ROOT ESTIMATORS FOR THE MEAN OF SKEWED DISTRIBUTIONS

L. J. Ringer, O. C. Jenkins, and H. O. Hartley
Texas A&M University

For a random sample, y_i, i = 1, 2, ..., n, from a finite population of size N, the most commonly used estimator of $\bar{Y}$ is the sample mean $\bar{y}$. In fact, Hartley and Rao [1968] have shown that $\bar{y}$ is the best estimator in any competition in which only unbiased estimators are allowed.

For certain distributions, it may be that there is a biased estimator which in some sense is preferable to $\bar{y}$. In this paper we examine the class of square root estimators when the distributions have positive skewness. The mean square error (MSE) is used as the basis for determining the best estimator.

Square root estimator.

Let y_i, i = 1, 2, ..., n denote a random sample without replacement from a finite population of size N, with all y > 0. Let $u = \sqrt{y}$. The two competitors for estimating the population mean are the usual sample mean

$$\bar{y} = \sum_{i=1}^{n} y_i/n$$

and the square root estimator

$$\hat{y} = C_1\bar{y} + C_2\bar{u}^2$$

where C_1 and C_2 are predetermined constants.

The comparison of the expectations and mean square errors of these two estimators will be facilitated by the use of the finite population cumulants κ_r and the sample k-statistics (e.g. Wishart [1952]) of u, where

$$k_1 = \Sigma u_i/n = \bar{u}$$

$$k_2 = \Sigma(u_i-\bar{u})/(n-1) \quad .$$

In terms of the κ parameters of the distribution of u

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})[\kappa_4 + 4\kappa_{31} + 4\kappa_{211}]$$
$$+ 2\,\{(\frac{n-1}{n})^2(\frac{1}{n-1} - \frac{1}{N-1}) + (\frac{n-1}{n})[\frac{1}{n(n-1)}$$
$$- \frac{1}{N(N-1)}] - (\frac{2n-1}{n(N-1)})(\frac{1}{n} - \frac{1}{N})\}\,\kappa_{22} \quad .$$

which, since $E(\bar{y}) = \bar{Y}$, is also $MSE(\bar{y})$.

Let us first consider the special case, $C_1 + C_2 = 1$, so that the estimator becomes

$$\hat{y} = (1-C)\bar{y} + C\bar{u}^2$$
$$= (1-C)\,\frac{n-1}{n}\,k_2 + k_1^2 \quad .$$

Then, the bias and mean square error of the estimator, $\hat{y}$, are

$$B(\hat{y}) = -\,C\kappa_2(1-1/n)$$

$$MSE(\hat{y}) = C(1-\frac{1}{n})\{C(1-\frac{1}{n})\kappa_2^2 + (\frac{1}{n}-\frac{1}{N})[C(1-\frac{1}{n}) - 2]\kappa_4$$
$$+ [2(C-2)(1-\frac{1}{n})(\frac{1}{n-1}-\frac{1}{N-1}) + \frac{4}{N-1}(\frac{1}{n}-\frac{1}{N})]\kappa_{22}$$
$$- 4(\frac{1}{n} - \frac{1}{N})\kappa_{31}\} + V(\bar{y}) \quad .$$

Note that the bracketed term must be negative for $MSE(\hat{y})$ to be less than $V(\bar{y})$. The $MSE(\hat{y})$ is minimized by choosing $C = C_o$ where

$$C_o = \frac{(\frac{1}{n-1})(\frac{N-n}{N})\;[\kappa_4 + 2\kappa_{22} + 2\kappa_{31}]}{\kappa_2^2 + (\frac{1}{n}-\frac{1}{N})\kappa_4 + 2(\frac{1}{n-1}-\frac{1}{N-1})\kappa_{22}} \;;\quad n > 1 \quad .$$

For N moderately large, there is little loss in looking at limits as $N\to\infty$. The corresponding quantities then become

$$\lim_{N\to\infty} MSE\;(\hat{y}) = \frac{C}{n}(1 - \frac{1}{n})\{[C(1 - \frac{1}{n}) - 2]\kappa_4$$
$$+ [C(n+1) - 4]\kappa_2^2 - 4\kappa_3\kappa_1\} + V(\bar{y})$$

$$\lim_{N\to\infty} V(\bar{y}) = \frac{1}{n}\,\{\kappa_4 + 4\kappa_3\kappa_1 + 4\kappa_2\kappa_1^2 + 2\kappa_2^2\}$$

$$\lim_{N\to\infty} C_o = \frac{(\frac{1}{n-1})[\kappa_4 + 2\kappa_2^2 + 2\kappa_3\kappa_1]}{\frac{1}{n}\,\kappa_4 + \frac{n+1}{n-1}\,\kappa_2^2} \quad .$$

It can be shown that for any distribution for which the third central moment is positive the square root estimator will be more efficient than $\bar{y}$ if and only if $0 < C < 2C_o$.

Example distributions for which the square root estimator is more efficient.

It should be apparent that both C_o and $R = MSE(\hat{y})/V(\bar{y})$ are not affected by changes in the scale of the observations. Therefore, in the discussion to follow the scale parameters of the distributions were set equal to one for simplicity.

To illustrate the usefulness of the square root estimator three families of distributions were examined, ranging from a fairly symmetric gamma distribution to a highly skewed Pareto distribution. The specific distributions are:

Family	Form	Curve Number
	$\frac{1}{6} y^3 e^{-y}$	1
	$\frac{1}{2} y^2 e^{-y}$	2
Gamma	$y e^{-y}$	3
	e^{-y}	4
	$\frac{1}{12} y e^{-\sqrt{y}}$	5
	$\frac{1}{4} \sqrt{y} e^{-\sqrt{y}}$	6
Wishart	$\frac{1}{2} e^{-\sqrt{y}}$	7
	$\frac{1}{2} y^{-\frac{1}{2}} e^{-\sqrt{y}}$	8
	y^{-2}	9
	$2y^{-3}$	10
Pareto	$3y^{-4}$	11
	$4y^{-5}$	12
	$5y^{-6}$	13

For sample sizes $n = 2(1)20$. the maximum value of $C(=2C_o)$ and the value of R at $C = C_o$ are given in Table 1.

Two points are obvious from the table. First, the efficiency of the square root estimator decreases with increasing sample size. Hence, the square root estimator would be appropriate only for small samples. Secondly, for a given value of n, a particular value of C will give near optimum results for a wide class of distributions. Consequently, it is not as important to know the exact distribution of $y_i(u_i)$ but knowledge of the general shape of the frequency distribution should be sufficient.

General square root estimators.

If we allow C_1 and C_2 to be arbitrary constants we find

$$B(\hat{y}) = -[(1-C_1-\frac{C_2}{n})\kappa_2 + (1-C_1-C_2)\kappa_{11}] \quad ,$$

$$V(\hat{y}) = C_1^2(\frac{n-1}{n})^2 V(k_2) + (C_1 + C_2)^2 V(k_1^2)$$

$$+ 2C_1(C_1 + C_2)(\frac{n-1}{n}) \operatorname{Cov}(k_2, k_1^2) \quad ,$$

where

$$V(k_2) = [\frac{1}{n} - \frac{1}{N}]\kappa_4 + 2[\frac{1}{n-1} - \frac{1}{N-1}]\kappa_{22}$$

$$V(k_1^2) = [\frac{1}{n} - \frac{1}{N}][\frac{1}{n^2}\kappa_4 + \frac{4}{n}\kappa_{31} + 4\kappa_{211}]$$

$$+ 2\{(\frac{n-1}{n})[\frac{1}{n(n-1)} - \frac{1}{N(N-1)}]$$

$$- \frac{1}{n(N-1)}[\frac{1}{n} - \frac{1}{N}]\}\kappa_{22}$$

$$\operatorname{Cov}(k_1^2, k_2) = [\frac{1}{n} - \frac{1}{N}][\frac{1}{n}\kappa_4 - \frac{2}{N-1}\kappa_{22} + 2\kappa_3\kappa_1] \quad ,$$

so that

$$MSE(\hat{y}) = V(\hat{y}) + [B(\hat{y})]^2 \quad ,$$

and the optimal values of the coefficients are

$$C_1 = \frac{DE_1 - BE_2}{AD - B^2}$$

$$C_2 = \frac{AE_2 - BE_1}{AD - B^2}$$

where

$$E_1 = (\kappa_2 + \kappa_{11})^2$$

$$E_2 = (\kappa_2 + \kappa_{11})(\frac{\kappa_2}{n} + \kappa_{11})$$

$$A = (\frac{n-1}{n})^2 V(k_2) + V(k_1^2) + 2(\frac{n-1}{n})\operatorname{Cov}(k_2, k_1^2) + E_1$$

$$B = V(k_1^2) + (\frac{n-1}{n})\operatorname{Cov}(k_2, k_1^2) + E_2$$

$$D = V(k_1^2) + (\frac{\kappa_2}{n} + \kappa_{11})^2 \quad .$$

The values of the optimal coefficients for the distributions considered earlier are given in Table 2. In this case, the values are dependent on the form of the distribution. The efficiency ratio of the square root estimator for optimal values of C_1 and C_2 will always be better (smaller) than the efficiency ratio for optimal C under the restricted situation. However, for general C_1 and C_2, the knowledge of the form of the distribution is much more critical for near optimal results.

Conclusions.

A marked gain in efficiency over the sample mean may be obtained, for small sample sizes, by using the square root estimator $\hat{y} = (1-C)\bar{y} + C\bar{u}^2$. While the value of C depends on the sample size and distribution type, for a given n a particular value of C can be determined which will give near optimal results over a wide class of distribution types.

REFERENCES

[1] Hartley, H. O. and Rao, J. N. K., "A New Estimation Theory for Sample Surveys," Biometrika, 55 (1968), 547-557.

[2] Wishart, J., "Moment Coefficients of the k-Statistics in Samples from a Finite Population," Biometrika, 39 (1952), 1-13.

TABLE 1. VALUES OF C FOR WHICH $MSE(\hat{y}) < V(\bar{y})$ FOR THIRTEEN DISTRIBUTIONS; n = 2, 20

Distribution	n= 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(1) $\frac{1}{6}y^3e^{-y}$	**.96	.96	.96	.97	.97	.97	.98	.98	.98	.98	.98	.98	.98	.99	.99	.99	.99	.99	.99
	*2.74	2.06	1.66	1.38	1.18	1.04	.92	.82	.76	.70	.64	.60	.56	.52	.49	.46	.43	.41	.39
(2) $\frac{1}{2}y^2e^{-y}$	.94	.94	.95	.96	.96	.96	.97	.97	.97	.97	.98	.98	.98	.98	.98	.98	.98	.98	.98
	2.77	2.08	1.67	1.39	1.19	1.04	.93	.83	.76	.70	.64	.60	.56	.52	.50	.46	.44	.42	.40
(3) $y\ e^{-y}$	.92	.92	.93	.93	.94	.95	.95	.96	.96	.96	.96	.97	.97	.97	.97	.97	.98	.99	.99
	2.83	2.13	1.70	1.42	1.22	1.07	.95	.85	.78	.71	.65	.61	.57	.53	.50	.47	.45	.43	.41
(4) e^{-y}	.84	.84	.86	.87	.88	.90	.91	.91	.92	.93	.93	.94	.94	.94	.95	.95	.95	.95	.96
	2.98	2.23	1.79	1.50	1.29	1.13	1.01	.91	.83	.76	.70	.65	.61	.57	.54	.51	.48	.46	.43
(5) $\frac{1}{12}ye^{-\sqrt{y}}$	.73	.73	.75	.77	.79	.81	.83	.84	.85	.86	.87	.88	.88	.89	.90	.90	.91	.91	.91
	4.00	3.00	2.45	2.08	1.81	1.61	1.45	1.32	1.21	1.12	1.04	.97	.91	.86	.81	.77	.73	.70	.67
(6) $\frac{1}{4}\sqrt{y}e^{-\sqrt{y}}$	.67	.67	.69	.72	.74	.76	.78	.80	.81	.82	.84	.84	.85	.86	.87	.87	.88	.89	.89
	4.00	3.00	2.46	2.10	1.85	1.65	1.49	1.35	1.25	1.16	1.08	1.01	.95	.90	.86	.82	.77	.74	.71
(7) $\frac{1}{2}e^{-\sqrt{y}}$	.57	.57	.60	.63	.66	.68	.71	.73	.75	.76	.78	.79	.80	.81	.82	.83	.83	.84	.85
	4.00	3.00	2.48	2.41	1.89	1.70	1.55	1.42	1.31	1.22	1.14	1.07	1.01	.95	.91	.86	.82	.79	.75
(8) $\frac{1}{2}y^{-\frac{1}{2}}e^{-\sqrt{y}}$	.40	.40	.43	.47	.50	.53	.56	.58	.60	.62	.64	.66	.68	.69	.70	.71	.72	.73	.74
	4.00	3.00	2.53	2.22	2.00	1.82	1.68	1.57	1.46	1.37	1.29	1.23	1.17	1.11	1.06	1.01	.97	.93	.89
(9) y^{-2} $(y \geq 1)$	.28	.28	.33	.36	.39	.42	.44	.46	.48	.50	.52	.54	.54	.55	.57	.58	.59	.60	.62
	5.26	3.85	3.38	3.03	2.79	2.59	2.43	2.29	2.17	2.07	1.97	1.89	1.81	1.74	1.67	1.62	1.56	1.51	1.46
(10) $2y^{-3}(y \geq 1)$	.22	.22	.23	.24	.26	.27	.29	.30	.32	.33	.34	.36	.37	.38	.39	.40	.41	.42	.43
	6.08	4.56	3.99	3.67	3.45	3.29	3.15	3.04	2.95	2.85	2.77	2.71	2.64	2.57	2.52	2.47	2.41	2.36	2.31
(11) $3y^{-4}(y \geq 1)$	.35	.35	.36	.37	.39	.40	.41	.42	.44	.45	.46	.47	.48	.49	.50	.51	.52	.52	.53
	7.20	5.40	4.72	4.34	4.08	3.89	3.73	3.59	3.48	3.37	3.28	3.19	3.12	3.04	2.97	2.91	2.85	2.79	2.73
(12) $4y^{-5}(y \geq 1)$	.45	.45	.47	.48	.50	.52	.53	.55	.56	.57	.59	.60	.61	.62	.63	.64	.65	.66	.66
	10.09	7.57	6.56	5.96	5.54	5.21	4.94	4.76	4.51	4.33	4.17	4.03	3.89	3.77	3.65	3.55	3.45	3.35	3.27
(13) $5y^{-6}(y \geq 1)$	.51	.51	.52	.54	.56	.57	.59	.61	.62	.63	.65	.66	.67	.68	.69	.70	.71	.71	.72
	12.89	9.66	8.32	7.51	6.93	6.48	3.11	5.79	5.51	5.27	5.05	4.84	4.66	4.49	4.34	4.20	4.07	3.94	3.82

*Maximum values of C for which $MSE(\hat{y}) < V(\bar{y})$.
Optimum value is ½ the listed value.

**Value of $R = \frac{MSE(\hat{y})}{V(\bar{y})}$ at C_0.
Minimum value of C for which $EMS(y) < V(\bar{y})$ is zero in each case.

TABLE 2. VALUES OF C_1 AND C_2 FOR OPTIMUM EFFICIENCY FOR THIRTEEN DISTRIBUTIONS; n = 2, 20

Distribution	n = 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
(1) $\frac{1}{6}y^3e^{-y}$	.89	.92	.94	.95	.96	.96	.97	.97	.97	.97	.98	.98	.98	.98	.98	.98	.98	.98	.98	C_1
	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_2
(2) $\frac{1}{2}y^2e^{-y}$	.86	.90	.92	.94	.95	.95	.96	.96	.97	.97	.97	.98	.98	.98	.98	.98	.98	.98	.98	C_1
	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_2
(3) $y\ e^{-y}$	.80	.86	.89	.91	.92	.93	.94	.95	.95	.96	.96	.96	.96	.97	.97	.97	.98	.98	.98	C_1
	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_2
(4) e^{-4}	.67	.75	.80	.83	.86	.88	.89	.90	.91	.92	.92	.93	.93	.94	.94	.94	.95	.95	.95	C_1
	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_2
(5) $\frac{1}{12}ye^{-\sqrt{y}}$	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_1
	.73	.85	.94	.99	1.03	1.05	1.08	1.09	1.11	1.12	1.13	1.14	1.15	1.15	1.16	1.16	1.17	1.17	1.18	C_2
(6) $\frac{1}{4}\sqrt{y}e^{-\sqrt{y}}$	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_1
	.67	.82	.91	.98	1.03	1.06	1.09	1.12	1.14	1.15	1.17	1.18	1.19	1.20	1.20	1.21	1.22	1.22	1.23	C_2
(7) $\frac{1}{2}e^{-\sqrt{y}}$	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_1
	.57	.75	.87	.96	1.03	1.08	1.12	1.16	1.19	1.21	1.23	1.25	1.26	1.28	1.29	1.30	1.31	1.32	1.33	C_2
(8) $\frac{1}{2}y^{-\frac{1}{2}}e^{-\sqrt{y}}$	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	C_1
	.40	.60	.80	.90	1.00	1.09	1.16	1.23	1.28	1.33	1.37	1.41	1.44	1.47	1.50	1.52	1.54	1.56	1.58	C_2
(9) y^{-2} $(y\geq 1)$	-1.17	-1.17	-1.18	-1.16	-1.15	-1.14	-1.13	-1.12	-1.11	-1.10	-1.10	-1.09	-1.09	-1.09	-1.08	-1.08	-1.07	-1.07	-1.07	C_1
	1.97	2.29	2.49	2.54	2.60	2.64	2.67	2.69	2.71	2.73	2.74	2.75	2.76	2.76	2.77	2.78	2.78	2.79	2.80	C_2
(10) $2y^{-3}(y\geq 1)$	-2.09	-1.48	-1.25	-1.13	-1.06	-1.01	- .98	- .95	- .93	- .91	- .90	- .88	- .87	- .87	- .86	- .85	- .85	- .84	- .84	C_1
	3.10	2.56	2.37	2.27	2.21	2.17	2.14	2.11	2.10	2.08	2.07	2.06	2.05	2.05	2.04	2.04	2.03	2.03	2.02	C_2
(11) $3y^{-4}(y\geq 1)$	-2.68	-1.88	-1.59	-1.45	-1.36	-1.30	-1.26	-1.23	-1.20	-1.18	-1.17	-1.15	-1.14	-1.13	-1.12	-1.11	-1.11	-1.10	-1.10	C_1
	3.69	2.92	2.64	2.51	2.47	2.37	2.33	2.30	2.27	2.25	2.24	2.23	2.22	2.21	2.20	2.19	2.18	2.18	2.17	C_2
(12) $4y^{-4}(y\geq 1)$	-4.37	-3.25	-2.85	-2.64	-2.52	-2.43	-2.37	-2.33	-2.29	-2.20	-2.24	-2.22	-2.20	-2.19	-2.18	-2.16	-2.15	-2.15	-2.14	C_1
	5.39	4.29	3.90	3.69	3.59	3.49	3.43	3.38	3.35	3.32	3.30	3.28	3.26	3.25	3.23	3.22	3.21	3.21	3.20	C_2
(13) $5y^{-6}(y\geq 1)$	-6.01	-4.61	-4.09	-3.83	-3.67	-3.56	-3.48	-3.42	-3.38	-3.34	-3.31	-3.28	-3.26	-3.24	-3.23	-3.21	-3.20	-3.19	-3.18	C_1
	7.04	5.64	5.13	4.87	4.71	4.60	4.53	4.47	4.42	4.39	4.36	4.33	4.31	4.29	4.27	4.26	4.25	4.24	4.23	C_2