The Empirical Bayes Approach: Estimating the Prior Distribution

J. R. Rutherford; R. G. Krutchkoff

Biometrika, Vol. 54, No. 1/2. (Jun., 1967), pp. 326-328.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28196706%2954%3A1%2F2%3C326%3ATEBAET%3E2.0.CO%3B2-W

Biometrika is currently published by Biometrika Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/bio.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Wed Dec 19 01:27:55 2007

The empirical Bayes approach: estimating the prior distribution

By J. R. RUTHERFORD* and R. G. KRUTCHKOFF

Virginia Polytechnic Institute

Summary

There is a random variable Λ distributed according to a specific but unknown prior distribution G from an appropriate class G_p. The random variable $\Lambda = \lambda$ is unobservable but another random variable $X = x$, distributed with known conditional distribution function $F(x|\lambda)$, is observable. We construct estimators $G_n(\lambda)$ of $G(\lambda)$ such that $\lim E[\{G_n(\lambda) - G(\lambda)\}^2] = 0$ and we use $G_n(\lambda)$ to estimate the posterior distribution $G(\lambda|x)$ and hence to construct consistent estimators of posterior confidence intervals.

1. Introduction

We assume that we are able to observe the conditionally independent random variables $X_1, X_2, \ldots, X_n$, which are distributed according to the known, single parameter, conditional density function $f(x_i|\lambda_i)$. The 'parameters' are realizations of the unobservable random variables $\Lambda_1, \Lambda_2, \ldots, \Lambda_n$ which are independently distributed according to the unknown prior distribution $G(\lambda)$. The problem considered here is the estimation of $G(\lambda)$: the need for a solution to this problem was pointed out by Robbins (1964). The constructive method presented here requires that:

(*a*) The density function $f(x|\lambda)$ be such that there exist known functions $h_k(x)$ $(k = 1, 2, 3, 4)$, for which

$$E\{h_k(X)|\lambda\} = \lambda^k.$$

(*b*) The prior distribution be some unspecified Pearson curve, with certain minor restrictions given in the next section.

For any two numbers λ_* and λ^* let

$$P(\lambda_* \leqslant \Lambda \leqslant \lambda^*) = G(\lambda^*) - G(\lambda_*).$$

The method developed in this note provides estimates of $P(\lambda_* \leqslant \Lambda \leqslant \lambda^*)$, a 'modernized' Bayes confidence interval and estimates of $P(\lambda_* \leqslant \Lambda \leqslant \lambda^*|X = x)$, a 'classical' Bayes confidence interval (see Neyman, 1952, p. 161).

2. Estimating the prior distribution

From condition (*a*) we have

$$E\{h_k(X)|\lambda\} = \lambda^k \quad (k = 1, 2, 3, 4). \tag{1}$$

Taking expectations of both sides of equation (1) we obtain

$$E\{h_k(X)\} = E(\Lambda^k).$$

Let us define the functions

$$M_{k,n}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} h_k(x_i) \quad (k = 1, 2, 3, 4), \tag{2}$$

where $\mathbf{x}$ represents the sequence of realizations $x_1, x_2, \ldots, x_n$.

If $E(\Lambda^4) < \infty$, then by the Kolmogorov strong law of large numbers we have that almost surely

$$M_{k,n}(\mathbf{X}) \to E(\Lambda^k). \tag{3}$$

Let $\boldsymbol{\mu} = \{E(\Lambda), E(\Lambda^2), E(\Lambda^3), E(\Lambda^4)\}$ and $\mathbf{M}_n(\mathbf{X}) = \{M_{1,n}(\mathbf{X}), M_{2,n}(\mathbf{X}), M_{3,n}(\mathbf{X}), M_{4,n}(\mathbf{X})\}$.

From condition (*b*), $G(\lambda)$ is a Pearson curve and has a density function $g(\lambda)$; we denote the dependence of these functions on their moments by writing $G(\lambda; \boldsymbol{\mu})$ and $g(\lambda; \boldsymbol{\mu})$. The domain of $\boldsymbol{\mu}$ is defined in terms of semi-open and open regions in the (β_1, β_2)-plane; $\beta_1 = \mu_3^2/\mu_2^3$, $\beta_2 = \mu_4/\mu_2^2$ if $\mu_1 = 0$. The restrictions mentioned in (*b*) are that the moments of the prior distribution must be such that the associated values

* Now at Royal Military College of Canada.

of β_1 and β_2 are in the regions defining the domain of $\boldsymbol{\mu}$. The regions are between the lines $\beta_2 - \beta_1 - 1 = 0$ and $8\beta_2 - 15\beta_1 - 36 = 0$ exclusive of the points on the biquadratic curves

$$\beta_1(\beta_2+3)^2(8\beta_2-9\beta_1-12)-4(4\beta_2-3\beta_1)(5\beta_2-6\beta_1-9) = 0$$

and

$$\beta_1(\beta_2+3)^2(5\beta_2-6\beta_1-9)-(4\beta_2-3\beta_1)(7\beta_2-9\beta_1-15)^2 = 0.$$

These curves are called curves of discontinuity by Dumas (1948).

The estimator of $g(\lambda; \boldsymbol{\mu})$ will be $g\{\lambda; \mathbf{M}_n(\mathbf{X})\}$, where $g\{\lambda; \mathbf{M}_n(\mathbf{X})\}$ represents the solution of Pearson's differential equation with $\mathbf{M}_n(\mathbf{X})$ substituted for $\boldsymbol{\mu}$. For n sufficiently large, with probability one $g\{\lambda; \mathbf{M}_n(\mathbf{X})\}$ will be well defined. The Pearson curves are continuous functions of $\boldsymbol{\mu}$ for every λ, hence from equation (3) we obtain that almost surely

$$g(\lambda; \mathbf{M}_n(\mathbf{X})) \to g(\lambda; \boldsymbol{\mu}), \quad \text{for every } \lambda. \tag{4}$$

From a result of Scheffé (1947) we obtain finally that almost surely

$$G(\lambda; \mathbf{M}_n(\mathbf{X})) \to G(\lambda; \boldsymbol{\mu}), \quad \text{uniformly in } \lambda. \tag{5}$$

Let λ_* and λ^* be two numbers. The posterior probability of an interval (λ_*, λ^*) is defined to be

$$P(\lambda_* \leqslant \Lambda \leqslant \lambda^* | X = x) = \int_{\lambda_*}^{\lambda^*} f(x|\lambda)\, dG(\lambda) \Big/ \int_{-\infty}^{\infty} f(x|\lambda)\, dG(\lambda).$$

Let $G_n(\lambda) = G\{\lambda; \mathbf{M}_n(\mathbf{X})\}$ and define

$$P_n(\lambda_* \leqslant \Lambda \leqslant \lambda^* | X = x) = \int_{\lambda_*}^{\lambda^*} f(x|\lambda)\, dG_n(\lambda) \Big/ \int_{-\infty}^{\infty} f(x|\lambda)\, dG_n(\lambda). \tag{6}$$

If $f(x|\lambda)$ is a continuous function of λ for every x then by the Helly–Bray lemma we obtain that almost surely

$$P_n(\lambda_* \leqslant \Lambda \leqslant \lambda^* | X = x) \to P(\lambda_* \leqslant \Lambda \leqslant \lambda^* | X = x),$$

and we have the required estimate.

3. Example

The data for the example are taken from Mosteller & Wallace (1963). A collection of a man's writings was broken up into 247 blocks of 200 words and the observed frequency of the word *may* was recorded; see row 1 and row 2 of Table 1.

We assumed that these observations were distributed according to a Poisson density with mean λ and that λ was distributed according to an unknown Pearson distribution. For the Poisson density the functions $h_k(x)$ are: $h_1(x) = x$, $h_2(x) = x(x-1)$, $h_3(x) = x(x-1)(x-2)$ and $h_4(x) = x(x-1)(x-2)(x-3)$.

The central moments, $\sqrt{\beta_1}$ and β_2 were found to be $\mu_1 = 0{\cdot}8097$, $\mu_2 = 0{\cdot}5834$, $\sqrt{\beta_1} = 0{\cdot}50$ and $\beta_2 = 2{\cdot}069$. Using tables of Pearson's curves provided by Johnson *et al.* (1963) we drew a graph of $G_n(\lambda)$, the estimate of $G(\lambda)$. We then evaluated numerically the fitted distribution $P_n(x)$, where

$$\begin{aligned} P_n(x) &= \int_{l_1}^{l_2} p(x|\lambda)\, dG_n(\lambda) \\ &= G_n(l_2)\, p(x|l_2) - G_n(l_1)\, p(x|l_1) + \int_{l_1}^{l_2} G_n(\lambda)\{p(x|\lambda) - p(x-1|\lambda)\}\, d\lambda. \end{aligned}$$

Here l_1 and l_2 are the estimated lower and upper limits of the prior distribution and $p(x|\lambda) = e^{-\lambda}\lambda^x/x!$. In the last row of Table 1 this fit is seen to be about as good as the negative binomial fit. The closeness

Table 1. *Observed, fitted Poisson, negative binomial and empirical Bayes distributions for the word* may

Occurrence	0	1	2	3	4	5	6	7
Observed	128	67	32	14	4	1	1	—
Poisson	109·9	88·9	36·0	9·7	2·0	0·3	0·1	—
Negative Binomial	128·2	69·4	30·1	12·1	4·6	1·7	0·6	0·3
Empirical Bayes	127·3	65·2	34·9	13·1	4·7	1·5	0·4	0·0

of fit is not affected very much by different assumptions about the prior distribution. We see this because a negative binomial random variable can be generated by a Poisson variable with a mean of λ which is the realization of a type III, or gamma, variable whereas the estimates of $\sqrt{\beta_1}$ and β_2 for the prior distribution indicate that the prior distribution is an L-shaped type I curve.

We also evaluated by numerical integration the posterior probability of the interval (0·1, 1·9), that is

$$P_n(0{\cdot}1 \leqslant \Lambda \leqslant 1{\cdot}9 | X = x) = G_n(1{\cdot}9)\frac{p(x|1{\cdot}9)}{P_n(x)} - G_n(0{\cdot}1)\frac{p(x|0{\cdot}1)}{P_n(x)} + \int_{0\cdot 1}^{1\cdot 9} G_n(\lambda)\frac{p(x|\lambda)-p(x-1|\lambda)}{P_n(x)}\,d\lambda.$$

Table 2. *Posterior probability of the interval* (0·1, 1·9)

x	0	1	2	3	4	5	6
$P_n\,(0{\cdot}1 \leqslant \Lambda \leqslant 1{\cdot}9 \| X = x)$	0·9424	0·9477	0·9043	0·8547	0·7906	0·7203	0·5625

The prior probability of the interval (0·1, 1·9) was estimated to be 0·93 from the graph of $G_n(\lambda)$.

4. Discussion

Other one-parameter density functions satisfying condition (a) are the binomial with unknown proportion λ; the gamma with unknown scale λ; the uniform with unknown range λ; and the normal with either mean or variance unknown. A density function not satisfying condition (a) is

$$f(x|\theta) = \theta\, e^{-x\theta} \quad (x, \theta > 0).$$

We emphasize that it is necessary to assume only that the prior distribution is a member of the Pearson family of curves: the continuity of the family with respect to $\boldsymbol{\mu}$ ensures this. The essential feature of condition (b) is that the prior distribution functions are continuous in the estimable moments and that the moments are finite. The Pearson family was chosen because of its size and the availability of tables.

The motivation for the technique developed here was introduced by von Mises (1942). For a discussion of some of the practical problems associated with estimating a distribution by moments see Pearson (1963).

This work was supported in part by the U.S. Office of Army Research.

References

Dumas, M. (1948). Sur les courbes de frequence de K. Pearson. *Biometrika* **35**, 113–7.

Johnson, N. L., Nixon, E., Amos, D. E. & Pearson, E. S. (1963). Tables of percentage points of Pearson curves for β_1 and β_2 expressed in standard measure. *Biometrika* **50**, 459–98.

Mosteller, F. & Wallace, D. L. (1963). Inference in an authorship problem. *J. Am. Statist. Ass.* **58**, 275–309.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*, pp. 162–5. U.S.D.A. Graduate School.

Pearson, E. S. (1963). Some problems arising in approximating to probability distributions, using moments. *Biometrika* **50**, 95–112.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**, 1–20.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18**, 434–8.

von Mises, R. (1942). On the correct use of Bayes formula. *Ann. Math. Statist.* **13**, 156–65.

[*Received February* 1966. *Revised October* 1966]