psychology of Education*, to appear.

Corroyer, D., Rouanet, H. (1994), "Sur l'Importance des Effets et ses Indicateurs dans l'Analyse Statistique des Données", *L'Année Psychologique*, 94, 607–624.

de Finetti, B. (1974/1975), *Theory of Probability*, Vols. 1 and 2, New York: John Wiley.

de Finetti, B. (1981), "Probabilités: Attention aux falsifications", *Revue d'Economie Politique*, 2, pp. 129–162.

Fang, K. T., Kotz, S., Ng, K. W. (1990), *Symetric Multivariate and Related Disributions*, New-York: Chapman and Hall.

Fisher, R. A. (1959), *Statistical Methods and Scientific Inference*, 2nd ed., London: Oliver and Boyd.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Monographs on Statistics and Applied Probability 55, New-York: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin D. B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Guilbaud, G.-Th. (1983), "Document du séminaire 1982-1983", Maison des Sciences de l'Homme, Paris.

Guttman, L. (1983), "What is not what in statistics?", *The Statistician*, 26, pp. 81–107.

Haldane, J. B. S. (1948), "The Precision of Observed Values of Small Frequencies," *Biometrika*, 35, pp. 297–300.

Hildebrand, D. K., Laing, J. D., Rosenthal, H. (1977), *Prediction Analysis of Cross Classifications*, New-York: Wiley.

Hoadley, B. (1969), "The compound multinomial distribution and Bayesian analysis of categorical data from finite population", *Journal of the American Statistical Association*, 64, pp. 216–229.

Jaynes, E. T. (1968), "Prior Probabilities", *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, September 1968, pp. 227–241.

Jaynes, E. T. (1976), "Confidence Intervals vs. Bayesian Intervals", in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W. L. Harper and C. A. Hooker, Editors, Dordrecht (Holland): D. Reidel Publishing Company.

Jeffreys, H. (1938/1961), *Theory of Probability*, 3rd ed., Oxford: Clarendon Press.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems", *Proceedings of the Royal Society of London*, Ser. A, 186, pp. 453–461.

Kass, R. E., Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules", *Journal of the American Statistical Association*, 91 n^o 435, pp. 1343–1370.

Laplace, P. S. (1825/1986), *Essai philosophique sur les probabilités*, re-edition of the 5th ed., Paris: C. Bourgois.

Lieberman, G. J., Owen D. B. (1961), *Tables of the hypergeometric probability distribution*, Stanford, CA: Stanford University Press.

Lindley, D. V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2 : Inference*, Cambridge: University Press.

Lindley, D. V., Phillips, L. D. (1976), "Inference for a Bernoulli Process (a Bayesian View)," *The American Statistician*, 30, pp. 112–119.

Loevinger, J. (1948), "The Technic of Homogeneous Tests Compared with some Aspects of Scale Analysis and Factor Analysis", *Psychological Bulletin*, 45, pp. 507–530.

Mosimann, J. E. (1962), "On the Compound Multinomial Distribution, the Multivariate β-distribution, and Correlations Among Proportions", *Biometrika*, 49 n^o 1–2, pp. 65–82.

Perks, F. J. A. (1947), "Some Observations on Inverse Probability Including a New Indifference Rule (with discussion)," *Journal of the Institute of Actuaries*, 73, pp. 285–334.

Poitevineau, J., Bernard, J.-M. (1986), "La série des programmes IBF", *Informatique et Sciences Humaines*, 68–69, pp. 135–137.

Raftery, A. E., Madigan, D., Volinsky, C. T. (1996), "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance", in *Bayesian Statistics V*, Bernardo J. M., Berger J. O., Dawid A. P. and Smith A. F. M. (eds.), Oxford: University Press, pp. 323–350.

Rouanet, H. (1996), "Bayesian Methods for Assessing Importance of Effects", *Psychological Bulletin*, 119, pp. 149–158.

Rouanet, H., Bernard, J.-M., Le Roux, B. (1990), *Statistiques en Sciences Humaines: Analyse Inductive des Données*, Paris: Dunod.

Rouanet, H., Bert, M.-P., Le Roux, B. (1987), *Statistiques en Sciences Humaines: Procédures Naturelles*, Paris: Dunod.

Steering Commitee of the Physicians' Health Study Research Group (1988), "Preliminary report: Findings from the aspirin component of the ongoing physicians' health study", *New England Journal of Medicine*, 318, pp. 262–264.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall.

Walley, P. (1996a), "Inferences from Multinomial Data: Learning about a Bag of Marbles", *Journal of the Royal Statistical Society*, Ser. B., 58, pp. 3–57.

Walley, P. (1996b), "Measures of Uncertainty in Expert Systems", *Artificial Intelligence*, 83, pp. 1–58.

Wilks, S. S. (1962), *Mathematical Statistics*, New York: John Wiley.