ON THE CHOICE OF THE NUMBER OF CLASS INTERVALS IN THE APPLICATION OF THE CHI SQUARE TEST

By H. B. Mann and A. Wald[1]

Columbia University

Introduction. To test whether a sample has been drawn from a population with a specified probability distribution, the range of the variable is divided into a number of class intervals and the statistic,

$$\sum_{i=1}^{i=k} \frac{(\alpha_i - Np_i)^2}{Np_i} = \chi^2, \tag{1}$$

computed. In (1) k is the number of class intervals, α_i the number of observations in the ith class, p_i the probability that an observation falls into the ith class (calculated under the hypothesis to be tested). It is known that under the null hypothesis (hypothesis to be tested) the statistic (1) has asymptotically the chi-square distribution with $k - 1$ degrees of freedom, when each Np_i is large. To test the null hypothesis the upper tail of the chi-square distribution is used as a critical region.

In the literature only rules of thumb are found as to the choice of the number and lengths of the class intervals. It is the purpose of this paper to formulate principles for this choice and to determine the number and lengths of the class intervals according to these principles.

If a choice is made as to the number of class intervals it is always possible to find alternative hypotheses with class probabilities equal to the class probabilities under the null hypothesis. The least upper bound of the "distances" of such alternative distributions from the null hypothesis distribution can evidently be minimized by making the class probabilities under the null hypothesis equal to each other. By the distance of two distribution functions we mean the least upper bound of the absolute value of the difference of the two cumulative distribution functions. We have therefore based this paper on a procedure by which the lengths of the class intervals are determined so that the probability of each class under the null hypothesis is equal to $1/k$ where k is the number of class intervals.[2]

Let $C(\Delta)$ be the class of alternative distributions with a distance $\geqq \Delta$ from the null hypothesis. Let $f(N, k, \Delta)$ be the greatest lower bound of the power of the chi-square test with sample size N and number of class intervals k with respect to alternatives in $C(\Delta)$. The maximum of $f(N, k, \Delta)$ with respect to k is a function $\Phi(N, \Delta)$ of N and Δ. It is most desirable to maximize $f(N, k, \Delta)$ for

[1] Research under a grant in aid from the Carnegie Corporation of New York.

[2] This procedure was first used by H. Hotelling. "The consistency and ultimate distribution of optimum statistics," *Trans. Am. Math. Soc.*, Vol. 32, pp. 851.) It has been advocated by E. J. Gumbel in a paper which will appear shortly.