8

Sufficiency, ancillarity, and all that

In the preceding five chapters we have examined the use of probability theory in problems that, although technically elementary, illustrated a fairly good sample of typical current applications. Now we are in a position to look back over these examples and note some interesting features that they have brought to light. It is useful to understand these features, for tactical reasons. Many times in the past when one tried to conduct inference by applying intuitive *ad hoc* devices instead of probability theory, they would not work acceptably unless some special circumstances were present, and others absent. Thus they were of major theoretical importance in orthodox statistics.

None of the material of the present chapter, however, is really needed in our applications; for us, these are incidental details that take care of themselves as long as we obey the rules. That is, if we merely apply the rules derived in Chapter 2, strictly and consistently in every problem, they lead us to do the right thing and arrive at the optimal inferences for that problem automatically, without our having to take any special note of these things. For us, they have rather a ‘general cultural value’ in helping us to understand better the inner workings of probability theory. One can see much more clearly why it is necessary to obey the Chapter 2 rules, and the predictable consequences of failure to do so.

8.1 Sufficiency

In our examples of parameter estimation, probability theory sometimes does not seem to use all the data that we offer it. In Chapter 6, when we estimated the parameter θ of a binomial distribution from data on n trials, the posterior pdf for θ depended on the data only through the number n of trials and the number r of successes; all information about the order in which success and failure occurred was ignored.

With a rectangular sampling distribution in $\alpha \le x \le \beta$, the joint posterior pdf for $\alpha,\ \beta$ used only the extreme data values $(x_{\rm min},\ x_{\rm max})$ and ignored the intermediate data.

Likewise, in Chapter 7, with a Gaussian sampling distribution and a data set $D \equiv \{x_1, \ldots, x_n\}$, the posterior pdf for the parameters $\mu,\ \sigma$ depended on the data only through n and their first two moments $(\bar{x}, \overline{x^2})$. The $(n-2)$ other properties of the data convey a great deal of additional information of some kind; yet our use of probability theory ignored them.

Is probability theory failing to do all it could here? No, the proofs of Chapter 2 have precluded that possibility; the rules being used are the only ones that can yield unique answers while agreeing with the qualitative desiderata of rationality and consistency. It seems, then, that the unused parts of the data must be *irrelevant* to the question we are asking.[1] But can probability theory itself confirm this conjecture for us in a more direct way?

This introduces us to a quite subtle theoretical point about inference. Special cases of the phenomenon were noted by Laplace (1812, 1824 edn, Supp. V). It was generalized and given its present name 100 years later by Fisher (1922), and its significance for Bayesian inference was noted by H. Jeffreys (1939). Additional understanding of its role in inference was achieved only recently, in the resolution of the 'marginalization paradox' discussed in Chapter 15.

If certain aspects of the data are not used when they are known, then presumably it would not matter (we should come to the same final conclusion) if they were unknown. Thus, if the posterior pdf for a parameter θ is found to depend on the data $D = \{x_1, \ldots, x_n\}$ only through a function $r(x_1, \ldots, x_n)$ (call it 'property R'), then it seems plausible that given r alone we should be able to draw the same inferences about θ. This would confirm that the unused parts of the data were indeed irrelevant in the sense just conjectured.

With a sampling density function $p(x_1 \ldots x_n|\theta)$ and prior $p(\theta|I) = f(\theta)$, the posterior pdf using all the data is

$$p(\theta|DI) = h(\theta|D) = \frac{f(\theta)p(x_1 \ldots x_n|\theta)}{\int \mathrm{d}\theta' \, f(\theta')p(x_1 \ldots x_n|\theta')}. \tag{8.1}$$

Note that we are not assuming independent or exchangeable sampling here; the sampling pdf need not factor in the form $p(x_1 \ldots x_n|\theta) = \Pi_i \, p(x_i|\theta)$ and the marginal probabilities $p(x_i|\theta) = k_i(x_i, \theta)$ and $p(x_j|\theta) = k_j(x_j, \theta)$ need not be the same function. Now carry out a change of variables $(x_1, \ldots, x_n) \to (y_1, \ldots, y_n)$ in the sample space S_x, such that $y_1 = r(x_1, \ldots, x_n)$, and choose $(y_2, \ldots, y_n)$ so that the Jacobian

$$J = \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} \tag{8.2}$$

is bounded and nonvanishing everywhere on S_x. Then the change of variables is a 1:1 mapping of S_x onto S_y, and the sampling density

$$g(y_1, \ldots, y_n|\theta) = J^{-1} p(x_1 \ldots x_n|\theta) \tag{8.3}$$

may be used just as well as $p(x_1 \ldots x_n|\theta)$ in the posterior pdf:

$$h(\theta|D) = \frac{f(\theta)g(y_1, \ldots, y_n|\theta)}{\int \mathrm{d}\theta' \, f(\theta')g(y_1, \ldots, y_n|\theta')} \tag{8.4}$$

since the Jacobian, being independent of θ, cancels out.

Then property R is the statement that for all $\theta \in S_\theta$, (8.4) is independent of $(y_2, \ldots, y_n)$. Writing this condition out as derivatives set to zero, we find that it defines a set of $n - 1$

[1] Of course, when we say that some information is 'irrelevant' we mean only that we don't need it *for our present purpose*; it might be crucially important for some other purpose that we shall have tomorrow.

simultaneous integral equations (actually, only orthogonality conditions) that the prior $f(\theta)$ must satisfy:

$$\int_{S_\theta} \mathrm{d}\theta' \, K_i(\theta, \theta') f(\theta') = 0 \qquad \left\{ \begin{array}{l} \theta \in S_\theta \\ 2 \le i \le n \end{array} \right\}, \tag{8.5}$$

where the ith kernel is

$$K_i(\theta, \theta') \equiv g(y|\theta) \frac{\partial g(y|\theta')}{\partial y_i} - g(y|\theta') \frac{\partial g(y|\theta)}{\partial y_i}, \tag{8.6}$$

and we used the abbreviation $y \equiv (y_1, \ldots, y_n)$, etc. It is antisymmetric: $K_i(\theta, \theta') = -K_i(\theta', \theta)$.

8.2 Fisher sufficiency

If (8.5) holds only for some particular prior $f(\theta)$, then $K_i(\theta, \theta')$ need not vanish; in its dependence on θ' it needs only to be orthogonal to that particular function. But if (8.5) is to hold for all $f(\theta)$, as Fisher (1922) required by implication – by failing to mention $f(\theta)$ – then $K_i(\theta, \theta')$ must be orthogonal to a complete set of functions $f(\theta')$; thus zero almost everywhere for $(2 \le i \le n)$. Noting that the kernel may be written in the form

$$K_i(\theta, \theta') = g(y|\theta)\, g(y|\theta') \frac{\partial}{\partial y_i} \log \left[\frac{g(y|\theta')}{g(y|\theta)} \right], \tag{8.7}$$

this condition may be stated as: given any (θ, θ'), then for all possible samples (that is, all values of $\{y_1, \ldots, y_n; \theta; \theta'\}$ for which $g(y|\theta)\, g(y|\theta') \neq 0$), the ratio $[g(y|\theta')/g(y|\theta)]$ must be independent of the components $(y_2, \ldots, y_n)$. Thus to achieve property R independently of the prior, $g(y|\theta)$ must have the functional form

$$g(y_1, \ldots, y_n|\theta) = q(y_1|\theta) m(y_2, \ldots, y_n). \tag{8.8}$$

Integrating $(y_2, \ldots, y_n)$ out of (8.8), we see that the function denoted by $q(y_1|\theta)$ is, to within a normalization constant, the marginal sampling pdf for y_1.

Transforming back to the original variables, Fisher sufficiency requires that the sampling pdf has the form

$$p(x_1 \ldots x_n|\theta) = p(r|\theta) b(x_1, \ldots, x_n), \tag{8.9}$$

where $p(r|\theta)$ is the marginal sampling density for $r(x_1, \ldots, x_n)$.

Equation (8.9) was given by Fisher (1922). If a sampling distribution factors in the manner (8.8), (8.9), then the sampling pdf for $(y_2, \ldots, y_n)$ is independent of θ. This being the case, he felt intuitively that the values of $(y_2, \ldots, y_n)$ can convey no information about θ; full information should be conveyed by the single quantity r, which he then termed a *sufficient statistic*. But Fisher's reasoning was only a conjecture referring to a sampling theory context. We do not see how it could be proved in that limited context, which did not use the concepts of prior and posterior probabilities.

Probability theory as logic can demonstrate this property directly without any need for conjecture. Indeed, using (8.9) in (8.1), the function $b(x)$ cancels out, and we find immediately the relation

$$h(\theta|D) \propto f(\theta)p(r|\theta). \tag{8.10}$$

Thus, if (8.10) holds, then $r(x_1, \ldots, x_n)$ is a sufficient statistic in the sense of Fisher, and in Bayesian inference with the assumed model (8.1), knowledge of the single quantity r does indeed tell us everything about θ that is contained in the full data set $(x_1, \ldots, x_n)$; and this will be true for all priors $f(\theta)$.

The idea generalizes at once to more variables. Thus, if the sampling distribution factors in the form $g(y_1, \ldots, y_n|\theta) = h(y_1, y_2|\theta)\, m(y_3, \ldots, y_n)$, we would say that $y_1(x_1, \ldots, x_n)$ and $y_2(x_1, \ldots, x_n)$ are jointly sufficient statistics for θ and, in this, θ could be multidimensional. If there are two parameters θ_1, θ_2 such that there is a coordinate system $\{y_i\}$ in which

$$g(y_1, \ldots, y_n|\theta_1\theta_2) = h(y_1|\theta_1)k(y_2|\theta_2)m(y_3, \ldots, y_n), \tag{8.11}$$

then $y_1(x_1, \ldots, x_n)$ is a sufficient statistic for θ_1, and y_2 is a sufficient statistic for θ_2; and so on.

8.2.1 Examples

Our discussion of the Gaussian distribution in Chapter 7 has already demonstrated that it has sufficient statistics [Eqs. (7.25)–(7.30)]. If the data $D = \{y_1, \ldots, y_n\}$ consist of n independent observations y_i, then the sampling distribution with mean and variance μ, σ^2 could be written as

$$p(D|\mu\sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{n}{2\sigma^2}[(\mu - \overline{y})^2 + s^2]\right\}, \tag{8.12}$$

where $\overline{y}, s^2$ are the observed sample mean and variance, Eq. (7.29). Since these are the only properties of the data that appear in the sampling distribution (8.12) – and therefore are the only properties of the data that occur in the joint posterior distribution $p(\mu\sigma|DI)$ – they are jointly sufficient statistics for estimation of μ, σ. The test for sufficiency via Bayes' theorem is often easier to carry out than is the test for factorization (8.11), although of course they amount to the same thing.

Let us examine sufficiency for the separate parameters. If σ is known, then we would find the posterior distribution for μ alone:

$$p(\mu|\sigma DI) = A\frac{p(\mu|I)p(D|\mu\sigma I)}{\int \mathrm{d}\mu\; p(\mu|I)p(D|\mu I)}, \tag{8.13}$$

$$\begin{aligned} p(x_1 \ldots x_n|\mu\sigma I) &= A\exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}, \\ &= A\exp\left\{-\frac{ns^2}{2\sigma^2}\right\} \times \exp\left\{-\frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right\}, \end{aligned} \tag{8.1}$$

where

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{x^2} \equiv \frac{1}{n}\sum_{i} x_i^2, \qquad s^2 \equiv \overline{x^2} - \overline{x}^2, \tag{8.15}$$

$$p(\mu|\sigma DI) \propto p(u|I)\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\mu)\right\} \tag{8.16}$$

are the sample mean, mean square, and variance, respectively. Since now the factor $\exp\{-ns^2/2\overline{s}^2\}$ appears in both numerator and denominator, it cancels out.

Likewise, if μ is known, then the posterior pdf for σ alone is found to be

$$p(\sigma|\mu DI) \propto p(\sigma|I)\sigma^{-n}\exp\left\{-\frac{n}{2\sigma^2}(\overline{x^2} - 2\mu\overline{x} + \mu^2)\right\}. \tag{8.17}$$

Fisher sufficiency was of major importance in orthodox (non-Bayesian) statistics, because it had so few criteria for choosing an estimator. It had, moreover, a fundamental status lacking in other criteria because, for the first time, the notion of *information* appeared in orthodox thinking. If a sufficient statistic for θ exists, it is hard to justify using any other for inference about θ. From a Bayesian standpoint one would be, deliberately, throwing away some of the information in the data that is relevant to the problem.[2]

8.2.2 The Blackwell–Rao theorem

Arguments in terms of information content had almost no currency in orthodox theory, but a theorem given by D. Blackwell and C. R. Rao in the 1940s did establish a kind of theoretical justification for the use of sufficient statistics in orthodox terms. Let $r(x_1, \ldots, x_n)$ be a Fisher sufficient statistic for θ, and let $\beta(x_1, \ldots, x_n)$ be any proposed estimator for θ. By (8.9) the joint pdf for the data conditional on r:

$$p(x_1 \ldots x_n|r\theta) = b(x)p(r|x\theta) = b(x)\delta(r - r(x)) \tag{8.18}$$

is independent of θ. Then the conditional expectation

$$\beta_0(r) \equiv \langle\beta|r\theta\rangle = E(\beta|r\theta) \tag{8.19}$$

is also independent of θ, so β_0 is a function only of the x_i, and so is itself a conceivable estimator for θ, which depends on the observations only through the sufficient statistic: $\beta_0 = E(\beta|r)$. The theorem is then that the ‘quadratic risk’

$$R(\theta, \beta) \equiv E[(\beta - \theta)^2|\theta] = \int \mathrm{d}x_1 \cdots \mathrm{d}x_n\,[\beta(x_1, \ldots, x_n) - \theta]^2 \tag{8.20}$$

satisfies the inequality

$$R(\theta, \beta_0) \leq R(\theta, \beta), \tag{8.21}$$

[2] This rather vague statement becomes a definite theorem when we learn that, if we measure information in terms of entropy, then zero information loss in going from the full data set D to a statistic r is equivalent to sufficiency of r. The beginnings of this appeared long ago, in the Pitman–Koopman theorem (Koopman, 1936; Pitman, 1936); we give a modern version in Chapter 11.

for all θ. If $R(\theta, \beta)$ is bounded, there is equality if and only if $\beta_0 = \beta$; that is, if β itself depends on the data only through the sufficient statistic r.

In other words, given any estimator β for θ, if a sufficient statistic r exists, then we can find another estimator β_0 that achieves a lower or equal risk and depends only on r. Thus the best estimator we can find by the criterion of quadratic risk can always be chosen so that it depends on the data only through r. A proof is given by de Groot (1975, 1986 edn, p. 373); the orthodox notion of risk is discussed further in Chapters 13 and 14. But if a sufficient statistic does not exist, orthodox estimation theory is in real trouble because it wastes information; no single estimator can take note of all the relevant information in the data.

The Blackwell–Rao argument is not compelling to a Bayesian, because the criterion of risk is a purely sampling theory notion that ignores prior information. But Bayesians have a far better justification for using sufficient statistics; it is straightforward mathematics, evident from (8.9) and (8.10) that, if a sufficient statistic exists, Bayes' theorem will lead us to it *automatically*, without our having to take any particular note of the idea. Indeed, far more is true: from the proofs of Chapter 2, Bayes' theorem will lead us to the optimal inferences,[3] whether or not a sufficient statistic exists. So, in Bayesian inference, sufficiency is a valid concept; but it is not a fundamental theoretical consideration, only a pleasant convenience affecting the amount of computation but not the quality of the inference.

We have seen that sufficient statistics exist for the binomial, rectangular, and Gaussian sampling distributions. But consider the Cauchy distribution

$$p(x_1 \dots x_n|\theta I) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}. \tag{8.22}$$

This does not factor in the manner (8.9), and so there is no sufficient statistic. With a Cauchy sampling distribution, it appears that no part of the data is irrelevant; every scrap of it is used in Bayesian inference, and it makes a difference in our inferences about θ (that is, in details of the posterior pdf for θ). Then there can be no satisfactory orthodox estimator for θ; a single function conveys only one piece of information concerning the data, and misses $(n - 1)$ others, all of which are relevant and used by Bayesian methods.

8.3 Generalized sufficiency

What Fisher could not have realized, because of his failure to use priors, is that the proviso *for all priors* is essential here. Fisher sufficiency, Eq. (8.9), is the strong condition necessary to achieve property R independently of the prior. But what was realized only recently is that property R may hold under weaker conditions that depend on which prior we assign. Thus, the notion of sufficiency, which originated in the Bayesian considerations of Laplace, actually has a wider meaning in Bayesian inference than in sampling theory.

[3] That is, optimal in the aforementioned sense that no other procedure can yield unique results while agreeing with our desiderata of rationality.

To see this, note that, since the integral equations (8.5) are linear, we may think in terms of linear vector spaces. Let the class of all priors span a function space (Hilbert space) H of functions on the parameter space S_θ. If property R holds only for some subclass of priors $f(\theta) \in H'$ that span a subspace $H' \subset H$, then in (8.5) it is required only that the projection of $K_i(\theta, \theta')$ onto that subspace vanishes. Then $K_i(\theta, \theta')$ may be an arbitrary function on the complementary function space $(H - H')$ of functions orthogonal to H'.

This new understanding is that, for some priors, it is possible to have 'effective sufficient statistics', even though a sufficient statistic in the sense of Fisher does not exist. Given any specified function $r(x_1, \ldots, x_n)$ and sampling density $p(x_1 \ldots x_n|\theta)$, this determines a kernel $K_i(\theta, \theta')$ which we may construct by (8.6). If this kernel is incomplete (i.e. as (θ, θ', i) vary over their range, the kernel, thought of as a set of functions of θ' parameterized by (θ, i), does not span the entire function space H), then the set of simultaneous integral equations (8.5) has nonvanishing solutions $f(\theta)$. If there are non-negative solutions, they will determine a subclass of priors $f(\theta)$ for which r would play the role of a sufficient statistic.

Then the possibility seems open that, for different priors, different functions $r(x_1, \ldots, x_n)$ of the data may take on the role of sufficient statistics. This means that *use of a particular prior may make certain particular aspects of the data irrelevant. Then a different prior may make different aspects of the data irrelevant*. One who is not prepared for this may think that a contradiction or paradox has been found.

This phenomenon is mysterious only for those who think of probability in terms of frequencies; as soon as we think of probability distributions as *carriers of information* the reason for it suddenly seems trivial and obvious. It really amounts to no more than the principle of Boolean algebra $AA = A$; redundant information is not counted twice. A piece of information in the prior makes a difference in our conclusions only when it tells us something that the data do not tell us. Conversely, a piece of information in the data makes a difference in our conclusions only when it tells us something that the prior information does not. Any information that is conveyed by both is redundant, and can be removed from either one without affecting our conclusions. Thus in Bayesian inference a prior can make some aspect of the data irrelevant simply by conveying some information that is also in the data.

But is this new freedom expressing trivialities, or potentially useful new capabilities for Bayesian inference, which Fisher and Jeffreys never suspected? To show that we are not just speculating about an empty case, note that we have already seen an extreme example of this phenomenon, in the strange properties that use of the binomial monkey prior had in urn sampling (Chapter 6); it made all of the data irrelevant, although with other priors all of the data were relevant.

8.4 Sufficiency plus nuisance parameters

In Section 8.2, the parameter θ might have been multidimensional, and the same general arguments would go through in the same way. The question becomes much deeper if we now suppose that there are two parameters θ, η in the problem, but we are not interested in η, so for us the question of sufficiency concerns only the marginal posterior pdf for θ. Factoring the prior $p(\theta\eta|I) = f(\theta)\, g(\eta|\theta)$, we may write the desired posterior pdf as

$$h(\theta|D) = \frac{\int \mathrm{d}\eta\, p(\theta\eta) f(x_1, \dots, x_n|\theta\eta)}{\int\int \mathrm{d}\theta \mathrm{d}\eta\, p(\theta\eta) f(x_1, \dots, x_n|\theta\eta)} = \frac{f(\theta) F(x_1, \dots, x_n|\theta)}{\int \mathrm{d}\theta\, f(\theta) F(x_1, \dots, x_n|\theta)}, \tag{8.23}$$

where

$$F(x_1, \dots, x_n|\theta) \equiv \int \mathrm{d}\eta\, p(\eta|\theta I) f(x_1, \dots, x_n|\theta, \eta). \tag{8.24}$$

Since this has the same mathematical form as (8.1), the steps (8.5)–(8.9) may be repeated and the same result must follow; given any specified $p(\eta|\theta I)$ for which the integral (8.24) converges, if we then find that the marginal distribution for θ has property R for all priors $f(\theta)$, then $F(x_1, \dots, x_n|\theta)$ must factorize in the form

$$F(x_1, \dots, x_n|\theta) = F^*(r|\theta) B(x_1, \dots, x_n). \tag{8.25}$$

But the situation is entirely different because $F(x_1, \dots, x_n|\theta)$ no longer has the meaning of a sampling density, being a different function for different priors $p(\eta|\theta I)$. Now $\{F, F^*, B\}$ are all functionals of $p(\eta|\theta I)$.[4] Thus the presence of nuisance parameters changes the details, but the general phenomenon of sufficiency is retained.

8.5 The likelihood principle

In applying Bayes' theorem, the posterior pdf for a parameter θ is always a product of a prior $p(\theta|I)$ and a likelihood function $L(\theta) \propto p(D|\theta I)$; the only place where the data appear is in the latter. Therefore it is manifest that

> *Within the context of the specified model*, the likelihood function $L(\theta)$ from data D contains all the information about θ that is contained in D.

For us, this is an immediate and mathematically trivial consequence of the product rule of probability theory, and is no more to be questioned than the multiplication table. Put differently, two data sets D, D' that lead to the same likelihood function to within normalization: $L(\theta) = aL'(\theta)$, where 'a' is a constant independent of θ, have just the same import for any inferences about θ, whether it be point estimation, interval estimation, or hypothesis testing. But for those who think of a probability distribution as a physical phenomenon arising from 'randomness' rather than a carrier of incomplete information, the above quoted statement – since it involves only the sampling distribution – has a meaning independent of the product rule and Bayes' theorem. They call it the 'likelihood principle', and its status as a valid principle of inference has been the subject of long controversy, still continuing today.

An elementary argument for the principle, given by George Barnard (1947), is that irrelevant data ought to cancel out of our inferences. He stated it thus: Suppose that in

[4] In orthodox statistics, $F^*(r|\theta)$ would be interpreted as the sampling density to be expected in a compound experiment in which θ is held fixed but η is varied at random from one trial to the next, according to the distribution $p(\eta|\theta I)$.

addition to obtaining the data D we flip a coin and record the result $Z = H$ or T. Then the sampling probability for all our data becomes, as Barnard would have written it,

$$p(DZ|\theta) = p(D|\theta)p(Z). \tag{8.26}$$

Then he reasoned that, obviously, the result of a coin flip can tell us nothing more about the parameter θ beyond what the data D have to say; and so inference about θ based on DZ ought to be exactly the same as inference based on D alone. From this he drew the conclusion that constant factors in the likelihood must be irrelevant to inferences; that is, inferences about θ may depend only on the ratios of likelihoods for different values:

$$\frac{L_1}{L_2} = \frac{p(DZ|\theta_1 I)}{p(DZ|\theta_2 I)} = \frac{p(D|\theta_1 I)}{p(D|\theta_2 I)}, \tag{8.27}$$

which are the same whether Z is or is not included. This is commonly held to be the first statement of the likelihood principle by an orthodox statistician. It is just what we considered obvious already back in Chapter 4, when we noted that a likelihood is not a probability because its normalization is arbitrary. But not all orthodoxians found Barnard's argument convincing.

Alan Birnbaum (1962) gave the first attempted proof of the likelihood principle to be generally accepted by orthodox statisticians. From the enthusiastic discussion following his paper, we see that many regarded this as a major historical event in statistics. He again appeals to coin tossing, but in a different way, through the principle of Fisher sufficiency plus a 'conditionality principle' which appeared to him more primitive:

Conditionality principle

Suppose we can estimate θ from either of two experiments, E_1 and E_2. If we flip a coin to decide which to do, then the information we get about θ should depend only on the experiment that was actually performed. That is, recognition of an experiment that might have been performed, but was not, cannot tell us anything about θ.

But Birnbaum's argument was not accepted by all orthodox statisticians, and Birnbaum himself seems to have had later doubts. One can criticize the conditionality principle by asking: 'How did you choose the experiments E_1, E_2?' Presumably, they were chosen with some knowledge of their properties. For example, we may know that one kind of experiment may be very good for small θ, a different one for large θ. Suppose that both E_1 and E_2 are most accurate for small θ and that there is a third experiment E_3 which is accurate for large θ. We assume that we chose E_1 and E_2, and the coin flip chose E_1. Then the fact that the coin flip did not choose E_2 need not make recognition of E_2 irrelevant to the inference; the very fact that we included it in our enumeration of experiments worth considering implies some prior knowledge favoring small θ.

In any event, Kempthorne and Folks (1971) and Fraser (1980) continued to attack the likelihood principle and deny its validity. From his failure to attack it when he was attacking almost every other principle of inference, we may infer that R. A. Fisher probably accepted the likelihood principle, although his own procedures did not respect it. But he continued to denounce the use of Bayes' theorem on other ideological grounds. For further discussion, see A.W. F. Edwards (1974), or Berger and Wolpert (1988). The issue becomes even more complex and confusing in connection with the notion of ancillarity, discussed below.

Orthodoxy is obliged to violate the likelihood principle for three different reasons: (1) its central dogma that 'The merit of an estimator is determined by its long-run sampling properties', which makes no reference to the likelihood function; (2) its secondary dogma that the accuracy of an estimate is determined by the width of the sampling distribution for the estimator, which again takes note of the likelihood principle; and (3) procedures in which 'randomization' is held to generate the probability distribution *used in the inference*! These are still being taught, and defended vigorously, by people who do not seem to comprehend that their conclusions are then determined, not by the relevant evidence in the data, but by irrelevant artifacts of the randomization. In Chapter 17 we shall examine the so-called 'randomization tests' of orthodoxy and see how Bayesian analysis deals with the same problems.

Indeed, even coin flip arguments cannot be accepted unconditionally if they are to be taken literally; particularly by a physicist who is aware of all the complicated things that happen in real coin flips, as described in Chapter 10. If there is any logical connection between θ and the coin, so that knowing θ would tell us anything about the coin flip, then knowing the result of the coin flip must tell us something about θ. For example, if we are measuring a gravitational field by the period of a pendulum, but the coin is tossed in that same gravitational field, there is a clear logical connection. Both Barnard's argument and Birnbaum's conditionality principle contain an implicit hidden assumption that this is not the case. Presumably, they would reply that, without saying so explicitly, they really meant 'coin flip' in a more abstract sense of some binary experiment totally detached from θ and the means of measuring it. But then, the onus was on them to define exactly what that binary experiment was, and they never did this.

In our view, this line of thought takes us off into an infinite regress of irrelevancies; in our system, the likelihood principle is already proved as an immediate consequence of the product rule of probability theory, independently of all considerations of coin flips or any other auxiliary experiment. But for those who ignore Cox's theorems, *ad hoc* devices continue to take precedence over the rules of probability theory, and there is a faction in orthodoxy that still militantly denies the validity of the likelihood principle.

It is important to note that the likelihood principle, like the likelihood function, refers only to the context of *a specified model which is not being questioned*; seen in a wider context, it may or may not contain all the information in the data that we need to make the best estimate of θ, or to decide whether to take more data or stop the experiment now. Is there additional external evidence that the apparatus is deteriorating? Or, is there reason to suspect that our model may not be correct? Perhaps a new parameter λ is needed. But to claim that the need for additional information like this is a refutation of the likelihood principle, is only to display a misunderstanding of what the likelihood principle is; it is a 'local' principle, not a 'global' one.

8.6 Ancillarity

Consider estimation of a location parameter θ from a sampling distribution $p(x|\theta I) = f(x - \theta|I)$.[5] Fisher (1934) perceived a strange difficulty with orthodox procedures. Choosing some function of the data $\theta^*(x_1, \ldots, x_n)$ as our estimator, two different data sets might yield the same estimate for θ, yet have very different configurations (such as range, fourth central moments, etc.), and must leave us in a very different state of knowledge concerning θ. In particular, it seemed that a very broad range and a sharply clustered one might lead us to the same actual estimate, but they ought to yield very different conclusions as to the accuracy of that estimate. Yet if we hold that the accuracy of an estimate is determined by the width of the sampling distribution for the estimator, one is obliged to conclude that all estimates from a given estimator have the same accuracy, regardless of the configuration of the sample.

Fisher's proposed remedy was not to question the orthodox reasoning which caused this anomaly, but rather to invent still another *ad hockery* to patch it up: use sampling distributions conditional on some 'ancillary' statistic $z(x_1, \ldots, x_n)$ that gives some information about the data configuration that is not contained in the estimator. In general, a single statistic cannot describe the data configuration fully; this could require as many as $(n - 1)$ ancillary statistics. But Fisher could not always supply them; often they do not exist, because he also demanded that the sampling distribution $p(z|\theta I) = p(z|I)$ for an ancillary statistic must be independent of θ. We do not know Fisher's private reason for imposing this independence, but from a Bayesian viewpoint we can see easily what it accomplishes.

The conditional sampling distribution for the data that Fisher would use is then $p(D|z\theta I)$. In orthodox statistics, this changed sampling distribution can in general lead to different conclusions about θ. But we process this by Bayes' theorem:

$$p(D|z\theta I) = \frac{p(zD|\theta I)}{p(z|\theta I)} = p(D|\theta I)\frac{p(z|D\theta I)}{p(z|\theta I)}. \tag{8.28}$$

Now if $z = z(D)$ is a function only of the data, then $p(z|D\theta I)$ is just a delta-function $\delta[z - z(D)]$; so, if $p(z|\theta I)$ is independent of θ, the conditioned sampling distribution $p(D|z\theta I)$ has the same θ dependence (that is, it yields the same likelihood function) as does the unconditional sampling distribution $p(D|\theta I)$. Put differently, from a Bayesian standpoint what Fisher's procedure accomplishes is nothing at all; the likelihood $L(\theta)$ is unchanged, so any method of inference – whether for point estimation, interval estimation, or hypothesis testing – that respects the likelihood principle will lead to just the same inferences about θ, whether or not we condition on an ancillary statistic. Indeed, in Bayesian analysis, if z is a function only of the data, then the value of z is known from the data, so it is redundant information; whether it is or is not included also in the prior information cannot matter. This is, again, just the principle $AA = A$ of elementary logic that we are obliged to stress so often because orthodoxy does not seem to comprehend its implications.

[5] For example, if the mean of a set of samples is used as the estimator, then, given a set of samples, the observed variation of the mean is called the sampling distribution of the mean.

The fact that Fisher obtained different estimates, depending on whether he did or did not condition on ancillary statistics, indicates only that his unconditioned procedure violated the likelihood principle. On the other hand, if we condition on a quantity Z that is not just a function of the data, then Z conveys additional information that is not in the data; and we must expect that in general this *will* alter our inferences about θ.

Orthodoxy, when asked for the accuracy of the estimate, departs from the likelihood principle a second time by appealing not to any property of the likelihood function from our data set, but rather to the width of the sampling distribution for the estimator – a property of that imaginary collection of data sets that one thought might have been observed but were not. For us, adhering to the likelihood principle, it is the width of the likelihood function, from the one data set that we actually have, that tells us the accuracy of the estimate from that data set; imaginary data sets that were not seen are irrelevant to the question we are asking.[6] Thus, for a Bayesian the question of ancillarity never comes up at all; we proceed directly from the statement of the problem to the solution that obeys the likelihood principle.

8.7 Generalized ancillary information

Now let us take a broader view of the notion of ancillary information, as referring not to Fisher ancillarity (in which the ancillary statistic z is part of the data), but to any additional quantity Z that we do not consider part of the prior information or the data. As before, we define

$$\begin{aligned} \theta &= \text{parameters (interesting or uninteresting)} \\ E &= e_1, \ldots, e_n, \qquad \text{noise} \\ D &= d_1, \ldots, d_n, \qquad \text{data} \\ d_i &= f(t_i\theta) + e_i, \qquad \text{model.} \end{aligned} \tag{8.29}$$

But now we add

$$Z = z_1, \ldots, z_m \qquad \text{ancillary data.} \tag{8.30}$$

We want to estimate θ from the posterior pdf, $p(\theta|DZI)$, and direct application of Bayes' theorem gives

$$p(\theta|DZI) = p(\theta|I)\frac{p(DZ|\theta I)}{p(DZ|I)}, \tag{8.31}$$

in which Z appears as part of the data. But now we suppose that Z has, by itself, no direct relevance to θ:

$$p(\theta|ZI) = p(\theta|I). \tag{8.32}$$

This is the essence of what Fisher meant by the term 'ancillary', although his ideology did not permit him to state it this way (since he admitted only sampling distributions, he was

[6] The width of the sampling distribution for the estimator is the answer to a very different question: How would the estimates vary over the class of all different data sets that we think might have been seen?

obliged to define all properties in terms of sampling distributions). He would say instead that ancillary data have a sampling distribution independent of θ:

$$p(Z|\theta I) = p(Z|I), \tag{8.33}$$

which he would interpret as: θ exerts no causal physical influence on Z. But from the product rule,

$$p(\theta Z|I) = p(\theta|ZI)p(Z|I) = p(Z|\theta I)p(\theta|I), \tag{8.34}$$

we see that from the standpoint of probability theory as logic, (8.32) and (8.33) are equivalent; either implies the other. Expanding the likelihood ratio by the product rule and using (8.33),

$$\frac{p(DZ|\theta I)}{p(DZ|I)} = \frac{p(D|\theta ZI)}{p(D|ZI)}. \tag{8.35}$$

Then, in view of (8.32), we can rewrite (8.31) equally well as

$$p(\theta|DZI) = p(\theta|ZI)\frac{p(D|\theta ZI)}{p(D|ZI)}, \tag{8.36}$$

and now the generalized ancillary information appears to be part of the prior information.

A peculiar property of generalized ancillary information is that the relationship between θ and Z is a reciprocal one; had we been interested in estimating Z but knew θ, then θ would appear as a 'generalized ancillary statistic'. To see this most clearly, note that the definitions (8.32) and (8.33) of an ancillary statistic are equivalent to the factorization:

$$p(\theta Z|I) = p(\theta|I)p(Z|I). \tag{8.37}$$

Now recall how we handled this before, when our likelihood was only

$$L_0(\theta) \propto p(D|\theta I). \tag{8.38}$$

Because of the model equation (8.29), if θ is known, then the probability of getting any datum d_i is just the probability that the noise would have made up the difference:

$$e_i = d_i - f(t_i, \theta). \tag{8.39}$$

So if the prior pdf for the noise is a function

$$p(E|\theta I) = u(e_1, \ldots, e_n, \theta) = u(\{e_i\}, \theta) \tag{8.40}$$

we have

$$p(D|\theta I) = u(\{d_i - f(t_i, \theta)\}, \theta), \tag{8.41}$$

the same function of $\{d_i - f(t_i, \theta)\}$. In the special case of a white Gaussian noise pdf independent of θ, this led to Eq. (7.28).

Our new likelihood function (8.35) can be dealt with in the same way, only in place of (8.41) we shall have a different noise pdf, conditional on Z. Thus the effect of ancillary data is simply to update the original noise pdf:

$$p(E|\theta I) \rightarrow p(E|\theta Z I), \tag{8.42}$$

and in general ancillary data that have any relevance to the noise will affect our estimates of all parameters through this changed estimate of the noise.

In (8.40)–(8.42) we have included θ in the conditioning statement to the right of the vertical stroke to indicate the most general case. But in all the cases examined in the orthodox literature, knowledge of θ would not be relevant to estimating the noise, so what they actually did was the replacement

$$p(E|I) \rightarrow p(E|ZI) \tag{8.43}$$

instead of (8.42).

Also, in the cases we have analyzed, this updating is naturally regarded as arising from a joint sampling distribution, which is a function

$$p(DZ|I) = w(e_1, \ldots, e_n, z_1, \ldots, z_m). \tag{8.44}$$

The previous noise pdf (8.40) is then a marginal distribution of (8.44):

$$p(D|I) = u(e_1 \cdots e_n) = \int \mathrm{d}z_1 \cdots \mathrm{d}z_m \, w(e_1, \ldots, e_n, z_1, \ldots, z_m), \tag{8.45}$$

the prior pdf for the ancillary data is another marginal distribution:

$$p(Z|I) = \int \mathrm{d}e_1 \cdots \mathrm{d}e_n \, w(e_1, \ldots, e_n, z_1, \ldots, z_m), \tag{8.46}$$

and the conditional distribution is

$$p(D|ZI) = \frac{p(DZ|I)}{p(Z|I)} = \frac{w(e_i, z_j)}{v(z_j)}. \tag{8.47}$$

Fisher's original application, and the ironic lesson it had for the relation of Bayesian and sampling theory methods, is explained in the Comments at the end of this chapter, Section 8.12.

8.8 Asymptotic likelihood: Fisher information

Given a data set $D \equiv \{x_1, \ldots, x_n\}$, the log likelihood is

$$\frac{1}{n} \log L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta). \tag{8.48}$$

What happens to this function as we accumulate more and more data? The usual assumption is that, as $n \rightarrow \infty$, the sampling distribution $p(x|\theta)$ is actually equal to the limiting relative frequencies of the various data values x_i. We know of no case where one could actually know this to be true in the real world; so the following heuristic argument is all that is justified. If this assumption were true, then we would have asymptotically, as $n \to \infty$,

$$\frac{1}{n} \log L(\theta) \to \int \mathrm{d}x \; p(x|\theta_0) \log p(x|\theta), \tag{8.49}$$

where θ_0 is the 'true' value, presumed unknown. Denoting the entropy of the 'true' density by

$$H_0 = -\int \mathrm{d}x \; p(x|\theta_0) \log p(x|\theta_0), \tag{8.50}$$

we have for the asymptotic likelihood function

$$\frac{1}{n} \log L(\theta) + H_0 = \int \mathrm{d}x \; p(x|\theta_0) \log \left[\frac{p(x|\theta)}{p(x|\theta_0)} \right] \le 0, \tag{8.51}$$

where, letting $q \equiv p(x|\theta_0)/p(x|\theta)$, we used the fact that, for positive real q, we have $\log(q) \le q - 1$, with equality if and only if $q = 1$. Thus we have equality in (8.51) if and only if $p(x|\theta) = p(x|\theta_0)$ for all x for which $p(x|\theta_0) > 0$. But if two different values θ, θ_0 of the parameter lead to identical sampling distributions, then they are confounded: the data cannot distinguish between them. If the parameter is always 'identified', in the sense that different values of θ always lead to different sampling distributions for the data, then we have equality in (8.51) if and only if $\theta = \theta_0$, so the asymptotic likelihood function $L(\theta)$ reaches its maximum at the unique point $\theta = \theta_0$.

Supposing the parameter multidimensional: $\theta \equiv \{\theta_1, \ldots, \theta_m\}$ and expanding about this maximum, we have

$$\log p(x|\theta) = \log p(x|\theta_0) - \frac{1}{2} \sum_{i,j=1}^{m} \frac{\partial^2 \log p(x|\theta)}{\partial\theta_i \partial\theta_j} \delta\theta_i \delta\theta_j \tag{8.52}$$

or

$$\frac{1}{n} \log \left[\frac{L(\theta)}{L(\theta_0)} \right] = -\frac{1}{2} \sum_{ij} I_{ij} \delta\theta_i \delta\theta_j, \tag{8.53}$$

where

$$I_{ij} \equiv \int \mathrm{d}^n x \; p(x|\theta_0) \frac{\partial^2 \log p(x|\theta)}{\partial\theta_i \partial\theta_j} \tag{8.54}$$

is called the *Fisher information matrix*. It is a useful measure of the 'resolving power' of the experiment; that is, considering two close values θ, θ', how big must the separation $|\theta - \theta'|$ be in order that the experiment can distinguish between them?

8.9 Combining evidence from different sources

We all know that there are good and bad experiments. The latter accumulate in vain. Whether there are a hundred or a thousand, one single piece of work by a real master – by a Pasteur, for example – will be sufficient to sweep them into oblivion.

Henri Poincaré (1904, p. 141)

We all feel intuitively that the totality of evidence from a number of experiments ought to enable better inferences about a parameter than does the evidence of any one experiment. But intuition is not powerful enough to tell us when this is valid. One might think naïvely that if we have 25 experiments, each yielding conclusions with an accuracy of $\pm 10\%$, then by averaging them we get an accuracy of $\pm 10/\sqrt{25} = \pm 2\%$. This seems to be supposed by a method currently in use in psychology and sociology, called meta-analysis (Hedges and Olkin, 1985). Probability theory as logic shows clearly how and under what circumstances it is safe to combine this evidence.

The classical example showing the error of uncritical reasoning here is the old fable about the height of the Emperor of China. Supposing that each person in China surely knows the height of the Emperor to an accuracy of at least ± 1 meter; if there are $N = 1\,000\,000\,000$ inhabitants, then it seems that we could determine his height to an accuracy at least as good as

$$\frac{1}{\sqrt{1\,000\,000\,000}}\,\text{m} = 3 \times 10^{-5}\,\text{m} = 0.03\,\text{mm}, \tag{8.55}$$

merely by asking each person's opinion and averaging the results.

The absurdity of the conclusion tells us rather forcefully that the $\sqrt{N}$ rule is not always valid, even when the separate data values are causally independent; it is essential that they be *logically* independent. In this case, we know that the vast majority of the inhabitants of China have never seen the Emperor; yet they have been discussing the Emperor among themselves, and some kind of mental image of him has evolved as folklore. Then, knowledge of the answer given by one does tell us something about the answer likely to be given by another, so they are not logically independent. Indeed, folklore has almost surely generated a systematic error, which survives the averaging; thus the above estimate would tell us something about the folklore, but almost nothing about the Emperor.

We could put it roughly as follows:

$$\text{error in estimate} = S \pm \frac{R}{\sqrt{N}}, \tag{8.56}$$

where S is the common systematic error in each datum, R is the RMS 'random' error in the individual data values. Uninformed opinions, even though they may agree well among themselves, are nearly worthless as evidence. Therefore sound scientific inference demands that, when this is a possibility, we use a form of probability theory (i.e., a probabilistic model) which is sophisticated enough to detect this situation and make allowances for it.

As a start on this, (8.56) gives us a crude but useful rule of thumb; it shows that, unless we *know* that the systematic error is less than about one-third of the random error, we cannot be sure that the average of one million data values is any more accurate or reliable than the average of ten. As Henri Poincaré put it: 'The physicist is persuaded that one good measurement is worth many bad ones.' Indeed, this has been well recognized by experimental physicists for generations; but warnings about it are conspicuously missing from textbooks written by statisticians, and so it is not sufficiently recognized in the ‘soft’ sciences whose practitioners are educated from those textbooks.

Let us investigate this more carefully using probability theory as logic. Firstly we recall the chain consistency property of Bayes’ theorem. Suppose we seek to judge the truth of some hypothesis H, and we have two experiments which yield data sets A, B, respectively. With prior information I, from the first we would conclude

$$p(H|AI) = p(H|I)\frac{p(A|HI)}{p(A|I)}. \tag{8.57}$$

Then this serves as the prior probability when we obtain the new data B:

$$p(H|ABI) = p(H|AI)\frac{p(B|AHI)}{p(B|AI)} = p(H|I)\frac{p(A|HI)p(B|AHI)}{p(A|I)p(B|AI)}. \tag{8.58}$$

But

$$\begin{aligned} p(A|HI)p(B|AHI) &= p(AB|HI) \\ p(A|I)p(B|AI) &= p(AB|I), \end{aligned} \tag{8.59}$$

so (8.58) reduces to

$$p(H|ABI) = p(H|I)\frac{p(AB|HI)}{p(AB|I)}, \tag{8.60}$$

which is just what we would have found had we used the total evidence $C = AB$ in a single application of Bayes’ theorem. This is the chain consistency property. We see from this that it is valid to combine the *evidence* from several experiments if:

(1) the prior information I is the same in all;
(2) the prior for each experiment includes also the results of the earlier ones.

To study one condition a time, let us leave it as an exercise for the reader to examine the effect of violating (1), and suppose for now that we obey (1) but not (2), but we have from the second experiment alone the conclusion

$$p(H|BI) = p(H|I)\frac{p(B|HI)}{p(B|I)}. \tag{8.61}$$

Is it possible to combine the conclusions (8.57) and (8.61) of the two experiments into a single more reliable conclusion? It is evident from (8.58) that this cannot be done in general; it is not possible to obtain $p(H|ABI)$ as a function of the form

$$p(H|ABI) = f\,[p(H|AI), p(H|BI)]\,, \tag{8.62}$$

because this requires information not contained in either of the arguments of that function. But if it is true that $p(B|AHI) = p(B|HI)$, then from the product rule written in the form

$$p(AB|I) = p(A|BHI)p(B|HI) = p(B|AHI)p(A|HI), \tag{8.63}$$

Table 8.1. *Experiment A.*

	Failures	Successes	Success (%)
Old	16 519	4343	20.8 ± 0.28
New	742	122	14.1 ± 1.10

Table 8.2. *Experiment B.*

	Failures	Successes	Success (%)
Old	3876	14 488	78.9 ± 0.30
New	1233	3907	76.0 ± 0.60

it follows that $p(A|BHI) = p(A|HI)$, and this will work. For this, the data sets A, B must be logically independent in the sense that, given H and I, *knowing either data set would tell us nothing about the other*.

If we do have this logical independence, then it is valid to combine the results of the experiments in the above naïve way, and we will in general improve our inferences by so doing. Meta-analysis, applied without regard to these necessary conditions, can be utterly misleading.

At this point, we are beginning to see the kind of dangerous nonsense that can be produced by those who fail to distinguish between causal independence and logical independence. But the situation is still more subtle and dangerous; suppose one tried to circumvent this by pooling all the data before analyzing them; that is, using (8.60). Let us see what could happen to us.

8.10 Pooling the data

The following data are real, but the circumstances were more complicated than supposed in the following scenario. Patients were given either of two treatments, the old one and a new one, and the number of successes (recoveries) and failures (deaths) were recorded. In experiment A the data were as given in Table 8.1. In which the entries in the last column are of the form $100 \times [p \pm \sqrt{p(1-p)/n}]$, indicating the standard deviation to be expected from binomial sampling. Experiment B, conducted two years later, yielded the data given in Table 8.2. In each experiment, the old treatment appeared slightly but significantly better (that is, the differences in p were greater than the standard deviations). The results were very discouraging to the researchers.

But then one of them had a brilliant idea: let us pool the data, simply adding up in the manner $4343 + 14\,488 = 18\,831$, etc. Then we have the contingency table, Table 8.3. Now the new treatment appears much better with overwhelmingly high significance (the difference is over 20 times the sum of the standard deviations)! They eagerly publish this

Table 8.3. *Pooled data.*

	Failures	Successes	Success (%)
Old	20 395	18 831	48.0 ± 0.25
New	1975	4029	67.1 ± 0.61

gratifying conclusion, presenting only the pooled data; and become (for a short time) famous as great discoverers.

How is such an anomaly possible with such innocent looking data? How can two data sets, each supporting the same conclusion, support the opposite conclusion when pooled? Let the reader, before proceeding, ponder these tables and form your own opinion of what is happening.

The point is that an extra parameter is clearly present. Both treatments yielded much better results two years later. This unexpected fact is, evidently, far more important than the relatively small differences in the treatments. Nothing in the data *per se* tells us the reason for this (better control over procedures, selection of promising patients for testing, etc.) and only prior information about further circumstances of the tests can suggest a reason.

Pooling the data under these conditions introduces a very misleading bias; the new treatment appears better simply because, in the second experiment, six times as many patients were given the new treatment, while fewer were given the old one. The correct conclusion from these data is that the old treatment remains noticeably better than the new one; but another factor is present that is vastly more important than the treatment.

We conclude from this example that pooling the data to estimate a parameter θ is not permissible if the separate experiments involve other parameters $(\alpha, \beta, \ldots)$ which can be different in different experiments. In (8.61)–(8.63) we supposed (by failing to mention them) that no such parameters were present, but real experiments almost always have nuisance parameters which are eliminated separately in drawing conclusions.

In summary, the meta-analysis *procedure* is not necessarily wrong; but when applied without regard to these necessary qualifications it can lead to disaster. But we do not see how anybody could have found all these qualifications by intuition alone. Without the Bayesian analysis there is almost no chance that one could apply meta-analysis safely; the safe procedure is not to mention meta-analysis at all as if it were a new principle, but simply to apply probability theory with *strict adherence to our Chapter 2 rules*. Whenever meta-analysis is appropriate, the full Bayesian procedure automatically reduces to meta-analysis.

8.10.1 Fine-grained propositions

One objection that has been raised to probability theory as logic notes a supposed technical difficulty in setting up problems. In fact, many seem to be perplexed by it, so let us examine the problem and its resolution.

The Venn diagram mentality, noted at the end of Chapter 2, supposes that every probability must be expressed as an additive measure on some set; or, equivalently, that every proposition to which we assign a probability must be resolved into a disjunction of elementary ‘atomic’ propositions. Carrying this supposition over into the Bayesian field has led some to reject Bayesian methods on the grounds that, in order to assign a meaningful prior probability to some proposition such as $W \equiv$ the dog walks, we would be obliged to resolve it into a disjunction $W = W_1 + W_2 + \cdots$ of every conceivable subproposition about how the dog does this, such as

$$\begin{aligned} &W_1 \equiv \text{first it moves the right forepaw, then the left hindleg, then} \ldots \\ &W_2 \equiv \text{first it moves the right forepaw, then the right hindleg, then}\ldots \\ &\ldots \end{aligned}$$

This can be done in any number of different ways, and there is no principle that tells us which resolution is ‘right’. Having defined these subpropositions somehow, there is no evident element of symmetry that could tell us which ones should be assigned equal prior probabilities. Even the professed Bayesian L. J. Savage (1954, 1961, 1962) raised this objection, and thought that it made it impossible to assign priors by the principle of indifference. Curiously, those who reasoned this way seem never to have been concerned about how the orthodox probabilist is to define *his* ‘universal set’ of atomic propositions, which performs for him the same function as would that infinitely fine-grained resolution of the dog’s movements.

8.11 Sam’s broken thermometer

If Sam, in analyzing his data to test his pet theory, wants to entertain the possibility that his thermometer is broken, does he need to enumerate every conceivable way in which it could be broken? The answer is not intuitively obvious at first glance, so let

$$\begin{aligned} A &\equiv \text{Sam's pet theory}, \\ H_o &\equiv \text{the thermometer is working properly}, \\ H_i &\equiv \text{the thermometer is broken in the } i\text{th way}, \ 1 \leq i \leq n, \end{aligned}$$

where, perhaps, $n = 1000$. Then, although

$$p(A|DH_0I) = p(A|H_0I)\frac{p(D|AH_0I)}{p(D|H_0I)} \tag{8.64}$$

is the Bayesian calculation Sam would like to do, it seems that honesty compels him to note 1000 other possibilities $\{H_1, \ldots, H_n\}$, and so he must do the calculation

$$p(A|DI) = \sum_{i=0}^{n} p(AH_i|DI) = p(A|H_0DI)p(H_0|I) + \sum_{i=1}^{n} p(A|H_iDI)p(H_i|DI). \tag{8.65}$$

Now expand the last term by Bayes' theorem:

$$p(A|H_iDI) = p(A|H_iI)\frac{p(D|AH_iI)}{p(D|H_iI)} \tag{8.66}$$

$$p(H_i|DI) = p(H_i|I)\frac{p(D|H_iI)}{p(D|I)}. \tag{8.67}$$

Presumably, knowing the condition of his thermometer does not in itself tell Sam anything about the status of his pet theory, so

$$p(A|H_iI) = p(A|I), \qquad 0 \le i \le n. \tag{8.68}$$

But if he knew the thermometer was broken, then the data would tell him nothing about his pet theory (all this is supposed to be contained in the prior information I):

$$p(A|H_iDI) = p(A|H_iI) = p(A|I), \qquad 1 \le i \le n. \tag{8.69}$$

Then from (8.66), (8.68) and (8.69) we have

$$p(D|AH_iI) = p(D|H_iI), \qquad 1 \le i \le n. \tag{8.70}$$

That is, if he knows the thermometer is broken, and as a result the data can tell him nothing about his pet theory, then his probability of getting those data cannot depend on whether his pet theory is true. Then (8.65) reduces to

$$p(A|DI) = \frac{p(A|I)}{p(D|I)}\left[p(D|AH_0I)p(H_0I) + \sum_{i=1}^{n} p(D|H_iI)p(H_i|I)\right]. \tag{8.71}$$

From this, we see that if the different ways of being broken do not in themselves tell him different things about the data,

$$p(D|H_iI) = p(D|H_1I), \qquad 1 \le i \le n, \tag{8.72}$$

then enumeration of the n different ways of being broken is unnecessary; the calculation reduces to finding the likelihood

$$L \equiv p(D|AH_0I)p(H_0|I) + p(D|H_1I)[1 - p(H_0|I)] \tag{8.73}$$

and only the total probability of being broken,

$$p(\overline{H}_0|I) = \sum_{i=1}^{n} p(H_i|I) = 1 - p(H_0|I), \tag{8.74}$$

is relevant. Sam does not need to enumerate 1000 possibilities. But if $p(D|H_iI)$ can depend on i, then the sum in (8.71) should be over those H_i that lead to different $p(D|H_iI)$. That is, information contained in the variations of $p(D|H_iI)$ would be relevant to his inference, and so they should be taken into account in a full calculation.

Contemplating this argument, common sense now tells us that this conclusion should have been ‘obvious’ from the start. Quite generally, enumeration of a large number of ‘fine-grained’ propositions and assigning prior probabilities to all of them is necessary only if the breakdown into those fine details contains information relevant to the question being asked. If they do not, then only the disjunction of all of the propositions is relevant to our problem, and we need only assign a prior probability directly to it.

In practice, this means that in a real problem there will be some natural end to the process of introducing finer and finer subpropositions; not because it is wrong to introduce them, but because it is unnecessary and it contributes nothing to the solution of the problem. The difficulty feared by Savage does not arise in real problems; and this is one of the many reasons why our policy of assigning probabilities on finite sets succeeds in the real world.

8.12 Comments

There are still a number of interesting special circumstances, less important technically but calling for short discussions.

Trying to conduct inference by inventing intuitive *ad hoc* devices instead of applying probability theory has become such a deeply ingrained habit among those with conventional training that, even after seeing the Cox theorems and the applications of probability theory as logic, many fail to appreciate what has been shown, and persist in trying to improve the results – without acquiring any more information – by adding further *ad hoc* devices to the rules of probability theory. We offer here three observations intended to discourage such efforts, by noting what *information* is and is not contained in our equations.

8.12.1 The fallacy of sample re-use

Richard Cox’s theorems show that, given certain data and prior information D, I, any procedure which leads to a different conclusion than that of Bayes’ theorem, will necessarily violate some very elementary desideratum of consistency and rationality. This implies that a *single* application of Bayes’ theorem with given D, I will extract all the information that is in D, I, relevant to the question being asked. Furthermore, we have already stressed that, if we apply probability theory correctly, there is no need to check whether the different pieces of information used are logically independent; any redundant information will cancel out and will not be used twice.[7]

The feeling persists that, somehow, using the same data again in some other procedure might extract still more information from D that Bayes’ theorem has missed the first time, and thus improve our ultimate inferences from D. Since there is no end to the conceivable arbitrary devices that might be invented, we see no way to prove once and for all that no such attempt will succeed, other than pointing to Cox’s theorems. But for any particular device we can always find a direct proof that it will not work; that is, the device cannot

[7] Indeed, this is a property of any algorithm, in or out of probability theory, which can be derived from a constrained variational principle, because adding a new constraint cannot change the solution if the old solution already satisfied that constraint.

change our conclusions unless it also violates one of our Chapter 2 desiderata of rationality. We consider one commonly encountered example.

Having applied Bayes' theorem with given D, I to find the posterior probability

$$p(\theta|DI) = p(\theta|I)\frac{p(D|\theta I)}{p(D|I)} \tag{8.75}$$

for some parameter θ, suppose we decide to introduce some additional evidence E. Then another application of Bayes' theorem updates that conclusion to

$$p(\theta|EDI) = p(\theta|DI)\frac{p(E|\theta DI)}{p(E|DI)}, \tag{8.76}$$

so the necessary and sufficient condition that the new information will change our conclusions is that, on some region of the parameter space of positive measure, the likelihood ratio in (8.76) differs from unity:

$$p(E|\theta DI) \neq p(E|DI). \tag{8.77}$$

But if the evidence E was something already implied by the data and prior information, then

$$p(E|\theta DI) = p(E|DI) = 1, \tag{8.78}$$

and Bayes' theorem confirms that re-using redundant information cannot change the results. This is really only the principle of elementary logic: $AA = A$.

There is a famous case in which it appeared at first glance that one actually did get important improvement in this way; this leads us to recognize that the meaning of 'logical independence' is subtle and crucial. Suppose we take $E = D$; we simply use the same data set twice. But we act as if the second D were logically independent of the first D; that is, although they are the same data, let us call them D^* the second time we use them. Then we simply ignore the fact that D and D^* are actually one and the same data set, and instead of (8.76)–(8.78) we take, in violation of the rules of probability theory,

$$p(D^*|DI) = p(D^*|I) \quad \text{and} \quad p(D^*|\theta DI) = p(D^*|\theta I). \tag{8.79}$$

Then the likelihood ratio in (8.76) is the same as in the first application of Bayes' theorem, (8.75). We have squared the likelihood function, thus achieving a sharper posterior distribution with apparently more accurate estimate of θ!

It is evident that a fraud is being perpetrated here; by the same argument we could re-use the same data any number of times, thus raising the likelihood function to an arbitrarily high power, and seemingly getting arbitrarily accurate estimates of θ – all from the same original data set D which might consist of only one or two observations.

If we actually had two different data sets D, D^* which were *logically independent*, in the sense that knowing one would tell us nothing about the other – but which happened to be numerically identical – then indeed (8.79) would be valid, and the correct likelihood function from the two data sets *would* be the square of the likelihood from one of them.

Therefore the fraudulent procedure is, in effect, claiming to have twice as many observations as we really have. One can find this procedure actually used and advocated in the literature, in the guise of a 'data dependent prior' (Akaike, 1980). This is also close to the topic of 'meta-analysis' discussed earlier, where ludicrous errors can result from failure to perceive the logical dependence of different data sets which are causally independent.

The most egregious example of attempted sample re-use is in the aforementioned 'randomization tests', in which every one of the $n!$ permutations of the data is thought to contain new evidence relevant to the problem! We examine this astonishing view and its consequences in Chapter 17.

8.12.2 A folk theorem

In ordinary algebra, suppose that we have a number of unknowns $\{x_1, \dots, x_n\}$ in some domain X to be determined, and are given the values of m functions of them:

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n) \\ y_2 &= f_2(x_1, \dots, x_n) \\ &\cdots \\ y_m &= f_m(x_1, \dots, x_n). \end{aligned} \tag{8.80}$$

If $m = n$ and the Jacobian $\partial(y_1, \dots, y_n)/\partial(x_1, \dots, x_n)$ is not zero, then we can in principle solve for the x_i uniquely. But if $m < n$ the system is underdetermined; one cannot find all the x_i because the information is insufficient.

It appears that this well-known theorem of algebra has metamorphosed into a popular folk theorem of probability theory. Many authors state, as if it were an evident truth, that from m observations one cannot estimate more than m parameters. Authors with the widest divergence of viewpoints in other matters seem to be agreed on this. Therefore we almost hesitate to point out the obvious; that nothing in probability theory places any such limitation on us. In probability theory, as our data tend to zero, the effect is not that fewer and fewer parameters can be estimated; given a single observation, nothing prevents us from estimating a million different parameters. What happens as our data tend to zero is that those estimates just relax back to the prior estimates, as common sense tells us they must.

There may still be a grain of truth in this, however, if we consider a slightly different scenario; instead of varying the amount of data for a fixed number of parameters, suppose we vary the number of parameters for a fixed amount of data. Then does the accuracy of our estimate of one parameter depend on how many other parameters we are estimating? We note verbally what one finds, leaving it as an exercise for the reader to write down the detailed equations. The answer depends on how the sampling distributions change as we add new parameters; are the posterior pdfs for the parameters independent? If so, then our estimate of one parameter cannot depend on how many others are present.

But if in adding new parameters they all get correlated in the posterior pdf, then the estimate of one parameter θ might be greatly degraded by the presence of others (uncertainty in the values of the other parameters could then 'leak over' and contribute to the uncertainty in θ). In that case, it may be that some function of the parameters can be estimated more accurately than can any one of them. For example, if two parameters have a high negative correlation in the posterior pdf, then their sum can be estimated much more accurately than can their difference.[8] All these subtleties are lost on orthodox statistics, which does not recognize even the concept of correlations in a posterior pdf.

8.12.3 Effect of prior information

As we noted above, it is obvious, from the general principle of non-use of redundant information $AA = A$, that our data make a difference only when they tell us something that our prior information does not. It should be (but apparently is not) equally obvious that prior information makes a difference only when it tells us something that the data do not. Therefore, whether our prior information is or is not important can depend on which data set we get. For example, suppose we are estimating a general parameter θ, and we know in advance that $\theta < 6$. If the data lead to a negligible likelihood in the region $\theta > 6$, then that prior information has no effect on our conclusions. Only if the data alone would have indicated appreciable likelihood in $\theta > 6$ does the prior information matter.

But consider the opposite extreme: if the data placed practically all the likelihood in the region $\theta > 6$, then the prior information would have overwhelming importance and the robot would be led to an estimate very nearly $\theta^* = 6$, determined almost entirely by the prior information. But in that case the evidence of the data strongly contradicts the prior information, and we would become skeptical about the correctness of the prior information, the model, or the data. This is another case where astonishing new information may cause resurrection of alternative hypotheses that we always have lurking somewhere in our minds.

The robot, by design, has no creative imagination and always believes literally what we tell it; and so, if we fail to tell it about any alternative hypotheses, it will continue to give us the best estimates based on unquestioning acceptance of the hypothesis space that we gave it – right up to the point where the data and the prior information become logically contradictory – at which point, as noted at the end of Chapter 2, the robot crashes.

In principle, a single data point could determine accurate values of a million parameters. For example, if a function $f(x_1, x_2, \ldots)$ of one million variables takes on the value $\sqrt{2}$ only at a single point, and we learn that $f = \sqrt{2}$ exactly, then we have determined one million variables exactly. Or, if a single parameter is determined to an accuracy of 12 decimal digits, a simple mapping can convert this into estimates of six parameters to two digits each. But this gets us into the subject of 'algorithmic complexity', which is not our present topic.

8.12.4 Clever tricks and gamesmanship

Two very different attitudes toward the technical workings of mathematics are found in the literature. In 1761, Leonhard Euler complained about isolated results which 'are not based

[8] We shall see this in Chapter 18, in the theory of seasonal adjustment in economics. The phenomenon is demonstrated and discussed in detail in Jaynes (1985e); conventional non-Bayesian seasonal adjustment loses important information here.

on a systematic method' and therefore whose 'inner grounds seem to be hidden'. Yet in the 20th century, writers as diverse in viewpoint as Feller and de Finetti are agreed in considering computation of a result by direct application of the systematic rules of probability theory as dull and unimaginative, and revel in the finding of some isolated clever trick by which one can see the answer to a problem without any calculation.

For example, Peter and Paul toss a coin alternately starting with Peter, and the one who first tosses 'heads' wins. What are the probabilities p, p' for Peter or Paul to win? The direct, systematic computation would sum $(1/2)^n$ over the odd and even integers:

$$p = \sum_{n=0}^{\infty} \frac{1}{2^{2n+1}} = \frac{2}{3}, \qquad p' = \sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{3}. \tag{8.81}$$

The clever trick notes instead that Paul will find himself in Peter's shoes if Peter fails to win on the first toss: *ergo*, $p' = p/2$, so $p = 2/3$, $p' = 1/3$.

Feller's perception was so keen that in virtually every problem he was able to see a clever trick; and then gave only the clever trick. So his readers get the impression that:

(1) probability theory has no systematic methods; it is a collection of isolated, unrelated clever tricks, each of which works on one problem but not on the next one;
(2) Feller was possessed of superhuman cleverness;
(3) only a person with such cleverness can hope to find new useful results in probability theory.

Indeed, clever tricks do have an aesthetic quality that we all appreciate at once. But we doubt whether Feller, or anyone else, was able to see those tricks on first looking at the problem.

We solve a problem for the first time by that (perhaps dull to some) direct calculation applying our systematic rules. *After* seeing the solution, we may contemplate it and see a clever trick that would have led us to the answer much more quickly. Then, of course, we have the opportunity for gamesmanship by showing others only the clever trick, scorning to mention the base means by which we first found the answer. But while this may give a boost to our ego, it does not help anyone else.

Therefore we shall continue expounding the systematic calculation methods, because they are the only ones which are guaranteed to find the solution. Also, we try to emphasize *general* mathematical techniques which will work not only on our present problem, but on hundreds of others. We do this even if the current problem is so simple that it does not require those general techniques. Thus we develop the very powerful algorithms involving group invariance, partition functions, entropy, and Bayes' theorem, that do not appear at all in Feller's work. For us, as for Euler, these are the solid meat of the subject, which make it unnecessary to discover a different new clever trick for each new problem.

We learned this policy from the example of George Pólya. For a century, mathematicians had been, seemingly, doing their best to conceal the fact that they were finding their theorems first by the base methods of plausible conjecture, and only afterward finding the 'clever trick' of an effortless, rigorous proof. Pólya (1954) gave away the secret in his *Mathematics and Plausible Reasoning*, which was a major stimulus for the present work.

Clever tricks are always pleasant diversions, and useful in a temporary way, when we want only to convince someone as quickly as possible. Also, they can be valuable in understanding a result; having found a solution by tedious calculation, if we can then see a simple way of looking at it that would have led to the same result in a few lines, this is almost sure to give us a greater confidence in the correctness of the result, and an intuitive understanding of how to generalize it. We point this out many times in the present work. But the road to success in probability theory goes first through mastery of the general, systematic methods of permanent value. For a teacher, therefore, maturity is largely a matter of overcoming the urge to gamesmanship.