Chapter 4. Statistical Inference using Bayes's Theorem.

To this point, we have been concerned with probability models specified by hypothesis, and with learning how to calculate with these models and how to describe, summarize, and measure them. Statistical inference involves the use of such models in empirical science, but the approach is in some sense the reverse of that adopted up to now: instead of discussing the probabilities of different outcomes of an experiment under a specified hypothesis, we will suppose that we observe the outcome, and then attempt to verify the hypothesis or to choose among several possible hypotheses. Instead of reasoning from cause to effect we attempt the reverse, to go from effect to cause.

4.1 Bayes's Theorem.

The ideal approach to statistical inference is based upon an elementary theorem involving conditional probability. There is some disagreement about where this ideal falls between the two extremes of, on the one hand, being a Platonic ideal, unattainable in practice, or on the other hand, one that with experience becomes available for all problems. There is no disagreement about the truth and usefulness of the theorem, however.

The most elementary situation we will be concerned with can be described as follows: We have a list of possible causes or states of nature. We know the relative probabilities of these causes, and for each cause we know the probabilities of the various outcomes of an experiment. The experiment is performed, and we observe the outcome or effect, F. How do we assess the probabilities of the causes now, after the fact? To be more specific, let the causes $E_1, E_2, \ldots, E_n$ be a comprehensive and mutually exclusive list. We know their probabilities before the experiment, $P(E_1), P(E_2), \ldots, P(E_n)$. We call these *a priori* probabilities, for "before the fact," or "prior to the experiment." We also know the probabilities of the effect F for each cause; that is, we know $P(F|E_1), \ldots, P(F|E_n)$. We observe F as the result of the experiment, and we want to know the conditional probabilities of the E_i's given F, $P(E_1|F), \ldots, P(E_n|F)$. We call these *a posteriori* probabilities, for "after the fact" or "posterior to the experiment." Thus we *have* the probabilities $P(F|E_i)$ of the effect given the causes and we *want* the probabilities $P(E_i|F)$ of the causes given the effect.

Example 4.A. A Diagnostic Test. A patient is to be given a blood test that is designed to detect the presence of cancer cells. Suppose there are but two possibilities, two "causes" or states of the patient's body:

$$E_1 = \text{The patient has cancer,}$$

$$E_2 = \text{The patient does not have cancer.}$$

Let F represent the "effect" that the result of the test is positive, that the test indicates the presence of cancer cells. Now if the test is performed and effect F observed, both doctor and patient will be interested in the probability that the patient indeed has the cancer suggested by the cells' presence. The question is, what is $P(E_1|F)$?

We know the characteristics of the test; in extensive laboratory testing it was found that of patients with the cancer, 90% had a positive reaction to the test, while of those known on the basis of extensive examination not to have the cancer, 20% had a positive reaction. (On the basis of these numbers the test has been advertised as "90% reliable".) That is, we know

$$P(F|E_1) = .9,$$

$$P(F|E_2) = .2.$$

We are willing to consider the patient as randomly selected from a population with a prevalence of this type of cancer of θ_o; that is, we have

$$P(E_1) = \theta_o,$$

$$P(E_2) = 1 - \theta_o,$$

There is disagreement on the exact value of θ_o, though. Given this information, what is $P(E_1|F)$?

Example 4.B. Forensic Identification. A law court is confronted with a case of identification. A man is accused of a crime, and the sole evidence at issue is evidence from a tiny segment of DNA found at the crime scene that matched the accused's DNA. Here we have

$$E_1 = \text{accused left the DNA at the crime scene,}$$

$$E_2 = \text{accused was not the source of the crime scene DNA,}$$

$$F = \text{accused's DNA matches that from the crime scene.}$$

Suppose that it is assumed that the DNA is accurately typed, that is,

$$P(F|E_1) = 1,$$

but that particular DNA pattern is found in 0.3% of the population,

$$P(F|E_2) = 0.003.$$

Ignoring the question of selection (that is, how many potential suspects were examined before a match was found?), we want to calculate $P(E_1|F)$.

The framework in which we consider such questions is called Bayes's Theorem, after the Reverend Thomas Bayes, who died in 1761 leaving an essay among his papers that investigated the use of this theorem in statistical inference. The theorem itself is elementary, little more than a minor twist on our definition of conditional probability.

Bayes's Theorem (Elementary Version).

If $E_1, E_2, \ldots, E_n$ partition the sample space S (that is, they are mutually exclusive and exhaustive: $E_1 \cup E_2 \cup \cdots \cup E_n = S$), then for each E_i and any event F with $P(F) > 0$,

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{\sum_{j=1}^{n} P(E_j)P(F|E_j)}. \tag{4.1}$$

Proof: The general rule of multiplication (1.8) tells us that for any events E and F

$$P(E \cap F) = P(F)P(E|F).$$

Applying this twice we get

$$P(E_i \cap F) = P(F)P(E_i|F)$$

$$P(E_i \cap F) = P(E_i)P(F|E_i)$$

and, since $P(F) > 0$, equating these gives

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{P(F)}. \tag{4.2}$$

This is essentially the Theorem; all that is left is to show

$$P(F) = \sum_{j=1}^{n} P(E_j)P(F|E_j).$$

It will be instructive to do this in two different ways. First, since the E_i partition S we must have $\sum_{i=1}^{n} P(E_i|F) = 1$. But from (4.2)

$$1 = \sum_{i=1}^{n} P(E_i|F) = \sum_{i=1}^{n} \left[\frac{P(E_i)P(F|E_i)}{P(F)} \right]$$
$$= \frac{1}{P(F)} \sum_{i=1}^{n} P(E_i)P(F|E_i),$$

or

$$P(F) = \sum_{i=1}^{n} P(E_i)P(F|E_i).$$

Alternatively, we note that since the E_i partition S,

$$F = \bigcup_{i=1}^{n} (E_i \cap F).$$

Also, since the E_i are mutually exclusive, the $(E_i \cap F)$ are too, and finite additivity (1.5) gives

$$P(F) = \sum_{i=1}^{n} P(E_i \cap F).$$

But for each i, $P(E_i \cap F) = P(E_i)P(F|E_i)$, so

$$P(F) = \sum_{i=1}^{n} P(E_i)P(F|E_i).$$

We note that, as the first of these alternatives emphasizes, the denominator in (4.1), namely $P(F) = \sum_{i=1}^{n} P(E_i)P(F|E_i)$, is playing the role of a normalizing constant; it is exactly the factor needed to ensure that $\sum_{i=1}^{n} P(E_i|F) = 1$.

Example 4.A (continued). We have

$$P(F|E_1) = .9$$
$$P(F|E_2) = .2$$

and the prevalence of cancer is

$$P(E_1) = \theta_o.$$

From (4.1) we have

$$P(E_1|F) = \frac{\theta_o(.9)}{\theta_o(.9) + (1 - \theta_o)(.2)}.$$

(It is worth reiterating that the denominator is a normalizing factor: from (4.2),

$$P(E_1|F) = \frac{\theta_o(.9)}{P(F)}$$

and

$$P(E_2|F) = \frac{(1 - \theta_o)(.2)}{P(F)};$$

requiring that these sum to 1 gives

$$P(F) = \theta_o(.9) + (1 - \theta_o)(.2).$$

This result makes it clear just how $P(E_1|F)$ depends upon the prevalence θ_o: We have

$$\begin{aligned} P(E_1|F) &= .82 \quad \text{if } \theta_o = .5 \\ &= .53 \quad \text{if } \theta_o = .2 \\ &= .19 \quad \text{if } \theta_o = .05 \\ &= .043 \quad \text{if } \theta_o = .01 \\ &= .0045 \quad \text{if } \theta_o = .001 \end{aligned}$$

[Figure 4.1]

If the positive reaction is considered in light of a high prevalence, if the patient is considered as a member of a high risk category, we find it likely that the patient has cancer. If the prevalence is low, if the patient is considered as a member of the general population, cancer remains unlikely, despite the positive reaction from a "90% reliable" test. The question, what is $P(E_1|F)$, evidently hinges upon determining the relevant prevalence, a difficult matter.

Example 4.B (Continued). Here $P(F|E_1) = 1$, and $P(F|E_2) = .003$. If $P(E_1)$ is the a priori (before the DNA test) probability the accused is guilty,

$$P(E_1|F) = \frac{P(E_1)}{P(E_1) + (1 - P(E_1))(.003)}.$$

If we are a priori certain of guilt or innocence, the test result does not change our mind. If $P(E_1) = 1/2$, then the test result increases our assessment of the probability of guilt to .994. If $P(E_1) = 1/20$, then the test increased our assessment to .943.

In neither of Examples 4.A and 4.B do we obtain a definitive answer without $P(E_1)$, but in both examples Bayes's Theorem shows how to incorporate the test result with the a priori probability $P(E_1)$. In the ideal situation where $P(E_1)$ is available, Bayes's Theorem gives the ideal answer, $P(E_1|F)$.

4.2 Bayes's Theorem More Generally.

Essentially the same argument that established the elementary version of Bayes's Theorem works more generally. We give it here for the continuous case, although it works as well for the discrete case or for the mixed discrete-continuous case. Suppose we have available the marginal distribution of a random variable Y, namely its density $f_Y(y)$, and the conditional distributions of X given Y, the conditional density $f(x|y)$. If we wish to find the conditional distribution of Y given X, $f(y|x)$, we proceed as follows. First, if $f_X(x) > 0$ we have (from (3.14))

$$f(y|x) = \frac{f(x,y)}{f_X(x)}.$$

Second, we know that

$$f(x,y) = f(x|y)f_Y(y). \tag{4.3}$$

Together these give

$$f(y|x) = \frac{f(x|y)f_Y(y)}{f_X(x)}. \tag{4.4}$$

All that remains is to find $f_X(x)$. As in the elementary case this can be done in either of two ways. First, we must have

$$\int_{-\infty}^{\infty} \frac{f(x|u)f_Y(u)du}{f_X(x)} = 1$$

or

$$f_X(x) = \int_{-\infty}^{\infty} f(x|u) f_Y(u) du. \tag{4.5}$$

This approach emphasizes that in (4.4) where x is considered as fixed, $f_X(x)$ is simply a normalizing factor, present to guarantee that $f(y|x)$ is a proper density in y, that it integrates to 1. Alternatively, (4.5) would follow directly from (4.3) and the definition of the marginal density $f_X(x)$ given by (3.12).

The equations (4.4) and (4.5) together give us the *general version of Bayes's Theorem* for the continuous case;

$$f(y|x) = \frac{f(x|y) f_Y(y)}{\int_{-\infty}^{\infty} f(x|u) f_Y(u) du}, \tag{4.6}$$

as long as the expression on the right is not $0/0$ (in which case $f(y|x)$ remains undefined). We reiterate that this is a distribution of Y given $X = x$; that is, for a fixed x this density gives the conditional density of Y, and the denominator in (4.6) (which depends only on the fixed x, not on y) is only there as a normalizing factor, present so the density integrates to 1. To emphasize this we will sometimes write (4.6) as

$$f(y|x) \propto f(x|y) f_Y(y), \tag{4.7}$$

meaning "$f(y|x)$ is proportional to the product $f(x|y)f_Y(y)$, though the constant of proportionality can depend upon x ." The "constant of proportionality" alluded to is just $1/\int_{-\infty}^{\infty} f(x|u) f_Y(u) du$. This way of looking at $f(y|x)$ will be of use later, where we find that we can often evaluate the product $f(x|y)f_Y(y)$ and recognize its form as a function of y, thereby avoiding the necessity of evaluating the integral in (4.6).

The discrete counterparts of (4.6) and (4.7) are

$$p(y|x) = \frac{p(x|y) p_Y(y)}{\sum\limits_{\text{all } k} p(x|k) p_Y(k)}. \tag{4.8}$$

$$p(y|x) \propto p(x|y) p_Y(y). \tag{4.9}$$

4.3 Inference about the Binomial Distribution.

The arguments of the previous section can be applied more generally, to the mixed case, and this will permit our first significant attack upon a basic problem of statistical inference. For the case where X is discrete and Y is continuous, Bayes's Theorem becomes

$$f(y|x) = \frac{p(x|y) f_Y(y)}{\int_{-\infty}^{\infty} p(x|u) f_Y(u) du}, \tag{4.10}$$

or alternatively,

$$f(y|x) \propto p(x|y) f_Y(y). \tag{4.11}$$

To see how this simple consequence of the definition of conditional probability can be developed into a powerful (and powerfully debated) technique that is sometimes referred to as *Bayesian Inference,* we consider an example.

Example 4.C. A Survey. A polling organization wishes to determine the fraction of registered Chicago Democrats in favor of the incumbent Governor, two weeks before the primary election. They select $n = 100$ names from the list of about a million registered Democrats to be interviewed. Assuming (which in Chicago is unlikely) that all n are interviewed and express an opinion, the poll will result in a count "X" for the incumbent and a count "$n - X$" against.

Formally, we could let

$$\theta = \text{fraction of Chicago Democrats for the incumbent,}$$

and if the selection of $n = 100$ for the survey is truly random, we would accept that given θ, (at least to a good approximation) X has a Binomial $(100, \theta)$ distribution:

$$\begin{aligned} p(x|\theta) &= b(x; n, \theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

But we are not "given θ;" in fact the object of the survey is to determine θ. We are uncertain about θ, and we express our uncertainty in the form of a probability distribution, $f(\theta)$, by treating θ as a random variable. But which distribution? For now, suppose we agree that our uncertainty about θ is well-represented as a Uniform $(0, 1)$ distribution,

$$\begin{aligned} f(\theta) &= 1 \quad \text{for } 0 < \theta < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

That is, θ is as likely to be above $\frac{1}{2}$ as not, as likely to be between $\frac{1}{4}$ and $\frac{3}{4}$ as not, and so forth.

Accepting this hypothesis for the time being, the problem of assessing our uncertainty about θ in the light of the outcome of the survey becomes a straightforward application of Bayes's Theorem. We observe $n = 100$, $X = 40$, say, and (4.11) gives us $f(\theta|x) \propto f(\theta)p(x|\theta)$, so the density of θ becomes

$$\begin{aligned} f(\theta|40) &\propto p(40|\theta) \cdot f(\theta) \\ &= \binom{100}{40} \theta^{40}(1-\theta)^{60} \cdot 1 \\ &= \binom{100}{40} \theta^{40}(1-\theta)^{60} \quad \text{for } 0 < \theta < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Now, this represents the density of θ given $X = 40$ up to a constant of proportionality, and we could evaluate that constant if we wish (it is the reciprocal of $\int_0^1 \binom{100}{40} \theta^{40}(1-\theta)^{60} d\theta$). But we recognize that the only portion of the density that depends upon θ, namely $\theta^{40}(1-\theta)^{60}$, is exactly equal to the variable part of a Beta $(41, 61)$ density (given by (2.10)). Since the variable parts agree and both integrate to 1, the constants must agree, and we conclude that $f(\theta|40)$ *is* a Beta $(41, 61)$ density.

To summarize: Our hypotheses were that

(a) θ is Uniform $(0, 1)$ a priori (before the survey)

(b) Given θ, X is Binomial $(100, \theta)$.

Our conclusion is that

(c) θ is Beta $(41, 61)$ a posteriori (after the survey), given that $X = 40$.

Figure 4.3 illustrates how the result of the survey changed our assessment of the uncertainty about the fraction of Chicago's Democrats for the incumbent. Bayes's theorem processes the survey data and alters our assessment of θ.

If we accept the hypotheses, there is no doubt about the conclusion. The second, (b), can be defended on the basis of the random selection from the voters list. But what of the first? Does a Uniform $(0, 1)$ distribution represent our pre-survey uncertainty about the fraction of Chicago's Democrats for the incumbent? To help understand the issue, let us reconsider Example 3.D., Bayes's Billiard Table. There we had the same mathematical structure, but with Y in place of θ: Y, the distance of the first ball from the left end of the table is random, and given $Y = y$, X is Binomial (n, y). Asking for a post-survey assessment of the

uncertainty about the fraction Y of Democrats for the incumbent is mathematically like asking the following question: Suppose we are not told the location of Y, we are only told that of $n = 100$ rolls of the second ball, $X = 40$ came to rest to the left of Y. What is the conditional distribution of Y? There the answer is clear, namely a Beta $(41, 61)$ distribution. Is the analogy a proper one, and is the specification of an a priori Uniform $(0, 1)$ distribution equally reasonable in the two cases?

In the Billiard Table example, $f_Y(y)$ represents a tangible model; a physical randomization is performed, and if the initial velocity of the first ball is sufficiently uncontrolled, we find it easy to obtain general consent that a Uniform $(0, 1)$ distribution represents at least approximately our a priori uncertainty about the position of Y. In the Survey example, $f(\theta)$ represents a less tangible process; it is an assessment of uncertainty about a process we understand less well than billiard tables. Is the political history of Chicago like rolling a ball across a flat table? Perhaps, but the lack of a common understanding of the process creates conceptual difficulties. Different social scientists may bring different degrees of past experience, of knowledge of the political process, to the study. There may well be no commonly acceptable a priori distribution $f(\theta)$ for the Survey example. That means that different scientists would be led to different conclusions. This need not be troubling, since different sets of past experiences, different frames of reference almost invariably lead to different perspectives, but it has led to disagreement about how it is best to present the analysis of situations such as the Survey example. Some scientists would prefer to isolate and present only that part of the analysis that can be commonly agreed upon, as a less personal scientific statement. We shall discuss ways of doing this later, as well as other aspects of what is a rather deep and occasionally philosophically divisive question, but first we look at ways of enriching the structure while keeping the same approach.

4.4 A Richer Class of Models for Binomial Inference.

One potentially significant limitation of the analysis of the previous section can be easily overcome; it is the restrictive hypothesis that the fraction θ is a priori distributed as Uniform $(0, 1)$. Bayes's Theorem, either as (4.10) or (4.11), can be applied to any specification of $f(\theta)$, and this flexibility can be exploited to represent any degree of uncertainty based upon prior experience that can be summarized as a density. A particularly rich class of possibilities for $f(\theta)$ that permits a simple analysis is the Beta (α, β) Distributions, already encountered in Chapter 2 (Example 2.E):

$$\begin{aligned} f(\theta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{4.12}$$

This class includes the Uniform $(0, 1)$ as a special case, with $\alpha = \beta = 1$.

If we accept a particular $f(\theta)$ from this class as an a priori density for θ, the analysis is straightforward. Subsequent to observing the number of successes $X = x$ in n independent trials, the conditional density of θ given $X = x$ is, from (4.11),

$$\begin{aligned} f(\theta|x) &\propto p(x|\theta)f(\theta) \\ &= \binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= C\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

Where C depends on n, x, α, and β, but not on θ. Just as in the previous section we recognize that the portion of $f(\theta|x)$ that depends upon θ, namely $\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$, is (but for the lack of the normalizing constant) a Beta $(x+\alpha, n-x+\beta)$ density. Thus $f(\theta|x)$ must be a Beta $(x+\alpha, n-x+\beta)$ density,

$$\begin{aligned} f(\theta|x) &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}, \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{4.13}$$

(Alternatively, the integral in (4.10) can be evaluated, using (2.12) and (2.13) and after some algebraic cancellation we also arrive at (4.13)).

The specification of an a priori density $f(\theta)$ from within the class (4.12) can be simplified by using three facts about the Beta (α, β) distributions. First, if θ has density (4.12),

$$\mu_\theta = \frac{\alpha}{\alpha + \beta}, \tag{4.14}$$

and

$$\begin{aligned} \sigma_\theta^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}. \end{aligned} \tag{4.15}$$

The third fact is that, if α and β are not very small, $f_\theta(y)$ is approximately like the density of a Normal $(\mu_\theta, \sigma_\theta^2)$ distribution, and in particular,

$$P(|Y - \mu_\theta| < \sigma_\theta) \approx 2/3.$$

Table 4.1 illustrates the degree of approximation, which, while crude, is adequate for present purposes.

α	β	$P(\lvert Y - \mu_\theta\rvert < \sigma_\theta)$
1	1	.577
1	2	.629
2	2	.626
1	3	.668
2	3	.640
1	4	.697
3	3	.644
1	9	.768
2	8	.682
3	7	.666
4	6	.661
5	5	.659
1	19	.812
2	18	.708
4	16	.681
6	14	.674
8	12	.671
10	10	.671

Table 4.1. Illustration of the adequacy of the approximation $P(|Y - \mu_\theta| < \sigma_\theta) \simeq 2/3$ for various beta densities. The approximation is worse for highly skewed densities.

To illustrate the use of these three facts, let us return to our example.

Example 4.C (Continued). Before conducting the survey, we specify $f(\theta)$ by asking, first, what is the expectation of θ, namely μ_θ? If α and β are not very small, $f(\theta)$ is close to being symmetric about μ_θ, and μ_θ is nearly equal to the median of $f(\theta)$. In that case we could as well ask, for what value is it even odds that θ is above it or below it? Suppose we agree that $\mu_\theta = .5$ answers these questions, at least approximately; that is, prior to the survey we consider θ equally likely to be above or below .5. Next we ask, for what interval of values of θ around $\mu_\theta = .5$ is the probability 2/3? For what value of σ_θ is it twice as likely that $.5 - \sigma_\theta < \theta < .5 + \sigma_\theta$ as not? Suppose we agree that a priori,

$$P(.4 < \theta < .6) \approx 2/3,$$

or $\sigma_\theta \approx .1$. Now (4.15) gives us

$$\sigma_\theta^2 = \mu_\theta(1-\mu_\theta)/(\alpha+\beta+1);$$

if $\mu_\theta = .5$ and $\sigma_\theta^2 = (.1)^2$, this gives us

$$\begin{aligned}\alpha+\beta+1 &= \mu_\theta(1-\mu_\theta)/\sigma_\theta^2 \\ &= (.5)^2/(.1)^2 = 25,\end{aligned}$$

or $\alpha+\beta = 24$. Together with $\mu_\theta = .5$ or

$$\frac{\alpha}{\alpha+\beta} = .5,$$

we have $\alpha = \beta = 12$.

This suggests that within the class (4.12) of Beta (α, β) distributions, we should take as our a priori distribution for θ the Beta (12, 12) density,

$$\begin{aligned}f(\theta) =& \frac{23!}{11!11!}\theta^{11}(1-\theta)^{11} \quad \text{for } 0<\theta<1 \\ =& 0 \quad \text{otherwise.}\end{aligned}$$

This distribution has $\mu_\theta = .5$, $\sigma_\theta^2 = (.1)^2$, and a precise evaluation finds

$$P(.4 < \theta < .6) = .6727 \approx 2/3.$$

After the survey, the distribution that describes the uncertainty about θ given $n = 100$ and $X = 40$ is then, from (4.13), the Beta (52, 72) density

$$\begin{aligned}f(\theta|40) &= \frac{123!}{51!71!}\theta^{51}(1-\theta)^{71}, \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

[Figure 4.4]

Figure 4.4 illustrates how Bayes's Theorem works here to incorporate the survey data with the a priori specification to produce the a posteriori distribution.

The effect of the sample information X upon the distribution of θ can be neatly summarized by the changes in expectation and variance. *Before* the data X is observed, if θ has density (4.12),

$$E(\theta) = \frac{\alpha}{\alpha+\beta} = \mu_\theta \tag{4.14}$$

$$\text{Var}(\theta) = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}. \tag{4.15}$$

After we observe $X = x$,

$$\begin{aligned}E(\theta|X=x) &= \frac{\alpha+x}{\alpha+\beta+n} \\ &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}\end{aligned} \tag{4.16}$$

$$\text{Var}(\theta|X=x) = \frac{E(\theta|X=x)(1-E(\theta|X=x))}{\alpha+\beta+n+1}. \tag{4.17}$$

Formula (4.16) is particularly informative; it gives $E(\theta|X=x)$ as a weighted average of the prior expectation μ_θ and the fraction of the sample that are successes; that is, $E(\theta|X=x)$ is a compromise between μ_θ, our

expectation with no data, and X/n, the sample fraction. The larger $\frac{n}{\alpha+\beta+n}$ is, the more weight we put on X/n, the sample fraction.

The size of $\frac{n}{\alpha+\beta+n}$ depends upon the relative sizes of n and $\alpha+\beta$. This suggests an interpretation of $\alpha+\beta$: it may be thought of as the amount of prior information summarized by $f(\theta)$, while n is the amount of information (number of trials) in the sample. When n is so large as to dwarf $\alpha+\beta$, $E(\theta|X=x) \approx \frac{X}{n}$, and the prior distribution $f(\theta)$ has little impact. When $\alpha+\beta$ is so large as to dwarf n, $E(\theta|X=x) \approx \mu_\theta$, and the sample information has little impact on the prior expectation. This interpretation extends to the variances, which, as measures of dispersion may be considered as measures of uncertainty. Then (4.15) says that, holding μ_θ constant, the prior uncertainty decreases inversely as the amount of information summarized by the prior increases. And (4.17) says that, holding $E(\theta|X=x)$ constant, the posterior uncertainty decreases inversely as the total information (prior plus sample, $\alpha+\beta$ plus n) increases.

Example 4.C (Continued). For the Survey example we have

$$E(\theta) = .5$$

$$\alpha + \beta = 24$$

$$n = 100$$

$$\frac{X}{n} = .4$$

and

$$\frac{n}{\alpha+\beta+n} = \frac{100}{124} = .806.$$

Thus the posterior expectation is

$$\begin{aligned} E(\theta|X=40) &= (.5)(.194) + (.4)(.806) \\ &= .419 \end{aligned}$$

a compromise between .5 and .4 that weights .4 heavily, because the amount of sample information n outweighs the amount of prior information $\alpha+\beta$ by about 4 to 1.

Speaking of $\alpha+\beta$ as the "amount of prior information," in effect equating the information in our $f(\theta)$ with $\alpha+\beta=24$ to the information in a survey sample of size $n=24$, can be motivated by considering the posterior density (4.13), a Beta $(x+\alpha, n-x+\beta)$ density, with expectation (4.16) and variance (4.17), where we see a direct trade off. For example, the following two situations give exactly the same $f(\theta|x)$.

(i) A Uniform $(0,1)$ prior, which is Beta with $\alpha=\beta=1$, and a sample of $n=10$ with $X=5$ successes.

(ii) A Beta $(5,5)$ prior, and a sample of $n=2$ with $X=1$ success.

In both cases, $\alpha+\beta+n=12$, $x+\alpha=6$, and $n-x+\beta=6$. An increase of $\alpha+\beta$ by 8 has exactly compensated for a decrease of n by 8, as long as the ratios of α to β and x to $n-x$ remain equal. In this sense a Beta $(5,5)$ prior summarizes past experience equivalent to a survey of size $n=10$ with $X=5$ successes. Note that there is no Beta density equivalent to "no information," for (4.12) does not represent a density (its integral diverges) if α or β is 0.

4.5 Bayesian Inference for the Normal Distribution.

The principles of applying Bayes's Theorem to problems of inference remain the same in other situations. Given an a priori distribution $f(\theta)$ for θ and the conditional distributions $f(x|\theta)$, (4.6) or (4.7) permits us to reverse our point of view and find $f(\theta|x)$. Given prior information about the relative uncertainties of the causes or states of nature θ and the probability of the data X given each possible cause θ, we find the posterior probabilities of the causes given the data. We illustrate for an example involving the normal distributions.

Example 4.D (Weighing with an Imperfect Scale). We wish to weigh an object, say a sack of nails, using a reasonably accurate scale. By "reasonably accurate," we mean that, while we have adjusted the scale so that for objects of this general character it will be correct on average (in a large number of weighings), it may, in any single instance, err by an unpredictable amount, due to temperature, humidity, where the object is placed on the scale, etc. We formalize this by letting

$\theta =$ the object's true weight, as would be found from a much more accurate scale

$X =$ the object's weight as recorded from single weighing of the "reasonably accurate" scale

We suppose that the errors made by the scale are approximately normally distributed, with standard deviation τ lb.; that is, $f(x|\theta)$ is a Normal (θ, τ^2) density,

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\tau}\, e^{-\frac{(x-\theta)^2}{2\tau^2}} \quad \text{for } -\infty < x < \infty$$

[Figure 4.5]

Our goal is to express the uncertainty about θ after observing a single reading X of the scale; we wish to find $f(\theta|x)$.

Bayes's Theorem requires the a priori distribution of θ, our "before weighing" uncertainty about the true weight. Suppose we are willing to choose $f(\theta)$ from the class of Normal densities, $N(\mu, \sigma^2)$, a fairly rich class of possibilities. What values should we give to μ and σ^2? Suppose the sack has "100 lbs." printed on it, and we believe the shipper aimed at that goal; we take $\mu = 100$. But past experience has shown considerable variability in actual weights. In fact, we think

$$P(|\theta - 100| < 8) \approx 2/3.$$

Since for a $N(\mu, \sigma^2)$ distribution,

$$P(|\theta - \mu| < \sigma) = .6826 \approx 2/3,$$

we take $\sigma = 8$ or $\sigma^2 = 64$;

[Figure 4.6]

$$f(\theta) = \frac{1}{\sqrt{2\pi}8} e^{-\frac{(\theta-100)^2}{128}} \quad \text{for } -\infty < \theta < \infty.$$

Given the "data" $X = x$ we wish $f(\theta|x)$; we apply Bayes's Theorem in the form (4.7):

$$\begin{aligned} f(\theta|x) &\propto f(\theta) f(x|\theta) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(\theta-\mu)^2}{\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2}\frac{(x-\theta)^2}{\tau^2}} \\ &= \frac{1}{2\pi\sigma\tau} e^{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\sigma^2} + \frac{(x-\theta)^2}{\tau^2}\right]} \end{aligned}$$

Now, expanding and completing the square, we find

$$\begin{aligned}\left[\frac{(\theta-\mu)^2}{\sigma^2}+\frac{(x-\theta)^2}{\tau^2}\right] &= \theta^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)-2\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\theta+\frac{\mu^2}{\sigma^2}+\frac{x^2}{\tau^2}\\ &= \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta^2-2\left(\frac{\tau^2\mu+\sigma^2 x}{\tau^2+\sigma^2}\right)\theta+\left(\frac{\tau^2\mu+\sigma^2 x}{\tau^2+\sigma^2}\right)^2\right]+C,\\ &= \frac{(\theta-A)^2}{B^2}+C\end{aligned}$$

where

$$A=\frac{\tau^2\mu+\sigma^2 x}{\tau^2+\sigma^2}$$

and

$$B^2=\frac{\tau^2\sigma^2}{\tau^2+\sigma^2},$$

and C depends upon μ, σ^2, τ^2 and x. Then

$$f(\theta|x)\propto\frac{1}{2\pi\sigma\tau}e^{-\frac{(\theta-A)^2}{2B^2}}.$$

But this means that

$$f(\theta|x)\propto e^{-\frac{(\theta-A)^2}{2B^2}}\quad\text{for } -\infty<\theta<\infty,$$

But we recognize this as, except for the unidentified normalizing constant, a Normal (A,B^2) density. This means that

$$f(\theta|x)=\frac{1}{\sqrt{2\pi}B}e^{-\frac{(\theta-A)^2}{2B^2}}\quad\text{for } -\infty<\theta<\infty.$$

(Alternatively, we could have evaluated the integral in (4.6) and applied Bayes's Theorem in that form, of course arriving at the same result.) What is more, we know that A and B^2 are respectively the expectation and variance of $f(\theta|x)$, and so

$$\begin{aligned}E(\theta|X=x) &= \frac{\tau^2\mu+\sigma^2 x}{\tau^2+\sigma^2}\\ &= \left(\frac{\tau^2}{\tau^2+\sigma^2}\right)\mu+\left(\frac{\sigma^2}{\tau^2+\sigma^2}\right)x,\\ \mathrm{Var}(\theta|X=x) &= \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\end{aligned}$$

Just as in the case of inference about the Binomial, the expectation of the a posteriori distribution of θ is a weighted average of the a priori expectation μ and the data x (the result of the single weighing), with weights $\frac{\tau^2}{\tau^2+\sigma^2}$ on μ and $\frac{\sigma^2}{\tau^2+\sigma^2}$ on x. The larger our a priori uncertainty σ^2, the more weight we put on the single weighing x. If $\mu=100$, $\sigma^2=8^2=64$, $\tau^2=1$, and $x=94$ lbs.,

$$E(\theta|X=x)=\frac{1}{65}\cdot 100+\frac{64}{65}\cdot 94=94.1\text{ lbs.}.$$

Had we had $\sigma^2=2^2=4$, we would have

$$E(\theta|X=x)=\frac{1}{5}\cdot 100+\frac{4}{5}\cdot 94=95.2\text{ lbs.}.$$

The point of compromise between what we see (x) and what we expected (μ) depends upon the relative dispersions of our prior distribution (σ^2) and the conditional distribution of X given θ (τ^2). Our a posteriori

uncertainty ($\tau^2\sigma^2/(\tau^2+\sigma^2)$) is always less than the uncertainty about the scale, but if we have little prior information about θ (that is, if σ^2 is large) it is not much less. If we are a priori certain that $\theta = \mu$ (that is, if $\sigma^2 = 0$), then we are also a posteriori certain ($E(\theta|X = x) = \mu$, $\text{Var}(\theta|X = x) = 0$), regardless of x.

If we agree that uncertainty, such as a priori uncertainty about θ in the past examples, should be expressible as a probability distribution, and we are able to specify that distribution as well as the possible conditional distributions of the data, $f(x|\theta)$, then Bayes's Theorem gives the solution to the inference problem. Some scientists deny the appropriateness of expressing all uncertainty in this form; they might argue, for example, that uncertainty about the "fraction of Chicago's Democrats for the incumbent two weeks before the primary" is not susceptible to interpretation as a long-run relative frequency. Others would not challenge the usefulness of the idea of the a priori distribution $f(\theta)$, but they would be unable or unwilling to agree on a specific distribution. Others would not share these qualms when considering simple problems, but they would be unable to come to grips with choosing $f(\theta)$ for complicated higher dimensional problems. And still others would fully accept the "Bayesian" approach, but would look elsewhere for ways of convincingly communicating their work to others who do not. Most of the remainder of this text will be concerned with such methods, although even there we will find the guidance given by the ideal case to be invaluable.

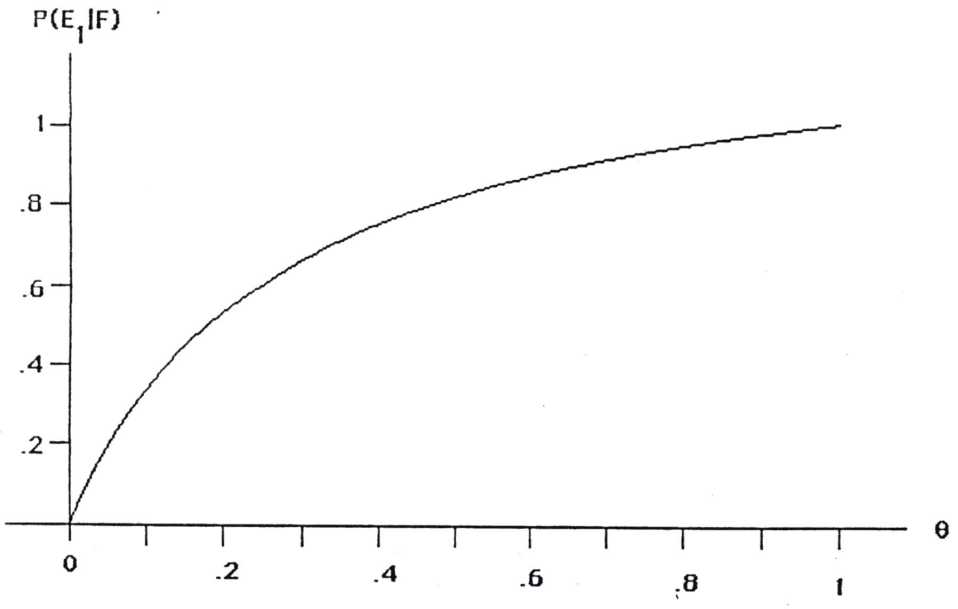
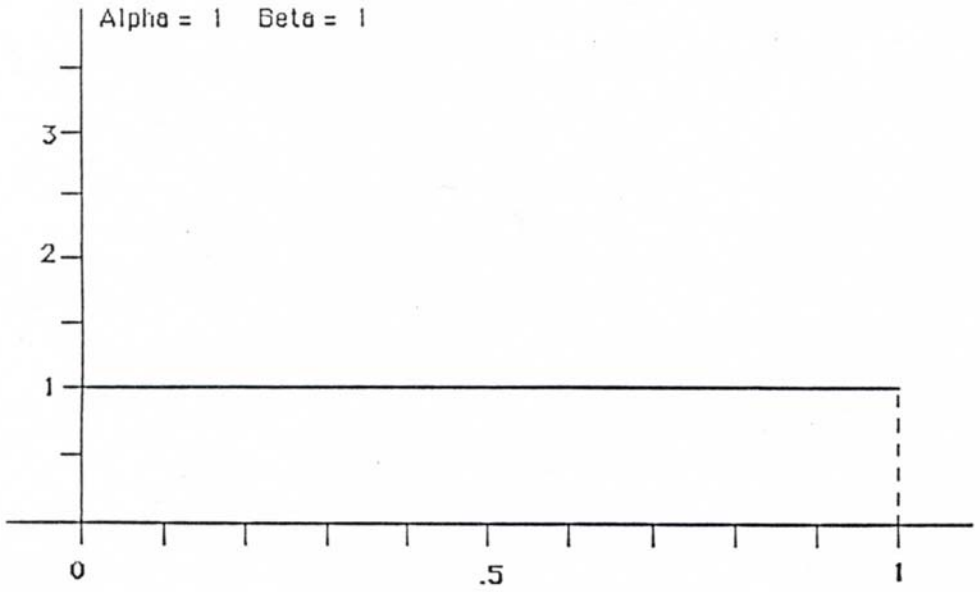


Figure 4.1

(a) Before Survey:



(b) After Survey:

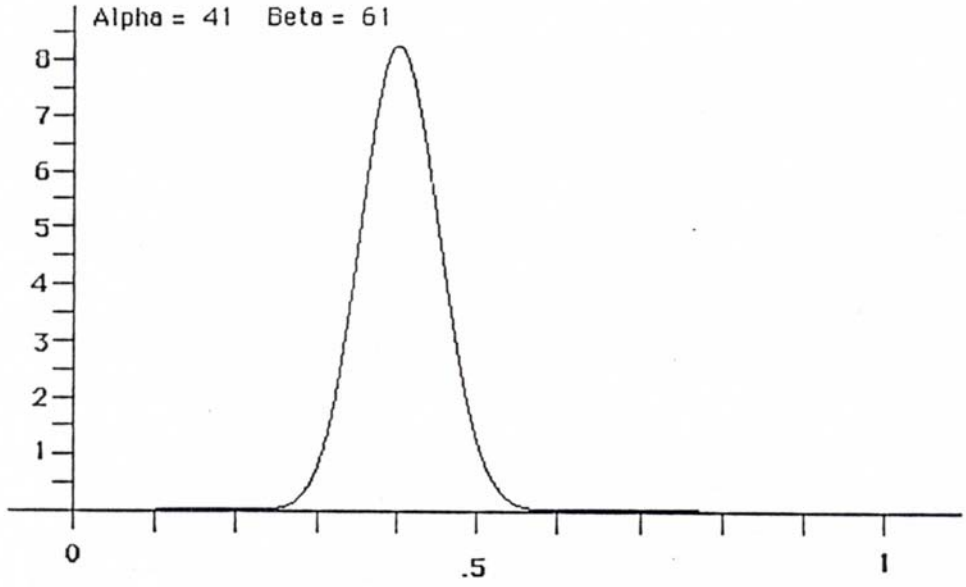


Figure 4.3

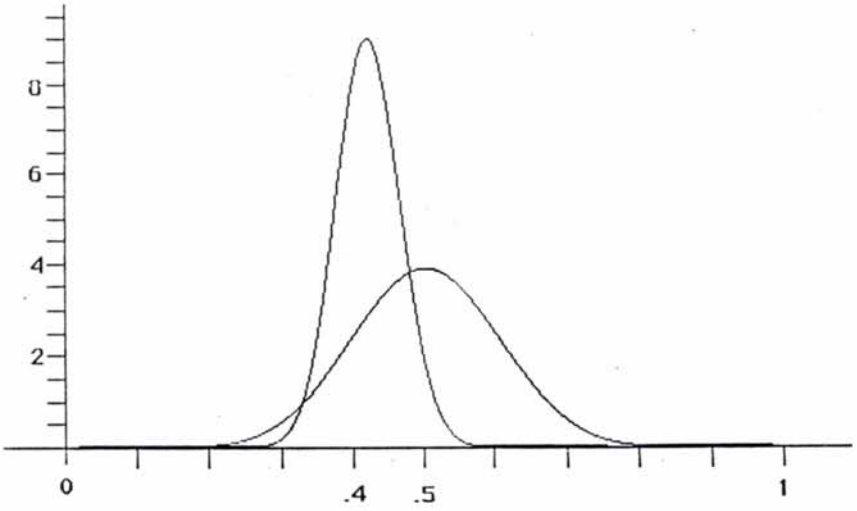


Figure 4.4

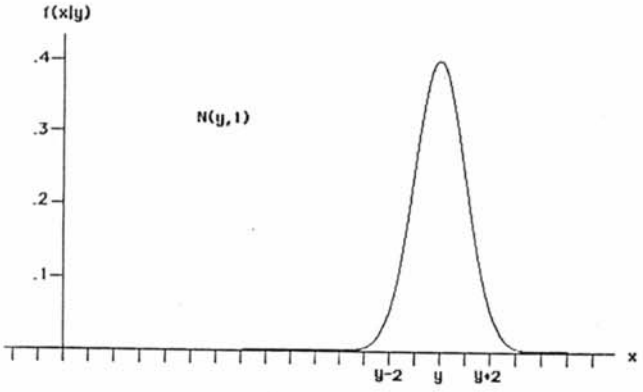


Figure 4.5

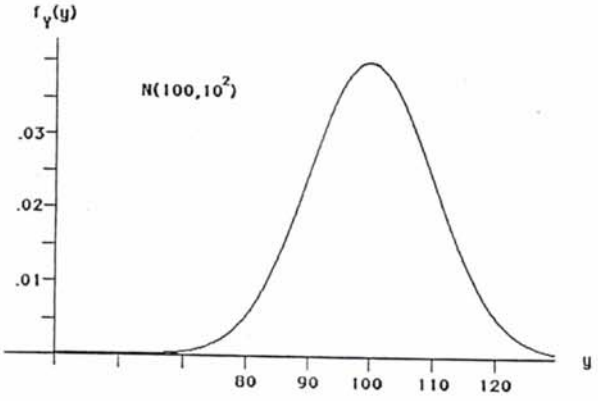


Figure 4.6