ABSTRACTS OF PAPERS

Presented Sept. 2–4, 1947, at the Yale meeting of the Institute

1. Estimation of Parameters in Truncated Pearson Frequency Distributions. A. C. Cohen, University of Georgia.

Given a truncated univariate Pearson frequency distribution, parameters of the complete distribution are required. Karl Pearson and Alice Lee, (*Biometrika*, Vol. 6 (1915), pp. 59–69) and R. A. Fisher, (*Introduction to Mathematical Tables*, Vol. 1, British Assn. Adv. Sci., 1931, pp. xxvi–xxxv), obtained solutions of the truncated normal distribution with a single *tail* missing. The present paper presents three general methods of solution applicable to any of the Pearson distributions. The first utilizes moments of a higher order than are required to characterize corresponding complete distributions. The order of the highest moment required is increased by one for each missing *tail*. The second method, applicable when only a single *tail* is missing, utilizes the terminal ordinate at the point of truncation and moments of the same order as required to characterize the complete distribution. The terminal ordinate is evaluated by successive approximations. The third method utilizes only the first two moments, but requires that the given distribution be further truncated and that moments be computed both before and after the additional truncations. This latter method can also be applied to complete distributions to avoid direct computation of third and fourth order moments.

2. Distribution of a Root of Determinantal Equation. D. N. Nanda, University of North Carolina.

The joint distribution of the roots of a determinantal equation was given by P. L. Hsu in 1939 and the distribution of any one of the roots was studied by S. N. Roy. The present paper, however, gives a different method of working out the distribution of any root, specified by its place in a monotonic arrangement. This method enables us to express the distribution of a root of a certain determinantal equation in terms of a linear combination of products of incomplete beta integrals and in terms of the distribution of a root of lower-order determinantal equations.

3. The Power of Certain Non-Parametric Tests of Independence. Wassily Hoeffding, University of North Carolina.

Several tests of independence have been proposed which are based on statistics depending only on the ranks of the sample values. Under the hypothesis H_0 of independence the distribution of such statistics does not depend on the form of the parent distribution. Two of these statistics, Spearman's rank correlation coefficient and Lindeberg-Kendall's statistic based on the number of inversions in the permutation of the ranks, are shown to be asymptotically normally distributed in samples from any population (the limiting normal distribution being singular in certain degenerate cases). The asymptotic distribution of these coefficients reveals that the corresponding tests of independence are inconsistent (in the sense that the probability of rejecting H_0 does not necessarily tend to 1 if H_0 is not true), and at least one of them is biased in the limit. It can be shown that at least for some sample sizes and some sizes of the critical region there do not exist unbiased tests of independence based on ranks. But there do exist rank tests of independence which are consistent, and hence unbiased in the limit. Examples of such tests are given.

4. Some Significance Tests for the Mean Using the Sample Range and Midrange. John E. Walsh, Princeton University.

Consider a sample of size n, $(2 \leq n \leq 10)$, drawn from a normal population with mean μ. Let x_n be the largest value and x_1 the smallest value of the sample. Significance tests are developed to compare μ with a given hypothetical value μ_0 by use of the sample. These significance tests are based on the quantity $D = [\frac{1}{2}(x_1 + x_n) - \mu_0]/(x_n - x_1) =$ [(sample midrange) − (hypothetical mean)]/(sample range). One-sided and symmetrical tests are considered. Values of D_α such that $Pr(D > D_\alpha \mid \mu = \mu_0) = \alpha$ are computed for $\alpha = .05, .025, .01, .005$. These values of D_α can be used to obtain one-sided tests at the .05, .025, .01, .005 significance levels and symmetrical tests at the .10, .05, .02, .01 significance levels. Efficiencies are computed for one-sided tests at the .05 and .01 significance levels. The efficiency is at least 90% for $n \leq 6$ at the .05 significance level and for $n \leq 8$ at the .01 level. The range-midrange test can be applied without computation through the use of an easily constructed graph. The application of a test requires only the plotting of the sample point (x_1, x_n) on this graph.

5. Testing Compound Symmetry in a Normal Multivariate Distribution. David F. Votaw, Jr., Princeton University.

Let $F(X)$ be the d.f. of a t-order vector variate $X(t \geq 3)$. Suppose the components of X are divided into mutually exclusive and exhaustive subsets. $F(X)$ is said to be *compound symmetric*, for the given division of its variates into subsets, if it is invariant over all permutations of its variates within these subsets. $F(X)$ is *completely symmetric* if the invariance holds over all permutations of its variates. If $F(X)$ is normal and compound symmetric, then within each subset of variates the means are equal, the variances are equal and the covariances are equal, and between any two subsets of variates the covariances are equal. Testing hypotheses of compound or complete symmetry in a normal $F(X)$ is of interest, for example, in studying psychological examinations and in medical research.

In this paper likelihood ratio criteria are developed for testing various hypotheses involving compound symmetry in regard to a normal distribution and to k normal distributions $(k \geq 2)$. Given that the corresponding null hypothesis is true, the moments of each criterion are obtained explicitly and the distribution of each criterion is identified as the product of independent beta variates (in the case of a single normal distribution, the distributions are given explicitly for $t = 3, 4$, and 5 for certain divisions of the variates into subsets). In a previous paper Wilks has given results on a very thorough study of the sampling theory of likelihood ratio criteria for various hypotheses involving complete symmetry in regard to a normal distribution.

6. Effects of Non-Normality at High Significance Levels. Harold Hotelling, University of North Carolina.

The effects of non-normality in the underlying population on the probabilities of *significance* by customary statistical tests are not well understood, in spite of numerous attacks, both mathematical and experimental, on the problem. Chung's recent proof that the distribution of the Student ratio t has in samples from an arbitrary population a distribution approaching normality for large samples tends to confirm the common idea that non-normality makes little difference if only the sample is fairly large, but this holds only for a fixed range of values of t while the sample number N increases. The tail areas beyond a deviation which increases with N in certain ways often behave quite differently than in sampling from a normal population. If p is the probability that $\mid t \mid > t_0$ in samples of N from a normal population and p' is the corresponding probability for another population, it is shown that $\lim_{N\to\infty} \left\{\lim_{t_0\to\infty} (p'/p)\right\}$ may be zero or infinite or may take any finite value, even when the non-normal distribution involved is of simple and realistic continuous forms. The conditions that this limit be unity are concerned only with the *shoulders* of the population histogram, and have nothing to do with its moments or its behavior at infinity or at its mean. This suggests that caution should be used in applying familiar tests with high significance levels; that further calculations should be directed toward making this caution quantitatively definite; and that the use of sample moments or cumulants cannot lead to the most appropriate criterion of non-normality for this purpose.

7. On the Problem of Similar Regions. E. L. Lehmann, University of California, Berkeley, and Henry Scheffé, University of California, Los Angeles.

If $X = (X_1, \cdots, X_n)$ is a set of random variables with a joint probability density depending on a set of parameters $\theta = (\theta_1, \cdots, \theta_m)$, and if $T = (T_1, \cdots, T_m)$ is a set of sufficient statistics for θ, then Neyman (*Phil. Trans. Roy. Soc. London*, Vol. 236 (1937), pp. 333–380) has proved that a region w in the space of X is similar with respect to θ if it has the following structure: The intersections $w(t)$ of w with the surfaces $T = t$ have the property that the conditional probability of the sample point X falling into w given that $T = t$ does not depend on t. In the present paper a necessary and sufficient condition is found for the regions with the above structure to be the only similar regions. This condition is shown to be satisfied for a certain class K of probability densities which contains as special cases all densities for which the totality of similar regions has been previously determined. In particular the partial differential equations which Neyman (*Annals of Math. Stat.*, Vol. 12 (1941), pp. 46–76) assumed were satisfied in his solution of the problem of similar regions are solved and it is shown that any density satisfying these equations belongs to the above class K.

8. Fourth Degree Exponential Function. L. A. Aroian and Marguerite Darkow, Hunter College.

It is shown that the fourth degree exponential function is supported by the Bernoulli probability function and the hypergeometric probability function as well as being the function for which the method of moments is the *best* method according to the criterion of maximum likelihood. In the general situation six moments, at most, are needed. The function is classified into two general groups depending on symmetry or asymmetry and each case is divided again into unimodal and bimodal distributions. Examples show that the function is very successful in graduating the main Pearson types and the Gram-Charlier Type A frequency function. Various generalizations of the exponential function are indicated. In addition to its wide generality, the greatest practical advantage of the new system is the simplicity of the numerical calculations.

9. A General Weak Limit Theorem for Independent Distributions. P. L. Hsu, University of North Carolina. (Read by title.)

For every positive integer n let there be n distribution functions $F_{n1}(x)$, $F_{n2}(x)$, $\cdots$, $F_{nn}(x)$. Assume that $\lim_{n \to \infty} \text{Max}_{1 \leq j \leq n}\{1 - F_{nj}(x) + F_{nj}(-x)\} = 0$. Let $F(x)$ be the convolution $F_{n1}(x) * F_{n2}(x) * \cdots * F_{nn}(x)$. Let $\psi(t) = mit + \int_{-\infty}^{+\infty} [e^{itx} - 1 - itx/(1 + x^2)](1 + x^2)/x^2 \, dG(x)$, with $G(x) \uparrow$ and $G(\infty) - G(-\infty) < \infty$. Let $F(x)$ be the (infinitely divisible) distribution law having $\exp \psi(t)$ as its characteristic function. In order to have $\lim_{n \to \infty} F_n(x) = F(x)$ at every continuity point of $F(x)$, it is necessary and sufficient that the following relations hold at every $x > 0$ such that $\pm x$ are continuity points of $G(y)$:

$$\text{(I)} \quad \lim_{n \to \infty} \sum_{j=1}^{n} \int_{|y| > x} dF_{nj}(y) = \int_{|y| > x} ((1 + y^2)/y^2) \, dG(y),$$

$$\text{(II)}\ \lim_{n\to\infty}\sum_{j=1}^{n}\left\{\int_{|y|>x} y^2\, dF_{nj}^{(y)} - \left(\int_{|y|<x} y\, dF_{nj}(y)\right)^2\right\} = \int_{|y|<x} (1+y^2)\, dG(y),$$

$$\text{(III)}\ \lim_{n\to\infty}\sum_{j=1}^{n}\int_{|y|<x} y\, dF_{nj}(y) = m + \int_{|y|<x} y\, dG(y) - \int_{|y|<x} (1/y)\, dG(y).$$

10. On the Maximum Partial Sums of Sequences of Independent Random Variables. K. L. Chung, Princeton University.

The asymptotic behavior of the maximum partial sums of a sequence of independent random variables is studied in this paper. Two groups of new limit theorems are established under general conditions. The first group deals with theorems of the *weak* type. The limiting distribution of the maximum partial sums is obtained with an estimate of the remainder, thus improving a recent result of Erdös and Kac. Another estimate is obtained for a different domain of variation, which plays an essential role in the sequel. These results correspond to the sharper forms of the central limit theorem. In the second group, theorems of the *strong* type are obtained, giving precise lower bounds (in the sense of probability) for the maximum partial sums. These results form the exact counterpart to the general form of the law of the iterated logarithm, due to Feller, which give the precise upper bounds. A summary of the main results and methods has appeared in *Proc. Nat. Acad. of Sci.*, Vol. 33 (1947), pp. 132–136.

11. Some Results on the Distribution of Quadratic Forms From Gaussian Stochastic Processes. (Preliminary report). Herman Rubin, Cowles Commission.

If one considers the estimation of the parameters of a Gaussian stochastic process whose elements are continuous functions from the functional values over a finite interval, one often finds that certain parameters can be estimated exactly, and certain parameters can not. This result often depends on the distribution of quadratic functionals whose arguments are elements of the stochastic process under consideration. In this paper, it is shown that the elements of a certain class of quadratic functionals have distributions concentrated at a point, and that the elements of a different class do not; in this latter case, the characteristic function is computed.

12. Some Significance Tests for the Median which are Valid under Very General Conditions. (Preliminary Report) John E. Walsh, Princeton University. (Read by title.)

Consider n independent values drawn from populations necessarily satisfying only: 1) Each population has a unique median. 2) The median has the same value φ for each population. 3) Each population is symmetrical. 4) Each population is continuous. (It is to be emphasized that no two of the values are necessarily drawn from the same population.) Significance tests are derived for φ on the basis of 1)–4). These significance tests are based on order statistics of certain combinations of order statistics, each combination being either a single order statistic of the n values or one-half the sum of two order statistics. The tests are invariant under permutation of the n values and reasonably efficient if the values represent a sample from a normal population. The significance levels are of the form $r/2^n$, $(r = 1, \cdots, 2^n - 1)$. Each value of r can be obtained for some one-sided significance test. Thus any significance level can be closely approximated if n is large. The major disadvantage of these tests is the limited number of suitable significance levels available for small values of n. This disadvantage is partially eliminated by the development of tests which have a specified significance level if the values are a sample from a normal population and a significance level bounded near this specified value if only 1)–4) necessarily hold. Results based on 1)–4) are applied to several well known statistical problems: Tests are obtained for the mean on the basis of a large number of independent values from populations having the mean but little else in common. Also generalized results are obtained for the Behrens-Fisher problem, quality control, slippage tests, the sign test and cases where some of the n values are dependent.

13. Loss of Information in t-tests with Unbalanced Samples. (Preliminary Report) JOHN E. WALSH, Princeton University. (Read by title.)

Consider two normal populations, the first with mean a_1 and variance σ_1^2, the second with mean a_2 and variance σ_2^2, while σ_1/σ_2 has a known value C. If the hypothesis $a_1 = a_2$ is to be tested by a t-test (one-sided or symmetrical) using n_1 sample values from the first population and n_2 values from the second population ($n_1 + n_2 = n$, fixed), it is shown that this experiment is most powerful when $n_1/n_2 = \sigma_1/\sigma_2$ (integer considerations neglected). The t-tests satisfying this condition will be referred to as balanced t-tests. Thus information will be lost by not using a balanced experiment. A quantitative measure of the information lost by using given values of n_1 and n_2 is determined by the total sample size m, ($m_1 + m_2 = m$), of the balanced t-test (same significance level) which has approximately the same power. Then $n - m$ sample values are wasted by using (n_1 , n_2) rather than (m_1 , m_2), i.e. only $100m/n\%$ of the information obtainable per observation is used by (n_1 , n_2). A symmetrical t-test with significance level 2α has the same value of m as a one-sided t-test with significance level α. For one-sided t-tests with significance level α: $m \doteq \frac{1}{2}(B + \sqrt{B^2 - 8A})$, where $B = 2 + A + K_\alpha^2/2$, $A = (C + 1)^2[1 - K_\alpha^2/2(n - 2)][C^2/n_1 + 1/n_2]^{-1}$, and K_α is the standardized normal deviate exceeded with probability α. This approximation to m is valid for $m \geq 5$ if $\alpha = .05$, $m \geq 6$ if $\alpha = .025$, $m \geq 7$ if $\alpha = .01$, $m \geq 8$ if $\alpha = .005$. (A fractional value of m represents an interpolated measure of the sample size of the equivalent balanced experiment.)

14. Some Theorems on the Bernoullian Multiplicative Process. T. E. HARRIS Princeton University. (Read by title.)

A single entity may have j descendents with probability P_j , ($j = 0, 1, 2, \cdots$). Each *first generation* entity has then the same procreative probabilities, etc. Let

$$f(s) = p_0 + p_1 s + \cdots$$

If z_n is the number of entities in the nth generation, it is known that $P(z_n = j)$ is given by the coefficient of s^j in the nth iterate $f[f \cdots (f)] = f_n(s)$. Let $Ez_1 = x, 1 < x < \infty$. Conditions are given insuring that as $n \to \infty$ the cumulative distribution of the variate z_n/x^n approaehes a limit-function which is absolutely continuous except for a possible single jump. Let $g(u)$ be the corresponding frequency function. If $f(s)$ is a polynomial of degree k, let $q = \log_x k/(\log_x k - 1)$. Otherwise, $q = 1$. Then $g(u) \cdot \exp\{u^{q+\epsilon}\}$ [is, is not] summable $(0, \infty)$ according as ϵ is [negative, positive]. Behavior of $g(u)$ near $u = 0$ is also considered. Special cases are considered where $g(u) = \text{constant} \cdot u^{1/m-1} \cdot e^{-u/m}$, m a positive integer. Maximum likelihood estimates for the parameters p_0 , p_1 , $\cdots$, and x are obtained as functions of n successive values z_1 , z_2 , $\cdots$, z_n . Consistency, in a certain sense, is proved. A specialized method is given for finding the moment-generating function of the variate N, the smallest value of n such that $z_n = 0$.