

A NEW FORMULA FOR PREDICTING THE SHRINKAGE OF THE COEFFICIENT OF MULTIPLE CORRELATION

By

DR. R. J. WHERRY
Cumberland University, Lebanon, Tennessee

With the perfection of the Doolittle Method for the solution of the constant values necessary for the multiple correlation and prediction technique, we may expect a constant increase in the use of this method in statistical practice. Theoretical statisticians have recognized for some time however that the multiple correlation coefficient, derived from a large number of independent variables, is apt to be deceptively large due to chance factors. When prediction equations derived in this manner are applied to subsequent sets of data, there is apt to be a rather large shrinkage in the resulting correlation coefficient obtained, as compared with the original observed multiple correlation coefficient. In order to avoid over optimism it is necessary to have some equation which will predict the most probable value of this shrinkage. The development of such a formula is the purpose of this paper.

The most promising formula of this type so far developed is the B. B. Smith formula, presented by M. J. B. Ezekial at the December, 1928, meeting of the American Mathematical Society held at Chicago. This formula is

$$(1) \quad \bar{R}^2 = 1 - \frac{1 - R^2}{1 - \frac{M}{N}} = \frac{NR^2 - M}{N - M}$$

where \bar{R} = the estimated correlation obtaining in the universe
 R = the observed multiple correlation coefficient
 M = the number of *independent* variables
 N = the number of observations (the statistical population).

This formula was evidently developed by B. B. Smith by an application of the method of least squares as follows (the derivation is that of the author, since he could not find it given elsewhere):

The customary formula for the coefficient of multiple correlation may be written in the form

$$(2) \quad R^2 = 1 - \frac{S_o^2}{\sigma_o^2}$$

where

$$(3) \quad S_o^2 = \frac{\sum v^2}{N}$$

where

$$(4) \quad v = x_o - \bar{x}_o$$

The method of least squares, however, says that the most probable value of the standard error of estimate is not that given in equation (3) but

$$(5)^1 \quad \bar{S}_o^2 = \frac{\sum v^2}{N-M} = \frac{N}{N-M} \cdot S_o^2$$

Now, if we substitute the value of (5) in place of (3) in equation (2), we have at once

¹See Merriman, *Method of Least Squares*, John Wiley & Sons, London, 8th Edition, pp. 80-82. Also see derivation later in this paper.

$$(6) \quad \bar{R}^2 = 1 - \frac{s_o^2}{\sigma_o^2} \cdot \frac{N}{N-M}$$

and since, by (2) above, we have $\frac{s_o^2}{\sigma_o^2}$ equal to $(1 - R^2)$, we have

$$(7) \quad \bar{R}^2 = 1 - \frac{N(1-R^2)}{N-M} = 1 - \frac{1-R^2}{1-\frac{M}{N}}$$

which is, exactly, the B. B. Smith formula (1).

This formula has been widely used during the last few years, but up until recently had not been subjected to much critical examination. However, in a recent article in the *Journal of Educational Psychology*¹, S. C. Larson actually tested the formula empirically on some data obtained from the Mississippi Survey conducted by M. V. O'Shea, obtaining the results indicated in the tables and graphs below, and on the basis of which he reached the following conclusion:

"The Smith Shrinkage-Reduction formula parallels all of the empirical findings but quite consistently gives values which are in excess of those obtained under present experimental conditions." This meant that the Smith formula predicted shrinkages consistently greater than those actually obtained.

It was in view of this reported empirical difference that the writer started his attempt to derive the Smith formula and hit on the method given above. The question at once arose in the writer's mind as to why, when the standard error of estimate had been corrected to correspond to the most probable value by a least squares criterion, the standard deviation of the dependent variable had not been treated in the same fashion.

¹"The Shrinkage of the Coefficient of Multiple Correlation," Jan., 1931, pp. 45-55.

Merriman, whose formula we used above in correcting the standard error of estimate (5), likewise, and by identical reasoning, shows that the most probable value of the standard deviation of the dependent variable existing in the universe, should really be represented by the following relationships:

Where

$$(8) \quad \sigma_o^2 = \frac{\sum x_o^2}{N}$$

we find

$$(9) \quad \bar{\sigma}_o^2 = \frac{\sum x_o^2}{N-1} = \sigma_o^2 \cdot \frac{N}{N-1}$$

which reduces formula (6) to the form

$$(10a) \quad \bar{R}^2 = 1 - \frac{s_o^2}{\sigma_o^2} \cdot \frac{\frac{N}{N-1}}{\frac{N}{N-1}}$$

and when the same substitution is made as in step (7) above, we have

$$(10b) \quad \bar{R}^2 = \frac{(N-1)R^2 - (M-1)}{N-M}$$

which is, by a more correctly applied criterion of least squares, the formula we have been seeking, and is a closer approximation than that given by the Smith formula.

The reasons for the substitutions made above in our formulae may not be entirely clear to all readers, so we now present the derivations of the formulae given in (5) and (9) above. The derivations given here are directly adapted from those of Merriman referred to above, but have been translated into the customary statistical notation whenever possible.

First, let us consider the derivation of the value in (9). As

stated in (8) the most customary form of Sigma is

$$(8) \quad \sigma_o^2 = \frac{\sum x_o^2}{N}$$

where

$$(11) \quad x_o = x - M_x.$$

Each value x_o has a certain error, however, due to the fact that the value of the mean is merely the most probable value, not the true value. So for each x_o value there is a small unknown error δx_o , so that if we take \bar{x}_o to be the true value of a deviation we have

$$(12) \quad \bar{x}_o = x_o + \delta x_o$$

and, squaring and summing, disregarding the terms involving second power delta terms as small in comparison with the first power terms, we have

$$(13) \quad \sum \bar{x}_o^2 = \sum x_o^2 + 2\sum x_o \delta x_o$$

Now, by the laws of probability, we know that the probability of the occurrence of an error \bar{x}_o , whose measure of precision is "h," is

$$(14) \quad \Pi = h d\bar{x} \pi^{-\frac{1}{2}} e^{-\bar{x}^2 h^2}$$

multiplying both sides of this equation by \bar{x}^2 and summing between the limits plus and minus infinity, we have

$$(15) \quad \sum \Pi \bar{x}^2 = \int_{-\infty}^{+\infty} h \bar{x}^2 \pi^{-\frac{1}{2}} e^{-h^2 \bar{x}^2} d\bar{x} = \frac{1}{2h^2}$$

and since $\sum \Pi \bar{x}^2$ is the same as $\frac{\sum \bar{x}^2}{N}$, since in our work we assume the weight of each value \bar{x} , for each of the N observations, to be $\frac{1}{N}$, we have

$$(16) \quad \frac{\sum \bar{x}^2}{N} = \frac{1}{2h^2}$$

or

$$(16a) \quad \sum \bar{x}^2 = \frac{N}{2h^2}$$

Likewise, if we let

$$(17) \quad 2\sum x_o \delta x_o = u^2$$

the probability of the system of errors, u^2 , is

$$(18) \quad \Pi' = h du \pi^{-\frac{1}{2}} e^{-u^2 h^2}$$

and the mean of all of the possible values of u^2 is

$$(19) \quad \frac{h}{\pi^{\frac{1}{2}}} \int_{-\infty}^{+\infty} u^2 e^{-h^2 u^2} du = \frac{1}{2h^2}$$

and this must be taken as the best attainable value of u^2 . But it was shown that the quantity $\frac{1}{2h^2}$ is equal to $\frac{\sum \bar{x}^2}{N}$ (16). Hence

$$(20) \quad \sum \bar{x}^2 = \sum x^2 + \frac{\sum \bar{x}^2}{N}$$

from which

$$(9) \quad \bar{\sigma}_o^2 = \frac{\sum \bar{x}^2}{N} = \frac{\sum x^2}{N-1} = \sigma_o^2 \cdot \frac{N}{N-1}$$

which was to be proved.

To derive (5) we proceed in much the same manner. After our normal equations have been solved for the most probable values of $\beta_{01}, \beta_{02}, \beta_{03}, \dots, \beta_{0m}$ for our set of data, we know that these are not the true values, but that they err by small unknown corrections $\delta\beta_{01}, \delta\beta_{02}, \delta\beta_{03}, \dots, \delta\beta_{0m}$, the corresponding true values for the universe being $(\beta_{01} + \delta\beta_{01}), (\beta_{02} + \delta\beta_{02}), (\beta_{03} + \delta\beta_{03}), \dots, (\beta_{0m} + \delta\beta_{0m})$.

Now, if we substitute the most probable values of the Betas in our original observation equations, they will not reduce to zero, but will leave small residuals v_1, v_2, \dots, v_N , thus

$$\bar{x}_{01} - x_{01} = \beta_{01} x_{11} + \beta_{02} x_{21} + \beta_{03} x_{31} + \dots + \beta_{0m} x_{m1} - x_{01} = v_1$$

$$\bar{x}_{02} - x_{02} = \beta_{01} x_{12} + \beta_{02} x_{22} + \beta_{03} x_{32} + \dots + \beta_{0m} x_{m2} - x_{02} = v_2$$

.....

$$\bar{x}_{0M} - x_{0M} = \beta_{01} x_{1N} + \beta_{02} x_{2N} + \beta_{03} x_{3N} + \dots + \beta_{0m} x_{mN} - x_{0N} = v_N$$

while if the corresponding true values be substituted, we obtain

$$(\beta_{01} + \delta\beta_{01})x_{11} + (\beta_{02} + \delta\beta_{02})x_{21} + \dots + (\beta_{0m} + \delta\beta_{0m})x_{m1} - x_{01} = \bar{v}_1$$

$$(\beta_{01} + \delta\beta_{01})x_{12} + (\beta_{02} + \delta\beta_{02})x_{22} + \dots + (\beta_{0m} + \delta\beta_{0m})x_{m2} - x_{02} = \bar{v}_2$$

.....

$$(\beta_{01} + \delta\beta_{01})x_{1N} + (\beta_{02} + \delta\beta_{02})x_{2N} + \dots + (\beta_{0m} + \delta\beta_{0m})x_{mN} - x_{0N} = \bar{v}_N$$

Subtracting each of the former equations from the latter, we obtain

$$v_1 + \delta\beta_{01} x_{1_1} + \delta\beta_{02} x_{2_1} + \dots + \delta\beta_{0m} x_{m_1} = \bar{v}_1$$

$$v_2 + \delta\beta_{01} x_{1_2} + \delta\beta_{02} x_{2_2} + \dots + \delta\beta_{0m} x_{m_2} = \bar{v}_2$$

.....

$$v_N + \delta\beta_{01} x_{1_N} + \delta\beta_{02} x_{2_N} + \dots + \delta\beta_{0m} x_{m_N} = \bar{v}_N$$

Now the principle of least squares provides that $\sum \bar{v}^2$ shall be made a minimum to give the most probable values of β_{01} , β_{02} , β_{0m} , and by the solution of the normal equations by the Doolittle method its minimum value is found to be $\sum v^2$. From the residual equations we may find the relationship existing between the values $\sum \bar{v}^2$ and $\sum v^2$. Thus, if we square each equation immediately above and then summate we have (if we neglect squares and products of the delta values as small in comparison with the first powers):

$$(21) \quad \sum v^2 + 2\delta\beta_{01} \sum x_{1_1} v + 2\delta\beta_{02} \sum x_{2_1} v + \dots + 2\delta\beta_{0m} \sum x_{m_1} v = \sum \bar{v}^2$$

which we may write as

$$(22) \quad \sum v^2 + u_1^2 + u_2^2 + \dots + u_m^2 = \sum \bar{v}^2$$

Now, by analogous reasoning to that in steps (14), (15), and

(16), we may set

$$(23) \quad \Sigma \bar{v}^2 = \frac{N}{2h^2}$$

Further, if there be but one independent variable, there will be but one $\dot{2}\delta\beta_{ox} \Sigma x_x v$ and its value by the same process used in steps (18) and (19) can be shown to be

$$(24) \quad u_x^2 = \frac{1}{2h^2}$$

and since that is true whichever unknown quantity be considered, the values of each u_x^2 value must be $\frac{1}{2h^2}$; and as there are M of these values the above equation (22) becomes

$$\Sigma v^2 + \frac{M}{2h^2} = \frac{N}{2h^2}$$

from which

$$(25) \quad h = \sqrt{\frac{N-M}{2\Sigma v^2}}$$

Therefore, from the constant relationship which exists between the value "h" and the Probable Error, we have

$$(26) \quad P.E._{\bar{v}} = 0.6745 \sqrt{\frac{\Sigma v^2}{N-M}}$$

and therefore, by the relationship existing between the probable error and the standard deviation we have at once

$$(5) \quad \sigma_{\bar{v}}^2 = \bar{s}_0^2 = \frac{\Sigma v^2}{N-M}$$

which was to be proved.

The next step was to test out the formula empirically. This was done by using Larson's material, with the results indicated in the tables below, and in the graphs which show the same

set of facts, but which make the results much more apparent.

An inspection of the tables and graphs will show at once that the new formula predicts what will actually happen much more accurately than the Smith formula did. In graph 1, for example, the agreement is so good that the results appear almost to have been a regression line fit to the particular set of data.

It was to have been expected that if the formula actually predicted the most probable values of the correlations obtaining in the universe that the errors incurred by the use of the formula would be normally distributed around zero as a mean value. Graph 3 presents a comparison of the error curves obtained by use of the Smith and the Wherry prediction formulae, together with an approximation to the normal curve. As a further and more scientific check the criteria for a normal curve as set forth by Rietz¹ were applied to the data. His criteria are

$$\mu_1 = 0, \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0, \beta_2 = \frac{\mu_4}{\mu_2^2} = 3; \quad \text{where } \mu_n = \frac{\sum x^n}{N}$$

The results for the two formulae are given below.

(Results based on an expectancy of zero)

	Smith Formula	Wherry Formula
μ_1	.00138	.00038
β_1	.223	.025
β_2	3.004	3.703

¹Rietz, H. L. *Mathematical Statistics*, Carus Mathematical Monograph No. 3, Mathematical Association of America, Chicago 1927, pp. 58-59.

It is apparent therefore that the Wherry formula gave much better results for both the first criterion (mean error) and the second criterion (skewness), but that the excess was greater for the Wherry formula than for the Smith formula. However, one cannot quarrel too much with getting errors actually smaller than would be expected by assuming normality. Even this superiority is seen to be fictitious if the distributions are measured from their own means rather than from an expected mean of zero. When this is done, which is the manner in which the criteria are customarily used, we have

(Results based on means of distributions)

	Smith Formula	Wherry Formula
μ_1	.000	.000
β_1	1.712	.025
β_2	5.524	3.753

Thus, we find that the Smith distribution has, in reality, even a greater excess than does the Wherry formula, but has it at a point farther removed from the desired value.

SUMMARY AND CONCLUSIONS

1. Larson has shown that the theoretically expected shrinkage is an empirical fact.
2. Larson has shown that the Smith formula, when tested empirically, consistently over-estimates this shrinkage as determined empirically.
3. It has been demonstrated that the new Wherry formula,

both by a least squares criterion and by actual application, is more nearly true than the corresponding Smith formula.

4. The correct formula for the shrunken coefficient of multiple correlation is

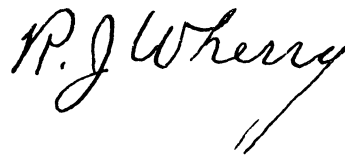
$$\bar{R}^2 = \frac{(N-1)R^2 - (M-1)}{N-M}$$

where \bar{R} = the estimated correlation obtaining in the universe

R = the observed coefficient of multiple correlation

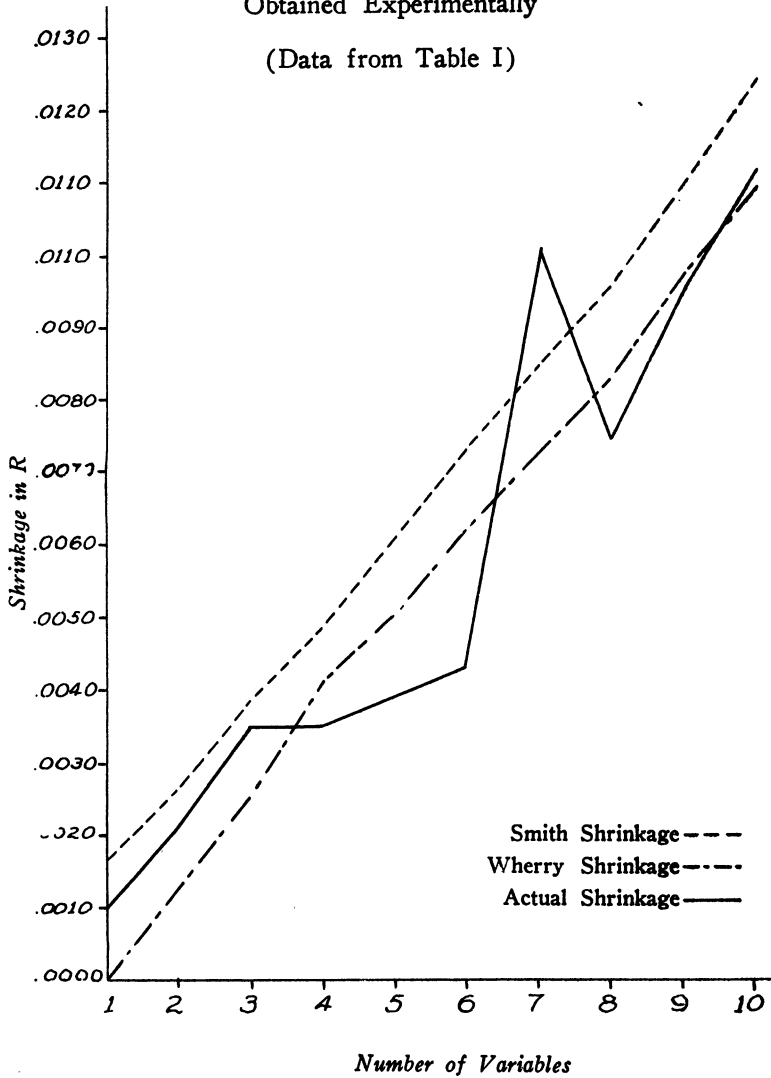
M = the number of *independent* variables

and N = the number of observations (statistical population).



GRAPH 1.

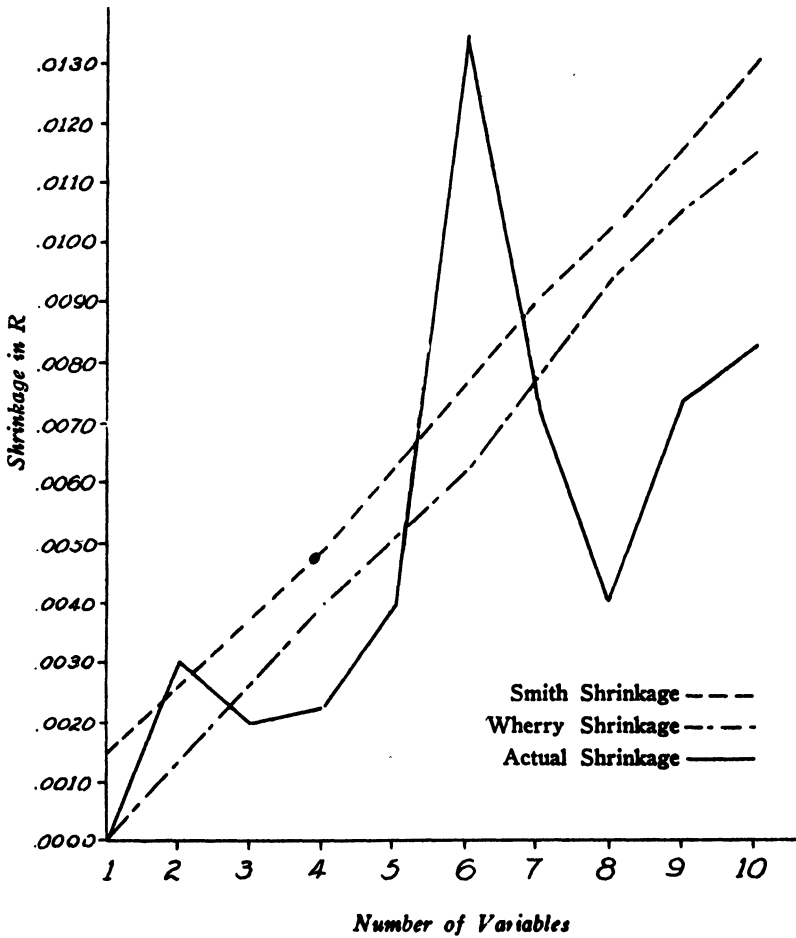
Shrinkage as Obtained by Use of the Formulae and Also as
Obtained Experimentally



GRAPH 2

Shrinkage as Obtained by Use of the Formulae and Also as
Obtained Experimentally

(Data from Table II)



GRAPH 3

Ogive Showing the Distribution of Error in Predicting Shrinkage

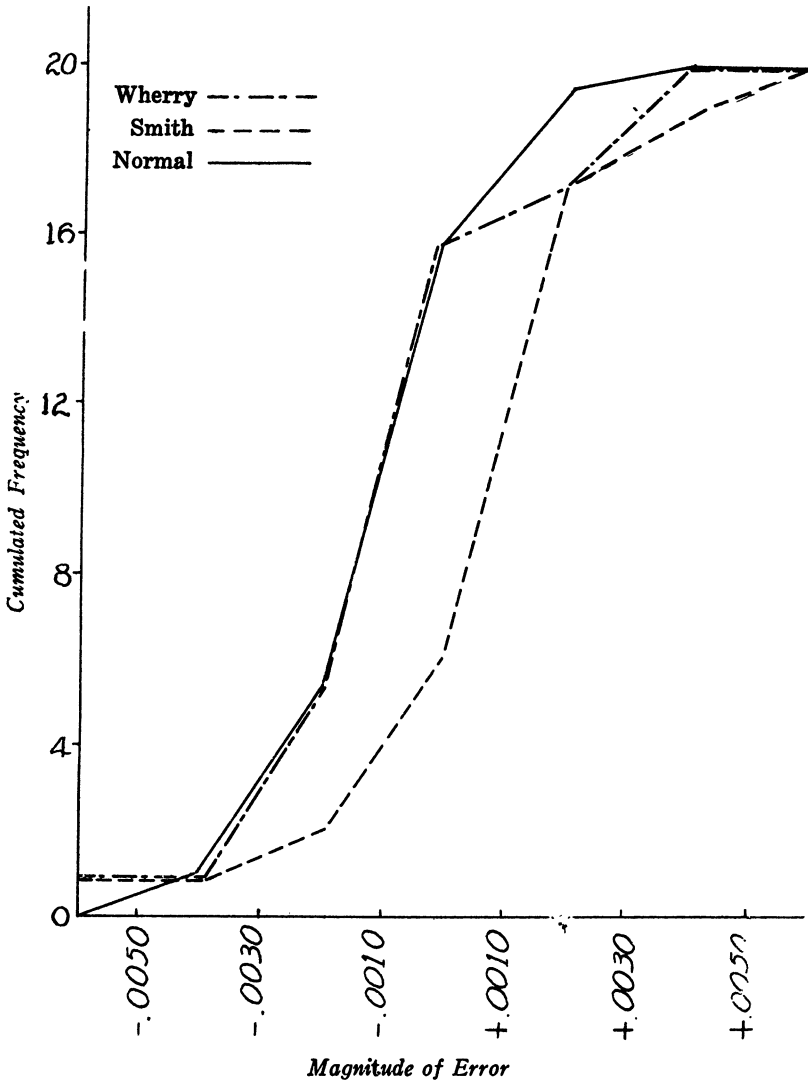


TABLE I*
 Showing the Actual Shrinkage in \mathcal{R} Found When the Prediction Equation Found on One Group of Subjects Is Applied to a Comparable Group, Together with the Shrinkage of \mathcal{R} as Indicated by the Smith and Wherry Formulae. The Statistical Population ($\gamma\gamma$) Is 200 Throughout.

$\gamma\gamma$	1	2	3	4	5	6	7	8	9	10
\mathcal{R}	.7042	.7794	.7834	.7872	.7880	.7907	.7929	.7941	.7944	.7945
Actual Shrinkage	.0000	.0021	.0036	.0036	.0060	.0044	.0102	.0075	.0097	.0113
Shrinkage by Smith formula	.0017	.0026	.0038	.0049	.0062	.0074	.0085	.0097	.0110	.0123
Shrinkage by Wherry formula	.0000	.0013	.0025	.0040	.0049	.0062	.0073	.0085	.0098	.0111

*The article by Larson reported the values for the Smith formula erroneously, due to a misconception of the meaning of $\gamma\gamma$. Those in the present tables are the correct values.

TABLE II

Showing for a Second Set of Groups the Same Facts as Obtain in Table I

m	1	2	3	4	5	6	7	8	9	10
R	.7402	.7759	.7813	.7826	.7847	.7858	.7859	.7863	.7868	.7869
Actual Shrinkage	.0000	.0031	.0019	.0023	.0041	.0133	.0073	.0042	.0074	.0083
Shrinkage by Smith formula	.0015	.0026	.0038	.0051	.0063	.0076	.0089	.0102	.0115	.0129
Shrinkage by Wherry formula	.0000	.0013	.0026	.0038	.0051	.0063	.0076	.0089	.0105	.0115

TABLE III

Showing the Mean Error Attained by the Use of the Smith and Wherry Shrinkage Formulae.

Formula	Table I	Table II	Tables I and II
Smith00097	.00180	.00138
Wherry00018	.00057	.00038
N	10	10	20