Estimating Event Probabilities using Zero Failure Data

Karl Breitung
Researcher, TU Munich, Germany

Marc A. Maes
Professor, University of Calgary, Canada, and UGent, Belgium

ABSTRACT: In a rather well publicized June 2013 CBC broadcast, the safety of a specific series of Canadian onshore gas pipeline joints was declared "absolute" by one proponent, since, historically, no incidents or failures had ever been reported for that site. An opponent then argued that the face value of this risk could never be zero, but "very small". The objective of the present paper is to review just how small, and at which confidence level, one can sensibly consider the actual incidence rate to be. A comparison of the most popular approaches and a comprehensive test for consistency, point to the superiority of the Bayesian estimator together with a non-central posterior probability interval.

1. INTRODUCTION

A common scenario in the context of data mining and explorative statistics of accidents/failures, is that the analyst runs into data segments or subsets that have zero reported incidents. Is it legitimate to construct, with a specified confidence, an upper limit p_U for which we can claim that the "true" incident probability p will not exceed this value?

A typical example concerns pipeline rupture and leak statistics (TSB, 2013): out of thousands of pipeline joints n a large majority of joints are observed to have no recorded incidents. There is often confusion about which confidence intervals apply in a case like this. Some analysts replace the zero incident case $X = 0$ by $X = 1$ because it is "easier" to analyze and because it is after all "conservative".

Others suggest that if, in the case of a binomial model, n consecutive trials have not resulted in any failure ($X = 0$), then we should prudently assume that the $(n+1)$-th will; subsequently this pessimistic estimate is used for prediction.

2. ESTIMATION FOR ZERO INCIDENTS

One fundamental problem is related to the setting of the problem and to its formulation. In this paper we consider a binomial model originating from a setup where n units are monitored and incidents are observed within each unit. It also applies to the case where n sections of a continuous system such as a pipeline are tested and X are found to be faulty. The analysis for a Poisson model is entirely similar.

Suppose we are to estimate for a sequence of n independent Bernoulli trials, the probability of failure p if X failures have been observed in this sequence. The (frequentist) estimator for p is:

$$\hat{p} = \frac{X}{n} \qquad (1)$$

which suggests that, if no failures were observed, the estimator is zero.

In a Bayesian setting, one starts from a prior distribution for the parameter p. Usually, the conjugate prior Beta distribution f_{pr} is used with pdf:

$$f_{\mathrm{pr}}(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1} \qquad (2)$$

where $B(a,b) = (\Gamma(a)\Gamma(b))/\Gamma(a+b)$ is the Beta function. The random variable with this pdf has a mean equal to $a/(a+b)$ and a standard variance equal to $ab/(a+b)^2(a+b+1)$. Its shape is determined by the parameters a and b. Taking $a = b = 1$ gives a uniform distribution over the unit interval. Sometimes also the Jeffrey's prior with

$a = b = 0.5$ is used, but here we will consider only the uniform prior, since it has some optimality properties as shown in the following. From this we then calculate the posterior using the likelihood of observations:

$$\ell(X|p) = \binom{n}{X} p^X (1-p)^{n-X} \tag{3}$$

yielding:

$$f_{\text{post}}(p|X,n) = \\ = \frac{1}{B(X+1,n-X+1)} p^X (1-p)^{n-X} \tag{4}$$

The Bayes estimator for a parameter is the mean of the posterior distribution f_{post}. Then the Bayes estimator for the probability p is in the case $X = 0$:

$$\hat{p} = \frac{1}{n+2} \tag{5}$$

These results can be found in any standard text about Bayesian methods, e.g. in Press (1989), p. 40.

So in the Bayesian setting the estimator is not equal to zero for $X = 0$ as in the frequentist case.

The next issue concerns the derivation of sensible confidence intervals for the probability p. This poses also significant problems in the case $X = 0$.

3. THE CLOPPER-PEARSON CONFIDENCE INTERVALS

The standard method for determining such intervals is described in detail in Clopper and Pearson (1934). They derive confidence intervals for the probability p in a frequentist setting. Unfortunately, this procedure does not give very satisfactory results for the case $X = 0$. In this case the confidence interval for the level α is in fact a confidence interval for the level $\alpha/2$, i.e. it is too large.

The original objective (Clopper and Pearson, 1934) is to calculate for a given confidence level α and an observed number of failures X as the lower bound p_L for the confidence interval, the value p_L for which:

$$\sum_{j=X}^{n} \binom{n}{j} p_L^j (1-p_L)^{n-j} = \frac{(1-\alpha)}{2} \tag{6}$$

and as the upper bound p_U the value for which:

$$\sum_{j=0}^{X} \binom{n}{j} p_U^j (1-p_U)^{n-j} = \frac{(1-\alpha)}{2} \tag{7}$$

One has for the sums using integration by parts:

$$\sum_{j=X}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \\ = \int_0^p \frac{t^{X-1}(1-t)^{n-X}}{B(X,n-X+1)} dt \tag{8}$$

as well as

$$\sum_{j=0}^{X} \binom{n}{j} p^j (1-p)^{n-j} = \\ = \int_p^1 \frac{(1-t)^{n-X-1} t^X}{B(X+1,n-X)} dt \tag{9}$$

So both bounds fulfilling the equations above can be found by inverting the incomplete Beta function with the respective parameters k and $n - k + 1$.

The authors do not consider in the paper the case of $X = 0$ separately, but from the diagrams one sees that the lower bound is simply set to zero, since the sum on the left side of Eq. (6) is equal to unity, so the equation cannot be fulfilled. Therefore the confidence level is determined by the second equation only resulting in an interval with confidence level $\alpha/2$. So its probability content is larger and it is longer.

The Bayesian probability interval $[p_L, p_U]$ for a specified "confidence" level α can be retrieved from the posterior pdf (4):

$$\Pr(p_L < p < p_U|X,n) = \\ = \int_{p_L}^{p_U} f_{\text{post}}(p|X,n)dp = \alpha \tag{10}$$

For $X = 0$, the lower bound p_L can be set to zero, without distorting the highest posterior density interval too much, so that p_U can be found from inverting the incomplete beta pdf with parameters $X + 1$ and $n - X + 1$ (Box and Tiao, 1992):

$$\int_0^{p_U} f_{\text{post}}(p|X,n)dp =$$
$$= \int_0^{p_U} \frac{p^X(1-p)^{n-X}}{B(X+1,n-X+1)}dp = \alpha \quad (11)$$

4. COMPARISON OF METHODS

In Pires and Amado (2008) a whole menagerie of about twenty different methods for calculating confidence intervals for the binomial proportions are presented. They are then compared with respect to criteria such as mean coverage probability and expected length.

For a given method with number *i*, which produces the confidence interval [$L_i(j)$, $U_i(j)$] for *j* successes in *n* experiments, the coverage probability CP for given *n* and given *p* is given by the summation:

$$\text{CP}(p,n,i) =$$
$$= \sum_{j=0} \binom{n}{j} p^j(1-p)^{n-j} I_{[L_i(j),U_i(j)]}(p) \quad (12)$$

with I_A the indicator function for the interval *A*. For a fixed *n*, the mean coverage probability is found as:

$$\mathbb{E}\big(\text{CP}(n,i)\big) = \int_0^1 \text{CP}(p,n,i)dp \quad (13)$$

Then the expected length of an interval calculated with method *i*, given *p* and *n* is:

$$\text{E}\ell(p,n,i) =$$
$$= \sum_{j=0} \binom{n}{j} p^j(1-p)^{n-j}\big(U_i(j) - L_i(j)\big) \quad (14)$$

Integrating over the parameter *p* gives then the overall expected length:

$$\mathbb{E}\big(\text{E}\ell(p,n,i)\big) = \int_0^1 \text{E}\ell(p,n,i)\,dp =$$
$$= \sum_{j=0} \binom{n}{j} \int_0^1 p^j(1-p)^{n-j}\big(U_i(j) - L_i(j)\big)\,dp \quad (15)$$

The coverage probabilities CP are shown for *n* = 40 and *n* = 400 in Pires and Amado (2008). These functions show quite erratic behavior as function of *p*.

The twenty methods given in Pires and Amado (2008) are compared considering mean coverage probability, maximum and minimum coverage probability and then also mean, minimum and maximum length of the intervals. A method which is optimal for all possible criteria cannot be found, so mean coverage probability and mean length seem to be the most useful ones. The mean coverage should be as close as possible to the nominal coverage for which the intervals are constructed and the length of the intervals should be minimal.

Consider now these two quality criteria for possible confidence intervals:

1. the mean coverage probability is to be as near as possible to the nominal coverage probability.
2. its length should be as short as possible.

Due to the definition of the Bayesian confidence interval its mean coverage probability is always equal to the nominal. Considering the second criterion of mean length the Bayesian highest posterior density interval is always optimal, see p. 190 in Pires and Amado (2008).

So the Bayesian approach using a posterior based on a uniform prior distribution gives "optimal" intervals if the criteria selected are the mean coverage probability and the mean length of the interval.

In their conclusions Pires and Amado (2008) consider only central intervals thereby excluding the Bayesian intervals citing computational challenges related to non-central intervals. The exclusion of non-central intervals is not really justifiable, since nowadays the computational effort in computing such intervals has largely become irrelevant. For these non-central intervals no clear optimal methods are singled out.

5. NEED FOR CONSISTENT METHODS

Consider once more the case where no failure has been observed till now. Let us assume two different methods for estimating the probability of such a failure have been applied, resulting in two different estimates of the probability:

$$\hat{p}_1 > \hat{p}_2 \quad (16)$$

From this follows that we can use for estimate of *p*:

$$p \approx \hat{p}_1 \quad (17)$$

$$p \approx \hat{p}_2 \quad (18)$$

Subtracting the equations and dividing by $\hat{p}_1 - \hat{p}_2$ gives:

$$0 \approx 1 \quad (19)$$

So one can derive from such inconsistent estimation methods false statements and *ex falso sequitur quodlibet*. How can we improve this situation?

There have been proposals to improve the estimator in the case of zero observations by adding "virtual" observations. For example, if in *n* trials no failure has been observed, then we might assume that the first failure is just around the corner, i.e. it happens at the next trial. So the estimator for *p* is then:

$$\hat{p}_1 = \frac{1}{n+1} = \frac{1}{n} - \frac{1}{n(n+1)} \quad (20)$$

But for large *n* this is approximately equal to 1/*n*, so we here give almost the same probability weight to a case where no failures have been observed as to a case where one failure has been observed in *n* trials for which the estimator is:

$$\hat{p}_2 = \frac{1}{n} \quad (21)$$

If this is now plugged into a series system with *k* identical components, we would have as estimate for the failure F_k of the system in case 1:

$$\widehat{\Pr(F_k^1)} \approx \frac{k}{n+1} = \frac{k}{n} - \frac{k}{n(n+1)} \quad (22)$$

whereas in the second case:

$$\widehat{\Pr(F_k^2)} \approx \frac{k}{n} \quad (23)$$

which for large *n* is almost the same. So for small *k* we see that the probability of failure here is estimated as almost the same in spite of the fact that we have no failure observed in the first case. This leads to an unjustified and "conservative" shift of resources.

The only reasonable recourse is to apply a fully Bayesian setting (Eqs. (2) to (5) above). Furthermore, one should also consider the actual context of the problem. For instance, if the data base is spatial in its structure – as in the case of pipeline incidents Canada-wide (TSB, 2013) – then a spatially hierarchical analysis can be performed where evidence of reported incidents is to a certain optimal extent "shared" over all assets resulting in a more relaxed treatment of "zero" incidents.

6. REFERENCES

Box G.E.P. and Tiao G.C. (1992) "*Bayesian Inference in Statistical Analysis*", Wiley, New York, 1992.

Clopper C.J. and Pearson E.S. (1934) "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial", *Biometrika*, 26:404–423, 1934.

Pires A. and Amado, C. (2008) "Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods", *REVSTAT – Statistical Journal*, 6(2):165–197, June 2008.

Press S.J. (1989) "*Bayesian Statistics: Principles, Models, and Applications*", Wiley, New York, 1989.

TSB (2013) "*Statistical Summary Pipeline Occurrences 2013*", Transportation Safety Board of Canada, http://www.tsb.gc.ca