THE TWENTY-SEVEN PER CENT RULE[1]

By John Ross[2] and R. A. Weitzman[3]

Princeton University and Educational Testing Service

1. Summary. A method is described for computing the asymptotic variance of the maximum likelihood correlation estimator $\tilde{\rho}$, which uses the number of observations in symmetrically placed corner regions, ignoring a middle section. It is shown that the optimum location of the regions varies with the true value of the correlation. Tables and figures are presented showing the optimal locations and the efficiency of the estimation procedure for various locations under two conditions: one, where cost depends on total sample size and two, where cost depends on the number of observations in the corner regions. In the second case, the method discussed is shown to be able to attain any given precision more cheaply than estimation by the product moment correlation coefficient.

2. Introduction. It is well known that the correlation of two normally distributed variables may be estimated from the number of cases falling in the corners above or below the median of one variable y, and outside the limits $x = \mu_x \pm h\sigma$ on the other variable (see Figure 1). There is an appropriate maximum likelihood estimation procedure which will be described in Section 2. It is also well known that when $\rho = 0$, the asymptotic variance of $\tilde{\rho}$, the estimator, is at a minimum for any given sample size n when $h = .6121$, [1], leaving .2702 of the population in the regions past $\mu_x \pm h\sigma$. This is one basis for the widely accepted "twenty-seven per cent rule", that observations lying in the upper and lower 27% on one of the two continuous variables should be used in four-fold correlation estimation.

A fact not so well recognized is that the variance of the estimator $\tilde{\rho}$ is a function both of h (or $\lambda = (2\pi)^{-\frac{1}{2}} \int_h^\infty \exp(-\frac{1}{2}x^2)\,dx$) and of ρ, the true correlation between x and y. Consequently, it is wrong to accept on faith that the h (or λ) value that minimizes the variance of $\tilde{\rho}$ when ρ is 0 will do so when it is not. It is important to recognize, too, that the variance of $\tilde{\rho}$ at points other than the minimizing value of h or λ (for a given ρ) may not differ from the value at the minimum by a great amount. It is therefore of practical importance to know the shape of the functions Var $(\tilde{\rho})$ rather than just where their minima are. Firstly, it may turn out that a round value of λ such as .10, .20, or .25 will give a result very little different from a value like .2702, the minimizing value of λ for

Received 17 April 1962; revised 27 August 1963.

[1] Thanks are due to S. S. Wilks whose presentation of the problem led to our realization of the need for the present investigation and to R. Pinkham who gave us invaluable mathematical aid. The study was carried out while both authors held Psychometric Fellowships at Princeton University and Educational Testing Service.

[2] Now at University of Western Australia.

[3] Now at Los Angeles State College.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

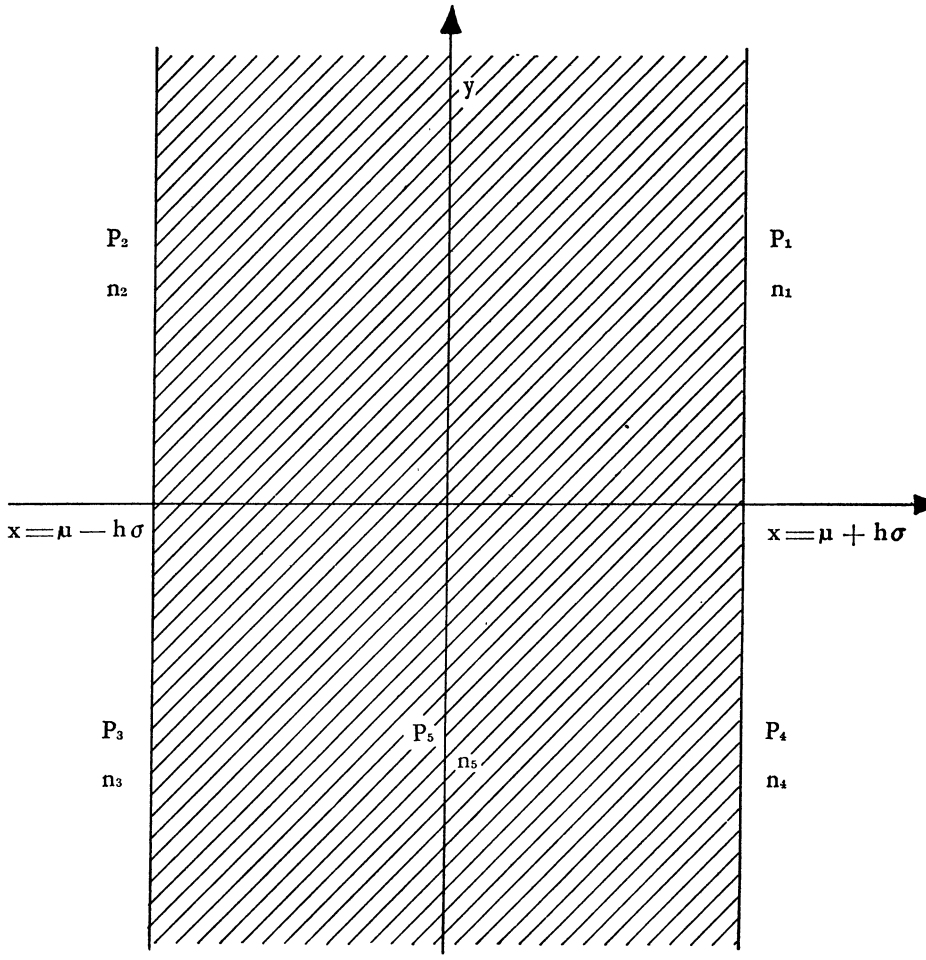

FIG. 1. Graph showing the locations of theoretical proportions (P_i) and empirical frequencies (n_i) in a center region (shaded) and four corner regions.

$\rho = 0$. Secondly, a knowledge of the variance function allows us to inspect the situation where the number of cases to be looked at on both variables is fixed, but the sample size of which they represent some fraction is free to vary. As it turns out the best λ values in situations where cost depends on cases looked at on both variables, and not on sample size, are considerably less than .2702, especially for small values of ρ.

In this paper we (i) describe a method of finding the asymptotic value Var ($\tilde{\rho}$) for a range of values of ρ and λ, (ii) present a table of values of Var ($\tilde{\rho}$) for $\rho = 0.00\ (0.10)\ 0.90$ and for $\lambda = 0.05\ (0.05)\ 0.50$, with more detailed information in the region of the optimal λ, (iii) present information on the relative efficiency of the maximum likelihood estimator $\tilde{\rho}$ as a function of ρ and λ (in rela-

tion to the product moment correlation coefficient), and (iv) discuss efficiency in two cases: (a) where cost depends on total sample size and (b) where cost depends on the number of bivariate observations made, rather than total sample size.

The practical value of the tables presented depends on the appropriateness of applying asymptotic values to empirical situations. We do not know how large n has to be before it is reasonable to take the asymptotic variance of $\tilde{\rho}$ as the variance of estimates made in practice.

The argument also supposes that the values h or λ are known population values. In practice we have to estimate them from samples and their variability would add to the variance of $\tilde{\rho}$, a fact not taken into account in the tables. However, the recommendation in this paper that, in the second cost situation, the λ value used be decreased from .2702 would seem to gain further strength from the fact that the larger samples used in so doing would stabilize the estimates of h and λ.

3. The asymptotic variance of the estimator. The maximum likelihood procedure and the expression for the asymptotic variance of the estimator have been described by Mosteller [4].

Let x and y be two standardized normal variables with the joint density function

$$f(x, y) = \frac{1}{2\pi(1 - \rho^2)^{\frac{1}{2}}} \exp\left\{\frac{-(x^2 + y^2 - 2\rho xy)}{2(1 - \rho^2)}\right\}.$$

Let h define five regions, as in Figure 1—four corner regions and a central region—supporting the five probabilities:

$$p_1 = \int_0^\infty \int_h^\infty f(x, y)\, dx\, dy \qquad p_2 = \int_0^\infty \int_{-\infty}^{-h} f(x, y)\, dx\, dy$$

$$p_3 = \int_{-\infty}^0 \int_{-\infty}^{-h} f(x, y)\, dx\, dy \qquad p_4 = \int_{-\infty}^0 \int_h^\infty f(x, y)\, dx\, dy$$

$$p_5 = 1 - p_1 - p_2 - p_3 - p_4.$$

Now, let n_i denote the number of observations falling in the region for which the probability of an observation is p_i. Then the joint probability distribution of the four values n_1, n_2, n_3 and n_4, the number of *observations* in each of the four corner regions, is

$$(1) \qquad g(n_1, n_2, n_3, n_4) = (n!/\prod_1^5 n_i!) \prod_1^5 p_i^{n_i},$$

where $n = \sum_{i=1}^5 n_i$. The maximum likelihood estimate of ρ comes simply from (1). Taking logarithms gives $\log g = \log c + \sum_{i=1}^5 n_i \log p_i$ where c is the multinomial coefficient in (1). Differentiation with respect to ρ gives

$$(2) \qquad d(\log g)/d\rho = \sum_{i=1}^{4} n_i \dot{p}_i / p_i \qquad \text{(since } p_5 \equiv 1 - 2\lambda \text{ for all } \rho\text{)}$$

where $\dot{p}_i$ is defined as $\dot{p}_i = dp_i/d\rho$. Then as Mosteller [4] shows, setting the right-hand side of (2) equal to zero and solving for ρ to get the maximum likelihood estimate $\tilde{\rho}$ leads to the condition that $(n_1 + n_3)/(n_1 + n_2 + n_3 + n_4) = p_1/\lambda$. It is generally possible to find a value of ρ for which the equality holds, and that value is the estimate. A practical procedure is to consult tables of the normal bivariate surface.

The estimator $\tilde{\rho}$ is shown by Mosteller [4] to have an asymptotically normal distribution with

$$(3) \qquad \text{Var } (\tilde{\rho}) = \frac{p_1(\lambda - p_1)}{2n\lambda \dot{p}_1^2} = \frac{1}{2n} \cdot \frac{p_1(\lambda - p_1)}{\lambda \dot{p}_1^2}$$

where, we recall (see Figure 1), $\lambda = p_1 + p_4 = (2\pi)^{-\frac{1}{2}} \int_h^\infty \exp(-\frac{1}{2}x^2)\, dx$. In order to evaluate the $p_1(\lambda - p_1)/\lambda \dot{p}_1^2$ part of expression (3) for the variance of $\tilde{\rho}$, we need first to find an expression for $\dot{p}_1$.

We have

$$p_1 = \int_0^\infty \int_h^\infty f(x, y)\, dx\, dy$$

$$= \int_0^\infty \int_h^\infty \frac{1}{2\pi(1 - \rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{x^2 + y^2 - 2\rho xy}{1 - \rho^2}\right\} dx\, dy$$

and $\dot{p}_1 = dp_1/d\rho$.

It is convenient to make a change of variables by defining α and β as follows: $\alpha = (x - \rho y)/(1 - \rho^2)^{\frac{1}{2}}$, $\beta = y$. Then, substituting, the expression for p_1 becomes

$$(4) \qquad p_1 = \frac{1}{2\pi} \int_0^\infty \int_{(h - \rho\beta)/(1-\rho^2)^{\frac{1}{2}}}^\infty \exp\left\{-\frac{1}{2}(\alpha^2 + \beta^2)\right\} d\alpha\, d\beta.$$

Define $\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^x \exp(-\frac{1}{2}t^2)\, dt$. Then, from (4)

$$(5) \qquad \begin{aligned} p_1 &= \frac{1}{(2\pi)^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\beta^2\right) \cdot \left\{1 - \Phi\left(\frac{h - \rho\beta}{(1 - \rho^2)^{\frac{1}{2}}}\right)\right\} d\beta \\ &= \frac{1}{2} - \frac{1}{(2\pi)^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\beta^2\right) \cdot \Phi\left(\frac{h - \rho\beta}{(1 - \rho^2)^{\frac{1}{2}}}\right) d\beta. \end{aligned}$$

From (5), we have

$$\frac{dp_1}{d\rho} = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\beta^2\right) \cdot \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{h - \rho\beta}{(1 - \rho^2)^{\frac{1}{2}}}\right)^2\right\}$$

$$\cdot \left\{-\frac{\rho(h - \rho\beta)}{[(1 - \rho^2)^3]^{\frac{1}{2}}} + \frac{\beta}{(1 - \rho^2)^{\frac{1}{2}}}\right\} d\beta.$$

We now make a further change of variable by defining $u = (\beta - \rho h)/(1 - \rho^2)^{\frac{1}{2}}$.

Then

$$\frac{dp_1}{d\rho} = \frac{1}{2\pi}\exp\left(-\frac{1}{2}h^2\right)\int_{-\rho h/(1-\rho^2)^{\frac{1}{2}}}^{\infty}\exp\left(-\frac{1}{2}u^2\right)\left\{\frac{\rho h + u(1-\rho^2)^{\frac{1}{2}}}{1} - \frac{\rho}{(1-\rho^2)}[h - \rho^2 h - \rho u(1-\rho^2)^{\frac{1}{2}}]\right\}du$$

$$= \frac{\exp(-\frac{1}{2}h^2)}{2\pi(1-\rho^2)^{\frac{1}{2}}}\int_{-\rho h/(1-\rho^2)^{\frac{1}{2}}}^{\infty}\exp\left(-\frac{1}{2}u^2\right)\cdot u\,du$$

$$= \frac{\exp(-\frac{1}{2}h^2)\exp(-\rho^2h^2/2(1-\rho^2))}{2\pi(1-\rho^2)^{\frac{1}{2}}}.$$

Finally, we have,

$$\frac{dp_1}{d\rho} = \frac{\exp(-h^2/2(1-\rho^2))}{2\pi(1-\rho^2)^{\frac{1}{2}}}. \tag{6}$$

Equation (6) provides a manageable expression for $dp_1/d\rho$ or $\dot{p}_1$, and allows ready evaluation of the expression for Var $(\tilde{p})$. We have computed the variance of $\tilde{p}$ as a function of λ (and, therefore, h) for a range of values of ρ from 0 to .90 in steps of .10. The method of computation is described in Section 7.

4. The shape of the variance functions. Values of $2n$ Var $(\tilde{p})$ are set out in Table 1 for values of ρ from 0 to .90 in steps of .10, and for selected values of λ between .05 and .50.

TABLE 1

Table of values of $p_1(\lambda - p_1)/\lambda\dot{p}_1^2$

λ	Correlation (ρ)									
	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
.05	7.385	7.300	7.061	6.711	6.246	5.719	5.186	4.669	n.a.	n.a.
.10	5.101	5.031	4.831	4.509	4.081	3.574	3.022	2.482	2.048	n.a.
.15	4.334	4.269	4.079	3.774	3.367	2.877	2.333	1.773	1.252	.620
.20	4.008	3.945	3.762	3.466	3.068	2.589	2.053	1.495	.964	.527
.25	3.889	3.827	3.646	3.352	2.957	2.479	1.941	1.379	.838	.391
.26	3.881	3.820	3.638	3.344	2.948	2.469	1.930			
.27	3.879	3.818	3.637	3.342	2.946	2.466	1.925			
.28	3.881	3.820	3.638	3.343	2.946	2.465	1.922	1.352		
.29	3.888	3.826	3.644	3.348	2.950	2.467	1.922	1.350		
.30	3.898	3.836	3.654	3.357	2.958	2.473	1.926	1.349	.790	.320
.31								1.352	.788	
.32								1.357	.788	
.33									.790	.302
.34									.793	.299
.35	4.007	3.945	3.759	3.456	3.048	2.550	1.985	1.386	.799	.298
.36										.299
.37										.299
.40	4.209	4.145	3.954	3.643	3.221	2.704	2.114	1.483	.857	.311
.45	4.512	4.446	4.248	3.925	3.485	2.944	2.322	1.649	.970	.360
.50	4.935	4.866	4.660	4.322	3.861	3.290	2.628	1.903	1.157	.461

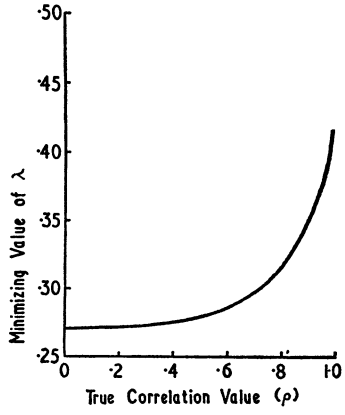

FIG. 2

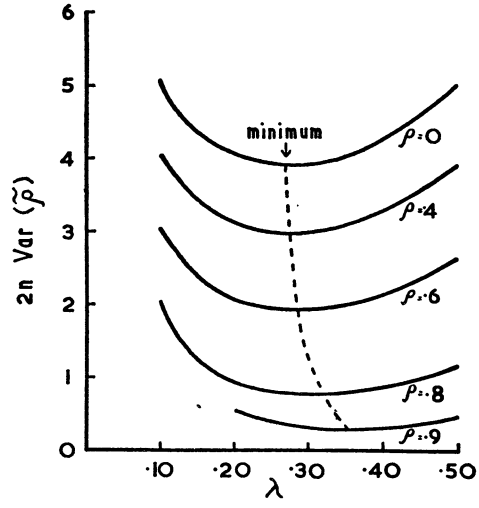

FIG. 3

FIG. 2. Graph showing the minimizing value of λ (proportion in each tail) for true correlation values (ρ) up to .98.

FIG. 3. The variance of $\tilde{\rho}$ as a function of λ for selected values of ρ.

Figure 2 shows the values of λ which minimize Var ($\tilde{\rho}$) for values of ρ up to .98. It is noteworthy that the value of λ for which Var ($\tilde{\rho}$) is a minimum is close to .27 for values from 0 to .6. After that point the optimal λ increases as ρ increases, becoming .35 for a ρ of .90. Figure 3 shows $2n$ Var ($\tilde{\rho}$) as a function of λ for various values of ρ. It is to be noted that the variance function is rather flat for all values of ρ over the range $\lambda = .50$ to $\lambda = .15$. This means that the optimal value of λ is not markedly superior to smaller or larger values.

5. Relative efficiency. Table 2 presents data on the efficiency of $\tilde{\rho}$ (inverse relative efficiency) as an estimator in comparison with the product moment coefficient, r. The entries in Table 2 are the values $p_1(\lambda - p_1)/2\lambda p_1^2$ divided by $(1 - \rho^2)^2$. It will be noted that relative efficiency decreases as ρ increases.

6. Efficiency for two cost situations. If the cost involved in making the estimate depends on n, the total sample size, the appropriate efficiency information is that given in Table 2. In many situations, e.g., educational, observations on one variable may be readily available (on file for example), and observations on the other may be costly to make. If information on the continuous selecting variable x is readily available and if the cost depends on the number of observations to be made on the second variable, y, the appropriate table is Table 3. Table 3 was computed to take account of the fact that Var ($\tilde{\rho}$) depends not on $\sum_{i=1}^{4} n_i$, the number of observations made on both variables x and y, but on n, the total sample, including n_5, the number of cases for which information about y is ignored. Table 3 presents efficiency as a function of $\sum_{i=1}^{4} n_i$ assuming $\sum_{i=1}^{4} n_i = 2n\lambda$. That is to say the variance of $\tilde{\rho}$ when $\sum_{i=1}^{4} n_i$ observations are made is

TABLE 2

Inverse relative efficiency of $\tilde{\rho}$

λ	Correlation (ρ)									
	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
.05	3.693	3.724	3.831	4.052	4.426	5.084	6.331	8.975	n.a.	n.a.
.10	2.551	2.567	2.621	2.723	2.892	3.177	3.689	4.771	7.901	n.a.
.15	2.167	2.178	2.213	2.279	2.386	2.557	2.848	3.408	4.830	8.587
.20	2.004	2.013	2.041	2.093	2.174	2.301	2.506	2.874	3.719	7.299
.25	1.945	1.953	1.978	2.024	2.095	2.204	2.369	2.651	3.233	5.416
.30	1.949	1.957	1.982	2.027	2.096	2.198	2.351	2.593	3.048	4.432
.35	2.004	2.013	2.039	2.087	2.160	2.267	2.423	2.664	3.083	4.127
.40	2.105	2.115	2.145	2.200	2.282	2.404	2.581	2.851	3.306	4.307
.45	2.256	2.268	2.305	2.370	2.469	2.617	2.834	3.170	3.742	4.986
.50	2.468	2.483	2.528	2.610	2.736	2.924	3.208	3.658	4.464	6.385

TABLE 3

Inverse relative efficiency of $\tilde{\rho}$ *where cost depends upon* $\sum_{i=1}^{4} n_i$ *, and where* n *is estimated as* $\sum_{i=1}^{4} n_i/2\lambda$

λ	Correlation (ρ)									
	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
.05	.369	.372	.383	.405	.443	.508	.633	.898	n.a.	n.a.
.10	.510	.513	.524	.545	.578	.635	.738	.954	1.580	n.a.
.15	.650	.653	.664	.684	.716	.767	.854	1.022	1.449	2.576
.20	.802	.805	.816	.837	.870	.920	1.002	1.150	1.488	2.920
.25	.973	.977	.989	1.012	1.048	1.102	1.185	1.326	1.617	2.708
.30	1.169	1.174	1.189	1.216	1.258	1.319	1.411	1.556	1.829	2.659
.35	1.403	1.409	1.427	1.461	1.512	1.587	1.696	1.865	2.158	2.889
.40	1.684	1.692	1.716	1.760	1.826	1.923	2.065	2.281	2.645	3.446
.45	2.030	2.041	2.075	2.133	2.222	2.355	2.551	2.853	3.368	4.487
.50	2.468	2.483	2.528	2.610	2.736	2.924	3.208	3.658	4.464	6.385

Note: Figures in this table are given by 2λ Var $[\tilde{\rho}]$/Var $[r]$.

compared with the variance of r with the same number of observations. In the case of r this is the sample size, since an observation is made on both variables in all cases.

It is of considerable interest that where cost depends on $\sum_{i=1}^{4} n_i$, not n, it is possible for low values of ρ to attain any given level of precision at less cost than with the product moment correlation coefficient (which involves looking at all n cases). Table 3 suggests that the optimum λ values are very small for low values of ρ, but increase with ρ. Table 3 shows that in an educational testing situation where item-test correlations are being estimated, and where the value of $\sum_{i=1}^{4} n_i$ is fixed, at say 100, the best sample size is 1000 or more (i.e., $\lambda = .10$ or less) for values of $\rho = .7$ or less. The variance of the estimated $\tilde{\rho}$ is less than with 100 observations selected from a sample of 185 where the 100 observations would correspond to a λ of .27.

Table 3 also shows that the variance of the estimator $\tilde{\rho}$ computed on 100 cases out of a sample of 1000 is about half the variance of r computed on 100 cases for values of ρ up to and including .5. One needs to be wary in taking Table 3 too literally, since with very small λ values it is likely that irregularities in the bivariate surface would be magnified. When considerations of robustness are brought in, probably the best procedure would be to select values of λ in the region .5 to .10.

7. Method of computation. The computations were carried out on an IBM 1620 computer. The value of the expression for the variance of the estimator, Var $(\tilde{\rho})$, was computed for values of λ ranging between .05 and .50 in steps of .01 for each value of ρ separately.

A number of approximation procedures were used in obtaining the values of the terms needed to evaluate the expression.

(1) The value of h was computed by Hastings', [2], approximation formula sheet: 67, Part II.

(2) The value of p_1 was computed by interpolating between the tabled values of the bivariate normal surface given by Owens [5].

(3) Exponential values and logarithms were obtained by standard FORTRAN Subroutines.

It should be noted that as ρ increases toward 1, p_1 increases toward λ, and the difference is very small for high values of ρ and h. For this reason the table of values of Var $(\tilde{\rho})$ is likely to be inaccurate for high values of ρ. Since the point of this report is to bring out general features, however, the inaccuracies should not be serious in their effects, and, in particular, should not affect the general shape of the curves showing Var $(\tilde{\rho})$ as a function of λ. It is this general shape which is central to the point made in this paper.

REFERENCES

[1] FLANAGAN, J. C. (1939). General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *J. Educ. Psych.* **30** 674–680.

[2] HASTINGS, C. (1955). *Approximations for Digital Computers.* Princeton Univ. Press.

[3] KELLEY, T. (1939). The selection of upper and lower groups for the validation of test items. *J. Educ. Psych.* **30** 17–24.

[4] MOSTELLER, F. (1946). On some useful "inefficient" statistics. *Ann. Math. Statist.* **17** 377–408.

[5] OWEN, D. B. (1956). Tables for computing bivariate normal probabilities. *Ann. Math. Statist.* **27** 1075–1090.