Stoch Environ Res Risk Assess (2009) 23:9–12
DOI 10.1007/s00477-007-0190-6

ORIGINAL PAPER

Generous statistical tests

T. V. Hromadka II · R. J. Whitley · S. B. Horton · M. J. Smith · J. M. Lindquist

Published online: 11 October 2007

Abstract A common statistical problem is deciding which of two possible sources, A and B, of a contaminant is most likely the actual source. The situation considered here, based on an actual problem of polychlorinated biphenyl contamination discussed below, is one in which the data strongly supports the hypothesis that source A is responsible. The problem approach here is twofold: One, accurately estimating this extreme probability. Two, since the statistics involved will be used in a legal setting, estimating the extreme probability in such a way as to be as generous as is possible toward the defendant's claim that the other site B could be responsible; thereby leaving little room for argument when this assertion is shown to be highly unlikely. The statistical testing for this problem is modeled by random variables $\{X_i\}$ and the corresponding sample mean $\bar{X} = \frac{1}{n}S_n,\ S_n = \sum_1^n X_i$, the problem considered is providing a bound ε for which $\text{Prob}(\bar{X} \geq a_0) \leq \varepsilon$, for a given number a_0. Under the hypothesis that the random variables $\{X_i\}$ satisfy $E(X_i) \leq \mu$, for some $0 < \mu < 1$, statistical tests are given, described as "generous", because ε is maximized. The intent is to be able to reject the hypothesis that a_0 is a value of the sample mean while eliminating any possible objections to the model distributions chosen for the $\{X_i\}$ by choosing those distributions which maximize the value of ε for the test used.

Keywords Testing unlikely events · Extreme deviations · PCB contamination

T. V. Hromadka II · S. B. Horton · M. J. Smith · J. M. Lindquist
Department of Mathematical Sciences,
United States Military Academy,
West Point, NY 10096, USA
e-mail: ted@phdphdphd.com

S. B. Horton
e-mail: steve.horton@usma.edu

M. J. Smith
e-mail: Mick.Smith@usma.edu

J. M. Lindquist
e-mail: aj0558@usma.edu

R. J. Whitley (✉)
P.O. Box 11133, Bainbridge Island, WA 98110, USA
e-mail: rwhitley@math.uci.edu

1 Introduction

A practical example of the problem under consideration involves multiple sources of conservative contamination, such as PCB contaminations, the question being whether or not certain measured concentrations at one particular location are the result of contaminant transport from another location. For example, in the transport of PCBs in sediment, the movement of water can transport both bedload and washload in the water and PCBs can adhere to a wide range of particle sizes of the sediment, and depending on the water flow characteristics, sediment can be entrained into the water and moved downstream where it can be deposited. The example considered in this paper is the comparison of concentrations of the contaminant between two locations where the sampling process involves a relatively large number of grains of sediment. The specific setting is the sampling of sediment in a water channel or water course, where some ten thousand or more grains of sediment are involved in the containment measurement process as is done when measuring PCBs. In the specific problem studied, a sample of 4,890 parts per million (ppm) of PCB sediment was measured at a location downstream of a possible source of PCB in sediment which had

concentrations not exceeding 360 ppm. The question is how likely is it that the 4,890 ppm concentration is but a part of the upstream sediment source.

The general mathematical approach taken was to formulate the problem in such a way as to be as generous as possible, in the sense that the distributional assumptions were made in such a way as to favor the hypothesis that the contamination came from the upstream site, and thereby to avoid unnecessary controversy over how the statistical bounds were computed when this hypothesis is rejected.

Two estimates are considered. In the first estimate, a probability distribution is developed for the population of sediment grains at the alleged source that maximizes the variance of the distribution of PCBs at the alleged source and also maximizes the variance of the particularly transportable sediment grains containing such PCBs to move downstream. The variance is maximized by using the standard Bernoulli distribution with two outcomes of concentration, either 0 or maximum, in the sediment where the population mean in this example is 360 ppm and the maximum concentration is pure PCB at 1,000,000 ppm. For the distribution with this maximum variance, the standard Chebyshev bounds for the probability that the upstream site was the source of the 4,890 ppm sample is shown to be maximized, and is small for the data under consideration. The procedure here is entirely elementary and gives a simple example of computing generous bounds.

The second estimate uses a large deviation inequality, with a correspondingly much smaller bound for the probability that the sample came from the upstream site. This estimate is also maximized by the Bernoulli distribution as before, but in this case the reason lies deeper than a simple variance maximization.

The data at the site consisted of, in round numbers, 90 samples, each of which being the mean of the PCB contaminant in a large number n of particles, say $n = 10{,}000$. These values for the mean were in general quite small, and it was assumed on the basis of the data that the distribution describing the PCB distribution on a particle had mean bounded by the value 360 ppm. The value to test was $a_0 =$ 4,890 ppm. The problem was to find a bound for the (obviously small) probability that a_0 came from this site.

The data above will be normalized by dividing the values by 10^6 so as to scale all the numbers to the interval [0,1]. Then each $E(X_i)$ (see next paragraph) is bounded by $3.60 \times 10^{-4} = \mu$; the value to be tested is $a_0 = 4.89 \times 10^{-3}$; and $X_i = 0$ if the particle has no PCB contamination, while $X_i = 1$ denotes a particle completely composed of PCB.

The general model is: an independent set of random variables $\{X_i : i = 1, \ldots, n\}$ are given, each with values in [0,1], and which are each constrained by $E(X_i) \le \mu$ for some given μ, $0 < \mu < 1$. The tests will be for the sample mean of the n test particles, $\bar{X} = \frac{S_n}{n}$, where $S_n = \sum_1^n X_i$, and will have the form

$$\text{Prob}(\bar{X} \ge a_0) \le \varepsilon. \tag{1}$$

The problem considered here is, given a specific method for computing a bound ε, how can this be done in a way which most favors the acceptance of the hypothesis that a_0 is a possible value of the sample mean, i.e. which maximizes ε.

2 Chebyshev's Inequality

One simple procedure would be to use Chebshev's inequality to obtain a bound as in Eq. (1). The following elementary theorem will determine the most generous such test.

Theorem 1 *Let X be a random variable with values in* [0,1] *and* $E(X) \le \mu$ *for some* $0 < \mu < 1$. *Then, letting* var (X) *denote the variance of X, if* $0 < \mu \le \frac{1}{2}$, *then*

$$\text{var}(X) \le \mu(1-\mu)$$

and equality is obtained if and only if $X = B$, B *the Bernoulli random variable defined by* Prob $(B = 0) = 1 - \mu$ and $P(B = 1) = \mu$.

If $\frac{1}{2} \le \mu < 1$, *then*

$$\text{var}(X) \le \frac{1}{4}$$

and equality is obtained if and only if $X = B'$, B' *the Bernoulli random variable defined by* $\text{Prob}(B' = 0) = P(B' = 1) = \frac{1}{2}$.

Proof Because $0 \le X \le 1$, $E(X^2) \le E(X) \le \mu$, and so var $(X) \le E(X) - E(X)^2$. For $\mu \le \frac{1}{2}$, the function $t(1-t)$ has its maximum on $[0,\mu]$ at μ, var $(X) \le \mu(1-\mu) =$ var (B), while for $\frac{1}{2} \le \mu < 1$, the maximum occurs at $\frac{1}{2}$ and $\text{var}(X) \le \frac{1}{4} = \text{var}(B')$.

Suppose $0 < \mu \le \frac{1}{2}$ and that X has the maximum variance $\mu(1-\mu)$. As

$$\mu(1-\mu) \ge E(X) - E(X)^2 \ge E\left(X^2\right) - E(X)^2 = \text{var}(X) = \mu(1-\mu),$$

it follows that $E(X) = \mu = E(X^2)$. Let $F_X(x) = \text{Prob}(X \le x)$ be the distribution function for X. Since $E(X) = E(X^2)$, the integral $\int_0^1 (x - x^2)\, dF_X$ is zero. The integrand $x(1-x)$ is positive on (0,1), so X must have mass only at 0 and 1, and, as $E(X) = \mu$, it must be that $X = B$.

If $\frac{1}{2} \le \mu < 1$, it follows as above that $E(X) = E(X^2) = \frac{1}{2}$, and that this maximum is attained only for B'. □

If $\mu \leq \frac{1}{2}$, then (with a similar argument for $\frac{1}{2} < \mu \leq 1$) use Chebyshev's inequality to bound the probability in (1):

$$\text{Prob}(\bar{X} \geq a_0) = \text{Prob}(\bar{X} - \mu \geq a_0 - \mu) \leq \frac{\text{var}(\bar{X})}{(a_0 - \mu)^2}. \quad (2)$$

By the independence of the X_i

$$\text{var}(\bar{X}) = \frac{1}{n^2}\sum_{1}^{n} \text{var}(X_i) \leq \frac{1}{n}\mu(1-\mu). \quad (3)$$

The result above shows that if each X_i is replaced by an independent Bernoulli random variable with mean μ, the resulting variance of the sample mean is maximized, which maximizes the right-hand side of (2) so that this test gives the largest value of ε.

Note that there is no assumption that the X_i have a common distribution. This leaves open the possibility of modeling the PCB problem by having the particles consist of several different types, each type with a different distribution of PCB.

For the PCB data, $n = 10{,}000$, $\mu = 360 \times 10^{-6}$, and $a_0 = 4890 \times 10^{-6}$,

$$\text{Prob}(\bar{X} \geq a_0) \leq 1.75 \times 10^{-3};$$

and it is unlikely that a_0 is a value of the sample mean. A sharper estimate will be given in the next section.

3 Chernoff's large deviation inequality

The appearance of the sample mean taken over a large number of terms suggests the use of the Central Limit Theorem. In the application of the Central Limit Theorem, the sample mean is scaled by subtracting its mean and dividing by its standard deviation, and considering the random variable

$$\hat{Z} = \frac{\bar{X} - \mu_X}{\sqrt{\text{var}(\bar{X})}},$$

which is distributed approximately like a normal random variable with mean 0 and standard deviation 1. With this normalization, the left-hand side of Eq. (1) becomes

$$\text{Prob}\left(\hat{Z} \geq \frac{a_0 - \mu_X}{\sqrt{\text{var}(\bar{X})}}\right). \quad (4)$$

The presence of the variance in the denominator shows that, if $\hat{Z}$ were exactly $N(0,1)$, the most generous choice would be for each X_i to have the Bernoulli distribution B of the previous section. However, there are problems with this approach. For one, there are rough rules-of-thumb for the application of the Central Limit Theorem approximation to the binomial distribution $B(n,p)$, for example that if $np(1-p) \geq 10$ the approximation will "generally be quite good". [4, p. 89] This criterion is not satisfied for the PCB example, but using $\mu = 3 \times 3.60 \times 10^{-4}$ takes care of that in a generous manner.

The essential problem is that however the phase "generally quite good" is interpreted, it surely is not meant to apply to a situation where the test variable is many standard variations from the mean; in the PCB case, more than ten. What is needed in this case is a large deviation result. Chernoff's Theorem [3, 5.4.6; 5, 1.2] gives an inequality of the form

$$\text{Prob}\left(\frac{S_n}{n} \geq a\right) \leq e^{-nh(a)} \quad (5)$$

where $h(a)$ is a function depending on the distribution function of the random variable X, the assumption in the theorem as stated in [3] and [5] being that each X_i has the same distribution as X. Since the statement of Chernoff's Theorem does not involve the variance of the sample mean, it is not easy, as it was in Theorem 1, to see what assumptions on the distributions X_i minimize $h(a)$. To establish this, a proof is given below, modeled on that of [3], of the more general situation where the X_i need not have a common distribution.

Theorem 2 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with values in* [0,1] *and $E(X_i) \leq \mu$ for each i for some $0 < \mu < 1$. Then for $a > \mu$,*

$$\text{Prob}\left(\frac{S_n}{n} \geq a\right) \leq e^{-nh_B(a)} \quad (6)$$

where $h_B(a)$ is the Cramer transform [2, 3.3.5] (*discussed below*) *for the Bernoulli random variable B.*

Proof Let $I(t)$ be the indicator function of the interval $[0,\infty)$, so that $I(t) = 1$ if $t \geq 0$ and $I(t) = 0$ if $t < 0$. Let a be given, $\mu < a < 1$. For $x > 0$, and $S_n = X_1 + X_2 + \ldots + X_n$,

$$I(S_n - na) \leq e^{x(S_n - na)}$$

and so

$$P\left(\frac{S_n}{n} \geq a\right) = P(S_n - na \geq 0) = E(I(S_n - na))$$
$$\leq E\left(e^{x(S_n - na)}\right) = e^{-xna}E\left(e^{xS_n}\right) = e^{-xna}\prod_{i=1}^{n} E\left(e^{xX_i}\right)$$
$$\leq e^{-xna}\prod_{i=1}^{n}\left(1 + \mu\sum_{n=1}^{\infty}\frac{x^n}{n!}\right).$$

Thus

$$P(S_n - na \geq 0) \leq e^{-xna} \prod_{i=1}^{n} \phi_B(x) = e^{-xna}[\phi_B(x)]^n = e^{-n[ax - \log \phi_B(x)]}.$$

The function $\phi_{X_i}(x) = E(e^{xX_i})$ is the moment gathering function of X_i. Since $0 \leq X_i \leq 1$,

$$E(X_i^n) \leq E(X_i^{n-1}) \leq \ldots \leq E(X_i) \leq \mu,$$

and

$$\phi_{X_i}(x) = E(e^{xX_i}) = E\left(\sum_0^\infty \frac{(xX_i)^n}{n!}\right) = \sum_0^\infty \frac{x^n}{n!} E(X_i^n)$$

which is less than or equal to

$$1 + \mu \sum_1^\infty \frac{x^n}{n!} = 1 - \mu + \mu e^x = E(e^{xB}) = \phi_B(x),$$

where $\phi_B(x)$ is the moment generating function for B. Thus

$$\text{Prob}(S_n \geq na) \leq e^{-n[ax - \log \phi_B(x)]},$$

from which it follows that Prob $(S_n \geq na)$ is bounded by

$$\inf_{x>0} e^{-n[xa - \log \phi_B(x)]} = e^{-n \sup_{x>0}[xa - \log \phi_B(x)]}.$$

The function $\sup_{x>0}[ax - \log \phi_B(x)] = h_B(a)$ is the Cramer transform of B [2, 3.3.5], [3, 5.4.1] which establishes (6). □

The computation of the general Cramer transform is discussed in [2, 3.3.5]. Let $g(x) = ax - \log \phi_B(x)$. This function is strictly concave, $g(0) = 0$, $g'(0) = a - \mu > 0$, and $g(x)$ tends to $-\infty$ as x tends to infinity, so $g(x)$ has a unique positive maximum at the point x_0 where $g'(x_0) = 0$. This point is $x_0 = \log\left(\frac{a(1-\mu)}{\mu(1-a)}\right)$, and

$$h_B(a) = a \log\left(\frac{a(1-\mu)}{\mu(1-a)}\right) - \log\left(\frac{1-\mu}{1-a}\right).$$

Applied to the PCB data this gives the bound

$$\text{Prob}(\bar{X} \geq a_0) \leq e^{10^4 h_B(a_0)} = 1.29 \times 10^{-36}$$

and it is extremely unlikely that a_0 is a value of the sample mean. It is interesting to compare this estimate with an application of the Central Limit Theorem; in this case the value a_0 lies approximately 25 standard deviations from the mean and the probability of the corresponding tail of the normal is about 3×10^{-137} [1, Table 26.2]. This number is much smaller than the large deviation estimate given above, but it is not an accurate bound as neither the Central Limit Theorem nor simulations can address what happens so far out on the tail of the distribution, and so is best regarded as a rough indication of the price paid for using the rigorous large deviation estimate.

4 Conclusion

The results obtained apply to random variables $\{X_i\}$ each with values in [0,1] and corresponding sample mean $\bar{X} = \frac{1}{n} S_n$, $S_n = \sum_1^n X_i$ under the hypothesis that the random variables $\{X_i\}$ satisfy $E(X_i) \leq \mu$, for some $0 < \mu < 1$. Statistical tests are given, described as "generous" because the bound ε is maximized in the expression $\text{Prob}(\bar{X} \geq a_0) \leq \varepsilon$ by choice of distribution, for a given number a_0 and for a given method of bounding this probability. For tests using Chebyshev's inequality and Chernoff's large deviation inequality there is a choice of statistical distribution which maximizes the above probability that a sample mean from that distribution was greater than or equal to a given (relatively large) value a_0. Using either of these tests, one can reject the hypothesis that a_0 is a value of the sample mean while at the same time eliminating any possible objections that the model distributions chosen for the $\{X_i\}$ were biased in favor of rejection. This is useful in legal applications. This "generosity" is only possible if the value a_0 is extremely unlikely. In such a case, the large deviation calculation given here provides a rigorous estimate under circumstances which would commonly be handled by an unjustified use of the Central Limit Theorem.

References

1. Abramowitz M, Stegun I (1965) Handbook of Mathematical Functions, National Bureau of Standards. Applied Math Series 55, US Government Printing Office, Washington
2. Dacunha-Castelle D, Duflo M (1986) Probability and statistics I. Springer, New York
3. Dacunha-Castelle D, Duflo M (1986) Probability and statistics II. Springer, New York
4. Ross S (2003) Introduction to orobability models. 8th edn. Academic, Boston
5. Schwartz A, Weiss A (1996) Large deviations for performance analysis. Chapman and Hall, London