Asymptotic Relative Efficiency in Estimation

Robert Serfling*
University of Texas at Dallas

October 2009

Prepared for forthcoming
INTERNATIONAL ENCYCLOPEDIA OF STATISTICAL SCIENCES,
to be published by Springer

Asymptotic relative efficiency of two estimators

For statistical estimation problems, it is typical and even desirable that several reasonable estimators can arise for consideration. For example, the mean and median parameters of a symmetric distribution coincide, and so the *sample mean* and the *sample median* become competing estimators of the point of symmetry. *Which is preferred? By what criteria shall we make a choice?*

One natural and time-honored approach is simply to compare the sample sizes at which two competing estimators meet a given standard of performance. This depends upon the chosen measure of performance and upon the particular population distribution F.

To make the discussion of sample mean versus sample median more precise, consider a distribution function F with density function f symmetric about an unknown point θ to be estimated. For $\{X_1, \ldots, X_n\}$ a sample from F, put $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\mathrm{Med}_n = \mathrm{median}\{X_1, \ldots, X_n\}$. Each of $\overline{X}_n$ and Med_n is a consistent estimator of θ in the sense of convergence in probability to θ as the sample size $n \to \infty$. To choose between these estimators we need to use further information about their performance. In this regard, one key aspect is *efficiency*, which answers: *How spread out about θ is the sampling distribution of the estimator?* The smaller the variance in its sampling distribution, the more "efficient" is that estimator.

Here we consider "large-sample" sampling distributions. For $\overline{X}_n$, the classical central limit theorem tells us: if F has finite variance σ_F^2, then the sampling distribution of $\overline{X}_n$ is approximately $N(\theta, \sigma_F^2/n)$, i.e., Normal with mean θ and variance σ_F^2/n. For Med_n, a similar

*Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083-0688, USA. Email: `serfling@utdallas.edu`. Website: `www.utdallas.edu/~serfling`. Support by NSF Grant DMS-0805786 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

classical result [11] tells us: if the density f is continuous and positive at θ, then the sampling distribution of Med_n is approximately $N(\theta, 1/4[f(\theta)]^2 n)$. On this basis, we consider $\overline{X}_n$ and Med_n to perform equivalently at respective sample sizes n_1 and n_2 if

$$\frac{\sigma_F^2}{n_1} = \frac{1}{4[f(\theta)]^2 n_2}.$$

Keeping in mind that these sampling distributions are only approximations assuming that n_1 and n_2 are "large", we define the *asymptotic relative efficiency (ARE)* of Med to $\overline{X}$ as the *large-sample limit* of the ratio n_1/n_2, i.e.,

$$\mathrm{ARE}(\mathrm{Med}, \overline{X}, F) = 4[f(\theta)]^2 \sigma_F^2. \tag{1}$$

Definition in the general case

For any parameter η of a distribution F, and for estimators $\widehat{\eta}^{(1)}$ and $\widehat{\eta}^{(2)}$ approximately $N(\eta, V_1(F)/n)$ and $N(\eta, V_2(F)/n)$, respectively, the *ARE of* $\widehat{\eta}^{(2)}$ *to* $\widehat{\eta}^{(1)}$ is given by

$$\mathrm{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) = \frac{V_1(F)}{V_2(F)}. \tag{2}$$

Interpretation. If $\widehat{\eta}^{(2)}$ is used with a sample of size n, the number of observations needed for $\widehat{\eta}^{(1)}$ to perform equivalently is $\mathrm{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) \times n$.

Extension to the case of multidimensional parameter. For a parameter $\boldsymbol{\eta}$ taking values in $\mathbb{R}^k$, and two estimators $\widehat{\boldsymbol{\eta}}^{(i)}$ which are k-variate Normal with mean $\boldsymbol{\eta}$ and nonsingular covariance matrices $\boldsymbol{\Sigma}_i(F)/n$, $i = 1, 2$, we use (see [11])

$$\mathrm{ARE}(\widehat{\boldsymbol{\eta}}^{(2)}, \widehat{\boldsymbol{\eta}}^{(1)}, F) = \left(\frac{|\boldsymbol{\Sigma}_1(F)|}{|\boldsymbol{\Sigma}_2(F)|} \right)^{1/k}, \tag{3}$$

the ratio of *generalized variances* (determinants of the covariance matrices), raised to the power $1/k$.

Connection with the maximum likelihood estimator

Let F have density $f(x \,|\, \eta)$ parameterized by $\eta \in \mathbb{R}$ and satisfying some differentiability conditions with respect to η. Suppose also that $I(F) = E_\eta\{[\frac{\partial}{\partial \eta} \log f(x \,|\, \eta)]^2\}$ (the *Fisher information*) is positive and finite. Then [5] it follows that (i) the *maximum likelihood estimator* $\widehat{\eta}^{(\mathrm{ML})}$ of η is approximately $N(\eta, 1/I(F)n)$, and (ii) for a wide class of estimators $\widehat{\eta}$ that are approximately $N(\eta, V(\widehat{\eta}, F)/n)$, a *lower bound* to $V(\widehat{\eta}, F)$ is $1/I(F)$. In this situation, (2) yields

$$\mathrm{ARE}(\widehat{\eta}, \widehat{\eta}^{(\mathrm{ML})}, F) = \frac{1}{I(F)V(\widehat{\eta}, F)} \leq 1, \tag{4}$$

making $\widehat{\eta}^{(\mathrm{ML})}$ (asymptotically) the most efficient among the given class of estimators $\widehat{\eta}$. We note, however, as will be discussed later, that (4) does not necessarily make $\widehat{\eta}^{(\mathrm{ML})}$ the estimator of choice, when certain other considerations are taken into account.

Detailed discussion of estimation of point of symmetry

Let us now discuss in detail the example treated above, with F a distribution with density f symmetric about an unknown point θ and $\{X_1, \ldots, X_n\}$ a sample from F. For estimation of θ, we will consider not only $\overline{X}_n$ and Med_n but also a third important estimator.

Mean versus median

Let us now formally compare $\overline{X}_n$ and Med_n and see how the ARE differs with choice of F. Using (1) with $F = N(\theta, \sigma_F^2)$, it is seen that

$$\text{ARE}(\text{Med}, \overline{X}, N(\theta, \sigma_F^2)) = 2/\pi = 0.64.$$

Thus, for sampling from a *Normal* distribution, the sample mean performs as efficiently as the sample median using only 64% as many observations. (Since θ and σ_F are location and scale parameters of F, and since the estimators $\overline{X}_n$ and Med_n are location and scale equivariant, their ARE does not depend upon these parameters.) The superiority of $\overline{X}_n$ here is no surprise since it is the MLE of θ in the model $N(\theta, \sigma_F^2)$.

As noted above, *asymptotic* relative efficiencies pertain to large sample comparisons and need not reliably indicate small sample performance. In particular, for F *Normal*, the *exact* relative efficiency of Med to $\overline{X}$ for sample size $n = 5$ is a very high 95%, although this decreases quickly, to 80% for $n = 10$, to 70% for $n = 20$, and to 64% in the limit.

For sampling from a *double exponential* (or *Laplace*) distribution with density $f(x) = \lambda e^{-\lambda|x-\theta|}/2$, $-\infty < x < \infty$ (and thus variance $2/\lambda^2$), the above result favoring $\overline{X}_n$ over Med_n is reversed: (1) yields

$$\text{ARE}(\text{Med}, \overline{X}, \text{Laplace}) = 2,$$

so that the sample mean requires 200% as many observations to perform equivalently to the sample median. Again, this is no surprise because for this model the MLE of θ is Med_n.

A compromise: the Hodges-Lehmann location estimator

We see from the above that the ARE depends dramatically upon the shape of the density f and thus must be used cautiously as a benchmark. For Normal versus Laplace, $\overline{X}_n$ is either greatly superior or greatly inferior to Med_n. This is a rather unsatisfactory situation, since in practice we might not be quite sure whether F is Normal or Laplace or some other type. A very interesting solution to this dilemma is given by an estimator that has excellent *overall performance*, the so-called *Hodges-Lehmann location estimator* [2]:

$$\text{HL}_n = \text{Median}\left\{\frac{X_i + X_j}{2}\right\},$$

the median of all pairwise averages of the sample observations. (Some authors include the cases $i = j$, some not.) We have [3] that HL_n is asymptotically $N(\theta, 1/12[\int f^2(x)dx]^2 n)$, which yields that $\text{ARE}(\text{HL}, \overline{X}, N(\theta, \sigma_F^2)) = 3/\pi = 0.955$ and $\text{ARE}(\text{HL}, \overline{X}, \text{Laplace}) = 1.5$. Also, for the *Logistic* distribution with density $f(x) = \sigma^{-1} e^{(x-\theta)/\sigma}/[1 + e^{(x-\theta)/\sigma}]^2$, $-\infty < x < \infty$, for which HL_n is the MLE of θ and thus optimal, we have $\text{ARE}(\text{HL}, \overline{X}, \text{Logistic}) = \pi^2/9 = 1.097$ (see [4]). Further, for $\mathcal{F}$ the class of all distributions symmetric about θ and having finite variance, we have $\inf_{\mathcal{F}} \text{ARE}(\text{HL}, \overline{X}, F) = 108/125 = 0.864$ (see [3]). The estimator HL_n is highly competitive with $\overline{X}$ at Normal distributions, can be infinitely more efficient at some other symmetric distributions F, and is never much less efficient at any distribution F in $\mathcal{F}$. The computation of HL_n appears at first glance to require $O(n^2)$ steps, but a much more efficient $O(n \log n)$ algorithm is available (see [6]).

Efficiency versus robustness trade-off

Although the asymptotically most efficient estimator is given by the MLE, the particular MLE depends upon the shape of F and can be drastically inefficient when the actual F departs even a little bit from the nominal F. For example, if the assumed F is $N(\mu, 1)$ but the actual model differs by a small amount ε of "contamination", i.e., $F = (1 - \varepsilon)N(\mu, 1) + \varepsilon N(\mu, \sigma^2)$, then

$$\text{ARE}(\text{Med}, \overline{X}, F) = \frac{2}{\pi}\left(1 - \varepsilon + \varepsilon\sigma^{-1}\right)^2\left(1 - \varepsilon + \varepsilon\sigma^2\right),$$

which equals $2/\pi$ in the "ideal" case $\varepsilon = 0$ but otherwise $\to \infty$ as $\sigma \to \infty$. A small perturbation of the assumed model thus can destroy the superiority of the MLE.

One way around this issue is to take a *nonparametric* approach and seek an estimator with ARE satisfying a favorable lower bound. Above we saw how the estimator HL_n meets this need.

Another criterion by which to evaluate and compare estimators is *robustness*. Here let us use finite-sample *breakdown point (BP)*: the minimal fraction of sample points which may be taken to a limit L (e.g., $\pm\infty$) without the estimator also being taken to L. A *robust* estimator remains stable and effective when in fact the sample is only partly from the nominal distribution F and contains some non-F observations which might be relatively extreme contaminants.

A single observation taken to ∞ (with n fixed) takes $\overline{X}_n$ with it, so $\overline{X}_n$ has BP = 0. Its optimality at Normal distributions comes at the price of a complete sacrifice of robustness. In comparison, Med_n has extremely favorable BP = 0.5 but at the price of a considerable loss of efficiency at Normal models.

On the other hand, the estimator HL_n appeals broadly, possessing *both* quite high ARE over a wide class of F and relatively high $\text{BP} = 1 - 2^{-1/2} = 0.29$.

As another example, consider the problem of estimation of scale. Two classical scale estimators are the *sample standard deviation* s_n and the *sample MAD* (median absolute deviation about the median) MAD_n. They estimate scale in different ways but can be regarded as competitors in the problem of estimation of σ in the model $F = N(\mu, \sigma^2)$, as follows. With both μ and σ unknown, the estimator s_n is (essentially) the MLE of σ and is asymptotically most efficient. Also, for this F, the population MAD is equal to $\Phi^{-1}(3/4)\sigma$, so that the estimator $\widehat{\sigma}_n = \mathrm{MAD}_n/\Phi^{-1}(3/4) = 1.4826\,\mathrm{MAD}_n$ competes with s_n for estimation of σ. (Here Φ denotes the standard normal distribution function, and, for any F, $F^{-1}(p)$ denotes the pth quantile, $\inf\{x : F(x) \geq p\}$, for $0 < p < 1$.) To compare with respect to robustness, we note that a single observation taken to ∞ (with n fixed) takes s_n with it, s_n has BP = 0. On the other hand, MAD_n and thus $\widehat{\sigma}_n$ have BP = 0.5, like Med_n. However, $\mathrm{ARE}(\widehat{\sigma}_n, s_n, N(\mu, \sigma^2)) = 0.37$, even worse than the ARE of Med_n relative to $\overline{X}$. Clearly desired is a more balanced trade-off between efficiency and robustness than provided by either of s_n and $\widehat{\sigma}_n$. Alternative scale estimators having the same 0.5 BP as $\widehat{\sigma}_n$ but much higher ARE of 0.82 relative to s_n are developed in [10]. Also, further competitors offering a range of trade-offs given by (BP, ARE) $= (0.29, 0.86)$ or $(0.13, 0.91)$ or $(0.07, 0.96)$, for example, are developed in [12].

In general, efficiency and robustness trade off against each other. Thus ARE should be considered in conjunction with robustness, choosing the balance appropriate to the particular application context. This theme is prominent in the many examples treated in [14].

A few additional aspects of ARE

Connections with confidence intervals

In view of the asymptotic normal distribution underlying the above formulation of ARE in estimation, we may also characterize the ARE given by (2) as the limiting ratio of sample sizes at which the *lengths of associated confidence intervals at approximate level* $100(1-\alpha)\%$,

$$\widehat{\eta}^{(i)} \;\pm\; \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{V_i(F)}{n_i}},\;\; i = 1, 2,$$

converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$. (In practice, of course, consistent estimates of $V_i(F)$, $i = 1, 2$, are used in forming the CI.)

Fixed width confidence intervals and ARE

One may alternatively consider confidence intervals of *fixed length*, in which case (under typical conditions) the noncoverage probability depends on n and tends to 0 at an exponential rate, i.e., $n^{-1}\log\alpha_n \to c > 0$, as $n \to \infty$. For fixed width confidence intervals of the form

$$\widehat{\eta}^{(i)} \;\pm\; d\,\sigma_F,\;\; i = 1, 2,$$

we thus define the *fixed width asymptotic relative efficiency (FWARE)* of two estimators as the limiting ratio of sample sizes at which the respective *noncoverage probabilities* $\alpha_n^{(i)}$, $i = 1, 2$, of the associated fixed width confidence intervals converge to zero at the same exponential rate. In particular, for Med versus $\overline{X}$, and letting $\eta = 0$ and $\sigma_F = 1$ without loss of generality, we obtain [13]

$$\text{FWARE}(\text{Med}, \overline{X}, F) = \frac{\log m(-d)}{\log[2(F(d) - F^2(d))^{1/2}]}, \tag{5}$$

where $m(-d)$ is a certain parameter of the moment generating function of F. The FWARE is derived using *large deviation theory* instead of the central limit theorem. As $d \to 0$, the FWARE converges to the ARE. Indeed, for F a Normal distribution, this convergence (to $2/\pi = 0.64$) is quite rapid: the expression in (5) rounds to 0.60 for $d = 2$, to 0.63 for $d = 1$, and to 0.64 for $d \leq 0.1$.

Confidence ellipsoids and ARE

For an estimator $\widehat{\boldsymbol{\eta}}$ which is asymptotically k-variate Normal with mean $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma}/n$, as the sample size $n \to \infty$, we may form (see [11]) an *associated ellipsoidal confidence region of approximate level* $100(1-\alpha)\%$ for the parameter $\boldsymbol{\eta}$,

$$E_{n,\alpha} = \{\boldsymbol{\eta} : n\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})'\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \leq c_\alpha\},$$

with $P(\chi^2_k > c_\alpha) = \alpha$ and in practice using a consistent estimate of $\boldsymbol{\Sigma}$. The *volume* of the region $E_{n,\alpha}$ is

$$\frac{\pi^{k/2}(c_\alpha/n)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}{\Gamma((k+1)/2)}.$$

Therefore, for two such estimators $\widehat{\boldsymbol{\eta}}^{(i)}$, $i = 1, 2$, the ARE given by (3) may be characterized as the limiting ratio of sample sizes at which the *volumes of associated ellipsoidal confidence regions at approximate level* $100(1-\alpha)\%$ converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$.

Under regularity conditions on the model, the maximum likelihood estimator $\widehat{\boldsymbol{\eta}}^{(\text{ML})}$ has a confidence ellipsoid $E_{n,\alpha}$ attaining the *smallest possible volume* and, moreover, lying wholly within that for any other estimator $\widehat{\boldsymbol{\eta}}$.

Connections with testing

Parallel to ARE in estimation as developed here is the notion of *Pitman ARE* for comparison of two hypothesis test procedures. Based on a different formulation, although the central limit theorem is used in common, the Pitman ARE agrees with (2) when the estimator and the hypothesis test statistic are linked, as for example $\overline{X}$ paired with the *t-test*, or Med_n paired with the *sign test*, or HL_n paired with the *Wilcoxon signed-rank test*. See [4], [7], [8], and [11].

Other notions of ARE

As illustrated above with FWARE, several other important approaches to ARE have been developed, typically using either moderate or large deviation theory. For example, instead of asymptotic variance parameters as the criterion, one may compare *probability concentrations* of the estimators in an ε-neighborhood of the target parameter η: $P(|\widehat{\eta}^{(i)} - \eta| > \varepsilon)$, $i = 1, 2$. When we have

$$\frac{\log P(|\widehat{\eta}_n^{(i)} - \eta| > \varepsilon)}{n} \to \gamma^{(i)}(\varepsilon, \eta),\ i = 1, 2,$$

as is typical, then the ratio of sample sizes n_1/n_2 at which these concentration probabilities converge to 0 at the same rate is given by $\gamma^{(1)}(\varepsilon, \eta)/\gamma^{(2)}(\varepsilon, \eta)$, which then represents another ARE measure for the efficiency of estimator $\widehat{\eta}_n^{(2)}$ relative to $\widehat{\eta}_n^{(1)}$. See [11, **1.15.4**] for discussion and [1] for illustration that the variance-based and concentration-based measures need not agree on which estimator is better. For general treatments, see [7], [9] [8], and [11, Chap. 10], as well as the other references cited below. A comprehensive bibliography is beyond the present scope. However, very productive is *ad hoc* exploration of the literature using a modern search engine.

References

[1] Basu, D. (1956). On the concept of asymptotic relative efficiency. *Sankhyā* **17** 193–196.

[2] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* **34** 598–611.

[3] Lehmann, E. L. (1998). *Elements of Large-Sample Theory.* Springer.

[4] Lehmann, E. L. (1998). *Nonparametrics: Statistical Methods Based on Ranks.* Prentice-Hall.

[5] Lehmann, E. L. and Casella, G. (1988). *Theory of Point Estimation*, 2nd edition. Springer.

[6] Monohan, J. F. (1984). Algorithm 616: fast computation of the Hodges-Lehmann location estimator. *ACM Transactions on Mathematical Software (TOMS)* **10** 265–270.

[7] Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests.* Cambridge University Press.

[8] Nikitin, Y. (2010). Asymptotic relative efficiency in testing. *International Encyclopedia of Statistical Sciences.* Springer.

[9] Puhalskii, A. and Spokoiny, V. (1998). On large-deviation efficiency in statistical inference. *Bernoulli* **4** 203–272.

[10] Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88** 1273–1283.

[11] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

[12] Serfling, R. (2002). Efficient and robust fitting of lognormal distributions. *North American Actuarial Journal* **4** 95–109.

[13] Serfling, R. and Wackerly, D. D. (1976). Asymptotic theory of sequential fixed-width confidence interval procedures. *Journal of the American Statistical Association* **71** 949–955.

[14] Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*. Wiley.