Appendix C

Tests of Significance and Sampling Errors

In this study, problems of sampling error may arise in at least three different connections: two samples drawn from the same population may erroneously appear to be different (an error of Type I); two samples drawn from different populations may erroneously appear to be identical (an error of Type II); and finally the sample estimates of some of the special measures introduced here, such as the efficiency index and the bad-loan relative, may deviate considerably from the true values. In Chapter 2 the Chi-square test and the t-test were mentioned in connection with the first of these sampling problems. These tests, which are adequately described in standard treatises,[1] need little further discussion. It is only necessary to point out that special procedures for calculating Chi-square may be appropriate when frequency distributions are presented in percentages, as they are in this study. (See pages 157-58.)

Both the Chi-square test and the t-test, if used as previously suggested, have the great disadvantage of testing the significance of only one variate at a time. This is unsatisfactory for two reasons. First, two samples may not differ significantly in respect to any one of p variates, and yet the combined difference for all p variates may be highly significant. Second, a significant difference may appear in one or two isolated variates when the combined difference for all p variates is not significant; for if 100 tests of significance were applied to 100 independent factors, five of these tests could exceed the 5 percent significance

[1] See footnotes 2 and 3, Chapter 2.

level, and one of them could exceed the one percent level, without discrediting the null hypothesis;[2] hence the singling out of the particular variates that happened to meet the specifications would be entirely erroneous. In a case entailing several factors, the ideal procedure is simultaneously to test the significance of all the factors under consideration; and the findings of the individual tests should then be reviewed in the light of the findings of the combined test.

A simultaneous test of significance can be accomplished in two ways. In the first place, an n-way cross classification may be made—if there are n factors—and the Chi-square test can be used to test the difference between the two n-way distributions just as it would be used to test the difference between two one-way distributions. This process requires considerable labor and rather large samples if the number of factors considered is more than four.[3] An alternative approach is the generalized t-test, which simultaneously tests the differences between a number of means. This test, which has been discussed by several writers, is extremely pertinent to some of the sampling problems encountered in this study.

The T^2-statistic, introduced by Hotelling,[4] is appropriate for determining whether an apparent difference between two samples is attributable to sampling error only (an error of Type I). T^2 is defined by

$$T^2 = \Sigma\Sigma A_{ij}(\bar{x}_i - \bar{x}'_i)(\bar{x}_j - \bar{x}'_j)\frac{(n+1)(n'+1)}{n+n'+2},$$

where $\bar{x}_i$ is the mean value of the i-th variate for one sample

[2] Here the null hypothesis is that both samples are drawn from the same population.

[3] If only two classification cells are used for each factor—with and without bank account, and more or less than six years of employment tenure, for example—the number of classification cells for n factors is 2^n. Thus five factors would entail 32 cells; and if the number of good plus bad-loan cases in each cell is to be at least 20, a sample of 320 good loans and 320 bad is the minimum, and probably a much larger sample will be required.

[4] Harold Hotelling, "The Generalization of 'Student's' Ratio," *Annals of Mathematical Statistics*, vol. 2, no. 3 (1931) pp. 360–78.

and $\bar{x}'_i$ is the mean value for the other. Moreover, the matrix A_{ij} is the inverse of the matrix of the covariances; i.e.,

$$A_{ij} = \frac{s^{ij}}{|s_{ij}|},$$

where $|s_{ij}|$ is the determinant of the s_{ij}'s and s^{ij} is the cofactor of s_{ij} in that determinant. For two samples s_{ij} is defined by

$$s_{ij} = \frac{1}{n+n'}\left[\Sigma(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \Sigma(x'_i - \bar{x}'_i)(x'_j - \bar{x}'_j)\right],$$

where n is the number of degrees of freedom in one sample and n′ is the number in the other. On the assumption that the two samples to be tested are drawn from the same multivariate normal population, T has the distribution

$$d(f) = \frac{2\Gamma\left(\frac{n+n'+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{n+n'+1-p}{2}\right)\left(n+n'\right)^{\frac{p}{2}}} \times \frac{T^{p-1}dT}{\left(1+\frac{T^2}{n+n'}\right)^{\frac{n+n'+1}{2}}} \qquad (1)$$

This is obviously equal to "Student's" ratio, t, for p equal to one. For large values of n or n′ d(f) approaches

$$\frac{(T^2)^{\frac{p-1}{2}}\, e^{-\frac{T^2}{2}}dT}{2^{\frac{p-2}{2}}\,\Gamma\left(\frac{p}{2}\right)},$$

which indicates that T is normally distributed for p equal to one if both positive and negative values of T are considered, and that T^2 has the Chi-square distribution for all values of p. For small values of n and n′, the significance of T^2 can be determined from the z-distribution by means of the transformation

$$z = \tfrac{1}{2}\log_e \frac{n+n'+1-p}{p(n+n')}T^2, \qquad (2)$$

where there are $n_1 = p$ and $n_2 = n + n' + 1 - p$ degrees of freedom.

The amount of clerical labor necessary to compute T increases rapidly as the number of variates considered increases.

This difficulty is not serious if the data can be punched on cards, so that the sums of squares and products can be computed by automatic multiplying punches, and if the necessary determinants can be solved mechanically; otherwise, it is serious. In this study we have frequently been able to economize labor by determining T for a small number of variates and by using this value as a test of significance for a larger number. The reason is that the samples used here are large enough to give very significant results for some of the individual factors. The generalized t-test is not needed to establish combined significance when individual significance is lacking; it is only necessary to confirm individual significance. Since the value of T for p variates cannot be less than the value of T for any p-h of the same variates,[5] a large value of T (or t) for a single variate may suffice to establish significance for all p variates; this value of t can be used in (2) in place of the true value of T, and if the resulting value of z is significant, the true value of z must also be significant. To establish significance in this way, the value of t would have to be distinctly higher than the value necessary to establish significance for one variate. If a single

[5] To prove this, it is only necessary to show that $T_p \geq T_{p-h}$, where T_p is determined for p variates and T_{p-h} is determined for $p - h$ of the original p variates. In Appendix A we mentioned (see footnote 3) that

$$T_p\sqrt{\frac{n+n'+2}{(n+1)(n'+1)}} = \frac{\Sigma l_i a_i}{\sqrt{\Sigma\Sigma l_i l_j s_{ij}}} = U_p \quad (i, j = 1 \ldots p),$$

where the fact that $l_i = \Sigma_j a_j \frac{s^{ij}}{|s_{ij}|}$ $\quad (i, j = 1 \ldots p),$

makes U_p the maximum of all ratios having the form (see page 111)

$$\frac{\Sigma l'_i a_i}{\sqrt{\Sigma\Sigma l'_i l'_j s_{ij}}} \quad (i, j = 1 \ldots p).$$

U_{p-h} can be written in the same form, i.e.,

$$U_{p-h} = \frac{\Sigma l''_i a_i}{\sqrt{\Sigma\Sigma l''_i l''_j s_{ij}}} \quad (i, j = 1 \ldots p)—$$

where $l''_i = \Sigma_i a_i \frac{s^{ij}}{|s_{ij}|}$ for $i, j = 1 \ldots p - h$

and $l''_i = 0$ for $i, j = p - h + 1 \ldots p.$

Therefore $U_p \geq U_{p-h}$, and $T_p \geq T_{p-h}$.

variate does not yield a sufficient value of t, a combination of two or three of the most likely variates may give a generalized T large enough for all other variates.

The generalized t-test was used in practice to establish significance for the four factors singled out for special analysis in connection with the used-car sample—down payment, cash purchase price, borrower's income, and length of contract. The value of T^2 obtained was 86.76. This is more than large enough to establish significance for the four factors in question; the value of z was 1.54 against the 1 percent value of less than .65. In fact, 86.76 for T^2 is large enough to establish significance for many more than four factors. The corresponding value of z for 24 factors ($n_1 = p = 24$), which is the largest finite number tabulated for n_1 by R. A. Fisher,[6] is .63; it is more than significant by the 1 percent criterion.

A similar determination of T^2 can be made for the seven factors included in the second credit-rating formula. This formula was originally determined from a subsample of 191 good loans and 190 bad loans; and the first problem is to establish significance within the subsample. The value of t in the subsample for stability of occupation is 5.29, which is more than sufficient to establish significance for one degree of freedom. Since t^2 (27.9) is necessarily less than T^2, and since the corresponding value of z (.682) is significant for seven factors ($n_1 = p = 7$), it follows that the seven factors are conjointly significant for the original subsample. Furthermore, after the formula had been determined for the subsample, it was tested on the entire commercial bank sample; then it was tested, with slight modifications, on the industrial bank sample. In both cases, an extremely significant difference between good and bad loans can be shown by means of the Chi-square test.

The sampling distribution of T in (1) is based on the assumption that the population value, τ, is 0. This distribution is appropriate only to determine the probability that two samples

[6] *Statistical Methods for Research Workers* (London and Edinburgh, 6th edition, 1936) Table VI.

showing an apparent discrepancy could have been drawn from a single universe (an error of Type I). Sometimes, however, it is desirable to determine the probability that no significant discrepancy will be observed between two samples drawn from different universes (an error of Type II). For this purpose the distribution of T must be determined on the assumption that τ is not 0. This problem has been investigated by Bose and Roy, Hsu, and Tang.[7] Tang has prepared tables of the distribution to permit the calculation of the probability of a Type II error.

When a discriminant function,

$$Z = l_1x_1 + l_2x_2 + \ldots,$$

is determined for several factors, the l-coefficients are naturally subject to sampling error. The problem of finding their sampling distribution, however, can be reduced to a more fundamental one—that of finding the sampling distribution of the ratio U. The l-coefficients are not unique. Although a unique set of constants will be determined from the solution of equation (1) (see Appendix A, p. 111), any other set of constants proportional to them will produce an equally effective discriminant function with the same value of U; that is, the l's will be uniquely determined only after one of them has been arbitrarily chosen. As a result it is meaningless to speak of the sampling error of one single l-coefficient, for an error in one coefficient implies an error in all the others. For most purposes a set of l's will be erroneous only if they jointly produce an unsatisfactory estimate of U; if U can be determined precisely, possible variations in the l's can usually be overlooked.[8]

The sampling distribution of U follows directly from the distribution of Hotelling's generalized T or from the distribution of the D^2-statistic of Bose and Roy. These distributions are admirably adapted to determining the probability of a Type I or a Type II error in a small sample, but sometimes another sampling problem presents itself. In a large sample, the value of U may be so large and its standard error may be so small that an error of either Type I or Type II is unthinkable. Here we are not interested in determining whether U departs significantly from 0; we want to know how reliable U is as an estimate of the population value Υ. If, for example, Υ is equal to one, is a value of U less than .9 or greater than 1.1 likely to occur? For problems like this the limiting value of the distribution of U will usually be a satisfactory approximation.

In the one-variate case, two populations have a standard deviation of σ and a mean difference of α. Two samples drawn from these populations will have a standard deviation of s and a mean difference of a. We require the limiting distribution of a/s for large samples. The difference a is normally distributed with variance $\sigma^2\ (n + n')/nn'$ where n is the number of cases in one sample and n′ is the number in the other. The standard deviation s has the Chi distribution with $n + n' - 2$ degrees of freedom, but for large values of either n or n′ the distribution approaches normal, with variance of $\sigma^2/2(n + n')$. The problem therefore reduces to the distribution of the quotient of two normal independent variates.

Geary has shown that if x and y are uncorrelated normal variates with 0 means, and if z is defined by

$$z = \frac{Y + y}{X + x} -$$

where Y and X are constants, and $X \geq 3\sigma_x$ —

then
$$t = \frac{Xz - Y}{\sqrt{\sigma_x{}^2z^2 + \sigma_y{}^2}}$$

[7] R. C. Bose and S. N. Roy, "The Distribution of the Studentised D^2-Statistic," *Sankhya*, vol. 4, no. 1 (Dec. 1938) pp. 19–38; S. N. Roy, "A Note on the Distribution of the Studentised D^2-Statistic," *Sankhya*, vol. 4, no. 3 (Sept. 1939) pp. 373–80; P. L. Hsu, "Notes on Hotelling's Generalized T," *Annals of Mathematical Statistics*, vol. 9, no. 4 (Dec. 1938) pp. 231–43; P. C. Tang, "The Power Function of the Analysis of Variance Tests with Tables and Illustrations of their Use," *Statistical Research Memoirs*, vol. 2 (1938) pp. 126–49.

[8] Occasionally the problem will arise of determining how much the l's can vary without unduly affecting U. We illustrated this sort of problem in Appendix A, where we investigated the effect of the arbitrary assumption that all correlation coefficients are 0.

will be approximately normally distributed with unit variance.[9] It can be shown that as σ_x and σ_y both approach 0,

$$t = \frac{X^2(z - Y/X)}{\sqrt{\sigma_x^2 Y^2 + \sigma_y^2 X^2}}$$

also approaches normal with unit variance.[10] From this it follows that the limiting distribution of a/s is normal with a variance of

$$\frac{n + n'}{nn'} + \frac{\alpha^2}{\sigma^2}\frac{1}{2(n + n')} = \frac{n + n'}{nn'}\left(1 + \frac{\alpha^2}{\sigma^2}\frac{nn'}{2(n + n')^2}\right),$$

where $\frac{\alpha^2}{\sigma^2}$ can be replaced by v^2. This result, moreover, can be generalized to any finite number of variates: in the limit the distribution of U is normal with variance of

$$\frac{n + n'}{nn'} + \Upsilon^2\frac{1}{2(n + n')} = \frac{n + n'}{nn'}\left(1 + \Upsilon^2\frac{nn'}{2(n + n')^2}\right)$$ [11]

[9] R. C. Geary, "The Frequency Distribution of the Quotient of Two Normal Variates," *Journal of the Royal Statistical Society*, vol. XCIII, part III (1930) pp. 442–46. The notation used here is not Geary's.

[10] To prove this, we have only to prove that

$$\frac{\sqrt{\sigma_x^2 z^2 + \sigma_y^2}}{\sqrt{\sigma_x^2\frac{Y^2}{X^2} + \sigma_y^2}}$$

approaches 1 as σ_x and σ_y approach 0. Squaring, we get

$$\frac{\sigma_x^2 z^2 + \sigma_y^2}{\sigma_x^2\frac{Y^2}{X^2} + \sigma_y^2} = 1 - \frac{\left(\frac{Y^2}{X^2} - z^2\right)\sigma_x^2}{\sigma_x^2\frac{Y^2}{X^2} + \sigma_y^2} = 1 - \frac{\frac{Y^2}{X^2} - z^2}{\frac{Y^2}{X^2} + \frac{\sigma_y^2}{\sigma_x^2}}$$

which clearly approaches 1 because $\frac{Y^2}{X^2} - z^2$ approaches 0, and $\frac{Y^2}{X^2} + \frac{\sigma_x^2}{\sigma_y^2}$ does not.

[11] Let $U = \Upsilon + \mathfrak{U}$, $s_{ij} = \sigma_{ij} + \delta_{ij}$, and $a_i = \alpha_i + \mathfrak{a}_i$, where the Greek letters represent population parameters, and the German letters represent random variations about them; as the size of sample increases, the random variations grow smaller and eventually approach zero. By definition

$$U^2 = (\Upsilon + \mathfrak{U})^2 = \frac{\Sigma\Sigma(\alpha_i + \mathfrak{a}_i)(\alpha_j + \mathfrak{a}_j)\text{ cofactor }(\sigma_{ij} + \delta_{ij})}{|(\sigma_{ij} + \delta_{ij})|}.$$

Since U remains invariant for all non-singular linear transformations, we can assume without loss of generality that $\sigma_{ij} = 0$ whenever $i \neq j$. We wish to reduce this to a linear function in the $\mathfrak{a}_{ij}$'s and δ_{ij}'s, which is possible because second order terms in $\mathfrak{a}_{ij}$ and δ_{ij} can be neglected as infinitesimals of higher order. We may therefore write:

$$\frac{(\alpha_i + \mathfrak{a}_i)(\alpha_j + \mathfrak{a}_j)\text{ cofactor }(\sigma_{ij} + \delta_{ij})}{|(\sigma_{ij} + \delta_{ij})|} \qquad [i \neq j]$$

$$= \frac{(\alpha_i + \mathfrak{a}_i)(\alpha_j + \mathfrak{a}_j)\delta_{ij}(\sigma_{11} + \delta_{11})(\sigma_{22} + \delta_{22}) \cdots (\sigma_{pp} + \delta_{pp})}{(\sigma_{ii} + \delta_{ii})(\sigma_{jj} + \delta_{jj})(\sigma_{11} + \delta_{11})(\sigma_{22} + \delta_{22}) \cdots (\sigma_{pp} + \delta_{pp})}$$

$$= \frac{(\alpha_i + \mathfrak{a}_i)(\alpha_j + \mathfrak{a}_j)\delta_{ij}}{(\sigma_{ii} + \delta_{ii})(\sigma_{jj} + \delta_{jj})} = (v_i + u_i)(v_j + u_j)r_{ij},$$

where $u_i = v_i + \mathfrak{u}_i = \frac{\alpha_i + \mathfrak{a}_i}{\sqrt{\sigma_{ii} + \delta_{ii}}}$ and $r_{ij} = \frac{\delta_{ij}}{\sqrt{(\sigma_{ii} + \delta_{ii})(\sigma_{jj} + \delta_{jj})}}$; moreover, $\frac{(\alpha_i + \mathfrak{a}_i)^2\text{ cofactor }(\sigma_{ii} + \delta_{ii})}{|(\sigma_{ii} + \delta_{ii})|} = \frac{(\alpha_i + \mathfrak{a}_i)^2}{\sigma_{ii} + \delta_{ii}} = (v_i + u_i)^2$.

Therefore, $(\Upsilon + \mathfrak{U})^2 = \Sigma\Sigma(v_i + u_i)(v_j + u_j)r_{ij}$, where $r_{ii} = 1$. Omitting all second-order terms in u_i and r_{ij} gives $\Upsilon^2 + 2\mathfrak{U}\Upsilon = \Sigma\Sigma v_i v_j r_{ij} + 2\Sigma v_i u_i$,

whence $\mathfrak{U} = \frac{\Sigma v_i u_i + \frac{1}{2}\Sigma\Sigma_{i\neq j} v_i v_j r_{ij}}{\Upsilon}$

since $\Upsilon^2 = \Sigma v_i^2$.

This last is a linear function in u_i and r_{ij}; it is therefore normally distributed in the limit.

Since $\sigma_{2u_i} = \frac{n + n_1}{nn'} + \frac{u_i^2}{2(n + n')}$ and since $\sigma_{2r_{ij}} = \frac{1}{n + n'}$, the variance of $\mathfrak{U}$ is equal to

$$\frac{1}{\Upsilon^2}\left[\Sigma\left(\frac{n + n'}{nn'}v_i^2 + \frac{v_i^4}{2(n + n')}\right) + \Sigma\Sigma_{i\neq j} v_i^2 v_j^2\frac{1}{n + n'}\right]$$

$$= \frac{1}{\Upsilon^2}\left[\frac{n + n'}{nn'}\Sigma v_i^2 + \frac{1}{2(n + n')}\Sigma\Sigma v_i^2 v_j^2\right]$$

$$= \frac{n + n'}{nn'} + \frac{\Upsilon^2}{2(n + n')} = \frac{n + n'}{nn'}\left(1 + \frac{\Upsilon^2 nn'}{2(n + n')^2}\right)$$

A single example will serve to illustrate the size of the errors to be expected in our good- and bad-loan samples. In a sample of about 825 good and 825 bad loans, the approximate standard error of U is $.049\sqrt{1 + \Upsilon^2/8}$. For a value of .5 for Υ, the standard error is .050, which suggests that there is about one chance in twenty that U will lie outside the range of .4 to .6.

STANDARD ERROR OF THE EFFICIENCY INDEX

Since the efficiency index is related to Υ by the relation

$$\text{Index} = \int_{-\Upsilon/2}^{\Upsilon/2} e^{-\frac{t^2}{2}}\,dt$$

for a normal population, sampling errors of the efficiency index can be estimated from the standard error of U. In the above example, a value of .5 for Υ corresponds to an efficiency index of about 20, and the sampling range of .4 to .6 for U corresponds to a range of approximately 16 to 24 for the efficiency index.

An alternative approach to the standard error of the efficiency index is worth pointing out. Consider the 2x2 contingency table

	Class A	*Class B*
Good loans	β	$100 - \beta$
Bad loans	β'	$100 - \beta'$

where β represents the population probability in percentage form that a good loan will belong to Class A, etc. The efficiency index is equal to the absolute value of $\beta - \beta'$. Since the standard error of b, the sampling estimate of β in a sample of N cases, is $\sqrt{\frac{\beta(100-\beta)}{N}}$, and since the standard error of b′ is $\sqrt{\frac{\beta'(100-\beta')}{N'}}$, the standard error of the difference is

$$\sqrt{\frac{\beta(100-\beta)}{N} + \frac{\beta'(100-\beta')}{N'}}.$$

This formula, derived for a 2x2 table, can also be used for a 2xp table, for a 2xp table can be reduced to a 2x2 table by the simple expedient of consolidating all better-than-average classes into one class, and all worse-than-average classes into another. When the formula is used, the sample estimates must be used in place of the population parameters. This is particularly unfortunate when a 2xp table is to be consolidated, for some better-than-average classes may be erroneously classed as worse than average, and vice versa.

STANDARD ERROR OF THE BAD-LOAN RELATIVE

The bad-loan relative, the ratio of the percent of bad loans in a particular class to the percent of good loans in that class, has been used as a means of comparing the risk merits of any class with those of any other class or with the average. This relative is, of course, subject to sampling error, and comparisons should be modified accordingly. An approximate expression for the standard error of this ratio is derived here.

Let α be the probability that a loan drawn at random from the good-loan population will belong to class A; let α' be the probability that a loan drawn from the bad-loan population will belong to class A; then $\frac{\alpha'}{\alpha}$ is the true bad-loan relative for class A. Let a, a′, and $\frac{a'}{a}$ be the estimates of α, α', and $\frac{\alpha'}{\alpha}$ derived from samples of n good loans and n′ bad ones. If n and n′ are large, a and a′ are both normally and independently distributed with variance

$$\frac{\alpha(1-\alpha)}{n} \quad \text{and} \quad \frac{\alpha'(1-\alpha')}{n'}.$$

From the previous discussion of the sampling error of a quotient, it will be seen that the limiting distribution of a′/a is normal with variance of

$$\frac{\sigma_{a'}^2}{\alpha^2} + \frac{\sigma_a^2 \alpha'^2}{\alpha^4}, \text{ which equals}$$

$$\frac{1}{\alpha^3}\left[\frac{\alpha\alpha'(1-\alpha')}{n'} + \frac{\alpha'^2(1-\alpha)}{n}\right] \tag{3}$$

The square root of (3) is the approximate expression for the standard error of the bad-loan relative.

To give some idea of the amount of error to be expected, the standard errors shown in Table C-1 were computed for sixteen assumed class intervals and two assumed sample sizes. In samples of this size the distribution of a′/a is not normal, but distinctly skewed. These standard errors are computed for a sufficient range of values to indicate fairly well the amount of error possible in the bad-loan relatives computed from the available samples. The standard errors quoted are probably not

TABLE C-1

STANDARD ERRORS FOR ASSUMED SET OF CASES

α (*percent*)	α' (*percent*)	$\frac{\alpha'}{\alpha}$	$\sigma_{a'/a}$ 1,000 cases in each sample	$\sigma_{a'/a}$ 500 cases in each sample
5	5	1.0	.195	.276
10	10	1.0	.134	.190
20	20	1.0	.089	.127
40	40	1.0	.055	.078
3	6	2.0	.438	.620
5	10	2.0	.334	.473
10	20	2.0	.228	.322
20	40	2.0	.148	.210
5	15	3.0	.471	.667
5	20	4.0	.606	.858
15	5	.33	.052	.074
20	5	.25	.038	.054
6	3	.50	.110	.155
10	5	.50	.084	.118
20	10	.50	.057	.081
40	20	.50	.037	.052

adequate to represent a satisfactory margin of error; twice the above standard errors is probably a better estimate, and even then about 5 percent of the sample estimates can be expected to differ from the true value by more than this margin. Since roughly 300 bad-loan relatives are quoted in the tables accompanying this report, some 15 of them are probably erroneous by more than two standard errors.

This discussion of error throws more light on the limitations of small samples in risk analysis. The samples used here are large enough—in many cases much larger than necessary—to demonstrate bona fide relations between bad-loan experience and certain credit factors; stability of employment is a prime example. Although the available samples are adequate to show that persons who have been engaged in the same employment

for 10 years or more are better-than-average risks, and much better than those employed for less than two years, they are not adequate to estimate precisely the degree of difference. To obtain a high degree of precision in estimating bad-loan relatives, much larger samples are necessary; for a sample containing as many as 10,000 good and 10,000 bad loans, the standard errors amount to about 31 percent of the errors for 1,000 cases, which are shown in the set of hypothetical errors presented above.

COMPUTATION OF CHI-SQUARE FOR PERCENTAGE DISTRIBUTIONS

The numerous common methods for computing Chi-square presuppose that the distribution of cases is given in actual frequencies and not in percentages. In the present study, where all distributions have been reduced to percentages, an alternative method designed for percentage distributions was found convenient. To apply this method, only the total number of cases in the samples need be known. The following formula is appropriate:

$$\chi^2 = \frac{n'n''}{10{,}000}\sum_{i=1}^{m}\frac{(a_i' - a_i'')^2}{\frac{a_i'n'}{100} + \frac{a_i''n''}{100}}$$

where n' and n'' are the total number of cases in the good and bad samples, m is the number of classes into which each sample is divided, and a_i' and a_i'' are the percentages of cases in the i^{th} class for the good and bad samples. The quantity $\frac{a_i'n'}{100} + \frac{a_i''n''}{100}$ is the total actual number of cases of both samples in class i. When n' and n'' are equal, or approximately equal, the above formula takes the very simple and convenient form

$$\chi^2 = \frac{n}{100}\sum_{i=1}^{m}\frac{(a_i' - a_i'')^2}{a_i' + a_i''}$$

where n is the number of cases in either sample.

Where n′ and n″ are only approximately equal, this second formula is still useful. If a significant value of χ^2 is obtained when the smaller of the two n's is substituted, the true χ^2 is obviously greater and also significant; and if a non-significant value is obtained with the larger of the two n's, the true value is also non-significant. An example may prove enlightening. The following is the percentage distribution of loans by sex and marital status in the sample submitted by one bank:

	Single Females	*Single Males*	*Married Females*	*Married Males*	*Others*
150 Good loans	30.0	9.3	12.7	40.7	7.3
100 Bad loans	5.0	24.0	2.0	59.0	10.0

In the first class the quantity $\frac{(30.0 - 5.0)^2}{(30.0 + 5.0)}$ is 17.86; the sum of this and four similar quantities for the other four classes is 35.89.[12] If we substitute 100, the smaller of the two n's, we still have 35.89, which is an underestimate of the true χ^2. Since the 1 percent value of χ^2 is only 13.28, 35.89 is clearly significant. Since the contribution of the first class to the total χ^2, 17.86, is itself greater than the 1 percent value of 13.28, the significance can be demonstrated from the first class alone, and additional computation is unnecessary.

[12] With the aid of a table of squares and a calculating machine, the calculation of χ^2 by this process is reasonably easy.

Index

AGE OF BORROWER—4, 74.
APPLICATIONS FOR LOANS—Data Provided by 15, 20.
ASSETS OF BORROWER—3, 62-65, 79-81, 132.*

BAD LOAN RELATIVE (INDEX OF BAD-LOAN EXPERIENCE)—27-28, 95-96; Sampling Error of 9-10, 36, 95-96, 154-57.*
BÔCHER, M.—114n.*
BORROWERS OF CONSUMER INSTALMENT LENDING INSTITUTIONS—Financial Characteristics of 2-5, 14-19, 45-65; Fundamental Requirements of 14-15; Geographical Distribution of 12; Income of, *See* Income; Non-Financial Characteristics of 2-4, 65-77; Vocational Composition of 12.
BOSE, R. C.—113,* 150,* 150n.*

CASH PRICE OF ARTICLE PURCHASED—4, 57-58, 80, 127-30.*
CHAPMAN, JOHN M.—17n, 38n, 40n, 44n, 50n, 56n, 57n, 63n, 77n.
CHI-SQUARE TEST—26n, 145-46,* 157-58.*
COLLATERAL FOR LOAN—10-11, 56-57.
CONSUMER INSTALMENT CREDIT—*See* Credit.
CONTRACT—Length of Loan Contract 53-56, 79-80, 127-30.*
COON, OWEN L.—83n.
COPPOCK, JOSEPH D.—48n.
CORRELATION—Effect of on Analysis 53, 85-90, 115-16,* 128-29,* 131-41.*
COSTS OF CONSUMER INSTALMENT FINANCING BUSINESS—Study of 94-99.
CREDIT—Characteristics of Consumer Instalment Credit 10-13; Classification of Transactions 11.
CREDIT ANALYSIS—Value of 99-101.
CREDIT FACTORS—Evaluation of 1-2, 6-7, 15-19, 90-91.
CREDIT INVESTIGATION—1, 9, 14.
CREDIT POLICY—Revision of 93-94; Social Implications of 8, 100-1.
CREDIT-RATING FORMULAE—7, 83-91, 125-42.*
CREDIT RISK—And Borrower's Financial Characteristics 2-5, 45-65; And Borrower's Non-Financial Characteristics 2-4, 65-77; Fundamentals of Risk Selection 1-2, 14-15; Method of Analyzing Risk Factors 22-43; Time Element as Cause of Variation in Risk Experience 40-41.
CREDIT TRANSACTIONS — Classification of 11.

DEPENDENTS OF BORROWER—4, 74, 77.
DOWN PAYMENT—4, 59-62, 79-81, 127-30.*
DUNHAM, H. L.—83n.

EFFICIENCY INDEX—5-6, 28-31, 107-8;* Sampling Error of 153-54.*
EMPLOYMENT OF BORROWER—Nature of 3-4, 69-74, 80-81, 132;* Stability of 2, 3, 23-26, 65-67, 80, 132.*
EVALUATION OF CREDIT FACTORS—*See* Credit Factors.

FINANCIAL CHARACTERISTICS OF BORROWERS—2-5, 14-19, 45-65.
FISHER, R. A.—24n, 26n, 33n, 111,* 111n,* 149,* 149n.*
FORMULAE, CREDIT-RATING—7, 83-91, 125-42.*
FUNDAMENTALS OF RISK SELECTION—1-2, 14-15.
FUNDS—Use of Funds Borrowed, As Risk Factor 4, 77-78.

GEARY, R. C.—151,* 152n.*
GREENBERG, JOSEPH M.—83n.

* This reference applies to technical edition.