ABSTRACTS OF PAPERS

I. Presented on December 27, 1941, at the New York Meeting of the Institute

A Generalized Analysis of Variance. FRANKLIN E. SATTERTHWAITE, University of Iowa and Aetna Life Insurance Company.

This paper examines the fundamental principals underlying designs for the analysis of variance. Given several statistics of the type, $\chi_i^2 = \Sigma_i \theta_{ij}^2$ where the θ's are arbitrary orthogonalized linear functions of certain underlying normal data, x_k; a rule is set up for determining a set of m_k as linear functions of the x_k such that $\chi_0^2 = \Sigma(x_k - m_k)^2$ will be independent of the remaining χ_i^2's. Further it is shown that simultaneously with the above, the x's and the θ's may be subjected to certain types of linear restrictions (for the purpose of estimating parameters or otherwise) without disturbing the distributions or the independence relations except for the appropriate reduction in degrees of freedom. The rule used to determine the m's gives results consistent with the standard designs for the analysis of variance. However, it goes further in that one may use weighted rather than simple averages in setting up his design. A practical application of this is the two way analysis of data which are averages and lack homogeniety of variance through constants of proportionality between the variances are known. The two way analysis of incomplete data is another practical problem which is solved by the simple expedient of a zero weight. The use of weighted averages frequently introduces difficulties in estimating parameters, particularly the mean. The combination of the linear restriction concept with standard analysis of variance methods solves this difficulty.

On the Power Function of the Analysis of Variance Test. ABRAHAM WALD, Columbia University.

It is known that the power function of the analysis of variance test depends only on a single parameter, say λ, where λ is a certain function of the parameters involved in the distribution of the sample observations. Let Z be any critical region (subset of the sample space) whose size does not depend on unknown parameters, i.e., it has the same size for all values of the parameters which are compatible with the hypothesis to be tested. It is shown that for any positive c the average power (a certain weighted integral of the power function) of the region Z over the surface $\lambda = c$ cannot exceed the power of the analysis of variance test on the surface $\lambda = c$ (the power of the latter test is constant on the surface $\lambda = c$). P. S. Hsu's result, *Biometrika*, January, 1941, pp. 62–68, follows from this as a corollary.

Definition of the Probable Error. E. J. GUMBEL, The New School for Social Research.

The probable error is usually defined either as the semi-interquartile range or as $\frac{2}{3}$ of the standard error. We define it as half of the smallest interval that has the probability $\frac{1}{2}$. For distributions which never increase (decrease), the beginning (end) of this interval is the origin (the median), and the end is the median (the end of the distribution). In general the probable error ρ is the solution of the equations $W(\xi + \rho) - W(\xi - \rho) = \frac{1}{2}$ and $w(\xi + \rho) = w(\xi - \rho)$ where ξ denotes the midpoint of the interval. For symmetrical distributions the first definition remains valid. For the Gaussian distribution the second definition holds besides. The numerical values for the midpoint ξ and the probable error ρ are given for some distributions usual in statistics. The calculation of the standard error of the probable error, which depends upon the distribution $w(x)$, determines whether the probable error is more or less precise than the standard error. For the asymmetrical exponential

distribution the mean and the median have the same precision, and the probable error is more precise than the standard error. For the first law of Laplace, and for Galton's reduced distribution the median and the probable error are more precise than the mean and the standard error. For Maxwell's distribution the mean and the probable error are more precise than the median and the standard error.

A Class of Multivariate Distributions. WALTER JACOBS, Security and Exchange Commission, Washington.

The multivariate normal distribution has the property that its probability density is constant along the surface of a hyper-ellipsoid. The class of distributions characterized by this property is considered. The form of the characteristic function of any distribution of the class is determined; in this way the parameters of the distribution are shown to be simply related to the first and second moments, when these exist.

Every distribution of the class is the n-variate extension of a univariate symmetrical distribution. The method of determining the form of the extension of such a univariate distribution is given. A number of properties of regression for the multivariate normal distribution are shown to hold for any distribution of the class. Among other properties considered is the form of some sampling distributions. Some special cases of interest, including the extensions of the Cauchy distribution and the median law, are discussed briefly.

Methods for Scanning Data to Determine the Significance of the Difference Between the Frequency of an Event in Contrasted Groups. JOSEPH ZUBIN, N. Y. S. Psychiatric Institute, New York.

In many investigations in Psychology, Sociology, Economics and Public Health, there is a need for a quick and ready method for scanning a mass of data in order to select the items that have a significant bearing on the problem under investigation. The statistical procedure for this item analysis consists essentially of evaluating the 2×2 tables which arise when two groups are contrasted for the presence and absence of a given character or event. The chi square method or its equivalent, the ratio of the difference between per cents to its standard error, require considerable labor and time and several methods have been proposed for shortening the work. Recently a method was developed which eliminates the need for computing percentages or expected values, the analysis being made with the absolute frequencies. This method depends upon transforming p, the per cent, to the inverse sine function of $\sqrt{p}$. The method is applicable not only to 2×2 tables but can also be made applicable to $2 \times n$ tables and $r \times n$ tables with the aid of simple formulae.

Compounding Probabilities from Independent Significance Tests. W. ALLEN WALLIS, Stanford University.

For combining the probabilities obtained from N independent tests of significance into a single measure, the product of the N independent probabilities provides a criterion which, though rarely ideal, is usually satisfactory. The probability that such a product will be less than Q always exceeds Q, and is the sum of the first N terms in a Poisson series whose parameter is $-\log_e Q$; since this sum is also the probability that a value of χ^2 based on $2N$ degrees of freedom will exceed $-2 \log_e Q$, existing tables of χ^2 may (as R. A. Fisher has pointed out in *Statistical Methods for Research Workers*, section 21.1) be used to test the significance of a product of probabilities. If any of the probabilities have been derived from discontinuous distributions, as is likely with small samples of non-metric data, this method of calculating the probablity of the product fails; in such instances it invariably overstates the probability of the product. Formulas are given for various special cases arising frequently in practice and also for the general case of $D + C$ tests of which D are

based on discontinuous distributions and C on continuous distributions. In several illustrative examples, the overstatement of the joint probability consequent upon neglect of discontinuities is of the order of 100 to 200 per cent.

A Method of Computing the Roots of the General Cubic Equation with Real or Complex Coefficients. ERNEST E. BLANCHE, Michigan State College.

The general cubic equation with real or complex coefficients may readily be reduced to the form $y^3 + 3Hy + G = 0$. Suitable substitutions for y in the reduced equation permit the use of the identities for hyperbolic functions and circular functions: $\sin 3x$, $\cos 3x$, $\sinh 3x$, $\cosh 3x$ and $\sin(u + iv)$. The following classifications may be set up: (A) If $G < 0$ and $H > 0$, only real root is $y = 2\sqrt{H}\sinh z$ where $\sinh 3z = G/2H\sqrt{H} = M$; (B-1) If $G < 0$, $H < 0$, $G/2H\sqrt{-H} \leqq 1$, three real roots, obtained by use of circular identity, $\cos 3x$; (B-2) If $G < 0$, $H < 0$, $G/2H\sqrt{-H} > 1$, only real root is $y = 2\sqrt{-H}\cosh z$ where $\cosh 3z = G/2H\sqrt{-H}$. Complex roots are $-\frac{1}{2}\, y_1 \pm bi$. The general cubic with complex coefficients has solutions $y_{n+1} = -2\sqrt{H}\sin(u + 2n\pi/3 + iv)$ for $n = 0, 1, 2$, where $\sin(3u + 3iv) = a + bi = M$. For M real, special cases are similar to (A), (B-1) and (B-2).

Limited Type of Probability Distribution Applied to Flood Flows (Preliminary Report). BRADFORD F. KIMBALL, Port Washington, N. Y.

Relative to Gumbel's recent paper on Flood Flows (E. J. Gumbel, "The return period of flood flows," *Annals of Mathematical Statistics*, Vol. 12 (1941)) the author points out that Gumbel's argument that the probability distribution of maximum values does not stem from a limited form of primary probability distribution of the stream flow, is misleading (see page 177, loc. cit.). One might argue for a primary probability distribution of stream flows of the type: $dV = \exp(-\frac{1}{2}u^2)du$ where $u = k[b - \log(a - x)]$, $0 \leqq x \leqq a$, where x is the measure of flow. This increment of x is related to normal probability increment by the linear equation $k\,dx = (a - x)du$. This distribution will not satisfy the condition that von Mises uses in his argument concerning a finite distribution, since the cumulative distribution V does not possess a positive derivative of finite order at $x = a$. Also, although x does not have infinite range, the transformed variate u has an infinite range to the right, and will satisfy von Mises' argument for the derivation of the cumulative distribution of the maxima, of the form $\exp[-\exp\{-\alpha(u - u_0)\}]$ in terms of u. The author finds that such a distribution more accurately describes the behavior of maximum annual flood flows than one which ignores the existence of an upper limit a.

Additive Partition Functions. J. WOLFOWITZ, New York City.

Let n_1 and n_2 be positive integers and let

$$m = \max\left(\frac{n_1}{n_1 + n_2}, \frac{n_2}{n_1 + n_2}\right).$$

Let the stochastic variable $V = (v_1, v_2, \cdots v_5)$ be any sequence of positive integers such that $v_1 + v_3 + v_5 + \cdots$ is equal to either one of n_1 and n_2, while $v_2 + v_4 + v_6 + \cdots$ is equal to the other. Two sequences V with the same elements arranged in different order are to be considered distinct and all sequences V are to be assigned the same probability. Such sequences are of statistical importance (Wald and Wolfowitz, *Annals of Math. Stat.*, Vol. 11 (1940). Let $f(x)$ be a function defined for all positive integral values of x which fulfills the following conditions:

1. There exists a pair of positive integers, a and b, such that that

$$\frac{f(a)}{f(b)} \neq \frac{a}{b}$$

2. The series

$$\sum_{i=1}^{\infty} |f(i)| m^{\frac{1}{2}i}$$

is convergent. Then, as n_1 and $n_2 \to \infty$, while n_1/n_2 remains constant, the distribution of the stochastic variable

$$F(V) = \sum_{i=1} f(v_i)$$

approaches the normal distribution. When $f(x) \equiv 1$, $F(V) \equiv U(V)$ (loc. cit., Theorem I). When $f(x) = \log\left(\frac{x^x}{x!}\right)$, $F(V)$ is a statistic introduced by the author (*Amer. Math. Soc. Bull.* (1941), p. 216).

A similar result holds for partitions of a single integer.

II. Presented on December 29, 1941, at the joint session of the Institute, The Econometric Society, and Section A of the A. A. A. S.

Certain Tests for Randomness Applied to Data Grouped into Small Sets. EDWARD L. DODD, University of Texas.

G. Udny Yule, in his paper *A Test of Tippett's Random Sampling Numbers* (*Roy. Stat. Soc. Jour.*, Vol. 101(1938), pp. 167–172), described tests applied to certain sums of the Tippett numbers. Yule regarded the Tippett numbers as not altogether satisfactory.

The tests now to be described, however, involve no summation. For sets of three digits, four classes may be distinguished: The middle number may be the largest, or it may be the least; or the sequence may be monotone increasing or monotone decreasing —here the sequence a, a, a, may be classified with the monotone increasing sequences when $a > 4$; otherwise, with the monotone decreasing sequences. Similarly, six consecutive digits in two sets of three digits each give rise to sixteen classes. On the basis of range, sets of two or more of the digits $0, 1, 2 \cdots, 9$ may be separated into ten classes.

Chi-square tests applied by the present author on the basis of the foregoing and similar classifications have not thus far indicated that the Tippett numbers are not satisfactorily random.

Stratified Sampling. A. M. MOOD, University of Texas.

When certain relations between the probabilities $p_1, p_2, \cdots, p_k$ of a multinomial population are known in advance, the technique of stratified sampling provides more efficient estimates of the probabilities than does random sampling. Under certain conditions of stratified sampling, however, the maximum likelihood estimates, n_i/n, of p_i are biased but are unbiased in the limit as the sample size increases. The methods and results of the theory of maximum likelihood require no modification to be made applicable to the problem of estimation in stratified sampling; in fact the results of this theory imply the use of stratified sampling when the conditions for its use obtain.

Advantages of Singling Out Degrees of Freedom in Analyses of Variance. WILLIAM DOWELL BATEN, Michigan Agriculture Experiment Station.

This paper pertains to an experiment involving dummy plots for analyzing effects of placements and fertilizers for cannery peas. Three fertilizers were used at different distances from the pea seeds at planting, the design being a randomized block layout. Advantages are given for breaking up the sum of squares, due to differences between "treatment" means, into sums of squares, each with one degree of freedom. Methods are given for securing the sum of squares involving dummy plots, and obtaining the variances due to main effects and interaction. Interpretations are given for each phase of the analysis.