UNBIASED ESTIMATE FOR b -VALUE OF MAGNITUDE FREQUENCY

Yosihiko OGATA* and Ken'ichiro YAMASHINA**

* *Institute of Statistical Mathematics, Tokyo, Japan*

** *Earthquake Research Institute,
The University of Tokyo, Tokyo, Japan*

(Received October 18, 1985; Revised May 8, 1986)

If we assume that magnitudes of earthquakes are distributed identically and independently according to a negative-exponential function, then the maximum likelihood estimate proposed by Utsu for the b -value is biased from the true value. We suggest an unbiased alternative estimate which is asymptotically equivalent to the maximum likelihood estimate. The relation between the unbiased estimate and the maximum likelihood estimate are presented from a Bayesian viewpoint. The two estimates are compared in order to show the superiority of the unbiased estimate over the maximum likelihood estimate, on the basis of the expected entropy maximization principle for the predictive distributions. From the same principle, the posterior with the noninformative improper prior is recommended for the inferential distribution of the b -value rather than the standardized likelihood.

1. *Bias of the Maximum Likelihood Estimate*

According to Gutenberg-Richter's law (GUTENBERG and RICHTER, 1944), the number $N(M)$ of earthquakes having magnitude equal to or larger than M can be expressed by the equation

$$\log_{10} N(M) = a - bM. \quad (1)$$

UTSU (1965) proposed the estimate for b with the equation

$$\hat{b} = \frac{N \log_{10} e}{\sum_{i=1}^N (M_i - M_0)} \quad (2)$$

where N is the number of earthquakes and M_0 is the minimum magnitude in a given sample. AKI (1965) suggested that this is nothing but the maximum likelihood estimate which maximizes the likelihood function

$$L(b) = \prod_{i=1}^N f(M_i | \beta), \quad (3)$$

where

$$f(M|\beta) = \beta e^{-\beta(M-M_0)}, \quad (M \geq M_0), \quad (4)$$

and $\beta = b/\log_{10}e$. Using the large sample theory for the maximum likelihood, AKI (1965) provided the asymptotic error band for $\hat{b}$. It is further seen that the maximum likelihood estimate Eq. (2) tends to be the true value b_0 and that $\hat{b}$ is the asymptotically unbiased estimate of b_0 ; i.e., $E(\hat{b}) - b_0 \rightarrow 0$ as N increases.

For fixed N , however, $\hat{b}$ is not unbiased. Indeed, using the exact distribution density of $\hat{b}$ in (2)

$$f(b) = \frac{N^N}{\Gamma(N)} \frac{1}{b_0} \left(\frac{b_0}{b}\right)^{N+1} \exp\left(-\frac{Nb_0}{b}\right), \quad (5)$$

which is derived from Eq. (4) by UTSU (1966), we have

$$E(\hat{b}) = \int_0^\infty bf(b)db = \frac{Nb_0}{N-1} = b_0 + \frac{b_0}{N-1}. \quad (6)$$

This suggests that the bias is not small when N is small or b_0 is large.

A natural way for correcting the bias of $\hat{b}$ arises immediately based on Eq. (6); that is,

$$\tilde{b} = \frac{N-1}{N} \hat{b} = \frac{(N-1) \log_{10}e}{\sum_{i=1}^N (M_i - M_0)}, \quad (7)$$

which is the unbiased estimate for b_0 . Although there are many unbiased estimators of b_0 besides $\tilde{b}$, we will show that the estimator Eq. (7) enjoys certain optimality.

2. Relation between $\hat{b}$ and $\tilde{b}$ from a Bayesian Viewpoint

Without loss of generality we shall consider the relationship between $\hat{\beta} = N / \sum_{i=1}^N X_i$ and $\tilde{\beta} = (N-1) / \sum_{i=1}^N X_i$ instead of $\hat{b}$ and $\tilde{b}$, respectively, where $X_i = M_i - M_0$ is distributed according to $f(x|\beta) = \beta e^{-\beta x}$. Suppose there is a prior distribution $\pi(\beta)$ for the parameter β . Then by applying the Bayes theorem we have the posterior distribution

$$p(\beta) = \frac{\pi(\beta)L(\beta)}{\int_0^\infty \pi(\beta)L(\beta)d\beta}. \quad (8)$$

Consider the case where

$$\pi(\beta) = 1, \quad 0 \leq \beta < \infty, \quad (9)$$

although this is not the probability density; this is called the uniform improper prior. Then Eq. (8) provides the standardized likelihood which is conventionally used for the confidence probability distribution;

$$p_{\text{SL}}(\beta) = \frac{\{\sum X_i\}^{N+1}}{N!} \beta^N e^{-\beta \sum X_i} \quad (10)$$

where the sum $\sum$ is taken from $i=1$ up to $i=N$, and the mode is given by the maximum likelihood estimate $\hat{\beta}$. In general, the uniform improper prior Eq. (9) does not seem to be appropriate when characterizing a situation where "nothing (or more realistically, little) is known a priori." JEFFEREYS (1961) suggested the rule that noninformative prior distribution is given by the square root of Fisher's information measure. The rule is justified on the grounds of its invariance under parameter transformations; see also BOX and TIAO (1973) and AKAIKE (1978) for examples. In the present case the Fisher's information measure is

$$I(\beta) = E \left[-\frac{\partial^2}{\partial \beta^2} \log f(M|\beta) \right] = 1/\beta^2.$$

Thus the noninformative prior is improper and $\pi(\beta) = 1/\beta$, ($0 < \beta < \infty$). Therefore, the posterior probability is given by

$$p_{\text{NI}}(\beta) = \frac{(\sum X_i)^N}{(N-1)!} \beta^{N-1} e^{-\beta \sum X_i} \quad (11)$$

where the sum $\sum$ is taken from $i=1$ up to $i=N$. It is easily seen that the mode of the posterior distribution Eq. (11) is $\hat{\beta}$, while the posterior mean (Bayes estimate) is

$$\int_0^\infty \beta p_{\text{NI}}(\beta) d\beta = \frac{N}{\sum X_i} \quad (12)$$

which is equal to the maximum likelihood estimate $\hat{\beta}$.

3. Entropy Maximization Principle and the Performance of the Estimates

AKAIKE (1977) introduced and formulated the entropy maximization principle, which may be specifically described for our purpose as follows. Denote by $\mathbf{x}$ the vector of observations $(x_1, x_2, \dots, x_N)$. Assume that the true distribution of $\mathbf{x}$ is specified by an unknown density function $f(\mathbf{z}) = f(z_1, z_2, \dots, z_N)$ and is considered to be included in a parameterized density $g(\mathbf{z}|\theta)$. Considering $\mathbf{y} = (y_1, y_2, \dots, y_N)$ as the future observation with the same distribution as $\mathbf{x}$, and supposing that we predict it based on the present data set $\mathbf{x}$, we may call $g(\mathbf{y}|\tilde{\theta})$ a predictive distribution, where $\tilde{\theta} = \tilde{\theta}(\mathbf{x})$ is an estimate of the true parameter θ_0 satisfying $f(\mathbf{y}) = g(\mathbf{y}|\theta_0)$.

As the measure of the correct fit of $g(\mathbf{y}|\tilde{\theta})$ to $f(\mathbf{y})$ we use the entropy of $f(\mathbf{y})$ with respect to $g(\mathbf{y}|\tilde{\theta})$, defined by

$$B(f; g) = - \int \left\{ \frac{f(\mathbf{y})}{g(\mathbf{y}|\tilde{\theta})} \right\} \log \left\{ \frac{f(\mathbf{y})}{g(\mathbf{y}|\tilde{\theta})} \right\} g(\mathbf{y}|\tilde{\theta}) d\mathbf{y}. \quad (13)$$

Note here that $B(f; g)$ is the function of $\mathbf{x}$. Thus we consider the expected entropy

$$\begin{aligned}
 J(\tilde{\theta}) &= \int B(f; g) f(x) dx = E_X[B(f; g)] \\
 &= -E_X E_Y[\log \{f(Y)/g(Y | \tilde{\theta}(X))\}], \quad (14)
 \end{aligned}$$

where E_X and E_Y are expectations with respect to the random vectors X and Y , respectively. AKAIKE (1977) justified the use of this measure based on the process of conceptual sampling experiments, and described the natural relation to the log likelihood. Further, the AIC was derived from this quantity (AKAIKE, 1973).

Here we use the expected entropy $J(\tilde{\theta})$ to measure the performance of the estimates; a larger value shows the better fit. Suppose $\{X_i\}$ and $\{Y_i\}$, $i=1, 2, \dots, N$ are identically and independently distributed according to the density function $g(x | \beta_0) = \beta_0 e^{-\beta_0 x}$. Then we have

$$J(\hat{\beta}) = N \left\{ \log N - \frac{1}{N-1} - \Psi(N) \right\} \quad (15)$$

and

$$J(\tilde{\beta}) = N \{ \log(N-1) - \Psi(N) \}. \quad (16)$$

Here we have used the equalities

$$E \left[1 / \sum_{i=1}^N X_i \right] = \beta_0 / (N-1) \text{ and } E \left[\log \sum_{i=1}^N X_i \right] = \Psi(N) - \log \beta_0,$$

and $\Psi(N)$ is the digamma function such that $\Psi(z) = d/dz \log \Gamma(z)$. Note that both $J(\hat{\beta})$ and $J(\tilde{\beta})$ are independent of β_0 . Since

$$\{J(\tilde{\beta}) - J(\hat{\beta})\} / N = \frac{N}{N-1} - 1 - \log \frac{N}{N-1} > 0,$$

we see the superiority of the estimates $\tilde{\beta} = (N-1) / \sum X_i$.

Furthermore, consider the family of estimators of the form $\xi / \sum X_i$ where ξ is a positive constant. Here we find the optimal ξ as follows: By substituting this for $\tilde{\theta}(X)$ in (14), we have

$$J = N \left[1 + \log \xi - \Psi(N) - \frac{\xi}{N-1} \right].$$

From this we see that $\xi = N-1$ maximizes the expected entropy Eq. (14).

There are many losses other than in Eq. (14). For example, one should consider the expected square loss $E_X[(\theta(X) - \theta_0)^2]$ to measure the performance of the estimators. For the estimators of the form $\xi / \sum X_i$ we have

$$E[(\xi / \sum X_i - \beta_0)^2] = \frac{\beta_0^2}{(N-1)(N-2)} \{ \xi^2 - 2(N-2)\xi + (N-1)(N-2) \}. \quad (17)$$

This equality suggests that the unbiased estimator $\tilde{\beta} = (N-1) / \sum X_i$ is better than the

Table 1. Comparison of the average absolute loss.

N	$b_0=0.5$			$b_0=1.0$			$b_0=1.5$		
	$\xi=N$	$\xi=N-1$	$\xi=N-2$	$\xi=N$	$\xi=N-1$	$\xi=N-2$	$\xi=N$	$\xi=N-1$	$\xi=N-2$
2	0.656	0.374	—	1.191	0.694	—	1.673	1.020	—
3	0.391	0.275	0.294	0.794	0.573	0.618	1.154	0.829	0.902
4	0.277	0.224	0.241	0.571	0.456	0.492	0.872	0.682	0.714
5	0.235	0.197	0.207	0.500	0.416	0.428	0.658	0.552	0.595
6	0.220	0.186	0.187	0.410	0.351	0.364	0.593	0.507	0.524
7	0.181	0.157	0.161	0.371	0.320	0.324	0.562	0.488	0.493
8	0.154	0.139	0.146	0.313	0.278	0.289	0.468	0.431	0.459
9	0.154	0.139	0.142	0.303	0.274	0.279	0.453	0.416	0.431
10	0.144	0.131	0.133	0.284	0.259	0.263	0.405	0.374	0.387
11	0.137	0.124	0.125	0.274	0.249	0.249	0.420	0.387	0.386
12	0.131	0.119	0.117	0.259	0.239	0.239	0.368	0.347	0.354
13	0.125	0.115	0.115	0.258	0.241	0.242	0.370	0.336	0.334
14	0.113	0.107	0.109	0.243	0.228	0.228	0.337	0.318	0.325
15	0.112	0.104	0.104	0.216	0.207	0.213	0.327	0.310	0.313
16	0.106	0.101	0.103	0.214	0.207	0.212	0.326	0.311	0.313
17	0.102	0.096	0.096	0.200	0.188	0.190	0.322	0.302	0.299
18	0.101	0.096	0.097	0.197	0.186	0.188	0.306	0.288	0.287
19	0.094	0.090	0.092	0.202	0.191	0.189	0.281	0.268	0.273
20	0.091	0.088	0.090	0.198	0.186	0.185	0.278	0.267	0.269

maximum likelihood estimator $\hat{\beta} = N/\sum X_i$, although the minimum value of Eq. (17) is attained by the estimator $(N-2)/\sum X_i$. Of further interest is the loss to see in the absolute value of the difference, $|\theta(X) - \theta_0|$. With the Monte Carlo experiment the comparison has been performed for the estimator $N/\sum X_i$, $(N-1)/\sum X_i$ and $(N-2)/\sum X_i$. The trial was repeated 10,000 times for β_0 such that $b_0 = \beta_0 \log_{10} e = 0.5$, 1.0, and 1.5. The sample averages of the loss are shown in Table 1, which shows that the unbiased estimator $\tilde{\beta}$ performs best.

Thus we see that different losses lead to different preferences with the estimators. Therefore we should be cautious about the meaning of losses. Our purpose here is to compare the performance of the estimator $\tilde{b}$ with the maximum likelihood estimator $\hat{b}$. The maximum likelihood method (as well as AIC) is viewed as an optimal realization of the entropy maximization principle (AKAIKE, 1973, 1977); we already know that the expected log likelihood is almost the probabilistic definition of entropy itself. Thus, we have used the loss Eq. (14) to examine the effect of $\tilde{b}$ as the refinement of the maximum likelihood estimator.

4. Predictive Distributions Using Posteriors

In the previous section we have only considered the case where the predictive distribution is given in the form of $g(y|\tilde{\theta})$ by substituting the point estimate $\tilde{\theta} = \tilde{\theta}(x)$.

Table 2. Comparison of the expected entropy.

N^*	$J(\hat{\beta})$	$J(\tilde{\beta})$	J_{SL}	J_{NI}
2	-1.45927	-0.845569	-0.527329	-0.387143
3	-0.972515	-0.688911	-0.464194	-0.374007
4	-0.812627	-0.630022	-0.433636	-0.367259
5	-0.733397	-0.599116	-0.415649	-0.363163
6	-0.686148	-0.580077	-0.403812	-0.360415
7	-0.654785	-0.567174	-0.395440	-0.358451
8	-0.632458	-0.557851	-0.389193	-0.356970
9	-0.615752	-0.550799	-0.384364	-0.355816
10	-0.602784	-0.545278	-0.380518	-0.354893
11	-0.592429	-0.540844	-0.377369	-0.354132
12	-0.583967	-0.537191	-0.374781	-0.353516
13	-0.576928	-0.534152	-0.372569	-0.352963
14	-0.570977	-0.531564	-0.370684	-0.352512
15	-0.565872	-0.529339	-0.369063	-0.352119
16	-0.561464	-0.527409	-0.367636	-0.351769
17	-0.557603	-0.525722	-0.366381	-0.351463
18	-0.554195	-0.524224	-0.365286	-0.351204
19	-0.551173	-0.522897	-0.364291	-0.350956
20	-0.548462	-0.521698	-0.363402	-0.350758

* N is the sample size.

AKAIKE (1978) suggested that the predictive distribution of the form $\int g(y|\theta)p(\theta|x) d\theta$ fitted better on average than a particular $g(y|\hat{\theta})$ with $\hat{\theta}$ randomly chosen according to the posterior distribution $p(\theta|x)$.

The implication of this is that the full use of posterior distribution performs better than the point estimates. For example the posteriors Eqs. (10) and (11) for the estimate of b -value provides the so-called confidence limits or error bars of the inferential distributions. Incidentally, both distributions Eqs. (10) and (11) tend to be normal distributions in a sense as the sample size increases, and the confidence limits provided in AKI (1965), for example, are obtained.

To see the performance of posteriors with a fixed sample size we calculate the expected entropy. For the predictive distribution with the posterior Eq. (10), we have

$$J_{SL} = N + \log \frac{\Gamma(2N)}{\Gamma(N)} - (2N+1)\Psi(2N) + (N+1)\Psi(N) + \log 2, \quad (18)$$

and for the predictive distribution with the posterior Eq. (11), we have

$$J_{NI} = N + \log \frac{\Gamma(2N)}{\Gamma(N)} - 2N\Psi(2N) + N\Psi(N). \quad (19)$$

Note that these quantities also are independent of the true value of β_0 . Therefore

$$J_{NI} - J_{SL} = \Psi(2N) - \Psi(N) - \log 2 = \sum_{k=N}^{2N-1} 1/k - \log 2 > 0, \quad (20)$$

which suggests that the posterior distribution with the noninformative prior works

better than one with the uniform prior from the predictive viewpoint.

To compare the values of Eqs. (15), (16), (18), and (19), we carried out numerical calculation using the approximation of the digamma function for the large x

$$\Psi(x) \sim \log x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6} + \frac{1}{240x^8} - \frac{1}{132x^{10}}$$

and also used the relation $\Psi(x+1) = \Psi(x) + 1/x$. Table 2 suggests that the predictive distribution used in averaging the posterior is more effective than those based on the point estimates.

5. Discussion

In personal communication Dr. Masuda pointed out that the different parameterization of Gutenberg-Richter's law $\log_{10} N(M) = a - M/s$ instead of Eq. (1), or the density function $f(M|\sigma) = \sigma^{-1} e^{-(M-M_0)/\sigma}$ where $\sigma = s/\log_{10} e$ instead of Eq. (4), leads to the unbiased maximum likelihood estimator for σ . Indeed, the standard likelihood is given by

$$P_{\text{SL}}(\sigma) = \frac{(\sum X_i)^{N-1}}{(N-2)!} \left(\frac{1}{\sigma} \right)^N e^{-\sum X_i/\sigma}, \quad (21)$$

and this mode $\hat{\sigma} = \sum X_i/N$ is unbiased. However, if we examine a family of estimators in the form of $\sigma_{\xi} = \sum X_i/\xi$ where ξ is a positive constant, then we have, as in section 3,

$$J(\sigma_{\xi}) = N \left[1 + \log \xi - \Psi(N) - \frac{\xi}{N-1} \right].$$

From this we see that $\xi = N-1$ maximizes the expected entropy, and $\tilde{\sigma} = \sum X_i/(N-1)$ performs best as the point estimator.

Furthermore, since the noninformative prior is $\pi(\sigma) = 1/\sigma$ ($0 < \sigma < \infty$), the corresponding posterior to Eq. (11) is given by

$$P_{\text{NI}}(\sigma) = \frac{(\sum X_i)^N}{(N-1)!} \left(\frac{1}{\sigma} \right)^{N+1} e^{-\sum X_i/\sigma}. \quad (22)$$

By a similar computation in the previous section we have

$$\begin{aligned} J_{\text{NI}} - J_{\text{SL}} &= -\Psi(2N) + \Psi(N) + \log \frac{2N-1}{N-1} \\ &= \log(2N-1) - \log(N-1) - \sum_{k=N}^{2N-1} \frac{1}{k} > 0, \quad (N \geq 2), \end{aligned} \quad (23)$$

instead of Eq. (20). This suggests that the same consequence as that in section 4 holds for the present parameterization of Gutenberg-Richter's law.

6. Concluding Remarks

The entropy maximization principle suggests that the full use of posterior Eq. (11) is the most effective among the considered estimations for the b -value. Thus the confidence limits, for example, should be made based on the posterior Eq. (11). As far as the point estimation is concerned the unbiased estimate $\bar{\beta} = (N-1)/\sum X_i$ is better than the maximum likelihood estimate $\hat{\beta} = N/\sum X_i$, and furthermore is the best among the $\xi/\sum X_i$ type family.

We are grateful to Sigeo Aki for helpful information in the numerical calculation of the digamma function. We also wish to thank Dr. T. Masuda of Tohoku University for his useful comments.

REFERENCES

- AKAIKE, H., Information theory and an extension of the maximum likelihood principle, in *2nd International Symposium on Information Theory*, ed. B. N. Petrov and F. Csaki, pp. 267-281, Akademiai Kiado, Budapest, 1973.
- AKAIKE, H., On entropy maximization principle, in *Applications of Statistics*, ed. P. R. Krishnaiah, pp. 27-41, North-Holland, Amsterdam, 1977.
- AKAIKE, H., A new look at the Bayes procedure, *Biometrika*, **65**, 53-59, 1978.
- AKI, K., Maximum likelihood estimate of b in the formula $\log N = a - bM$ and its confidence limits, *Bull. Earthq. Res. Inst.*, **43**, 237-239, 1965.
- BOX, G. E. P. and G. C. TIAO, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts, 1973.
- GUTENBERG, B. and C. F. RICHTER, Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.*, **34**, 185-188, 1944.
- JEFFEREYS, H., *Theory of Probability*, 3rd ed., Clarendon Press, Oxford, 1961.
- UTSU, T., A method for determining the value of b in a formula $\log n = a - bM$ showing the magnitude-frequency relation for earthquakes, *Geophys. Bull. Hokkaido Univ.*, **13**, 99-103, 1965 (in Japanese with English summary).
- UTSU, T., A statistical significance test of the difference in b -value between two earthquake groups, *J. Phys. Earth*, **14**, 37-40, 1966.

Note Added in Proof

Exact form of AIC can be derived from Eq. (14) for the estimates $b_{\xi} = \xi \log_{10} e / \sum_{i=1}^N (M_i - M_0)$; that is to say, $AIC(\xi) = (-2) \log L(b_{\xi}) + 2c$, where $c = \xi/(N-1)$. When we use the unbiased estimate for the comparison of b -values between two seismic activities (in time or space), we can adopt $AIC(N-1)$, where $c=1$ always hold. The $AIC(\xi)$ is derived by the similar way to AKAIKE (1973, 1977).