AN EMPIRICAL STUDY OF ALTERNATIVE ESTIMATORS FOR μ AND σ^2

Donald T. Searls, Westat Research, Inc.

In two previous papers [1], [2] estimators for μ were developed which tended to minimize the effect of large true observations occurring in small samples. Proofs that these estimators could have smaller mean-squared errors than $\bar{y}$ were presented.

This paper will consider analagous estimators for σ^2, however since the proofs become extremely cumbersome the gains will be demonstrated with an empirical sampling study.

The first estimator considered is the one where observations larger than a predetermined cutoff point, t, are discarded. This procedure leads to the estimators $\bar{y}_1$, and s_1^2 .

$$\bar{y}_1 = \frac{\sum_{j=1}^{r} y_j}{r} \qquad (y_j < t)$$

$$(r > 0)$$

$$= t ; \qquad (r = 0)$$

$$s_1^2 = \frac{\sum_{j=1}^{r} (y_j - \bar{y}_1)^2}{r - 1} , \qquad (y_j < t)$$

$$(r \geq 2)$$

$$= 0 , \qquad (r < 2)$$

where r is the number of observations less than t.

The next estimators are formed by substituting the value of t for those observations greater than t. For the estimator of σ^2 only one t is used.

$$\bar{y}_t = \frac{\sum_{j=1}^{r} y_j + (n - r)t}{n} \qquad (0 \leq r \leq n)$$

$$(y_j \leq t)$$

$$\text{Let } T = \sum_{j=1}^{r} y_j + t ,$$

$$S = \sum_{j=1}^{r} y_j^2 + t^2 .$$

then

$$s_t^2 = \frac{(r + 1) S - T^2}{r(r + 1)} , \qquad (r \geq 1)$$

$$= 0 . \qquad (r = 0)$$

If one is unwilling to discard observations an optimum weight for the observations [2] and squared deviations can be derived.

$$\bar{y}_w = [1/(n + v^2)] \sum_{j=1}^{n} y_j$$

where $v^2 = \sigma^2/\mu^2$.

A similar development for estimating σ^2 gives:

$$s_w^2 = W \sum_{j=1}^{n} (y_j - \bar{y})^2$$

$$MSE\ (s_w^2) = W^2 (n-1)^2 (\mu_4 - \sigma^4)/n + \sigma^4[1-(n-1)W]^2 .$$

Differentiating with respect to W and solving gives approximately

$$W = 1/\ [(n-1) + \beta_2] .$$

Terms with small contributions as n increases have been deleted.

Hence

$$s_w^2 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{(n-1) + \beta_2}$$

where $\beta_2 = \mu_4/\sigma^4$.

Since $\beta_2 = 3$ for the normal distribution,

$$s_w^2 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n+2}$$ for this situation.

Positively skewed populations were used to demonstrate the effectiveness of the alternatives. See table 1 for the tabulation of the 1000 observations obtained from 200 samples of size 5 from each population.

Table 1. Number of observations by intervals and other characteristics for populations used.

	Distribution	
Interval	I	II
0-1	667	134
1-2	187	562
2-3	68	171
3-4	41	75
4-5	19	38
5-6	9	9
6-7	4	5
7-8	4	2
8-9	1	2
9+	0	2
Total	1000	1000

Parameter	I	II
μ	.959	1.866
σ^2	1.556	1.328
CV	1.301	.618

The 1000 observations were considered as the complete populations for comparison purposes. Distribution I has a coefficient of variation greater than one (1.301) and distribution II has a coefficient of variation less than one (.618) .

Tables 2 and 3 present the mean-squared errors for different cutoff points and sample sizes for the two distributions. Results for samples of sizes 10 and 20 were inferred from the results for samples of size 5 by adding together the bias and the variance with the appropriate division.

Figures 1 and 2 demonstrate graphically the gains achieved with the alternatives s_1^2 and s_t^2 for samples of size 20. Similar patterns are shown in the tables for $\bar{y}_1$ and $\bar{y}_t$.

It is evident that s^2_t and $\bar{y}_t$ achieve gains over a wider region than do s_1^2 and $\bar{y}_1$. Also, lower mean-squared errors are achieved by s_t^2 and $\bar{y}_t$. However s_1^2 and $\bar{y}_1$ are superior in the region which would normally be of the most interest and they are simpler to work with since they are the estimators formed by ignoring the large observations and using only those remaining. For these reasons it would appear to be preferable to use s_1^2 and $\bar{y}_1$ when sampling from distributions of the types used in this study. Of course if no information is available beforehand as to what might constitute a "large" observation the choosing of the cutoff point becomes very difficult. The results indicate that gains are achieved for a rather large region, however if a mistake is made and the point is chosen too small, rather disastrous consequences are obtained. In practice the sampler frequently has some very general ideas concerning minimum and maximum possible values and these plus information concerning the general shape of the distribution will generally provide sufficient information to intelligently pick a t value.

Figure 3 presents the distribution of s_t^2 (t=4) and s^2 for the 200 samples of size 5 from distribution II.

REFERENCES

[1] Searls, Donald T., "Some Alternative Estimators for a Population Mean," 1964 Social Statistics Proceedings of the American Statistical Association.

[2] Searls, Donald T., " The Utilization of a Known Coefficient of Variation in the Estimation Procedure, " Journal of the American Statistical Association, 59(1964), 1225-1226.

Table 2. Mean-squared errors for alternative estimators of μ and σ^2 for distribution I.

n	Cutoff point t	Mean squared errors					
		$\bar{y}$	$\bar{y}_1$	$\bar{y}_t$	s^2	s^2_1	s^2_t
5	9	.319	.319	.319	4.363	4.363	4.363
	7	.319	.273	.309	4.363	2.519	3.524
	6	.319	.264	.296	4.363	1.880	2.799
	5	.319	.233	.274	4.363	1.348	1.879
	4	.319	.216	.244	4.363	1.141	1.174
	3	.319	.209	.200	4.363	1.265	.864
	2	.319	.244	.156	4.363	1.633	1.178
	1	.319	.465	.222	4.363	2.201	1.930
	0	.319	.920	.920	4.363	2.421	2.421
10	9	.160	.160	.160	2.181	2.181	2.181
	7	.160	.137	.155	2.181	1.287	1.764
	6	.160	.134	.148	2.181	1.003	1.410
	5	.160	.121	.138	2.181	.819	.971
	4	.160	.122	.123	2.181	.879	.683
	3	.160	.143	.105	2.181	1.182	.691
	2	.160	.207	.100	2.181	1.609	1.142
	1	.160	.450	.204	2.181	2.198	1.927
	0	.160	.920	.920	2.181	2.421	2.421
20	9	.080	.080	.080	1.091	1.091	1.091
	7	.080	.069	.077	1.091	.670	.883
	6	.080	.069	.074	1.091	.565	.711
	5	.080	.066	.069	1.091	.554	.517
	4	.080	.076	.063	1.091	.748	.438
	3	.080	.110	.058	1.091	1.141	.605
	2	.080	.189	.072	1.091	1.597	1.124
	1	.080	.442	.196	1.091	2.197	1.925
	0	.080	.920	.920	1.091	2.421	2.421

Table 3. Mean-squared errors for alternative estimators of μ and σ^2 for distribution II.

n	Cutoff point t	Mean squared errors $\bar{y}$	$\bar{y}_1$	$\bar{y}_t$	s^2	s^2_1	s^2_t
5	11	.240	.240	.240	3.941	3.941	3.941
	9	.240	.221	.236	3.941	2.319	3.332
	8	.240	.212	.231	3.941	1.718	2.824
	7	.240	.205	.220	3.941	1.278	2.037
	6	.240	.197	.209	3.941	.954	1.449
	5	.240	.173	.194	3.941	.803	.954
	4	.240	.164	.163	3.941	.839	.649
	3	.240	.201	.129	3.941	1.106	.755
	0	.240	3.483	3.483	3.941	1.764	1.764
10	11	.120	.120	.120	1.971	1.971	1.971
	9	.120	.110	.118	1.971	1.168	1.666
	8	.120	.106	.115	1.971	.882	1.414
	7	.120	.104	.110	1.971	.681	1.026
	6	.120	.101	.105	1.971	.557	.743
	5	.120	.091	.097	1.971	.526	.523
	4	.120	.103	.084	1.971	.721	.451
	3	.120	.164	.077	1.971	1.070	.706
	0	.120	3.483	3.483	1.971	1.764	1.764
20	11	.060	.060	.060	.985	.985	.985
	9	.060	.055	.059	.985	.593	.834
	8	.060	.054	.058	.985	.464	.708
	7	.060	.053	.055	.985	.383	.520
	6	.060	.052	.053	.985	.359	.390
	5	.060	.050	.049	.985	.386	.307
	4	.060	.072	.044	.985	.662	.352
	3	.060	.141	.051	.985	1.051	.681
	0	.060	3.483	3.483	.985	1.764	1.764

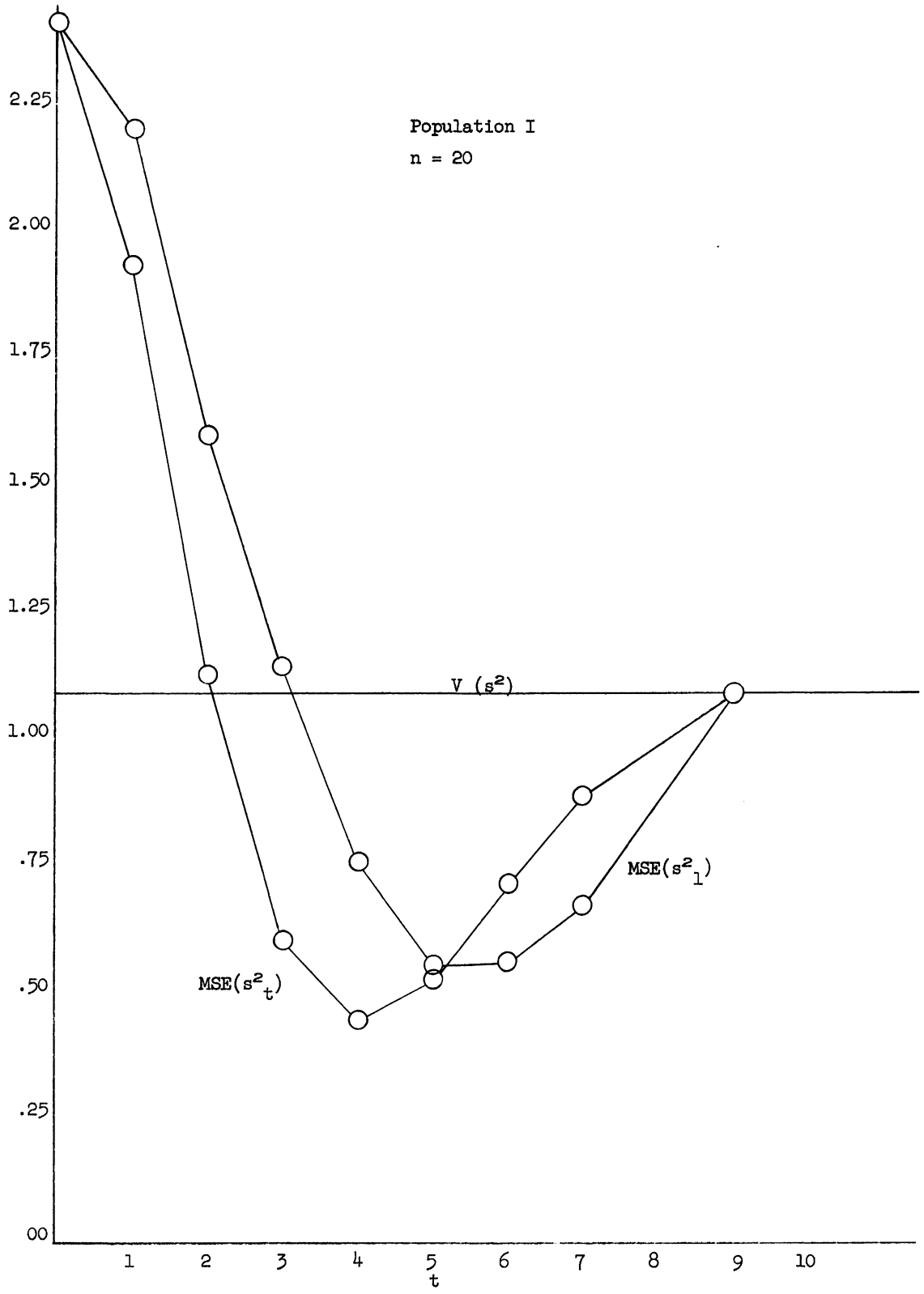

Figure 1. Mean-squared errors versus cutoff point - population I.

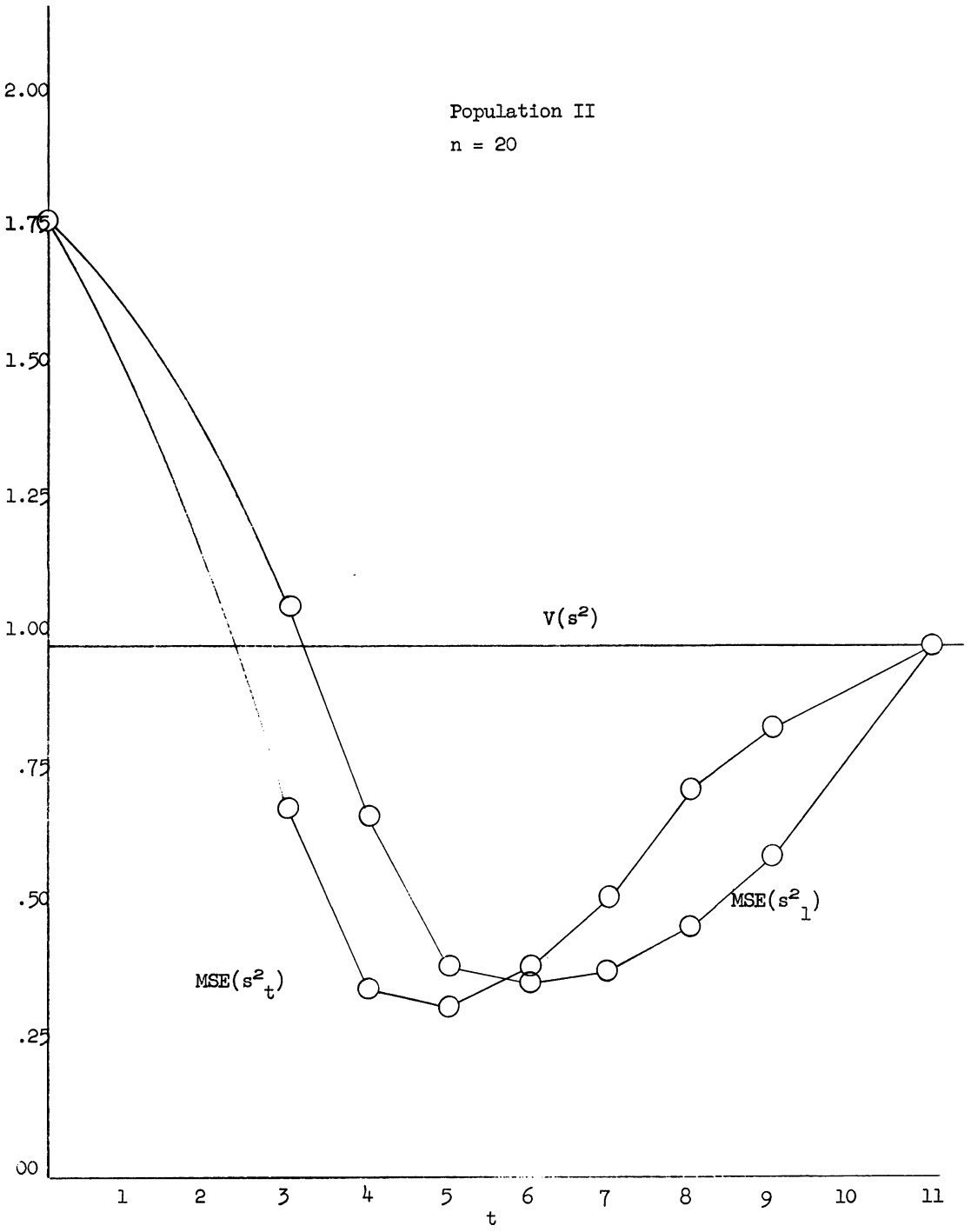

Figure 2. Mean-squared errors versus cutoff point - population II.

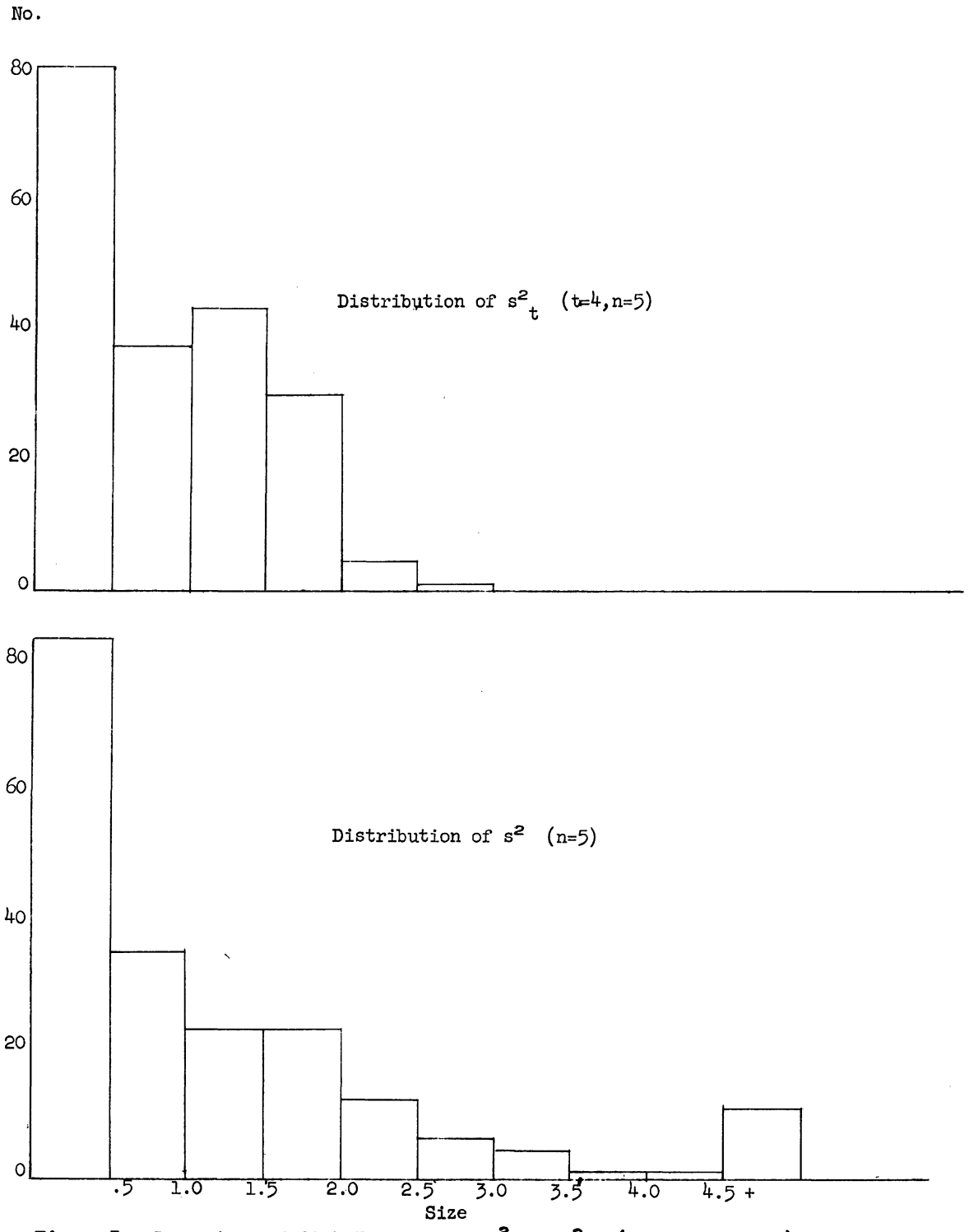

Figure 3. Comparison of distributions of s_t^2 and s^2. (Distribution II)