Sankhyā : The Indian Journal of Statistics
Special issue in memory of D. Basu
2002, Volume 64, Series A, Pt. 3, pp 722-739

A HOLISTIC ANALYSIS OF POISSON DATA WITH APPLICATION TO A TRIAL OF SELENIUM AND CANCER DEATHS

By DENNIS V. LINDLEY

SUMMARY. We consider a comparative trial of a new treatment against a standard, where the data are Poisson. The novelty is that the likelihood for the data is placed in the scientific context and the model incorporates both data uncertainty and initial uncertainty about the parameters. Both estimation and hypothesis-testing are discussed. The analysis is applied to a clinical trial of the effect of selenium on cancer deaths. Several assumptions have been made to simplify the analysis; in a final section it is shown how these might be relaxed.

DEDICATION. Dev Basu's work was marked by profound simplicity; he would take a deep problem and analyse it with outstanding clarity, revealing a solution that was correct and, above all, had that great virtue of simplicity. In offering this paper as a tribute to the memory of one of the finest statisticians it has been my good fortune to meet, I have tried to take a simple problem and analyse it in its complicated, real-world context, in the hope that the offered solution will have a few of the features that he would have produced.

1. Introduction

In almost all analyses of data that are reported in the statistical literature, the data are considered in isolation from the general scientific environment in which they were obtained, and a serious effort is made to 'let the data speak for themselves'. In many cases scientific knowledge is invoked in the construction of the statistical model; but there are many analyses in which the model is merely selected 'off the shelf' from the collection of standard models with which the statistician is familiar, a tendency which is encouraged by the availability of statistical packages. Even when the model

Paper received December 2001.
AMS (2000) *subject classification.* Primary 62F15; secondary 62P10, 62P30.
Keywords and phrases. Reliability, clinical trials, Poisson, exponential, negative binomial, F-distribution, Bayes factor, elimination of nuisance parameters, likelihood, χ^2-distribution, estimation, hypothesis tests, conjugate distributions.

is thoughtfully constructed with the help of scientific advice, once settled on, the statistics proceeds entirely within the purview of the model for the data. We argue here that this attitude is unsound and that at every stage, the statistics should take cognizance of the milieu in which the data were obtained, and be in close contact with scientific opinion at the time. This approach we term 'holistic' in that it attempts to present the data analysis only as part of the whole scientific study. The general philosophy has been discussed by Lindley (2000); here we take a specific situation, in which a novel treatment is being compared with the standard, and model it in a wider sense than statisticians ordinarily mean when they refer to modelling, incorporating parameter, as well as data, uncertainty. After constructing the holistic model, we next eliminate nuisance parameters, and then describe the information about the treatment effect, referring both to estimation and to tests of hypotheses. The ideas are then applied to some data on a clinical trial of selenium which were obtained under interesting, but not exceptional, circumstances, with results that differ from those proceeding from more conventional procedures like maximum likelihood. One interesting point to emerge is that scientific knowledge about nuisance parameters can have a substantial influence on one's conclusions about the parameter of primary interest, here the treatment effect. The methods make some restrictive assumptions and in a final section we describe, in outline, how these might be weakened without affecting the basic methodology.

2. The Statistical Model

The situation studied is one in which "units" are put under test in standard conditions and the number of "failures" observed; other units similarly tested under modified conditions, the object being to see whether the modification is effective in reducing the failure rate. The obvious application is to engineering reliability but the data to be discussed here arises from medicine and the terminology appropriate to that field will be used; where the units are patients, either given a placebo or a treatment in a clinical trial, and failure is death. It is supposed that the trial has been carefully designed so that any differences found between placebo and treatment can only be attributed to genuine effects of the treatment and not confounded with other factors. It is supposed that, conditional on parameters yet to be described, the observations with the placebo are independent and identically distributed; and similarly those for the treatment though for possibly different parameters values; and finally these two distributions are independent,

again conditional on the parameters. These are standard assumptions and scarcely require any discussion. The next assumption is more limiting.

ASSUMPTION 1. *The numbers of deaths, both with the placebo and treatment, have Poisson distributions.*

Assumption 1 has been made because the resulting mathematical analysis is rather simple, leading, we hope, to clearer appreciation of the holistic approach. The methodology is very general and may be used in more complicated cases, enabling a wide class of distributions to be accommodated.

Denote by λ the death rate with the placebo, and $\lambda\theta$ that for the treatment, so that $\theta < 1$, $\theta = 1$ and $\theta > 1$ correspond to an improvement, no change, or a deterioration as a result of the treatment. If (T, U) are respectively the total times on test and (r, s) the observed numbers of deaths for placebo and treatment, the likelihood is

$$\lambda^{r+s}\theta^s \exp\{-\lambda(T + \theta U)\}. \tag{1}$$

In a clinical trial, like that discussed below, there will normally be a balance between placebo and treatment with $T = U$.

Most conventional statistical analyses would effectively regard (1) as the complete model upon which the statistical conclusions would be based. For the likelihood school, (1) is literally the total model; the frequentist school would consider (1) but extend it to include values of r and s (and perhaps T and U) beyond those actually observed in the experiment, to embrace other possibilities. This extension is necessary to provide a sample space needed, for example, to provide a tail-area, significance test. An example will occur at the end of Section 4. The holistic approach requires more than the likelihood provided by (1).

Essentially (1) gives the uncertainty, described through probability, of the data (r, s, T, U) conditional on the parameters (λ, θ). Because only data uncertainty is involved, whether by the frequentist or likelihood schools, this is often referred to as letting the data 'speak for themselves'. The holistic attitude says that uncertainty resides not only in the data but also in the parameters, here (λ, θ), and consequently their uncertainty must also be described, and that by probability. This description can only be provided in the clinical trial by the collective medical experience of those involved. (Reports of clinical trials typically have many authors; the one used below had 16.) These considerations lead us to describe probability distributions for λ and θ, in addition to those for the data, and hence to the Bayesian paradigm.

3. The Basic Failure Rate

The designers of a clinical trial will ordinarily have information about the death-rate under normal circumstances that can be extrapolated to those patients receiving the placebo, so that some knowledge of λ is reasonable. We suppose

ASSUMPTION 2. *For some positive values a and b termed hyperparameters,*

$$p(\lambda) = \lambda^{a-1} \exp(-\lambda b) b^a / (a-1)!. \tag{2}$$

(An alternative interpretation is that $2\lambda b$ is χ^2 on $2a$ degrees of freedom.) The assumption is again made partly for mathematical convenience and because, by choice of a and b, it allows the clinician to select any mean and variance. These are

$$E(\lambda) = a/b \text{ and } var(\lambda) = a/b^2 \tag{3}$$

Although the hyperparameters must necessarily be strictly positive, $a > 0$, $b > 0$, for (2) to integrate to 1, the limiting case, $a = b = 0$, has often occurred in the literature. This gives $p(\lambda) \propto 1/\lambda$ and is improper in the sense that it is not integrable over the whole range $(0, \infty)$ of possible values of λ. In particular, it has been advocated as a reference distribution to be used by those who wish to employ the methods of this paper, but also want to describe the information in the data without reference to its context. No clinician could possibly have opinions about the failure rate that were represented by $1/\lambda$ and its use here seems inappropriate.

Even in the present, simple context, it may be hard to think about a rather abstract concept like λ, whose value is unlikely ever to be determined precisely, and it may be preferable to think about a quantity, familiar to the scientist, whose value will ultimately be realized. The technique is not new and had been used most effectively by Kadane et al. (1980). A real possibility here is to use r, the number of deaths amongst patients in the trial receiving the placebo. We have

$$\begin{aligned} p(r) &= \int_0^\infty p(r|\lambda) p(\lambda) d\lambda = \int_0^\infty \frac{\exp(-\lambda T)(\lambda T)^r}{r!} \cdot \frac{\lambda^{a-1} \exp(-\lambda b) b^a}{(a-1)!} \cdot d\lambda \\ &= \frac{b^a (r+a-1)!\, T^r}{r!\, (a-1)!\, (T+b)^{r+a}}, \end{aligned}$$

a negative binomial density of index a and parameter $b/(T+b)$, with

$$E(r) = Ta/b \text{ and } var(r) = Ta(T+b)/b^2. \tag{4}$$

Again assessment of these by the engineer will provide values for the hyperparameters. This method will be used in the selenium example in Section 7. Notice that the variance of r in (4) exceeds the mean by a factor $(T+b)/b$, whereas in the Poisson density conditional of λ, the variance and mean are equal. The amount of this excess depends on T, a fact that will be exploited in the analysis of the selenium data in Section 7.

4. Elimination of the Basic Failure Rate

With the hyperparameters, a and b, of the density (2) determined either directly, or through r, the analysis can proceed. Assumption 2 deals only with the marginal distribution of λ, whereas the complete analysis will need to incorporate the uncertainty about θ and therefore the joint density of (λ, θ). To handle this we use

ASSUMPTION 3. *λ and θ are independent.*

Again this is a strong assumption and needs careful consideration. To write the two rates, for standard and new components, as λ and $\lambda\theta$ respectively, is of no significance; it is merely notational. But to assert the independence of λ and θ amounts to saying that, in the clinician's judgement, the possible effect of the new component is multiplicative in the sense that whatever the death rate with the placebo, the treatment alters the rate by a constant multiple θ. If the effect were thought to be additive, changing λ to $\lambda+\phi$ say, then λ and θ would not be independent. Assumption 3 simplifies the analysis but if thought inappropriate then a bivariate distribution, or an alternative parameterization, would be required. Again our methodology will apply but with greater complexity.

The parameter of interest is θ, with λ a nuisance parameter. From the likelihood (1), the density for λ, (2), and assumption 3, the nuisance parameter can be eliminated from the likelihood by standard operations of the calculus of probability, to obtain, for data (r, s, T, U) :

$$\begin{aligned} p(D|\theta) &= \int_0^\infty p(D|\theta,\lambda)p(\lambda|\theta)d\lambda = \int_0^\infty p(D|\theta,\lambda)p(\lambda)d\lambda \\ &= \int_0^\infty \lambda^{r+s}\theta^s \exp\{-\lambda(T+\theta U)\}\lambda^{a-1}\exp(-\lambda b)d\lambda \\ &= \theta^s/(T+\theta U+b)^{r+s+a}. \end{aligned} \tag{5}$$

This is sometimes called the posterior likelihood for θ and the integration that provides it is one of the many ways for adherents of likelihood methods

to eliminate a nuisance parameter. It is not uncommon in that context to use the reference density for λ with $a = b = 0$. Inferences from the posterior likelihood (5) are of interest in their own right but also to compare with the methods to be used here to calculate the complete inference $p(\theta|D)$. It is usual to consider the maximum of the posterior likelihood as a reasonable point estimate of θ. Differentiation of the logarithm of (5) easily shows that the maximum occurs at

$$\hat{\theta} = \frac{s(T+b)}{(r+a)U} = \frac{s/U}{r/T} \cdot \frac{T+b}{T+\rho b}, \tag{6}$$

where $\rho = aT/br$.

To appreciate this estimate, notice that r/T and s/U are the observed death rates of placebo and treatments respectively, in contrast to the theoretical rates λ and $\lambda\theta$. Their ratio is therefore not an unreasonable, naive estimate of θ. $\hat{\theta}$ in (6) is this ratio multiplied by a factor $(T+b)/(T+\rho b)$, and is exactly this ratio in the reference case $a = b = 0$. Otherwise the factor is only one when $\rho = 1$. Now $E(r) = Ta/b$, from (4), so $\rho = E(r)/r$, the ratio of expected, to observed, number of failures. Consequently, aside from the limiting case $a = b = 0$, $\hat{\theta}$ is only equal to the ratio of observed rates if the observed number of failures exactly agrees with the expected number. It is clearly possible for expectation and realization to be different without offending the assumptions of the model; indeed, a departure of one from the other of up to about two standard deviations would be reasonable. If this happens, ρ will differ from one and the multiplying factor in (6); $(T+b)/(T+\rho b)$, will also differ from one. We shall see this effect when we consider the selenium data, but one can see intuitively that this modification to the naive ratio of observed failure rates is sensible. Suppose r is substantially in excess of $E(r)$, and therefore $\rho < 1$, without offending the negative binomial distribution of r, and, when $T = U$, s is less than r, say about $E(r)$. Then the apparent reduction from r to s may not be due to the treatment effect but merely due to the chance excess of r over its expectation. This is just when the factor $(T+b)/(T+\rho b)$ in (6) will exceed one and $\hat{\theta}$ will exceed the ratio of observed rates.

Returning to the posterior likelihood (5), in addition to the value at the maximum it is usual to consider the second derivative there, whose negative inverse provides a measure of spread for θ comparable to a variance. Straightforward calculation when $a = b = 0$ and $T = U$ shows that this measure is $s(r+s)/r^3$. It is better to work in terms of $\phi = \ln\theta$ because the posterior likelihood is then more symmetric about its maximum. The corresponding maximum is, of course, $\ln\hat{\theta}$, but the spread is $(r+s)/rs$. Within

the probability calculus these results are not open to the usual interpretation since they are based on a likelihood, not a probability, but are recorded here for subsequent comparison with the complete distribution of θ, given the data, in Theorem 1.

There is an alternative likelihood approach that might be mentioned here, again for comparison with later results. For simplicity we confine ourselves to the case with $T = U$ where the naive estimate with $a = b = 0$ is $\hat{\theta} = s/r$. If $r + s$, the total number of failures with both types of component, is held fixed, it is easy to see that the conditional distribution of s is binomial with index $(r + s)$ and parameter $\theta/(1 + \theta)$, which is free of λ and leads immediately to a likelihood $\theta^s/(1+\theta)^{r+s}$ identical to the posterior likelihood with $a = b = 0$ (and $T = U$) and hence to the same estimate. The difficulty with this method is to justify keeping $(r + s)$ fixed. It is sometimes said to be ancillary but this is not so as its distribution depends on θ (and λ). The conditional, binomial distribution of s, with $(r + s)$ fixed, also permits the construction of a tail-area significance test of the null hypothesis that the new component is not different from the old, or that the medical treatment is ineffectual, $\theta = 1$. If $\theta = 1$, $\theta/(1 + \theta) = 1/2$ and s, given $(r + s)$, is binomial with parameter $1/2$. The null hypothesis is rejected if s lies in the appropriate left-hand tail of the binomial with index $(r + s)$ and parameter $1/2$.

5. Inference about θ, Estimation

The holistic analysis now requires the clinician's views about θ to be expressed through a probability distribution. Two cases are considered; in the first, $\theta = 1$ where the treatment is ineffectual, is taken as a serious possibility with $p(\theta = 1) > 0$. Following Jeffreys (1961) this is a model for testing the null hypothesis that $\theta = 1$. The other case introduces a smooth density $p(\theta)$ in which no value of θ is singled out; this will be referred to as the estimation case and we begin with it. What is needed therefore is a density that can reflect reasonable opinion and, at the same time, allows different views to be expressed through a range of values of hyperparameters. Recall that $p(D|\theta)$ is, from (5), proportional to

$$\theta^s/(c + \theta)^{r+s+a} \text{ with } c = (T + b)/U. \tag{7}$$

The simplest possibility is to use the conjugate density. We therefore make

Assumption 4. *For some positive values* u, v

$$p(\theta) = \frac{(u + v - 1)!}{(u - 1)!(v - 1)!} c^v \theta^{u-1}/(c + \theta)^{u+v}. \tag{8}$$

(An alternative interpretation is that $v\theta/cu$ has an F-distribution on $2u$ and $2v$ degrees of freedom.) Standard results show that for $v > 2$

$$\begin{array}{rcl} E(\theta) & = & cu/(v-1) \text{ and} \\ var(\theta) & = & u(u+v-1)c^2/(v-2)(v-1)^2 \;=\; E(\theta)\{E(\theta)+c\}/(v-2), \end{array} \tag{9}$$

so that, as before, the free hyperparameter u, v enable the mean and variance of θ to be selected. The restriction here to F has similar advantages and disadvantages to that of χ^2 for λ but there is an additional unfortunate feature here, that the opinion about θ involves that for λ. This happens because (8) involves c, and hence b, from (7), a hyperparameter for λ. However, this is not as serious as it might at first appear, because all that is being asked of the engineer or clinician is that they consider a class of densities which depends on b, and to choose u and v within that class.

We now have $p(D|\theta)$ and $p(\theta)$, so that an application of Bayes's theorem provides our final inference for θ.

THEOREM 1. *Assumptions 1 to 4, with choice of hyperparameters a, b, u and v yield*

$$p(\theta|D) = \frac{(r+s+a+u+v-1)!}{(s+u-1)!(r+a+v-1)!} \cdot \frac{c^{r+a+v}\theta^{s+u-1}}{(c+\theta)^{r+s+a+u+v}}$$

for data $D = (r, s, T, U)$, where $c = (T+b)/U$. Alternatively expressed,

$$(r+a+v)\theta/c(s+u) \tag{10}$$

has an F-distribution with $2(s+u)$ and $2(r+a+v)$ degrees of freedom.

Notice, from the expression in terms of F, that the final inference only depends on three quantities, c, $r+a+v$ and $s+u$. This feature will not persist when the test of $\theta = 1$ is considered in Theorem 2.

The theorem provides a complete inference about θ on the basis of the data and relevant knowledge about the experimental conditions, here expressed in terms of the hyperparameters. Complete, in the sense that there is nothing more to be said about the effect of the treatment; unlike a point estimate, to which a standard error must be added and, even then, considerations of skewness etc. are ignored. It is often convenient to describe important features of $p(\theta|D)$; the mean and the standard deviation, for example, might replace estimate and standard error. As with (10), these are

$$E(\theta|D) = \frac{(s+u)U}{(r+1+v-1)/(T+b)},$$

to be compared with (6), and

$$var(\theta|D) = \frac{(s+u)(r+s+a+u+v-1)(T+b)^2}{(r+a+v-2)(r+a+v-1)^2U^2}.$$

The mean agrees with (6) when $u = 0$, $v = 1$. There is no generally accepted reference distribution for θ; $u = v = 0$, $u = v = 1/2$, $u = v = 1$ have all been proposed. As will be seen in Section 7, rather large values of u, v may be required to reflect accurately the practical circumstances. In any case, the degrees of freedom in the final inference will surely be large, when it will be convenient to use the approximation that, for an F-distribution with ν_1, ν_2 degrees of freedom, $\ln F$ is normal with

$$\text{mean: } 1/\nu_2 - 1/\nu_1 \quad \text{ and variance: } 2(1/\nu_1 + 1/\nu_2). \tag{11}$$

It was suggested in Section 3 that, rather than assess a and b through consideration of λ, it might be more realistic to use quantities that could be realized like r, the number of failures, with the standard component (or placebo). Similarly here, with u and v reflecting opinion about θ, one might use s, the number of deaths with the treatment. Unfortunately $p(s)$ is not available in closed form. However, the mean and variance are available and are given here for reference. The calculation is tedious but straightforward.

$$E(s) = \frac{Uacu}{b(v-1)} \quad \text{ and } var(s) = E(s) + E(s)^2 \cdot \frac{1 + (v-1)(a+u+1)/au}{v-2}. \tag{12}$$

Notice that the variance of s exceeds the Poisson variance, equal to the mean. The results are valid if a, u and v all exceed 2.

6. Inference about θ, Hypothesis Test

We now pass to the case where the scientist seriously contemplates the possibility that the treatment is ineffectual; sufficiently strongly to attach a positive probability to $\theta = 1$. To test this against the possibility that it is different, $\theta \neq 1$, it is necessary to express the uncertainty about how different it might be; that is, to consider, in addition to $p(\theta = 1)$, $p(\theta|\theta \neq 1)$. We shall suppose that this density has the same form as in the estimation case, equation (8), leaving the hyperparameters, u and v, to be selected.

Interest now centers on the probability of the null hypothesis, $\theta = 1$, given the data, which is most easily studied through the odds form of Bayes's theorem,

$$\frac{p(\theta = 1|D)}{p(\theta \neq 1|D)} = \frac{p(D|\theta = 1)}{p(D|\theta \neq 1)} \cdot \frac{p(\theta = 1)}{p(\theta \neq 1)},$$

where the first ratio on the right side is the Bayes factor, which is now calculated. Consider first the numerator; from (7) with $\theta = 1$, it is

$$p(D|\theta = 1) = \kappa/(c+1)^{r+s+a},$$

where κ is some constant. The denominator is more complicated:

$$\begin{aligned} p(D|\theta \neq 1) &= \int_{\theta \neq 1} p(D|\theta) p(\theta|\theta \neq 1) d\theta \\ &= \int \frac{\kappa \theta^s}{(c+\theta)^{r+s+a}} \cdot \frac{(u+v-1)!}{(u-1)!\ (v-1)!} \cdot \frac{c^v \theta^{u-1}}{(c+\theta)^{u+v}} \cdot d\theta \end{aligned}$$

from (7) and (8). The integration has already been performed in obtaining $p(\theta|D)$ in Theorem 1. Consequently we have

THEOREM 2. *Assumptions 1 to 4, with choice of hyperparameters a, b, u and v imply that the Bayes factor for testing $\theta = 1$ against the alternative $\theta \neq 1$ is*

$$\frac{p(D|\theta = 1)}{p(D|\theta \neq 1)} = \frac{(r+s+a+u+v-1)!}{(s+u-1)!\ (r+a+v-1)!} \cdot \frac{(u-1)!\ (v-1)!}{(u+v-1)!} \cdot \frac{c^{r+a}}{(1+c)^{r+s+a}}. \tag{13}$$

It is only necessary to multiply this by the initial odds to obtain the final odds on the null hypothesis, given the data. The factorials are most easily calculated by taking logarithms and using Stirling's formula, the constants and the exponentials in that formula conveniently cancelling to leave only the logarithms, with the result that the logarithm of the Bayes factor is

$$\sum_{\pm n} \ln(n + 1/2) + (r+a)\ln c - (r+s+a)\ln(1+c), \tag{14}$$

with n running over all the arguments in the six factorials in (13), and taking the plus (minus) sign for those in the numerator (denominator). In most applications the values of n will be so large that Stirling's formula is almost exact and (14) most accurate.

This completes the methodological development of the clinical trial, placing it in the context within which it was conducted by incorporating natural knowledge of λ and θ. The development includes both the estimation and hypothesis-testing scenarios. Notice that if, in the latter $p(\theta = 1|D)$ is small, usually because the Bayes factor is, supporting an effect, this effect may be estimated by using, with minor modifications, the estimation distribution $p(\theta|D, \theta \neq 1)$. These results are next applied to some data on the possible influence of selenium on cancer deaths.

7. Selenium and Cancer

The data to be analysed come from a clinical trial on the possible effect of selenium on cancer, Clark et al. (1996), which has also been discussed by Colditz (1996) in an editorial in the same journal. The part of the data used here has been re-analysed by Efron and Gous (1997). The trial was originally set up to investigate the possible effect of selenium on skin cancer, the trial being carefully balanced with equal numbers of matched patients on placebo and receiving treatment and $T = U$. At the proposed termination of the trial, no effect had been found, but data had also been kept on deaths from other types of cancer, where there did appear to be an effect with 16 deaths amongst patients receiving a placebo and only 7 for those on a selenium regime. Using the binomial test mentioned in Section 4, with $n = 23$ patient deaths and $p = 1/2$ on the null hypothesis of no effect, the probability of 7 or fewer deaths with selenium is 0.047, just significant at 5%. Alternatively, the likelihood approach of Section 4 with $a = b = 0$ yields an estimate $7/16 = 0.44$ for θ with a standard error of 0.20. The logarithmic transformation gives the same estimate and a 95% interval of (0.18, 1.07). On this evidence, the trial was continued, now looking at the other types of cancer, and not just skin; with the result that at the end of the trial there were 57 deaths with the placebo (including the 16 in the first part of the trial) and only 29 with selenium. The binomial test shows great significance at three standard deviations. The estimate of θ is 0.51 and a 95% interval is (0.32, 0.80). This is an impressive result suggesting at least a 20% reduction in deaths as a result of selenium, and reasonably as high as 50%, possibly about 70%. This was a well-conducted trial, with all reasonable checks and balances incorporated, and many careful statistical analyses performed, in addition to those just mentioned. What these analyses did not do was to place the data in the context of our knowledge of cancer and selenium at the time of the trial, so we use the methods of this paper to provide a holistic analysis. One point to recognize is that all the patients that took part in the trial had had some experience of skin cancer, and that therefore the inferences about θ described here and in the original paper could only apply directly to other people with some experience of cancer of the skin.

The trial is of interest here for two reasons. First, whereas most trials begin with the investigators anticipating — even hoping for — an effect, as this did with skin cancer, here the effect subsequently observed, on other cancers, was not expected, at least on the scale that subsequently appeared. Or, to put it another way, it was the data itself which drew attention to the effect, rather than the effect being anticipated, and it is known that most

data sets can produce a significant effect if mined sufficiently extensively, the theory behind such tests being based on a decision to do the test before seeing the results. It will be seen how holistic methods handle this point by incorporating initial opinion. A second reason for our interest in this trial is that the analysis reveals how knowledge about a nuisance parameter can influence opinions about the parameter of interest. Here the investigators, as later correspondence has revealed, had knowledge about the death rates to be expected for the patients receiving the placebo, our λ, which will affect our view of θ. The paper makes no mention of this and does not provide the value of T, the exposure length of the trial.

The methodology of this paper requires, in addition to the data (r, s), the specification of four hyperparameters (a, b, u, v). Since accounts of the trial do not mention the latter, we have had to 'invent' values that might reasonably reflect the investigators' views as far as they are revealed in the paper, somewhat in the style of Dickey (1973), a paper that has become a classic of scientific reporting. Three sets of hyperparameters have been chosen and each applied both to the intermediate data (16,7) and the complete set (57,29), making six cases in all, numbered 1-6. The first case is described in detail, the same ideas being used for the other five, which are therefore described more cursorily. The cases are summarized in the Table 1. The various cases may be thought of as illustrating the robustness of our analysis to changes in the hyperparameters.

Case 1. The trial was balanced, so that the exposure times of the two sets of patients were equal, $T = U$, and the observations are reasonably Poisson with means λT for placebo and $\lambda\theta T$ for treatment. Case 1 deals with the intermediate data, $r = 16$, $s = 7$. Consider first reasonable values for the hyperparameters a, b applying to the placebo. In Section 3 it was suggested that, instead of thinking about the initial uncertainty of λ, it might be better to contemplate r, the actual number of death to be anticipated with the placebo. In Case 1 it is supposed that the actual number agrees well with that expected, taking $E(r) = 18$, where 16 were experienced. In addition it is necessary to think about how much r might vary. On the Poisson assumption, the variance would also be 18 but the marginal distribution is negative binomial, equation (4), with larger variance, so it was doubled to give 36. From (4) it follows that

$$Ta/b = 18 \text{ and } TA(T + b)/b^2 = 36,$$

yielding $T = b$, $a = 18$ and $c = (T + b)/T = 2$. This provides values for a and c. It does not provide one for b but this does not matter since b does

not occur in the final results except through c, which is determined and is essential in the assessment of θ.

We next consider the hyperparameters u, v for the treatment effect θ, taking first the estimation aspect where the family of densities in Assumption 4 is used, dependent on c, here equal to 2. As already explained, the scientists did not expect any effect of selenium on cancers other than skin, so it is supposed that $E(\theta) = 1$, following a similar proposal by Press and Shigemasu (1989), giving, from (9), $2u = (v - 1)$. To discuss $var(\theta)$ it is necessary to put oneself into the minds of the trial organizers, before any data had been collected, where they were being asked about an effect they did not anticipate. We can only try a few values and start with a standard deviation for θ of 0.2, allowing roughly, at two standard deviations, values reasonably in the range (0.6, 1.4). This may appear too wide for an unanticipated effect but general experience shows that it is not wise to be too firm about quantities for which there is little knowledge. Some would argue for a wider range, even a reference distribution, but recall we are looking at an unanticipated effect; had values seriously different from $\theta = 1$ been contemplated, the trial might have been designed with other aims. A variance of 0.04, with $E(\theta) = 1$ and $c = 2$, in equation (9) gives $v = 77$ and hence $u = 38$. Thus the suggested initial opinion of θ is that $77\theta/76$ is F on (76, 154) degrees of freedom. With such large degrees of freedom, the normal approximation for $\ln F$ is reasonable and a 95% interval for θ, from (12), works out to be (0.67, 1.46), a little different from that based on the standard deviation, reflecting the skewness of F. The effect of selenium could hardly result in more than a 33% reduction in death rates, although it could produce an increase of almost 50%. The right-hand tail is perhaps too long but, within the limitations of the F-distribution, could only be shortened by making $E(\theta) < 1$ and a small effect anticipated.

As the opinion about λ was expressed through r, so s could have been used indirectly to determine that about θ. As it is, equation (12) may be used to find the mean and standard deviation of s implied by the above choice of u and v as a check on coherence. The expectation is naturally the same as that for r since $S(\theta) = 1$, namely 18; and the standard deviation is calculated to be 7.05. The observed value 7, a deviation from expectation of 11, is will within two standard deviations. The observations do not conflict with the initial views.

All the hyperparameters are now determined ($a = 18$, $c = 2$, $u = 38$, $v = 77$) with data ($r = 16$, $s = 7$). Theorem 1 shows that the distribution of θ, given the data, is such that $111\theta/90$ is F on (90, 222) degrees of freedom. Using the logarithmic approximation, θ is centered at 0.80 with a

95% interval (0.57, 1.14). The probability is 0.89 that $\theta < 1$. These results differ seriously from those based on maximum likelihood, where $\hat{\theta} = 0.44$, and the tail-area test, giving significance at 5%, though they do provide evidence of a beneficial effect and might be enough to suggest continuation of the trial.

This completes Case 1 from the estimation viewpoint, but some clinicians may feel that selenium is ineffectual and therefore that a significance test of the null hypothesis $\theta = 1$ may, for them, present a more sensible statistical analysis. In developing the test of the hypothesis in Section 6, the same density for θ under the alternative $\theta \neq 1$ as in estimation was used, so that the same hyperparameters may be used. Theorem 2 gives the Bayes factor to be 0.612. If the original odds had been 20-1 against any effect, which is reasonable for an unsuspected effect, they will have been reduced to about 12-1 against, hardly enough to warrant the trial's extension. Unlike conventional statistical analyses, there can, as here, be a substantial difference between the conclusion with estimation and that with a significance test, due to the appreciable differences in opinion before the trial was started and the violation of the likelihood principle in using the test.

Case 2. This is exactly as Case 1 except that the placebo deaths were expected to be lower than observed, taking $E(r) = 10$ and $var(r) = 20$, again twice the Poisson value, but still with the observed value of 16 reasonably within anticipated limits. These yield $a = 10$, $c = 2$ and u, v remain unaltered. In the estimation form, the data provides a distribution for θ centered at 0.87 with 95% limits of $(0.61, 1.23)$. The probability is 0.79 that $\theta < 1$. For the test that $\theta = 1$, the Bayes factor is 0.925, demonstrating that the data has hardly any effect on the belief that selenium is ineffectual in reducing death due to cancer. Thus the apparent effect could be accounted for by the deaths with the placebo exceeding expectation, without being exceptionally large.

Case 3. This is as Case 1 with the exception that $var(\theta) = 0.01$, suitable for a person who doubts any substantial effect of selenium. Then $a = 18$, $c = 2$ as before, but $u = 150$, $v = 300$. The estimation conclusion is that θ is centered at 0.94 (instead of 0.80) with 95% limit of (0.78, 1.13). The probability that $\theta < 1$ is 0.74 (reduced from 0.89). The Bayes factor is 0.870, demonstrating little change in the clinician's opinion about the null hypothesis as a result of the data. The firmer initial opinion about θ does not conflict with the observed value of s since $E(s) = 18$, with standard deviation 6.27, making $s = 7$ within two standard deviations. The large differences

between methods, based either on likelihood or tail-area significance tests, and holistic treatments just observed, may be due to the small numbers of patients involved, so now the analyses are applied to the whole data, $r = 57$, $s = 29$.

CASE 4. This parallels Case 1 with $E(r) = 60$, $var(r) = 100$ giving $a = 90$, $c = 2.5$. With $E(\theta) = 1$, $var(\theta) = 0.04$ as before, $u = 35$, $v = 90$. With all the hyperparameters determined, $237\theta/160$ is F on 128 and 474 degrees of freedom. The result is that the distribution of θ, given the data, is centered on 0.67 and the 95% interval is (0.51, 0.89). The Bayes factor for the test of $\theta = 1$ is 0.031, so that if the original odds were 20-1 against an effect of selenium, the odds are reduced to 0.62 and the probability that $\theta \neq 1$ is 0.62. The estimation results differ from those provided by maximum likelihood, principally in concluding that the effect is around 33% reduction, rather than almost 50%. The significance test using the Bayes factor is dramatically different from the conclusion using a tail-area approach, replacing the 1 in 1000 level of the latter with a modest 62% probability of an effect, which is hardly enough to justify enthusiastic claims or to take important decisions.

CASE 5. This parallels Case 2 with $E(r) = 40$, lower than Case 4, $var(r) = 80$, giving $a = 40$, $c = 2$ and hence $u = 38$, $v = 77$. Given the data, $174\theta/134$ is F on 134 and 348 degrees of freedom, yielding a central value of 0.76 and a 95% interval of (0.57,1.02). This is different from the last case, reducing the improvement from 33% to 24% and admitting no improvement, $\theta = 1$, to be a real possibility. The probability that $\theta < 1$ is about 0.97. The Bayes factor is 10 times as large at 0.346, so that the original odds of 20-1 against are only reduced to about 7-1. Comparison of Cases 4 and 5 demonstrates the importance of information about a nuisance parameters λ on inference about the parameter θ of interest.

CASE 6. This parallels Case 3 with tighter views about θ. With $E(r) = 60$, $var(r) = 100$, as in Case 3, $a = 90$, $c = 2.5$. Doubts concerning the efficacy of selenium are expressed through $E(\theta) = 1$ and $var(r) = 0.01$, giving new degrees of freedom, $u = 140$, $v = 352$. In the estimation approach, the central value of θ, given the full data, is 0.85 with 95% interval (0.71, 1.01). Comparison with Case 4 not surprisingly reveals a marked shift to the right. In the hypothesis-testing approach, the Bayes factor is 0.210 (as against 0.031 in Case 4), so only reducing the initial odds of 20-1 against to 4-1. This is because the null hypothesis effectively being compared with a suggestion from the data of θ being about 29/57 which, on the alternative, $\theta \neq 1$, has been assigned very little credibility.

TABLE 1. VALUES OF VARIOUS QUANTITIES FOR THE SIX CASES

Cases	1	2	3	4	5	6
Data (r, s)	16,7	16,7	16,7	57,29	57,29	57,29
$E(r)$	18	10	18	60	40	60
$var(r)$	36	20	36	100	80	100
$E(\theta)$	1	1	1	1	1	1
$var(\theta)$	0.04	0.04	0.01	0.04	0.04	0.01
a	18	10	18	90	40	90
c	2	2	2	2.5	2	2.5
u	38	38	150	35	38	140
v	77	77	300	90	77	352
θ/D	0.80	0.87	0.94	0.67	0.76	0.85
95% limits	0.57,1.14	0.61,1.23	0.78,1.13	0.51,0.89	0.57,1.02	0.71,1.01
Bayes factor	0.612	0.925	0.870	0.031	0.346	0.210

Cases 1—3 cover the intermediate data set, 4—6 the complete data. Cases 1 and 4 are the basic cases with r and $E(r)$ in good agreement and a 'liberal' view of possible effect of selenium. Cases 2 and 5 are those where the placebo deaths exceeded expectation, $r > E(r)$. Cases 3 and 6 describe a more 'sceptical' view of the treatment, reducing the prior plausible range for θ. The naive estimates of θ, given D, from Section 4 are 0.44 with the intermediate data and 0.51 with the complete data; both suggesting larger effects than the holistic analysis in the Table. The estimates θ/D above are estimates of θ obtained using the mean of $\ln\theta$ as normal, rather than the exact F distribution.

The statistician's task is to help the organizers to express the views they undoubtedly hold in terms of probability. This is a major task which has not received the attention it merits. All that has modestly been done here is to describe six cases that might cover a range of opinions. If these are reasonable, then the results of the six cases demonstrate that the holistic view explored in this paper reduces the apparent effect of selenium from that presented by the naive attitude and leads to a more sceptical view of its value. There is also a clear indication that information about a nuisance parameter, that was not even mentioned in the otherwise admirable and complete account of the trial in Clark et al. (1996), can influence conclusions about selenium.

8. Possible Extensions

In this section it is shown how the restrictions that have been placed on the analysis, mainly to simplify the mathematics, like the use of conjugate

densities, can be removed at the cost of increased numerical work. In other words, we are arguing that the methods used are very general, only the details need changing for many applications.

Assumption 1 of a Poisson distribution does not always apply either in reliability studies or in clinical trials. It may be better to use a wider class of distributions with more than one nuisance parameter. This will affect the parameter distributions and the range of hyperparameters. Rarely will this permit analytic integration, and numerical procedures will be required. The key point is that the methodology remains the same as that exhibited here in that the probability calculus, and it alone, provides a solution; in contrast to ideas based on likelihood, which need variants like marginal likelihood or tail areas, which need ancillaries or similar devices, both using ad hoc concepts. The methods used here apply generally, numerical integration replacing the analytical techniques used here. One expects that the inclusion of an additional nuisance parameter would increase the spread of the distribution of θ, given the data, thereby casting further doubts on the efficacy of selenium.

Acknowledgements. This research was supported in part by Subcontract no. 35352-6085 between the George Washington University and the Cornell University under WO8333-04 from the Electric Power Institute and the US Army Research Office. I would like to thank Bradley Efron for letting me see his paper with Gous, thereby drawing my attention to the selenium trial, and for helpful comments on a first draft of this paper. Nozer Singpurwala, Jim Press and Bruce Turnbull (one of the 16 authors of the report on the selenium trial) have helped improve the original analysis by making pertinent remarks about the first draft, for which I am most appreciative. A referee provided valuable comments and I hope that the shortening of the paper necessitated by them has not increased the readers' difficulty in understanding my thesis.

References

Clark, L.C., Combs, G.F., Turnbull, B.W., Slate, E.H., Chalker, D.K., Chow, J., Davis, L.S., Glover, R.A., Graham, G.F., Gross, E.G., Krongrad, A., Lesher, J.L., Park, H.K., Sanders, B.B., Smith, C.L. and Taylor, J.R. (1996). Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: a randomized controlled trial, *J. Amer. Med. Assoc.*, **276**, 1957-1963.

Colditz, G.A. (1996). Selenium and cancer prevention, *J. Amer. Med. Assoc.*, **276**, 1983-1985.

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis, *J.R. Statist. Soc. Ser. B*, **35**, 285-305.

Efron, B. and Gous, A. (1997). Bayesian and frequentist model selection. Technical Report 193, Dept. Biostatistics, Stanford University.

Jeffreys, H. (1961). *Theory of Probability*, third edition. Oxford University Press, Oxford.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S. and Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model, *J. Amer. Statist. Assoc.*, **75**, 845-854.

Lindley, D.V. (2000). The philosophy of statistics, *J.R. Statist. Soc. Ser. D*, **49**, 293-337 (with comments).

Press, S.J. and Shigemasu, K. (1989). Bayesian inference in factor analysis. In *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, L.J. Glasser, M.D. Perlman, S.J. Press and A.R. Sampson, eds., Springer-Verlag, New York, 271-287.

Dennis V. Lindley
"Woodstock"
Quay Lane, Minehead
Somerset TA24 5QU
England, UK