Risk Elements in Consumer Instalment Financing

BY DAVID DURAND

Technical Edition

Financial Research Program

Studies in Consumer Instalment Financing 8

NATIONAL BUREAU OF ECONOMIC RESEARCH

isk Elements in onsumer Instalment Financing

echnical Edition

---

his study presents a statistical approach the analysis of certain pertinent risk ements in consumer instalment financ- g. It attempts to analyze the significance some credit factors generally con- dered important by credit men, to dis- rn which of these factors have been sociated in the past with bad loans, and determine whether or not this informa- on can be used to predict the course of ture transactions. Highly refined statis- al methods have been employed in der to assure precise results and also test the applicability of such methods the problems involved. The study ovides information on:

- the characteristics of consumer instalment credit;
- details of procedure to be followed in an analysis of borrower characteristics as factors in credit risk;
- the relative importance of various credit factors in risk selection as revealed by the analysis;
- credit-rating formulae developed by means of a specialized statistical procedure;
- technical aspects of the statistical methods evolved;
- the application of findings to credit policy and to a study of costs.

he basic data for the study were otained from actual loan applications id were contributed by lending con- rns operating in various phases of nsumer instalment credit — commer- al banks, industrial banking com- anies, personal finance companies, itomobile finance companies, and ap- iance finance companies.

FINANCIAL RESEARCH PROGRAM OF THE
NATIONAL BUREAU OF ECONOMIC RESEARCH

Studies in Consumer Instalment Financing:
Number Eight

DIRECTORS BY UNIVERSITY APPOINTMENT

WILLIAM L. CRUM, *Harvard*
E. E. DAY, *Cornell*
GUY STANTON FORD, *Minnesota*
H. M. GROVES, *Wisconsin*
WALTON H. HAMILTON, *Yale*
WESLEY C. MITCHELL, *Columbia*
A. H. WILLIAMS, *Pennsylvania*
THEODORE O. YNTEMA, *Chicago*
E. W. ZIMMERMANN, *North Carolina*

DIRECTORS APPOINTED BY OTHER ORGANIZATIONS

C. REINOLD NOYES, *American Economic Association*
WINFIELD W. RIEFLER, *American Statistical Association*

RESEARCH STAFF

WESLEY C. MITCHELL, *Director*

MOSES ABRAMOVITZ
ARTHUR F. BURNS
SOLOMON FABRICANT
MILTON FRIEDMAN
SIMON KUZNETS
FREDERICK R. MACAULAY
FREDERICK C. MILLS
RAYMOND J. SAULNIER
LEO WOLMAN
RALPH A. YOUNG

Relation of the Directors to the Work of the National Bureau of Economic Research

1. The object of the National Bureau of Economic Research is to ascertain and to present to the public important economic facts and their interpretation in a scientific and impartial manner. The Board of Directors is charged with the responsibility of ensuring that the work of the Bureau is carried on in strict conformity with this object.

2. To this end the Board of Directors shall appoint one or more Directors of Research.

3. The Director or Directors of Research shall submit to the members of the Board, or to its Executive Committee, for their formal adoption, all specific proposals concerning researches to be instituted.

4. No study shall be published until the Director or Directors of Research shall have submitted to the Board a summary drawing attention to the character of the data and their utilization in the study, the nature and treatment of the problems involved, the main conclusions and such other information as in their opinion would serve to determine the suitability of the report for publication in accordance with the principles of the Bureau.

5. A copy of any manuscript proposed for publication shall also be submitted to each member of the Board. For each subject to be so submitted a special committee shall be appointed by the President, or at his designation by the Executive Director, consisting of three Directors selected as nearly as may be one from each general division of the Board. The names of the special manuscript committee shall be stated to each Director when the summary and report described in paragraph (4) are sent him. It shall be the duty of each member of the committee to read the manuscript. If each member of the special committee signifies his approval within thirty days, the manuscript may be published. If each member of the special committee has not signified his approval within thirty days of the transmittal of the report and manuscript, the Director of Research shall then notify each member of the Board, requesting approval or disapproval of publication, and thirty additional days shall be granted for this purpose. The manuscript shall then not be published unless at least a majority of the entire Board and a two-thirds majority of those members who shall have voted on the proposal within the time fixed for the receipt of votes on the publication proposed shall have approved.

6. No manuscript may be published, though approved by each member of the special committee, until forty-five days have elapsed from the transmittal of the summary and report. The interval is allowed for the receipt of any memorandum of dissent or reservation, together with a brief statement of his reasons, that any member may wish to express; and such memorandum of dissent or reservation shall be published with the manuscript if he so desires. Publication does not, however, imply that each member of the Board has read the manuscript, or that either members of the Board in general or of the special committee have passed upon its validity in every detail.

7. A copy of this resolution shall, unless otherwise determined by the Board, be printed in each copy of every National Bureau book.

(Resolution adopted October 25, 1926 and revised February 6, 1933 and February 24, 1941)

Financial Research Program: Committee

In the conduct of this and other studies under its program of research in finance the National Bureau of Economic Research has benefited from the advice and guidance of its Committee on Research in Finance. The functions of this committee are to review and supervise the specific research plans of the staff of the Financial Research Program. The membership includes:

WINFIELD W. RIEFLER, Chairman—*Institute for Advanced Study*

RALPH A. YOUNG, Secretary—*University of Pennsylvania; Director, Financial Research Program*

WILLIAM J. CARSON—*University of Pennsylvania; Executive Director, National Bureau of Economic Research*

THOMAS JEFFERSON COOLIDGE—*Chairman, United Fruit Company*

DAVID FRIDAY—*Chairman, National Bureau of Economic Research; Consulting Economist*

E. A. GOLDENWEISER—*Director, Division of Research and Statistics, Board of Governors of the Federal Reserve System*

F. CYRIL JAMES—*Principal and Vice-Chancellor, McGill University*

WALTER L. MITCHELL, JR.—*Director of Surveys, Research and Statistical Division, Dun and Bradstreet, Inc.*

WESLEY C. MITCHELL—*Columbia University; Director of Research, National Bureau of Economic Research*

SHEPARD MORGAN—*Vice-President, Chase National Bank; Treasurer, National Bureau of Economic Research*

RAYMOND J. SAULNIER—*Barnard College, Columbia University; Research Staff, National Bureau of Economic Research*

DONALD S. THOMPSON—*Chief, Division of Research and Statistics, Federal Deposit Insurance Corporation*

ROBERT B. WARREN—*Institute for Advanced Study*

JOHN H. WILLIAMS—*Littauer School, Harvard University; Vice-President, Federal Reserve Bank of New York*

LEO WOLMAN—*Columbia University; Research Staff, National Bureau of Economic Research*

DONALD WOODWARD—*Research Assistant to the President, Mutual Life Insurance Company*

Risk Elements in Consumer Instalment Financing

(Technical Edition)

BY DAVID DURAND

Financial Research Program
Studies in Consumer Instalment Financing 8

NATIONAL BUREAU OF ECONOMIC RESEARCH

PRINTED IN THE UNITED STATES OF AMERICA BY
THE HADDON CRAFTSMEN, INC., CAMDEN, N. J.

Preface

THIS study presents an analysis of certain factors which are relevant to the selection of credit risks and the determination of credit standards in the field of consumer instalment financing. It constitutes one phase of the investigation in this field, initiated in 1938 by the National Bureau of Economic Research and supported by special grants from the Association of Reserve City Bankers and the Rockefeller Foundation. A study of consumer instalment financing was originally recommended by the National Bureau's Exploratory Committee on Financial Research in its report submitted in 1937, and the broad purposes of such a study were set forth as follows:

"Instalment financing of consumer purchasers withstood the strain of the depression so well and showed such relatively small losses throughout the crisis as compared with many other types of credit instrument that banks and other financial agencies, pushed to find outlets for surplus funds, are now expanding rapidly in this field. This expansion, moreover, is assuming a competitive form, with respect not only to interest rates and other financial charges, but also to the down payment, the term of loan, the security, and the amount extended in relation to the income of the borrower. As a result, pressure is being brought to bear to relax the strictness of the procedures that tended to safeguard instalment financing during the depression. The Committee feels that, in view of its potentialities, this situation deserves careful analysis. At present, it is impossible to decide with any

confidence whether these modifications of procedure are justified or whether they constitute introduction of credit standards which are far too lax and which may have serious repercussions. In the present state of knowledge, such judgments cannot be based on data drawn from broad experience; they must be largely expressions of opinion. It is essential, the Committee holds, that an effort be made to gather all the available data on this type of financing for the purpose of identifying those credit standards which are sound and have stood the test of experience."

In the five institutional studies previously prepared and published under the consumer instalment financing project—dealing with personal finance companies, sales finance companies, industrial banking companies, consumer financing departments of commercial banks, and government agencies of consumer instalment credit—we presented separate analyses of credit experience in the several areas represented by these agencies. The present study brings together the findings of the individual studies, and makes an integrated analysis of risk factors in the entire field of consumer financing.

The raw materials for this study consisted of about 7,200 reports on loans actually made by 37 firms engaged in consumer instalment financing. These firms included 21 personal loan departments of commercial banks, 2 personal finance companies, 10 industrial banking companies, 3 automobile finance agencies and 1 appliance finance company. Although the basic data were supplied by a variety of firms in different areas, certain tendencies appeared consistently in most of the samples supplied.

Highly refined statistical methods were employed in this study, in order to assure precise results as well as to test the applicability of such methods to the problems involved. But since many companies may not find feasible the use of elaborate statistical methods, we have limited the discussion in the main text to procedures which are simpler, easier, and

less expensive, and which any company can apply to its own records in order to test its risk experience. The technical discussion of statistical theory and methods has been confined to three appendices. Since these appendices will be of interest chiefly to statisticians with specialized mathematical training, the study has been published in two editions, and the appendices have been eliminated from one of them. This is the technical edition, with appendices.

We welcome the opportunity to express indebtedness to the following firms, which cooperated, at considerable expense to themselves, in furnishing data or other assistance for this study:

Bank of the Manhattan Company
The City National Bank and Trust Company, Columbus, Ohio
The City National Bank and Trust Company, Kansas City, Missouri
Corn Exchange National Bank and Trust Company, Philadelphia
The Equitable Trust Company, Baltimore
The First National Bank of Boston
The First National Bank of Kansas City, Missouri
The First National Bank and Trust Company in Macon, Georgia
First National Bank and Trust Company of Minneapolis
First Wisconsin National Bank of Milwaukee
The Fourth National Bank, Columbus, Georgia
The Liberty National Bank and Trust Company of Savannah
Manufacturers Trust Company, New York
Midland National Bank and Trust Company, Minneapolis
National Bank of Tulsa
The National City Bank of New York

The National Exchange Bank of Augusta, Georgia
The Pennsylvania Company for Insurances on Lives and Granting Annuities, Philadelphia
Security-First National Bank of Los Angeles
Springfield National Bank, Springfield, Massachusetts
Trust Company of Georgia, Atlanta

Associates Investment Company, South Bend, Indiana
General Motors Acceptance Corporation, New York, New York
The National Shawmut Bank of Boston
Reserve Discount Company, St. Louis, Missouri

American Investment Company of Illinois, St. Louis, Missouri
Beneficial Industrial Loan Corporation, Newark, New Jersey
Household Finance Corporation, Chicago, Illinois

Citizens Savings and Loan Corporation, Chattanooga, Tennessee
The Community Consumer Discount Company, Warren, Pennsylvania
Community Savings and Loan Company, Parkersburg, West Virginia
Indianapolis Morris Plan, Indianapolis, Indiana
The Morris Plan Bank of Virginia, Richmond
The Morris Plan Industrial Bank of New York
Peoria Finance and Thrift Company, Peoria, Illinois
Progressive Company, Incorporated, New Orleans, Louisiana
Royal Industrial Bank, Louisville, Kentucky
Thrift, Incorporated, Des Moines, Iowa
Thrift, Incorporated, Evansville, Indiana

The collection and analysis of the data presented many difficult technical problems, and much experimental statistical work was required to determine the most appropriate treatment of the material. Mr. Durand, who has been in charge of the analysis from its beginning, has resolved these problems with great skill, patience, and resourcefulness.

By pointing the way to a recurrent statistical testing of credit experience by institutions engaged in consumer instalment financing, Mr. Durand has made a unique contribution to credit practices in the field, and we hope that the completion of this study will stimulate further investigation into the problem of such credit standards. In modern interest theory, much emphasis is placed on credit risk as a factor affecting the gross charge to the borrower, but little attention is given to the elements that comprise or affect risk. By identifying and indicating the role of some of these elements in the field of consumer instalment credit, Mr. Durand's study affords an empirical basis for the elaboration of the risk problem in this single sphere of interest theory.

RALPH A. YOUNG
Director, Financial Research Program

April 1941

Author's Acknowledgments

I WISH to express sincere appreciation to those who have contributed data for this study and to those who have given valuable assistance in its organization and development. I am particularly indebted to James W. Angell, Columbia University; Milan V. Ayres, American Finance Conference; Wilfred Helms, Household Finance Corporation; Ross I. Hewitt, General Motors Acceptance Corporation; Harold Hotelling, Columbia University; Frederick C. Mills, Columbia University; M. R. Neifeld, Beneficial Industrial Loan Corporation; L. M. Robitshek, American Investment Company of Illinois; and Theodore Yntema, University of Chicago.

I am also most grateful to the members of the financial research staff of the National Bureau of Economic Research who assisted in the planning and execution of the study; to R. J. Saulnier, John M. Chapman, Sidney S. Alexander, and Carl Kaysen, for suggestions and advice; to Dorothy Wescott, who edited the manuscript; to Aileen Barry, Catherine Connolly, and Mary Deeley, for assistance in tabulation.

Finally, I wish to extend thanks to Dr. R. A. Young, Director, and Dr. Winfield W. Riefler, Chairman, of the Financial Research Program, who have been a constant source of assistance and inspiration.

DAVID DURAND,
Financial Research Staff
(National Bureau of Economic Research)

Contents

Tables

APPENDIX A

A Note on the Theory of Discriminant Functions

Appendix A

A Note on the Theory of Discriminant Functions

VIEWED in the abstract, the present problem of statistical analysis is one of differentiating two species by means of a set of measurements; it is analogous to some of the problems of biology in which two varieties of plants or other organisms are differentiated on the basis of length of leaf, breadth of stem, etc. The two species under consideration in this study are the good and bad loans of consumer instalment lending, or rather the borrowers who repay their loans and those who fail to repay. This twofold classification, as we have pointed out, is somewhat artificial because loans or borrowers vary considerably in quality; but the distinction is useful and, roughly speaking, reasonably valid. The set of measurements includes information concerning borrower's income, occupation, sex, stability of residence, and the like. Again, to speak of measuring characteristics such as occupation, which is classified qualitatively and not quantitatively, may not be strictly correct, but in a broad sense the concept is satisfactory.

Statistical theorists have given considerable attention to the problem of differentiating two species by a set of measurements, and they have advanced the method of discriminant functions to solve it. This method permits an investigator to weight several credit factors according to their relative importance, and to allow for interrelationships between factors, which are extremely hard to account for by other approaches. A brief discussion of the theory underlying the method will be useful background for the study of good- and bad-loan samples.

Unfortunately, discriminant functions are usually determined

on the rather restrictive assumptions that each species considered has the multivariate normal distribution, and that the two species differ only in the average values of the measurements or variates—in other words, that the standard deviation of the variates and the coefficients of correlation between them are the same for each species. These conditions are not met in the good- and bad-loan samples; hence the method in question is not strictly applicable. Nevertheless, for illustrative purposes, its value is sufficiently great to warrant detailed attention.

The problem of differentiating two species by a set of measurements may be introduced by a discussion of the one-variate case. Assume the two species are normally distributed with respect to the distinguishing criterion. Each distribution has variance σ^2; but the means are different—say $+\frac{a}{2}$ and $-\frac{a}{2}$, so that the difference between them is a. The two species then have the probability distributions

$$P(A) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x-\frac{a}{2}\right)^2}{2\sigma^2}} dx, \qquad P(B) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x+\frac{a}{2}\right)^2}{2\sigma^2}} dx.$$

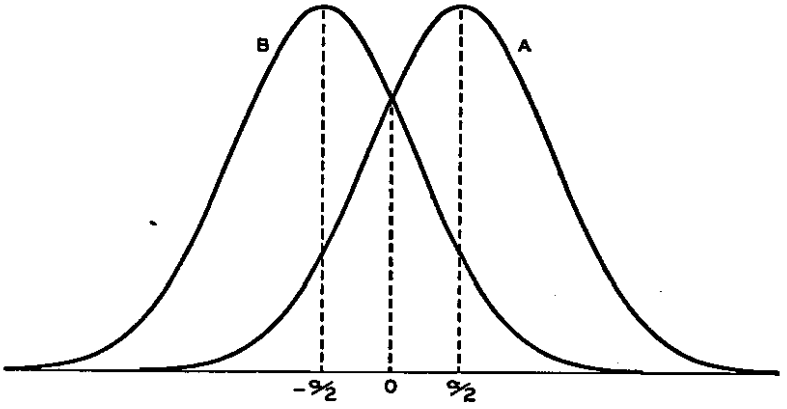
If species A and species B are equally numerous, the distributions may be represented by two congruent curves, as in Figure 1. To make the example concrete, imagine that A represents good loans, that B represents bad loans, and that the distinguishing criterion is number of years at present occupation.

The ratio $\frac{P(A)}{P(B)} = e^{\frac{ax}{\sigma^2}}$ is the ratio of the relative frequency of A's to B's in a small region around x. The ratio is an increasing function, approaching 0 as x approaches negative infinity, and approaching positive infinity as x approaches positive infinity. When x equals 0, the ratio equals one, indicating that in this region A's and B's are equally numerous. Because the ratio is an increasing function, all regions to the right of 0

contain more A's than B's, and conversely for all regions to the left of 0.

If species A and species B are to be differentiated on the basis of the value of x, several schemes are possible. One common scheme is to use the point 0, the midpoint between the means, as a criterion; values greater than 0 are classified as probably belonging to group A, and vice versa. Under this scheme the probability of misclassifying either an A or a B, P(Mis), is the ratio of the area of the portion of the A curve

Figure 1



left of 0 to the total A area, which is the same as the area of the B portion right of 0 to the total B area. P(Mis) is therefore equal to one-half the probability that the absolute value of a normal variate will exceed the absolute value of the ratio $a/2\sigma$. The ratio a/σ, or v, will be used in the future as a measure of the effectiveness of a criterion as a means of differentiating the two species. P(Mis) = $\frac{1}{2}$ when v is 0; it decreases as v becomes larger, approaching 0 as v becomes infinite. The quantity 1 − P(Mis), the probability of classifying correctly, varies from $\frac{1}{2}$ to 1 as v varies from 0 to infinity. Earlier in this study we have used the quantity 1 − 2P(Mis), which we have called the

efficiency index, to measure the effectiveness of the variate x as a means of distinction. This index, which varies from 0 to 1, can be expressed in terms of the ratio v by the following integral:

$$\text{Index} = \frac{1}{\sqrt{2\pi}} \int_{-v/2}^{v/2} e^{-\frac{t^2}{2}} dt$$

Equally numerous species differentiated by the midpoint between the means are a special case of a much more general situation. In credit analysis the generalization is desirable, for the special case is far from realistic. Good loans and bad loans are not equally numerous. If the ratio of good to bad—i.e., A to B—is k, then the relative frequency ratio

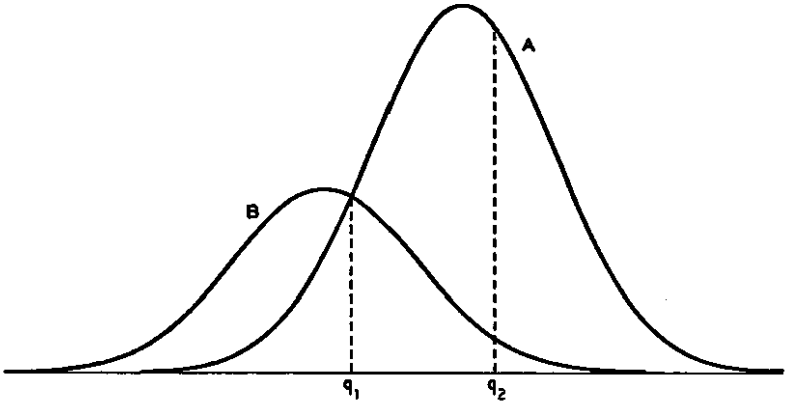
$$\frac{P(A)}{P(B)} = ke^{\frac{ax}{\sigma^2}}$$

is no longer equal to unity when x is zero; it is equal to unity at some other point q_1, which depends on k, a, and σ. But the point q_1, where $\frac{P(A)}{P(B)}$ is unity, is not a satisfactory point of demarcation because the net loss on a bad loan is likely to be considerably greater than the net profit on a good loan; the suitable point, q_2, is determined by equating $\frac{P(A)}{P(B)}$ to the ratio of the average profit on good loans to the average loss on bad loans. In risk selection, two points of demarcation, $q_2'' > q_2'$, may be required in place of only one. For example, applicants to the right of q_2'' could be accepted unconditionally; applicants to the left of q_2' could be rejected unconditionally; and those between q_2'' and q_2' could be given a more rigorous investigation and be required to furnish additional collateral.

For the general case, the concept of the probability of misclassification is substantially altered. Instead of one simple quantity, there are now four as follows: (I) the probability that species A will be misclassified; (II) the probability that species B will be misclassified; (III) the probability that an observation

with a value of x greater than the critical value (q_2) will be misclassified; (IV) the probability that an observation with a value less than q_2 will be misclassified. In Figure 2, (I) is represented by the fraction of curve A to the left of the critical value q_2; (II) by the fraction of B to the right of q_2; (III) by the ratio of the tail of B (to the right of q_2) to the sum of the tails of A and B; and (IV) by the ratio of the main portion of A (to the left of q_2) to the sum of the main portion of A and the main portion of B.

Figure 2



In practice, all these values can be determined from tables of the normal curve. These four quantities are not entirely independent; they can be reduced to two quantities. For example,

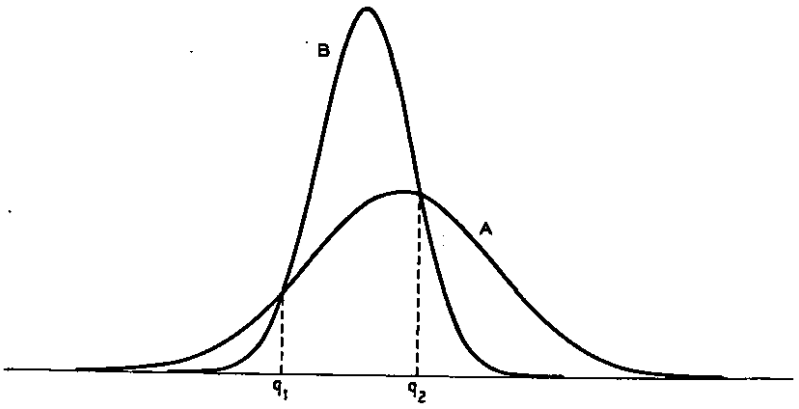
$$(III) = \frac{(II)}{K[1 - (I)] + (II)}$$

$$(IV) = \frac{K(I)}{1 - (II) + K(I)},$$

where K is the ratio of A's to B's. In the special case, where the two species are equally numerous and where 0 is the point of demarcation, P(Mis) = (I) = (II) = (III) = (IV).

A new set of complications is introduced when the two species have different variances as well as different means. The situation is illustrated in Figure 3, where the A variance is larger than the B variance. For the case of equal variances, the logarithm of the ratio $\frac{P(A)}{P(B)}$ $\left(\text{equal to } \frac{\alpha x}{\sigma^2}\right)$ is the equation of an upward sloping straight line through the origin; all values are possible from negative infinity to positive infinity. This means that the ratio of A's to B's can be increased indefinitely

Figure 3



by taking a region to the right of a sufficiently large value of x, and conversely. With unequal variances, however, the situation is entirely changed. The logarithm of the probability ratio represents a second degree parabola. In general, the relative frequency ratio is unity at two points, q_1 and q_2. In all regions between these two points, B's are preponderant, but the ratio of B's to A's is everywhere bounded. In the two external regions, the A's are preponderant, and the ratio of A's to B's can be increased indefinitely by taking sufficiently large or sufficiently small values of x.

When several variates or criteria are available for differenti-

ating the two species, the one dimensional case, already discussed, can be generalized. The appropriate method is by means of discriminant functions, which have been developed by R. A. Fisher and a few other writers.[1] Fisher's discriminant function is a linear function of n-variables,

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

where the x's represent the p criteria available for differentiation. This function has a mean for the A species of $\overline{Z}_A = \Sigma l_i\overline{x}_i$ where $\overline{x}_i$ is the mean of the i^{th} variate for the A species; the function has a similar mean $\overline{Z}_B$ for the B species, and a pooled variance (based on both species) of $s_z^2 = \Sigma\Sigma l_i l_j s_{ij}$ where the s_{ij}'s are the pooled variances and covariances of the x's. Here the means, the variances, and the covariances refer to some specific sample. The problem is to determine the coefficients l_i so that the ratio $U^2 = \dfrac{(\overline{Z}_A - \overline{Z}_B)^2}{s_z^2}$ will be maximized. This is accomplished by solving the following set of equations for the l's:[2]

$$\begin{aligned} s_{11}l_1 + s_{12}l_2 + \ldots\ldots + s_{1p}l_p &= a_1 \\ s_{21}l_1 + s_{22}l_2 + \ldots\ldots + s_{2p}l_p &= a_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ s_{p1}l_1 + s_{p2}l_2 + \ldots\ldots + s_{pp}l_p &= a_p \end{aligned} \qquad (1)$$

In these equations a_i is the mean difference $\overline{x}_i - \overline{x}_i'$, and

$$s_{ij} = \frac{1}{n+n'}\,[\Sigma(x_i - \overline{x}_i)\,(x_j - \overline{x}_j) + \Sigma(x_{i'} - \overline{x}_{i'})\,(x_{i'} - \overline{x}_{j'})],$$

where n is the number of degrees of freedom in one sample and n′ is the number in the other sample. The solution is

$$l_i = \sum_j \frac{a_j s^{ij}}{|s_{ij}|},$$

[1] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, part 2 (September 1936) pp. 179–88; and "The Statistical Utilization of Multiple Measurements," *ibid.*, vol. 8, part 4 (August 1938) pp. 376–86.

[2] Fisher presents these equations in terms of the actual sums S_{ij} instead of the covariances s_{ij}; the result is to multiply the l's by a constant.

where $|s_{ij}|$ is the determinant of the s_{ij}'s and s^{ij} is the cofactor of s_{ij} in that determinant.

A somewhat different approach, which yields the same results with the proper assumptions, is to investigate the relative frequency of species A to species B in various regions of the p-dimensional variate space. Assume two multivariate normal distributions

$$P(A) = Ce^{-1/2 \Sigma\Sigma Q_{ij}\left(x_i-\frac{\alpha_i}{2}\right)\left(x_j-\frac{\alpha_j}{2}\right)} dx_1 \ldots dx_p \qquad (2)$$
$$P(B) = Ce^{-1/2 \Sigma\Sigma Q_{ij}\left(x_i+\frac{\alpha_i}{2}\right)\left(x_j+\frac{\alpha_j}{2}\right)} dx_1 \ldots dx_p,$$

which are identical except for the mean values of the variates. The Q_{ij}'s and the α_i's are supposed to be true population parameters and not sample estimates. In this particular form, which entails no loss of generality, α_i is the difference between the i-mean of the A's and the i-mean of the B's, and 0 is the midpoint between those means; but other forms in which the midpoints are not 0 are sometimes convenient. The ratio $\frac{P(A)}{P(B)}$ has the form $e^{\Sigma\Sigma Q_{ij}x_i\alpha_j}$, which may also be written $e^{\Sigma x_i\lambda_i}$ where $\lambda_i = \sum_j Q_{ij}\alpha_j$.

The equation $\frac{P(A)}{P(B)} = e^{\Sigma x_i\lambda_i} = K$ is the locus of all points in the vicinity of which the ratio of A's to B's is K. This can be transformed into

$$\Sigma x_i\lambda_i = \log_e K,$$

which is the equation of a hyperplane. In particular $\Sigma x_i\lambda_i = 0$ is the equation of a hyperplane through the origin, which is the locus of all points in whose vicinity A's and B's are equally numerous. Since the matrix of Q_{ij} is the inverse of that of σ_{ij}, the covariances of the x's,

$$\lambda_i = \sum_j \alpha_j \frac{\sigma^{ij}}{|\sigma_{ij}|}$$

This is the same as the solution of (1) if $s_{ij} = \sigma_{ij}$ and $a_j = \alpha_j$.

The function $Z = \Sigma x_i\lambda_i$ provides a unique means of differentiating the two species. According to (2), the function Z is nor-

mally distributed with variance $\sigma_z^2 = \Sigma\Sigma\lambda_i\lambda_j\sigma_{ij}$; it has a mean for the A species of $\overline{Z}_A = \sum_{i=1}^{p}\frac{\lambda_i\alpha_i}{2}$ and for the B species o $\overline{Z}_B = -\sum_{i=1}^{p}\frac{\lambda_i\alpha_i}{2}$, where $\frac{\alpha_i}{2}$ is the A-mean of x_i and $-\frac{\alpha_i}{2}$ is the B-mean. The function Z therefore transforms the multivariate problem into a one-variate problem exactly analogous to that considered earlier.

If A and B are equally numerous, all regions for which Z is greater than 0, which is the midpoint between $\overline{Z}_A$ and $\overline{Z}_B$, contain a preponderance of A's, and conversely. If A's are K times as numerous as B's, and if some adjustment must be made to equate the average loss on bad loans to the average profit on good loans, then an alternative point of demarcation Z_q can be determined.

In the one-variate case with normal distributions and equal variances, the ratio v was advanced as a measure of the effectiveness of the variate as a differentiator. Two other measures, the probability of misclassification and the efficiency index, were also introduced, but for the case in point these measures depend only on v and are merely supplementary to it. For the multivariate case, the ratio T is exactly analogous to v in the one variate case; it serves as a measure of the effectiveness of the discriminant function as a differentiator. The probability of misclassification and the efficiency index for a discriminant function are determined by T just as they were determined by v for one variate. It is interesting to note that U, the sample estimate of T, is related to Hotelling's generalized T^2 and to the D^2-statistic of Bose and Roy by the following:

$$U = T\sqrt{\frac{n + n' + 2}{(n+1)(n'+1)}} = \sqrt{pD^2}\ ^3$$

[3] By definition $U = \frac{\Sigma l_i a_i}{\sqrt{\Sigma\Sigma l_i l_j s_{ij}}}$. The numerator of this fraction can be rewritten $\frac{\Sigma\Sigma a_i a_j s^{ij}}{|s_{ij}|}$ since $a_i l_i = a_i \Sigma_j a_j \frac{s^{ij}}{|s_{ij}|}$; moreover, the quadratic form in the denomi-

(where $n+1$ is the number of cases in one sample; $n'+1$ is the number in the other samples; and p is the number of variates).

The ratio T cannot be smaller than any of the individual ratios v, and in general it will be larger. It may be considerably or only slightly larger; and if it is only slightly larger, the necessary labor of computing the discriminant function may be hardly worthwhile. Consideration of the conditions that make for a larger ratio and those that make for a small one is therefore pertinent.

In general, the computation of the discriminant function and of the ratio T is a difficult task, which grows more difficult as the number of variates increases; but for the special case of complete independence of variates, the computation is almost simple. For the case of complete independence $\sigma_{ij} = 0$ except when $i = j$; therefore, $\lambda_i = \frac{\alpha_i}{\sigma_i^2}$. This means that the discriminant function can be computed as soon as the α's and σ's are known. The ratio T, equal to

$$\frac{\Sigma\lambda_i\alpha_i}{\sqrt{\Sigma\Sigma\lambda_i\lambda_j\sigma_{ij}}},$$

simplifies to

$$\frac{\Sigma\frac{\alpha_i^2}{\sigma_i^2}}{\sqrt{\Sigma\frac{\alpha_i^2}{\sigma_i^2}}}$$

and thence to

$$\sqrt{\Sigma\frac{\alpha_i^2}{\sigma_i^2}},$$

---

nator, $|l_{ij}||s_{ij}|$, is equal to its inverse, $\Sigma\Sigma a_i a_j \frac{s^{ij}}{|s_{ij}|}$, for the same reason (Cf. Bôcher, *Introduction to Higher Algebra*, 1936, p. 160). Therefore, $U = \sqrt{\frac{\Sigma\Sigma a_i a_j s^{ij}}{|s_{ij}|}}$. Since $T^2 = \frac{\Sigma\Sigma a_i a_j s^{ij}}{|s_{ij}|} \cdot \frac{n'+n+2}{(n'+1)(n+1)}$, and since $D^2 = \frac{1}{p}\Sigma\Sigma a_i a_j \frac{s^{ij}}{|s_{ij}|}$ (cf. Appendix C, pp. 146, 148, 150-51) the relation of U to T^2 and D^2 follows easily.

which will be written hereafter $\sqrt{\Sigma v_i^2}$. This also is extremely easy to compute when the ratios $\frac{\alpha_i}{\sigma_i} = v_i$ are known.

It would be very convenient if the expression $\sqrt{\Sigma v_i^2}$ could be used as a first approximation for the true value of T. One might be able to predict whether the actual computation of a discriminant function would be justified by the results obtainable. The following pertinent relation has been worked out for the case of two variates; but a simple generalization for more than two variates appears to be impossible.

The true ratio T is equal to $\sqrt{v_1^2 + v_2^2}$ at two points, $\rho = 0$ and $\frac{2}{\frac{v_1}{v_2} + \frac{v_2}{v_1}}$ (where ρ is the correlation coefficient between x_1 and x_2). The ratio reaches a minimum value of v_1 or v_2, whichever is larger, at the point $\rho = \frac{v_1}{v_2}$ or $\frac{v_2}{v_1}$, whichever is less than one in absolute value. Naturally the minimum point lies between 0 and $\frac{2}{\frac{v_1}{v_2} + \frac{v_2}{v_1}}$. On either side of the minimum point, the ratio increases steadily, approaching infinity as ρ approaches ± 1.[4]

[4] For two variates $T = \left[\frac{\alpha_1^2\sigma_{22} - 2\alpha_1\alpha_2\sigma_{12} + \alpha_2^2\sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right]^{\frac{1}{2}}$ (see footnote 3). Dividing both numerator and denominator by $\sigma_{11}\sigma_{22}$, and writing $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, $v_1 = \alpha_1/\sqrt{\sigma_{11}}$, $v_2 = \alpha_2/\sqrt{\sigma_{22}}$, we get

$$T = \left[\frac{v_1^2 - 2v_1v_2\rho + v_2^2}{1 - \rho^2}\right]^{\frac{1}{2}}$$

When $|\rho|$ approaches unity, T becomes infinite except in two special cases: when $v_1 = v_2$ and ρ approaches one, or when $v_1 = -v_2$ and ρ approaches minus one, then $|T|$ approaches $|v_1| = |v_2|$. The derivative of T^2 with respect to ρ, which is

$$\frac{2\rho(v_1^2 + v_2^2) - 2v_1v_2(1 + \rho^2)}{(1 - \rho^2)^2}$$

is equal to zero at the point v_1/v_2 or v_2/v_1, whichever is less than one in absolute value. At this point T has a minimum value of v_1 or v_2, whichever is larger.

We now inquire: At what values of ρ will $T = \sqrt{v_1^2 + v_2^2}$? We get $v_1^2 - 2v_1v_2\rho + v_2^2 = (1 - \rho^2)(v_1^2 + v_2^2)$, whence $\rho = 0$ or $\frac{2}{v_1/v_2 + v_2/v_1}$.

There are, then, four different types of cases, which are illustrated in Figure 4. To make the example concrete, imagine that A represents good loans, that B represents bad loans, and the two correlated criteria for differentiation are number of years at present address and number of years at present occupation. In the first two of these (4a and 4b), the true ratio is higher than $\sqrt{v_1^2 + v_2^2}$; in the second two it may be higher or lower depending on the value of ρ.

A few concrete applications of this theory may be in order. Suppose that for stability of occupation $v = .5$, which corresponds to an efficiency index of about 20; and that for stability of residence $v = .4$, which corresponds to an efficiency index of 16. (These are approximately the efficiency indices actually obtained in the commercial bank samples.) If there is no correlation between stability of residence and stability of employment, the ratio T will be .64, which corresponds to an efficiency index of 25. But actually a positive correlation is to be expected. The situation is like that of Figure 4c below; if the correlation lies between 0 and $.976 = \frac{2}{\frac{.4}{.5} + \frac{.5}{.4}}$, the actual ratio will be less than .64. Furthermore, since the actual correlation is very likely to lie between 0 and .976, it is a fairly safe prediction that T will actually be less than .64. In the commercial bank samples an estimate of the correlation between stability of residence and stability of occupation was made from a small number of cases. The result, .15, was well within the limits of 0 and .976. (See Table B-3, p. 132.)

In the commercial bank samples no appreciable difference was found between the good- and bad-loan samples in connection with either borrower's income or amount borrowed. What then can be inferred about the ratio of amount borrowed to income? Under the assumptions of normality and of equal standard deviations and correlation coefficients, two definite conclusions are possible: (1) as a means of differentiating good

and bad loans, the ratio of the amount of the loan to borrower's income, which is just one possible way of combining amount and income, will be inferior to a linear discriminant function; (2) the discriminant function will not show any appreciable difference between good and bad loans. Under the assumed conditions, an independent study of the amount/income ratio, or any other combination of income and amount, would not be warranted. Actually, the distribution of loans according to the amount/income ratio was determined, and the results were negative.

Conclusions such as the above rest on the assumption of normality and the equality of standard deviations and correlation coefficients. Since these assumed conditions do not exist in the loan samples, any of the foregoing conclusions may be invalid. Situations that will upset almost any conclusions based on the theory of this chapter are easily invented. No standardized procedure can be worked out for handling such cases, for each one presents its own problem. A few examples will be shown.

Although a linear discriminant function is entirely appropriate for multivariate normal distributions with equal variances and covariances, it is not so appropriate for most other forms of distributions. For example, when the logarithms of the variates are distributed normally with equal variances and covariances, the appropriate discriminant function has the form

$$Z = \lambda_1 \log x_1 + \lambda_2 \log x_2 + \ldots,$$

for which we may conveniently substitute

$$Z' = e^{z} = x_1^{\lambda_1} x_2^{\lambda_2} \ldots.$$

A very interesting case occurs when there are only two variates. If $\lambda_1 = \pm\lambda_2$, as will be the case when $\alpha_1(\sigma_{22} \pm \sigma_{12}) = \alpha_2(\sigma_{12} \pm \sigma_{11})$, then the appropriate discriminant function will be x_1x_2 or $\frac{x_1}{x_2}$.

When the distributions are normal but with the variances and covariances of A unequal to those of B, the appropriate

Figure 4a

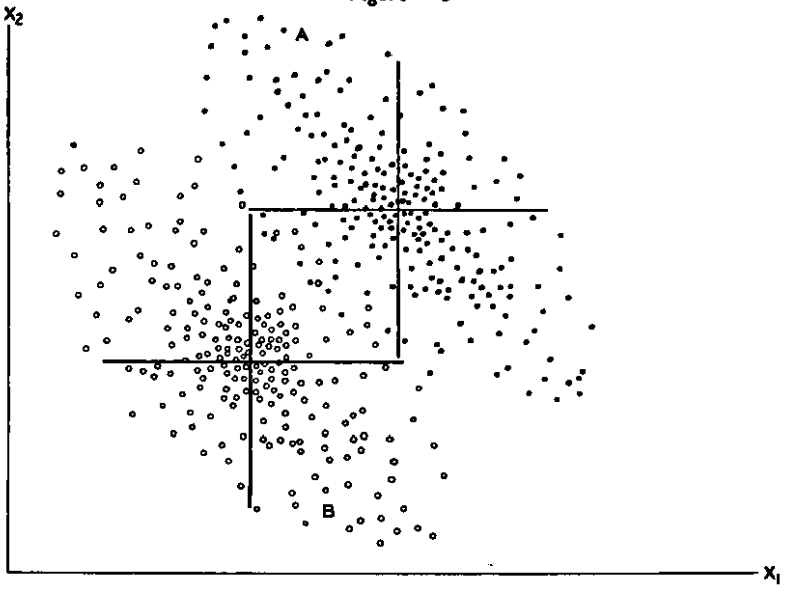


Figure 4b

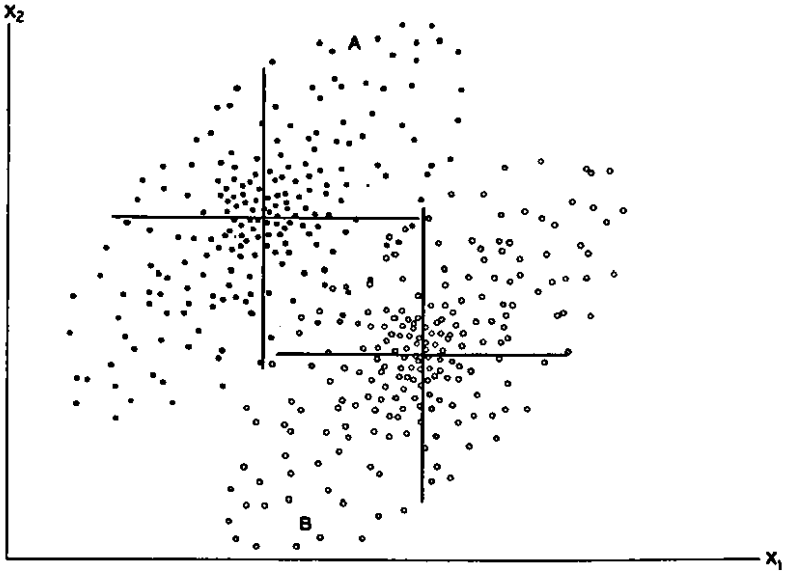


Figure 4c

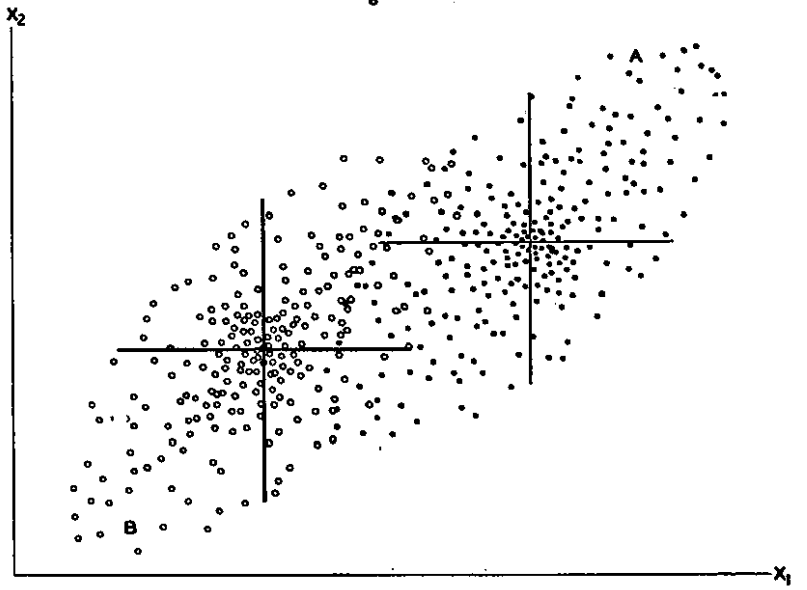
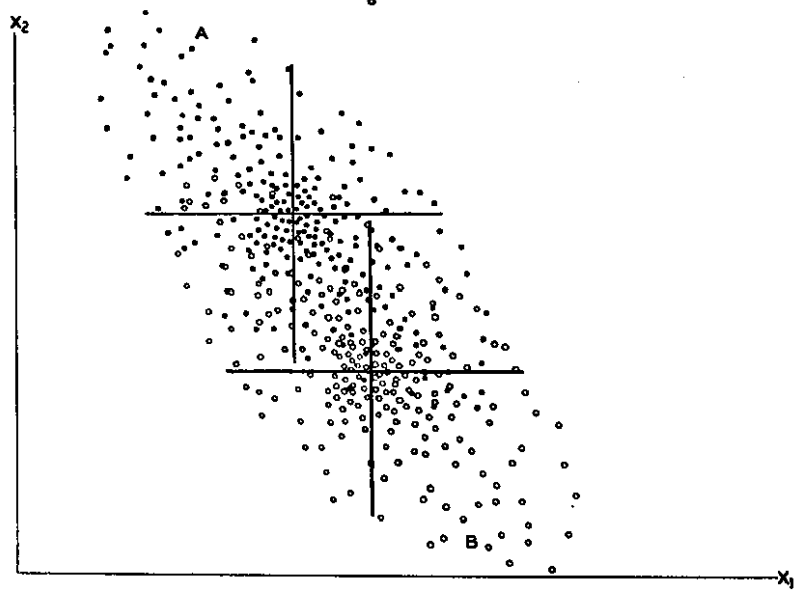


Figure 4d



discriminant function is a general second degree function. We have

$$\frac{P(A)}{P(B)} = \frac{C_A}{C_B} \frac{e^{-\Sigma\Sigma A_{ij}(x_i-\alpha_i)(x_j-\alpha_j)}}{e^{-\Sigma\Sigma B_{ij}(x_i-\beta_i)(x_j-\beta_j)}} =$$

$$\frac{C_A}{C_B} e^{-\Sigma\Sigma[(A_{ij}-B_{ij})x_ix_j-2x_i(\alpha_jA_{ij}-\beta_jB_{ij})+A_{ij}\alpha_i\alpha_j-B_{ij}\beta_i\beta_j]},$$

which indicates a discriminant function of the form

$$\Sigma\Sigma\lambda_{ij}x_ix_j + \Sigma\lambda_ix_j.$$

Such a function will not be normally distributed.

It is even conceivable that the means of the sample may be equal and that the only differences may be in the variances or covariances. A single example is cited by way of illustration. Assume only two variables, and assume the distributions are given by

$$P(A) = Ce^{-\frac{1}{2(1-\rho^2)}(x_1^2-2\rho x_1x_2+x_2^2)}dx_1dx_2$$

$$P(B) = Ce^{-\frac{1}{2(1-\rho^2)}(x_1^2+2\rho x_1x_2+x_2^2)}dx_1dx_2;$$

in other words, the means are equal; the variances are both unity; and the correlation coefficients are equal in absolute magnitude but opposite in sign, the A correlation being positive. (See Figure 5.) The probability ratio is

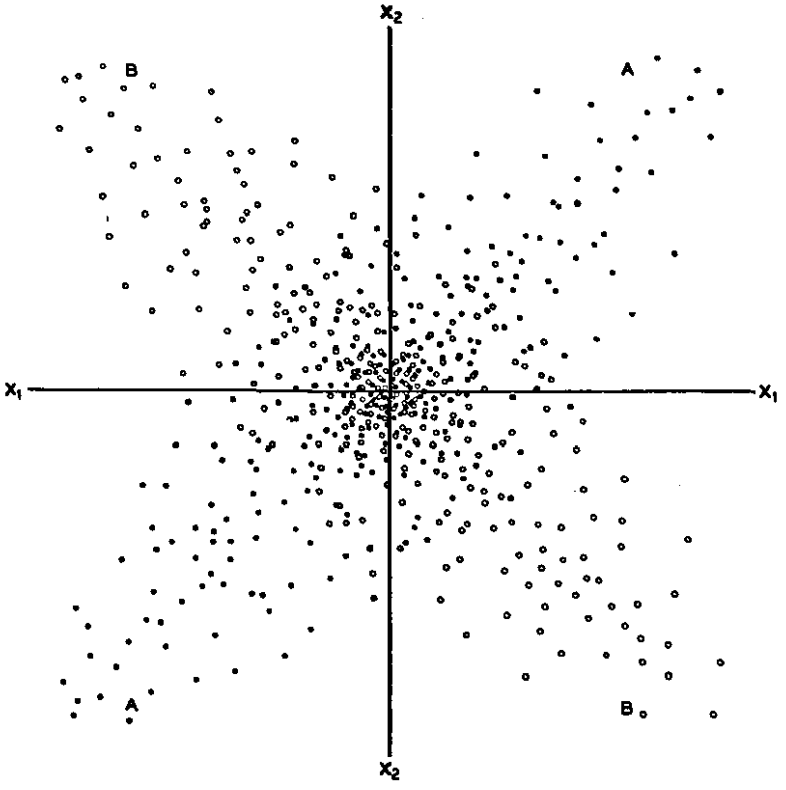
$$K = e^{\frac{2\rho x_1x_2}{(1-\rho^2)}}, \text{ whence } \log K = \frac{2\rho x_1x_2}{(1-\rho^2)}.$$

When K is greater (less) than one, the above equation represents a pair of hyperbolas lying in the lower right (left) and upper left (right) quadrants; and as K approaches ± 1, the hyperbolas approximate the coordinate axis. Thus, when A's and B's are equally numerous, all regions in the upper right and lower left quadrants contain a preponderance of A's.

Enough examples have been presented to show that for departures from ideal conditions a linear discriminant function is less appropriate than some other form, the precise nature of

which depends on the nature of the distribution. For special cases like the above, the task of determining the appropriate function would not be unduly onerous; but for more general cases the task would be next to impossible. Most practical

Figure 5



investigators will probably prefer to determine a linear function, even when the ideal conditions do not exist; and in many instances the resulting approximations will probably be satisfactory.

APPENDIX B

Application of the Method of Discriminant Functions to the Good- and Bad-Loan Samples

Appendix B

Application of the Method of Discriminant Functions to the Good- and Bad-Loan Samples

CONSIDERING the fundamental assumption of a dichotomous classification of loans, the problem of analysis is to discover differences between the good-loan and bad-loan distributions. The factors analyzed in this report fall into two rough categories: the qualitative attributes like occupation and marital status, and the quantitatively measurable variates like income and number of years at present address. Analysis of the qualitative attributes may be made by comparing the proportion of good loans in a given occupational group, for example, with the proportion of bad loans. Analysis of the quantitatively measurable factors can, of course, be carried out by the same process. The proportion of good loans in any income class can be compared with the proportion of bad loans; but one further step in the analysis is generally desirable and possible. A difference between the income distributions of the good and bad loans usually can be translated into a difference in mean or average income, a difference in the standard deviation about the mean, a difference in skewness, a difference in kurtosis, etc.

In the non-technical sections of this report, the distributions of all factors, quantitative and qualitative alike, are shown on the same basis; in all cases the percentage of good and bad loans in each of a small number of class intervals is determined; and for the quantitative factors no attempt is made to measure mean value, variance, skewness, kurtosis, etc. Nevertheless, a differ-

ence in mean values is frequently obvious. For example, the good-loan samples in Table 11 undoubtedly have a longer average tenure of employment than the bad-loan samples, although the amount of the difference is not readily apparent. Differences in other measures, such as variance or skewness or kurtosis, are much more obscure; and the difficulty of analyzing these differences is often great. On the whole, the analysis of the quantitative factors thus far has consisted of a rough attempt to determine differences between the means of the good and bad loans.

The analysis in Chapter 3 consists of a set of individual treatments of separate factors. The differences that were discovered between the samples of good and bad loans related only to separate factors—income distribution, occupational distribution, etc. This individualistic approach has its shortcomings, however. A more satisfactory approach would be to consider each of the samples as a single distribution in a number of variates. Any difference between two distributions could be used for the purpose of differentiation; for example, the correlation coefficient between tenure of residence and tenure of occupation might be one value for the good loans and another for the bad. In practice, however, differences between means are the most obvious and by far the easiest to handle, i.e., when quantitatively measurable factors are concerned. For this purpose the use of discriminant functions, described in Appendix A, has two distinct advantages; it provides a means by which a number of credit factors can be weighted and combined into an index of credit risk; and it helps to indicate when individual analyses may be specious because of correlation between factors.

The method of linear discriminant functions is the ideal method of analysis when the two populations have multivariate normal distributions with equal variances and covariances but differing means. In the good- and bad-loan samples, where the assumed conditions are not actually met, this method is no longer ideal, but it may be a useful approximation.

An experiment with discriminant functions was carried out for the used-car sample. Four factors were singled out for analysis: cash price, actual down payment in dollars, purchaser's monthly income, and length of contract. These factors were chosen because they are fundamentals from which a number of other factors can be derived. From the ratio of down payment to price, the percent down payment is derived. The difference between price and down payment is the unpaid balance, which is usually a fair approximation to the amount of the note; the ratio of the unpaid balance to contract length is an approximation of the amount of the monthly payment; and the ratio of this last factor to income is an index of the burden of the debt upon the borrower's purchasing power. Instead of a separate investigation of all these derivative factors, a single discriminant function analysis of the four basic factors appears to be more systematic and more expedient.

The four selected factors have already received separate analysis. The distribution of cases was presented in Tables 4, 6, 7, and 8. A summary of these analyses is presented in Table B-1,

TABLE B-1

MEANS AND STANDARD DEVIATIONS OF NON-REPOSSESSED AND REPOSSESSED USED-CAR SAMPLE, BY PRICE, DOWN PAYMENT, INCOME, AND MATURITY

	Price	*Down Payment*	*Income*	*Maturity*
Mean (non-repossessions)	$410	$166	$172	13½ mos.
Mean (repossessions)	$344	$119	$148	13½ mos.
Difference	$66	$47	$24	0 mos.
Standard deviation (both samples)	$195	$89	$93	3.4 mos.
Ratio $\frac{\text{mean difference}}{\text{standard deviation}}$	.34	.53	.26	.00
Theoretical efficiency index[a]	13	21	10	0

[a] This index was not determined from the actual distribution of loans; it was computed from the ratio of mean difference to standard deviations by means of a table of the normal curve.

which presents means and standard deviations instead of percentage distributions. These values have been determined from the entire used-car sample of 484 non-repossessions and 485 repossessions of which 439 of the non-repossessions and 448 of the repossessions reported full data on price, contract length, down payment, and income.

This tabulation suggests that the first three variates are related to risk; that the order of importance is down payment, price, income; and that the last variate, contract length, is not related. As we have pointed out before, these conclusions may be specious if the correlation between the variates is high; and the correlation coefficients in Table B-2 indicate considerable correlation between some of the variates.

TABLE B-2

CORRELATION COEFFICIENTS FOR SELECTED FACTORS

Factor	*Price*	*Down Payment*	*Income*
Down payment	.87		
Income	.33	.29	
Length of contract	.62	.47	.05[a]

[a] Does not differ significantly from zero.

The discriminant function for these four factors was found to be $Z = d - .174p + .124i - 6.45m$, where d is the down payment in dollars, p is the price in dollars, i is the monthly income in dollars, and m is the length of contract in months.

The effectiveness of the function Z can be measured by the ratio of the difference between its two means (the mean for the good sample and the mean for the bad sample) to its standard deviation. The value of this ratio can be estimated without the actual computation and tabulation of the value of Z for each loan. The ratio is .63, which is an appreciable though not startling increase over the value of .53 for down payment alone; the corresponding efficiency indices computed from a table of the normal curve are 25 and 21. This increase is not striking;

if the factors had been independent, the ratio would have been .68, and the efficiency index would have been 27.

On the basis of these data we can now show that the individual analyses and their indications of the relative importance of factors are sometimes misleading. In the individual analysis, length of contract does not appear to be related to risk, for the good- and bad-loan samples have the same mean value. In the discriminant function Z, however, relation between contract length and risk does appear. Owing to the correlation between factors, the coefficient for length of contract is −6.45, which indicates that risks tend to improve as length becomes shorter. This inconsistency, as we have explained earlier, is attributable to the fact that few of the lower-priced used cars are financed on contracts of more than 12 months; for cars of the same price, the short terms are distinctly superior.

In the individual analysis, a high price appears to indicate good risk; but in the discriminant function, the price coefficient, −.174, indicates exactly the opposite. This apparent inconsistency can be explained by the high correlation between price and down payment. High price indicates good risk as long as it is accompanied by a high down payment, which is usually the case; but when down payments remain constant, the higher prices indicate poorer risks.

Nevertheless, the coefficients of the various factors are not entirely reliable as indices of the relative importance of the various factors. If the function Z is transformed to express the measurement of each variate in units of one standard deviation—a process analogous to the computation of the Beta-coefficients in multiple correlation—the transformed coefficients are somewhat more reliable, but they are not yet ideal. Transformed to units of one standard deviation, the discriminant function found above becomes $Z' = d - .382p + .131i - .246m$.

One possible way of measuring the relative importance of the factors is to determine discriminant functions for a number of combinations based on fewer than four factors. For combina-

TABLE B-3

CORRELATION COEFFICIENTS, MEAN DIFFERENCES, AND STANDARD DEVIATIONS, FOR SEVEN RISK FACTORS, COMPUTED FROM A COMMERCIAL BANK SUBSAMPLE OF 191 GOOD LOANS AND 190 BAD LOANS[a]

	Sex[b]	*Real Estate*[c]	*Number of Years at Present Address*	*Nature of Occupation*[d]	*Number of Years at Occupation*	*Bank Account*[e]	*Life Insurance*[f]
Correlation coefficients							
Real estate	−.09[g] .03[g]						
Number of years at present address	.17 −.03[g]	.16 .30					
Nature of occupation	.56 .11[g]	.09[g] .05[g]	.08[g] −.01[g]				
Number of years at occupation	.00[g] .03[g]	.24 .22	.12[g] .15	.09[g] .07[g]			
Bank account	−.04[g] .11[g]	.26 .21	.09[g] .02[g]	.21 .15	.25 .01[g]		
Life insurance	.20 .03[g]	.05[g] −.13[g]	−.04[g] −.12[g]	.11[g] .06[g]	.13[g] −.20	−.13[g] .12[g]	
Mean	.241 .090	.241 .090	5.937 5.160	1.094 .695	8.785 5.884	.492 .205	.235 .305
Difference	.151	.151	.777	.399	2.901	.287	.070
Standard Deviation	.303 .202	.303 .202	5.15 5.40	.562 .503	5.81 4.88	.353 .285	.300 .326

Footnotes will be found on page 133.

tions of down payment with each of the other factors, the results are as follows:

$$Z_1 = d + .214i$$
$$Z_2 = d - .232p$$
$$Z_3 = d - 12.5m$$

The theoretical ratios of mean difference to standard deviation were determined; they are .54, .58, and .60, respectively, and naturally enough, they lie between .53 for down payment alone and .63 for all four factors. Of these three combinations, that containing contract length has the highest ratio, but it is not strikingly better than the one containing price.

A second attempt with discriminant functions was made with the commercial bank sample. Here the number of available cases was so large that the drawing of a random subsample of 191 good loans and 190 bad seemed expedient. The computations were made entirely on the basis of this subsample.

Seven factors were selected for analysis: sex, stability of residence, stability of occupation, nature of occupation, bank account, life insurance, and real estate. Five of these factors are merely qualitative attributes incapable of quantitative measurement. As such, they are not directly subject to the discriminant function analysis; however, by assigning arbitrary numerical values to the qualitative categories the mechanical process of computing a discriminant function can still be followed. Thus women were given a value of 1, and men a value of 0; occupations were divided into three groups, with the poorest risk group having a value of 0, the middle group a value of 1, and the best a value of 2; cases with bank account were given a value of 1 compared with 0 for those with no bank account; and a similar process was used for life insurance and real estate. The means of the two samples, the standard deviations, and the correlation coefficients, are shown in Table B-3.

The form of the discriminant function obtained has been shown on page 86. This formula agrees well with the result

of the individual analyses except in regard to stability of residence; the negative weight given to stability of residence suggests that risk increases as residence becomes more stable, which is a direct contradiction of the individual analysis. This discrepancy seems to be traceable, however, to a substantial sampling error in the subsample.[1]

Values of Z were computed and tabulated for all loans in the commercial bank sample; and with slight modifications, the process was then extended to the industrial banking company sample. The efficiency index based on combined factors was in each case noticeably higher than that for any one of the individual factors. Since these efficiency indices were obtained from actual distributions and not from theoretical estimates, they are particularly important. The assumptions underlying the classical discriminant function approach were sadly lacking, and the function itself was determined from a relatively small subsample of the total available cases. Despite these serious drawbacks, the method produced concrete results.

SHORT-CUT METHODS FOR COMPUTATION, ON THE ASSUMPTION OF INDEPENDENCE

Ordinarily the process of computing a discriminant function is arduous; but when the factors in question are independent, the process is simplified. If the distributions are normal or approxi-

[1] The following percentage distribution of loans in the subsample, with an efficiency index of 8.3, is distinctly at variance with the corresponding distribution in the total sample, with an index of 13.8. The difference, which is not excessive in a sample of this size, is large enough to affect the discriminant function considerably.

	0–2 years	*2–6 years*	*6–10 years*	*10 years and over*
Subsample				
Good	29.8	34.0	7.9	28.3
Bad	32.6	36.9	10.5	20.0
Total sample				
Good	28.0	34.8	10.1	27.1
Bad	40.4	36.2	7.2	16.2

mately normal, a mere simplification of the standard procedure is appropriate. The equations (see p. 111)

$$\begin{aligned} s_{11}l_1 + s_{12}l_2 + \ldots &= a_1 \\ s_{12}l_1 + s_{22}l_2 + \ldots &= a_2 \\ \ldots\ldots\ldots\ldots & \end{aligned}$$

become

$$\begin{aligned} s_{11}l_1 &= a_1 \\ s_{22}l_2 &= a_2 \\ \ldots\ldots & \end{aligned}$$

in case of complete mutual independence. If a state approaching independence is suspected, the l's can be computed directly from the mean differences and the variances; and the resulting function will probably be a good approximation. If, however, the distributions depart markedly from normality, an alternative procedure may be preferable. This second short-cut method is based on the simple principle that the probability of two or more events may be computed, in the case of independence, merely by multiplying together the individual probabilities of the occurrence of the events.

Suppose that as far as factor A is concerned, the good and bad loans are distributed among p discrete classes. Let a'_i represent the percentage of good loans in the A_i class ($i = 1 \ldots p$); let a''_i represent the percentage of bad loans; then $\frac{a''_i}{a'_i}$ is the bad-loan relative. Similarly for factor B with q discrete classes, b'_j and b''_j represent the percentage of good and bad loans in

Footnotes for Table B-3 on page 130

[a] The upper figure of each pair refers to the good-loan sample; the lower figure refers to the bad-loan sample. The correlation coefficients and standard deviations can be appropriately averaged by pairs to obtain a pooled estimate of the supposedly equal value for both distributions. Since the numbers of cases in each sample are virtually equal, an unweighted arithmetic average will suffice.
[b] Males given a value of 0, females of 1.
[c] Non-owners of real estate given a value of 0, owners of 1.
[d] Better than average given a value of 2, average of 1, worse than average of 0.
[e] Those with bank accounts given a value of 1, those without of 0.
[f] Those with life insurance given a value of 1, those without of 0.
[g] Does not differ significantly from 0. See also footnote 3, p. 135.

class B_i, and $\frac{b''_j}{b'_j}$ is the bad-loan relative. On the assumption of independence, the expected percentages of loans belonging to both class A_i and class B_j are $a'_i b'_j$ and $a''_i b''_j$ with a bad-loan relative of $\frac{a''_i b''_j}{a'_j b'_i}$. The result can be generalized to any number of factors.

The generalized bad-loan relative $\frac{a''_i b''_j c''_k \ \ldots}{a'_i b'_j c'_k \ \ldots}$ will serve as a sort of discriminant function; if it is greater than one, it signifies a worse-than-average loan, and conversely. In actual practice, modifications of this procedure will be found convenient. The logarithm of the reciprocal of the bad-loan relative, which equals

$$\log \frac{a''_i}{a'_i} + \log \frac{a''_j}{a'_j} + \log \frac{a''_k}{a'_k} + \ldots$$

is probably the most fundamental. This function is positive for better-than-average loans and negative for worse-than-average.

This short-cut method may be combined with the classical method of discriminant functions. Suppose three variates a, b, c are normally distributed and highly correlated. A discriminant function

$$z = L_a a + L_b b + L_c c$$

would be determined. There would be two normal distributions, one for the good loans and one for the bad. A transformation[2] can be made so that these distributions take the form

[2] Probably the most convenient transformation is of the form

$$A = a - \frac{\bar{a}' + \bar{a}''}{2}, \quad B = b - \frac{\bar{b}' + \bar{b}''}{2}, \text{ etc.}$$

where $\bar{a}'$ is the a-mean of the good loans and $\bar{a}''$ is the a-mean of the bad loans, etc. The effect is to make the origin the midpoint between the means. Any transformation, however, that makes

$$L_a \bar{A}' + L_b \bar{B}' + L_c \bar{C}' = \frac{Dz}{2}$$

$$\frac{1}{\sigma_z\sqrt{2\pi}}e^{-\frac{(z+Dz/2)}{2\sigma_z^2}}dz,\ \frac{1}{\sigma_z\sqrt{2\pi}}e^{-\frac{(z-Dz/2)^2}{2\sigma_z^2}}dz$$

where Dz is the mean difference and 0 is the dividing line between better-than-average and worse-than-average cases. The bad-loan relative for any particular case is the ratio of the two

$$\frac{e^{-\frac{z^2+zDz+Dz^2/4}{2\sigma_z^2}}}{e^{-\frac{z^2-zDz+Dz^2/4}{2\sigma_z^2}}} = e^{-\frac{zDz}{\sigma_z^2}}$$

The natural logarithm of the reciprocal of this is $\frac{zDz}{\sigma_z^2}$; it will be positive for better-than-average and negative for worse-than-average loans. If some additional factors D, E, . . . are not correlated, the discriminant function for all factors will be

$$\frac{zDz}{\sigma_z^2} + \log_e\frac{d'_1}{d''_1} + \log_e\frac{e'_i}{e''_i} + \ldots$$

APPLICATION OF THE SECOND SHORT-CUT METHOD TO COMMERCIAL BANK SAMPLE

The evidence obtained from the available samples indicates that the factors under investigation are not entirely independent, but the degree of interdependence is surprisingly small. In Table B-3, which refers to the subsample of 191 good and 190 bad loans, the highest correlation coefficient (.56) is between occupation and sex in the good-loan sample, and the next highest (.30) is between stability of address and ownership of real estate in the bad-loan sample. These particular coefficients are more than large enough to be statistically significant, but most of the others are not.[3] Even the significant coefficients, however,

---

and

$$L_a\bar{A}'' + L_b\bar{B}'' + L_c\bar{C}'' = -\frac{Dz}{2}$$

will suffice.

[3] On the assumption of true independence in the parent universe, the standard error of the correlation coefficient is $\frac{1}{\sqrt{189}} = .073$ in a sample of 190 cases. Since the 5 percent significance level is .143 (.073 × 1.96, where 1.96 is the 5 percent value of t for the normal curve), all values of .14 or less for the coefficient may be considered non-significant.

are not sufficiently high to suggest a particularly close relationship; hence a situation approximating independence may, perhaps, be indicated.

Further evidence on independence is obtained from a series of 21 2 x 2 breakdowns of the commercial bank loans, one for each of the 21 possible pairs of the seven factors shown in Table B-3. For each of these factors the entire sample may be divided into two parts. In the case of some factors, like ownership of bank account, only two classifications are possible; for others like occupation, an arbitrary division is made so that the better risks are included in one classification and all the rest in another. For each pair of factors a two-way distribution may then be arranged by distributing all loans among four classes. Table B-4, which presents these data, will require some explanation. The first column is a percentage distribution of both good and bad loans by sex and real estate. The top figure (4.14) is the percent of females owning real estate among the good loans; beneath this figure is a similar percent (1.52) for the bad loans, followed by the bad-loan relative (.37). The next group of three figures (14.85; 6.99; .47) gives the percent of females not owning real estate and the bad-loan relative; the third group refers to males owning real estate; and the fourth refers to males not owning real estate. The second column gives the situation that would exist in a state of complete independence. The top figure (5.20) represents the expected proportion of females owning real estate among the good loans. This figure is determined by multiplying the total proportion of females among the good loans (4.14 + 14.85 = 18.99) by the total proportion of all persons owning real estate (4.14 + 23.23 = 27.37). Below this top figure is the expected proportion of females owning real estate among the bad loans (1.21), followed by the expected bad-loan relative (1.21 ÷ 5.20 = .23). All these expected figures can be calculated easily from the summary totals at the end of Table B-4.

This table permits comparison of the actual proportion of good or bad loans in any class with the proportion that would

be expected in case of complete independence; it also permits comparison of the actual bad-loan relative with the expected bad-loan relative. This last comparison is important; for as long as the actual and expected relatives are approximately equal, the second short-cut method of computing the discriminant function can be used with assurance.

In Table B-4 the expected and actual values of the bad-loan relatives are surprisingly similar in most cases. The four most noticeable exceptions are for females owning real estate, females in the bad occupations, owners of real estate not owning life insurance, and persons having both bank account and real estate. Interestingly enough, the first three of these four cases include only a small proportion of all borrowers.

Although the evidence indicates that complete independence does not exist in the good- and bad-loan samples, we feel that the use of the second short-cut method is amply warranted in the case of the commercial bank sample. The standard discriminant function approach, which accounts for correlations between variates, is based on assumptions of normality that are not supported by the available evidence. The second short-cut method, which assumes independence but makes no assumption of normality, may be quite as realistic as the standard approach.

When the second short-cut method was tried for the commercial bank sample, the two factors age and income were added to the seven used in the previous experiment. The formula resulting from the experiment appears on page 85; and the distribution of loans is shown in Table 18. To illustrate the computation procedure, we shall show how some of the terms of this formula were computed.

The bad-loan relative for persons having bank accounts is .5 (see summary of Table B-4); the reciprocal is 2.0; and the common logarithm of the reciprocal is .30. For persons not having bank accounts, the relative is 1.4; the reciprocal, .715; the logarithm, $\bar{1}.85$ or $-.15$. At this point two alternative proce-

TABLE B-4

PERCENTAGE DISTRIBUTIONS FOR THE COMMERCIAL BANK SAMPLE, SHOWING INTERDEPENDENCE AMONG THE FOLLOWING SEVEN CREDIT FACTORS: SEX, POSSESSION OF LIFE INSURANCE, OWNERSHIP OF REAL ESTATE, POSSESSION OF BANK ACCOUNT, STABILITY OF RESIDENCE, STABILITY OF OCCUPATION, AND NATURE OF OCCUPATION[a]

Classification[b]	SEX: FEMALE+, MALE−						LIFE INSURANCE: OWNED+, NOT OWNED−							
	Real Estate		*Occupation*[c]		*Bank Account*		*Real Estate*		*Occupation*[c]		*Sex*		*Stab. of Res.*	
	Owned X / Not Owned O		Good X / Bad O		Owned X / Not Owned O		Owned X / Not Owned O		Good X / Bad O		Female X / Male O		3 Yrs–Up X / 0–3 Yrs[d] O	
	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.
+X														
Good loans	4.14	5.20	16.92	12.28	8.29	8.51	23.61	22.40	53.34	52.92	14.07	15.54	49.63	47.48
Bad loans	1.52	1.21	6.08	3.82	2.43	1.91	12.47	10.64	32.42	33.42	4.46	6.34	33.84	33.20
Relative	.37	.23	.36	.31	.29	.22	.53	.48	.61	.63	.32	.41	.68	.70
+O														
Good loans	14.85	13.79	2.07	6.71	10.70	10.48	58.22	59.43	28.49	28.91	67.76	66.29	32.20	34.35
Bad loans	6.99	7.30	2.43	4.69	6.08	6.60	62.00	63.83	42.05	41.05	70.01	68.13	40.63	41.27
Relative	.47	.53	1.17	.70	.57	.63	1.06	1.07	1.48	1.42	1.03	1.03	1.26	1.20
−X														
Good loans	23.23	22.17	47.75	52.39	36.53	36.31	3.76	4.97	11.33	11.75	4.92	3.45	8.39	10.54
Bad loans	12.77	13.08	38.80	41.06	20.06	20.58	1.82	3.65	12.46	11.46	4.05	2.17	10.74	11.38
Relative	.55	.59	.81	.78	.55	.57	.48	.73	1.10	.98	.82	.63	1.28	1.08
−O														
Good loans	57.78	58.84	33.26	28.62	44.48	44.70	14.41	13.20	6.84	6.42	13.25	14.72	9.78	7.63
Bad loans	78.72	78.41	52.69	50.43	71.43	70.91	23.71	21.88	13.07	14.07	21.48	23.36	14.79	14.15
Relative	1.36	1.33	1.58	1.76	1.61	1.59	1.65	1.66	1.91	2.19	1.62	1.59	1.51	1.85

(continued on next page)

TABLE B-4

PERCENTAGE DISTRIBUTIONS FOR THE COMMERCIAL BANK SAMPLE[a] (*continued*)

Classification[b]	REAL ESTATE: OWNED+, NOT OWNED−						BANK ACCOUNT: OWNED+, NOT OWNED−							
	Stab. of Res.		*Occupation*[a]		*Stab. of Occup.*		*Real Estate*		*Stab. of Res.*		*Occupation*[a]		*Life Insurance*	
	3 Yrs–Up X 0–3 Yrs[d] O		Good X Bad O		6 Yrs–Up X 0–6 Yrs[d] O		Owned X Not Owned O		3 Yrs–Up X 0–3 Yrs[d] O		Good X Bad O		Owned X Not Owned O	
	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.
+X														
Good loans	19.81	15.88	17.16	17.70	19.90	16.30	15.66	12.27	28.14	26.00	30.07	28.99	39.04	36.68
Bad loans	9.73	6.37	7.80	6.41	8.21	5.76	5.47	3.21	10.74	10.03	12.77	10.09	18.34	16.75
Relative	.49	.40	.45	.36	.41	.35	.35	.26	.38	.39	.42	.35	.47	.46
+O														
Good loans	7.56	11.49	10.21	9.67	7.47	11.07	29.16	32.55	16.68	18.82	14.75	15.83	5.78	8.14
Bad loans	4.56	7.92	6.49	7.88	6.08	8.53	17.02	19.28	11.75	12.46	9.72	12.40	4.15	5.74
Relative	.60	.69	.64	.81	.81	.77	.58	.59	.70	.66	.66	.78	.72	.71
−X														
Good loans	38.21	42.14	47.51	46.97	39.67	43.27	11.71	15.10	29.88	32.02	34.60	35.68	42.79	45.15
Bad loans	34.85	38.21	37.08	38.47	32.11	34.56	8.82	11.08	33.84	34.55	32.11	34.79	56.13	57.72
Relative	.91	.91	.78	.82	.81	.80	.75	.73	1.13	1.08	.93	.98	1.31	1.28
−O														
Good loans	34.42	30.49	25.12	25.66	32.96	29.36	43.47	40.08	25.30	23.16	20.58	19.50	12.39	10.03
Bad loans	50.86	47.50	48.63	47.24	53.60	51.15	68.69	66.43	43.67	42.96	45.40	42.72	21.38	19.79
Relative	1.48	1.56	1.94	1.84	1.63	1.74	1.58	1.66	1.73	1.85	2.21	2.19	1.73	1.97

(*continued on next page*)

TABLE B-4

PERCENTAGE DISTRIBUTIONS FOR THE COMMERCIAL BANK SAMPLE[a] *(continued)*

Classification[b]	STAB. OF RES.: 3 YRS–UP+, 0–3 YRS–						STABILITY OF OCCUPATION: 6 YRS–UP+, 0–6 YRS–							
	Occupation[c]		*Stab. of Occup.*		*Sex*		*Occupation*[c]		*Sex*		*Life Insurance*		*Bank Account*	
	Good X Bad O		6 Yrs–Up X 0–6 Yrs[d] O		Female X Male O		Good X Bad O		Female X Male O		Owned X Not Owned O		Owned X Not Owned O	
	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.	Actual	Exptd.
+X														
Good loans	38.50	37.52	37.40	34.56	11.42	11.02	39.27	38.52	11.76	11.31	51.04	48.75	27.86	26.70
Bad loans	21.58	20.01	21.38	17.97	4.46	3.79	19.45	18.09	3.75	3.43	32.11	30.03	9.02	9.07
Relative	.56	.53	.57	.52	.39	.34	.50	.47	.32	.30	.63	.62	.32	.34
+O														
Good loans	19.52	20.50	20.62	23.46	46.60	47.00	20.30	21.05	47.81	48.26	8.53	10.82	31.71	32.87
Bad loans	23.00	24.57	23.20	26.61	40.12	40.79	20.87	22.23	36.57	36.89	8.21	10.29	31.30	31.25
Relative	1.18	1.20	1.13	1.13	.86	.87	1.03	1.06	.76	.76	.96	.95	.99	.95
–X														
Good loans	26.17	27.15	22.17	25.01	7.57	7.97	25.40	26.15	7.23	7.68	30.79	33.08	16.96	18.12
Bad loans	23.30	24.87	18.94	22.35	4.05	4.72	25.43	26.79	4.76	5.08	42.36	44.44	13.47	13.42
Relative	.89	.92	.85	.89	.54	.59	1.00	1.02	.66	.66	1.38	1.34	.79	.74
–O														
Good loans	15.81	14.83	19.81	16.97	34.41	34.01	15.03	14.28	33.20	32.75	9.64	7.35	23.47	22.31
Bad loans	32.12	30.55	36.48	33.07	51.37	50.70	34.25	32.89	54.92	54.60	17.32	15.24	46.21	46.26
Relative	2.03	2.06	1.84	1.95	1.49	1.49	2.28	2.30	1.65	1.67	1.80	2.07	1.97	2.07

(concluded on next page)

TABLE B-4

PERCENTAGE DISTRIBUTIONS FOR THE COMMERCIAL BANK SAMPLE[a] (*continued*)

Classification	SUMMARY													
	Sex		*Life Insurance*		*Real Estate*		*Bank Account*		*Stab. of Res.*		*Stab. of Occup.*		*Occupation*[c]	
	Male	Female	Owned	Not Owned	Owned	Not Owned	Owned	Not Owned	3 Yrs –Up	0–3 Years[d]	6 Yrs –Up	0–6 Years[d]	Good	Bad
Good loans	81.01	18.99	81.83	18.17	27.37	72.63	44.82	55.18	58.02	41.98	59.57	40.43	64.67	35.33
Bad loans	91.49	8.51	74.47	25.53	14.29	85.71	22.49	77.51	44.58	55.42	40.32	59.68	44.88	55.12
Relative	1.13	.45	.91	1.41	.52	1.18	.50	1.40	.77	1.32	.68	1.48	.69	1.56

[a] This table is based on 1,179 good loans and 987 bad loans, all of which reported complete information for all seven factors. For explanation of table, see text, p. 136.

[b] The meanings of the symbols in this column are indicated by corresponding symbols in the captions over the columns of percentages.

[c] The following occupational groups of Table 13, pp. 70–71, were considered good: professional (1a and 1b); clerical, except outside salesmen and commercial representatives (2a, 2b, 2d); policemen and firemen (3); and proprietors (4). All others were classed as bad.

[d] Upper limit excluded from this class interval.

dures are possible. One is to add .30 to the score of all cases with bank account and to subtract .15 from those without; the other is to add the difference, .45, to those having bank accounts and to subtract nothing from the others. With the first scheme, the point 0 is the dividing line between the better-than-average and worse-than-average cases; with the second, the point .15 is the dividing line. The second scheme, which may be a little easier for computing actual scores, was used here. The dividing line for the entire scoring system was 1.25, the sum of .15 for bank account plus eight similar quantities for the other factors.

A rough job of curve fitting was done in the case of stability of residence. The bad-loan relative is 1.6 for the class of less than one year; and it decreases more or less regularly to .6 for the class of 10 years or over (Table 12). The common logarithm or the reciprocals increase from $-.20$ to .22 so that the difference between the extremes is .42. For each year up to 10 at present address the loan was rated one-tenth of .42 or .042. Since the class of 10 years and over was not subdivided, we have no evidence to show whether the bad-loan relatives continue to fall as the length of residence increases above 10 years. For this reason the total score was limited to .42 no matter how long the tenure of residence. Some readers may take exception to this conservative policy; they may feel that an additional score of .042 should be added for each year over 10. While this point of view may be justified, we merely suggest that such a policy may give too high a rating to the young person of 25 who has never been away from home.

APPENDIX C

Tests of Significance and Sampling Errors

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)

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

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

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

where $l''_i = \Sigma_j a_j \frac{s^{ij}}{|s_{ij}|}$ 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 dis-

[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

tribution 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 T. If, for example, T 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
$$t = \frac{Xz - Y}{\sqrt{\sigma_x^2 z^2 + \sigma_y^2}}$$

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

$$\frac{n + n'}{nn'} + T^2\frac{1}{2(n + n')} = \frac{n + n'}{nn'}\left(1 + T^2\frac{nn'}{2(n + n')^2}\right)$$ [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 $U = T + \mathfrak{U}$, $s_{ij} = \sigma_{ij} + \mathfrak{s}_{ij}$, 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

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

Since U remains invariant for all non-singular linear transformations, we can

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 $.049\sqrt{1 + T^2/8}$. For a value of .5 for T, 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 T by the relation

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

---

assume without loss of generality that $\sigma_{ij} = 0$ whenever $i \neq j$. We wish to reduce this to a linear function in the a_{ij}'s and δ_{ij}'s, which is possible because second order terms in a_{ij} and δ_{ij} can be neglected as infinitesimals of higher order. We may therefore write:

$$\frac{(\alpha_i + a_i)(\alpha_j + a_j)\ \text{cofactor}\ (\sigma_{ij} + \delta_{ij})}{|(\sigma_{ij} + \delta_{ij})|} \qquad [i \neq j]$$

$$= \frac{(\alpha_i + a_i)(\alpha_j + a_j)\delta_{ij}(\sigma_{11} + \delta_{11})(\sigma_{22} + \delta_{22}) \ldots (\sigma_{pp} + \delta_{pp})}{(\sigma_{11} + \delta_{11})(\sigma_{jj} + \delta_{jj})}$$

$$= \frac{(\alpha_i + a_i)(\alpha_j + a_j)\delta_{ij}}{(\sigma_{ii} + \delta_{ii})(\sigma_{jj} + \delta_{jj})} = (v_i + u_i)(v_j + u_j)r_{ij},$$

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

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

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

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

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

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

$$\frac{1}{T^2}\left[\Sigma\left(\frac{n + n'}{nn'} v_i^2 + \frac{v_i^4}{2(n + n')}\right) + \Sigma\Sigma_{i \neq j} v_i^2 v_j^2 \frac{1}{n + n'}\right]$$

$$= \frac{1}{T^2}\left[\frac{n + n'}{nn'}\Sigma v_i^2 + \frac{1}{2(n + n')}\Sigma\Sigma v_i^2 v_j^2\right]$$

$$= \frac{n + n'}{nn'} + \frac{T^2}{2(n + n')} = \frac{n + n'}{nn'}\left(1 + \frac{T^2 nn'}{2(n + n')^2}\right)$$

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 r 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	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

* This reference applies to technical edition.

* This reference applies to technical edition.

PUBLICATIONS OF THE NATIONAL BUREAU OF ECONOMIC RESEARCH

INCOME IN THE UNITED STATES
Wesley C. Mitchell, W. I. King, F. R. Macaulay and *O. W. Knauth*

*1	Volume I, SUMMARY (1921)	152 pp.
2	Volume II, DETAILS (1922)	420 pp., $5.15
3	DISTRIBUTION OF INCOME BY STATES IN 1919 (1922) *O. W. Knauth*	30 pp., $1.30
*4	BUSINESS CYCLES AND UNEMPLOYMENT (1923) *National Bureau Staff* and *Sixteen Collaborators*	405 pp.
*5	EMPLOYMENT, HOURS AND EARNINGS, UNITED STATES, 1920-22 (1923) *W. I. King*	147 pp.
6	THE GROWTH OF AMERICAN TRADE UNIONS, 1880-1923 (1924) *Leo Wolman*	170 pp., $2.50
7	INCOME IN THE VARIOUS STATES: ITS SOURCES AND DISTRIBUTION, 1919, 1920 AND 1921 (1925) *Maurice Levin*	306 pp., $3.50
8	BUSINESS ANNALS (1926) *W. L. Thorp*, with an introductory chapter, "Business Cycles as Revealed by Business Annals," by *Wesley C. Mitchell*	380 pp., $2.50
9	MIGRATION AND BUSINESS CYCLES (1926) *Harry Jerome*	256 pp., $2.50
10	BUSINESS CYCLES: THE PROBLEM AND ITS SETTING (1927) *Wesley C. Mitchell*	489 pp., $5.00
*11	THE BEHAVIOR OF PRICES (1927) *Frederick C. Mills*	598 pp.
12	TRENDS IN PHILANTHROPY (1928) *W. I. King*	78 pp., $1.00
13	RECENT ECONOMIC CHANGES (1929) *National Bureau Staff* and *Fifteen Collaborators*	2 vols., 950 pp., $7.50
	INTERNATIONAL MIGRATIONS	
14	Volume I, STATISTICS (1929), compiled by *Imre Ferenczi* of the International Labour Office, and edited by *W. F. Willcox*	1112 pp., $7.00
18	Volume II, INTERPRETATIONS (1931), edited by *W. F. Willcox*	715 pp., $5.00
*15	THE NATIONAL INCOME AND ITS PURCHASING POWER (1930) *W. I. King*	394 pp.
16	CORPORATION CONTRIBUTIONS TO ORGANIZED COMMUNITY WELFARE SERVICES (1930) *Pierce Williams* and *F. E. Croxton*	347 pp., $2.00
17	PLANNING AND CONTROL OF PUBLIC WORKS (1930) *Leo Wolman*	260 pp., $2.50
*19	THE SMOOTHING OF TIME SERIES (1931) *Frederick R. Macaulay*	172 pp.
20	THE PURCHASE OF MEDICAL CARE THROUGH FIXED PERIODIC PAYMENT (1932) *Pierce Williams*	308 pp., $3.00
*21	ECONOMIC TENDENCIES IN THE UNITED STATES (1932) *Frederick C. Mills*	639 pp.

* *Out of print.*

22 SEASONAL VARIATIONS IN INDUSTRY AND TRADE (1933)
Simon Kuznets 455 pp., $4.00

23 PRODUCTION TRENDS IN THE UNITED STATES SINCE 1870 (1934)
A. F. Burns 363 pp., $3.50

24 STRATEGIC FACTORS IN BUSINESS CYCLES (1934)
J. Maurice Clark 238 pp., $1.50

25 GERMAN BUSINESS CYCLES, 1924-1933 (1934)
C. T. Schmidt 288 pp., $2.50

26 INDUSTRIAL PROFITS IN THE UNITED STATES (1934)
R. C. Epstein 678 pp., $5.00

27 MECHANIZATION IN INDUSTRY (1934)
Harry Jerome 484 pp., $3.50

28 CORPORATE PROFITS AS SHOWN BY AUDIT REPORTS (1935)
W. A. Paton 151 pp., $1.25

29 PUBLIC WORKS IN PROSPERITY AND DEPRESSION (1935)
A. D. Gayer 460 pp., $3.00

30 EBB AND FLOW IN TRADE UNIONISM (1936)
Leo Wolman 251 pp., $2.50

31 PRICES IN RECESSION AND RECOVERY (1936)
Frederick C. Mills 561 pp., $4.00

32 NATIONAL INCOME AND CAPITAL FORMATION, 1919-1935 (1937)
Simon Kuznets 100 pp., 8¼ x 11¾, $1.50

33 SOME THEORETICAL PROBLEMS SUGGESTED BY THE MOVEMENTS OF INTEREST RATES, BOND YIELDS AND STOCK PRICES IN THE UNITED STATES SINCE 1856 (1938)
F. R. Macaulay 586 pp., $5.00
"The Social Sciences and the Unknown Future," a reprint of the introductory chapter to Dr. Macaulay's volume: 35 cents; in orders of 10 or more, 25 cents.

34 COMMODITY FLOW AND CAPITAL FORMATION, Volume I (1938)
Simon Kuznets 500 pp., 8¼ x 11¾, $5.00

35 CAPITAL CONSUMPTION AND ADJUSTMENT (1938)
Solomon Fabricant 271 pp., $2.75

36 THE STRUCTURE OF MANUFACTURING PRODUCTION, A CROSS-SECTION VIEW (1939)
C. A. Bliss 234 pp., $2.50

37 THE INTERNATIONAL GOLD STANDARD REINTERPRETED, 1914-34 (1940)
William Adams Brown, Jr. 2 vols., 1420 pp., $12

38 RESIDENTIAL REAL ESTATE, ITS ECONOMIC POSITION AS SHOWN BY VALUES, RENTS, FAMILY INCOMES, FINANCING, AND CONSTRUCTION, TOGETHER WITH ESTIMATES FOR ALL REAL ESTATE (1940)
D. L. Wickens 320 pp., 8¼ x 11¾, $3.50

39 THE OUTPUT OF MANUFACTURING INDUSTRIES, 1899-1937 (1940)
Solomon Fabricant 685 pp., $4.50

FINANCIAL RESEARCH PROGRAM

I *A Program of Financial Research*

1 REPORT ON THE EXPLORATORY COMMITTEE ON FINANCIAL RESEARCH (1937)
91 pp., $1.00

2 INVENTORY OF CURRENT RESEARCH ON FINANCIAL PROBLEMS (1937)
253 pp., $1.50

II *Studies in Consumer Instalment Financing*

1 PERSONAL FINANCE COMPANIES AND THEIR CREDIT PRACTICES (1940)
Ralph A. Young and *Associates* 170 pp., $2.00

2 SALES FINANCE COMPANIES AND THEIR CREDIT PRACTICES (1940)
Wilbur C. Plummer and *Ralph A. Young* 298 pp., $3.00

3 COMMERCIAL BANKS AND CONSUMER INSTALMENT CREDIT (1940)
John M. Chapman and *Associates* 318 pp., $3.00

4 INDUSTRIAL BANKING COMPANIES AND THEIR CREDIT PRACTICES (1940)
Raymond J. Saulnier 192 pp., $2.00

5 GOVERNMENT AGENCIES OF CONSUMER INSTALMENT CREDIT (1940)
Joseph D. Coppock 216 pp., $2.50

6 THE PATTERN OF CONSUMER DEBT, 1935-36 (1940)
Blanche Bernstein 238 pp., $2.50

7 THE VOLUME OF CONSUMER INSTALMENT CREDIT, 1929-38 (1940)
Duncan McC. Holthausen in collaboration with *Malcolm L. Merriam* and *Rolf Nugent* 137 pp., $1.50

8 RISK ELEMENTS IN CONSUMER INSTALMENT FINANCING (1941)
David Durand 106 pp., $1.50
Technical edition, 163 pp., $2.00

CONFERENCE ON RESEARCH IN NATIONAL INCOME AND WEALTH

STUDIES IN INCOME AND WEALTH (Volumes I-III together, $7.50)
Volume I (1937) 368 pp., $2.50
Volume II (1938) 342 pp., $3.00
Volume III (1939) 500 pp., $3.50

CONFERENCE ON PRICE RESEARCH

1 REPORT OF THE COMMITTEE ON PRICES IN THE BITUMINOUS COAL INDUSTRY (1938) 144 pp., $1.25

2 TEXTILE MARKETS—THEIR STRUCTURE IN RELATION TO PRICE RESEARCH (1939) 304 pp., $3.00

3 PRICE RESEARCH IN THE STEEL AND PETROLEUM INDUSTRIES (1939)
224 pp., $2.00

NATIONAL BUREAU OF ECONOMIC RESEARCH
1819 Broadway, New York, N. Y.
European Agent: Macmillan & Co., Ltd.
St. Martin's Street, London, W. C. 2.

Studies in Consumer Instalment Financing

These studies are part of a broad program of research in finance inaugurated by the National Bureau of Economic Research under grants from the Association of Reserve City Bankers and the Rockefeller Foundation. The program has been undertaken in cooperation with public agencies, private enterprises and university specialists.

Risk Elements in Consumer Instalment Financing, the eighth volume in the series, presents an analysis of certain factors which are relevant to the selection of credit risks and the determination of credit standards in the field of consumer instalment financing. The study makes an integrated analysis of risk factors in the entire field of consumer financing, bringing together the findings of five institutional studies previously published. These five studies are:

Personal Finance Companies and Their Credit Practices (January 1940);
Sales Finance Companies and Their Credit Practices (July 1940);
Commercial Banks and Consumer Instalment Credit (June 1940);
Industrial Banking Companies and Their Credit Practices (October 1940);
Government Agencies of Consumer Instalment Credit (November 1940).

The sixth and seventh volumes in the series—*The Pattern of Consumer Debt, 1935-36* and *The Volume of Consumer Instalment Credit, 1929-38*—were undertaken as special statistical studies and were published in September 1940.

The following studies are in preparation: a comparative analysis of the operating experience of instalment financing agencies in 1929, 1933, and 1936; a study of the relation between consumer instalment financing and economic fluctuations; and a summary of the findings of the entire series on consumer instalment financing.