MASTER

BNL 30554
AMD 885

CONF-8106199--1

BNL--30554

DE82 006936

Estimating Many Variances*

by Herbert Robbins, Columbia University and
Brookhaven National Laboratory

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

* The submitted manuscript has been authored under contract
DE-AC02-76CH00016 with the U.S. Department of Energy.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Suppose that an unknown random parameter θ with distribution function G is such that given θ , an observable random variable x has conditional probability density $f(x|\theta)$ of known form. If a function $t = t(x)$ is used to estimate θ , then the expected squared error with respect to the random variation of both θ and x is

$$(1) \quad E(t-\theta)^2 = \int \int (t(x)-\theta)^2 f(x|\theta) dx dG(\theta).$$

For fixed G we can seek to minimize (1) within any desired class of functions t , such as the class of all linear functions $A + Bx$, or the class of all Borel functions whatsoever. The minimum of (1) within the class of linear functions is achieved for the function $\tilde{t}$ defined by

$$(2) \quad \tilde{t}(x) = E\theta + \frac{\text{Cov}(\theta, x)}{\text{Var } x} \cdot (x - Ex),$$

with

$$(3) \quad E(\tilde{t}-\theta)^2 = \text{Var } \theta - \frac{\text{Cov}^2(\theta, x)}{\text{Var } x},$$

while the unrestricted minimum of (1) is achieved for the function t^* defined by

$$(4) \quad t^*(x) = E(\theta|x) = \int \theta f(x|\theta) dG(\theta) / \int f(x|\theta) dG(\theta),$$

with

$$(5) \quad E(t^* - \theta)^2 = E \text{Var}(\theta|x).$$

When G is unknown to the statistician, the empirical Bayes approach is to approximate (2) or (4) by observing the x -components of a large number of independent pairs (θ_i, x_i) ($i=1, \dots, n$) with the same joint distribution as the pair (θ, x) of the preceding paragraph. It is not obvious how this approximation is to be effected in practice, nor whether one should try to approximate (2), (4), or perhaps something else, for a given n .

We begin by trying to approximate (2), and assume for simplicity that $f(x|\theta)$ is such that for known constants a, b, c

$$(6) \quad E(x|\theta) = \theta, \quad \text{Var}(x|\theta) = a + b\theta + c\theta^2.$$

From the first of these relations it follows that

$$(7) \quad E x = E \theta, \quad \text{Cov}(\theta, x) = \text{Var } \theta, \quad E(x - \theta)^2 = \text{Var } x - \text{Var } \theta = E \text{Var}(x|\theta),$$

and when the second is added that

$$(8) \quad \text{Var } \theta = [\text{Var } x - a - b E x - c E^2 x] / (1 + c).$$

We can therefore write (2) in the form

$$(9) \quad \tilde{t}(x) = Ex + \left[1 - \frac{c(\text{Var } x + E^2 x) + a + b Ex}{(1 + c) \text{Var } x} \right] (x - Ex),$$

with

$$(10) \quad E(\tilde{t} - \theta)^2 = \frac{\text{Var } \theta [\text{Var } x - \text{Var } \theta]}{\text{Var } x} \\ = \frac{\text{Var } x - a - b Ex - cE^2 x}{(1 + c) \text{Var } x} E(x - \theta)^2.$$

The usual estimators of Ex and $\text{Var } x$ are

$$(11) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

and substituting them (or statistics asymptotically equivalent to them as $n \rightarrow \infty$) into (9) we obtain the l.e.B. (linear empirical Bayes) function $\tilde{t}_n$ defined by

$$(12) \quad \tilde{t}_n(x) = \bar{x} + \left[1 - \frac{c(s^2 + \bar{x}^2) + a + b\bar{x}}{(1 + c) s^2} \right] (x - \bar{x})$$

as a consistent approximation to $\tilde{t}$. We then estimate θ_i by $\tilde{t}_n(x_i)$ ($i=1, \dots, n$), and hope that, under mild assumptions on the nature of G ,

$$(13) \quad E(\tilde{t}_n(x_i) - \theta_i)^2 = \frac{1}{n} \sum_{i=1}^n E(\tilde{t}_n(x_i) - \theta_i)^2 \rightarrow E(\tilde{t} - \theta)^2 \\ \sim \frac{s^2 - a - b\bar{x} - c\bar{x}^2}{(1 + c)s^2} E(x - \theta)^2 \text{ as } n \rightarrow \infty.$$

The usual example is that in which for some known $\sigma > 0$

$$(14) \quad f(x|\theta) = \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right),$$

where $\phi(x)$ is the standard normal p.d.f. (or more generally, any p.d.f. with mean 0 and variance 1), so that

$$(15) \quad E(x|\theta) = \theta, \text{Var}(x|\theta) = \sigma^2,$$

and (12) and (10) become

$$(16) \quad \tilde{t}_n(x) = \bar{x} + \left[1 - \frac{\sigma^2}{s^2}\right]^+ (x - \bar{x}), \quad \frac{E(x-\theta)^2}{E(\tilde{t}-\theta)^2} = \frac{\text{Var } x}{\text{Var } x - \sigma^2} \sim \frac{s^2}{s^2 - \sigma^2}.$$

Consider now the less familiar problem of estimating many variances (instead of means), in samples of the same size from normal populations. To define the basic pair (θ, x) , for some $r = k/2$ ($k = 1, 2, \dots$) let $y_1, \dots, y_{2r+1}$ be i.i.d. normal with mean μ and variance θ , and let

$$(17) \quad x = \frac{1}{2r} \sum_{i=1}^{2r+1} (y_i - \bar{y})^2 = \text{sample variance.}$$

Then $2rx/\theta$ has the chi-squared distribution with $2r$ degrees of freedom, and the conditional p.d.f. of x given θ is of the gamma family with known shape parameter r and unknown scale parameter θ ,

$$(18) \quad f(x|\theta) = \frac{r^r}{\Gamma(r)} \theta^{-r} e^{-\frac{rx}{\theta}} x^{r-1} \quad (r, x, \theta > 0),$$

for which

$$(19) \quad E(x|\theta) = \theta, \text{ Var } (x|\theta) = \theta^2/r,$$

so that

$$(20) \quad \begin{aligned} \tilde{t}_n(x) &= \bar{x} + \left[1 - \frac{s^2 + \bar{x}^2}{(1+r)s^2}\right]^+ (x - \bar{x}), \\ \frac{E(x-\theta)^2}{E(\tilde{t}-\theta)^2} &= 1 + \frac{1}{r} \left(1 + \frac{E^2\theta}{\text{Var } \theta}\right) \sim \frac{(r+1)s^2}{rs^2 - \bar{x}^2}. \end{aligned}$$

Thus, for n independent samples of size $2r + 1$ from normal populations with unknown means μ_i and unknown variances θ_i , if x_i denotes the i th sample variance then a l.e.B. estimator of θ_i is

$$(21) \quad \tilde{t}_n(x_i) = \bar{x} + \left[1 - \frac{s^2 + \bar{x}^2}{(1+r)s^2}\right]^+ (x_i - \bar{x}) \quad (i=1, \dots, n).$$

For large n this will always be better, and in some cases much better, than the usual estimator x_i . For example, if $P(\theta=99) = P(\theta=101) = 1/2$, so that $E\theta = 100$, $\text{Var } \theta=1$, then

$$(22) \quad \frac{E(x-\theta)^2}{E(\tilde{t}-\theta)^2} = 1 + \frac{1}{r} (10,001),$$

which is $> 1,000$ for $r < 10$; i.e., sample size $= 2r+1 < 21$.

We now consider, for the same gamma family (18), the problem of approximating the unrestricted or general empirical Bayes (g.e.B.) estimator (4), of which the denominator is

$$(23) \quad f(x) = \int f(x|\theta) dG(\theta) = \text{marginal p.d.f. of } x.$$

Now, $f(x)$ can be estimated from $x_1, \dots, x_n$ by various techniques of density estimation. The numerator of (4), however, bears no immediate relation to $f(x)$, and it is not obvious how to approximate it from the observed $x_1, \dots, x_n$. Happily, it can be shown for (18), by interchanging the order of integration, that for all $x > 0$ the identity

$$(24) \quad t^*(x) = E(\theta|x) = rx^{r-1} \int_x^\infty y^{1-r} f(y) dy / f(x)$$

holds, and this gives t^* in terms of f without explicit reference to G , which is what we want. The numerator of $t^*(a)$ for any fixed $a > 0$ is the expectation of the random variable

$$(25) \quad \psi(x) = \begin{cases} r a^{r-1} x^{1-r} & \text{if } x > a \\ 0 & \text{if } x < a, \end{cases}$$

and can be estimated by $\sum_{i=1}^n \psi(x_i) / n$.

For the gamma family (18), the l.e.B. and g.e.B. estimators (2) and (4) are equal when G is such that

$$(26) \quad G'(\theta) = g(\theta) = \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\frac{\alpha}{\theta}} \theta^{-\beta-1}$$

for some $\alpha, \beta > 0$ with $\beta + r > 1$; their common value is the linear function

$$(27) \quad \frac{\alpha + rx}{\beta + r - 1}.$$

Density estimation can be avoided by changing the problem from that of the simultaneous estimation of many unknown variances θ_i to that of estimating something else. We mention three closely related versions of a different problem.

I. For any fixed $a > 0$ define

$$\begin{aligned}
 (28) \quad H = H(a) = E(\theta | x > a) &= \frac{\int_a^\infty E(\theta | x) f(x) dx}{P(x > a)} = \frac{\int_a^\infty r x^{r-1} \int_x^\infty y^{1-r} f(y) dy dx}{P(x > a)} \\
 &= \frac{\int_a^\infty y^{1-r} f(y) \int_a^y r x^{r-1} dx dy}{P(x > a)} = \frac{\int_a^\infty x [1 - (\frac{a}{x})^r] f(x) dx}{P(x > a)} = \frac{Ev}{Eu},
 \end{aligned}$$

where by definition

$$(29) \quad u = \begin{cases} 1 & \text{if } x > a \\ 0 & \text{if } x < a \end{cases}, \quad v = ux [1 - (\frac{a}{x})^r].$$

We can estimate H consistently as $n \rightarrow \infty$ by the ratio T_n/N_n , where

$$(30) \quad N_n = \sum_{i=1}^n u_i, \quad T_n = \sum_{i=1}^n v_i = \sum_{i=1}^n u_i x_i [1 - (\frac{a}{x_i})^r].$$

In fact, by the central limit theorem, as $n \rightarrow \infty$

$$(31) \quad \sqrt{n} \left(\frac{T_n}{N_n} - H \right) = \frac{\frac{1}{n} \sum_{i=1}^n (v_i - Hu_i)}{(\sum_{i=1}^n u_i / n) \sqrt{n}} \xrightarrow{D} N(0, \sigma_1^2),$$

where

$$(32) \quad \sigma_1^2 = \frac{E(v - Hu)^2}{E^2 u} \sim \frac{n}{N_n} \left\{ \frac{1}{N_n} \sum_{i=1}^n v_i^2 - \left(\frac{T_n}{N_n} \right)^2 \right\},$$

so that

$$(33) \quad \frac{\frac{T_n}{N_n} - H}{\frac{1}{\sqrt{N_n}} \sqrt{\frac{1}{N_n} \sum_{i=1}^n v_i^2 - \left(\frac{T_n}{N_n} \right)^2}} \xrightarrow{D} N(0,1).$$

Thus, for large n , an approximately 95% confidence interval for H is given by

$$(34) \quad \frac{T_n}{N_n} \pm \frac{1.96}{\sqrt{N_n}} \sqrt{\frac{1}{N_n} \sum_{i=1}^n v_i^2 - \left(\frac{T_n}{N_n} \right)^2}.$$

The width of this interval tends almost surely to 0 as $n \rightarrow \infty$.

An interesting question: in what sense, if any, is T_n/N_n an asymptotically optimal estimator of H ? Another: find a confidence band for the function $H(a) = E(\theta | x > a)$ for all $0 < a_0 < a < a_1 < \infty$.

II. Let

$$(35) \quad R_n = \sum_{i=1}^n u_i \cdot \theta_i = \text{sum of all the } N_n \text{ values } \theta_i \text{ for which } x_i > a.$$

We note that

$$(36) \quad Eu\theta = P(x > a) E(\theta | x > a) = Eu (Ev/Eu) = Ev,$$

so that as $n \rightarrow \infty$

$$(37) \quad \frac{T_n - R_n}{\sqrt{n}} \xrightarrow{D} N(0, \sigma_2^2),$$

where

$$(38) \quad \sigma_2^2 = E(v - u\theta)^2 = E(v^2 - 2v\theta + \theta^2).$$

III. Suppose that given θ_i , x_i and x'_i are independent, each with p.d.f. $f(x|\theta_i)$. For the case (18), from the i th of n normal populations we take two independent samples, each of size $2r+1$, and denote the two sample variances by x_i and x'_i . Let

$$(39) \quad S_n = \sum_{i=1}^n u_i x'_i = \text{sum of all the } N_n \text{ values } x'_i \text{ for which } x_i > a.$$

Since

$$(40) \quad E(ux'|\theta) = E(u|\theta) \cdot \theta = E(u\theta|\theta),$$

it follows from (36) that

$$(41) \quad Eux' = Eu\theta = Ev,$$

so that as $n \rightarrow \infty$

$$(42) \quad \frac{T_n - S_n}{\sqrt{n}} \xrightarrow{D} N(0, \sigma_3^2),$$

where

$$(43) \quad \sigma_3^2 = E(v - ux')^2 = E(v^2 - 2v x' + ux'^2).$$

To complete the discussion of II and III above, we state without proof that $\sigma_2^2 = Ew$ and $\sigma_3^2 = Ez$, where

$$(44) \quad \begin{aligned} w &= u \left[\frac{1}{1+r} \cdot x^2 + (r-2)a^r x^{2-r} + a^{2r} x^{2-2r} - \frac{r^2}{1+r} a^{1+r} x^{1-r} \right], \\ z &= u \left[\frac{2}{1+r} \cdot x^2 + (r-3)a^r x^{2-r} + a^{2r} x^{2-2r} - \frac{r(r-1)}{1+r} a^{1+r} x^{1-r} \right]. \end{aligned}$$

Thus, in analogy with (33),

$$(45) \quad \frac{\frac{T_n}{N_n} - \frac{R_n}{N_n}}{\frac{1}{\sqrt{N_n}} \sqrt{\frac{1}{N_n} \sum_{i=1}^n w_i}} \xrightarrow{D} N(0,1), \quad \frac{\frac{T_n}{N_n} - \frac{S_n}{N_n}}{\frac{1}{\sqrt{N_n}} \sqrt{\frac{1}{N_n} \sum_{i=1}^n z_i}} \xrightarrow{D} N(0,1).$$

For $r=1$, when (18) reduces to $f(x|\theta) = \theta^{-1} e^{-x/\theta}$, the negative exponential density with mean θ , (44) reduces to

$$(46) \quad z = 2w = u(x-a)^2.$$

Version III is particularly interesting because the second relation of (45) can be used not only to provide a prediction interval for S_n/N_n if the x_i^1 have not yet been observed, but also to provide a test of the null hypothesis that the sample variances x_i and x_i^1 for which $x_i > a$ are in fact from normal populations with the same unknown population variance σ_i^2 . To invent an example, suppose that $2r + 1$ vital measurements of some sort are taken on each of n persons at different times. It is considered that large variability of the measurements is

unhealthy, so to each of the N_n persons with sample variance $x_i > a$ (and only to such persons) a purportedly variance reducing treatment is administered. Then $2r + 1$ new measurements are taken on each of these N_n people, yielding sample variances x_i' . It is found that S_n/N_n is less than T_n/N_n , so that the treatment seems to reduce variance. It is objected that this is precisely what would be expected from a placebo -- regression towards mediocrity. But if

$$(47) \quad \frac{S_n}{N_n} < \frac{T_n}{N_n} - \frac{d}{\sqrt{N_n}} \sqrt{\frac{1}{N_n} \sum_{i=1}^n z_i^2}$$

for some d equal to 2 or more, say, it would indicate that the treatment does reduce variance when applied to people with high x_i values, even though no control group was used. An analogous result holds for other g.e.B. problems; e.g. for means instead of variances.

Remarks

A. Problems I - III, along with various generalizations of them, may be more important in practice than the problem of simultaneous inference about many parameters with which e.B. and compound decision theory were introduced in 1950.

B. As indicated in Robbins [1950] and Copas [1969], there is no basic distinction in practice between the e.B. and the compound formulations, in the latter of which we regard $\theta_1, \dots, \theta_n$ as unknown constants that are held fixed in repetitions of the experiment that gives rise to $x_1, \dots, x_n$. The James-Stein estimator of (16) was originally devised to show the inadmissibility of the usual estimator of the multivariate normal mean in the compound sense, while Efron and Morris (see Efron [1975] for references) have emphasized the e.B. aspect of (16).

C. Whether Neyman [1962] was right in thinking that the e.B. and compound approaches will have some effect on statistical practice remains to be seen. The following quotation from Efron [1975] emphasizes one oddity of the compound theory that may account for its neglect in practice: "If the different θ_i refer to obviously disjoint problems (e.g., θ_1 is the speed of light, θ_2 is the price of tea in China, θ_3 is the efficacy of a new treatment for psoriasis, etc.), combining the data can produce a definitely uncomfortable feeling in the statistician." (Compare Robbins [1950], page 133, lines 1-4.) A most amusing example of combining different sorts of data is given in E.S. Pearson [1925]. Everyone who is interested in Bayesianism should read

this paper, together with K. Pearson [1920, 1924], keeping in mind the warning by D.V. Lindley (page 421 of Copas [1969]) on the effect of even a seemingly slight deviation from orthodoxy: "there is no one less Bayesian than an empirical Bayesian".

D. We can combine the two problems that led to (16) and (20) respectively, in order to estimate simultaneously many means and variances. Adopting a new notation, let x_{ij} be independent and normal for $i=1, \dots, N$ and $j=1, \dots, n_i = 2r_i + 1 \geq 2$, with unknown $\mu_i = E x_{ij}$ and $\sigma_i^2 = \text{Var } x_{ij}$. Define

$$\begin{aligned} \bar{x}_i &= \frac{1}{n_i} \sum_j x_{ij}, \quad s_i^2 = \frac{1}{n_i-1} \sum_j (x_{ij} - \bar{x}_i)^2, \quad q^2 = \frac{1}{N} \sum_i s_i^2, \\ (48) \quad \bar{x} &= \frac{1}{N} \sum_i \bar{x}_i, \quad S^2 = \frac{1}{N} \sum_i (\bar{x}_i - \bar{x})^2, \quad d^4 = \left[\frac{1}{N} \sum_i \left(\frac{r_i}{1+r_i} \right) s_i^4 - q^4 \right]^+. \end{aligned}$$

Then one of the possible l.e.B. ways of estimating the σ_i^2 and μ_i is to use

$$(49) \quad \tilde{\sigma}_i^2 = q^2 + \left[1 - \frac{d^4 + q^4}{(1 + r_i) d^4 + q^4} \right] (s_i^2 - q^2),$$

and

$$(50) \quad \tilde{\mu}_i = \bar{x} + \left[1 - \frac{(\tilde{\sigma}_i^2/n_i)}{(\tilde{\sigma}_i^2/n_i) - \frac{1}{N} \sum_i (\sigma_i^2/n_i) + S^2} \right]^+ (\bar{x}_i - \bar{x}).$$

References

- Copas, J. B. [1969]. Compound decisions and empirical Bayes. Jour. Royal Stat. Soc. B, 31, 397-423.
- Efron, B. [1975]. Biased versus unbiased estimation. Advances in Mathematics, 16, 259-277.
- Neyman, J. [1962]. Two breakthroughs in the theory of statistical decision making. Rev. Int. Stat. Inst., 30, 11-27.
- Pearson, E. S. [1925]. Bayes' theorem examined in the light of experimental sampling. Biometrika, 17, 388-442.
- Pearson, K. [1920]. The fundamental problem of practical statistics. Biometrika, 13, 1-16 and 300-301.
- Pearson, K. [1924]. Note on Bayes' theorem. Biometrika, 16, 190-193. (See also W. Burnside's note on page 189.)
- Robbins, H. [1950]. Asymptotically subminimax solutions of compound statistical decision problems. Proc. Second Berkeley Symposium Math. Stat. and Prob., 131-148. U. of C. Press.
- Robbins, H. [1955]. An empirical Bayes approach to statistics. Proc. Third Berkeley Symposium Math. Stat. and Prob., Vol. 1, 157-164. U. of C. Press.