Probability in the Engineering and Informational Sciences, **15**, 2001, 351–367. Printed in the U.S.A.

VARIANCE ESTIMATION FOR THE NORMAL DISTRIBUTION UNDER LOG SYMMETRIC LOSS

MARK BROWN
Department of Mathematics
The City College, CUNY
New York, NY
E-mail: CYBERGARF@aol.com

For a normal sample with unknown mean, the almost universally used estimator of the variance, σ^2, is "the sample variance," $\hat{\sigma}^2(0) = (n-1)^{-1}S^2$, where

$$S^2 = \sum_1^n (X_i - \bar{X})^2.$$

This estimator is the minimum variance unbiased estimator of σ^2, but it is inadmissible under square error loss. It is dominated by the maximum likelihood estimator, $\hat{\sigma}^2(1) = n^{-1}S^2$, which is also inadmissible. We consider the class of estimators,

$$\hat{\sigma}^2(c) = (n-1+c)^{-1}S^2, \qquad -(n-1) < c < \infty,$$

and compare these estimators under a class of loss functions which we call "log symmetric." This class includes the loss function corresponding to the total variation distance,

$$l(\sigma_1^2, \sigma_2^2) = \sup_B |\Pr(N(\mu, \sigma_1^2) \in B) - \Pr(N(\mu, \sigma_2^2) \in B)|.$$

Let $L(c)$ denote the loss of the estimator $\hat{\sigma}^2(c)$, under an arbitrary log symmetric loss function. It is shown that $L(c)$ is stochastically increasing in $c \geq 0$. Thus, for example, the minimum variance unbiased estimator ($\hat{\sigma}^2(0)$), stochastically dominates the maximum likelihood estimator ($\hat{\sigma}^2(1)$). Furthermore, no estimator in the above class stochastically dominates $\hat{\sigma}^2(0)$. Asymptotic approximations to the distribution of $L(c)$ are derived to quantify the advantage of $\hat{\sigma}^2(0)$ over $\hat{\sigma}^2(1)$, as well as to study asymptotic expected loss.

© 2001 Cambridge University Press 0269-9648/01 $12.50

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

1. INTRODUCTION AND SUMMARY

For a single normal sample with unknown mean, the almost universally used estimator of the variance σ^2 is "the sample variance"

$$\hat{\sigma}^2(0) = \frac{1}{n-1}\sum_1^n (X_i - \bar{X})^2.$$

As all statisticians know, $\hat{\sigma}^2(0)$ is the minimum variance unbiased estimator of σ^2, but under square error loss, $\hat{\sigma}^2$ is inadmissible, being dominated by the maximum likelihood estimator

$$\hat{\sigma}^2(1) = \frac{1}{n}\sum_1^n (X_i - \bar{X})^2,$$

which is inadmissible also. In fact,

$$\hat{\sigma}^2(2) = \frac{1}{n+1}\sum_1^n (X_i - \bar{X})^2$$

is the minimum mean square error estimator among constant multiples of $\hat{\sigma}^2(0)$. These very well known results can be found in Bickel and Doksum [1, pp. 123–124, 143] or Lehmann [8, pp. 85, 113].

The preference for $\hat{\sigma}^2(0)$ would seem to suggest that unbiasedness is a highly valued attribute. Nevertheless, $\hat{\sigma}(0)$, the sample standard deviation, is the estimator of choice for σ and is biased. It is a routine exercise to find a constant multiple of $\hat{\sigma}(0)$ that is unbiased, but there do not appear to be advocates for this estimator.

The use of mean square error is widely acknowledged (going back to Gauss) to be primarily motivated by mathematical convenience. In this example, perhaps a more intrinsic measure of loss would be an increasing function of the total variation distance between the true distribution, $N(\mu, \sigma^2)$, and the distribution with estimated variance $\hat{\sigma}^2$, $N(\mu, \hat{\sigma}^2)$. This total variation distance is defined by

$$l(\hat{\sigma}^2, \sigma^2) = \sup_B |\Pr(N(\mu, \hat{\sigma}^2) \in B) - \Pr(N(\mu, \sigma^2) \in B)|. \tag{1.1}$$

This quantity is independent of μ and satisfies

(i) $l(\hat{\sigma}^2, \sigma^2) = g(\hat{\sigma}^2/\sigma^2)$,

where

(ii) $g(r) = g(r^{-1}), \qquad 0 < r < \infty,$

and

(iii) g is strictly increasing on $[1,\infty)$.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

The g function corresponding to total variation distance (1.1) is given by

$$g_v(r) = 2\left|\Phi\left(\left(\frac{r\log r}{r-1}\right)^{1/2}\right) - \Phi\left(\left(\frac{\log r}{r-1}\right)^{1/2}\right)\right|, \tag{1.2}$$

where Φ is the standard normal cumulative distribution function (c.d.f.). The derivation of (1.2) is found in Brown [3, p. 377].

A loss function with properties (i)–(iii) will be called a log symmetric loss function. This class coincides with the class of loss functions which are strictly increasing functions of total variation distance; that is,

$$l(\sigma_1^2, \sigma_2^2) = y(g_v(\sigma_1^2/\sigma_2^2)),$$

where y is strictly increasing on [0,1).

If the purpose in estimating σ^2 is (along with an estimate of μ) to estimate numerous probabilities under the unknown normal distribution, then a log symmetric loss function may be appropriate.

Other examples of g functions corresponding to log symmetric loss functions, relevant to this variance estimation problem, are

$$g_1(r) = |\log(r)| \tag{1.3}$$

and

$$g_2(r) = \left[1 - \left(\frac{2r^{1/2}}{1+r}\right)^{1/2}\right]^{1/2}. \tag{1.4}$$

The loss function corresponding to g_1 is proportional to an arc length distance as discussed in Brown [3, p. 375]. The loss corresponding to g_2 is the Hellinger distance, often used in large sample theory. LeCam [7, Chap. 7] defines and explores the connections between total variation and the Hellinger distance.

Consider the class of estimators

$$\hat{\sigma}^2(c) = \frac{1}{n-1+c}\sum_1^n (X_i - \bar{X})^2, \qquad -(n-1) < c < \infty. \tag{1.5}$$

This class includes the sample variance ($c = 0$), the maximum likelihood estimator of variance ($c = 1$), and the minimum mean square error scalar multiple estimator ($c = 2$). Denote by $L(c)$ a random variable whose distribution is that of the loss incurred by the use of $\hat{\sigma}^2(c)$, under a fixed but arbitrary log symmetric loss function. This distribution is independent of σ^2. In Section 2 (Thm. 2.1), we show that $L(c)$ is stochastically increasing in $c \geq 0$. More specifically,

$$\Pr(L(c) \leq x)$$

is nonincreasing in $c \geq 0$ for all x and strictly decreasing in $c \geq 0$ for any $x \in (g(1+), g(\infty))$.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

This stochastic dominance property is considerably stronger than increasing risk ($EL(c)\uparrow$) when this expectation exists.

We conclude that under log symmetric loss, for $0 \le c_1 < c_2$, $\hat{\sigma}^2(c_1)$ is better than $\hat{\sigma}^2(c_2)$. In particular, the minimum variance unbiased estimator ($c = 0$) is preferable to the maximum likelihood estimator ($c = 1$), which, in turn, is preferable to the minimum mean square scalar multiple estimator ($c = 2$). This conclusion provides strong support for the conventional wisdom on this issue.

The above stochastic monotonicity does not extend to $c \le 0$. Under a log symmetric loss function, no $\hat{\sigma}^2(c)$ with $-(n-1) < c \le 0$ will stochastically dominate another member of this subclass. This is proved in Corollaries 2.1 and 2.2. Within this subclass, we need to discriminate among estimators by the weaker criteria of risk (expected loss). Our asymptotic analysis in Section 3.5 suggests that $c = -\frac{2}{3}$ gives minimal asymptotic risk for a wide class of log symmetric loss functions.

To get an idea of the magnitude of the differences between estimators, we resort to asymptotic analysis via Edgeworth series expansions for the distribution of $L(c)$. These results are derived in Sections 3.4 and 3.5.

In comparing loss c.d.f.'s between the minimum variance unbiased estimator ($\hat{\sigma}^2(0)$) and the maximum likelihood estimator ($\hat{\sigma}^2(1)$) under total variation distance, we find that

$$\begin{aligned}&\Pr(\sqrt{2\pi e(n-1)}L(0) \le x) - \Pr(\sqrt{2\pi e(n-1)}L(1) \le x)\\ &\quad = \frac{\sqrt{2}\phi\left(\dfrac{x}{\sqrt{2}}\right)}{12(n-1)}\, x(3+x^2) + o(n^{-3/2}),\end{aligned} \tag{1.6}$$

where n is the sample size and ϕ the standard normal probability density function (p.d.f.). The $O(n^{-1})$ term is maximized at $\hat{x} = 2.091$, the maximum value being $0.24293(n-1)^{-1}$. For example, if $n = 20$, we estimate the maximum difference to be approximately 0.0128.

From a comparison of the expected total variation distance between $\hat{\sigma}^2(0)$ and $\hat{\sigma}^2(1)$, we find

$$EL(1) - EL(0) = \frac{7}{12\pi}\sqrt{\frac{2}{e}}(n-1)^{-3/2} + o(n^{-2}) \sim 0.15927(n-1)^{-3/2}. \tag{1.7}$$

For $n = 20$, this difference is approximately 0.0019.

Comparing the risks between the asymptotically optimal, $\hat{\sigma}^2(-\frac{2}{3})$, and $\hat{\sigma}^2(0)$, under total variation distance, we find

$$\begin{aligned}EL(0) - EL\left(-\frac{2}{3}\right) &= \frac{1}{9\pi}\sqrt{\frac{2}{e}}(n-1)^{-3/2} + o(n^{-2})\\ &\sim 0.03034(n-1)^{-3/2} \sim (5.25)^{-1}(EL(1) - EL(0)).\end{aligned} \tag{1.8}$$

The asymptotic improvement of $\hat{\sigma}^2(-\frac{2}{3})$ over $\hat{\sigma}^2(0)$ is only about 19% that of $\hat{\sigma}^2(0)$ over $\hat{\sigma}^2(1)$.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Thus, for small sample sizes, there is an advantage in $\hat{\sigma}^2(-\frac{2}{3})$ over $\hat{\sigma}^2(0)$, but it is not sufficiently dramatic to recommend a change in conventional practice.

Our results also apply to the variance estimation for a class of linear models, as well as the estimation of an exponential parameter. These cases are briefly discussed in Section 4.

2. COMPARISON OF ESTIMATORS

For $-(n-1) < c < \infty$, define $Z(c)$ to be a random variable distributed as $\chi^2_{n-1}/n - 1 + c$. This is the distribution of $\hat{\sigma}^2(c)/\sigma^2$. Recall that if X and Y are nonnegative random variables with absolutely continuous distributions, then X is larger than Y under monotone likelihood ratio ordering (abbreviated by $X \overset{l}{\geq} Y$) if there exist versions of their p.d.f.'s f_X and f_Y, such that $f_X(t)/f_Y(t)$ is nondecreasing in $t > 0$.

LEMMA 2.1: *For all $c \geq 0$, $(Z(c))^{-1} \overset{l}{\geq} Z(c)$. Moreover,*

$$R_c(a) \overset{\text{def}}{=} \frac{f^{(a)}_{(Z(c))^{-1}}}{f^{(a)}_{Z(c)}} > 1 \quad \textit{for } a > 1,\ c \geq 0. \tag{2.1}$$

PROOF: We calculate

$$R_c(a) = a^{-(n-1)} \exp\left(\frac{n-1+c}{2} (a - a^{-1}) \right) \quad \text{for } a > 0, \tag{2.2}$$

and after some algebra,

$$\frac{d}{da} \log R_c(a) = \frac{n-1+c}{2a^2} \left[(a-1)^2 + \frac{2ca}{n-1+c} \right]. \tag{2.3}$$

For $a > 0, c > 0$, the derivative in (2.3) is strictly positive. For $a > 0$ and $c = 0$, the derivative is strictly positive for $a \neq 1$, and zero for $a = 1$. It follows that in both cases ($c > 0$ and $c = 0$), $R_c(a)$ is strictly increasing in $a > 0$. Next, since $R_c(1) = 1$ and R_c is strictly increasing, it follows that $R_c(a) > 1$ for all $a > 1$ and $c \geq 0$. ■

Next, define $Y(c)$ to be a random variable distributed as $\max((Z(c))^{-1}, Z(c))$. $Y(c)$ takes values in $[1,\infty)$. If g is log symmetric, then

$$L(c) \sim g(\hat{\sigma}^2(c)/\sigma^2) \sim g(Z(c)) \sim g(Y(c)), \tag{2.4}$$

the last relationship holding because $g(r^{-1}) = g(r)$.

Recall that a log symmetric loss function g is strictly increasing on $[1,\infty)$. Define

$$g^{-1}(t) = \begin{cases} \sup\{x \geq 1 : g(x) \leq t\}, & g(1+) < t < g(\infty) \\ 1, & t \leq g(1+) = \lim_{y \downarrow 1} g(y) \\ \infty, & t \geq g(\infty) = \lim_{y \to \infty} g(y). \end{cases} \tag{2.5}$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

THEOREM 2.1: *For g log symmetric, $L(c)$ is stochastically increasing in $c \geq 0$. Moreover,*

$$\Pr(L(c) \leq x)$$

is strictly decreasing in $c \geq 0$, for $x \in (g(1+), g(\infty))$.

PROOF: First note that

$$\Pr(Y(c) \leq a) = F^{(a)}_{Z(c)} - F^{(a^{-1})}_{Z(c)} = F^{(a)}_{Z(c)} - \bar{F}^{(a)}_{(Z(c))^{-1}}, \tag{2.6}$$

where $F_{Z(c)}(F_{Z(c)^{-1}})$ is the c.d.f. of $Z(c)$ $(Z(c)^{-1})$ and $\bar{F} = 1 - F$.

It follows from (2.6) that

$$\frac{d}{dc}\Pr(Y(c) \leq a) = \frac{a}{n-1+c} f_{Z(c)}(a)[1 - R_c(a)], \tag{2.7}$$

where R_c is the likelihood ratio (2.1).

From (2.7) and Lemma 2.1, it follows that $\Pr(Y(c) \leq a)$ is strictly decreasing in $c \geq 0$ for all $a > 1$.

For $x \in (g(1+), g(\infty))$, $g^{-1}(x) \in (1, \infty)$, thus $\Pr(L(c) \leq x) = \Pr(Y(c) \leq g^{-1}(x))$ is strictly decreasing in $c \geq 0$. (For $x \leq g(1+)$, $\Pr(L(c) \leq x) \equiv 0$, and for $x \geq g(\infty)$, $\Pr(L(c) \leq x) \equiv 1$.) ■

COROLLARY 2.2: *Given log symmetric loss function g and $c_0 \in (-(n-1), 0)$, there exists an x such that*

$$\Pr(L(c_0) \leq x) > \Pr(L(c) \leq x) \quad \textit{for all } c \neq c_0.$$

Thus, no $\hat{\sigma}^2(c)$ with $c \in (-(n-1), 0)$ can be stochastically dominated by a member of the class $\{\hat{\sigma}^2(c), -(n-1) < c < \infty\}$.

PROOF: From (2.2), we see that for $a > 1$, $R_c(a)$ equals 1 if and only if

$$c = c(a) \stackrel{\text{def}}{=} (n-1)\left[\frac{2\log a}{a - a^{-1}} - 1\right]. \tag{2.8}$$

Furthermore, $R_c(a) > 1$ for $c > c(a)$, and less than 1 for $c < c(a)$. The function $c(a)$ is continuous and strictly decreasing. It decreases from zero to $-(n-1)$ as a increases from 1 to ∞. Given $c_0 \in (-(n-1), 0)$, there is a unique a, call it a^*, with $c(a^*) = c_0$.

It follows from (2.7) that $\Pr(Y(c) \leq a)$ achieves its unique maximum at $c = c(a)$. Given g and $c_0 \in (-(n-1), 0)$, choose $x = g(a^*)$ (recall $c(a^*) = c_0$). Then,

$$\begin{aligned}\Pr(L(c_0) \leq x) &= \Pr(L(c_0) \leq g(a^*)) = \Pr(Y(c_0) \leq a^*) \\ &> \Pr(Y(c) \leq a^*) = \Pr(L(c) \leq x)\end{aligned}$$

for all $c \neq c_0$. ■

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

COROLLARY 2.2: *Given $c \neq 0$ and log symmetric loss function g, there exists an x (which depends on c) such that*

$$\Pr(L(0) \leq x) > \Pr(L(c) \leq x);$$

thus, $\hat{\sigma}^2(0)$ is not stochastically dominated by any estimator in the class $\{\hat{\sigma}^2(c), -(n-1) < c < \infty\}$.

PROOF: Consider $f^{(1)}_{Y(c)}$, the p.d.f. of $Y(c)$ at $a = 1$. Note that

$$f^{(1)}_{Y(c)} = k(n-1+c)^{(n-1)/2} \exp\left[-\left(\frac{n-1+c}{2}\right)\right],$$

where k is a constant. By differentiating $\log f^{(1)}_{Y(c)}$ with respect to c, we find that $f^{(1)}_{Y(c)}$ achieves its unique maximum at $c = 0$. As each $Y(c)$ is concentrated on $[1,\infty)$ and possesses a continuous p.d.f., it follows that given $c \neq 0$ (recall that $f^{(1)}_{Y(0)} > f^{(1)}_{Y(c)}$), there exists an $a(c) \in (1,\infty)$, such that

$$\Pr(Y(0) \leq a) > \Pr(Y(c) \leq a) \quad \text{for } 1 < a \leq a(c). \tag{2.9}$$

Define $x = g(a(c))$. Then, from (2.9),

$$\Pr(L(0) \leq x) = \Pr(Y(0) \leq a(c)) > \Pr(Y(c) \leq a(c)) = \Pr(L(c) \leq x). \quad \blacksquare$$

3. ASYMPTOTICS

3.1. Total Variation Distance

The total variation distance between $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$ is given by $g_v(\sigma_1^2/\sigma_2^2)$, where g_v is defined in (1.2). The function $g_v(r)$ is not smooth at $r = 1$, but, rather,

$$g_v(r) = |h(r)|,$$

where h is infinitely differentiable in a neighborhood of $r = 1$. For our asymptotic expansions, we will need the first four derivatives of h at $r = 1$. Toward this end, define

$$k(x) = \Phi(x^{1/2}), \tag{3.1}$$

where Φ is the standard normal c.d.f. and

$$b(r) = \frac{\ln(r)}{r-1} = \sum_0^\infty \frac{(-1)^k (r-1)^k}{k+1}, \qquad 0 < r < 2. \tag{3.2}$$

Then,

$$\tfrac{1}{2} h(r) = k(rb(r)) - k(b(r)), \tag{3.3}$$

a form convenient for differentiation. We find from (3.2) that

$$b^{(j)}(1) = \left.\frac{d^j b}{dr^j}\right|_{r=1} = \frac{(-1)^j j!}{j+1}, \qquad j = 0,1,\ldots. \tag{3.4}$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Furthermore,

$$k^{(v)}(1) = \left.\frac{d^{(v)}k}{dx^{(v)}}\right|_{x=1} = \alpha(v)\phi(1), \qquad v = 1,2,3,4, \tag{3.5}$$

where $\alpha(1) = \frac{1}{2}, \alpha(2) = -\frac{1}{2}, \alpha(3) = \frac{3}{4}, \alpha(4) = -\frac{7}{4}$, and ϕ is the standard normal p.d.f. From (3.3)–(3.5), we find that

$$h(r) = \phi(1)[(r-1) - \tfrac{1}{2}(r-1)^2 + \tfrac{5}{16}(r-1)^3 - \tfrac{7}{32}(r-1)^4 + o(r-1)^4] \tag{3.6}$$

in a neighborhood of $r = 1$. Furthermore,

$$h^{-1}(\phi(1)x) = 1 + x + \frac{x^2}{2} + \frac{3x^3}{16} + \frac{x^4}{16} + o(x^4) \tag{3.7}$$

in a neighborhood of $x = 0$.

3.2. A Class of Log Symmetric Loss Functions

Define the class $\mathcal{G}$ to consist of functions on $(0,\infty)$ satisfying

(i) $g(x^{-1}) = g(x)$.
(ii) g strictly increasing on $[1,\infty)$.
(iii) $g(x) = |h(x)|$, with h possessing a continuous fifth derivative in a neighborhood of $x = 1$.
(iv) $h'(1) > 0$.

The class $\mathcal{G}$ includes $g_v(1.2)$, as well as $g_1(x) = |\log x|$ and g_2 (Hellinger distance). It follows that $g \in \mathcal{G}$ satisfies $h(1) = 0$ and

$$h(r^{-1}) = -h(r). \tag{3.8}$$

Differentiating both sides of (3.8) four times, we conclude that

$$h^{(2)}(1) = -h^{(1)}(1) \tag{3.9}$$

and

$$h^{(4)}(1) = 6(h^{(1)}(1) - h^{(3)}(1)). \tag{3.10}$$

Defining $\gamma = h^{(1)}(1) = g'(1+)$ and $\delta = h^{(3)}(1)$, it follows from (3.9) and (3.10) that

$$h(r) = \gamma(r-1) - \frac{\gamma}{2}(r-1)^2 + \frac{\delta}{6}(r-1)^3 + \frac{\gamma-\delta}{4}(r-1)^4 + o(r-1)^4 \tag{3.11}$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

in a neighborhood of $r = 1$. Moreover,

$$h^{-1}(\gamma x) = 1 + x + \frac{x^2}{2} + \frac{1}{6}\left(3 - \frac{\delta}{\gamma}\right)x^3 + \frac{1}{24}\left(9 - \frac{4\delta}{\gamma}\right)x^4 + o(x^4) \qquad \textbf{(3.12)}$$

in a neighborhood of $x = 0$. For total variation distance, $\gamma = \phi(1)$ and $\delta = (15/8)\phi(1)$. For the Hellinger distance, $\gamma = \frac{1}{4}$ and $\delta = 57/128$.

3.3. Symmetric Invertibility

A sequence of the form

$$z_n = 1 + an^{-1/2} + bn^{-1} + cn^{-3/2} + dn^{-2} + o(n^{-2})$$

is defined [4, p. 346] to be symmetrically invertible if

$$z_n^{-1} = 1 - an^{-1/2} + bn^{-1} - cn^{-3/2} + dn^{-2} + o(n^{-2}).$$

It is easily shown [4, p. 347] that z_n is symmetrically invertible if and only if both

$$b = \frac{a^2}{2} \quad \text{and} \quad d = ac - \frac{a^4}{8}. \qquad \textbf{(3.13)}$$

The following result will be useful in our asymptotic approximations in Section 3.5.

LEMMA 3.1: *Let $g(r) = |h(r)|$ belong to the class $\mathcal{G}$, defined in Section 3.2. Define $\gamma = h^{(1)}(1) > 0$ and $\delta = h^{(3)}(1)$. Then, $h^{-1}(\gamma x n^{-1/2})$ can be represented as a symmetrically invertible sequence. Specifically,*

$$h^{-1}(\gamma x n^{-1/2}) = 1 + xn^{-1/2} + \left(\frac{x^2}{2}\right)n^{-1} + cn^{-3/2} + dn^{-2} + o(n^{-2}) \qquad \textbf{(3.14)}$$

where $c = (x^3/6)(3 - \delta/\gamma)$, $d = (x^4/24)(9 - 4\delta/\gamma)$, and

$$\frac{1}{h^{-1}(\gamma x n^{-1/2})} = 1 - xn^{-1/2} + \left(\frac{x^2}{2}\right)n^{-1} - cn^{-3/2} + dn^{-2} + o(n^{-2}). \qquad \textbf{(3.15)}$$

PROOF: Equation (3.14) follows directly from (3.12). The coefficients satisfy the necessary and sufficient conditions for symmetric invertibility (3.13). ■

3.4. Edgeworth Expansions

For a sample of size n, the χ^2 statistic of interest has $n - 1$ degrees of freedom. In two other problems to be discussed in Section 4, the χ^2 statistic based on a sample of size n has $n - p$ degrees of freedom [one-way analysis of variance (ANOVA)] or $2n$ degrees of freedom (exponential). We will thus find it convenient to base our expansions on the degrees of freedom, m, with $m \to \infty$, and then convert to the sample size, n, if desired.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Let S_m denote a random variable distributed as χ^2_m (χ^2 with m degrees of freedom). The Edgeworth series expansion for the c.d.f. of S_m is well known [1, p. 33; 5, p. 45; 6, p. 40]:

$$\Pr\left(\frac{S_m - m}{\sqrt{2m}} \le x\right) = \Phi(x) - \phi(x)m^{-1/2}q_1(x) - \phi(x)m^{-1}q_2(x) - \phi(x)m^{-3/2}q_3(x) + o(m^{-3/2}), \quad \textbf{(3.16)}$$

where

$$q_1(x) = \frac{\sqrt{2}}{3}(x^2 - 1) \quad \text{with } q_1(-x) = q_1(x), \quad \textbf{(3.17)}$$

$$q_2(x) = \tfrac{1}{18}(2x^5 - 11x^3 + 3x) \quad \text{with } q_2(-x) = -q_2(x), \quad \textbf{(3.18)}$$

and q_3 is an even function ($q_3(-x) = q_3(x)$), whose terms we will not need.

Replacing x by $x + um^{-1/2} + vm^{-1} + wm^{-3/2}$ and using the delta method in (3.16), similarly to Hall [5, p. 69], we obtain

$$\begin{aligned}\Pr&\left(\frac{S_m - m}{\sqrt{2m}} \le x + um^{-1/2} + vm^{-1} + wm^{-3/2}\right)\\ &= \Phi(x) + \phi(x)m^{-1/2}[u - q_1(x)]\\ &\quad + \phi(x)m^{-1}\left[v - \frac{u^2x}{2} - u(q_1'(x) - xq_1(x)) - q_2(x)\right]\\ &\quad + \phi(x)m^{-3/2}q_3^*(u,v,w,x), \end{aligned} \quad \textbf{(3.19)}$$

where $q_3^*(u,-v,w,-x) = q_3^*(u,v,w,x)$.

In (3.19), substitute $(x/\sqrt{2}, u/\sqrt{2}, v/\sqrt{2}, w/\sqrt{2})$ for (x,u,v,w) and define $R_m = S_m/m \sim \chi^2_m/m$, obtaining

$$\begin{aligned}\Pr&(R_m \le 1 + xm^{-1/2} + um^{-1} + vm^{-3/2} + wm^{-2})\\ &= \Phi\left(\frac{x}{\sqrt{2}}\right) + \phi\left(\frac{x}{\sqrt{2}}\right)m^{-1/2}\tilde{q}_1(u,x) + \phi\left(\frac{x}{\sqrt{2}}\right)m^{-1}\tilde{q}_2(u,v,x)\\ &\quad + \phi\left(\frac{x}{\sqrt{2}}\right)m^{-3/2}\tilde{q}_3(u,v,w,x) + o(m^{-3/2}), \end{aligned} \quad \textbf{(3.20)}$$

where

$$\tilde{q}_1(u,x) = \frac{\sqrt{2}}{6}(3u + 2 - x^2), \qquad \tilde{q}_1(u,-x) = \tilde{q}_1(u,x), \quad \textbf{(3.21)}$$

$$\tilde{q}_2(u,v,x) = \left(\frac{\sqrt{2}}{72}\right)(36v - 9u^2x - 36ux + 6ux^3 - x^5 + 11x^3 - 6x),$$
$$\tilde{q}_2(u,-v,-x) = -\tilde{q}_2(u,v,x), \quad \textbf{(3.22)}$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

and $\tilde{q}_3$ satisfies

$$\tilde{q}_3(u,-v,w,-x) = \tilde{q}_3(u,v,w,x). \tag{3.23}$$

It follows from (3.20)–(3.23), that

$$\begin{aligned}&\Pr(R_m \in 1 \pm xm^{-1/2} + um^{-1} \pm vm^{-3/2} + wm^{-2}) \\ &\quad = 2\Phi\left(\frac{x}{\sqrt{2}}\right) - 1 + 2\phi\left(\frac{x}{\sqrt{2}}\right)m^{-1}\tilde{q}_2(u,v,x) + o(m^{-3/2}).\end{aligned} \tag{3.24}$$

We now have the tools to approximate the distribution of loss, $L(c)$, for the estimator $\hat{\sigma}^2(c)$, for a log symmetric loss function with $g \in \mathcal{G}$.

THEOREM 3.1: *Consider $g \in \mathcal{G}$, with $g(x) = |h(x)|$, $h^{(1)}(1) = \gamma > 0$, and $h^{(3)}(1) = \delta$. Then,*

(i) $$\begin{aligned}&\Pr(\sqrt{m}\gamma^{-1}L(c) \le x) \\ &= \left(2\Phi\left(\frac{x}{\sqrt{2}}\right) - 1\right) \\ &\quad + \frac{\sqrt{2}\phi\left(\frac{x}{\sqrt{2}}\right)}{36m}\left[-\frac{x^5}{4} + \left(11 - 6\frac{\delta}{\gamma}\right)x^3 - 6x - 3cx^3 - 9c^2x\right] \\ &\quad + o(m^{-3/2}).\end{aligned}$$

(ii) $$\begin{aligned}&\lim_{m\to\infty} m[\Pr(\sqrt{m}\gamma^{-1}L(0) \le x) - \Pr(\sqrt{m}\gamma^{-1}L(c) \le x)] \\ &= \frac{\sqrt{2}\phi\left(\frac{x}{\sqrt{2}}\right)}{12}(cx^3 + 3c^2x).\end{aligned}$$

For $c > 0$ and $x > 0$, the right-hand side is strictly positive and gives the asymptotic improvement of $\hat{\sigma}^2(0)$ over $\hat{\sigma}^2(c)$, with $c > 0$.

(iii) *For fixed x, the $O(m^{-1})$ term in (i) is maximized at $\hat{c} = -(x^2/6)$.*

PROOF: (i) Since $g(r^{-1}) = g(r) = |h(r)|$, it follows that for $0 < a < h(\infty)$,

$$g(r) \le a \Leftrightarrow \frac{1}{h^{-1}(a)} \le r \le h^{-1}(a), \tag{3.25}$$

where $h^{-1}(a) = g^{-1}(a)$, defined in (2.5).

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Recall that $\hat{\sigma}^2(c)/\sigma^2 \sim \chi_m^2/(m+c) \sim (1+c/m)^{-1}R_m$, where $R_m \sim x_m^2/m$. Consequently,

$$\Pr(m^{1/2}\gamma^{-1}L(c) \le x) = \Pr(m^{1/2}\gamma^{-1}g(\hat{\sigma}^2(c)/\sigma^2) \le x)$$
$$= \Pr\left(\frac{1}{h^{-1}(m^{-1/2}\gamma x)} \le \left(1+\frac{c}{m}\right)^{-1} R_m \le h^{-1}(m^{-1/2}\gamma x)\right). \tag{3.26}$$

Recall that $h^{-1}(m^{-1/2}\gamma x)$ is symmetrically invertible (Lemma 3.1). Thus, (3.26) reduces to

$$\Pr\left(R_m \in (1+cm^{-1})\left(1 \pm xm^{-1/2} + \left(\frac{x^2}{2}\right)m^{-1} \pm \frac{1}{6}\,m^{-3/2}\left(3-\frac{\delta}{\gamma}\right)x^3 + \frac{1}{24}\,m^{-2}\left(9-\frac{4\delta}{\gamma}\right)x^4 + o(m^{-2})\right)\right). \tag{3.27}$$

By (3.24), (3.26), and (3.27), with

$$u = c + \frac{x^2}{2}, \qquad v = cx + \frac{1}{6}\left(3-\frac{\delta}{\gamma}\right)x^3,$$

and

$$w = \frac{cx^2}{2} + \frac{1}{24}\left(9-\frac{4\delta}{\gamma}\right)x^4,$$

the desired result follows.

(ii) Follows immediately from (i).
(iii) For fixed x, to maximize the $O(m^{-1})$ term, we need to minimize $3cx^3 + 9c^2x$. A minimum value of $-(x^5/4)$ is attained at $\hat{c} = -(x^2/6)$.

Remark 3.1: Result (1.6) follows from (ii), with $c = 1$.

Remark 3.2: Theorem 2.1 and Corollaries 2.1 and 2.2 all have asymptotic $O(m^{-1})$ analogs. For example, corresponding to Theorem 2.1, we have the comparison

$$\lim_{m\to\infty} m[\Pr(m^{1/2}L(c_1) \le x) - \Pr(m^{1/2}L(c_2) \le x)] > 0$$

for all $x > 0$ and $0 \le c_1 < c_2$.

There are similar asymptotic versions of Corollaries 2.1 and 2.2.

An asymptotic $O(m^{-1})$ comparison not suggested by the results of Section 2 is

$$\lim_{m\to\infty} m[\Pr(m^{1/2}L(c_1) \le x) - \Pr(m^{1/2}L(c_2) \le x)] > 0$$

for all $x > 0$ and $c_1 < 0 < |c_1| < c_2$. This follows from part (i) of Theorem 3.1.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Remark 3.3: Part (iii) of Theorem 3.1 also follows from Lemma 2.1. The optimizing c for $\gamma m^{1/2}x$ is given by

$$c(g^{-1}(\gamma m^{-1/2}x)),$$

where $c(\cdot)$ is defined by (2.8) and $g^{-1}(\gamma m^{-1/2}x)$ can be expanded by (3.14). Substituting and passing to the limit, we find that

$$c(g^{-1}(\gamma m^{-1/2}x)) = -\frac{x^2}{6} + o(1).$$

3.5. Minimizing the Expected Loss

We will obtain an expansion for $L(c) = Eg(\hat{\sigma}^2(c)/\sigma^2)$ by integrating the expansion for $\Pr(L(c) > x)$, obtained in Theorem 3.1(i). This procedure is heuristic in the sense that it assumes that a certain $o(n^{-2})$ sequence of functions of x integrates to a $o(n^{-2})$ sequence of constants. To justify this, we would need to develop a nonuniform error bound on the Edgeworth expansion for $\Pr(L(c) > x)$, similar, for example, to the proofs of Theorems 19.2 and 22.1 of Bhattacharya and Ranga Rao [2].

For all $g \in \mathcal{G}$ for which the above integration requirement holds, we find that the value of c which minimizes the expected loss is $c = -\frac{2}{3}$. Thus,

$$\hat{\sigma}^2\left(-\frac{2}{3}\right) = \frac{1}{n - 5/3}\sum_1^n (X_i - \bar{X})^2$$

is asymptotically optimal within the class $\{\hat{\sigma}^2(c), -\infty < c < \infty\}$, in the sense of minimizing the asymptotic expected loss.

The computation follows from Theorem 3.1(i), as

$$\begin{aligned} Eg(\hat{\sigma}^2(c)/\sigma^2) = m^{-1/2}\gamma \int_0^\infty 2\bar{\Phi}\left(\frac{x}{\sqrt{2}}\right) + \frac{\sqrt{2}\phi\left(\frac{x}{\sqrt{2}}\right)}{36m} \\ \times \left[\frac{x^5}{4} + \left(6\frac{\delta}{\gamma} - 11\right)x^3 + 6x + 3cx^3 + 9c^2x\right] dx \\ + o(m^{-2}). \end{aligned} \tag{3.28}$$

Next, we compute

$$\int_0^\infty \bar{\Phi}\left(\frac{x}{\sqrt{2}}\right) dx = \pi^{-1/2} \tag{3.29}$$

and

$$\int_0^\infty x^{2m+1}\phi\left(\frac{x}{\sqrt{2}}\right) dx = 2^{a_m}(2\pi)^{-1/2}, \qquad m = 0,1,2, \tag{3.30}$$

with $a_0 = 1$, $a_1 = 3$, and $a_2 = 6$.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Substituting these values into (3.28) leads to

$$EL(c) = \frac{2\gamma}{\sqrt{\pi m}}\left[1 + \frac{1}{12m}\left(8\frac{\delta}{\gamma} - 10 + 3c^2 + 4c\right)\right] + o(m^{-2}). \qquad \textbf{(3.31)}$$

From (3.31), it follows that $EL(c)$ is asymptotically minimized at $c = -\frac{2}{3}$. Expressions (1.7) and (1.8), discussed in Section 1, follow directly from (3.31). Note that the range of c for which $\hat{\sigma}^2(c)$ has smaller asymptotic risk than $\hat{\sigma}^2(0)$ is $(-\frac{4}{3},0)$.

For the total variation distance loss function, $\gamma = \phi(1)$ and $\delta/\gamma = 15/8$; thus,

$$Eg_v(\hat{\sigma}^2(c)/\sigma^2) = \frac{1}{\pi}\sqrt{\frac{2}{me}}\left[1 + \frac{1}{12m}(3c^2 + 4c + 5)\right] + o(m^{-2}). \qquad \textbf{(3.32)}$$

4. EXAMPLES

4.1. One-Way ANOVA

Consider p independent normal samples, each with k observations, the ith sample consisting of

$$X_{i,1},\ldots,X_{i,k}, \quad \text{independent } N(\mu_i,\sigma^2).$$

The variance in each of the k samples is assumed to be a common unknown σ^2. Define

$$S^2 = \sum_{i,j}(X_{ij} - X_{i.})^2 \quad \text{with } X_{i.} = \frac{1}{k}\sum_{j=1}^{k} X_{ij}.$$

The minimum variance unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2(0) = \frac{S^2}{n-p}, \quad \text{where } n = kp,$$

and the maximum likelihood estimator is

$$\hat{\sigma}^2(p) = \frac{S^2}{n}.$$

Consider the total variation distance between $N(\mu^{p\times 1},\sigma_1^2 I)$ (p independent normals with mean vector μ, and variance σ^2) and $N(\mu^{p\times 1},\sigma_2^2 I)$,

$$l(\sigma_1^2,\sigma_2^2) = \sup_B |\Pr(N(\mu,\sigma_1^2 I) \in B) - \Pr(N(\mu,\sigma_2^2 I) \in B)|.$$

The loss function l is of the form $l(\sigma_1^2,\sigma_2^2) = g(\sigma_1^2/\sigma_2^2)$, where g is log symmetric and is given by

$$g(r) = \left|\Pr\left(\chi_p^2 \le \frac{pr\log r}{r-1}\right) - \Pr\left(\chi_p^2 \le \frac{p\log r}{r-1}\right)\right|.$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

It follows from Theorem 2.1 that the loss $L(c)$ (of $\hat{\sigma}^2(c)$) is stochastically increasing in $c \geq 0$. The obvious analogs of Lemmas 2.1 and 2.2 also hold.

Furthermore, the expansion (3.31) holds, with $m = n - p = k(p-1)$. As $k \to \infty$ (and thus $n \to \infty$), the member of the class with smallest asymptotic expected loss is

$$\hat{\sigma}^2(-2/3) = \frac{S^2}{n - p - 2/3}.$$

4.2. Linear Models

Consider the model

$$X^{n\times 1} \sim N(\mu, \sigma^2\Sigma),$$

where Σ is a known positive-definite matrix, σ^2 is unknown, and μ is known to lie in $\mathcal{M}$, a p-dimensional subspace of R^n.

Consider, next, the loss function

$$l(\sigma_1^2, \sigma_2^2) = \sup_B |\Pr(N(\mu, \sigma_1^2\Sigma) \in B) - \Pr(N(\mu, \sigma_2^2\Sigma) \in B)|.$$

Again, $l(\sigma_1^2, \sigma_2^2) = g(\sigma_1^2/\sigma_2^2)$, where g is log symmetric. We compute g,

$$g(r) = \left| \Pr\left(\chi_n^2 \leq \frac{nr \log r}{r-1}\right) - \Pr\left(\chi_n^2 \leq \frac{n \log r}{r-1}\right) \right|.$$

In the one-way ANOVA, there was a natural asymptotic situation ($k \to \infty$). We fixed p and the loss function g. Here, the function g depends on n, and there may not be an asymptotic problem to consider. Nevertheless, the results of Section 2 apply.

Define

$$S^2 = \|X - P_{\mathcal{M}} X\|^2,$$

where $\|x\|^2 = x'\Sigma^{-1}x$, and $P_{\mathcal{M}}X$ is the projection of X on $\mathcal{M}$ under the inner product, $(x, y) = x'\Sigma^{-1}y$. Define

$$\hat{\sigma}^2(c) = \frac{S^2}{n - p + c}, \qquad -(n-p) < c < \infty.$$

Then, once again, $L(c)$ is stochastically increasing in $c \geq 0$, and the minimum variance unbiased estimator, $\hat{\sigma}^2(0)$, stochastically dominates the maximum likelihood error (MLE), $\hat{\sigma}^2(p)$.

4.3. Exponential Distribution

Let $X_1, \ldots, X_n$ denote a sample from an exponential distribution with parameter θ,

$$f_\theta(x) = \theta e^{-\theta x}, \qquad x \geq 0, \theta > 0.$$

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

Define $S = \sum_1^n X_i$. The minimum variance unbiased estimator of θ is given by

$$\hat{\theta}(-1) = \frac{n-1}{S},$$

and the MLE is

$$\hat{\theta}(0) = \frac{n}{S}.$$

Define the class of estimators

$$\hat{\theta}(c) = \frac{n+c}{S}, \qquad -n < c < \infty, \tag{4.1}$$

and note that for log symmetric g,

$$g(\hat{\theta}(c)/\theta) = g(\theta/\hat{\theta}(c)) \sim g(x_{2n}^2/2(n+c)).$$

The g function corresponding to the total variation distance between exponential distributions with parameters θ_1 and θ_2 with $\theta_1/\theta_2 = r$ is given by

$$g(r) = |r-1|r^{r/(1-r)},$$

as is derived in [3, p. 378].

The results of Sections 2 and 3 apply. Once again, $L(c)$ is stochastically increasing in $c \geq 0$. In this case, the minimum variance unbiased estimator ($c = -1$) and the MLE ($c = 0$) do not stochastically dominate one another.

The expansion (3.31) holds, and in this example, $m = 2n$, so that $m - \frac{2}{3} = 2(n - \frac{1}{3})$. Consequently, the member of the class (4.1) which asymptotically minimizes the expected risk is

$$\hat{\theta}(-\tfrac{1}{3}) = \frac{n-\frac{1}{3}}{S} = \frac{1-(3n)^{-1}}{\bar{x}},$$

where $\bar{x} = S|n$. Further asymptotic work in this exponential example can be found in Brown [4].

5. COMMENTS AND ADDITIONS

1. Total variation distance is independent of the labeling of the distributions. Thus, if $\hat{\theta}_1$ is preferred to $\hat{\theta}_2$ on the basis of better total variation distance loss characteristics, then $\nu(\hat{\theta}_1)$ will be preferred to $\nu(\hat{\theta}_2)$ for any 1–1 function ν. For example, in the normal variance case, $\nu(\hat{\sigma}^2(0))$ stochastically dominates $\nu(\hat{\sigma}^2(1))$ for any 1–1 function, ν, on $(0,\infty)$.
2. I have been unable to show that the total variation distance between $N(\bar{X}, \hat{\sigma}^2(0))$ and $N(\mu, \sigma^2)$ tends to be smaller than that between $N(\bar{X}, \hat{\sigma}^2(1))$ and $N(\mu, \sigma^2)$. I believe this to be the case (at least asymptotically), and I pose it as a hopefully interesting research problem.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press

3. The key observation in developing expressions for total variation distance is that if $f_{\theta_1}(t) > f_{\theta_2}(t)$ if and only if $t \in B$, then the total variation distance between the two distributions is $P_{\theta_1}(B) - P_{\theta_2}(B)$. Expressions (1.2), (4.1), (4.2), and (4.4) follow easily from this observation.

Acknowledgment

This work was sponsored by the Office of Equal Opportunity Employment of the National Security Agency.

References

1. Bickel, P.J. & Doksum, K. (1977). *Mathematical statistics: Basic ideas and selected topics.* San Francisco: Holden-Day.
2. Bhattacharya, R.N. & Ranga Rao, R. (1976). *Normal approximations and asymptotic expansions.* New York: Wiley.
3. Brown, M. (1995). The distribution of total variation distance, with applications to simultaneous confidence intervals. *Computers and Operations Research* 22(4): 373–381.
4. Brown, M. (1997). Estimation of an exponential distribution. *Probability in the Engineering and Informational Sciences* 11: 341–359.
5. Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York: Springer-Verlag.
6. Jensen, J.L. (1995). *Saddlepoint approximations.* Oxford: Clarendon Press.
7. LeCam, L. (1974). *Notes on asymptotic methods in statistical decision theory.* Publication CRM 245. Montréal: Centre De Recherches Mathématiques.
8. Lehmann, E.L. (1983). *Theory of point estimation.* New York: Wiley.

https://doi.org/10.1017/S0269964801153052 Published online by Cambridge University Press