AN EMPIRICAL BAYES APPROACH TO STATISTICS

HERBERT ROBBINS
COLUMBIA UNIVERSITY

Let X be a random variable which for simplicity we shall assume to have discrete values x and which has a probability distribution depending in a known way on an unknown real parameter Λ,

$$p(x \mid \lambda) = Pr[X = x \mid \Lambda = \lambda], \tag{1}$$

Λ *itself being a random variable* with *a priori* distribution function

$$G(\lambda) = Pr[\Lambda \leqq \lambda]. \tag{2}$$

The unconditional probability distribution of X is then given by

$$p_G(x) = Pr[X = x] = \int p(x \mid \lambda)\, dG(\lambda), \tag{3}$$

and the expected squared deviation of any estimator of Λ of the form $\varphi(X)$ is

$$\begin{aligned} E[\varphi(X) - \Lambda]^2 &= E\{E[(\varphi(X) - \Lambda)^2 \mid \Lambda = \lambda]\} \\ &= \int \sum_x p(x \mid \lambda)[\varphi(x) - \lambda]^2 dG(\lambda) \\ &= \sum_x \int p(x \mid \lambda)[\varphi(x) - \lambda]^2 dG(\lambda), \end{aligned} \tag{4}$$

which is a minimum when $\varphi(x)$ is defined for each x as that value $y = y(x)$ for which

$$I(x) = \int p(x \mid \lambda)(y - \lambda)^2 dG(\lambda) = \text{minimum}. \tag{5}$$

But for any fixed x the quantity

$$\begin{aligned} I(x) &= y^2 \int p\,dG - 2y \int p\lambda\, dG + \int p\lambda^2 dG \\ &= \int p\,dG \left(y - \frac{\int p\lambda\, dG}{\int p\,dG}\right)^2 + \left[\int p\lambda^2 dG - \frac{(\int p\lambda\, dG)^2}{\int p\,dG}\right] \end{aligned} \tag{6}$$

is a minimum with respect to y when

$$y = \frac{\int p\lambda\, dG}{\int p\,dG}, \tag{7}$$

the minimum value of $I(x)$ being

$$I_G(x) = \int p(x \mid \lambda)\,\lambda^2 dG(\lambda) - \frac{[\int p(x \mid \lambda)\,\lambda\, dG(\lambda)]^2}{\int p(x \mid \lambda)\, dG(\lambda)}. \tag{8}$$

Research supported by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command, and by the Office of Ordnance Research, U.S. Army, under Contract DA-04-200-ORD-355.

Hence the *Bayes estimator* of Λ corresponding to the *a priori* distribution function G of Λ [in the sense of minimizing the expression (4)] is the random variable $\varphi_G(X)$ defined by the function

$$\varphi_G(x) = \frac{\int p(x \mid \lambda)\, \lambda dG(\lambda)}{\int p(x \mid \lambda)\, dG(\lambda)}, \tag{9}$$

the corresponding minimum value of (4) being

$$E[\varphi_G(X) - \Lambda]^2 = \sum_x I_G(x). \tag{10}$$

The expression (9) is, of course, the expected value of the *a posteriori* distribution of Λ given $X = x$.

If the *a priori* distribution function G is known to the experimenter then φ_G defined by (9) is a computable function, but if G is unknown, as is usually the case, then φ_G is not computable. This trouble is not eliminated by the adoption of arbitrary rules prescribing forms for G (as is done, for example, by H. Jeffreys [1] in his theory of statistical inference). It is partly for this reason—that even when G may be assumed to exist it is generally unknown to the experimenter—that various other criteria for estimators (unbiasedness, minimax, etc.) have been proposed which have the advantage of not requiring a knowledge of G.

Suppose now that the problem of estimating Λ from an observed value of X is going to occur repeatedly with a fixed and known $p(x \mid \lambda)$ and a fixed but unknown $G(\lambda)$, and let

$$(\Lambda_1, X_1), (\Lambda_2, X_2), \cdots, (\Lambda_n, X_n), \cdots \tag{11}$$

denote the sequence so generated. [The Λ_n are independent random variables with common distribution function G, and the distribution of X_n depends only on Λ_n and for $\Lambda_n = \lambda$ is given by $p(x \mid \lambda)$.] If we want to estimate an unknown Λ_n from an observed X_n and if the previous values $\Lambda_1, \cdots, \Lambda_{n-1}$ are by now known, then we can form the empirical distribution function of the random variable Λ,

$$G_{n-1}(\lambda) = \frac{\text{number of terms } \Lambda_1, \cdots, \Lambda_{n-1} \text{ which are} \leqq \lambda}{n-1}, \tag{12}$$

and take as our estimate of Λ_n the quantity $\psi_n(X_n)$, where by definition

$$\psi_n(x) = \frac{\int p(x \mid \lambda)\, \lambda dG_{n-1}(\lambda)}{\int p(x \mid \lambda)\, dG_{n-1}(\lambda)}, \tag{13}$$

which is obtained from (9) by replacing the unknown *a priori* $G(\lambda)$ by the empirical $G_{n-1}(\lambda)$. Since $G_{n-1}(\lambda) \rightarrow G(\lambda)$ with probability 1 as $n \rightarrow \infty$, the ratio (13) will, under suitable regularity conditions on the kernel $p(x \mid \lambda)$, tend for any fixed x to the Bayes function $\varphi_G(x)$ defined by (9) and hence, again under suitable conditions, the expected squared deviation of $\psi_n(X_n)$ from Λ_n will tend to the Bayes value (10).

In practice, of course, it will be unusual for the previous values $\Lambda_1, \cdots, \Lambda_{n-1}$ to be known, and hence the function (13) will be no more computable than the true Bayes function (9). *However, in many cases the previous values $X_1, \cdots, X_{n-1}$ will be available to the experimenter at the moment when Λ_n is to be estimated,* and the question then arises whether it is possible to infer from the set of values $X_1, \cdots, X_n$ the approximate form of the unknown G, or at least, in the present case of quadratic estimation, to approximate the value of the functional of G defined by (9). To this end we observe that for any fixed x the empirical frequency

$$(14) \qquad p_n(x) = \frac{\text{number of terms } X_1, \cdots, X_n \text{ which equal } x}{n}$$

tends with probability 1 as $n \rightarrow \infty$ to the function $p_G(x)$ defined by (3), no matter what the *a priori* distribution function G. Hence there arises the following mathematical problem: from an approximate value (14) of the integral (3), where $p(x \mid \lambda)$ is a known kernel, to obtain an approximation to the unknown distribution function G, or at least, in the present case, to the value of the Bayes function (9) which depends on G. (This problem was posed in [4].) The possibility of doing this will depend on the nature of the kernel $p(x \mid \lambda)$ and on the class, say $\mathcal{G}$, to which the unknown G is assumed to belong. In order to fix the ideas we shall consider several special cases, the first being that of the *Poisson* kernel

$$(15) \qquad p(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}; \qquad x = 0, 1, \cdots; \lambda > 0;$$

$\mathcal{G}$ being the class of all distribution functions on the positive real axis.

In this case we have

$$(16) \qquad p_G(x) = \int p(x \mid \lambda)\, dG(\lambda) = \int_0^\infty e^{-\lambda} \lambda^x dG(\lambda) / x!$$

and

$$(17) \qquad \varphi_G(x) = \frac{\int_0^\infty e^{-\lambda} \lambda^{x+1} dG(\lambda)}{\int_0^\infty e^{-\lambda} \lambda^x dG(\lambda)},$$

and we can write the fundamental relation

$$(18) \qquad \varphi_G(x) = (x+1) \cdot \frac{p_G(x+1)}{p_G(x)}.$$

If we now define the function

$$(19) \quad \varphi_n(x) = (x+1) \frac{p_n(x+1)}{p_n(x)} = (x+1) \cdot \frac{\text{number of terms } X_1, \cdots, X_n \text{ which are equal to } x+1}{\text{number of terms } X_1, \cdots, X_n \text{ which are equal to } x}$$

then no matter what the unknown G we shall have for any fixed x

$$(20) \qquad \varphi_n(x) \rightarrow \varphi_G(x) \text{ with probability 1 as } n \rightarrow \infty.$$

This suggests using as an estimate of the unknown Λ_n in the sequence (11) the computable quantity

$$(21) \qquad \varphi_n(X_n),$$

in the hope that as $n \rightarrow \infty$,

$$(22) \qquad E[\varphi_n(X_n) - \Lambda_n]^2 \rightarrow E[\varphi_G(X) - \Lambda]^2.$$

We shall not investigate here the question of whether (22) does actually hold for the particular function (19) or whether (19) represents the best possible choice for minimizing in some sense the expected squared deviation. (See [8].)

It is of interest to compute the value of (10) for various *a priori* distribution functions G in order to compare its value with the expected squared deviation of the usual (maximum likelihood, minimum variance unbiased) Poisson estimator, X itself, for which

$$E(X-\Lambda)^2=E\Lambda=\int_0^\infty \lambda dG(\lambda)\,. \tag{23}$$

Suppose, for example, that G is a gamma type distribution function with density

$$G'(\lambda)\;=C\lambda^{b-1}e^{-h\lambda}\,;\qquad \lambda,\ b,\ h>0;\ C=h^b/\Gamma(b)\,. \tag{24}$$

By elementary computation we find that

$$E\Lambda=\frac{b}{h},\qquad \operatorname{Var}\Lambda=\frac{b}{h^2} \tag{25}$$

and

$$\varphi_G(x)\;=\frac{x+b}{1+h},\qquad E\,[\varphi_G(X)\;-\Lambda]^2=\frac{b}{h(1+h)}\,; \tag{26}$$

hence

$$\frac{E\,[\varphi_G(X)\;-\Lambda]^2}{E(X-\Lambda)^2}=\frac{1}{1+h}\,. \tag{27}$$

For example, if b = 100, h = 10 then

$$E\Lambda=10,\ \operatorname{Var}\Lambda=1,\ \varphi_G(x)\;=\frac{x+100}{11},\ \frac{E\,[\varphi_G(X)\;-\Lambda]^2}{E(X-\Lambda)^2}=\frac{1}{11}\,. \tag{28}$$

An even simpler case occurs when Λ has all its probability concentrated at a single value λ. In this case, of course, the Bayes function is

$$\varphi_G(x)\;=\lambda\;, \tag{29}$$

not involving x at all, and

$$E\,[\varphi_G(X)\;-\Lambda]^2=0\;, \tag{30}$$

while as before

$$E(X-\Lambda)^2=E\Lambda=\lambda\;. \tag{31}$$

Here the sequence (11) consists of observations $X_1,\cdots,X_n,\cdots$ from the same Poisson population (although this fact may not be apparent to the experimenter at the beginning); the traditional estimator $\varphi(x) = x$ does not take advantage of this favorable circumstance and continues to have the expected squared deviation λ after any number n of trials.

As a second example we take the *geometric* kernel

$$p(x\,|\,\lambda)=(1-\lambda)\,\lambda^x\,;\qquad x=0,\ 1,\cdots;\ 0<\lambda<1\;; \tag{32}$$

for which

$$p_G(x)\;=\int_0^1 (1-\lambda)\,\lambda^x dG(\lambda)\;,\ \varphi_G(x) \tag{33}$$

$$\varphi_G\,'x)=\frac{\int_0^1 (1-\lambda)\,\lambda^{x+1}dG(\lambda)}{\int_0^1 (1-\lambda)\,\lambda^x dG(\lambda)}=\frac{p_G(x+1)}{p_G(x)}\,.$$

Here it is natural to estimate Λ_n by (21) with the definition

$$\varphi_n(x) = \frac{\text{number of terms } X_1, \cdots, X_n \text{ which are equal to } x+1}{\text{number of terms } X_1, \cdots, X_n \text{ which are equal to } x}. \tag{34}$$

Our third example will be the *binomial* kernel

$$p_r(x \mid \lambda) = \binom{r}{x} \lambda^x (1-\lambda)^{r-x}; \qquad x = 0, 1, \cdots, r;\ 0 \leqq \lambda \leqq 1. \tag{35}$$

Here r is a fixed positive integer representing the number of trials, X the number of successes, and Λ the unknown probability of success in each trial. G may be taken as the class of all distribution functions on the interval (0, 1). In this case

$$\begin{cases} p_{G,\,r}(x) = \int p_r(x \mid \lambda)\, dG(\lambda) = \binom{r}{x} \int_0^1 \lambda^x (1-\lambda)^{r-x} dG(\lambda), \\ \varphi_{G,\,r}(x) = \dfrac{\int_0^1 \lambda^{x+1} (1-\lambda)^{r-x} dG(\lambda)}{\int_0^1 \lambda^x (1-\lambda)^{r-x} dG(\lambda)}, \end{cases} \tag{36}$$

so that we can write the fundamental relation

$$\varphi_{G,\,r}(x) = \frac{x+1}{r+1} \cdot \frac{p_{G,\,r+1}(x+1)}{p_{G,\,r}(x)}; \qquad x = 0, 1, \cdots, r. \tag{37}$$

Let

$$p_{n,\,r}(x) = \frac{\text{number of terms } X_1, \cdots, X_n \text{ which are equal to } x}{n}; \tag{38}$$

then $p_{n,\,r}(x) \rightarrow p_{G,\,r}(x)$ with probability 1 as $n \rightarrow \infty$. Now consider the sequence of random variables

$$X_1', X_2', \cdots, X_n', \cdots \tag{39}$$

where X_n' denotes the number of successes in, say, the first $r-1$ out of the r trials which produced X_n successes, and let

$$p_{n,\,r-1}(x) = \frac{\text{number of terms } X_1', \cdots, X_n' \text{ which are equal to } x}{n}; \tag{40}$$

then $p_{n,\,r-1}(x) \rightarrow p_{G,\,r-1}(x)$ with probability 1 as $n \rightarrow \infty$. Thus if we set

$$\varphi_{n,\,r}(x) = \frac{x+1}{r} \cdot \frac{p_{n,\,r}(x+1)}{p_{n\ r-1}(x)}, \tag{41}$$

then

$$\varphi_{n,\,r}(x) \rightarrow \frac{x+1}{r} \cdot \frac{p_{G,\,r}(x+1)}{p_{G,\,r-1}(x)} = \varphi_{G,\,r-1}(x) \tag{42}$$

with probability 1 as $n \rightarrow \infty$. If we take as our estimate of Λ_n the value

$$\varphi_{n,\,r}(X_n') \tag{43}$$

then for large n we will do about as well as if we knew the *a priori* G but confined ourselves to the first $r-1$ out of each set of r trials. For large r this does not sacrifice much information, but it is by no means clear that (43) is the "best" estimator of Λ_n that could be devised in the spirit of our discussion.

As a final example consider any kernel of the "Laplacian" type

$$p(x \mid \lambda) = e^{\lambda x} f(x)\, h(\lambda)\,. \tag{44}$$

We have

$$\begin{aligned} p_G(x) &= f(x) \int e^{\lambda x} h(\lambda)\, dG(\lambda)\,, \\ \varphi_G(x)\, p_G(x) &= f(x) \int \lambda e^{\lambda x} h(\lambda)\, dG(\lambda) = f(x) \frac{d}{dx}\left\{\frac{p_G(x)}{f(x)}\right\}, \end{aligned} \tag{45}$$

provided the differentiation under the integral sign is justified. Hence

$$\varphi_G(x) = \frac{d}{dx} \log\left\{\frac{p_G(x)}{f(x)}\right\}. \tag{46}$$

Perhaps a satisfactory approximation to $\varphi_G(x)$ might be obtained by replacing $p_G(x)$ in (46) by a smoothed interpolation based on $p_n(x)$. The kernel (44) has been considered by M. C. K. Tweedie [6] and I am indebted to him for this example.

Until now we have been concerned only with approximating to the Bayes function $\varphi_G(x)$ defined by (9). In many cases we shall want an approximation to some other functional of the unknown *a priori* distribution function G; in particular to G itself. We shall make a few remarks about this problem in the general case in which X is not restricted to discrete values but has a distribution function

$$F(x \mid \lambda) = Pr\,[X \leqq x \mid \Lambda = \lambda] \tag{47}$$

depending on the random variable Λ whose distribution function G is unknown. The unconditional distribution function of X is then

$$F_G(x) = Pr\,[X \leqq x] = \int F(x \mid \lambda)\, dG(\lambda)\,, \tag{48}$$

and there is assumed to be available an infinite sequence $X_1, X_2, \cdots$ of independent random variables with the common distribution function F_G. The empirical distribution function

$$F_n(x) = \frac{\text{number of terms } X_1, \cdots, X_n \text{ which are} \leqq x}{n} \tag{49}$$

is known to converge uniformly to $F_G(x)$ with probability 1 as $n \to \infty$.

Problem: to find in terms of $F_n(x)$ a distribution function $G_n(\lambda)$ which will converge as $n \to \infty$ to the unknown $G(\lambda)$.

Let $\mathcal{G}$ denote some class of distribution functions to which the unknown G is assumed to belong. ($\mathcal{G}$ might, for example, be the class of all distribution functions, or all those with total mass distributed on some fixed finite interval.) The correspondence

$$F_G(x) = \int F(x \mid \lambda)\, dG(\lambda) \tag{50}$$

maps $\mathcal{G}$ onto some class of distribution functions which we shall denote by $\mathscr{G}$. We shall assume that the kernel $F(x \mid \lambda)$ is such that this mapping is one-to-one. Now, since we know an approximation F_n to F_G, it would be natural to seek an approximation to G by solving the functional equation (50) for G with F_G replaced by F_n. Unfortunately, in general this will be impossible since F_n will not belong to the class $\mathscr{G}$. [For example, if $F(x \mid \lambda)$ is continuous in x then all elements of $\mathscr{G}$ will be continuous, whereas F_n is a step function.] However, we may proceed as follows. Let F_n^* be any element of $\mathscr{G}$ whose distance (in the sense of maximum absolute value of the difference for all x) from F_n is within $\epsilon_n \to 0$ of the minimum distance of F_n from $\mathcal{G}$ (this is the "minimum distance" method of Wolfowitz), and let G_n be defined by the relation

$$F_n^* (x) = \int F (x \mid \lambda) \, dG_n (\lambda) . \tag{51}$$

Then $F_n^* \to F_G$ in the maximum difference metric, and under suitable conditions on the kernel $F(x \mid \lambda)$ it will follow that $G_n \to G$. We shall go into this question in more detail elsewhere, but at least it indicates one possible way of obtaining an empirical approximation to a "mixing" distribution G from observations on the "mixed" distribution F_G. (See [5] also.) This problem, special cases of which have occurred several times in the statistical literature (see, for example, [2], [4], [7] and pp. 84–102 in [3]), awaits a satisfactory solution and seems to be of considerable importance.

I should like to express my appreciation to A. Dvoretzky, J. Neyman, and H. Raiffa for helpful discussions and suggestions.

REFERENCES

[1] H. JEFFREYS, *Theory of Probability*, 2d ed., Oxford, Clarendon Press, 1948.

[2] P. F. LAZARSFELD, "A conceptual introduction to latent structure analysis," *Mathematical Thinking in the Social Sciences*, Glencoe, Ill., Free Press, 1954, pp. 349–387.

[3] J. NEYMAN, *Lectures and Conferences on Mathematical Statistics and Probability*, 2d ed., Washington, U.S. Department of Agriculture Graduate School, 1952.

[4] H. ROBBINS, "Asymptotically subminimax solutions of compound statistical decision problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1951, pp. 131–148.

[5] ———, "A generalization of the method of maximum likelihood: estimating a mixing distribution," (abstract), *Annals of Math. Stat.*, Vol. 21 (1950), pp. 314–315.

[6] M. C. K. TWEEDIE, "Functions of a statistical variate with given means, with special reference to Laplacian distributions," *Proc. Camb. Phil. Soc.*, Vol. 43 (1947), pp. 41–49.

[7] R. VON MISES, "On the current use of Bayes' formula," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 156–165.

[8] M. V. JOHNS, JR., "Contributions to the theory of empirical Bayes procedures in statistics," doctoral thesis at Columbia University (unpublished), 1956.