ELSEVIER

Journal of Statistical Planning
and Inference 43 (1995) 215–228

journal of
statistical planning
and inference

Bootstrapping an empirical Bayes estimator of the distribution of historical controls in carcinogen bioassay

Daniel Krewski[a,b,*], Robert T. Smythe[c], Anup Dewanji[d], Mieczyslaw Szyszkowicz[a]

[a] *Health Protection Branch, Ottawa, Ontario, Health and Welfare Canada, Canada K1A 0L2*
[b] *Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6*
[c] *Department of Statistics, George Washington University, Washington, DC USA*
[d] *Computer Science Unit, Indian Statistical Institute, Calcutta-700035, India*

Abstract

Previous applications of tests for trend in tumor occurrence rates in laboratory studies of carcinogenicity have frequently employed a beta-binomial distribution for the historical control response rates. In this paper, we develop generalized score tests for trend based on a logistic dose response model and an arbitrary distribution for the historical control tumor response probabilities. A closed-form expression for the score statistic is presented in the special case of a discrete uniform distribution with mass points corresponding to the observed historical control series. The asymptotic distribution of this statistic is shown to be a finite mixture of normal distributions. A discrete empirical Bayes shrinkage estimator of the historical control response rates is proposed, along with a bootstrap variance estimator for the corresponding score statistic which takes into account sampling variability in the historical data. The application of this test is illustrated using bioassay data taken from the literature.

1. Introduction

The first statistical procedure incorporating historical control data into tests for increasing trend in tumor occurrence rates observed in laboratory studies of carcinogenicity was given by Tarone (1982). In this approach, a beta-binomial model is used to describe the extra-binomial variation often seen in historical control data (Haseman et al., 1984), along with a logistic dose response model for the experimental

* Corresponding author.

SSDI 0378-3758(94)00021-M

data. Under these assumptions, the score test for trend is a simple modification of the Cochran–Armitage test which is widely employed without historical controls (cf. Gart et al., 1986).

Subsequently, other tests based on the beta-binomial distribution have been proposed (Yanagawa and Hoel, 1985; Hoel and Yanagawa, 1986; Krewski et al., 1987; Smythe et al., 1987; Prentice et al., 1992; Krewski et al., 1991). A detailed discussion of these procedures has been given by Krewski et al. (1988) and Smythe (1991). Dempster et al. (1993) adopt a Bayesian approach in which the logits of the historical control tumor response probabilities satisfy a normal prior. Prentice et al. (1992) also use an estimating equation approach which requires specification only of the first two moments of the data.

In this paper, we show that the score test for trend under an arbitrary distribution for the historical data is also a simple modification of the Cochran–Armitage test statistic. In particular, the use of a discrete uniform distribution with mass points corresponding to the historical control series leads to a closed form expression for the score statistic. This test statistic is developed in Section 2 along with its distributional properties. A discrete empirical Bayes estimator of the historical control distribution is given in Section 3. A bootstrap variance estimator which takes into account the sampling variability in the historical data is proposed in Section 4. The application of these methods is illustrated in Section 5 using bioassay data taken from the literature.

2. Tests for trend with historical controls

Consider a bioassay with $k+1$ dose levels $0=d_0<\cdots<d_k$ with x_i of the n_i animals at risk in group $i=0,1,\ldots,k$ developing tumors by the end of the study. Suppose further that the probability of developing a tumor in the ith group is given by

$$p_i=[1+\exp\{-(a+bd_i)\}]^{-1}, \tag{2.1}$$

indicating a logistic dose–response relationship. We wish to test the null hypothesis H_0: $b=0$ against the one-sided alternative $H_1: b>0$.

To accommodate historical controls, suppose that the spontaneous response rate $p\equiv p_0$ varies from study to study in accordance with some distribution

$$F(\delta)=Pr\{p\leqslant\delta\} \quad (0<\delta<1). \tag{2.2}$$

Since $a=-\log((1-p)/p)$ is a random variable, we will use the marginal likelihood

$$L(b|\boldsymbol{x})=\int_0^1 L(b|\boldsymbol{x},p)\,dF(p) \tag{2.3}$$

for b given F and the data $\boldsymbol{x}=(x_0, x_1, \ldots, x_k)$ to construct a score statistic for testing H_0: $b=0$. Here, the conditional likelihood $L(b|\boldsymbol{x}, p)$ is

$$L(b|\boldsymbol{x}, p)=\prod_{i=0}^{x}\binom{n_i}{x_i}p_i^{x_i}(1-p_i)^{n_i-x_i}. \tag{2.4}$$

The score statistic is

$$T=\left.\frac{\partial \log L}{\partial b}\right|_{b=0}=\sum x_i d_i-\tilde{p}\sum n_i d_i, \tag{2.5}$$

where $\tilde{p}=E(p|x)$ with $x=\sum x_i$.

Taking F to be a beta distribution with parameters α and β leads to the statistic proposed by Tarone (1982), with $\tilde{p}=(x+\alpha)/(n+\alpha+\beta)$ where $n=\sum n_i$. Here, we consider a discrete uniform distribution

$$F(\delta)=t^{-1}\sum I_{\{\delta_j\leqslant\delta\}}, \tag{2.6}$$

where $0<\delta_j<1\,(j=1,\ldots,t)$. Under (2.6), $\tilde{p}=\sum c_j\delta_j$, where $c_j(x)=\gamma_j(x)/\gamma_{\cdot}(x)$ and $\gamma_j(x)=\delta_j^x(1-\delta_j)^{n-x}$ with $\gamma_{\cdot}=\sum\gamma_j$. The discrete distribution (2.6) for p thus yields a closed form expression for $\tilde{p}$, and will approximate a continuous distribution when t is large. By using a discrete distribution with mass points corresponding to an observed historical control series, further assumptions concerning the distribution of other unobserved historical controls are avoided.

Under the null hypothesis, $E(T)=0$ and

$$V(T)=(\sum n_i d_i^2)\,(t^{-1}\sum\delta_j(1-\delta_j))+(\sum n_i d_i)^2\,\{E(\tilde{p})^2-t^{-1}\sum\delta_j^2\}. \tag{2.7}$$

Since $\tilde{p}=\tilde{p}(x)$ depends on the data $\boldsymbol{x}$ only through x, $E(\tilde{p})^2=\sum_{x=0}^{n}pr\{x\}\,[\tilde{p}(x)]^2$ can be computed using the distribution $pr(x)=\gamma_{\cdot}(x)\binom{n}{x}$. Noting that $\tilde{p}(x)=E(p|x)$ with $[E(p|x)]^2\leqslant E(p^2|x)$ by Jensen's inequality, we have $E(\tilde{p})^2\leqslant E(p)^2$ so that the second term in (2.7) is nonpositive. Thus, a conservative approximation to the variance of T is

$$V_c(T)=(\sum n_i d_i^2)\,(t^{-1}\sum\delta_j(1-\delta_j))\geqslant V(T). \tag{2.8}$$

As demonstrated in the appendix, the asymptotic null distribution of the standardized statistics $S=T/[V(T)]^{1/2}$ is given by

$$\lim_{n\to\infty}Pr\{S\leqslant s\}=t^{-1}\sum\Phi(s/r_j), \tag{2.9}$$

where $r_j^2=\delta_j(1-\delta_j)/\{t^{-1}\sum\delta_j(1-\delta_j)\}$ and Φ denotes the standard normal cumulative distribution function. Thus, the limiting distribution of S is a finite mixture of normal distributions with mean 0 and variance $t^{-1}\sum r_j^2=1$. This leads to an approximate p-value

$$p_{\text{obs}}=Pr\{S\geqslant S_{\text{obs}}\}\sim 1-t^{-1}\sum\Phi(S_{\text{obs}}/r_j), \tag{2.10}$$

where S_{obs} denotes the observed value of S. Since $V_c(T)/V(T) \to 1$ (see the appendix), this same result holds for $S_c = T/[(V_c(T)]^{1/2}$.

In small samples, the null distribution of $\boldsymbol{x}$ given by

$$pr\{\boldsymbol{x}\} = \gamma(x) \prod_{i=1}^{k} \binom{n_i}{x_i} \tag{2.11}$$

may be used to obtain the exact sampling distribution of T, and a corresponding p-value

$$p_{obs} = Pr\{T \geqslant T_{obs}\} = \sum_{x: T \geqslant T_{obs}} pr\{x\}, \tag{2.12}$$

where T_{obs} denotes the observed value of T. In particular, p_{obs} can be calculated by complete enumeration of (2.12) or estimated using computer simulation. In the latter case, simulation of experimental outcomes $\boldsymbol{x} = (x_0, x_1, \ldots, x_k)$ may be accomplished by first generating a value of x_0 from the marginal distribution

$$pr\{x_0\} = t^{-1} \binom{n_0}{x_0} \sum \delta_j^{x_0} (1-\delta_j)^{n_0 - x_0}. \tag{2.13}$$

After generating $x_0, x_1, \ldots, x_{l-1}$ $(l = 1, \ldots, k)$, x_l can be generated from the conditional distribution

$$pr\{x_l | x_0, x_1, \ldots, x_{l-1}\} = \frac{\prod_{u=0}^{l} \binom{n_u}{x_u} \sum \delta_j^{x_{(l)}} (1-\delta_j)^{n_{(l)} - x_{(l)}}}{\prod_{u=0}^{l-1} \binom{n_u}{x_u} \sum \delta_j^{x_{(l-1)}} (1-\delta_j)^{n_{(l-1)} - x_{(l-1)}}}, \tag{2.14}$$

where $x_{(l)} = \sum_{i=0}^{l} x_i$ and $n_{(l)} = \sum_{i=0}^{l} n_i$. Calculation of the test statistic T for a sequence of experimental outcomes simulated in this fashion provides a basis for estimating an exact p-value using (2.12).

3. Estimation of F

In order to implement the score test developed in Section 2, in practice it is necessary to estimate the distribution F in (2.2). This may be done using the observed historical control response rates $z_j = y_j/m_j$, where y_j of the m_j animals in group $j = 1, \ldots, t$ were diagnosed as having the lesion of interest.

Although the empirical distribution function

$$F_t(\delta) = t^{-1} \sum I_{\{z_j \leqslant \delta\}} \tag{3.1}$$

might be considered for this purpose, it will be subject to greater dispersion than F because of sampling variability in the z_j. The estimator F_t is also undesirable when

some of the z_j are zero, a situation which often arises in practice with historically rare lesions. This is because a historical control response probability of $p \equiv 0$ is incompatible with the observation of any tumors in the experiment at hand under the null hypothesis, and results in a weight of $c_j = 0$ being assigned to a value of $\delta_j = 0$.

Another possible approach is to use the nonparametric maximum likelihood estimator (NPMLE) of the mixing distribution F in (2.3) (Laird, 1978). However, our implementation of the NPMLE using the algorithm given by der Simonian (1986) generally resulted in a discrete distribution with a small number of distinct mass points (often only two). Much of this mass is concentrated at the origin when several $z_j = 0$.

These difficulties may be overcome through the use of linear Bayes estimators of the δ_j. Suppose that $\delta_1, \ldots, \delta_t$ are independent identically distributed random variables having mean $0 < \mu < 1$ and variance $\sigma^2 > 0$. Conditional on the δ_j, the y_j are independent binomial random variables with $E(y_j|\delta_j) = m_j\delta_j$ and $V(y_j|\delta_j) = m_j\delta_j(1-\delta_j)$. The linear Bayes estimator of the δ_j are of the form

$$\hat{\delta}_j = \mu + A_j(z_j - \mu), \tag{3.2}$$

where the A_j are chosen to minimize $\sum E(\hat{\delta}_j - \delta_j)^2$. This is done by choosing $A_j = \sigma^2/(D_j + \sigma^2)$, where

$$D_j = m_j^{-1} E\{\delta_j(1-\delta_j)) = m_j^{-1}\{\mu(1-\mu) - \sigma^2\} > 0 \tag{3.3}$$

is the expected value of the conditional variance of z_j (Efron and Morris, 1973). Since $0 < A_j < 1$, $\hat{\delta}_j$ is obtained by shrinking z_j towards its expected value μ, with $0 < \hat{\delta}_j < 1$.

Louis (1984) has shown that when the z_j are normally distributed, the $\hat{\delta}_j$ are shrunk too far. This also occurs here, since

$$E\{(t-1)^{-1}\sum(\hat{\delta}_j - \bar{\hat{\delta}}_\cdot)^2\} = \sigma^2/C^2 < \sigma^2, \tag{3.4}$$

where $\bar{\hat{\delta}}_\cdot = t^{-1}\sum\hat{\delta}_j$ and $C = (t^{-1}\sum A_j)^{-1/2} > 1$. It follows from (3.4) that this underdispersion can be avoided using the modified estimators

$$\hat{\delta}_j = \mu + \hat{A}_j(z_j - \mu), \tag{3.5}$$

where $\hat{A}_j = CA_j$. (Although it is possible that $\hat{A}_j = CA_j > 1$ when the m_j vary markedly, this has not been observed in the applications we have considered to date. In the balanced case $m_j \equiv m_0$ with $A_j \equiv A_0$, we have $\hat{A}_j = CA_j \equiv A_0^{1/2}$ with $0 < A_0^{1/2} < 1$.) A Lagrange multiplier argument may also be used to show that $\tilde{\delta}_j$ is the linear Bayes estimator of δ_j subject to the constraint

$$E\{(t-1)^{-1}\sum(\hat{\delta}_j - \bar{\hat{\delta}}_\cdot)^2\} = \sigma^2, \tag{3.6}$$

which has been used by Spjotvoll and Thomsen (1987) in small area estimation.

Calculation of the linear Bayes estimators requires estimates of μ and σ^2. Ghosh and Lahiri (1987) show that unbiased estimators of these two parameters are given by

$$\tilde{\mu} = \sum w_j z_j \equiv \bar{z}. \tag{3.7}$$

and

$$\tilde{\sigma}^2 = (1 - \sum w_j^2)^{-1} \{\sum w_j (z_j - \bar{z}.)^2 - (m-t)^{-1}(t-1) \sum w_j z_j (1 - z_j)\}, \tag{3.8}$$

where $m_j = m_j/m$ with $m = \sum m_j$. These estimators may be used to obtain an empirical Bayes estimator of $\tilde{A}_j$ and subsequently determine $\tilde{\delta}_j$. The corresponding estimator of F obtained by replacing δ_j by $\tilde{\delta}_j$ in (2.6) is denoted by $\tilde{F}_t$. Lahiri (1990) shows that the empirical Bayes estimators of the $\tilde{\delta}_j$ will minimize the Bayes risk in large samples.

4. Variance estimation

An estimator $\hat{V}(T)$ of $V(T)$ may be obtained by replacing δ_j with $\tilde{\delta}_j$ in (2.7). This approach may also be used to obtain asymptotic or exact p-values using (2.10) and (2.12), respectively. This will work well when the error in estimating F is small in comparison with the sampling error associated with the experimental data x. Since this cannot be assured, we seek an estimator of $V(T)$ which provides for this additional source of variability.

In the absence of a more direct approach to this problem, we propose a bootstrap technique similar to the Type II bootstrap successfully employed by Laird and Louis (1987) in a closely related context. This involves the generation of a bootstrap distribution $\tilde{F}_t^*$, which is simply the empirical distribution function of a random sample $\tilde{\delta}_1^*, \ldots, \tilde{\delta}_t^*$ from $\tilde{F}_t$. The bootstrap distribution $\tilde{F}_t^*$ in turn is used to generate bootstrap samples y^* and x^*, from which the test statistic T^* is computed. Here, y_j^* follows a binomial distribution $\text{Bin}(m_j, \tilde{\delta}_j^*)$ with parameters m_j and $\tilde{\delta}_j^*$. Conditional on δ, x_i^* is $\text{Bin}(n_i, \delta)$, where δ follows the discrete uniform distribution $\tilde{F}_t^*$. In the event that all $y_j^* = 0$, T^* is calculated with $\tilde{p} = x^*/(n+m)$.

Repeating this procedure yields a sequence of test statistics $T_1^*, \ldots, T_b^*$ from which the bootstrap variance estimator

$$\hat{V}^*(T) = (b-1)^{-1} \sum (T_l^* - \bar{T}^*)^2 \tag{4.1}$$

is calculated, where $\bar{T}^* = b^{-1} \sum T_l^*$. The empirical distribution of the T_l^* also provides a bootstrap p-value

$$p_{\text{obs}}^* = Pr\{T^* \geq T_{\text{obs}}\} \tag{4.2}$$

for evaluating the null hypothesis.

5. Applications

In order to illustrate the use of the procedures proposed in this paper, consider the data shown in Table 1 on the occurrence of lung tumors in mice (Example 1) considered previously by Tarone (1982). Ignoring the historical control data, application of the Cochran–Armitage test for increasing tumor occurrence with increasing dose leads to an observed significance level of 0.022, providing somewhat equivocal evidence against the null hypothesis. With historical controls, however, Tarone's test provides strong evidence against the null hypothesis ($p<0.001$).

Using $\tilde{F}_t$ to estimate F, our test statistic with historical controls is $T=5.3$, with a standardized value of $S=T/[\hat{V}(T)]^{1/2}=5.2$ (Table 2). (Note that the computationally simpler conservative variance estimator $\hat{V}_c(T)$ provides a sharp bound on $\hat{V}(T)$ in this case.) Using either the mixed normal approximation to the null distribution of S or the exact p-value based on the finite sample distribution of the experimental data leads to a highly significant result. The intrastudy correlation is 0.007, reflecting a fair degree of homogeneity among historical controls (cf. Fig. 1). The intrastudy correlation coefficient $\rho=\sigma^2/[\mu(1-\mu)-\sigma^2]$ measures the degree of variability among the historical control response rates, with small values of ρ indicating a relatively homogeneous historical control series.

The asymptotic and exact tests applied here are based on the assumption that the distribution of historical control tumor response rates is known without error. The bootstrap variance estimator $\hat{V}^*(T)$ is close to $\hat{V}(T)$, indicating that $\tilde{F}_t$ is relatively well determined in this example (Table 2). With only 12 of 10 000 bootstrap values of the score statistic T^* exceeding the observed value of $T=5.33$, the bootstrap p-value in (4.2) is $p^*=0.0012$. The bootstrap distribution of T^* is shown in Fig. 2, where the small number of cases (0, 43, 4, and 38 of the 10 000 bootstrap values in Examples 1–4, respectively) in which $|T^*|>6$ have been excluded from the graphical display. Note that the bootstrap p-value, which acknowledges the sampling error in $\tilde{F}_t$, is somewhat larger than p-values obtained assuming F_t is known. The bootstrap distribution of the score statistic T shown in Fig. 2 is slightly skewed to the right.

As a second example, consider the data on aveolar-bronchiolar tumors examined previously by Smythe et al. (1986). As in the first example, these lung tumors occur with a relatively high frequency in control animals (Fig. 1), with $\tilde{\mu}=0.085$. Despite an apparent increasing trend in the experimental data, the Cochran–Armitage test leads to a p-value of 0.103 without historical controls due to the relatively small size of the concurrent control group ($n_0=20$). Because of the high degree of homogeneity among the historical controls ($\tilde{p}=0.002$), the asymptotic and exact tests for trend with historical controls lead to much smaller observed significance levels ($p=0.005$ in both cases) than the Cochran–Armitage test. The bootstrap distribution of the score statistic T is nearly symmetric (Fig. 2), with a bootstrap p-value of $p^*=0.01$. Note that in this example the bootstrap variance estimator $\hat{V}^*(T)$ is slightly smaller than the estimator $\hat{V}(T)$, an observation which was confirmed by independently replicating the bootstrap distribution of T.

Table 1
Four examples of bioassay data with historical controls (y_i/m_j)

Example 1. *Lung tumors* (*Tarone*, 1982)

Historical controls (y_j/m_j)

0/50	0/50	0/50	0/49	0/49	0/49
0/47	0/47	0/25	0/25	0/24	0/24
0/22	0/20	0/20	0/20	0/20	0/20
0/20	0/20	0/20	0/20	0/20	0/20
0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/19	0/19	0/18	0/18	0/18
0/18	0/10	1/53	1/50	1/50	1/49
1/49	1/47	1/23	1/23	1/20	1/20
1/20	1/20	1/20	1/20	1/20	1/20
1/20	1/20	1/20	1/20	1/20	1/20
1/19	1/18	2/20	2/20	2/20	2/20
2/20	2/20	2/19	2/18		

Experimental data

Dose(d_i)[a]:	0	0.5	1.0
Response (x_i/n_i):	0/15	3/49	7/46

Example 2: Aveolar–Bronchiolar tumors (*Smythe et al., 1986*)

Historical controls (y_j/m_j)

0/20	0/20	0/19	0/17	0/12
0/12	0/10	1/20	1/19	1/19
1/17	1/15	2/25	4/47	2/22
2/20	1/10	6/54	3/20	3/20
8/49	3/18	4/20		

Experimental data

Dose(d_i)[a]:	0	0.5	1.0
Response (x_i/n_i):	2/20	6/49	10/49

Example 3: Follicular Cell Adenomas (*Bickis and Krewski, 1989*)

Historical controls (y_j/m_j)

0/48	0/42	0/39	0/23	0/20	0/20
0/20	0/20	0/19	0/18	0/17	0/17
0/17	0/17	0/14	0/13	0/12	0/11
0/10	0/9	1/21	1/14	2/19	

Experimental data

Dose(d_i)[a]:	0	0.5	1.0
Response (x_i/n_i):	0/8	0/23	4/39

Example 4: Fibrosarcomas (*Bickis and Krewski, 1989*)

Historical controls (y_j/m_j)

0/54	0/33	0/25	0/25	0/20	0/20
0/20	0/20	0/20	0/20	0/20	0/20
0/20	0/20	0/20	0/15	0/14	0/14
0/10	0/9	0/9	0/8	2/50	2/20

Experimental data

Dose(d_i)[a]:	0	0.5	1.0
Response (x_i/n_i):	0/20	0/50	2/50

[a] Expressed as a fraction of the highest dose used.

Table 2
Tests for trend with historical controls

Component	Example			
	1	2	3	4
Shrinkage parameters				
$\tilde{\mu}$	0.02216	0.08515	0.00870	0.00791
$\tilde{\sigma}^2$	0.00016	0.00013	0.00025	0.00015
Intrastudy correlation				
$\tilde{\rho}=\tilde{\sigma}^2/[\tilde{\mu}(1-\tilde{\mu})-\tilde{\sigma}^2]$	0.007	0.002	0.030	0.019
Test statistics[a]				
T	5.330	5.777	1.527	1.454
Variance estimators				
$\hat{V}(T)$	1.059	4.902	0.238	0.489
$\hat{V}_c(T)$	1.153	4.911	0.343	0.536
$\hat{V}^*(T)$	1.214	4.498	0.524	0.697
Significance levels				
Mixed normal approximation	<0.001	0.005	0.010	0.019
Exact[b]	<0.001	0.005	0.016	0.027
Bootstrap[c]	0.001	0.010	0.036	0.044
Cochran–Armitage test[d]	0.022	0.103	0.048	0.069

[a] Based on the empirical Bayes estimator $\tilde{F}_t$.
[b] Estimated by computer simulation of 10 000 samples from the exact null distribution of the experimental data.
[c] Based on 10 000 samples from $\tilde{F}_t$.
[d] Excluding historical controls.

To assess the performance of these tests with historically rare lesions, consider now the data shown in Table 1 on the occurrence of follicular cell adenomas and fibrosarcomas (Examples 3 and 4, respectively). These data were extracted from the data base examined by Bickis and Krewski (1989), which involves results from 25 randomly selected bioassays conducted under the US National Cancer Institute/National Toxicology Program carcinogenesis bioassay program. In both examples, the historical incidence of the lesion of interest is less than 1% ($\tilde{\mu}=0.009$ and 0.008 in Examples 3 and 4, respectively). In both cases, the p-values for the trend tests are notably lower than those for the Cochran–Armitage test. Note that the p-values based on the mixed normal approximation are close to those based on the exact distribution of the experimental data, confirming the accuracy of the large sample approximation. The bootstrap variances $\hat{V}^*(T)$ are appreciably larger than the V(T), indicating that the empirical Bayes estimators $\tilde{F}_t$ are not well determined in these two examples. The bootstrap p-values are thus greater than those for tests which treat F as known. The bootstrap distribution of T also demonstrated greater positive skewness in Examples

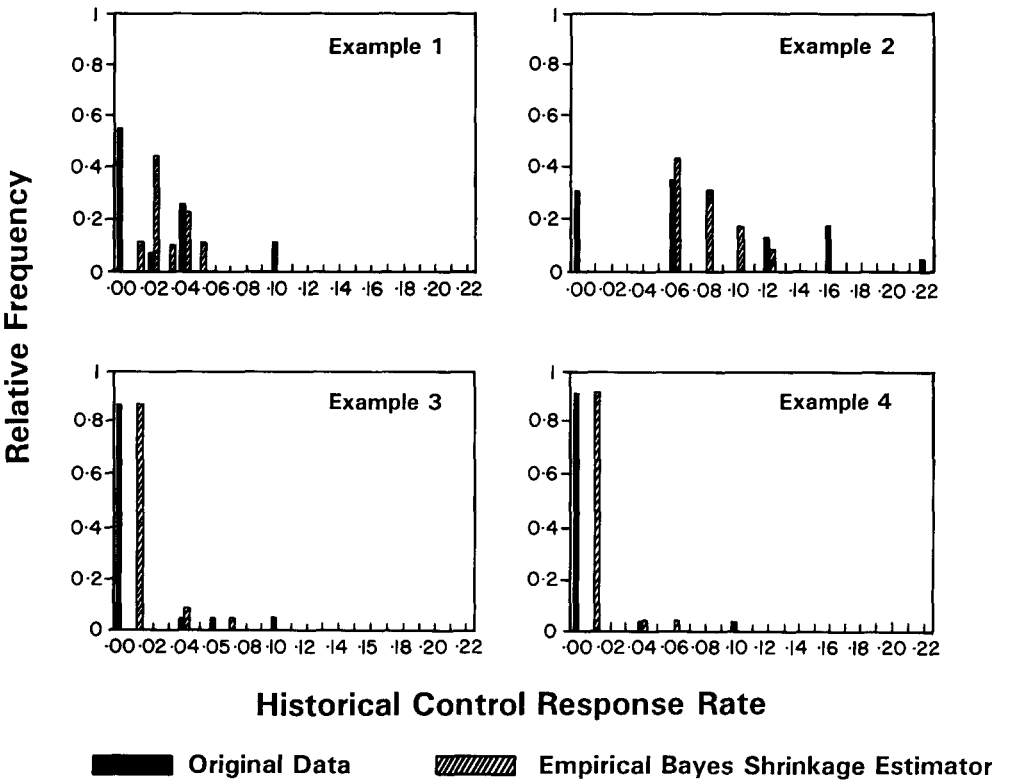

Fig. 1. Empirical Bayes shrinkage estimators of the distribution of historical control response rates.

3 and 4 where the lesion of interest occurs less frequently than in Examples 1 and 2 (Fig. 2).

6. Summary and conclusions

In this paper, we have developed a generalized score test for increasing trend in tumor occurrence rates in carcinogen bioassay utilizing historical controls based on a logistic dose response model and an arbitrary distribution for the historical controls. Using a beta-binomial model for the historical data leads to the test statistic proposed by Tarone (1982). Here, we examined the use of a discrete uniform distribution for the historical control tumor response probabilities. For practical purposes, this discrete distribution will approximate a continuous distribution when the number of historical controls is moderately large. Treating the historical control distribution as known, the asymptotic distribution of the score statistic is a finite mixture of normal distributions.

To apply this procedure in practice, a discrete empirical Bayes shrinkage estimator of the distribution of historical control response rates was introduced. This estimator

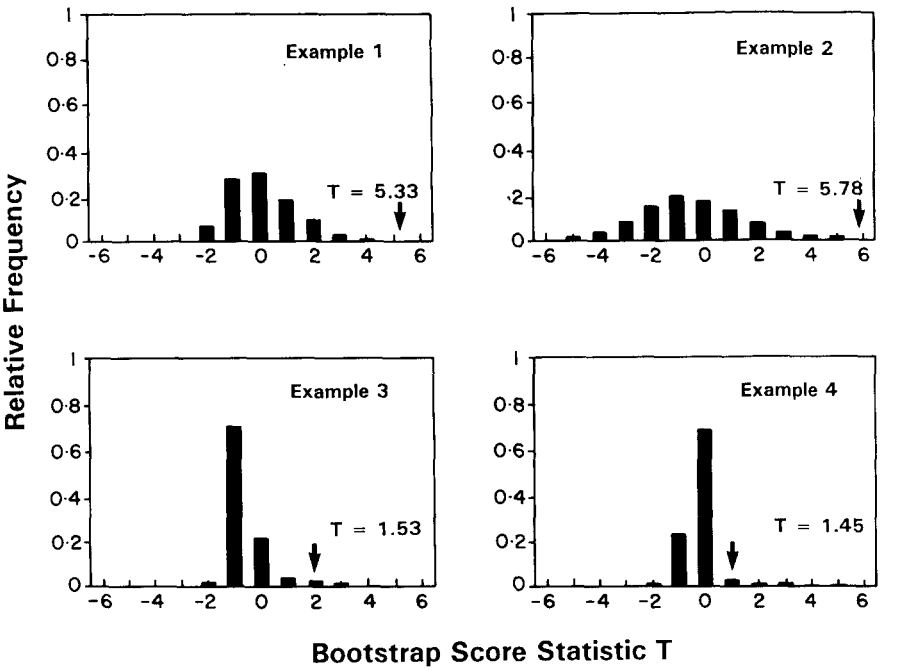

Fig. 2. Bootstrap distributions of the score statistic T (based on 10000 bootstrap samples).

is essentially in closed form, and is easier to compute than the maximum likelihood estimators of parameters of the beta-binomial distribution. Bootstrap methods for variance estimation and significance level determination were proposed to accommodate estimation of the historical control distribution.

The application of the proposed procedures was illustrated using three sets of bioassay data taken from the literature. When the lesion of interest occurred historically with a frequency in excess of 1%, the uncertainty in the empirical Bayes estimator of the historical control distribution did not appear to contribute greatly to the variance of the test statistic. With rarer lesions, however, the historical control distribution appeared to be less well determined, with the bootstrap variance estimator being notably larger than the variance estimator based on the assumption that the historical control distribution is known. In general, the bootstrap method appears to provide a reasonable approach to obtaining a p-value for testing the hypothesis of increasing trend in tumor response rates with increasing dose.

Appendix

The purpose of this appendix is to examine the large sample behavior of the score statistic T under the null hypothesis as $n \to \infty$ with k (the number of dose groups) and

t (the number of historical controls) fixed. To establish the asymptotic distribution of T, we first condition on p and write

$$(\sum n_i d_i^2)^{-1/2} T = Z_p - R_p, \tag{A.1}$$

where we claim that

$$Z_p = (\sum n_i d_i^2)^{-1/2} (\sum x_i d_i - p \sum n_i d_i) \to_d \mathrm{N}(0, p(1-p)) \tag{A.2}$$

and

$$R_p = (\sum n_i d_i^2)^{-1/2} [\tilde{p}(x) - p] (\sum n_i d_i) \to_p 0. \tag{A.3}$$

The result in (7.2) follows immediately from the fact that conditional on p, the x_i are independent binomial random variables with $E(x_i|p) = n_i p$ and $V(x_i|p) = n_i p(1-p)$.

To establish (7.3), we note that for $p = \delta_l$,

$$|c_l(x) - 1| = \sum_{j \neq l} c_j(x), \tag{A.4}$$

where $0 < |c_l(x) - 1| < 1$. Defining

$$\chi = \{x: [n\delta_l(1-\delta_l)]^{-1/2} |x - n\delta_l| \leqslant n^{1/2}\varepsilon\} \tag{A.5}$$

for $\varepsilon > 0$, it follows by direct calculation that

$$\begin{aligned} nE[|c_l(x) - 1)||p = \delta_l] &= n \sum_{x=0}^{n} |c_l(x) - 1| pr\{x|p = \delta_l\} \\ &\leqslant n \sum_{j \neq l} \sum_{x \in \chi} pr\{x|p = \delta_l\} + nPr\{\chi^{\mathrm{C}}|p = \delta_l\}. \end{aligned} \tag{A.6}$$

By Tchebycheff's inequality, we have

$$nPr\{\chi^{\mathrm{C}}|p = \delta_l\} \leqslant \frac{nE\{[n\delta_l(1-\delta_l)]^{-1/2}|x - n\delta_l|\}^4}{n^2\varepsilon^4} = \mathrm{O}(n^{-1}) \tag{7.7}$$

since $E|x - n\delta_l|^4 = \mathrm{O}(n^2)$ (Lamperti, 1966, p. 27). For ε sufficiently small, $x \in \chi$ implies $[n\delta_j(1-\delta_j)]^{-1/2}|x - n\delta_j| > n^{1/2}\varepsilon^*$ for some $\varepsilon^* > 0$ and $j \neq l$. Hence

$$\begin{aligned} n \sum_{x \in \chi} pr\{x|p = \delta_l\} &= nPr\{x \in \chi|p = \delta_l\} \\ &\leqslant nPr\{[n\delta_j(1-\delta_j)]^{-1/2}|x - n\delta_j| > n^{1/2}\varepsilon^*\} \to 0 \end{aligned} \tag{7.8}$$

as in (7.7).

Since $|c_l(x) - 1|^2 < |c_l(x) - 1|$, it follows from (7.6)–(7.8) that

$$nE[|c_l(x) - 1|^2|p = \delta_l] \to 0. \tag{7.9}$$

Noting that

$$n[\tilde{p}(x)-\delta_l]^2 \leqslant 2n(\delta_l+1)^2(c_l(x)-1)^2, \tag{7.10}$$

we now have

$$nE\{[\tilde{p}(x)-\delta_l]^2 | p=\delta_l\} \to 0 \tag{7.11}$$

so that

$$n^{1/2}[\tilde{p}(x)-\delta_l\} \to_p 0 \tag{7.12}$$

conditional on $p=\delta_l$ as required.

To establish the unconditional limiting distribution of $S=T/[V(T)]^{1/2}$, we first note that as a consequence of (7.11),

$$(\sum n_i d_i)^2 \{E[\tilde{p}(x)]^2 - t^{-1}\sum \delta_j^2\} = o(n) \tag{7.13}$$

so that $V(T)/V_c(T) \to 1$. It follows that

$$\lim_{n\to\infty} Pr\{S \leqslant s\} = t^{-1}\sum \Phi\left(\frac{s}{r_j}\right), \tag{7.14}$$

where $r_j^2 = \delta_j(1-\delta_j)/\{t^{-1}\sum \delta_l(1-\delta_l)\}$.

Acknowledgements

This work was supported in part by grant A8664 from the Natural Sciences and Engineering Research Council of Canada (D. Krewski), by grant NSF-DMS-8503774 from the US National Science Foundation (R. Smythe), and by grants CA-39949 and CA-47658 from the US National Cancer Institute (A. Dewanji). Additional support was provided by a grant from the ILSI Risk Sciences Institute (R. Smythe). A Fortran computer program implementing the procedures proposed here is available on request.

References

Bickis, M. and D. Krewski (1989). Statistical issues in the analysis of long term carcinogenicity bioassay in small rodents: an empirical evaluation of statistical decision rules. *Fund. Appl. Toxicology* **12**, 202–221.

Dempster, A.P., M.D. Selwyn and B.J. Weeks (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78**, 221–227.

der Simonian, R. (1986). Maximum likelihood estimation of a mixing distribution. *Appl. Statist.* **35**, 302–309.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitors — an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

Gart. J.J., D. Krewski, P.N. Lee, R.E. Tarone and J. Wahrendorf (1986). *Statistical Methods in Cancer Research, Vol. III: The Design and Analysis of Long-Term Animal Experiments.* Internat. Agency for Research on Cancer, Lyon.

Ghosh, M. and P. Lahiri (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* **82**, 1153–1162.

Haseman, J.K., J. Huff and G.A. Boorman, (1984). Use of historical control data in carcinogenicity studies in rodents. *Toxicological Pathology* **12**, 126–135.

Hoel, D.G. and T. Yanagawa (1986). Incorporating historical controls in testing for a trend in proportions. *J. Amer. Statist. Assoc.* **81**, 1095–1099.

Krewski, D., R.T. Smythe and D. Colin (1987). Tests for trend in binomial proportion: a proposed two-stage procedure. In: I.B. MacNeill and G.J. Umphrey, Eds., *Advances in the Statistical Sciences, Vol. V: Biostatistics*, Reidel, Boston, 61–69.

Krewski, D., R.T. Smythe, A. Dewanji and D. Colin (1988). Statistical tests with historical controls. In: H. Grice and J.L. Ciminera, Eds., *Carcinogenicity: The Design, Analysis and Interpretation of Long-Term Animal Studies*, Springer, New York, 23–38.

Krewski, D., R.T. Smythe, K.Y. Fung and R. Burnett (1991). Conditional and unconditional tests with historical controls. *Canad. J. Statist.* **19**, 407–423.

Lahiri, P. (1990). "Adjusted" Bayes and empirical Bayes estimation in finite population sampling. *Sankhyá Ser. B* **52**, 50–66.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73**, 805–811.

Laird, N.M. and T.A. Louis (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82**, 739–750.

Lamperti, J. (1966). *Probability*. Benjamin, New York.

Louis, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79**, 393–398.

Prentice, R.L., R.T. Smythe, D. Krewski and M. Mason (1992). On the use of historical control data to estimate dose response trends in quantile bioassay. *Biometrics* **48**, 459–478.

Smythe, R.T. (1991). The use of historical controls in bioassay for carcinogens. In: D. Krewski and C.A. Franklin, Eds., *Statistics in Toxicology*. Gordon and Breach, New York, and 591–606.

Smythe, R.T, D. Krewski and A. Dewanji (1987). Robust tests for trend in binomial proportions. In: R. Viertl, Ed., *Probability and Bayesian Statistics*, Plenum, New York, 443–454.

Smythe, R.T., D. Krewski and D. Murdoch (1986). The use of historical control information in modelling dose response relationships in carcinogenesis. *Statist. Probab. Lett.* **4**, 87–93.

Spjotvoll, E. and I. Thomsen (1987). Application of some empirical Bayes methods to small area statistics. *Bull. Internat. Statist. Institute LII* (Book 4), 435–450.

Tarone, R.E. (1982). The use of historical control information in testing for a trend in proportions. *Biometrics* **38**, 215–220.

Yanagawa, T. and D.G. Hoel (1985). Use of historical controls for animal experiments. *Environmental Health Perspectives* **63**, 217–224.