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.

such values of Δ for which $\Phi(N, \Delta)$ is neither too large nor too small and in this paper we propose to determine Δ so that $\Phi(N, \Delta)$ is equal to $\frac{1}{2}$.

Hence we introduce the following definitions:

DEFINITION 1. *A positive integer k is called best with respect to the number of observations N if there exists a Δ such that $f(N, k, \Delta) = \frac{1}{2}$ and $f(N, k', \Delta) \leqslant \frac{1}{2}$ for any positive integer k'.*

DEFINITION 2. *A positive integer k is called ϵ-best ($0 \leqslant \epsilon \leqslant 1$) with respect to the number of observations N if ϵ is the smallest number in the interval [0, 1] for which the following condition is fulfilled: There exists a Δ such that $f(N, k, \Delta) \geqslant \frac{1}{2} - \epsilon$ and $f(N, k', \Delta) \leqslant \frac{1}{2} + \epsilon$ for any positive integer k'.*

It is obvious that an ϵ-best k is a best k if $\epsilon = 0$. If ϵ is very small an ϵ-best k is for all practical purposes equivalent to a best k.

Since $f(N, k, \Delta)$ is a continuous function of Δ it is easy to see that for any pair of positive integers k and N there exists exactly one value ϵ such that k is ϵ-best with respect to the number of observations N. Since the value of this ϵ is a function of k and N we will denote it by $\epsilon(k, N)$.

DEFINITION 3. *A sequence $\{k_N\}$ of positive integers is called best in the limit if* $\lim_{N=\infty} \epsilon(k_N, N) = 0$.

In this paper the following theorem is proved:

THEOREM 1. *Let* $k_N = 4\sqrt[5]{\frac{2(N-1)^2}{c^2}}$ *where c is determined so that* $\frac{1}{\sqrt{2\pi}}\int_c^\infty e^{-x^2/2}\,dx$ *is equal to the size of the critical region (probability of the critical region under the null hypothesis) then the sequence $\{k_N\}$ is best in the limit. Furthermore* $\lim_{N=\infty} f(N, k_N, \Delta_N) = \frac{1}{2}$ *for* $\Delta_N = \frac{5}{k_N} - \frac{4}{k_N^2}$.

It is further shown that for $N \geqslant 450$, if the 5% level of significance is used, and for $N \geqslant 300$, if the 1% level of significance is used, the value of $\epsilon(k_N, N)$ is small so that for practical purposes k_N can be considered as a best k. The authors are convinced although no rigorous proof has been given that $\epsilon(k_N, N)$ is quite small for $N \geqslant 200$ and is very likely to be small even for considerably lower values of N.

1. Mean value and standard deviation of the statistic under alternative hypotheses. It is well known that every continuous distribution can by a simple transformation be transformed into a rectangular distribution with range [0, 1]. We may therefore for convenience assume that the hypothesis to be tested is that of a rectangular distribution with the range [0, 1]. Moreover as mentioned earlier we assume that a procedure is chosen by which the class probabilities under the null hypothesis are equal to each other.

The statistic whose mean value and standard deviation is to be determined is

$$\sum_{i=1}^{i=k} x_i^2 = \chi'^2 \quad \text{where} \quad x_i = \sqrt{\frac{k}{N}}\left(\alpha_i - \frac{N}{k}\right).$$

Let p_i be the probability under the alternative hypothesis that one observation will fall into the ith class. The probability of obtaining certain specified values $\alpha_1, \alpha_2, \cdots, \alpha_k$ is given by

$$f(\alpha_1, \alpha_2, \cdots \alpha_k) = \frac{N!}{\alpha_1! \alpha_2! \cdots \alpha_k!} p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}.$$

Since $\sum_{i=1}^{i=k} \alpha_i = N$ we have

$$\sum_{i=1}^{i=k} x_i^2 = \frac{k}{N} \sum_{i=1}^{i=k} \alpha_i^2 - N.$$

We consider the function

$$(p_1 e^{t_1} + p_2 e^{t_2} + \cdots p_k e^{t_k})^N = \Sigma f(\alpha_1, \alpha_2, \cdots \alpha_k) e^{\alpha_1 t_1 + \alpha_2 t_2 + \cdots \alpha_k t_k}.$$

Differentiating twice and then setting $t_i = 0$ for $i = 1, 2, \cdots k$ we obtain

$$(2) \quad N(N-1)p_i^2 + Np_i = E(\alpha_i^2), \quad N(N-1)p_i p_j = E(\alpha_i \alpha_j) \text{ for } i \neq j.$$

Hence

$$E\left(\sum_{i=1}^{i=k} a_i^2\right) = N(N-1) \sum_{i=1}^{i=k} p_i^2 + N,$$

and

$$(3) \qquad E(\chi'^2) = k(N-1) \sum_{i=1}^{i=k} p_i^2 + k - N.$$

To compute the standard deviation of χ'^2 we put

$$\mu_i = \left(Np_i - \frac{N}{k}\right)\sqrt{\frac{k}{N}} = \sqrt{Nk}\left(p_i - \frac{1}{k}\right),$$

$$y_i = (\alpha_i - Np_i)\sqrt{\frac{k}{N}} \qquad \text{hence} \qquad y_i = x_i - \mu_i, \qquad E(y_i) = 0.$$

We have

$$\sigma_{\chi'^2}^2 = E\left[\sum_{i=1}^{i=k} (y_i + \mu_i)^2 - E\left(\sum_{i=1}^{i=k} (y_i + \mu_i)^2\right)\right]^2$$

$$= E\left(\sum_{i=1}^{i=k} y_i^2 + 2\sum_{i=1}^{i=k} y_i \mu_i - E\left(\sum_{i=1}^{i=k} y_i^2\right)\right)^2.$$

Let

$$\sqrt{\frac{N}{k}}\, y_i = z_i, \qquad \sqrt{\frac{N}{k}}\, \mu_i = \nu_i;$$

then

$$\nu_i = N\left(p_i - \frac{1}{k}\right), \qquad z_i = \alpha_i - Np_i.$$

We now assume that N is so large that the joint distribution of the z_i is sufficiently well approximated by a multivariate normal distribution. Then

$E(z_i^2 z_j) = 0, \quad E(z_i^4) = 3[E(z_i^2)]^2, \quad E(z_i^2 z_j^2) = E(z_i^2)E(z_j^2) + 2[E(z_i z_j)]^2$ for $i \neq j$.

We have the well known relations

$$E(z_i^2) = E(\alpha_i^2) - N^2 p_i^2 = Np_i(1 - p_i),$$
$$E(z_i z_j) = E(\alpha_i \alpha_j) - N^2 p_i p_j = -Np_i p_j .$$

Using the above equations we obtain

$$\sigma_{\chi'^2}^2 = \frac{k^2}{N^2}\left\{E\left(\sum_{i=1}^{i=k} z_i^2\right)^2 - \left(E\sum_{i=1}^{i=k} z_i^2\right)^2 + 4E\left(\sum_{i=1}^{i=k} z_i \nu_i\right)^2\right\},$$

$$E\left(\sum_{i=1}^{i=k} z_i^2\right)^2 - \left(E\sum_{i=1}^{i=k} z_i^2\right)^2$$
$$= N^2\left\{3\sum_{i=1}^{i=k} p_i^2(1-p_i)^2 + \sum_{i \neq j} [p_i p_j(1-p_i)(1-p_j) + 2p_i^2 p_j^2] - \left[\sum_{i=1}^{i=k} p_i(1-p_i)\right]^2\right\}$$
$$= 2N^2\left[\sum_{i=1}^{i=k} p_i^2(1-p_i)^2 + \sum_{i \neq j} p_i^2 p_j^2\right]$$
$$= 2N^2\left[\sum_{i=1}^{i=k} p_i^2 - 2\sum_{i=1}^{i=k} p_i^3 + \left(\sum_{i=1}^{i=k} p_i^2\right)^2\right].$$

Further

$$E\left(\sum_{i=1}^{i=k} z_i \nu_i\right)^2 = E\left(\sum_{i=1}^{i=k} z_i^2 \nu_i^2\right) + E\left(\sum_{i \neq j} z_i z_j \nu_i \nu_j\right)$$
$$= N^3\left[\sum_{i=1}^{i=k} p_i(1-p_i)\left(p_i - \frac{1}{k}\right)^2 - \sum_{i \neq j} p_i p_j\left(p_i - \frac{1}{k}\right)\left(p_j - \frac{1}{k}\right)\right]$$
$$= N^3\left[\sum_{i=1}^{i=k} p_i\left(p_i - \frac{1}{k}\right)^2 - \left[\sum_{i=1}^{i=k} p_i\left(p_i - \frac{1}{k}\right)\right]^2\right]$$
$$= N^3\left[\sum_{i=1}^{i=k} p_i^3 - \frac{2}{k}\sum_{i=1}^{i=k} p_i^2 + \frac{1}{k^2} - \left[\sum_{i=1}^{i=k} p_i^2 - \frac{1}{k}\right]^2\right]$$
$$= N^3\left[\sum_{i=1}^{i=k} p_i^3 - (\sum p_i^2)^2\right].$$

Substituting this into the formula for $\sigma_{\chi'^2}^2$ we finally obtain

$$\sigma_{\chi'^2}^2 = 2k^2\left\{\sum_{i=1}^{i=k} p_i^2 + 2(N-1)\sum_{i=1}^{i=k} p_i^3 - (2N-1)\left(\sum_{i=1}^{i=k} p_i^2\right)^2\right\}. \tag{4}$$

2. The Taylor expansion of the power. Let C be determined so that the probability under the null hypothesis that $\sum_{i=1}^{i=k} x_i^2 \geqslant C$ is equal to the size λ_0 of

the critical region. Let $P\left(\sum_{i=1}^{i=k} x_i^2 \geqq C\right)$ be the probability under the alternative hypothesis that $\sum_{i=1}^{i=k} x_i^2 \geqq C$. Then the power P is given by

$$P\left(\sum_{i=1}^{i=k} x_i^2 \geqq C\right), \tag{5}$$

where

$$x_i = \frac{\alpha_i - \frac{N}{k}}{\sqrt{\frac{N}{k}}}.$$

Hence

$$\sum_{i=1}^{i=k} x_i^2 = \frac{k}{N}\left(\sum_{i=1}^{i=k} \alpha_i^2 - \frac{N^2}{k}\right),$$

and (5) can be written in the form

$$P\left(\sum_{i=1}^{i=k} \alpha_i^2 \geqq C'\right) \tag{6}$$

where C' is a certain function of N and k. Let $\delta_i = p_i - \frac{1}{k}$, where p_i is the probability of the ith class interval under the alternative hypothesis.

Expanding P into a power series we obtain (in this and the following derivations, we take all partial differential quotients at the point $\delta_1 = \delta_2 = \cdots = \delta_k = 0$)

$$P = \lambda_0 + \sum_{i=1}^{i=k} \delta_i \frac{\partial P}{\partial \delta_i} + \frac{1}{2}\left\{\sum_{i=1}^{i=k} \delta_i^2 \frac{\partial^2 P}{\partial \delta_i^2} + \sum_{i \neq j} \delta_i \delta_j \frac{\partial^2 P}{\partial \delta_i \partial \delta_j}\right\} + \cdots.$$

Since P is a symmetric function of the δ_i we have for $\delta_1 = \delta_2 = \cdots = \delta_k = 0$

$$\frac{\partial^2 P}{\partial \delta_i^2} = \frac{\partial^2 P}{\partial \delta_1^2}, \qquad \frac{\partial^2 P}{\partial \delta_i \partial \delta_j} = \frac{\partial^2 P}{\partial \delta_1 \partial \delta_2} \qquad \text{for } i \neq j.$$

Furthermore $\sum_{i=1}^{i=k} \delta_i = 0$. Therefore

$$P = \lambda_0 + \frac{1}{2}\left\{\frac{\partial^2 P}{\partial \delta_1^2} \sum_{i=1}^{i=k} \delta_i^2 + \frac{\partial^2 P}{\partial \delta_1 \partial \delta_2} \sum_{i \neq j} \delta_i \delta_j\right\} + \cdots.$$

We shall first show that the terms of second order are always positive. This shows that the test is unbiased and justifies again the choice of equal class probabilities under the null hypothesis since this assures unbiasedness and mini-

mizes among all unbiased tests the g.l.b. of the distances of such alternatives whose power is equal to the size of the critical region.
The power is given by

$$P = \sum_{\alpha_1^2+\alpha_2^2+\cdots\alpha_k^2 \geq C'} \frac{N!}{\alpha_1!\alpha_2!\cdots\alpha_k!} p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}.$$

Since $\sum_{i=1}^{i=k} \delta_i^2 = -\sum_{i \neq j} \delta_i \delta_j$ we obtain for the second order terms

$$\frac{\partial^2 P}{\partial \delta_1^2} \sum_{i=1}^{i=k} \delta_i^2 + \frac{\partial^2 P}{\partial \delta_1 \partial \delta_2} \sum_{i \neq j} \delta_i \delta_j = \left(\frac{\partial^2 P}{\partial \delta_1^2} - \frac{\partial^2 P}{\partial \delta_1 \partial \delta_2}\right) \sum_{i=1}^{i=k} \delta_i^2 \tag{7}$$

$$= \sum_{\alpha_1^2+\alpha_2^2+\cdots+\alpha_k^2 \geq C'} (\alpha_1^2 - \alpha_1 - \alpha_1\alpha_2) p(\alpha_1, \alpha_2 \cdots \alpha_k) \sum_{i=1}^{i=k} \delta_i^2$$

where

$$p(\alpha_1, \cdots \alpha_k) = \frac{N!}{\alpha_1!\alpha_2!\cdots\alpha_k!} \frac{1}{k^N}.$$

In the following derivation extend all sums if not otherwise stated over all terms for which $\sum_{i=1}^{i=k} \alpha_i^2 \geqslant C'$ and use the relation $\sum_{i=1}^{i=k} \alpha_i = N$. We have because of the symmetry

$$\sum \alpha_1 p(\alpha_1, \alpha_2, \cdots \alpha_k) = \frac{N}{k} \sum p(\alpha_1, \alpha_2 \cdots \alpha_k) = \frac{N}{k} \lambda_0,$$

$$\sum \alpha_1\alpha_2 p(\alpha_1, \alpha_2, \cdots \alpha_k) = \frac{1}{k(k-1)} \sum \left(N^2 - \sum_{i=1}^{i=k} \alpha_i^2\right) p(\alpha_1, \alpha_2, \cdots \alpha_k)$$

$$= \frac{N^2\lambda_0}{k(k-1)} - \frac{1}{k-1} \sum \alpha_1^2 p(\alpha_1, \alpha_2, \cdots \alpha_k).$$

Hence the coefficient of the second order term becomes

$$\frac{k}{k-1} \sum \alpha_1^2 p(\alpha_1, \alpha_2, \cdots \alpha_k) - \frac{N}{k}\lambda_0 - \frac{N^2}{k(k-1)}\lambda_0$$

$$= \frac{1}{k-1} \sum \sum_{i=1}^{i=k} \alpha_i^2 p(\alpha_1, \alpha_2, \cdots \alpha_k) - \frac{N}{k}\lambda_0 - \frac{N^2}{k(k-1)}\lambda_0.$$

But

$$\frac{\sum \sum_{i=1}^{i=k} \alpha_i^2 p(\alpha_1, \alpha_2, \cdots \alpha_k)}{\lambda_0} > E\left(\sum_{i=1}^{i=k} \alpha_i^2\right),$$

since the conditional mean for values of $\sum_{i=1}^{i=k} \alpha_i^2 \geqslant C'$ must be larger than the

mean of all values of $\sum_{i=1}^{i=k} \alpha_i^2$. Since $E\left(\sum_{i=1}^{i=k} \alpha_i^2\right) = \frac{N^2}{k} - \frac{N}{k} + N$, we obtain

$$\frac{1}{k-1} \sum \sum_{i=1}^{i=k} \alpha_i^2 p(\alpha_1, \alpha_2 \cdots \alpha_k)$$

$$> \frac{\lambda_0}{k-1}\left(\frac{N^2}{k} + \frac{N(k-1)}{k}\right) = \lambda_0\left(\frac{N^2}{k(k-1)} + \frac{N}{k}\right)$$

and hence the coefficient of $\sum_{i=1}^{i=k} \delta_i^2$ is larger than 0.

To prove Theorem 1, we will have to determine the alternative distribution for which $\sum_{i=1}^{i=k} \delta_i^2$ becomes a minimum subject to the condition that the distance from the null hypothesis should be greater than or equal to a given Δ.

Hence we have to find a distribution function $F(x)$ such that $| F(x) - x | \geqslant \Delta$ for at least one value x and $\sum_{i=1}^{i=k} \delta_i^2 = \sum_{i=1}^{i=k} \left(p_i - \frac{1}{k}\right)^2 = \sum_{i=1}^{i=k} p_i^2 - \frac{1}{k}$ is a minimum where $p_i = F\left(\frac{i}{k}\right) - F\left(\frac{i-1}{k}\right)$. Instead of minimizing $\sum_{i=1}^{i=k} \delta_i^2$ we may minimize $\sum_{i=1}^{i=k} p_i^2$, since the two expressions differ merely by a constant. There will be two different solutions for $F(x)$ depending on whether $F(x) - x \geqslant \Delta$ or $F(x) - x \leqslant -\Delta$ for at least one value x. Because of symmetry we restrict ourselves to the case in which $F(x) - x \geqslant \Delta$ for at least one value of x.

Let a be a value for which $F(a) - a \geqslant \Delta$ and suppose that

$$\frac{l-1}{k} < a \leqslant \frac{l}{k}$$

then

$$F(a) \geqslant a + \Delta,$$

$$F\left(\frac{l}{k}\right) = \frac{l}{k} + \epsilon.$$

We prove first

$$\epsilon \geqslant \Delta - \frac{1}{k}.$$

PROOF: Since $F\left(\frac{l}{k}\right) - F(a) \geqslant 0$ we have

$$F\left(\frac{l}{k}\right) = F(a) + F\left(\frac{l}{k}\right) - F(a) \geqslant a + \Delta$$

and

$$\epsilon = F\left(\frac{l}{k}\right) - \frac{l}{k} \geqslant a + \Delta - \frac{l}{k} \geqslant \frac{l-1}{k} + \Delta - \frac{l}{k} = \Delta - \frac{1}{k}.$$

If $\Delta \leqslant \frac{1}{k}$ we can always find a distribution function in $C(\Delta)$ for which $p_i = \frac{1}{k}$. Hence we consider only the case $k > \frac{1}{\Delta}$. We must minimize $\sum_{i=1}^{i=k} p_i^2$ under the condition $\sum_{i=1}^{i=l} p_i = \frac{l}{k} + \epsilon, \sum_{i=l+1}^{i=k} p_i = \frac{k-l}{k} - \epsilon$. We therefore minimize

$$\Phi = \sum_{i=1}^{i=k} p_i^2 - 2\lambda_1 \sum_{i=1}^{i=l} p_i - 2\lambda_2 \sum_{i=l+1}^{i=k} p_i.$$

This leads to

$$p_i = \begin{cases} \frac{1}{k} + \frac{\epsilon}{l} & \text{for } i = 1, \cdots l \\ \frac{1}{k} - \frac{\epsilon}{k-l} & \text{for } i = (l+1), \cdots k. \end{cases}$$

We then have

$$\sum_{i=1}^{i=k} p_i^2 = l\left(\frac{1}{k} + \frac{\epsilon}{l}\right)^2 + (k-l)\left(\frac{1}{k} - \frac{\epsilon}{k-l}\right)^2 = \frac{1}{k} + \frac{\epsilon^2 k}{l(k-l)}.$$

This is smallest if $\epsilon = \Delta - \frac{1}{k}$ and $l = \frac{k}{2}$. The following discontinuous distribution function gives these values for ϵ, l and p_i and has the distance Δ from the rectangular distribution.

$$(8)\quad \begin{aligned} F(x) &= x\left[1 + 2\left(\Delta - \frac{1}{k}\right)\right] && \text{for } 0 \leqslant x \leqslant \frac{1}{2} - \frac{1}{k}, \\ F(x) &= \frac{1}{2} + \Delta - \frac{1}{k} && \text{for } \frac{1}{2} - \frac{1}{k} < x \leqslant \frac{1}{2}, \\ F(x) &= x\left[1 - 2\left(\Delta - \frac{1}{k}\right)\right] + 2\left(\Delta - \frac{1}{k}\right) && \text{for } \frac{1}{2} \leqslant x \leqslant 1, \\ F(x) &= 0 && \text{for } 0 \leqslant x, \\ F(x) &= 1 && \text{for } x \geqslant 1. \end{aligned}$$

3. Solution for large N. Denote by $F(\Delta, k)$ the distribution function (8) of $C(\Delta)$ which makes $\sum_{i=1}^{i=k} \delta_i^2$ a minimum if the test is made with k class intervals.

Assume that k is large enough that χ'^2 can be taken as normally distributed. The power of the test is then given by

$$(9)\qquad \frac{1}{\sqrt{2\pi}\,\sigma'}\int_{(k-1)+c\sqrt{2(k-1)}}^{\infty} e^{-\frac{1}{2\sigma'^2}\left(\sum_{i=1}^{i=k} x_i^2 - E\left(\sum_{i=1}^{i=k} x_i^2\right)\right)^2} d\left(\sum_{i=1}^{i=k} x_i^2\right) = \frac{1}{\sqrt{2\pi}}\int_{\frac{k-1-E\left(\sum_{i=1}^{i=k} x_i^2\right)+c\sqrt{2(k-1)}}{\sigma'}}^{\infty} e^{-\frac{1}{2}y^2}\,dy,$$

where σ' is the standard deviation of $\sum_{i=1}^{i=k} x_i^2$ and c is determined so that $\frac{1}{\sqrt{2\pi}}\int_c^\infty e^{-\frac{1}{2}y^2}\,dy$ is equal to the size of the critical region. Hence to maximize the power with respect to k is equivalent to maximizing

$$\psi(k) = \frac{E\left(\sum_{i=1}^{i=k} x_i^2\right) - (k-1) - c\sqrt{2(k-1)}}{\sigma'}$$

with respect to k.

Under the alternative $F(\Delta, k)$ we obtain

$$E\left(\sum_{i=1}^{i=k} x_i^2\right) - (k-1) = k(N-1)\sum_{i=1}^{i=k} p_i^2 + k - N - k + 1 = 4(N-1)\left(\Delta - \frac{1}{k}\right)^2$$

Hence

$$\psi(k) = \frac{4(N-1)\left(\Delta - \frac{1}{k}\right)^2 - c\sqrt{2(k-1)}}{\sigma'}.$$

We choose Δ so that this maximum power is exactly $\frac{1}{2}$, that is, so that $\psi(k) = 0$ for that k which maximizes $\psi(k)$. Denote this value of Δ by Δ_N and let k_N be the value of k which maximizes $\psi(k)$. The differential-quotient of the numerator of $\psi(k)$ with respect to k is then equal to 0 for $k = k_N$. Hence

$$(10)\qquad 8(N-1)\left(\Delta_N - \frac{1}{k_N}\right)\frac{1}{k_N^2} = \frac{c}{\sqrt{2(k_N-1)}}.$$

Furthermore since $\psi(k_N) = 0$ we have

$$(11)\qquad 4(N-1)\left(\Delta_N - \frac{1}{k_N}\right)^2 = c\sqrt{2(k_N-1)}.$$

Solving equations (10) and (11) we obtain

$$(12)\qquad \Delta_N = \frac{5}{k_N} - \frac{4}{k_N^2}$$

and

$$\sqrt[5]{\frac{k_N^8}{(k_N-1)^5}} = 4\sqrt[5]{\frac{2(N-1)^2}{c^2}}$$

or since $k_N > 3$,

$$k_N < 4\sqrt[5]{\frac{2(N-1)^2}{c^2}} < k_N + 1.$$

Hence

$$(13) \quad \text{either } k_N = \left[4\sqrt[5]{\frac{2(N-1)^2}{c^2}}\right] \quad \text{or} \quad k_N = \left[4\sqrt[5]{\frac{2(N-1)^2}{c^2}}\right] + 1,$$

is the value of k for which the power with respect to $F(\Delta_N, k)$ becomes a maximum. We have merely to show that $\psi''(k)$ is negative for $k = k_N$.

Using the fact that $\psi(k_N) = \psi'(k_N) = 0$ we obtain

$$\sigma'\psi''(k_N) = \frac{-16(N-1)}{k_N^3}\Delta_N + \frac{24(N-1)}{k_N^4} + \frac{c}{(\sqrt{2(k_N-1)})^3}.$$

Substituting for Δ_N the right hand side of (12) we obtain on account of (10)

$$\sigma'\psi''(k_N) = \frac{-56(N-1)}{k_N^4} + \frac{64(N-1)}{k_N^5} + \frac{8(N-1)}{2(k-1)}\left(\frac{4}{k_N^3} - \frac{4}{k_N^4}\right).$$

Using $2(k-1) > k$ we obtain

$$\psi''(k_N) < \frac{1}{k^4\sigma'}\left(-24(N-1) + \frac{32}{k}(N-1)\right)$$

which is negative. σ' can be shown to be of order $k_N^{\frac{1}{2}}$; $\psi''(k_N)$ is, therefore, of order $\frac{N}{k_N^{4+\frac{1}{2}}} = 0\left(\frac{1}{N^{\frac{1}{2}}}\right)$. The maximum is, therefore, rather flat for large values of N.

We shall now show that if k is large enough to assume χ'^2 to be normally distributed then $F(\Delta, k)$ is the alternative which gives the smallest power compared with all alternatives in the class $C(\Delta)$ provided the power for the alternative $F(\Delta, k)$ equals $\frac{1}{2}$.

We know that $E\left(\sum_{i=1}^{i=k} x_i^2\right)$ is smallest for $F(\Delta, k)$. Since the power with respect to $F(\Delta, k)$ equals $\frac{1}{2}$ we have

$$E\left(\sum_{i=1}^{i=k} x_i^2\right) - (k - 1 - c\sqrt{2(k-1)} = 0.$$

Thus the lower limit of the integral in (9) becomes negative for every other alternative and the power will be larger than $\frac{1}{2}$.

The power with respect to $F(\Delta_N, k_N)$ is equal to $\frac{1}{2}$, hence if we choose $k = k_N$ the power of the test will be $\geqslant \frac{1}{2}$ for all alternatives in the class $C(\Delta_N)$. On the other hand if we choose $k \neq k_N$ then there will be at least one alternative in

[3] Cantelli's formula and its proof are given by Fréchet in his book *Recherches Théoriques Modernes sur la Théorie de Probabilités*, Paris (1937), pp. 123–126.

$C(\Delta_N)$ for which the power is $<\frac{1}{2}$. (For instance $F(\Delta_N, k)$ is such an alternative.)

The above statements have been derived under the assumption that χ'^2 is normally distributed. Hence if the distribution of χ'^2 were exactly normal $k_N = 4\sqrt[5]{\frac{2(N-1)^2}{c^2}}$ would be a best k and for this k_N and $\Delta_N = \frac{5}{k_N} - \frac{4}{k_N^2}$ the greatest lower bound of the power in the class $C(\Delta_N)$ would be exactly $\frac{1}{2}$. Since the distribution of χ'^2 approaches the normal distribution with $k \to \infty$ the sequence $\{k_N\}$ is best in the limit and Theorem 1 stated in the introduction is proved.

For the purposes of practical applications, it is not enough to know that $\{k_N\}$ is best in the limit. We have to know for what values of N k_N can be considered practically as a best k, i.e. for what values of N the quantity $\epsilon(k_N, N)$ defined in the introduction is sufficiently small. The quantity $\epsilon(k_N, N)$ is certainly small if for the number of class intervals k_N the distribution of χ'^2 is near to normal and if the power with respect to at least one alternative of the class $C(\Delta_N)$ is smaller than $\frac{1}{2}$ also in the case when the number of class intervals is too small to assume a normal distribution for χ'^2.

We shall in the following assume that for $k > 13$ the normal distribution is a sufficiently good approximation. Actually we need not assume a normal distribution but only that the probability is close to $\frac{1}{2}$ that the statistic will exceed its mean value.

Cantelli[3] gave the following formula. Let M_r be the rth moment of a distribution about x_0. Let d be any arbitrary positive number. Let $P(|x - x_0| \leqslant d)$ be the probability that $|x - x_0| \leqslant d$ then the following inequalities hold:

$$\text{If} \quad \frac{M_r}{d^r} \leqslant \frac{M_{2r}}{d^{2r}} \quad \text{then} \quad P(|x - x_0| \leqslant d) \geqslant 1 - \frac{M_r}{d^r}.$$

$$\text{If} \quad \frac{M_r}{d^r} \geqslant \frac{M_{2r}}{d^{2r}} \quad \text{then} \quad P(|x - x_0| \leqslant d) \geqslant 1 - \frac{M_{2r} - M_r^2}{(d^r - M_r)^2 + M_{2r} - M_r^2}.$$

Since χ'^2 can only take positive values we have

$$(14) \quad \text{If} \quad \frac{E(\chi'^2)}{c_k} \leqslant \frac{\sigma_{\chi'^2}^2 + [E(\chi'^2)]^2}{c_k^2} \quad \text{then} \quad P(\chi'^2 \leqslant c_k) \geqslant 1 - \frac{E(\chi'^2)}{c_k}.$$

$$(15) \quad \text{If} \quad \frac{E(\chi'^2)}{c_k} \geqslant \frac{\sigma_{\chi'^2}^2 + [E(\chi'^2)]^2}{c_k^2}$$

$$\text{then} \quad P(\chi'^2 \leqslant c_k) \geqslant 1 - \frac{\sigma_{\chi'^2}^2}{(c_k - E(\chi'^2))^2 + \sigma_{\chi'^2}^2}.$$

Where c_k is determined so that $P(\chi^2 \geqslant c_k)$ equals the size of the critical region if the null hypothesis is true and the number of class intervals equals k. c_k can be obtained from a table of the chi-square distribution.

For $F(\Delta_N, k)$ we obtain with $\Delta_N' = \frac{5}{k_N} - \frac{4}{k_N^2} - \frac{1}{k}$ from (3) and (4)

$$E(\chi'^2) = (k-1) + 4(N-1)\Delta_N'^2,$$

$$\sigma_{\chi'^2}^2 = 2(k-1) + 8\Delta_N'^2(k+2N-4) - 32(2N-1)\Delta_N'^4.$$

By numerically calculating $E(\chi'^2)$ and $\sigma_{\chi'^2}$ for $N = 450$ and a 5% level of significance, for $N = 300$ and a 1% level of significance, and for $k = 13, 12 \cdots \left[\frac{1}{\Delta_N}\right] + 1$ it can be shown that for these values of N and k

$$\frac{E(\chi'^2)}{c_k} \geqslant \frac{\sigma_{\chi'^2}^2 + [E(\chi'^2)]^2}{c_k^2}. \tag{16}$$

Hence we have to use (15). From (16) it follows that $c_k > E(\chi'^2)$. If $P(\chi'^2 \leqslant c_k \leqslant \frac{1}{2}$ we obtain on account of (15) and (16)

$$\frac{\sigma_{\chi'^2}^2}{(c_k - E(\chi'^2))^2 + \sigma_{\chi'^2}^2} \geqslant \frac{1}{2}, \qquad \sigma_{\chi'^2} + E(\chi'^2) \geqslant c_k.$$

Numerical calculation shows that for the values of N and k and the significance levels considered

$$\sigma_{\chi'^2} + E(\chi'^2) < c_k. \tag{17}$$

It can then be shown that for $N \geqslant 450$ and $N \geqslant 300$ respectively $N\Delta_N'$ decreases with N. A simple argument then shows that (16) and (17) are also true for all values $N \geqslant 450$ and $N \geqslant 300$ respectively. Hence the power with respect to $F(\Delta_N, k)$ is $<\frac{1}{2}$ for these values of N. Thus we see: For $N \geqslant 450$ if the 5% level is used, and for $N \geqslant 300$ if the 1% level is used, the value $k_N = 4\sqrt[5]{\frac{2(N-1)^2}{c^2}}$ can be considered for practical purposes as a best k. The value c is determined so that $\frac{1}{\sqrt{2\pi}}\int_c^\infty e^{-\frac{1}{2}t^2}\,dt$ is equal to the size of the critical region.