BOOK REVIEWS

Methods of correlation analysis. By Mordecai Ezekiel. 2d edition. New York, Wiley, 1941. 19+531 pp. $5.00.

The first edition of this well known work appeared in 1930 and had a marked influence upon the users of correlation theory in this country. Before its appearance attention had been centered upon the calculation of correlation coefficients to many decimal places with too much reliance upon probable errors and too much faith in Blakeman's tests. Ezekiel's book emphasized the regression side of correlation with particular stress on nonlinear regression functions. The use of free hand fitting of regression curves in simple and multiple correlation is clearly and quite completely discussed. As a result of the simplicity of explanations and the careful description of calculation procedures, it soon became one of the most important books in the field of correlation.

The difference between the two editions is a survey of the history of the advances in correlation theory during the intervening decade. As stated in the preface of the second edition these major changes have been "first, in the interpretation of the meaning of standard errors and, second in the application of logical limitations to the flexibility of graphic curves. Other significant developments have been in the perfection of new and speedier methods of estimating the reliability of an individual estimate or forecast." Remaining portions of the subject matter are left practically the same.

In general these changes are well made. Treatment of the reliability of an individual forecast is given in Chapter 19. Probability statements arising in the interpretations of standard errors have been correctly made, but the author did not introduce the terminology of "confidence interval" and "fiducial limits." This would be advisable. In discussing logical limitations of graphic curves it is carefully noted that extrapolation is based on these logical considerations rather than the statistical analysis. Samples give most reliable information for the ranges of the variables included in the sample.

Since the general features of the first edition are so well known, it seems most important to mention here some of the detail in the new edition where particular comment is pertinent.

The method of identifying classes by open class limits is superior to the one used. For example (p. 5) change the notation from 22.5–25.4 bushels to 23–25 bushels. The latter form gives the lowest and highest values included in the class and indicates the accuracy of the measurements.

On p. 10 the text states "Most of the reports come at about the middle values and then thin out to both ends (that is, the distribution approximates normality). In such cases the standard deviation gives a measure of the range within which a definite proportion of the cases will be included." This is followed by a statement of the common area values for the normal curve of error. The author should explain that these area values are only approximate for non-normal distributions which satisfy the given conditions, and that normality requires a particular functional relationship, not merely a graphic similarity.

On p. 156 the following statement appears. "Since the index of determination is simply $\bar{\rho}^2_{yx}$, it is 71.0 per cent." According to the definition given it should be .710.

In another discussion involving the coefficient of determination (p. 159) the following statement is made. "Since the coefficient is a ratio, it is a "pure number," that is, it is an arbitrary mathematical measure, whose values fall within a certain limited range, and it can be compared only with other constants like itself, derived from similar problems." The arithmetic mean is a ratio, and it depends on unit! Furthermore not all ratios are arbitrary! The meaning of the quotation is vague. The point in question is an important one, and a clear explanation should be given.

On pp. 160–162 the author discusses the interpretation of the three types of measure of correlation; regression coefficient, correlation coefficient, and standard error of estimate. He states that the most accurate estimate of values of the dependent variable made from the regression equation calculated from a sample will be made from that sample for which S_y, the standard error of estimate, is the smallest. The accuracy of an estimate of y should be interpreted with reference to the importance of a unit change in y. In a sample with small variation in the dependent variable two situations may be considered: (a) a small variation may be important in case the population itself has small variation, or (b) the dependent variable has been so controlled in sampling that the variation is materially·biased more than the usual sampling bias. It is well known that sampling under case (b) tends to lower the correlation coefficient and thus to make the whole correlation analysis unreliable. In case (a) it is clear that S_y must be interpreted in relation to σ_y. The factor $(1-r^2)^{1/2}$ measures the improvement in estimation of y due to the use of regression, and the smaller the value of r, disregarding sign, the smaller is the amount of information supplied by the regression.

In introducing linear multiple regression (p. 164) it should be pointed out that the form of the equation assumes the relationship between the variables to be additive. The discussion given might lead the reader to believe that the linear equation includes all cases.

Statisticians interested in factor analysis will question the recommendation that not more than ten, usually five, variables be used in a study.

To obtain such quantities as $\sum XY$ and $\sum X^2$ it should be mentioned that the extensions as made in the text are not needed. An explanation of the use of a calculating machine to shorten the work might well be given in the text at an early point.

One rather general criticism of the book should be made. There is throughout the tendency to over-correct calculated constants and to over-refine tests of significance. More care should be taken to explain the fact that experimental data frequently do not justify the use of many of these refinements. To an untrained reader they may imply an accuracy of analysis not actually present. In some cases the significance tests suggested are actually incorrect, as for example the use of the standard error of the coefficient of multiple correlation.

In spite of the above criticisms the reviewer considers this book still to be the best in its field.

E. L. Welker

Finite dimensional vector spaces. By Paul R. Halmos. (Annals of Mathematics Studies, no. 7.) Princeton University Press, 1942. 5+196 pp.

In this book the author presents the topics covered usually in an introductory course in algebra (matrices, linear equations, linear transformations, and so on) from the point of view of a modern analyst interested in general vector spaces.

The ever-growing interest in Hilbert and more general linear spaces makes the appearance of the book very timely, especially since it furnishes an excellent introduction to the subject certainly within the grasp of a first-year graduate student or even a good senior or junior.

The topics are treated in such a manner as to make future generalizations look both natural and suggestive. This sometimes is done at the expense of the shortness of exposition. Some theorems, as the author himself confesses, could be proved in fewer lines. He prefers, however, longer proofs that admit a generalization to shorter ones that do not.

The reviewer finds himself in complete agreement with this method