11-1-2002

On the Estimation of Binomial Success Probability With Zero Occurrence in Sample

Mehdi Razzaghi
Bloomsburg University, Bloomsburg, PA

On the Estimation of Binomial Success Probability With Zero Occurrence in Sample

Mehdi Razzaghi
Mathematics, Computer Science, & Statistics
Bloomsburg University

The problem of estimating the probability of a rare event when the sample shows no incidence of the event is considered. Several methodologies based on various statistical techniques are described and their relative performances are investigated. A decision theoretic approach for estimation of response probability when the sample contains zero responses is examined in depth. The properties of each method are discussed and an example from teratology is used to provide illustration and to demonstrate the results.

Key words: Binomial distribution, response probability estimation.

Introduction

There are many instances in practice that an estimate of the probability of occurrence of a rare event is desired. Because of the low probability of the event, however, the experimental data may conceivably indicate no occurrence of that event. For example, in cancer risk estimation with laboratory animals, often at low doses, data may exhibit no animals with tumors, even though there is a nonzero probability of response at that dose. More specifically, suppose that X is the number of occurrences of an event in a sample of n independent and identical Bernoulli trials. Then X has a binomial distribution with

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \cdots, n \quad (1)$$

where p is the probability of occurrence in each trial. It is well known that the maximum

Professor Mehdi Razzaghi's area of interest is environmental statistics with applications of statistical modeling and risk assessment in toxicological experiments. Address: Mathematics, Computer Science, & Statistics, Bloomsburg University, Bloomsburg, PA 17815. E-mail: razzaghi@bloomu.edu. The author is grateful to the Editor and the anonymous referees for their helpful comments.

likelihood estimate of p is x/n. But when x = 0, this estimate is often unrealistic and alternative methods should be utilized to estimate p. Observation of zero occurrence in a sample is not uncommon in practice. Table 1 provides numerical values of the probability of zero successes in binomial experiments for different sample sizes.

Table 1. Probability of zero response for varying sample sizes and different true response probabilities.

p \ n	0.01	0.02	0.05	0.07	0.10	0.15	0.20
1	0.990	0.980	0.950	0.930	0.900	0.850	0.800
2	0.980	0.960	0.902	0.865	0.810	0.722	0.640
4	0.961	0.922	0.814	0.748	0.656	0.522	0.410
10	0.904	0.817	0.599	0.484	0.349	0.197	0.107
20	0.818	0.668	0.358	0.234	0.122	0.039	0.011
30	0.740	0.545	0.215	0.113	0.423	0.008	0.001

Note that even when p is as high as 0.05 and the sample is as high as twenty, there is still a 36% chance of no response in the data. Bailey (1997) considered the problem of estimating p when the sample has no occurrence and proposed a method currently used in risk analysis of energetic initiation in the explosive testing field. This estimator is given by

$$\hat{p} = 1 - (0.5)^{1/n} \quad (2)$$

which is obtained by setting the probability of observing n failures equal to 0.5 and solving for p. Bailey noted that this estimator is nearly identical to the median of the Bayesian posterior distribution for p, derived with respect to a uniform distribution using the absolute error loss (AEL) function.

The problem of Bayesian estimation of p with respect to the more general class of a conjugate beta prior distribution but using the squared error loss (SEL) was considered by Basu et at. (1996). By comparing (2) with a few other estimates, Bailey (1997) concluded that $\hat{p}$ performs relatively well in practice and can be used in certain circumstances. It is also worth noting that because the upper $100(1 - \alpha)\%$ confidence limit for p is (see Bickel & Doksum, 2001) given by

$$u = 1 - \alpha^{1/n}$$

then (2) can be interpreted as the median of the sampling distribution of the random variable X/n. Moreover, as mentioned in Louis (1981), u may be thought of as the proportion of the number of successes in a future experiment of the same size and it is the upper $100(1 - \alpha)\%$ Bayesian prediction interval based on a uniform prior distribution.

In this paper, the problem of point estimation of p when a sample shows no occurrence is considered from a more general viewpoint. Several potential estimates based on statistical methods in addition to those suggested in Bailey (1997) and Basu et al. (1996) will be proposed and their properties will be discussed. Next, I review the Bayesian approach and consider the use of other loss functions, and then discuss the properties of an estimate derived from information theory. The next section is devoted to the discussion of a decision theoretic approach for estimating p, and the use of minimax estimation of p is considered. In the final section of this article, I give an example from teratology to provide further illustration of the results.

Bayesian Estimation

It is well known that when the prior distribution of p belongs to the family of a beta distribution $\beta(a, b)$,

$$g(p) = \frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1} \quad a, b > 0, 0 < p < 1 \tag{3}$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

then the posterior distribution of p belongs to the beta family $\beta(a + x, b + n - x)$ and the Bayes estimate p^* of p based on the SEL function $L(p,p^*) = (p - p^*)^2$, is given by (Basu et al., 1996)

$$p^* = \frac{(a+x)}{(a+b+n)} \tag{4}$$

Thus, if x = 0, then the Bayes estimator for a zero occurrence is

$$p^* = \frac{a}{a+b+n} \tag{5}$$

and in particular if a = b = 1, then the Bayes estimator under a uniform prior is derived. Also, when Jeffreys' non-informative prior, for which a = b = 0.5 is used, then the Bayes estimator of no response is given by

$$p_{ni}^* = \frac{1}{2(n+1)} \tag{6}$$

Basu et al. (1996) compared (5) and (6) with the classical approach based on upper confidence limits and conclude that the Bayes estimate under an informative prior is best. Both estimates (5) and (6), however, are derived using the SEL function which is but one of several possible loss functions that may be used to derive the Bayes estimate of p. In practice, there are many instances that other functions may be preferred.

Actually the SEL is a special case of a larger class of weighted quadratic loss functions

$$L(p, p^*) = w(p)(p - p^*)^2 \tag{7}$$

where $w(p) \geq 0$ is an appropriate weight function. For the class (7) the posterior expected loss is minimized when

$$p^* = \frac{E(pw(p))}{E(w(p))} \qquad (8)$$

where the expectation is with respect to the posterior distribution of p. In particular if $w(p)$ is of the form

$$w(p) = p^\alpha (1-p)^\beta \qquad (9)$$

for some α and β, then from (8)

$$P^* = \frac{E(p^{\alpha+1}(1-p)^\beta)}{E(p^\alpha (1-p)^\beta)} \qquad (10)$$

which for the family of beta prior, yields

$$p^* = \frac{a+\alpha+x}{a + b + n + \alpha + \beta} \qquad (11)$$

Now, if $\alpha = \beta = 0$, then (4) is obtained as a special case of this larger class of estimates. Another special case, and possibly more appropriate for the purpose of risk assessment, in (11) is when $\alpha = \beta = -1$, corresponding to the scaled square error loss (SSEL) function

$$L(p, p^*) = \frac{(p - p^*)^2}{p(1-p)}$$

In this case, however, it is easy to see that when x = 0, and a = 1, then p^* is the only estimate which produces an infinite posterior expected loss. Hence, when there is no occurrence in the sample the SSEL function does not produce a useful solution. Indeed, when x = 0, the SSEL function produces a negative estimate of p for a < 1. Note also from (11) that in this case the Bayes estimate with respect to a uniform distribution is identical to the maximum likelihood estimate.

Aside from the class of squared error loss functions, a class of functions often used in Bayesian estimation is the absolute error loss (AEL) given by

$$L(p, p^*) = | p - p^* | ,$$

for which the Bayes estimate is the median of the posterior distribution. Hence for the family of beta prior (3), when x = 0, we seek p_1^* such that

$$I_{p_1^*}(a, b+n) = \int_0^{p_1^*} \frac{1}{B(a,b+n)} p^{a-1}(1-p)^{b+n-1}dp = 0.5 \qquad (12)$$

which for given values of a and b can be evaluated using tables of incomplete beta functions (e.g. Pearson & Hartley, 1956) or any standard numerical technique. Specifically, if a = b = 1, then (12) yields

$$p_1^* = 1 - (0.5)^{1/(n+1)} \qquad (13)$$

which, as noted earlier, is for large n approximately equal to the Bailey (1997) estimate. Also, when Jeffrey's non-informative prior (a = b = 0.5) is used, an approximation to the solution of (12) may be obtained by using a procedure described in Johnson and Kotz (1995) regarding the approximations to the beta function ratio. Accordingly, if $p_{1.n_i}^*$ denotes the solution of (12) for a = b = 0.5, then an approximate value of p_{1,n_i}^* can be obtained as the solution of

$$n + \frac{1}{6} - \left(n + \frac{7}{3}\right)(1-x) + \frac{1}{5}\frac{2x}{2n+1} + \frac{x-1/2}{n+1} = 0 \qquad (14)$$

where the error of approximation is generally below .001.

Another choice of a loss function for Bayesian estimation is the so-called zero-one loss defined as

$$L(p,p^*) = \begin{cases} 0 & \text{if } | p - p^* | \leq \varepsilon \\ 1 & \text{if } | p - p^* | > \varepsilon \end{cases}$$

which amounts to no loss if the estimate p^* is within a distance ε from p. For this loss function, the expected posterior is given by

$$P(|p - p^*|) > \varepsilon \mid x) = 1 - P(|p - p^*| \le \varepsilon \mid x).$$

Consequently, if a modal interval of length 2ε is defined as an interval with center at the mode of the distribution, then as $\varepsilon \to 0$, the Bayes estimate with respect to the zero-one loss approaches the mode of the posterior distribution, provided that a mode exists. This in turn implies that the Bayes estimate in this case becomes the maximum likelihood estimate.

Maximum Information Estimation

Good(1965) and Typlados and Brimley (1962) showed that Shannon's information content of the observation x from the binomial distribution (1) is given by

$$I(p) = -p\ln(p) - (1-p)\ln(1-p) + \ln\left[\binom{n}{x}p^x(1-p)^{n-x}\right]$$
$$(15)$$

By maximizing I(p), one obtains the maximum information (MIE) estimate p_{MIE} of p as the solution of the equation

$$\ln\left(\frac{p}{1-p}\right) = \frac{x}{p} - \frac{n-x}{1-p} \qquad (16)$$

In particular when x = 0, the MIE of p is the solution of

$$\sqrt[n]{\frac{p}{1-p}} = \exp\left(-\frac{1}{1-p}\right). \qquad (17)$$

Chew (1971) pointed out that for n > 7, the solution of (17) is up to 3 decimals equal to zero and, once again, it is seen that this method fails to produce a reasonable estimate for p.

Minimax Estimation

The minimax criterion stems from the general theory of two-person zero-sum games of von Neuman and Morgenstern (1944). Loosely,

instead of averaging the risk as in Bayesian estimation, one looks at the least favorable scenario for each decision, that is the worst possible risk for that decision, and chooses a decision which gives the least value of the worst risk. Thus, the minimax rule minimizes the maximum risk. Although the methodology ignores all references to prior knowledge, but in the absence of any information regarding p, the minimax estimator provides a Bayesian estimate without knowing the prior distribution. As pointed out by Cox and Hinkley (1974), the minimax rule is defensible when the risk is small, since it ensures that, whatever the true parameter value, the expected loss is small. Although there may be an apparently better rule, any improvement can only be small and may carry with it the danger of a seriously bad performance for some values of the parameter.

Now, for the binomial parameter p in (1), it can be shown that the minimax decision rule, based on the SEL function, is given by (Bickel and Doksum, 2001)

$$\tilde{p} = \frac{x + \sqrt{n}/2}{n + \sqrt{n}} \qquad (18)$$

with variance bounded by

$$v = \left[2(1 + \sqrt{n})\right]^{-2} \qquad (19)$$

The minimax estimator (18) is Bayes with respect to a beta prior with parameters

$\sqrt{n}/2$ and $\sqrt{n}/2$. If x = 0, then from (18),

$$\tilde{p} = \left[2(1 + \sqrt{n})\right]^{-1} \qquad (20)$$

which can be used to estimate the probability of a rare event. In order to compare the minimax estimator given in (20) with those considered in Bailey (1997), $\tilde{p}$ was evaluated for several values of n. Table 2 presents these numerical values, where for comparison, the values of $\hat{p}$ in (2), the estimator suggested by Bailey and the Bayes estimator p_{ni}^* based on a noninformative prior given in (6) are also included. As the sample size

increases, the minimax method appears to produce numerically larger point estimates.

Table 2. Numerical values of minimax $(\widetilde{p})$, Bayes $\left(p_{ni}^{*}\right)$ and Bailey $(\hat{p})$ estimator.

n	1	2	4	10	20	30	40
$\widetilde{p}$	.250	.207	.167	.120	.091	.077	.062
p_{ni}^{*}	.250	.167	.100	.045	.024	.016	.010
$\hat{p}$	.500	.293	.159	.067	.034	.023	.014

Because the binomial distribution $E(X) = np$, it is clear from (4) and (18) that

$$E(p_{ni}^{*}) = \frac{2np+1}{2(n+1)} \qquad (21)$$

and

$$E(\widetilde{p}) = \frac{2\sqrt{n}p+1}{2(\sqrt{n}+1)} \qquad (22)$$

Table 3 provides the numerical values of (21) and (22) for selected values of n and p where for completeness we also include a crude estimate of $E(\hat{p})$, computed by using (2) for x = 0.

p	.01			.05			.10		
n	$\widetilde{p}$	p_{ni}^{*}	$\hat{p}$	$\widetilde{p}$	p_{ni}^{*}	$\hat{p}$	$\widetilde{p}$	p_{ni}^{*}	$\hat{p}$
4	0.173	0.108	0.169	0.200	0.140	0.209	0.233	0.180	0.259
10	0.128	0.054	0.077	0.158	0.091	0.117	0.196	0.136	0.167
20	0.099	0.033	0.044	0.132	0.071	0.084	0.173	0.119	0.134
30	0.086	0.026	0.033	0.119	0.064	0.073	0.162	0.113	0.123
40	0.077	0.022	0.027	0.111	0.061	0.067	0.155	0.110	0.117
50	0.071	0.020	0.023	0.105	0.059	0.064	0.149	0.108	0.114

Table 3. Expected values of minimax $(\widetilde{p})$, Bayes $\left(p_{ni}^{*}\right)$ and Bailey $(\hat{p})$ estimators for varying sample sizes and for different true response probabilities.

Example

Kochhar et al. (1992) describes an experiment to examine the developmental toxicity of two retinoylamino acids, RG and RL in IRC mice and compare them with other retinamides. One of the observed effects was the incidence of cleft palate in the viable fetuses. Table 4 presents the percentage of fetuses with cleft palate for different doses together with the number of implants per dose group as a result of maternal exposure to retinoic acid (RA).

Table 4. Incidence of cleft palate in offspring of mice exposed to retinoic acid (RA). Source: Kochhar et al. (1992).

Dose mg/kg	0	5	10	25	100
Number of Implants	152	98	78	86	164
% with Cleft Palate	1	0	13	33	82

It is observed that even though there was 1% response rate in the control group, there was no occurrence of cleft palate in the 5 mg/kg dose group. The incidence rate in other dose groups showed a statistically significant difference from the control group. For risk assessment purposes, in practice one would fit a suitable dose-response model to these data and extrapolate to low exposure levels to obtain an upper confidence limit for the risk at a fixed low dose.

The model can equivalently be used to obtain a benchmark dose, which is the lower confidence limit for dose corresponding to a given low negligible level of risk. However, because of no incidence at the lowest non-zero dose level, one might erroneously consider fitting a non-monotonic dose-response function.

That is, the analysis might lead to the conclusion that the chemical has a hormetic effect, i.e. it is low dose stimulative and high dose inhibitive. For a discussion on the concept of

chemical hormesis we refer to Calabrese and Baldwin (2000). However, as shown in Razzaghi and Loomis (2001), in developmental toxicology, more than a single replication of an experiment must be considered before a chemical can be declared as being hormetic. For the present data, therefore, in order to fit a monotonic dose-response function, one might consider replacing the observed incidence of zero by an estimate of it. In such a situation, it would seem unreasonable to estimate the probability of response in the 5 mg/kg dose group as 0, as given by the maximum likelihood method. In this case, because n = 98, from (2), (6), (14) and (20),

$$\hat{p} = .007, \quad p_{ni}^* = .005, \quad p_{1,ni}^* = .021, \quad \tilde{p} = 0.046$$

are four different point estimates for the probability of response at the first nonzero dose level.

In order to further investigate the properties of these estimates, a probit model was used to fit the response probability p as a function of the natural logarithm of dose, i.e.

$$p = \Phi(a + b \log d) \qquad (23)$$

Using PROC PROBIT in SAS (1996), it was found that the maximum likelihood estimates of the model parameters are $\hat{a} = 03.601$ and $\hat{b} = 0.987$. Using these parameter estimates, it is found that the point estimate of p when d = 5 mg/kg is .022. Furthermore, the standard deviation of $\hat{a} + \hat{b} \log 5$ is 0.163. Based on these quantities, if the 95% confidence interval is evaluated for the predicted proportion, one finds that this range is (.010, .046). Interestingly, although the minimax estimator $\tilde{p}$ is equal to the upper bound in this range, both the Bailey estimator $\hat{p}$ and the Bayesian estimator p_{ni}^* are outside this range and far too small to be plausible. Therefore, in this instance, $p_{1,ni}^*$ and the minimax procedure appear to produce more realistic estimates of p compared to other methods.

Discussion

Lack of occurrence of rare events in biological and physical experiments is not uncommon. In such situations, the maximum likelihood estimate becomes unusable and one needs to resort to alternative statistical methods. Here, I have considered this problem and investigated the use of several other statistical techniques and the minimax estimator.

It is immediately noted from (2) that for the Bailey estimator, $\hat{p} = 0\left(\dfrac{1}{n}\right)$. This property also holds for the Bayesian estimator considered by Basu et al. (1996). However, for the minimax estimator, from (18) $\tilde{p} = 0\left(\dfrac{1}{\sqrt{n}}\right)$. This means that for relatively large values of n, both $\hat{p}$ and the Bayes estimate lead to numerically smaller values than the minimax estimator. Actually, it can be shown (Roussas, 1997) that the Bayes estimate for the family of beta prior and SEL has the same asymptotic distribution as the maximum likelihood estimate for arbitrary fixed values of α and β, while the asymptotic distribution of $\sqrt{n}(\tilde{p} - p)$ is normal with mean $\dfrac{1}{2} - p$ and variance p(1-p). Thus, I can say that the minimax estimator is comparatively more conservative.

However, as discussed by Carlin and Louis (1996), although informative priors enable more precise estimation, extreme care must be taken in their use because they also carry the risk of disastrous performance when their informative content is in error. Although using a non-informative prior leads to a more conservative Bayes estimate, there may be situations when Bayes and other methods underestimate the value of this rare event. This result is demonstrated through an example in developmental toxicology.

The conclusion of this paper is not necessary to recommend the minimax or any other estimator in all situations when there is a zero response. Rather, the goal is to increase awareness and recommend that more caution should be taken when any single method is used to estimate the success probability when sample shows zero occurrence. The choice of the estimate should to a large extent depend on which kind of optimality is judged to be most appropriate for the case in question.

References

Bailey, R. T. (1997). Estimation from zero-failure data. *Risk Analysis*, *17*, 375-380.

Basu, A. P., Gaylor, D. W. & Chen, J. J. (1996). Estimating the probability of occurrence of tumor for a rare cancer with zero occurrence in a sample. *Regulatory Toxicology and Pharmacology*, *23*, 139-144.

Bickel, P.J., & Doksum, K. A.(2001). *Mathematical Statistics*. (2nd ed.) Oakland, CA: Holden-Day.

Calabrese, E. J., & Baldwin, L. A. (2000). Chemical hormesis: Its historical foundations as a biological hypothesis. *Human and Experimental Toxicology*, *19*, 2-31.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data* analysis. (2nd ed.). NY: Chapman and Hall.

Chew, V. (1971) Point estimation of the parameter of the binomial distribution. *The American Statistician*, *25*, 47-50.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.

Good, I. J. (1965). *The estimation of probabilities: An essay on modern Bayesian methods*. Research Monograph No. 30. Cambridge: The M.I.T. Press, p. 15-19.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*, Volume 2. (2nd ed.) NY: Wiley.

Kochhar, D. M., Shealy, Y. F., Penner, J. D., & Jiang, H. (1992). Retinamides: Hydrolic conversion of retinoylglycine to retinoic acid in pregnant mice contributes to teratogenicity. *Teratology*, *45*, 175-185.

Louis, T. A.(1981). Confidence intervals for a binomial parameter after observing no successes. *The American, 35*, 154.

Neuman, J. von, &Morgenstern, O. (1994). *Theory of games and economic behavior*. (3rd ed.). Princeton, NJ: Princeton University Press.

Pearson, E. S., & Hartley, H. O. (1956). *Biometrika tables for statisticians*. London: Cambridge University Press.

Razzaghi, M., & Loomis, P. (2001). The concept of hormesis in developmental toxicology. *Human and Ecological Risk Assessment*, *7*, 933-942.

Roussas, G. G. (1997). *A course in mathematical statistics*. (2nd ed.) NY: Academic Press.

Typaldos, Z. A., & Brimley, D. E. (1962). Point estimation of reliability from results of a small number of trials. *Memorandom RM-3044-PR*, Santa Monica, CA: The Rand Corporation.

SAS(1999). Statistical analysis system, Version 8. Cary, NC: SAS Institute.