CHAPTER 6

ELEMENTARY PARAMETER ESTIMATION

> "*A distinction without a difference has been introduced by certain writers who distinguish 'Point estimation', meaning some process of arriving at an estimate without regard to its precision, from 'Interval estimation' in which the precision of the estimate is to some extent taken into account.*" — R. A. Fisher (1956)

Probability theory as logic agrees with Fisher in spirit; that is, it gives us automatically both point and interval estimates from a single calculation. The distinction commonly made between hypothesis testing and parameter estimation is considerably greater than that which concerned Fisher; yet it too is, from our point of view, not a real difference. When we have only a small number of discrete hypotheses $\{H_1 \cdots H_n\}$ to consider, we usually want to pick out a specific one of them as the most likely in that set, in the light of the prior information and data. The cases $n = 2$ and $n = 3$ were examined in some detail in Chapter 4, and larger n is in principle a straightforward and rather obvious generalization.

However, when the hypotheses become very numerous, a different approach seems called for. A set of discrete hypotheses can always be classified by assigning one or more numerical indices which identify them, as in H_t, $1 \le t \le n$, and if the hypotheses are very numerous one can hardly avoid doing this. Then deciding between the hypotheses H_t and estimating the index t are practically the same thing, and it is a small step to regard the index, rather than the hypotheses, as the quantity of interest; then we are doing parameter estimation. We consider first the case where the index remains discrete.

Inversion of the Urn Distributions

In Chapter 3 we studied a variety of sampling distributions that arise in drawing from an Urn. There the number N of balls in the Urn, and the number R of red balls and $N - R$ white ones, were considered known in the statement of the problem, and we were to make "pre–data" inferences about what kind of mix of r red, $n - r$ white we were likely to get on drawing n of them. Now we want to invert this problem, in the way envisaged by Bayes and Laplace, to the "post–data" problem: the data $D \equiv (n, r)$ are known but the contents (N, R) of the Urn are not. From the data and our prior information about what is in the Urn, what can we infer about its true contents? It is probably safe to say that every worker in probability theory is surprised by the results – almost trivial mathematically, yet deep and unexpected conceptually – that one finds in this inversion. In the following we note some of the surprises already well known in the literature, and add to them.

We found before [Eq. (3–18)] the sampling distribution for this problem; in our present notation this is the hypergeometric distribution

$$p(D|N, R, I) = h(r|N, R, n) = \binom{N}{n}^{-1} \binom{R}{r} \binom{N-R}{n-r} \tag{6–1}$$

where I now denotes the prior information, the general statement of the problem as given above.

Both N and R Unknown

In general neither N nor R is known initially, and the robot is to estimate both of them. If we succeed in drawing n balls from the Urn, then of course we know deductively that $N \ge n$. It seems to us intuitively that the data could tell us nothing more about N; how could the number r of red balls drawn, or the order of drawing, be relevant to N? But this intuition is using a hidden assumption that we can hardly be aware of until we see the robot's answer to the question.

The joint posterior probability distribution for N and R is

$$p(NR|DI) = p(N|I)\, p(R|NI)\, \frac{p(D|NRI)}{p(D|I)} \tag{6–2}$$

where we have factored the joint prior probability by the product rule: $p(NR|I) = p(N|I)\, p(R|NI)$, and the normalizing denominator is a double sum:

$$p(D|I) = \sum_{N=0}^{\infty} \sum_{R=0}^{N} p(N|I)\, p(R|NI)\, p(D|NRI) \tag{6–3}$$

in which, of course, the factor $p(D|NRI)$ is zero when $N < n$, or $R < r$, or $N - R < n - r$. Then the marginal posterior probability for N alone is

$$p(N|DI) = \sum_{R=0}^{N} p(NR|DI) = p(N|I)\, \frac{\sum_R p(R|NI)\, p(D|NRI)}{p(D|I)}\,. \tag{6–4}$$

We could equally well apply Bayes' theorem directly:

$$p(N|DI) = p(N|I)\, \frac{p(D|NI)}{p(D|I)} \tag{6–5}$$

and of course (6–4) and (6–5) must agree, by the product and sum rules.

These relations must hold whatever prior information I we may have about N, R that is to be expressed by $p(NR|I)$. In principle, this could be arbitrarily complicated and conversion of verbally stated prior information into $p(NR|I)$ is an open–ended problem; you can always analyze your prior information more deeply. But usually our prior information is rather simple, and these problems are not difficult mathematically.

Intuition might lead us to expect further that, whatever prior $p(N|I)$ we had assigned, the data can only truncate the impossible values, leaving the relative probabilities of the possible values unchanged:

$$p(N|DI) = \left\{ \begin{array}{cc} Ap(N|I), & N \geq n \\ 0, & 0 \leq N < n \end{array} \right\} \tag{6–6}$$

where A is a renormalization constant. Indeed, the rules of probability theory tell us that this must be true if the data tell us only that $N \geq n$ and nothing else about N. For, define the proposition:

$$Z \equiv \text{“}N \geq n\text{”} \tag{6–7}$$

Then

$$p(Z|NI) = \left\{ \begin{array}{ll} 1, & n \leq N \\ 0, & n > N \end{array} \right\} \tag{6–8}$$

and Bayes' theorem reads:

$$p(N|ZI) = p(N|I)\,\frac{p(Z|NI)}{p(Z|I)} = \left\{ \begin{array}{cc} Ap(N|I), & N \ge n \\ 0, & N < n \end{array} \right\} \tag{6–9}$$

so if the data tell us only that Z is true, then we have (6–6) and the above renormalization constant is $A = 1/p(Z|I)$. Bayes' theorem confirms that if we learn only that $N \ge n$, the relative probability of the possible values of N are not changed by this information; only the normalization must be readjusted to compensate for the values $N < n$ that now have zero probability. Laplace considered this result intuitively obvious, and took it as a basic principle of his theory.

However, the robot tells us in (6–5) that this will not be the case unless $p(D|NI)$ is independent of N for $N \ge n$. And on second thought we see that (6–6) need not be true if we have some kind of prior information linking N and R. For example, it is conceivable that one might know in advance that $R < 0.06N$. Then necessarily, having observed the data $(n, r) = (10, 6)$ we would know not only that $N \ge 10$; but that $N > 100$ (and that we had drawn a very improbable sample). Any prior information that provides a logical link between N and R makes the datum r relevant to estimating N after all. But usually we lack any such prior information, and so estimation of N is uninteresting, reducing to the same result (6–6).

From (6–5), the general condition that the data can tell us nothing about N except to truncate values less than n, is a nontrivial condition on the prior probability $p(R|NI)$:

$$p(D|NI) = \sum_{R=0}^{N} p(D|NRI)\,p(R|NI) = \left\{ \begin{array}{cc} f(n,r), & N \ge n \\ 0, & N < n \end{array} \right\} \tag{6–10}$$

where $f(n, r)$ may depend on the data, but is independent of N. Since we are using the standard hypergeometric Urn sampling distribution (6–1), this is explicitly,

$$\sum_{R=0}^{N} \binom{R}{r}\binom{N-R}{n-r}\,p(R|NI) = f(n,r)\binom{N}{n}, \qquad N \ge n \tag{6–11}$$

This is that hidden assumption that our intuition could hardly have told us about. It is a kind of discrete integral equation‡ which the prior $p(R|NI)$ must satisfy as the necessary and sufficient condition for the data to be uninformative about N. The sum on the left–hand side is necessarily always zero when $N < n$, for the first binomial coefficient is zero when $R < r$, and the second is zero when $R \ge r$ and $N < n$. Therefore the mathematical constraint on $p(R|NI)$ is only, rather sensibly, that $f(n, r)$ in (6–11) must be independent of N when $N \ge n$.

In fact, most "reasonable" priors do satisfy this condition, and as a result estimation of N is relatively uninteresting. Then, factoring the joint posterior distribution (6–2) in the form

$$p(NR|DI) = p(N|I)\,p(R|NDI)\,, \tag{6–12}$$

our main concern is with the factor $p(R|N, D, I)$, drawing inferences about R or about the ratio R/N with N supposed known. The posterior probability distribution for R is then, by Bayes' theorem,

‡ This peculiar name anticipates what we shall find later, in connection with marginalization theory; very general conditions of 'uninformativeness' are expressed by similar integral equations that the prior for one parameter must satisfy in order to make the data uninformative about another parameter.

$$p(R|D,N,I) = p(R|N,I)\,\frac{p(D|N,R,I)}{p(D|N,I)}\,. \tag{6–13}$$

Different choices of the prior probability $p(R|N,I)$ will yield many quite different results, and we now examine a few of them.

Uniform Prior

Consider the state of prior knowledge denoted by I_0, in which we are, seemingly, as ignorant as we could be about R while knowing N: the uniform distribution

$$p(R|N,I_0) = \begin{Bmatrix} \dfrac{1}{N+1}\,, & 0 \le R \le N \\ 0\,, & R > N \end{Bmatrix}\,. \tag{6–14}$$

Then a few terms cancel out and (6–13) reduces to

$$p(R|D,N,I_0) = S^{-1}\binom{R}{r}\binom{N-R}{n-r}\,, \tag{6–15}$$

where S is a normalization constant. For several purposes, we need the general summation formula

$$S \equiv \sum_{R=0}^{N}\binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1}\,, \tag{6–16}$$

whereupon the correctly normalized posterior distribution for R is

$$p(R|D,N,I_0) = \binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r}\,. \tag{6–17}$$

This is not a hypergeometric distribution like (6–1) because the variable is now R instead of r.

The prior (6–14) yields, using (6–16),

$$\sum_{R=0}^{N}\frac{1}{N+1}\binom{R}{r}\binom{N-R}{n-r} = \frac{1}{N+1}\binom{N+1}{n+1} = \frac{1}{n+1}\binom{N}{n} \tag{6–18}$$

so the integral equation (6–11) is satisfied; with this prior the data can tell us nothing about N beyond the fact that $N \ge n$.

Let us check (6–17) so see whether it satisfies some obvious common–sense requirements. We see that it vanishes when $R < r$, or $R > N - n + r$, in agreement with what the data tell us by deductive reasoning. If we have sampled all the balls, $n = N$, then (6–17) reduces to $\delta(R,r)$, again agreeing with deductive reasoning. This is another illustration of the fact that probability theory as logic automatically includes deductive logic as a special case.

But if we obtain no data at all, $n = r = 0$, then (6–17) reduces, as it should, to the prior distribution: $p(R|D,N,I_0) = p(R|N,I_0) = 1/(N+1)$. If we draw only one ball which proves to be red, $n = r = 1$, then (6–17) reduces to

$$p(R|D,N,I_0) = \frac{2R}{N(N+1)}\,. \tag{6–19}$$

The vanishing when $R = 0$ again agrees with deductive logic. From (6–1) the sampling probability $p(r = 1|n = 1, N, R, I_0) = R/N$ that our one ball would be red is our original Bernoulli Urn result, proportional to R; and with a uniform prior the posterior probability for R must also be proportional to R. The numerical coefficient in (6–19) gives us an inadvertent derivation of the elementary sum rule

$$\sum_{R=0}^{N} R = \frac{N(N+1)}{2} . \tag{6–20}$$

These results are only a few of thousands now known, indicating that probability theory as logic is an exact mathematical system. That is, results derived from correct application of our rules without approximation have the property of exact results in any other area of mathematics; you can subject them to arbitrary extreme conditions and they continue to make sense.†

What value of R does the robot estimate in general? The most probable value of R is found within one unit by setting $p(R') = p(R' - 1)$ and solving for R', which yields

$$R' = (N+1)\,\frac{r}{n} \tag{6–21}$$

which is to be compared to (3–22) for the peak of the sampling distribution. If R' is not an integer, the most probable value is the next integer below R'. The robot anticipates that the fraction of red balls in the original Urn should be about equal to the fraction in the observed sample, just as you and I would from intuition.

For a more refined calculation let us find the mean value, or expectation of R over this posterior distribution:

$$\langle R \rangle = E(R|D, N, I_0) = \sum_{R=0}^{N} R\, p(R|D, N, I_0) . \tag{6–22}$$

To do the summation, note that

$$(R+1)\binom{R}{r} = (r+1)\binom{R+1}{r+1} \tag{6–23}$$

and so, using (6–16) again,

$$\langle R \rangle + 1 = (r+1)\binom{N+1}{n+1}^{-1}\binom{N+2}{n+2} = \frac{(N+2)(r+1)}{(n+2)} . \tag{6–24}$$

When (n, r, N) are large, the expectation of R is very close to the most probable value (6–21), indicating either a sharply peaked posterior distribution or a symmetric one. This result becomes more significant when we ask: "What is the expected fraction F of red balls left in the Urn after this drawing?" This is

$$\langle F \rangle = \frac{\langle R \rangle - r}{N - n} = \frac{r+1}{n+2} . \tag{6–25}$$

Now let us ask a different question: after having drawn a sample of r red balls in n draws, what is the probability that the next one drawn will be red? Defining the propositions:

$$R_i \equiv \text{"Red on the } i\text{'th draw"}, \qquad 1 \le i \le N$$

this is

† By contrast, the intuitive *ad hockeries* of current "orthodox" statistics generally give reasonable results within some 'safe' domain for which they were invented; but invariably they are found to yield nonsense in some extreme case. This is what one expects of results which are only approximations to an exact theory; as one varies the conditions the quality of the approximation varies.

$$p(R_{n+1}|D,N,I_0) = \sum_{R=0}^{N} p(R_{n+1}\, R|D,N,I_0) = \sum_R p(R_{n+1}|R,D,N,I_0) \cdot p(R|D,n,I_0) \qquad (6\text{–}26)$$

or,

$$p(R_{n+1}|D,N,I_0) = \sum_{R=0}^{N} \frac{R-r}{N-n} \cdot \binom{N+1}{n+1}^{-1} \binom{R}{r} \binom{N-R}{n-r} \qquad (6\text{–}27)$$

Using the summation formula (6–16) again, we find after some algebra,

$$p(R_{n+1}|D,N,I_0) = \frac{r+1}{n+2}, \qquad (6\text{–}28)$$

the same as (6–25). This agreement is another example of the rule noted before: a probability is not the same thing as a frequency; but under quite general conditions the probability of an event at a single trial is numerically equal to the *expectation* of its frequency in some specified class of trials.

Eq. (6–28) is a famous old result known as *Laplace's Rule of Succession.* It has played a major role in the history of Bayesian inference, and in the controversies over the nature of induction and inference. We shall find it reappearing many times; finally, in Chapter 18 we examine it carefully to see how it became controversial, but also how easily the controversies can be resolved today.

The result (6–28) has a greater generality than would appear from our derivation. Laplace first obtained it, not in consideration of drawing from an Urn, but from considering a mixture of binomial distributions, as we shall do presently in (6–70). The above derivation in terms of Urn sampling had been found as early as 1799 (see Zabell, 1989), but became well known only through its rediscovery in 1918 by C. D. Broad of Cambridge University, England, and its subsequent emphasis by Wrinch and Jeffreys (1919), W. E. Johnson (1924, 1932), and Jeffreys (1939). It was initially a great surprise to find that the Urn result (6–28) is independent of N.

But this is only the point estimate; what accuracy does the robot claim for this estimate of R? The answer is contained in the same posterior distribution (6–17) that gave us (6–28); we may find its variance $\langle R^2\rangle - \langle R\rangle^2$. Extending (6–23), note that

$$(R+1)(R+2)\binom{R}{r} = (r+1)(r+2)\binom{R+2}{r+2}. \qquad (6\text{–}29)$$

The summation over R is again simple, yielding

$$\langle (R+1)(R+2)\rangle = (r+1)(r+2)\binom{N+1}{n+1}^{-1}\binom{N+3}{n+3} = \frac{(r+1)(r+2)(N+2)(N+3)}{(n+2)(n+3)} \qquad (6\text{–}30)$$

Then noting that $\mathrm{var}(R) = \langle R^2\rangle - \langle R\rangle^2 = \langle (R+1)^2\rangle - \langle (R+1)\rangle^2$ and writing for brevity $p = \langle F\rangle = (r+1)/(n+2)$, from (6–24), (6–30) we find

$$\mathrm{var}(R) = \frac{p(1-p)}{n+3}\,(N+2)\,(N-n). \qquad (6\text{–}31)$$

Therefore, our (mean) $\pm$ (standard deviation) combined point and interval estimate of R is

$$(R)_{est} = r + (N-n)p \pm \sqrt{\frac{p(1-p)}{n+3}\,(N+2)\,(N-n)}\,. \tag{6–32}$$

The factor $(N-n)$ inside the square root indicates that, as we would expect, the estimate becomes more accurate as we sample a larger fraction of the contents of the Urn. Indeed, when $n = N$ the contents of the Urn are known and (6–32) reduces as it should to $(r \pm 0)$, in agreement with deductive reasoning.

But looking at (6–32) we note that $R - r$ is the number of red balls remaining in the Urn, and $N - n$ is the total number of balls left in the Urn; so an analytically simpler expression is found if we ask for the (mean) $\pm$ (standard deviation) estimate of the fraction of red balls remaining in the Urn after the sample is drawn. This is found to be

$$(F)_{est} = \frac{(R-r)_{est}}{N-n} = p \pm \sqrt{\frac{p(1-p)}{n+3}\,\frac{N+2}{N-n}}\,, \qquad 0 \le n < N \tag{6–33}$$

and this estimate gets less accurate as we sample a larger portion of the balls. In the limit $N \to \infty$ this goes into

$$(F)_{est} = p \pm \sqrt{\frac{p(1-p)}{n+3}}\,, \tag{6–34}$$

which corresponds to the binomial distribution result.

As an application of this, while preparing this Chapter we heard a news report that a "random poll" of 1600 voters was taken, indicating that 41% of the population favored a certain candidate in the next election, and claiming a $\pm 3\%$ margin of error for this result. Let us check the consistency of these numbers against our theory. To obtain $(F)_{est} = \langle F \rangle (1 \pm .03)$ we require according to (6–34) a sample size n given by

$$n + 3 = \frac{1-p}{p}\,\frac{1}{(.03)^2} = \frac{1-.41}{.41} \times 1111 = 1598.9 \tag{6–35}$$

or, $n = 1596$. The close agreement suggests that the pollsters are using this theory (or at least giving implied lip service to it in their public announcements).

These results, found with a uniform prior for $p(R|N, I_0)$ over $0 \le R \le N$, correspond very well with our intuitive common–sense judgments. Other choices of the prior can affect the conclusions in ways which often surprise us at first glance; then after some meditation we see that they were correct after all. Let us put probability theory to a more severe test by considering some increasingly surprising examples.

Truncated Uniform Priors

Suppose our prior information had been different from the above I_0; our new prior information I_1 is that we know from the start that $0 < R < N$; there is at least one red and one white ball in the Urn. Then the prior (6–14) must be replaced by

$$p(R|N, I_1) = \left\{ \begin{array}{ll} \dfrac{1}{N-1}\,, & 1 \le R \le N-1 \\ 0\,, & \text{otherwise} \end{array} \right\} \tag{6–36}$$

and our summation formula (6–16) must be corrected by subtracting off the two terms $R = 0$, $R = N$. Note that if $R = 0$, then

$$\binom{R}{r} = \binom{R+1}{r+1} = \delta(r, 0)$$

and if $R = N$, then

$$\binom{N-R}{n-r} = \delta(r,n)\,,$$

so we have the summation formulas

$$S = \sum_{R=1}^{N-1} \binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1} - \binom{N}{n}\,\delta(r,n) - \binom{N}{n}\,\delta(r,0) \qquad (6\text{–}37)$$

$$\sum_{R=1}^{N-1} \binom{R+1}{r+1}\binom{N-R}{n-r} = \binom{N+2}{n+2} - \binom{N+1}{n+1}\,\delta(r,n) - \binom{N}{n}\,\delta(r,0) \qquad (6\text{–}38)$$

What seems surprising at first is that as long as the observed r is in $0 < r < n$ the new terms vanish, and so the previous posterior distribution (6–17) is unchanged:

$$p(R|D,N,I_1) = p(R|D,N,I_0)\,, \qquad 0 < r < n\,. \qquad (6\text{–}39)$$

Why does the new prior information make no difference? Indeed, it would certainly make a difference in any form of probability theory that uses only sampling distributions; for the sample space is changed by the new information.

Yet on meditation we see that the result (6–39) is correct, for in this case the data tell us by deductive reasoning that R cannot be 0 or N; so whether the prior information does or does not tell us the same thing should not matter; our state of knowledge about R is the same and probability theory as logic so indicates. We discuss this further under "optional stopping" below.

But suppose that our data were $r = 0$; now the sum S in (6–15) is different:

$$S = \binom{N+1}{n+1} - \binom{N}{n} \qquad (6\text{–}40)$$

and in place of (6–17) the posterior probability distribution for R is found to be, after some calculation,

$$p(R|r=0,N,I_1) = \binom{N}{n+1}\binom{N-R}{n}\,, \qquad 1 \le R \le N-1 \qquad (6\text{–}41)$$

and zero outside that range. But still, within that range the relative probabilities of different values of R are not changed; we readily verify that the ratio

$$\frac{p(R|r=0,N,I_1)}{p(R|r=0,N,I_0)} = \frac{N+1}{N-n}\,, \qquad 1 \le R \le N-1 \qquad (6\text{–}42)$$

is independent of R. What has happened here is that the datum $r = 0$ gives no evidence against the hypothesis that $R = 0$ and some evidence for it; so on prior information I_0 which allows this, $R = 0$ is the most probable value. But the prior information I_1 now makes a decisive difference; it excludes just that value, and thus forces all the posterior probability to be compressed into a smaller range, with an upward adjustment of the normalization coefficient. We learn from this example that different priors do not necessarily lead to different conclusions; and whether they do or do not can depend on which data set we happen to get – which is just as it should be.

Exercise 6.1. Find the posterior probability distribution $p(R|r = n, N, I_1)$ by a derivation like the above. Then find the new (mean) $\pm$ (standard deviation) estimates of R from this distribution, and compare it with the above results from $p(R|r = n, N, I_0)$. Explain the difference so that it seems obvious intuitively. Now show how well you understand this problem by describing in words, without doing the calculation, how the result would differ if we had prior information that $(3 \le R \le N)$; the Urn had initially at least three red balls, but there was no prior restriction on large values.

A Concave Prior

The rule of succession, based on the uniform prior $\{p(R|NI) \propto \text{const.}, \quad 0 \le R \le N\}$, leads to a perhaps surprising numerical result, that the expected fraction (6–25) of red balls left in the Urn is not the fraction r/n observed in the sample drawn, but slightly different, $(r+1)/(n+2)$. What is the reason for this small difference? The following argument is hardly a derivation, but only a line of free association. Note first that Laplace's rule of succession can be written in the form

$$\frac{r+1}{n+2} = \frac{n \cdot (r/n) + 2 \cdot (1/2)}{n+2} \tag{6–43}$$

which exhibits the result as a weighted average of the observed fraction r/n and the prior expectation $1/2$, the data weighted by the number n of observations, the prior expectation by 2. It seems that the uniform prior carries a weight corresponding to two observations. Then could that prior be interpreted as a posterior distribution resulting from two observations $(n, r) = (2, 1)$? If so, it seems that we must start from a still more uninformative prior than the uniform one. But is there any such thing as a still more uninformative prior?

Mathematically, this suggests that we try to apply Bayes' theorem backwards, to find whether there is any prior that would lead to a uniform posterior distribution. Denote this conjectured still more primitive state of "pre–prior" information by I_{00}. Then Bayes' theorem would read:

$$p(R|DI_{00}) = p(R|I_{00}) \frac{p(D|RI_{00})}{p(D|I_{00})} = \text{const.}, \qquad 0 \le R \le N \tag{6–44}$$

and the sampling distribution is still the hypergeometric distribution (6–1), because when R is specified it renders any vague information like I_{00} irrelevant: $p(D|RI_0) = p(D|RI_{00})$. With the assumed sample, $n = 2, r = 1$ the hypergeometric distribution reduces to

$$h(r = 1|N, R, n = 2) = \frac{R(N-R)}{N(N-1)}, \qquad 0 \le R \le N \tag{6–45}$$

from which we see that there is no pre–prior that yields a constant posterior distribution over the whole range $(0 \le R \le N)$; it would be infinite for $R = 0$ and $R = N$. But we have just seen that the truncated prior, constant in $(1 \le R \le N-1)$, yields the same results if it is known that the Urn contains initially at least one red and one white ball. Since our presupposed data $(n, r) = (2, 1)$ guarantees this, we see that we have a solution after all: consider the prior that emphasizes extreme values:

$$p(R|I_{00}) \equiv \frac{A}{R(N-R)}, \qquad 1 \le R \le N-1 \tag{6–46}$$

where A stands for a normalization constant, not necessarily the same in all the following equations. Given new data $D \equiv (n, r)$, if $1 \le r \le n-1$ this yields, using (6–1), the posterior distribution

$$p(R|DNI_{00}) = \frac{A}{R(N-R)}\binom{R}{r}\binom{N-R}{n-r} = \frac{A}{r(n-r)}\binom{R-1}{r-1}\binom{N-R-1}{n-r-1}. \qquad (6\text{–}47)$$

From (6–16) we may deduce the summation formula

$$\sum_{r=1}^{N-1}\binom{R-1}{r-1}\binom{N-R-1}{n-r-1} = \binom{N-1}{n-1}, \qquad \begin{array}{l} 1 \le R \le N-1, \\ 1 \le r \le n-1 \end{array} \qquad (6\text{–}48)$$

so the correctly normalized posterior distribution is

$$p(R|DNI_{00}) = \binom{N-1}{n-1}^{-1}\binom{R-1}{r-1}\binom{N-R-1}{n-r-1} \qquad \begin{array}{l} 1 \le R \le N-1, \\ 1 \le r \le n-1 \end{array} \qquad (6\text{–}49)$$

which is to be compared with (6–17). As a check, if $n = 2, r = 1$ this reduces to the desired prior (6–36):

$$p(R|DNI_{00}) = p(R|NI_1) = \frac{1}{N-1}, \qquad 1 \le R \le N-1 \qquad (6\text{–}50)$$

At this point, we can leave it as an exercise for the reader to complete the analysis for the concave prior with derivations analogous to (6–22) – (6–34):

Exercise 6.2. Using the general result (6–49), repeat the calculations analogous to (6–22) – (6–34) and prove two exact results: (a) The integral equation (6–11) is satisfied, so (6–6) still holds. (b) For general data compatible with the prior in the sense that $0 \le n \le N$, $\quad 1 \le r \le n-1$ (so that the sample drawn includes at least one red and one white ball), the posterior mean estimated fractions $R/N, (R-r)/(N-n)$ are both equal simply to the observed fraction in the sample, $f = r/n$; the estimates now follow the data exactly and the concave prior (6–46) is given zero weight. (c) The (mean) $\pm$ (standard deviation) estimate is given by

$$\frac{(R)_{est}}{N} = f \pm \sqrt{\frac{f(1-f)}{n+1}\left(1-\frac{n}{N}\right)} \qquad (6-51)$$

also a simpler result than the analogous (6–32) found previously for the uniform prior.

Exercise 6.3. Now note that if $r = 0$ or $r = n$, the step (6–47) is not valid. Go back to the beginning and derive the posterior distribution for these cases. Show that if we draw one ball and find it not red, the estimated fraction of red in the Urn now drops from $1/2$ to approximately $1/\log N$ (whereas with the uniform prior it drops to $(r+1)/(n+2) = 1/3$).

The exercises show that the concave prior gives many results simpler than those of the uniform one, but has also some near instability properties that become more pronounced with large N. Indeed, as $N \to \infty$ the concave prior approaches an improper (non–normalizable) one, which must give absurd answers to some questions, although it still gives reasonable answers to most questions (those in which the data are so informative that they remove the singularity associated with the prior).

The Binomial Monkey Prior

Suppose prior information I_2 is that the Urn was filled by a team of monkeys who tossed balls in at random, in such a way that each ball entering had independently the probability g of being red. Then our prior for R will be the binomial distribution (3–79): in our present notation,

$$p(R|N, I_2) = \binom{N}{R} g^R (1-g)^{N-R}, \qquad 0 \le R \le N \tag{6–52}$$

and our prior estimate of the fraction of red ones in the Urn will be the (mean) $\pm$ (standard deviation) over this distribution:

$$(R)_{est} = Ng \pm \sqrt{Ng(1-g)} \tag{6–53}$$

The sum (6–10) is readily evaluated for this prior, with the result that

$$p(D|NI) = \binom{n}{r} g^r (1-g)^{n-r}, \qquad N \ge n \tag{6–54}$$

Since this is independent of N, this prior also satisfies our integral equation (6–11), so

$$p(NR|DI_2) = p(N|DI_2)\, p(R|NDI_2) \tag{6–55}$$

in which the first factor is the relatively uninteresting standard result (6–6). We are interested in the factor $p(R|NDI_2)$ in which N is considered known. We are interested also in the other factorization

$$p(NR|DI_2) = p(R|DI_2)\, p(N|RDI_2) \tag{6–56}$$

in which $p(R|DI)$ tells us what we know about R, regardless of N (here let the reader try to guess intuitively how $p(R|DNI)$ and $p(R|DI)$ would differ, before seeing the calculations). Likewise, the difference between $p(N|RDI_2)$ and $p(N|DI)$ tells us how much we would learn about N if we were to learn the true R; and again our intuition can hardly anticipate the result of the calculation.

We have set up quite an agenda of calculations to do. Using (6–52) and (6–1), we find

$$p(R|D, N, I_2) = A \binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r} \binom{N-R}{n-r} \tag{6–57}$$

where A is another normalization constant. To evaluate it, note that we can rearrange the binomial coefficients:

$$\binom{N}{R} \binom{R}{r} \binom{N-R}{n-r} = \binom{N}{n} \binom{n}{r} \binom{N-n}{R-r} \tag{6–58}$$

Therefore we find the normalization by

$$\begin{aligned} 1 = \sum_R p(R|D, N, I_2) &= A \binom{N}{n} \binom{n}{r} \sum_R \binom{N-n}{R-r} g^R (1-g)^{N-R} \\ &= A \binom{N}{n} \binom{n}{r} g^r (1-g)^{n-r}, \qquad r \le R \le N-n+r \end{aligned} \tag{6–59}$$

and so our normalized posterior distribution for R is

$$p(R|D,N,I_2) = \binom{N-n}{R-r} g^{R-r} (1-g)^{N-R-n+r} \tag{6–60}$$

from which we would make the (mean) $\pm$ (standard deviation) estimate

$$(R)_{est} = r + (N-n)g \pm \sqrt{g(1-g)(N-n)} \tag{6–61}$$

But the resemblance to (6–32) suggests that we again look at it this way: we estimate the fraction of red balls left in the Urn to be

$$\frac{(R-r)_{est}}{N-n} = g \pm \sqrt{\frac{g(1-g)}{N-n}}\,. \tag{6–62}$$

At first glance, (6–61) and (6–62) seem to be so much like (6–32) and (6–33) that it was hardly worth the effort to derive them. But on second glance we notice an astonishing fact: the parameter p in the former equations was determined entirely by the data; while g in the present ones is determined entirely by the prior information. In fact, (6–62) is exactly the prior estimate we would have made for the fraction of red balls in any subset of $N-n$ balls in the Urn, *without any data at all*. It seems that the binomial prior has the magical property that it nullifies the data! More precisely, with that prior the data can tell us nothing at all about the unsampled balls.

Such a result will hardly commend itself to a survey sampler; the basis of his profession would be wiped out. Yet the result is correct and there is no escape from the conclusion; if your prior information about the population is correctly described by the binomial prior, then sampling is futile (it tells you practically nothing about the population) unless you sample practically the whole population.

How can such a thing happen? Comparing the binomial prior with the uniform prior, one would suppose that the binomial prior, being moderately peaked, expresses more prior information about the proportion R/N of red balls; therefore by its use one should be able to improve his estimates of R. Indeed, we have found this effect; for the uncertainties in (6–61) and (6–62) are smaller than that those in (6–32) and (6–33) by a factor of $\sqrt{(n+3)/(N+2)}$. What is intriguing is not the magnitude of the uncertainty; but the fact that (6–33) depends on the data; while (6–62) does not.

It is not surprising that the binomial prior is more informative about the unsampled balls than are the data of a small sample; but actually it is more informative about them than are *any amount* of data; even after sampling 99% of the population, we are no wiser about the remaining 1%.

So what is the invisible strange property of the binomial prior? It is in some sense so "loose" that it destroys the logical link between different members of the population. But on meditation we see that this is just what was implied by our scenario of the Urn being filled by monkeys tossing in balls in such a way that each ball had *independently* the probability g of being red. Given that filling mechanism, then knowing that any given ball is in fact red, gives one no information whatsoever about any other ball. That is, $P(R_1R_2|I) = P(R_1|I)\,P(R_2|I)$. This logical independence in the prior is preserved in the posterior distribution.

Exercise 6.4. Investigate this apparent "law of conservation of logical independence". If the propositions: "$\{i$'th ball is red, $1 \le i \le N\}$" are logically independent in the prior information, what is the necessary and sufficient condition on the sampling distribution and the data, that the factorization property is retained in the posterior distribution: $P(R_1R_2|DI) = P(R_1|DI)\,P(R_2|DI)$?

This sets off another line of deep thought. In conventional probability theory, the binomial distribution is derived from the premise of causal independence of different tosses. In Chapter 3 we found that consistency requires one to reinterpret this as logical independence. But now, can we reason in the opposite direction? *Does the appearance of a binomial distribution already imply logical independence of the separate events*? If so, then we could understand the weird result just derived, and anticipate many others like it. We shall return to these questions in a later Chapter, after acquiring some more clues.

Metamorphosis into Continuous Parameter Estimation

As noted in the Introduction, if our hypotheses become so "dense" that neighboring hypotheses (*i.e.*, hypotheses with nearly the same values of the index t) are barely distinguishable in their observable consequences, then whatever the data, their posterior probabilities cannot differ appreciably. So there cannot be one sharply defined hypothesis that is strongly favored over all others. Then it may be appropriate and natural to think of t as a continuously variable parameter θ, and to interpret the problem as that of making an estimate of the parameter θ, and a statement about the accuracy of the estimate.

A common and useful custom is to use Greek letters to denote continuously variable parameters, Latin letters for discrete indices or data values. We shall adhere to this except when it would conflict with a more deeply entrenched custom.‡

The hypothesis testing problem has thus metamorphosed into a parameter estimation one. But it can equally well metamorphose back; for the hypothesis that a parameter θ lies in a certain interval $a < \theta < b$ is, of course, a compound hypothesis as defined in Chapter 4, so an interval estimation procedure (*i.e.*, one where we specify the accuracy by giving the probability that the parameter lies in a given interval) is automatically a compound hypothesis testing procedure.

Indeed, we followed just this path in Chapter 4 and found ourselves, at Eq. (4–57), doing what is really parameter estimation. It seemed to us natural to pass from testing simple discrete hypotheses, to estimating continuous parameters, and finally to testing compound hypotheses at Eq. (4–64), because probability theory as logic does this automatically. As in our opening remarks, we do not see parameter estimation and hypothesis testing as fundamentally different activities – one aspect of the greater unity of probability theory as logic.

But this unity has not seemed at all natural to some others. Indeed, in orthodox statistics parameter estimation appears very different from hypothesis testing, both mathematically and conceptually, largely because it has no satisfactory way to deal with compound hypotheses or prior information. We shall see some specific consequences of this in Chapter 17. Of course, parameters need not be one–dimensional; but let us consider first some simple cases where they are.

Estimation with a Binomial Sampling Distribution

We have already seen an example of a binomial estimation problem in Chapter 4, but we did not note its generality. There are hundreds of real situations in which each time a simple measurement or observation is made, there are only two possible results. The coin will show either heads or tails, the battery will or will not start the car, the baby will be a boy or a girl, the check will or will not arrive in the mail today, the student will pass or flunk the examination, *etc.*. As we noted in Chapter 3, the first comprehensive sampling theory analysis of such an experiment was by James Bernoulli (1713) in terms of drawing balls from an Urn, so such experiments are commonly called *Bernoulli trials*.

Traditionally, for any such binary experiment we call one of the results, arbitrarily, a "success" and the other a "failure". Generally, our data will be a record of the number of successes and

‡ Thus for generations the charge on the electron and the velocity of light have been denoted by e, c respectively. No scientist or engineer could bring himself to represent them by Greek letters, even when they are the parameters being estimated.

the number of failures;‡ the order in which they occur may or may not be meaningful, and if it is meaningful, it may or may not be known; and if it is known, it may or may not be relevant to the question we are asking. Presumably, the conditions of the experiment will tell us whether the order is meaningful or known; and we expect probability theory to tell us whether it is relevant.

For example, if we toss 10 coins simultaneously, then we have performed 10 Bernoulli trials, but it is not meaningful to speak of their 'order'. If we toss one coin 100 times and record each result, then the order of the results is meaningful and known; but in trying to judge whether the coin is 'honest', common sense probably tells us that the order is not relevant. If we are observing patient recoveries from a disease and trying to judge whether resistance to the disease was improved by a new medicine introduced a month ago, this is much like drawing from an Urn whose contents may have changed. Intuition then tells us that the order in which recoveries and non–recoveries occur is not only highly relevant; it is the crucial information without which no inference about a change is possible.†

To set up the simple general binomial sampling problem, define

$$x_i \equiv \left\{ \begin{array}{ll} 1, & \text{if the } i'\text{th trial yields success} \\ 0, & \text{otherwise} \end{array} \right\} . \tag{6–63}$$

Then our data are $D \equiv \{x_1, \cdots, x_n\}$. The prior information I specifies that there is a parameter θ such that at each trial we have, independently of anything we know about other trials, the probability θ of a success, therefore probability $(1-\theta)$ of a failure. As discussed before, by 'independent' we mean logical independence. There may or may not be causal independence, depending on further details of I that do not matter at the moment. The sampling distribution is then (mathematically, this is our *definition* of the model to be studied):

$$p(D|\theta, I) = \prod_{i=1}^{n} p(x_i|\theta, I) = \theta^r\,(1-\theta)^{n-r} \ , \tag{6–64}$$

in which r is the number of successes observed, $(n-r)$ the number of failures. Then with any prior probability density function $p(\theta|I)$ we have immediately the posterior *pdf*

$$p(\theta|D, I) = \frac{p(\theta|I)\,p(D|\theta, I)}{\int p(\theta|I)\,p(D|\theta, I)\,d\theta} = A\,p(\theta|I)\,\theta^r\,(1-\theta)^{n-r} \ , \tag{6–65}$$

where A is a normalizing constant. With a uniform prior for θ,

$$p(\theta|I) = 1\,, \qquad 0 \le \theta \le 1 \tag{6–66}$$

the normalization is determined by an Eulerian integral:

$$A^{-1} = \int_0^1 \theta^r\,(1-\theta)^{n-r}\,d\theta = \frac{r!\,(n-r)!}{(n+1)!} \tag{6–67}$$

and the normalized *pdf* is

‡ However, there are important problems involving censored data, to be considered later, in which only the successes can be recorded (or only the failures), and one does not know how many trials were performed. For example, a highway safety engineer knows from the public record how many lives were lost in spite of his efforts; but not how many were saved because of them.

† Of course, the final arbiter of relevance is not our intuition, but the equations of probability theory. But as we shall see later, judging this can be a tricky business. Whether a given piece of information is or is not relevant depends not only on what question we are asking, but also on the totality of all of our other information.

$$p(\theta|D,I) = \frac{(n+1)!}{r!\,(n-r)!}\,\theta^r\,(1-\theta)^{n-r} \tag{6–68}$$

identical with Bayes' original result, noted in Chapter 4, Eq. (4–57). Its moments are

$$\begin{aligned}\langle\theta^m\rangle = E(\theta^m|D,I) = A\int_0^1 \theta^{r+m}\,(1-\theta)^{n-r}\,d\theta &= \frac{(n+1)!}{(n+m+1)!}\,\frac{(r+m)!}{r!}\\ &= \frac{(r+1)(r+2)\cdots(r+m)}{(n+2)(n+3)\cdots(n+m+1)}\end{aligned} \tag{6–69}$$

leading to the predictive probability of success at the next trial of

$$p \equiv \langle\theta\rangle = \int_0^1 \theta\,p(\theta|DI)\,d\theta = \frac{r+1}{n+2} \tag{6–70}$$

in which we see Laplace's rule of succession in its original derivation. Likewise the (mean $\pm$ standard deviation) estimate of θ is:

$$(\theta)_{est} = \langle\theta\rangle \pm \sqrt{\langle\theta^2\rangle - \langle\theta\rangle^2} = p \pm \sqrt{\frac{p(1-p)}{n+3}} \tag{6–71}$$

Indeed, the continuous results (6–70) and (6–71) must be derivable from the discrete ones (6–28) and (6–34) by passage to the limit $N \to \infty$; but since the latter equations are independent of N, the limit has no effect.

In this limit the concave pre–prior distribution (6–46) would go into an improper prior for θ:

$$\frac{A}{R(N-R)} \to \frac{d\theta}{\theta(1-\theta)} \tag{6–72}$$

for which some sums or integrals would diverge; but that is not the strictly correct method of calculation. For example, to calculate the posterior expectation of any function $f(R/N)$ in the limit of arbitrarily large N, we should take limit of the ratio $\langle f(R/N)\rangle = Num/Den$, where

$$\begin{aligned}Num &\equiv \sum_{R=1}^{N-1} \frac{f(R/N)}{R(N-R)}\,p(D|N,R,I),\\ Den &\equiv \sum_{R=1}^{N-1} \frac{1}{R(N-R)}\,p(D|N,R,I)\end{aligned} \tag{6–73}$$

and under very general conditions this limit is well–behaved, leading to useful results. The limiting improper pre–prior (6–72) was advocated by Haldane (1932) and Jeffreys (1939), in the innocent days before the marginalization paradox, when nobody worried about such fine points. We were almost always lucky in that our integrals converged in the limit, so we used them directly, thus actually calculating the ratio of the limits rather than the limit of the ratio; but nevertheless getting the right answers. With this fine point now clarified, all this and its obvious generalizations seem perfectly straightforward; however, note the following point, important for a current controversy.

Digression on Optional Stopping

We did not include n in the conditioning statements in $p(D|\theta, I)$ because, in the problem as defined, it is from the data D that we learn both n and r. But nothing prevents us from considering a different problem in which we decide in advance how many trials we shall make; then it is proper to add n to the prior information and write the sampling probability as $p(D|n, \theta, I)$. Or, one might decide in advance to continue the Bernoulli trials until we have achieved a certain number r of successes, or a certain log–odds $u = \log[r/(n-r)]$; then it would be proper to write the sampling probability $p(D|r, \theta, I)$ or $p(D|u, \theta, I)$; and so on. Does this matter for our conclusions about θ?

Now in deductive logic (Boolean algebra) it is a triviality that $AA = A$; if you say: "A is true" twice, this is logically no different from saying it once. This property is retained in probability theory as logic, since it was one of our basic desiderata that propositions with the same truth value are always assigned the same probability. In practice this means that there is no need to ensure that the different pieces of information given to the robot are independent; our formalism has automatically the property that redundant information is not counted twice.

Thus in our present problem the data, as defined, tell us n. Then, since $p(n|n, \theta, I) = 1$, the product rule may be written

$$p(n, r, \text{order}|n, \theta, I) = p(r, \text{order}|n, \theta, I)\, p(n|n, \theta, I) = p(r, \text{order}|n, \theta, I)\,. \tag{6–74}$$

If something is known already from the prior information, then whether the data do or do not tell us the same thing cannot matter; the likelihood function is the same. Likewise, write the product rule as

$$p(\theta, n|D, I) = p(\theta|n, D, I)\, p(n|D, I) = p(n|\theta, D, I)\, p(\theta|D, I) \tag{6–75}$$

or, since $p(n|\theta, D, I) = p(n|D, I) = 1$,

$$p(\theta|n, D, I) = p(\theta|D, I) \tag{6–76}$$

In this argument we could replace n by any other quantity [such as r, or $(n-r)$, or $u \equiv \log[r/(n-r)]$ that was known from the data; if any part of the data happens to be included also in the prior information, then that part is redundant and it cannot affect our final conclusions.

Yet some statisticians (for example, Armitage, 1960) who look only at sampling distributions, claim that the stopping rule *does* affect our inference. Apparently, they believe that if a statistic such as r is not known in advance, then parts of the sample space referring to false values of r remain relevant to our inferences even after the true value of r becomes known from the data D, although they would not be relevant (they would not even be in the sample space) if the true value were known before seeing the data. Of course, that does violence to the principle $AA = A$ of elementary logic; it is astonishing that such a thing could be controversial in the twentieth Century.

It is evident that this same comment applies with equal force to any function $f(D)$ of the data. Therefore that principle of logic tells us that, quite generally, data sets that were not realized cannot be relevant to any inference. Although this conclusion might have seemed intuitively obvious from the start, it is not recognized at all in orthodox statistics; we shall see in Chapter 9 not only some irrational conclusions, but some absolutely spooky consequences (psychokinesis, black magic) this has had, and in Chapter 17 how much real damage this has caused, particularly in medical testing. This is a cogent lesson showing the importance of deriving the rules of inference from the requirements of logical consistency, instead of intuitive guesswork.

But what if part of the data set was actually generated by the phenomenon being studied, but for whatever reason we failed to observe it? We show next that probability theory as logic tells us uniquely how to deal with true but unobserved data; they must be relevant, but with a mathematical status somewhat like that of a set of nuisance parameters.

Compound Estimation Problems

We now consider in some depth a class of problems more complicated in structure, where more than one process is occurring but not all the results are observable. We want to make inferences not only about parameters, but about the unobserved data. The mathematics to be developed next is applicable to a large number of quite different real problems. To form an idea of the scope of the theory, consider these scenarios:

(A) In the general population, there is a probability p that any given person will contract a certain disease within the next year; and then a probability θ that anyone with the disease will die of it within a year. From the observed variations $\{c_1, c_2, \ldots\}$ of deaths from the disease in successive years (which is a matter of public record), estimate how the incidence of the disease $\{n_1, n_2, \ldots\}$ is changing in the general population (which is not a matter of public record).

(B) Each week, a large number N of mosquitos is bred in a stagnant pond near this campus, and we set up a trap on the campus to catch some of them. Each mosquito lives less than a week, during which it has a probability p of flying onto the campus, and once on the campus, it has a probability θ of being caught in our trap. We count the numbers $\{c_1, c_2, \ldots\}$ caught each week. From these data and whatever prior information we have, what can we say about the numbers $\{n_1, n_2, \ldots\}$ on the campus each week, and what can we say about N?

(C) We have a radioactive source (say Sodium 23 for example) which is emitting particles of some sort (say the positrons from Na^{23}). Each radioactive nucleus has the probability p of sending a particle through our counter in one second; and each particle passing through has the probability θ of producing a count. From measuring the number $\{c_1, c_2, \ldots\}$ of counts in different seconds, what can we say about the numbers $\{n_1, n_2, \ldots\}$ actually passing through the counter in each second, and what can we say about the strength of the source?

The common feature in these problems is that we have two "binary games" played in succession, and we can observe only the outcome of the last one. From this, we are to make the best inferences we can about the original cause and the intermediate conditions. This could be described also as the problem of trying to recover, in one special case, censored data.

We want to show particularly how drastically these problems are changed by various changes in the prior information. For example, our estimates of the variation in incidence of a disease are greatly affected, not only by the data, but by our prior information about the process by which one contracts that disease.†

In our estimates we will want to (1) state the "best" estimate possible on the data and prior information; and (2) make a statement about the accuracy of the estimate, giving again our versions of "point estimation" and "interval estimation" about which Fisher commented. We shall use the language of the radioactive source scenario, but it will be clear enough that the same arguments and the same calculations apply in a hundred others.

A Simple Bayesian Estimate: Quantitative Prior Information

First, we discuss the parameter θ, which a scientist would call the "efficiency" of the counter. By this we mean that, if θ is known, then each particle passing through the counter has *independently* the probability θ of making a count. Again we emphasize that this is not mere causal independence (which surely always holds, as any physicist would assure us); we mean *logical* independence; *i.e.* if

† Of course, in this first venture into the following kind of analysis, we shall not take into account all the factors that operate in the real world, so some of our conclusions may be changed in a more sophisticated analysis. However, nobody would see how to do that unless he had first studied this simple introductory example.

θ is known, then knowing anything about the number of counts produced by other particles would tell us nothing more about the probability of the next particle making a count.‡

We have already stressed the distinction between logical and causal dependence many times; and now we have another case where failure to understand it could lead to serious errors. The point is that causal influences operate in the same way independently of your state of knowledge or mine; thus if θ is not known, then everybody still believes that successive counts are *causally* independent. But they are no longer *logically* independent; for then knowing the number of counts produced by other particles tells us something about θ, and therefore modifies our probability that the next particle will produce a count. The situation is much like that of sampling with replacement, discussed above, where each ball drawn tells us something more about the contents of the Urn.

From the independence, the probability that n particles will produce exactly c counts in any specified order, is $\theta^c (1-\theta)^{n-c}$, and there are $\binom{n}{c}$ possible sequences producing c counts, so the probability of getting c counts regardless of order is the binomial distribution

$$p(c|n,\theta) = \binom{n}{c} \theta^c (1-\theta)^{n-c} , \qquad 0 \le c \le n \tag{6–78}$$

From the standpoint of logical presentation in the real world, however, we have to carry out a kind of bootstrap operation with regard to the quantity θ; for how could it be known? Intuitively, you may have no difficulty in seeing the procedure you would use to determine θ from measurements with the counter. But logically, we need to have the calculation about to be given before we can justify that procedure. So, for the time being we'll just have to suppose that θ is a number given to us by our teacher in assigning us this problem; and have faith that in the end we shall understand how our teacher determined it.

Now let us introduce a quantity p which is the probability, in any one second, that any particular nucleus will emit a particle that passes through the counter. We assume the number of nuclei N so large and the half – life so long, that we need not consider N as a variable for this problem. So there are N nuclei, each of which has independently the probability p of sending a particle through our counter in any one second. The quantity p is also, for present purposes, just a number given to us in the statement of the problem, because we have not yet seen in terms of probability theory, the line of reasoning by which we could convert measurements into a numerical value of p (but again, you see intuitively without any hesitation, that p is a way of describing the half – life of the source).

Suppose we were given N and p; what is the probability, on this evidence, that in any one second exactly n particles will pass through the counter? That is the same binomial distribution problem, so the answer is

$$b(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n} \tag{6–79}$$

But in this case there's a good approximation to the binomial distribution, because the number N is enormously large and p enormously small. In the limit $N \to \infty, \quad p \to 0$ in such a way that $Np \to s = \text{const.}$, what happens to (6–79)? To find this, write $p = s/N$, and pass to the limit $N \to \infty$. Then

‡ In practice, there is a question of resolving time; if the particles come too close together we may not be able to see the counts as separate, because the counter experiences a "dead time" after a count, during which it is unable to respond to another particle. We have disregarded those difficulties for this problem and imagined that we have infinitely good resolving time (or, what amounts to the same thing, that the counting rate is so low that there is negligible probability of missing a count). After we have developed the theory, the reader will be asked (Exercise 6.6) to generalize it to take these factors into account.

$$\frac{N!}{(N-n)!}\,p^n = N(N-1)\ldots(N-n+1)\left(\frac{s}{N}\right)^n = s^n\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)$$

which goes into s^n in the limit. Likewise,

$$(1-p)^{N-n} = \left(1-\frac{s}{N}\right)^{N-n} \to e^{-s}$$

and so the binomial distribution (6–79) goes over into the simpler *Poisson distribution:*

$$p(n|N,p) \to p(n|s) = e^{-s}\,\frac{s^n}{n!} \tag{6–80}$$

and it will be handy for us to take this limit. The number s is essentially what the experimenter calls his "source strength," the expectation of number of particles per second.

Now we have enough "formalism" to start solving useful problems. Suppose we are not given the number of particles n in the counter, but only the source strength s. What is the probability, on this evidence, that we shall see exactly c counts in any one second? Using our method of resolving the proposition c into a set of mutually exclusive alternatives, then applying the sum rule and the product rule:

$$p(c|s) = \sum_{n=0}^{\infty} p(cn|s) = \sum_n p(c|ns)\,p(n|s) = \sum_n p(c|n)\,p(n|s) \tag{6–81}$$

since $p(c|ns) = p(c|n)$; i.e. if we knew the actual number n of particles in the counter, it would not matter what s was. This is perhaps made clearer by a diagram, Fig. 6.1 rather like the logic flow diagrams of Fig. (4.3). In this case, we think of the diagram as indicating not only the logical connections, but also the causal ones; s is the physical cause which partially determines n; and then n in turn is the physical cause which partially determines c. Or, to put it another way, s can influence c only through its intermediate influence on n. We saw the same logical situation in the Chapter 5 horseracing example.

Since we have worked out both $p(c|n)$ and $p(n|s)$, we need only substitute them into (8–4); after some algebra we have

$$p(c|s) = \sum_{n=c}^{\infty}\left[\frac{n!}{c!\,(n-c)!}\,\theta^c\,(1-\theta)^{n-c}\right]\left[\frac{e^{-s}\,s^n}{n!}\right] = \frac{e^{-s\theta}\,(s\theta)^c}{c!} \tag{6–82}$$

This is again a Poisson distribution with expectation

$$\langle c \rangle = \sum_{c=0}^{\infty} c\, p(c|s) = s\theta \tag{6–83}$$

Our result is hardly surprising. We have a Poisson distribution with a mean value which is the product of the source strength times the efficiency of the counter. Without going through the analysis, that is just the estimate of c that we would make intuitively, although it is unlikely that anyone could have guessed from intuition that the distribution still has the Poissonian form.

In practice, it is c that is known, and n that is unknown. If we knew the source strength s, and also the number of counts c, what would be the probability, on that evidence, that there were exactly n particles passing through the counter during that second? This is a problem which arises all the time in physics laboratories, because we may be using the counter as a "monitor", and have it set up so that the particles, after going through the counter, then initiate some other reaction which is the one we're really studying. It is important to get the best possible estimates of n, because that is one of the numbers we need in calculating the cross–section of this other reaction. Bayes' theorem gives

$$p(n|cs) = p(n|s)\,\frac{p(c|ns)}{p(c|s)} = \frac{p(n|s)\,p(c|n)}{p(c|s)} \tag{6–84}$$

and all these terms have been found above, so we just have to substitute (6–80) – (6–82) into (6–84). Some terms cancel, and we are left with:

$$p(n|cs) = \frac{e^{-s(1-\theta)}\,[s\,(1-\theta)]^{n-c}}{(n-c)!} \tag{6–85}$$

It is interesting that we *still* have a Poisson distribution, now with parameter $s(1-\theta)$, but shifted upward by c; because of course, n could not be less than c. The expectation over this distribution is

$$\langle n \rangle = \sum_{n} n\, p(n|cs) = c + s(1-\theta) \tag{6–86}$$

So, now what is the best guess the robot can make as to the number of particles responsible for those c counts? Since this is the first time we have faced this issue in a serious way, let us take time for some discussion.

From Posterior Distribution Function to Estimate

Given its posterior *pdf* for some general parameter θ, continuous or discrete, what "best" estimate of θ should the robot make, and what accuracy should it claim? There is no one "right" answer; the problem is really one of decision theory which asks, "What should we do?" This involves value judgments and therefore goes beyond the principles of inference, which ask only "What should we believe?" We shall return to this in Chapters 13 and 14, but for now we give a preliminary discussion adequate for the simple problems being considered.

Laplace (1774) already encountered this problem. The unknown true value of a parameter is θ, and given some data D and prior information I we are to make an estimate $\theta^*(D, I)$ which depends on them in some way. In the jargon of the trade, θ^* is called an "estimator", and nothing prevents one from considering any function of (D, I) whatsoever as a potential estimator. But which estimator is 'best'? Our estimate will have an error $e = (\theta^* - \theta)$, and Laplace gave as a criterion that we should make that estimate which minimizes the expected magnitude $|e|$. He called this the "most advantageous" method of estimation.

Laplace's criterion was generally rejected for 150 years in favor of the least squares method of Gauss and Legendre; we seek the estimate that minimizes the expected square of the error. In these early works it is not always clear whether this means expected over the sampling *pdf* for θ^* or over the posterior *pdf* for θ; the distinction was not always recognized, and the confusion was encouraged by the fact that in some cases considerations of symmetry lead us to the same final conclusion from either. Some of the bad consequences of using the former are noted in Chapter 13. However, it is clear today that in general the former ignores all prior information about θ while the latter takes it into account and is therefore what we want; taking expectations over the posterior *pdf* for θ, the expected squared error of the estimate is

$$\begin{aligned} \langle(\theta-\theta^*)^2\rangle &= \langle\theta^2\rangle - 2\theta^*\langle\theta\rangle + \theta^{*2} \\ &= (\theta^* - \langle\theta\rangle)^2 + (\langle\theta^2\rangle - \langle\theta\rangle^2) \end{aligned} \tag{6–87}$$

The choice

$$\theta^* = \langle\theta\rangle = \int \theta\, p(\theta|D,I)\, d\theta \tag{6–88}$$

that is, the posterior mean value, therefore always minimizes the expected square of the error, over the posterior *pdf* for θ, and the minimum achievable value is the variance of the posterior *pdf*. The second term is the expected square of the deviation from the mean:

$$var(\theta) \equiv \langle(\theta - \langle\theta\rangle)^2\rangle = (\langle\theta^2\rangle - \langle\theta\rangle^2)\,, \tag{6–89}$$

often miscalled the *variance of* θ; of course, it is really the variance of the *probability distribution* that the robot assigns to θ. In any event, the robot can do nothing to minimize it. But the first term can be removed entirely by taking as the estimate just the mean value $\theta^* = \langle\theta\rangle$, which is the optimal estimator by the mean square error criterion.

Evidently, this result holds generally whatever the form of the posterior distribution $p(\theta|DI)$; provided only that $\langle\theta\rangle$ and $\langle\theta^2\rangle$ exist, the mean square error criterion always leads to taking the mean value $\langle\theta\rangle$, (i.e., the "center of gravity" of the posterior distribution) as the "best" guess. The posterior (mean $\pm$ standard deviation) then recommends itself to us as providing a more or less reasonable statement of what we know and how accurately we know it; and it is almost always the easiest to calculate. Furthermore, if the posterior *pdf* is sharp and symmetrical, this cannot be very different pragmatically from any other reasonable estimate. So in practice we use this more than any other. In the Urn inversion problems we simply adopted this procedure without comment.

But this may not be what we really want. We should be aware that there are valid arguments against the posterior mean, and cases where a different rule would better achieve what we want. The squared error criterion says that an error twice as great is considered four times as serious. Therefore, the mean value estimate in effect concentrates its attention most strongly on avoiding the very large (but also very improbable) errors, at the cost of possibly not doing as well as it might with the far more likely small errors.

Because of this, the posterior mean value estimate is quite sensitive to what happens far out in the tails of the *pdf*. If the tails are very unsymmetrical, our estimate could be pulled far away from the central region where practically all the probability lies and common sense tells us the parameter is most likely to be. In a similar way, a single very rich man in a poor village would pull the average wealth of the population far away from anything representative of the real wealth of the people. If we knew this was happening, then that average would be a quite irrational estimate of the wealth of any particular person met on the street.

This concentration on minimizing the large errors leads to another property that we might consider undesirable. Of course, by "large errors" we mean errors that are large *on the scale of* *the parameter* θ. If we redefined our parameter as some nonlinear function $\lambda = \lambda(\theta)$ (for example, $\lambda = \theta^3$, or $\lambda = \log\theta$), an error that is large on the scale of θ might seem small on the scale of λ; and vice versa. But then the posterior mean estimate

$$\lambda^* \equiv \langle\lambda\rangle = \int \lambda\, p(\lambda|D,I)\, d\lambda = \int \lambda(\theta)\, p(\theta|D,I)\, d\theta \tag{6–90}$$

would not in general satisfy $\lambda^* = \lambda(\theta^*)$. Minimizing the mean square error in θ is not the same thing as minimizing the mean square error in $\lambda(\theta)$.

Thus the posterior mean value estimates lack a certain consistency under parameter changes. When we change the definition of a parameter, if we continue to use the mean value estimate, then we have changed the criterion of what we mean by a "good" estimate.

Now let us examine Laplace's original criterion. If we choose an estimator $\theta^+(D,I)$ by the criterion that it minimizes the expected absolute error

$$E \equiv \langle|\theta^+ - \theta|\rangle = \int_{-\infty}^{\theta^+} (\theta^+ - \theta) f(\theta)\, d\theta + \int_{\theta^+}^{\infty} (\theta - \theta^+) f(\theta)\, d\theta \tag{6–91}$$

we require

$$\frac{dE}{d\theta^+} = \int_{-\infty}^{\theta^+} f(\theta)\, d\theta - \int_{\theta^+}^{\infty} f(\theta)\, d\theta = 0 \tag{6–92}$$

or, $P(\theta > \theta^+|DI) = 1/2$; Laplace's "most advantageous" estimator is the *median* of the posterior *pdf*.

But what happens now on a change of parameters $\lambda = \lambda(\theta)$? Suppose that λ is a strict monotonic increasing function of θ (so that θ is in turn a single–valued function of λ and the transformation is reversible). Then it is clear from the above equation that the consistency is restored: $\lambda^+ = \lambda(\theta^+)$.

More generally, all the percentiles have this invariance property: for example, if θ_{35} is the 35 percentile value of θ:

$$\int_{-\infty}^{\theta_{35}} f(\theta)\, d\theta = 0.35 \tag{6–93}$$

then we have at once

$$\lambda_{35} = \lambda(\theta_{35}) \tag{6–94}$$

Thus if we choose as our point estimate and accuracy claim the median and interquartile span over the posterior *pdf*, these statements will have an invariant meaning, independent of how we have defined our parameters. Note that this remains true even when $\langle\theta\rangle$ and $\langle\theta^2\rangle$ diverge, so the mean square estimator does not exist.

Furthermore, it is clear from their derivation from variational arguments, that the median estimator considers an error twice as great to be only twice as serious, so it is less sensitive to what happens far out in the tails of the posterior *pdf* than is the mean value. In current technical jargon, one says that the median is more *robust* with respect to tail variations. Indeed, it is obvious that the median is entirely independent of all variations that do not move any probability from one side of the median to the other; and an analogous property holds for any percentile. One very rich man in a poor village has no effect on the median wealth of the population.

Robustness, in the general sense that the conclusions are insensitive to small changes in the sampling distribution or other conditions, is often held to be a desirable property of an inference procedure, and some authors criticize Bayesian methods, because they suppose that they lack robustness. However, robustness in the usual sense of the word can always be achieved merely by *throwing away* cogent information! It is hard to believe that anyone could really want this if he were aware of it; but those with orthodox training do not think in terms of information content and so do not realize when they are wasting information. Evidently, the issue requires a much more careful discussion, to which we return later in connection with Model comparison.†

In at least some problems, then, Laplace's "most advantageous" estimates have indeed two significant advantages over the more conventional (mean $\pm$ standard deviation). But before the days of computers they were prohibitively difficult to calculate numerically, so the least squares philosophy prevailed as a matter of practical expedience.

Today, the computation problem is relatively trivial, and we can have whatever we want; it is easy to write computer programs which give us the option of displaying either the first and second moments or the quartiles (x_{25}, x_{50}, x_{75}) and only the force of long habit makes us continue to cling to the former.‡

Still another principle for estimation is to take the peak $\hat{\theta}$; or as it is called, the "mode" of the posterior *pdf*. If the prior *pdf* is a constant (or is at least constant in a neighborhood of this peak and not sufficiently greater elewhere), the result is identical with the "maximum likelihood" estimate (MLE) θ' of orthodox statistics. It is usually attributed to R. A. Fisher, who coined that name in the 1920's, although Laplace and Gauss used the method routinely 100 years earlier without feeling any need to give it a special name other than "most probable value". As explained in Chapter 16, Fisher's ideology would not permit him to call it that. The merits and demerits of the MLE are discussed further in Chapters 13 and 17; for the present we are not concerned with philosophical arguments, but wish only to compare the pragmatic results of MLE and other procedures.⋆

Back to the Problem

At this point, a statistician of the "orthodox" school of thought pays a visit to our laboratory. We describe the properties of the counter to him, and invite him to give us *his* best estimate as to the number of particles. He will, of course, use maximum likelihood because his textbooks have told him that (Cramér, 1946; p. 498): "From a theoretical point of view, the most important general method of estimation so far known is the method of maximum likelihood." His likelihood function is, in our notation, $p(c|n)$. The value of n which maximizes it is found, within one unit, from setting

---

† But to anticipate our final conclusion: robustness with respect to sampling distributions is desirable only when we are not sure of the correctness of our model. But then a full Bayesian analysis will take into account all the models considered possible and their prior probabilities. The result automatically achieves the robustness previously sought in intuitive *ad hoc* devices; and some of those devices, such as the 'jackknife' and the 'redescending Psi function' are derived from first principles, as first order approximations to the Bayesian result. The Bayesian analysis of such problems gives us for the first time a clear statement of the circumstances in which robustness is desirable; and then, because Bayesian analysis never throws away information, it gives us more powerful algorithms for achieving robustness.

‡ But in spite of all these considerations, the neat analytical results found in our posterior moments from Urn and binomial models, contrasted with the messy appearance of calculations with percentiles, show that moments have some kind of theoretical significance that percentiles lack. This appears more clearly in Chapter 7.

⋆ One evident pragmatic result is that the MLE fails altogether when the likelihood function has a flat top. But this is just the case we have in the "generalized inverse" problems of current importance in applications; only prior information can give us a reason for preferring one point over another on that flat top.

$$\frac{p(c|n)}{p(c|n-1)} = \frac{n(1-\theta)}{n-c} = 1$$

or

$$(n)_{MLE} = \frac{c}{\theta} \tag{6–95}$$

You may find the difference between the two estimates (6–86) and (6–95) rather startling, if we put in some numbers. Suppose our counter has an efficiency of 10 percent; in other words, $\theta = 0.1$, and the source strength is $s = 100$ particles per second, so that the expected counting rate according to Equation (6–83) is $\langle c \rangle = s\theta = 10$ counts per second. But in this particular second, we got 15 counts. What should we conclude about the number of particles?

Probably the first answer one would give without thinking is that, if the counter has an efficiency of 10 per cent, then in some sense each count must have been due to about 10 particles; so if there were 15 counts, then there must have been about 150 particles. That is, as a matter of fact, exactly what the maximum likelihood estimate (6–95) would be in this case. But what does the robot tell us? Well, it says the best estimate by the mean–square error criterion is only

$$\langle n \rangle = 15 + 100(1 - 0.1) = 15 + 90 = 105. \tag{6–96}$$

More generally, we could write Equation (6–86) this way:

$$\langle n \rangle = s + (c - \langle c \rangle), \tag{6–97}$$

so if you see k more counts than you 'should have' in one second, according to the robot that is evidence for only k more particles, not $10k$.

This example turned out to be quite surprising to some experimental physicists engaged in work along these lines. Let's see if we can reconcile it with our common sense. If we have an average number of counts of 10 per second with this counter, then we would guess, by rules well known, that a fluctuation in counting rate of something like the square root of this, ± 3, would not be at all surprising even if the number of incoming particles per second stayed strictly constant. On the other hand, if the average rate of flow of particles is $s = 100$ per second, the fluctuation in this rate which would not be surprising is $\pm\sqrt{100} = \pm 10$. But this corresponds to only ± 1 in the number of counts.

This shows that you cannot use a counter to measure fluctuations in the rate of arrival of particles, unless the counter has a very high efficiency. If the efficiency is high, then you know that practically every count corresponds to one particle, and you are reliably measuring those fluctuations. If the efficiency is low and you know that there is a definite, fixed source strength, then fluctuations in counting rate are much more likely to be due to things happening in the counter than to actual changes in the rate of arrival of particles.

The same mathematical result, in the disease scenario, means that if a disease is mild and unlikely to cause death, then variations in the observed number of deaths are not reliable indicators of variations in the incidence of the disease. If our prior information tells us that there is a constantly operating basic cause of the disease (such as a contaminated water supply), then a large change in the number of deaths from one year to the next is not evidence of a large change in the number of people having the disease. But if practically everyone who contracts the disease dies immediately, then of course the number of deaths tells us very reliably what the incidence of the disease was, whatever the means of contracting it.

What caused the difference between the Bayes and maximum likelihood solutions? It's due to the fact that we had *prior information* contained in this source strength s. The maximum likelihood estimate simply maximized the probability of getting c counts, given n particles, and that gives you 150. In Bayes' solution, we will multiply this by a prior probability $p(n|s)$ which represents our knowledge of the antecedent situation, before maximizing, and we'll get an entirely different value for the estimate. As we saw in the inversion of Urn distributions, *simple prior information can make a big change in the conclusions that we draw from a data set.*

Exercise 6.5. Generalize the above calculation to take the dead time effect into account; that is, if we know that two or more particles incident on the counter within a short time interval Δt can produce at most only one count, how is our estimate of n changed? These effects are important in many practical situations and there is a voluminous literature on the application of probability theory to them (see the works of Takacs and Bortkiewicz in the References).

Now let's extend this problem a little bit. We are now going to use Bayes' theorem in four problems where there is no quantitative prior information, but only one qualitative fact; and again see the effect that prior information has on our conclusions.

Effects of Qualitative Prior Information.

The situation is depicted in Fig. 6.2:

Two robots, which we shall humanize by naming them Mr. A and Mr. B, have different prior information about the source of the particles. The source is hidden in another room which they are not allowed to enter. Mr. A has no knowledge at all about the source of particles; for all he knows, it might be an accelerating machine which is being turned on and off in an arbitrary way, or the other room might be full of little men who run back and forth, holding first one radioactive source, then another, up to the exit window. Mr. B has one additional qualitative fact; he knows that the source is a radioactive sample of long lifetime, in a fixed position. But he does not know anything about its source strength (except, of course, that it is not infinite because, after all, the laboratory is not being vaporized by its presence. Mr. A is also given assurance that he will not be vaporized during the experiment). They both know that the counter efficiency is 10 per cent: $\theta = 0.1$. Again, we want them to estimate the number of particles passing through the counter, from knowledge of the number of counts. We denote their prior information by I_A, I_B respectively. All right, we commence the experiment. During the first second, $c_1 = 10$ counts are registered. What can Mr. A and Mr. B say about the number n_1 of particles? Bayes' theorem for Mr. A reads,

$$p(n_1|c_1 I_A) = p(n_1|I_A)\,\frac{p(c_1|n_1 I_A)}{p(c_1|I_A)} = \frac{p(n_1|I_A)\,p(c_1|n_1)}{p(c_1|I_A)} \qquad (6\text{–}98)$$

The denominator is just a normalizing constant, and could also be written,

$$p(c_1|I_A) = \sum_{n_1} p(c_1|n_1)\,p(n_1|I_A). \qquad (6\text{–}99)$$

But now we seem to be stuck, for what is $p(n_1|I_A)$? The only information about n_1 contained in I_A is that n_1 is not large enough to vaporize the laboratory. How can we assign prior probabilities on this kind of evidence? This has been a point of controversy for a long time, for in any theory which regards probability as a real physical phenomenon, Mr. A has no basis at all for determining the 'true' prior probabilities $p(n_1)$.

Choice of a Prior. Now, of course, Mr. A is programmed to recognize that there is no such thing as a 'true' probability. As the notation $p(n_1|I_A)$ indicates, the purpose of assigning a prior is to describe his own state of knowledge, and on this he is the final authority. So he does not need to argue the philosophy of it with anyone. We consider in Chapters 11 and 12 some of the general formal principles available to him for translating verbal prior information into prior probability assignments, but in the present discussion we wish only to demonstrate some pragmatic facts, by a prior that represents reasonably the information that n_1 is not infinite, and that for small n_1 there is no prior information that would justify any great variations in $p(n_1|I_A)$. For example, if as a function of n_1 the prior $p(n_1|I_A)$ exhibited features such as oscillations or sudden jumps, that would imply some very detailed prior information about n_1 that Mr. A does not have.

Mr. A's prior should, therefore, avoid all such structure; but this is hardly a formal principle, and so the result is not unique. But it is one of the points to be made from this example, first noted by Jeffreys (1939), that it does not need to be unique because, in a sense, "almost any" prior which is smooth in the region of high likelihood, will lead to substantially the same final conclusions.†

So Mr. A assigns a uniform prior probability out to some large but finite number N,

$$p(n_1|I_A) = \left\{ \begin{array}{cc} 1/N, & 0 \le n_1 < N \\ 0, & N \le n_1 \end{array} \right\}, \qquad (6\text{–}100)$$

which seems to represent his state of knowledge tolerably well. The finite upper bound N is an admittedly *ad hoc* way of representing the fact that the laboratory is not being vaporized. How large could it be? If N were as large as 10^{60}, then not only the laboratory, but our entire galaxy, would be vaporized by the energy in the beam (indeed, the total number of atoms in our galaxy is of the order of 10^{60}). So Mr. A surely knows that N is very much less than that. Of course, if his final conclusions depend strongly on N, then Mr. A will need to analyze his exact prior information and think more carefully about the value of N and whether the abrupt drop in

† We have seen already that in some circumstances, a prior can make a very large difference in the conclusions; but to do this it necessarily 'modulates' the likelihood function in the region of its peak, not its tails.

$p(n_1|I_A)$ at $n_1 = N$ should be smoothed out. Such careful thinking would not be wrong, but it turns out to be unnecessary, for it will soon be evident that details of $p(n_1|I_A)$ for large n_1 are irrelevant to his conclusions.

On With the Calculation! Nicely enough, the $1/N$ cancels out of Equations (6–98), (6–99), and we are left with

$$p(n_1|c_1 I_A) = \left\{ \begin{array}{cc} A\, p(c_1|n_1)\,, & 0 \le n_1 < N \\ 0\,, & N \le n_1 \end{array} \right\} . \tag{6–101}$$

where A is a normalization factor:

$$A^{-1} = \sum_{n=0}^{N-1} p(c|n)\,. \tag{6–102}$$

We have noted, in Equation (6–95), that as a function of n, $p(c|n)$ attains its maximum at $n = c/\theta$ (=100, in this problem). For $n\theta >> c$, $p(c|n)$ falls off like $n^c(1-\theta)^n \simeq n^c e^{-n\theta}$. Therefore, the sum (6–102) converges so rapidly that if N is as large as a few hundred, there is no appreciable difference between the exact normalization factor (6–102) and the sum to infinity.

In view of this, we may as well take advantage of a simplification; *after* applying Bayes' theorem, pass to the limit $N \to \infty$. But let us be clear about the rationale of this; we pass to the limit, not because we believe that N is infinite; we know that it is not. We pass to the limit rather because we know that this will simplify the calculation without affecting the final result; after this passage to the limit, all our calculations pertaining to this model can be performed exactly with the aid of the general summation formula

$$\sum_{m=0}^{\infty} \binom{m+a}{m} m^n x^m = \left(x \frac{d}{dx} \right)^n \frac{1}{(1-x)^{a+1}}\,, \qquad |x| < 1 \tag{6–103}$$

Thus, writing $m = n - c$, we replace (6–102) by

$$A^{-1} \simeq \sum_{n=0}^{\infty} p(c|n) = \theta^c \sum_{m=0}^{\infty} \binom{m+c}{m} (1-\theta)^m = \theta^c \left\{ \frac{1}{[1-(1-\theta)]^{(c+1)}} \right\} = \frac{1}{\theta} \tag{6–104}$$

Exercise (6.6). To better appreciate the quality of this approximation, denote the 'missing' terms in (6–102) by

$$S(N) \equiv \sum_{n=N}^{\infty} p(c|n)$$

and show that the fractional discrepancy between (6–102) and (6–104) is about

$$\delta \equiv S(N)/S(0) \simeq \frac{e^{-N\theta}\,(N\theta)^c}{c!}, \qquad \text{if} \qquad N\theta >> 1\,.$$

From this, show that in the present case ($\theta = 0.1$, $c = 10$), unless the prior information can justify an upper limit N less than about 270, the exact value of N – or indeed, all details of $p(n_1|I_A)$ for $n_1 > 270$ – can make less than one part in 10^4 difference in his conclusions. But it is hard to see how anyone could have any serious use for more than three figure accuracy in the final results; and so this discrepancy would have no effect at all on that final result. What happens for $n_1 \ge 340$, can affect the conclusions less than one part in 10^6, and for $n_1 \ge 400$ it is less than one part in 10^8.

This is typical of the way prior range matters in real problems, and it makes ferocious arguments over this seem rather silly. It is a valid question of principle, but its pragmatic consequences are almost always not just negligibly small; but strictly nil. Yet some writers have claimed that a fundamental qualitative change in the character of the problem occurs between $N = 10^{10}$ and $N = \infty$. The reader may be amused to estimate how much difference this makes in the final numerical results; to how many figures would we need to calculate before it made any difference at all?

Of course, if the prior information should start encroaching on the region $n_1 < 270$, it would then make a difference in the conclusions; but in that case the prior information was indeed cogent for the question being asked, and this is as it should be. Being thus reassured and using the approximation (6–104), we get the result

$$p(n_1|c_1 I_A) = \theta\, p(c_1|n_1) = \binom{n_1}{c_1} \theta^{c_1+1} (1-\theta)^{n_1-c_1}. \tag{6–105}$$

So, for Mr. A, the most probable value of n_1 is the same as the maximum–likelihood estimate:

$$(\hat{n}_1)_A = \frac{c_1}{\theta} = 100 \tag{6–106}$$

while the posterior mean value estimate is calculated as follows:

$$\langle n_1\rangle_A - c_1 = \sum_{n_1=c_1}^{\infty} (n_1 - c_1)\, p(n_1|c_1, I_A) = \theta^{c_1+1}(1-\theta)(c_1+1) \sum_{n_1} \binom{n_1}{n_1 - c_1 - 1} (1-\theta)^{n_1-c_1-1}$$

From (6–103) the sum is equal to

$$\sum_{m=0}^{\infty} \binom{m + c_1 + 1}{m} (1-\theta)^m = \frac{1}{\theta^{c_1+2}} \tag{6–107}$$

and, finally, we get

$$\langle n_1\rangle_A = c_1 + (c_1+1)\frac{1-\theta}{\theta} = \frac{c_1 + 1 - \theta}{\theta} = 109\,. \tag{6–108}$$

Now, how about the other robot, Mr. B? Does his extra knowledge help him here? He knows that there is some definite fixed source strength s. And, because the laboratory is not being vaporized, he knows that there is some upper limit S_0. Suppose that he assigns a uniform prior probability density for $0 \le s < S_0$. Then he will obtain

$$p(n_1|I_B) = \int_0^\infty p(n_1|s)p(s|I_B)ds = \frac{1}{S_0}\int_0^{S_0} p(n_1|s)ds = \frac{1}{S_0}\int_0^{S_0} \frac{s^{n_1} e^{-s}}{n_1!} ds. \tag{6–109}$$

Now, if n_1 is appreciably less than S_0, the upper limit of integration can for all practical purposes, be taken as infinity, and the integral is just unity. So, we have

$$p(n_1|I_B) = p(s|I_B) = \frac{1}{S_0} = \text{const.}, \qquad n_1 < S_0. \tag{6–110}$$

In putting this into Bayes' theorem with $c_1 = 10$, the significant range of values of n_1 will be of the order of 100, and unless his prior information indicates a value of S_0 lower than about 300, we will have the same situation as before; Mr. B's extra knowledge didn't help him at all, and he comes out with the same posterior distribution and the same estimates:

$$p(n_1|c_1 I_B) = p(n_1|c_1 I_A) = \theta\, p(c_1|n_1). \tag{6–111}$$

The Jeffreys Prior. Harold Jeffreys (1939; Chap. 3) proposed a different way of handling this problem. He suggests that the proper way to express "complete ignorance" of a continuous variable known to be positive, is to assign uniform prior probability to its logarithm; *i.e.*, the prior probability density is

$$p(s|I_J) = \frac{1}{s}\,, \qquad (0 \le s < \infty)\,. \tag{6–112}$$

Of course, you can't normalize this, but that doesn't stop you from using it. In many cases, including the present one, it can be used directly because all the integrals involved converge. In almost all cases we can approach this prior as the limit of a sequence of proper (normalizable) priors, with mathematically well–behaved results. If even that does not yield a proper posterior distribution, then the robot is warning us that the data are too uninformative about either very large s or very small s to justify any definite conclusions, and we need to get more evidence before any useful inferences are possible.

Jeffreys justified (6–112) on the grounds of invariance under certain changes of parameters; *i.e.* instead of using the parameter s, what prevents us from using $t \equiv s^2$, or $u \equiv s^3$? Evidently, to assign a uniform prior probability density to s, is not at all the same thing as assigning a uniform prior probability to t; but if we use the Jeffreys prior, we are saying the same thing whether we use s or any power s^m as the parameter.

There is the germ of an important principle here; but it was only recently that the situation has been fairly well understood. When we take up the theory of transformation groups in Chapter 12, we will see that the real justification of Jeffreys' rule cannot lie merely in the fact that the parameter is positive; but that our desideratum of consistency in the sense that equivalent states of knowledge should be represented by equivalent probability assignments, uniquely determines the Jeffreys rule in the case when s is a "scale parameter." Then marginalization theory will reinforce this by deriving it uniquely – without appealing to any principles beyond the basic product and sum rules of probability theory – as the only prior for a scale parameter that is completely uninformative about other parameters that may be in the model.

These arguments and others equally cogent all lead to the same conclusion: the Jeffreys prior is the only correct way to express complete ignorance of a scale parameter. The question then reduces to whether s can properly be regarded as a scale parameter in this problem. However, this line of thought has taken us beyond the present topic; in the spirit of our current problem, we shall just put (6–112) to the test and see what results it gives. The calculations are all very easy, and we find these results:

$$p(n_1|I_J) = \frac{1}{n_1}\,, \qquad (c_1|I_J) = \frac{1}{c_1}\,, \qquad p(n_1|c_1 I_J) = \frac{c_1}{n_1} p(c_1|n_1). \tag{6–113}$$

This leads to the most probable and mean value estimates:

$$(\hat{n}_1)_J = \frac{c_1 - 1 + \theta}{\theta} = 91\,, \qquad \langle n_1 \rangle_J = \frac{c}{\theta} = 100. \tag{6–114}$$

The amusing thing emerges that Jeffreys' prior probability rule just lowers the most probable and posterior mean value estimates by 9 each, bringing the mean value right back to the maximum likelihood estimate!

This comparison is valuable in showing us how little difference there is numerically between the consequences of different prior probability assignments which are not sharply peaked, and helps to put arguments about them into proper perspective. We made a rather drastic change in the prior probabilities, in a problem where there was really very little information contained in the meager data, and it still made less than 10 per cent difference in the result. This is, as we shall see, small compared to the probable error in the estimate which was inevitable in any event. In a more realistic problem where we have more data, the difference would be even smaller.

A useful rule of thumb, illustrated by the comparison of (6–106), (6–108) and (6–114), is that changing the prior probability $p(\alpha|I)$ for a parameter by one power of α has in general about the same effect on our final conclusions as does having one more data point. This is because the likelihood function generally has a relative width $1/\sqrt{n}$, and one more power of α merely adds an extra small slope in the neighborhood of the maximum, thus shifting the maximum slightly. Generally, if we have effectively n independent observations, then the fractional error in an estimate that was inevitable in any event is about $1/\sqrt{n}$,† while the fractional change in estimate due to one more power of α in the prior is about $1/n$.

In the present case, with ten counts, thus ten independent observations, changing from a uniform to Jeffreys prior made just under ten percent difference. If we had 100 counts, the error which is inevitable in any event would be about ten percent, while the difference from the two priors would be less than one percent.

So, from a pragmatic standpoint, arguments about which prior probabilities correctly express a state of "complete ignorance", like those over prior ranges, usually amount to quibbling over pretty small peanuts.⋆ From the standpoint of principle, however, they are important and need to be thought about a great deal, as we shall do in Chapter 12 after becoming familiar with the numerical situation. While the Jeffreys prior is the theoretically correct one, it is in practice a small refinement that makes a difference only in the very small sample case. In the past these issues were argued back and forth endlessly on a foggy philosophical level, without taking any note of the simple facts of actual performance; that is what we are trying to correct here.

The Point of It All

Now we are ready for the interesting part of this problem. For during the next second, we see $c_2 = 16$ counts. What can Mr. A and Mr. B now say about the numbers n_1, n_2 of particles responsible for c_1, c_2? Well, Mr. A has no reason to expect any relation between what happened in the two time intervals, and so to him the increase in counting rate is evidence only of an increase in the number of incident particles. His calculation for the second time interval is the same as before, and he will give us the most probable value

$$(\hat{n}_2)_A = \frac{c_2}{\theta} = 160 \qquad (6\text{–}115)$$

and his mean value estimate is

$$\langle n_2 \rangle_A = \frac{c_2 + 1 - \theta}{\theta} = 169. \qquad (6\text{–}116)$$

Knowledge of c_2 doesn't help him to get any improved estimate of n_1, which stays the same as before.

But now, Mr. B is in an entirely different position than Mr. A; his extra qualitative information suddenly becomes very important. For knowledge of c_2 enables him to improve his previous estimate of n_1. Bayes' theorem now gives

† However, as we shall see later, there are two special cases where the $1/\sqrt{n}$ rule fails: if we are trying to estimate the location of a discontinuity in an otherwise continuous probability distribution, and if different data values are strongly correlated.

⋆ This is most definitely *not* true if the prior probabilities are to describe a definite piece of prior knowledge, as the next example shows.

$$p(n_1|c_2c_1I_B) = p(n_1|c_1I_B)\,\frac{p(c_2|n_1c_1I_B)}{p(c_2|c_1I_B)} = p(n_1|c_1I_B)\,\frac{p(c_2|n_1I_B)}{p(c_2|c_1I_B)} \tag{6–117}$$

Again, the denominator is just a normalizing constant, which we can find by summing the numerator over n_1. We see that the significant thing is $p(c_2|n_1, I_B)$. Using our method of resolving c_2 into mutually exclusive alternatives, this is

$$p(c_2|n_1I_B) = \int_0^\infty p(c_2s|n_1I_B)\,ds = \int_0^\infty p(c_2|sn_1)\,p(s|n_1)\,ds = \int_0^\infty p(c_2|s)\,p(s|n_1)\,ds\,. \tag{6–118}$$

We have already found $p(c|s)$ in (6–82), and we need only

$$p(s|n_1) = p(s|I_B)\,\frac{p(n_1|s)}{p(n_1|I_B)} = p(n_1|s), \quad \text{if } n_1 \ll S_0 \tag{6–119}$$

where we have used Equation (6–110). We have found $p(n_1|s)$ in Equation (6–80), so we have

$$p(c_2|n_1I_B) = \int_0^\infty \left[\frac{e^{-s\theta}(s\theta)^{c_2}}{c_2!}\right]\left[\frac{e^{-s}s^{n_1}}{n_1!}\right]ds = \binom{n_1+c_2}{c_2}\frac{\theta^{c_2}}{(1+\theta)^{n_1+c_2+1}}. \tag{6–120}$$

Substituting (6–111) and (6–120) into (6–117) and carrying out an easy summation to get the denominator, the result is (*not* a binomial distribution):

$$p(n_1|c_2c_1I_B) = \binom{n_1+c_2}{c_1+c_2}\cdot\left(\frac{2\theta}{1+\theta}\right)^{c_1+c_2+1}\cdot\left(\frac{1-\theta}{1+\theta}\right)^{n_1-c_1}\,. \tag{6–121}$$

Note that we could have derived this equally well by direct application of the resolution method:

$$p(n_1|c_2c_1I_B) = \int_0^\infty p(n_1s|c_2c_1I_B)ds = \int_0^\infty p(n_1|sc_1)\,p(s|c_2c_1)ds. \tag{6–122}$$

We have already found $p(n_1|sc_1)$ in (6–85), and it is easily shown that $p(s|c_2c_1) \propto p(c_2|s)\,p(c_1|s)$, which is therefore given by the Poisson distribution (6–82). This, of course, leads to the same rather complicated result (6–121); thus providing another – and rather severe – test of the consistency of our rules.

To find Mr. B's new most probable value of n_1, we set

$$\frac{p(n_1|c_2c_1I_B)}{p(n_1-1|c_2c_1I_B)} = \frac{n_1+c_2}{n_1-c_1}\,\frac{1-\theta}{1+\theta} = 1$$

or,

$$(\hat{n}_1)_B = \frac{c_1}{\theta} + (c_2-c_1)\,\frac{1-\theta}{2\theta} = \frac{c_1+c_2}{2\theta} + \frac{c_1-c_2}{2} = 127 \tag{6–123}$$

His new posterior mean value is also readily calculated, and is equal to

$$\langle n_1\rangle_B = \frac{c_1+1-\theta}{\theta} + (c_2-c_1-1)\frac{1-\theta}{2\theta} = \frac{c_1+c_2+1-\theta}{2\theta} + \frac{c_1-c_2}{2} = 131.5 \tag{6–124}$$

Both estimates are considerably raised, and the difference between most probable and mean value is only half what it was before, suggesting a narrower posterior distribution as we shall confirm presently. If we want Mr. B's estimates for n_2, then from symmetry we just interchange the subscripts 1 and 2 in the above equations. This gives for his most probable and mean value estimates, respectively,

$$(\hat{n}_2)_B = 133\,, \qquad \langle n_2 \rangle_B = 137.5 \tag{6–125}$$

Now, can we understand what is happening here? Intuitively, the reason why Mr. B's extra qualitative prior information makes a difference is that knowledge of both c_1 and c_2 enables him to make a better estimate of the source strength s, which in turn is relevant for estimating n_1. The situation is indicated more clearly by the diagrams, Fig. (6.2). By hypothesis, to Mr. A each sequence of events $n_i \to c_i$ is logically independent of the others, so knowledge of one doesn't help him in reasoning about any other. In each case he must reason from c_i directly to n_i, and no other route is available. But to Mr. B, there are two routes; he can reason directly from c_1 to n_1 as Mr. A does, as described by $p(n_1|c_1 I_A) = p(n_1|c_1 I_B)$; but because of his knowledge that there is a fixed source strength s "presiding over" both n_1 and n_2, he can also reason along the route $c_2 \to n_2 \to s \to n_1$. If this were the *only* route available to him (*i.e.*, if he didn't know c_1), he would obtain the distribution

$$p(n_1|c_2 I_B) = \int_0^\infty p(n_1|s)\, p(s|c_2 I_B)\, ds = \frac{\theta^{c_2+1}}{c_2!(1+\theta)^{c_2+1}} \frac{(n_1+c_2)!}{n_1!(1+\theta)^{n_1}} \tag{6–126}$$

and, comparing the above relations, we see that Mr. B's final distribution (6–121) is, except for normalization, just the product of the ones found by reasoning along his two routes:

$$p(n_1|c_1 c_2 I_B) = (const.) \times p(n_1|c_1 I_B)\, p(n_1|c_2 I_B) \tag{6–127}$$

in consequence of the fact that $p(c_1, c_2|n_1) = p(c_1|n_1)\, p(c_2|n_1)$. The information (6–126) about n_1 obtained by reasoning along the new route $c_2 \to n_2 \to s \to n_1$ thus introduces a "correction factor" in the distribution obtained from the direct route $c_1 \to n_1$, enabling Mr. B to improve his estimates.

This suggests that, if Mr. B could obtain the number of counts in a great many different seconds, $(c_3, c_4, \ldots, c_m)$, he would be able to do better and better; and perhaps in the limit $m \to \infty$ his estimate of n_1 might be as good as the one we found when source strength was considered known exactly. We will check this surmise presently by working out the degree of reliability of these estimates, and by generalizing these distributions to arbitrary m, from which we can obtain the asymptotic forms.

Interval Estimation.

There is still an essential feature missing in the comparison of Mr. A and Mr. B in our particle-counter problem. We would like to have some measure of the degree of reliability which they attach to their estimates, especially in view of the fact that their estimates are so different. Clearly, the best way of doing this would be to draw the entire probability distributions

$$p(n_1|c_2 c_1 I_A) \qquad \text{and} \qquad p(n_1|c_2 c_1 I_B)$$

and from this make statements of the form, "90 per cent of the posterior probability is concentrated in the interval $\alpha < n_1 < \beta$." But, for present purposes, we will be content to give the standard deviations [i.e., square root of the variance as defined in Eq. (6–89)] of the various distributions we have found. An inequality due to Tchebycheff then asserts that, if σ is the standard deviation of any probability distribution over n_1, then the amount P of probability concentrated between the limits $\langle n_1 \rangle \pm t\sigma$ satisfies†

$$P \geq 1 - \frac{1}{t^2} \tag{6–128}$$

This tells us nothing when $t \leq 1$, but it tells us more and more as t increases beyond unity. For example, in any distribution with finite $\langle n \rangle$ and $\langle n^2 \rangle$, at least 3/4 of the probability is contained in the interval $\langle n \rangle \pm 2\sigma$, and at least 8/9 is in $\langle n \rangle \pm 3\sigma$.

Calculation of Variance. The variances σ^2 of all the distributions we have found above are readily calculated. In fact, calculation of any moment of these distributions is easily performed by the general formula (6–103). For Mr. A and Mr. B, and the Jeffreys prior probability distribution, we find the variances

$$\mathrm{Var}(n_1|c_1 I_A) = \frac{(c_1+1)(1-\theta)}{\theta^2} \tag{6–129}$$

$$\mathrm{Var}(n_1|c_2 c_1 I_B) = \frac{(c_1+c_2+1)(1-\theta^2)}{4\theta^2} \tag{6–130}$$

$$\mathrm{Var}(n_1|c_1 I_J) = \frac{c_1(1-\theta)}{\theta^2} \tag{6–131}$$

and the variances for $p(n_2|c_1 c_2 I_X)$ are found from symmetry.

This has been a rather long discussion, so let's summarize all our results so far in a table. We give, for problem 1 and problem 2, the most probable values of number of particles found by Mr. A and Mr. B, and also the (mean value) $\pm$ (standard deviation) estimates.

From Table 6.1 we see that Mr. B's extra information not only has led him to change his estimates considerably from those of Mr. A, but it has enabled him to make an appreciable decrease in his probable error. *Even purely qualitative prior information can greatly alter the conclusions we draw from a data set, and their reliability.* Now in virtually every real problem of scientific inference, we do have qualitative prior information of more or less the kind supposed here. Therefore, any method of inference which fails to take prior information into account is capable of misleading us, in a potentially dangerous way. The fact that it yields a reasonable result in one problem is no guarantee that it will do so in the next.

It is also of interest to ask how good Mr. B's estimate of n_1 would be if he knew only c_2; and therefore had to use the distribution (6–126) representing reasoning along the route $c_2 \to n_2 \to s \to n_1$ of Fig. (6.2). From (6–126) we find the most probable, and the (mean) $\pm$ (standard deviation) estimates

$$\hat{n}_1 = \frac{c_2}{\theta} = 160 \tag{6–132}$$

† Proof: Let $p(x)$ be a probability density over $(-\infty < x < \infty)$, a any real number, and $y \equiv x - \langle x \rangle$. Then

$$a^2(1-P) = a^2 p(|y| > a) = a^2 \int_{|y|>a} p(x)dx \leq \int_{|y|>a} y^2 p(x)dx \leq \int_{-\infty}^{\infty} y^2 p(x)dx = \sigma^2 \ .$$

Writing $a = t\sigma$, this is $t^2(1-P) \leq 1$, the same as Eq. (6–128). This proof includes the discrete cases, since then $p(x)$ is a sum of delta–functions. A large collection of useful Tchebycheff–type inequalities is given by I. R. Savage (1961).

		Problem 1 $c_1 = 10$	Problem 2 $c_1 = 10$ $c_2 = 16$	
		n_1	n_1	n_2
A	most prob. mean ± s.d.	100 109±31	100 109±31	160 169±39
B	most prob. mean ± s.d.	100 109±31	127 131.5±25.9	133 137.5±25.9
J	most prob. mean±s.d.	91 100±30	121.5 127±25.4	127.5 133±25.4

Table 6.1. The Effect of Prior Information on Estimates of n_1 and n_2

$$\text{mean} \pm \text{s.d.} = \frac{c_1+1}{\theta} \pm \frac{\sqrt{(c_2+1)(\theta+1)}}{\theta} = 170 \pm 43.3 \tag{6–133}$$

In this case he would obtain slightly poorer estimate (i.e., a larger probable error) than Mr. A even if the counts $c_1 = c_2$ were the same, because the variance (6–129) for the direct route contains a factor $(1-\theta)$, which gets replaced by $(1+\theta)$ if we have to reason over the indirect route. Thus, if the counter has low efficiency, the two routes give nearly equal reliability for equal counting rates; but if it has high efficiency, $\theta \simeq 1$, then the direct route $c_1 \to n_1$ is far more reliable. Your common sense will tell you that this is just as it should be.

Generalization and Asymptotic Forms.

We conjectured above that Mr. B might be helped a good deal more in his estimate of n_1 by acquiring still more data $\{c_3, c_4, \ldots, c_m\}$. Let's investigate that further. The standard deviation of the distribution (6–85) in which the source strength was known exactly, is only $\sqrt{s(1-\theta)} = 10.8$ for $s = 130$; and from the table, Mr. B's standard deviation for his estimate of n_1 is now about 2.5 times this value. What would happen if we gave him more and more data from other time intervals, such that his estimate of s approached 130? To answer this, note that, if $1 \le k \le m$, we have (now dropping the I_B except in priors because we will be concerned only with Mr. B from now on):

$$p(n_k|c_1 \ldots c_m) = \int_0^\infty p(n_k s|c_1 \ldots c_m)\, ds = \int_0^\infty p(n_k|sc_k)\, p(s|c_1 \ldots c_m)\, ds \tag{6–134}$$

in which we have put $p(n_k|sc_1 \ldots c_m) = p(n_k|sc_k)$ because, from Fig. (6.2), if s is known, then all the c_i with $i \neq k$ are irrelevant for inferences about n_k. The second factor in the integrand of (6–134) can be evaluated by Bayes' theorem:

$$p(s|c_1 \ldots c_m) = p(s|I_B)\,\frac{p(c_1 \ldots c_m|s)}{p(c_1 \ldots c_m|I_B)} = (const.) \times p(s|I_B) p(c_1|s) p(c_2|s) \cdots p(c_m|s)$$

Using (6–82) and normalizing, this reduces to

$$p(s|c_1 \ldots c_m) = \frac{(m\theta)^{c+1}}{c!} s^c e^{-ms\theta} \tag{6–135}$$

where $c \equiv c_1 + \cdots + c_m$ is the total number of counts in the m seconds. The most probable, mean, and variance of the distribution (6–135) are respectively

$$\hat{s} = \frac{c}{m\theta}\,, \qquad \langle s\rangle = \frac{c+1}{m\theta}\,, \qquad \text{var}(s) = \langle s^2\rangle - \langle s\rangle^2 = \frac{c+1}{m^2\theta^2} = \frac{\langle s\rangle}{m\theta} \tag{6–136}$$

So it turns out, as we might have expected, that as $m \to \infty$, the distribution $p(s|c_1 \dots c_m)$ becomes sharper and sharper, the most probable and mean value estimates of s get closer and closer together, and it appears that in the limit we would have just a δ-function:

$$p(s|c_1 \dots c_m) \to \delta(s - s') \tag{6–137}$$

where

$$s' \equiv \lim_{m\to\infty} \frac{c_1 + c_2 + \cdots + c_m}{m\theta} \tag{6–138}$$

But the limiting form (6–137) was found a bit abruptly, as was James Bernoulli's first limit theorem. We might like to see in more detail how the limit is approached, in analogy to the de Moivre–Laplace limit theorem for the binomial (5–10), or the limit (4–62) of the Beta distribution.

For example, expanding the logarithm of (6–135) about its peak $\hat{s} = c/m\theta$, and retaining only through the quadratic terms, we find for the asymptotic formula a Gaussian distribution:

$$p(s|c_1 \dots c_m) \to A \exp\left[-\frac{c(s-\hat{s})^2}{2\hat{s}^2}\right] \tag{6–139}$$

which is actually valid for all s, in the sense that the difference between the left–hand side and right–hand side is small for all s (although their ratio is not close to unity for all s). This leads to the estimate, as $c \to \infty$,

$$(s)_{est} = \hat{s}\left(1 \pm \frac{1}{\sqrt{c}}\right) \tag{6–140}$$

Quite generally, posterior distributions go into a Gaussian form as the data increases, because any function with a single rounded maximum, raised to a higher and higher power, goes into a Gaussian function. In the next Chapter we shall explore the basis of Gaussian distributions in some depth.

So, in the limit, Mr. B does indeed approach exact knowledge of the source strength. Returning to (6–134), both factors in the integrand are now known from (6–85) and (6–135), and so

$$p(n_k|c_1 \dots c_m) = \int_0^\infty \frac{e^{-s(1-\theta)}[s(1-\theta)]^{n_k - c_k}}{(n_k - c_k)!}\,\frac{(m\theta)^{c+1}}{c!}\,s^c e^{-ms\theta}\,ds \tag{6–141}$$

or

$$p(n_k|c_1 \dots c_m) = \frac{(n_k - c_k + c)!}{(n_k - c_k)!c!}\,\frac{(m\theta)^{c+1}(1-\theta)^{n_k - c_k}}{(1 + m\theta - \theta)^{n_k - c_k + c + 1}} \tag{6–142}$$

which is the promised generalization of (6–127). In the limit $m \to \infty$, $c \to \infty$, $(c/m\theta) \to s' = \text{const.}$, this goes into the Poisson distribution

$$p(n_k|c_1 \dots c_m) \to \frac{e^{-s'(1-\theta)}}{(n_k - c_k)!}\,[s'(1-\theta]^{n_k - c_k} \tag{6–143}$$

which is identical with (6–85). We therefore confirm that, given enough additional data, Mr. B's standard deviation can be reduced from 26 to 10.8, compared to Mr. A's value of 31. For finite m, the mean value estimate of n_k from (6–142) is

$$\langle n_k \rangle = c_k + \langle s \rangle (1-\theta) \qquad (6\text{–}144)$$

where $\langle s \rangle = (c+1)/m\theta$ is the mean value estimate of s from (6–136). Equation (6–144) is to be compared to (6–86). Likewise, the most probable value of n_k according to (6–142), is

$$\hat{n}_k = c_k + \hat{s}(1-\theta) \qquad (6\text{–}145)$$

where $\hat{s}$ is given by (6–136).

Note that Mr. B's revised estimates in problem 2 still lie within the range of reasonable error assigned by Mr. A. It would be rather disconcerting if this were not the case, as it would then appear that probability theory is giving Mr. A an over–optimistic picture of the reliability of his estimates. There is, however, no theorem which guarantees this; for example, if the counting rate had jumped to $c_2 = 80$, then Mr. B's revised estimate of n_1 would be far outside Mr. A's limits of reasonable error. But in this case, Mr. B's common sense would lead him to doubt the reliability of his prior information I_B; we would have another example like that in Chapter 4, of a problem where one of those 'something else' alternative hypotheses down at $-100\,db$, which we don't even bother to formulate until they are needed, is resurrected by very unexpected new evidence.

Exercise (6.7). The above results were found using the language of the particle counter scenario. Summarize the final conclusions in the language of the disease incidence scenario, as one or two paragraphs of advice for a medical researcher who is trying to judge whether public health measures are reducing the incidence of a disease in the general population, but has data only on the number of deaths from it. This should, of course, include something about judging under what conditions our model corresponds well to the real world; and what to do if it does not.

Now we turn to a different kind of problem to see some new features that can appear when we use a sampling distribution that is continuous except at isolated points of discontinuity.

Rectangular Sampling Distribution

The following "taxicab problem" is found in several places in the literature. You are traveling on a night train; on awakening from sleep, you notice that the train is stopped at some unknown town, and all you can see is a taxicab with the number 27 on it. What is then your guess as to the number N of taxicabs in the town, which would in turn give a clue as to the size of the town? Almost everybody answers intuitively that there seems to be something about the choice $N_{est} = 2\times 27 = 54$ that recommends itself; but few can offer a convincing rationale for this. The obvious "model" that forms in our minds is that there will be N taxicabs, numbered respectively $(1, \cdots, N)$, and given N, ours is equally likely to be any of them. Given that model, we would then know deductively that $N \geq 27$; but from that point on, one's reasoning depends on one's statistical indoctrination.

Here we study a continuous version of the same problem, in which more than one taxi may be in view, leaving it as an exercise for the reader to find the exact relation between the continuous and discrete problems. We illustrate this by the example of a rectangular sampling distribution in $[0, \alpha]$ where α is the parameter to be estimated, and suggest exercises for the reader which will extend what we learn from it.

We have a data set $D \equiv \{x_1 \cdots x_n\}$ of n observations thought of as "drawn from" this distribution, urn–wise; that is, each datum x_i is assigned independently the *pdf*

$$p(x_i|\alpha, I) = \begin{Bmatrix} \alpha^{-1}, & 0 \leq x_i \leq \alpha < \infty \\ 0\ , & \text{otherwise} \end{Bmatrix} \qquad (6\text{–}146)$$

Then our entire sampling distribution is

$$p(D|\alpha, I) = \alpha^{-n}\,. \qquad\qquad 0 \le \{x_1 \cdots x_n\} \le \alpha \tag{6–147}$$

where for brevity we suppose, in the rest of this section, that when the inequalities following an equation are not all satisfied, the left–hand side is zero. It might seem at first glance that this situation is too trivial to be worth analyzing; yet if one does not see in advance exactly how every detail of the solution will work itself out, there is always something to be learned from studying it. In probability theory, the most trivial–looking problems reveal deep and unexpected things.

The posterior *pdf* for α is by Bayes' theorem,

$$p(\alpha|D, I) = p(\alpha|I)\,\frac{p(D|\alpha, I)}{p(D|I)} \tag{6–148}$$

where $p(\alpha|I)$ is our prior. Now it is evident that any Bayesian problem with a proper (normalizable) prior and a bounded likelihood function must lead to a proper, well–behaved posterior distribution, whatever the data – as long as the data do not themselves contradict any of our other information. If any datum was found to be negative, $x_i < 0$, the model (6–147) would be known deductively to be wrong (put better, the data contradict the prior information I that led us to choose that model). Then the robot crashes, both (6–147) and (6–148) vanishing identically. But any data set for which the inequalities in (6–147) are satisfied is a possible one *according to the model*. Must it then yield a reasonable posterior *pdf*?

Not necessarily! The data could be compatible with the model, but still incompatible with the other prior information. Consider a proper rectangular prior

$$p(\alpha|I) = (\alpha_1 - \alpha_{00})^{-1}\,, \qquad\qquad \alpha_{00} \le \alpha \le \alpha_1 \tag{6–149}$$

where α_{00}, α_1 are fixed numbers satisfying $0 \le \alpha_{00} \le \alpha_1 < \infty$, given to us in the statement of the problem. If any datum were found to exceed the upper prior bound: $x_i > \alpha_1$, then the data and the prior information would again be logically contradictory.

But this is just what we anticipated already in Chapters 1 and 2; we are trying to reason from two pieces of information D, I, each of may be actually a logical conjunction of many different propositions. If there is a contradiction hidden anywhere in the totality of this, there can be no solution (in a set theory context, the set of possibilities that we have prescribed is the empty set) and the robot crashes, in one way or another. So in the following we suppose that the data are consistent with all the prior information – including the prior information that led us to choose this model.† Then the above rules should yield the correct and exact answer to the question we have posed.

The denominator of (6–148) is

$$p(D|I) = \int_R (\alpha_1 - \alpha_{00})^{-1}\, \alpha^{-n}\, d\alpha \tag{6–150}$$

† Of course, in the real world we seldom have prior information that would justify such sharp bounds on x and α and so such sharp contradictions would not arise; but that signifies only that we are studying an ideal limiting case. There is nothing strange about this; in elementary geometry, our attention is directed first to such things as perfect triangles and circles, although no such things exist in the real world. There, also, we are really studying ideal limiting cases of reality; but what we learn from that study enables us to deal successfully with thousands of real situations that arise in such diverse fields as architecture, engineering, astronomy, godesy, stereochemistry, and the artist's rules of perspective. It is the same here.

where the region R of integration must satisfy two conditions:

$$R \;\equiv\; \left\{ \begin{array}{c} \alpha_{00} \le \alpha \le \alpha_1 \\ x_{max} \le \alpha \le \alpha_1 \end{array} \right\} \tag{6–151}$$

and $x_{max} \equiv \max\{x_1 \cdots x_n\}$ is the greatest datum observed. If $x_{max} \le \alpha_{00}$, then in (6–151) we need only the former condition; the numerical values of the data x_i are entirely irrelevant (although the number n of observations remains relevant). If $\alpha_{00} \le x_{max}$, then we need only the latter inequality; the prior lower bound α_{00} has been superceded by the data, and is irrelevant to the problem from this point on.

Substituting (6–147), (6–149) and (6–150) into (6–148) the factor $(\alpha_1 - \alpha_{00})$ cancels out, and if $n > 1$ our general solution reduces to

$$p(\alpha|D,I) = \frac{(n-1)\,\alpha^{-n}}{\alpha_0^{1-n} - \alpha_1^{1-n}}\,, \qquad \alpha_0 \le \alpha \le \alpha_1\,, \qquad n > 1 \tag{6–152}$$

where $\alpha_0 \equiv \max(\alpha_{00},\, x_{max})$.

Small samples. Small values of n often present special situations that might be overlooked in a general derivation. In orthodox statistics, as we shall see in Chapter 17, they can lead to weird pathological results (like an estimator for a parameter which lies outside the parameter space, and so is known deductively to be impossible). In any other area of mathematics, when a contradiction appears one concludes at once that an error has been made. But curiously, in the orthodox literature such pathologies are never interpreted as revealing an error in the orthodox reasoning. Instead they are simply passed over; one proclaims his concern only with large n. But small n proves to be very interesting for us, just because of the fact that Bayesian analysis has no pathological, exceptional cases. As long as we avoid outright logical contradictions in the statement of a problem and use proper priors, the solutions do not break down but continue to make good sense.

It is very instructive to see how Bayesian analysis always manages to accomplish this, which also makes us aware of a subtle point in practical calculation. Thus, in the present case, if $n = 1$ (6–152) appears indeterminate, reducing to $(0/0)$. But if we repeat the derivation from the start for the case $n = 1$, the properly normalized posterior *pdf* for α is found to be, instead of (6–152),

$$p(\alpha|D,I) = \frac{\alpha^{-1}}{\log(\alpha_1/\alpha_0)} \qquad \alpha_0 \le \alpha \le \alpha_1\,, \qquad n = 1\,. \tag{6–153}$$

The case $n = 0$ can hardly be of any use; nevertheless, Bayes' theorem still gives the obviously right answer. For then $D =$ "No data at all", and $p(D|\alpha, I) = p(D|I) = 1$; that is, if we take no data, we shall have no data, whatever the value of α. Then (6–148) reduces, as common sense demands, to

$$p(\alpha|DI) = p(\alpha|I) \qquad \alpha_0 \le \alpha \le \alpha_1\,, \qquad n = 0\,. \tag{6–154}$$

Mathematical Trickery. But now we see a subtle point; the last two results are contained already in (6–152) without any need to go back and repeat the derivation from the start. We need to understand the distinction between the real world problem and the abstract mathematics. For although *in the real problem*, n is by definition a non–negative integer, the *mathematical expression* (6–152) is, from its derivation, necessarily meaningful when n is any complex number. Furthermore, as long as $\alpha_1 < \infty$. it is an entire function of n (that is, bounded and analytic everywhere except the point at infinity). Now in a purely mathematical derivation we are free to make use of whatever analytical properties our functions have, whether or not they would make sense in the real problem. Therefore, since (6–152) can have no singularity at any finite point, we may evaluate it at $n = 1$ by taking the limit as $n \to 1$. But

$$\begin{aligned}\frac{n-1}{\alpha_0^{1-n} - \alpha_1^{1-n}} &= \frac{n-1}{\exp[-(n-1)\log\alpha_0] - \exp[-(n-1)\log\alpha_1]} \\ &= \frac{n-1}{[1-(n-1)\log\alpha_0 + \cdots] - [1-(n-1)\log\alpha_1 + \cdots]} \\ &\to \frac{1}{\log(\alpha_1/\alpha_0)}\ .\end{aligned} \qquad (6\text{–}155)$$

leading to (6–153). Likewise, putting $n = 0$ into (6–152), it reduces to (6–154) because now we have necessarily $\alpha_0 = \alpha_{00}$. Even in extreme, degenerate cases, Bayesian analysis continues to yield the correct results.‡ And it is evident that all moments and percentiles of the posterior distribution are also entire functions of n, so they may be calculated once and for all for all n, taking limiting values whenever the general expression reduces to $(0/0)$ or (∞/∞); this will always yield the same result that we obtain by going back to the beginning and repeating the calculation for that particular value of n.*

If $\alpha_1 < \infty$, the posterior distribution is confined to a finite interval, and so it has necessarily moments of all orders. In fact,

$$\langle \alpha^m \rangle = \frac{n-1}{\alpha_0^{1-n} - \alpha_1^{1-n}} \int_{\alpha_0}^{\alpha_1} \alpha^{m-n}\, d\alpha = \frac{n-1}{n-m-1}\, \frac{\alpha_0^{1+m-n} - \alpha_1^{1+m-n}}{\alpha_0^{1-n} - \alpha_1^{1-n}} \qquad (6\text{–}156)$$

and when $n \to 1$ or $m \to n-1$, we are to take the limit of this expression in the manner of (6–155), yielding the more explicit forms:

‡ The writer became fully convinced of this only after many years of experimentation with hundreds of such cases, and his total failure to produce any pathology as long as the Chapter 2 rules were followed strictly.

* Recognizing this, we see that whenever some mathematical expression is an analytic function of some parameter, we can exploit that fact as a tool for calculation with it, whatever meaning it might have in the original problem. For example, the *number* 2 often appears, and it is almost always in an expression $Q(2)$ which is an analytic function of the *symbol* '2'. Then, if it is helpful, we are free to replace '2' by 'x' and evaluate quantities involving Q by such operations as differentiating with respect to x, or complex integration in the x–plane, *etc*, setting $x = 2$ at the end; and this is perfectly rigorous. Once we have distilled the real problem into one of abstract mathematics, our symbols mean whatever we say they mean; the writer learned this trick from Professor W. W. Hansen of Stanford University, who would throw a class into an uproar when he evaluated an integral, correctly, by differentiating another integral with respect to π.

$$\langle \alpha^m \rangle = \left\{ \begin{array}{ll} \dfrac{\alpha_1^m - \alpha_0^m}{m \log(\alpha_1/\alpha_0)}, & n = 1 \\ \\ \dfrac{(n-1)\,\log(\alpha_1/\alpha_0)}{\alpha_0^{1-n} - a_1^{1-n}}, & m = n - 1 \end{array} \right\} \tag{6–157}$$

In the above results, the posterior distribution is confined to a finite region ($a_0 \le \alpha \le \alpha_1$) and there can be no singular result. Consider now what happens as $\alpha_1 \to \infty$ and we pass to an infinite domain. Then some moments must cease to exist. They are well–behaved in the cases

$$\langle \alpha^m \rangle \to \left\{ \begin{array}{ll} \dfrac{(n-1)\,\alpha_0^m}{n - m - 1}, & m < n - 1, \quad n > 1 \\ \\ \dfrac{(n-1)\,(\alpha_1 - \alpha_0)}{\alpha_0^{1-n} - \alpha_1^{1-n}}, & m = n - 1, \quad n > 1 \end{array} \right\} \tag{6–158}$$

COMMENTS

The calculations which we have done here with ease – in particular, (6–121) and (6–140) – cannot be done with any version of probability theory which does not permit the use of the prior and posterior probabilities needed, and consequently does not allow one to integrate out a nuisance parameter with respect to a prior. It appears to us that Mr. B's results are beyond the reach of orthodox methods. Yet at every stage probability theory as logic has followed the procedures that are determined uniquely by the basic product and sum rules of probability theory; and it has yielded well–behaved, reasonable, and useful results. In some cases, the prior information was absolutely essential, even though it was only qualitative. Later we shall see even more striking examples of this.

But it should not be supposed that this recognition of the need to use prior information is a new discovery. It was emphasized very strongly by J. Bertrand (1889); he gave several examples, of which we quote the last (he wrote in very short paragraphs):

> "The inhabitants of St. Malo [a small French town on the English channel] are convinced; for a century, in their village, the number of deaths at the time of high tide has been greater than at low tide. We admit the fact.
>
> "On the coast of the English channel there have been more shipwrecks when the wind was from the northwest than for any other direction. The number of instances being supposed the same and equally reliably reported, still one will not draw the same conclusions.
>
> "While we would be led to accept as a certainty the influence of the wind on shipwrecks, common sense demands more evidence before considering it even plausible that the tide influences the last hour of the Malouins.
>
> "The problems, again, are identical; the impossibility of accepting the same conclusions shows the necessity of taking into account the prior probability of the cause."

Clearly, Bertrand cannot be counted among those who advocate R. A. Fisher's maxim: "Let the data speak for themselves!" which has so dominated statistics in this Century. The data *cannot* speak for themselves; and they never have, in any real problem of inference.

For example, Fisher advocated the method of maximum likelihood for estimating a parameter; in a sense, this is the value that is indicated most strongly by the data alone. But that takes note of only one of the factors that probability theory (and common sense) requires. For, if we do not supplement the maximum likelihood method with some prior information about which hypotheses we shall consider possible, then it will always lead us inexorably to favor the 'sure thing' hypothesis ST, according to which every tiny detail of the data was inevitable; nothing else could possibly have happened. For the data always have a much higher probability [namely $p(D|ST) = 1$], on ST than on any other hypothesis; ST is always the maximum likelihood solution over the class of all hypotheses. Only our extremely low prior probability for ST can justify our rejecting it.†

Orthodox practice deals with this in part by the device of specifying a model, which is, of course, a means of incorporating some prior information about the phenomenon being observed. But this is incomplete, defining only the parameter space within which we shall seek that maximum; without a prior probability over that parameter space one has no way of incorporating further prior information about the likely values of the parameter, which we almost always have and which is often highly cogent for any rational inference. For example, although a parameter space may extend formally to infinity, in virtually every real problem we know in advance that the parameter is enormously unlikely to be outside some finite domain. This may or may not be crucial, depending on what data set we happen to get.

As the writer can testify from his student days, steadfast followers of Fisher often interpret 'Let the data speak for themselves' as implying that it is somehow unethical – a reprehensible breach of 'scientific objectivity' – to allow one's self to be influenced at all by prior information. It required a few years to perceive, with Bertrand, what a disastrous error this is in real problems.

Prior information is essential also for a different reason, if we are trying to make inferences concerning which mechanism is at work. Fisher would, presumably, insist as strongly as any other scientist that a cause–effect relation requires a physical mechanism to bring it about. But as in St. Malo, the data alone are silent on this; they do not speak for themselves.‡ Only prior information can tell us whether some hypothesis provides a possible mechanism for the observed facts, consistent with the known laws of physics. If it does not, then the fact that it accounts well for the data may give it a high likelihood, but cannot give it any credence. A fantasy that invokes the labors of hordes of little invisible elves and pixies to generate the data would have just as high a likelihood.

It seems that it is not only orthodox statisticians who have denigrated prior information in the twentieth Century. The fantasy writer H. P. Lovecraft once defined 'common sense' as "*merely a stupid absence of imagination and mental flexibility*." Indeed, it is just the accumulation of unchanging prior information about the world that gives the mature person the mental stability that rejects arbitrary fantasies (although we may enjoy diversionary reading of them).

Today, the question whether our present information does or does not provide credible evidence for the existence of a causal effect is a major policy issue, arousing bitter political, commercial, medical, and environmental contention, resounding in courtrooms and legislative halls.* Yet cogent prior information – without which the issue cannot possibly be judged – plays little role in the testimony of 'expert witnesses' with orthodox statistical training, because their standard procedures have no place to use it. We note that Bertrand's clear and correct insight into this appeared the year before Fisher was born; the progress of scientific inference has not always been forward.

---

† Psychologists have noted that small children, when asked to account for some observed fact such as the exact shape of a puddle of spilled milk, have a strong tendency to invent 'sure thing' hypotheses; they have not yet acquired the worldly experience that makes educated adults consider them too unlikely to be considered seriously.

‡ Statisticians, even those who profess themselves disciples of Fisher, have been obliged to develop adages about this, such as '*Correlation does not imply causation.*' or '*A good fit is no substitute for a reason.*' to discourage the kind of thinking that comes automatically to small children, and to adults with untrained minds.

* For some frightening examples, see Gardner (1981). Deliberate suppression of inconvenient prior information is also the main tool of the scientific charlatan.

Thus this Chapter begins and ends with a glance back at Fisher, about whom the reader may find more in Chapter 16.