

Title	Some combinatorial tests of goodness of fit
Author(s)	Okamoto, Masashi
Citation	Osaka Mathematical Journal. 1952, 4(2), p. 215-228
Version Type	VoR
URL	https://doi.org/10.18910/10130
rights	
Note	

Osaka University Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

Osaka University

Some Combinatorial Tests of Goodness of Fit

By Masashi OKAMOTO

1. Introduction. We have recently [1] considered a test of goodness of fit, i.e., a test whether a random sample has come from the population with the specified continuous distribution. We now present a new approach to the same problem.

Let X_1, \dots, X_N be random variables distributed independently and identically according to the *d.f.* $F(x)$. To simplify the situation it is assumed that X 's range from 0 to 1. The hypothesis H_0 to be tested is that $F(x)$ is identical with the *d.f.* $F_0(x)$ of uniform distribution on the interval $(0, 1]$. We divide the interval in n small intervals $((i-1)/n, i/n]$, $i = 1, \dots, n$. In the sequel the word "interval" means if not stated otherwise any of these small intervals. Among $\binom{N}{k}$ k -tuples $(X_{\alpha_1}, \dots, X_{\alpha_k})$, $1 \leq \alpha_1 < \dots < \alpha_k \leq N$, we denote by M_k the number of those such that $X_{\alpha_1}, \dots, X_{\alpha_k}$ fall in the same interval. When we consider one observation, the more uniformly are X_1, \dots, X_N (observed values) distributed among the n intervals, the smaller becomes M_k , as shown in section 7. On account of this the following test (called M_k -test) of H_0 will be useful: we accept H_0 when M_k is sufficiently small.

It is proved in this paper that when the population distribution satisfies a certain condition M_k is asymptotically normally distributed as N and n tend to infinity (Theorems 1, 2, 1', 2'). Furthermore M_k -test is shown to be consistent (Theorem 3) and unbiased (Theorem 4) against a rather general class of alternatives. The statistics M_k are closely related with David's test (cf. [1], [2]) and can be considered as a generalisation of the chi-square test in the case of equal probability.

2. Definition of U_k . For real numbers t_1, \dots, t_k such that $0 < t_1 \leq 1$, $i = 1, \dots, k$, we define

$$\begin{aligned} \Theta_k(t_1, \dots, t_k) &= 1, \text{ if } t_1, \dots, t_k \text{ fall in the same interval,} \\ &= 0, \text{ otherwise,} \end{aligned}$$

where the word "interval" means by convention any of intervals $((i-1)/n, i/n]$, $i = 1, \dots, n$. Then

$$(2.1) \quad M_k = \sum \Theta_k(X_{\alpha_1}, \dots, X_{\alpha_k}).$$

Throughout this section the sum \sum is extended over all k -tuples $(\alpha_1, \dots, \alpha_k)$, $1 \leq \alpha_1 < \dots < \alpha_k \leq N$.

Denoting by E_0 and D_0^2 the expectation and the variance, respectively, under H_0 , we have

$$(2.2) \quad E_0 \Theta_k(X_1, \dots, X_k) = n^{-(k-1)}.$$

Therefore, letting

$$(2.3) \quad \begin{aligned} \Phi_k(t_1, \dots, t_k) &= n^{k-1} \Theta_k(t_1, \dots, t_k) \\ &= \begin{cases} n^{k-1}, & \text{if } t_1, \dots, t_k \text{ fall in the same interval,} \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

we have

$$(2.4) \quad E_0 \Phi_k(X_1, \dots, X_k) = 1.$$

Furthermore, defining

$$(2.5) \quad \begin{aligned} \psi_k(t_1, \dots, t_k) &= \Phi_k(t_1, \dots, t_k) - 1 \\ &= \begin{cases} n^{k-1} - 1, & \text{if } t_1, \dots, t_k \text{ fall in the same interval,} \\ -1, & \text{otherwise,} \end{cases} \end{aligned}$$

we obtain

$$(2.6) \quad E_0 \psi_k(X_1, \dots, X_k) = 0,$$

and

$$(2.7) \quad M_k = n^{-(k-1)} \binom{N}{k} (U_k + 1),$$

where

$$(2.8) \quad U_k = \binom{N}{k}^{-1} \sum \psi_k(X_{\alpha_1}, \dots, X_{\alpha_k}).$$

Formally U_k is the same as U -statistics of W. Hoeffding [3], but substantially they are quite different because the definition of $\psi_k(t_1, \dots, t_k)$ here contains n which tends to infinity with the sample size N . We can not therefore apply Hoeffding's results to our case and so we have to prove once again the asymptotic normality of U_k .

3. Expectation and variance under H_0 . From (2.6), (2.7), (2.8) we have

$$(3.1) \quad E_0(U_k) = 0,$$

$$(3.2) \quad E_0(M_k) = n^{-(k-1)} \binom{N}{k}.$$

In order to evaluate the variance of U_k it becomes necessary to prepare several computations. To begin with, by (2.2) we have

$$E_0 [\Theta_k(X_1, \dots, X_k)]^2 = n^{-(k-1)},$$

whence by (2.3), (2.5)

$$(3.3) \quad \begin{aligned} E_0 [\Phi_k(X_1, \dots, X_k)]^2 &= n^{k-1}, \\ E_0 [\Psi_k(X_1, \dots, X_k)]^2 &= n^{k-1} - 1. \end{aligned}$$

It follows easily that

$$(3.4) \quad E_0 \Psi_2(t_1, X_2) = 0,$$

$$(3.5) \quad E_0 \Psi_k(t_1, \dots, t_{k-1}, X_k) = \Psi_{k-1}(t_1, \dots, t_{k-1}), \quad k \geq 3,$$

and by induction

$$(3.6) \quad E_0 \Psi_k(t_1, X_2, \dots, X_k) = 0, \quad k \geq 2,$$

where the expectation is always with respect to X 's.

By means of these equations we can compute $D_0^2(U_k)$. That is

$$(3.7) \quad \begin{aligned} D_0^2(U_k) &= E_0(U_k^2) = \binom{N}{k}^{-2} E_0 [\sum \Psi_k(X_{\alpha_1}, \dots, X_{\alpha_k})]^2 \\ &= \binom{N}{k}^{-2} \sum_c \sum^{(c)} E_0 \Psi_k(X_{\alpha_1}, \dots, X_{\alpha_k}) \Psi_k(X_{\beta_1}, \dots, X_{\beta_k}), \end{aligned}$$

where $\sum^{(c)}$ stands for summation over all subscripts such that

$$1 \leq \alpha_1 < \dots < \alpha_k \leq N, \quad 1 \leq \beta_1 < \dots < \beta_k \leq N$$

and exactly c equations

$$\alpha_i = \beta_j$$

are satisfied. According this definition c must be greater than or equal to $2k - N$. By (2.6) and (3.6) every term in $\sum^{(0)}$ and $\sum^{(1)}$ vanishes. Therefore \sum_c in (3.7) may be extended from $c = c_0 \equiv \max(2, 2k - N)$ to $c = k$. Furthermore by (3.5) each term in $\sum^{(c)}$, $c \geq c_0$, is equal to $E_0 [\Psi_c(X_1, \dots, X_c)]^2 = n^{c-1} - 1$, and the number of terms in $\sum^{(c)}$ is

$$\frac{N!}{c!(k-c)!(k-c)!(N-2k+c)!} = \binom{N}{k} \binom{k}{c} \binom{N-k}{k-c}.$$

Hence

$$(3.8) \quad D_0^2(U_k) = \binom{N}{k}^{-1} \sum_{c=c_0}^k \binom{k}{c} \binom{N-k}{k-c} (n^{c-1} - 1).$$

Under the limiting condition

$$(3.9) \quad n \rightarrow \infty \quad \text{and} \quad N/n \rightarrow r \quad (\text{const})$$

we have

$$(3.10) \quad D_0^2(U_k) \sim \frac{1}{n} \cdot k! \sum_{c=2}^k \binom{k}{c} r^{-c} = \frac{1}{n} \sigma_{ok}^2 \quad (\text{say}).$$

From (3.2), (2.7), (3.10) it follows that

$$(3.11) \quad E_0(M_k/n) \sim r^k/k!,$$

$$(3.12) \quad D_0^2(M_k/n) \sim \frac{1}{n} \cdot \frac{r^{2k}}{k!} \sum_{c=2}^k \binom{k}{c} r^{-c} = \frac{1}{n} \sigma_{ok}^2 \text{ (say).}$$

In particular,

$$(3.13) \quad \begin{cases} D_0^2(U_2) = \frac{2(n-1)}{N(N-1)} \sim \frac{1}{n} \cdot \frac{2}{r^2}, \\ E_0(M_2/n) = \frac{N(N-1)}{2n^2} \sim \frac{r^2}{2}, \\ D_0^2(M_2/n) = \frac{(n-1)N(N-1)}{2n^4} \sim \frac{1}{n} \cdot \frac{r^2}{2}. \end{cases}$$

4. Asymptotic distribution under H_0 .

Theorem 1. *When $n \rightarrow \infty$ and $N/n \rightarrow r$ (const), U_k and M_k/n are asymptotically normal $(0, n^{-1}\sigma_{ok}^2)$ and $(r^k/k!, n^{-1}\sigma_{ok}^2)$, respectively, where the first term in each parenthesis refers to the asymptotic mean and the second term variance.*

Proof. As the proofs are almost similar for various values of k , we shall deal with only the case $k=2$ to avoid the excessive complication of subscripts. Thus we shall prove that $\sqrt{n} U_2$ is asymptotically normally distributed with mean zero and variance $2/r^2$, whence the asymptotic distribution of M_2/n is readily inferred.

Since $E_0(U_2) = 0$, we have only to show that the m -th moment μ_m ($m = 2, 3, \dots$) of $\sqrt{n} U_2$ tends to that of the normal distribution $(0, 2/r^2)$. Now

$$(4.1) \quad \begin{aligned} \mu_m &= E_0 \left[\sqrt{n} \binom{N}{2}^{-1} \sum \Psi_2(X_i, X_j) \right]^m \\ &= n^{m/2} \binom{N}{2}^{-m} \sum E_0 \Psi_2(X_{i_1}, X_{j_1}) \dots \Psi_2(X_{i_m}, X_{j_m}), \end{aligned}$$

where summation is extended over all sets of pairs $(i_1, j_1), \dots, (i_m, j_m)$, $1 \leq i_g < j_g \leq N$, $g = 1, \dots, m$. Denote by d the number of different ciphers among

$$(4.2) \quad i_1, j_1; \dots; i_m, j_m.$$

Classifying them by the equivalency relations $i_1 \simeq j_1, \dots, i_m \simeq j_m$, let e be the number of equivalence classes. Then

$$(4.3) \quad \mu_m = \sum_{e=1}^m \sum_{d=1}^{2m} A_{ed},$$

where

$$(4.4) \quad A_{ea} = n^{m/2} \binom{N}{2}^{-m} \sum^{\langle e, a \rangle} E_0 \Psi_2(X_{i_1}, X_{j_1}) \dots \Psi_2(X_{i_m}, X_{j_m}),$$

$\sum^{\langle e, a \rangle}$ standing for summation over all sets of pairs $(i_1, j_1), \dots, (i_m, j_m)$ such that the number of different ciphers is d and the number of equivalence classes is e .

We shall first evaluate

$$(4.5) \quad E_0 \psi_{r_2}(X_{i_1}, X_{j_1}) \dots \psi_{r_2}(X_{i_m}, X_{j_m})$$

in A_{ea} . Let e equivalence classes consist of m_1, \dots, m_e pairs. Obviously

$$(4.6) \quad m = m_1 + \dots + m_e.$$

In order to evaluate (4.5) we may assume without any loss of generality that these classes are (to avoid the typographical difficulty we put the subscripts of i, j 's in the parentheses after them),

$$(4.7.1) \quad i_{(1)}, j_{(1)}; \dots; i_{(m_1)}, j_{(m_1)};$$

$$(4.7.2) \quad i_{(m_1+1)}, j_{(m_1+1)}; \dots; i_{(m_1+m_2)}, j_{(m_1+m_2)};$$

.....

$$(4.7.e) \quad i_{(m_1 + \dots + m_{e-1} + 1)}, j_{(m_1 + \dots + m_{e-1} + 1)}; \dots; i_{(m)}, j_{(m)}.$$

Then, E_0 in (4.5) is distributed to e classes and (4.5) becomes the product of e expectations

$$(4.8.1) \quad E_0 \Psi_2(X_{i_{(1)}}, X_{j_{(1)}}) \dots \Psi_2(X_{i_{(m_1)}}, X_{j_{(m_1)}}),$$

$$(4.8.2) \quad E_0 \Psi_2(X_{i_{(m_1+1)}}, X_{j_{(m_1+1)}}) \dots \Psi_2(X_{i_{(m_1+m_2)}}, X_{j_{(m_1+m_2)}}),$$

.....

$$(4.8.e) \quad E_0 \Psi_2(X_{i_{(m_1 + \dots + m_{e-1} + 1)}}, X_{j_{(m_1 + \dots + m_{e-1} + 1)}}) \dots \Psi_2(X_{i_{(m)}}, X_{j_{(m)}}).$$

Denoting by d_g the number of different ciphers in the class (4.6.g), $g = 1, \dots, e$, we have

$$(4.9) \quad d = d_1 + \dots + d_e.$$

Since from (2.5) $\Psi_2(t_1, t_2) = n - 1$ or -1 and the probability that ν X 's fall in the same interval is $O(n^{-\nu+1})$, the order in n of (4.8.g) is

$$m_g - d_g + 1.$$

By (4.6) and (4.9), the order in n of (4.5) is

$$\sum_{g=1}^e (m_g - d_g + 1) = m - d + e.$$

Since $\sum^{\langle e, a \rangle}$ in (4.4) contains $O(N^a)$ terms of this magnitude, we have

$$(4.10) \quad A_{ea} = n^{m/2} \binom{N}{2}^{-m} O(N^a) O(n^{m-a+e}) = O(n^{e-m/2}).$$

If $e > m/2$, then from (4.6) there is at least one g such that $m_g = 1$ and (4.8.g) vanishes on account of (3.4), whence (4.5) also vanishes so that $A_{e,d} = 0$. After all

$$(4.11) \quad A_{e,d} = \begin{cases} 0, & \text{if } e > m/2, \\ O(n^{e-m/2}), & \text{if } e \leq m/2. \end{cases}$$

From (4.3) and (4.11) it follows that

$$(4.12) \quad \mu_m = o(1) \quad \text{for odd } m.$$

As for the case when m is even, we have only to consider $A_{e,d}$ for $e = m/2$ because of (4.11), i.e.,

$$(4.13) \quad \mu_m \sim \sum_{a=2}^{2m} A_{m/2, a} = A \quad (\text{say}).$$

(In the present case when $k=2$ $A = A_{m/2, m}$. In the general case, however, the definition above of A is necessary.) From the same reasoning above we may suppose $m_g = d_g = 2$, $g = 1, \dots, m/2$ and thus each class (4.7g) is of the form $i, j; i, j$. In order to evaluate A it is required to consider the classification of m pairs $(i_1, j_1), \dots, (i_m, j_m)$ into $m/2$ sets, each consisting of two pairs. It is easily seen that there are

$$(4.14) \quad \varphi(m) = \frac{m!}{2^{m/2}(m/2)!}$$

ways of such classification. Thus

$$(4.15) \quad A = \varphi(m) n^{m/2} \binom{N}{2}^{-m} \sum E_0 \Psi_2^2(X_{i(1)}, X_{j(1)}) \dots \Psi_2^2(X_{i(m/2)}, X_{j(m/2)}),$$

the sum extending over all sets of pairs $(i(1), j(1)), \dots, (i(m/2), j(m/2))$, where all i, j 's are different.

On the other hand we have

$$(4.16) \quad \varphi(m) \left[D_0^2(\sqrt{n} U_2) \right]^{m/2} \\ = \varphi(m) n^{m/2} \binom{N}{2}^{-m} \sum E_0 \Psi_2^2(X_{i(1)}, X_{j(1)}) \dots \Psi_2^2(X_{i(m/2)}, X_{j(m/2)}),$$

where the summation is extended over all subscripts such that $1 \leq i(g) < j(g) \leq N$, $g = 1, \dots, m/2$. ($i(g)$ and $j(g')$, $g \neq g'$, may take the same value.)

As two sums in (4.15) and (4.16) are equal in the highest order of N , we obtain consequently

$$A \sim \varphi(m) \left[D_0^2(\sqrt{n} U_2) \right]^{m/2} \sim \varphi(m) (2/r^2)^{m/2}.$$

This and (4.12) complete the proof.

5. **Expectation and variance in the general case.** In this and the following section we shall assume that the population *d.f.* $F(x)$ has the density function $f(x)$ which is continuous except for a finite number of points and such that $\int_0^1 f^m(x) dx$ ($m = 2, 3, \dots, 2k-1$) exist.

Putting

$$p_i = F\left(\frac{i}{n}\right) - F\left(\frac{i-1}{n}\right), \quad i = 1, \dots, n,$$

we have according to the mean value theorem

$$(5.1) \quad p_i = n^{-1} f\left(\frac{i}{n} - \frac{\theta_i}{n}\right), \quad 0 \leq \theta_i \leq 1.$$

Letting

$$(5.2) \quad p_{(k)} = \sum_{i=1}^n p_i^k,$$

we obtain from (5.1)

$$(5.3) \quad p_{(k)} \sim n^{-(k-1)} f_k,$$

where

$$(5.4) \quad f_k = \int_0^1 f^k(x) dx.$$

Define further

$$(5.5) \quad q_{(k)} = p_{(k)}^{-1} \sim n^{k-1} f_k^{-1}.$$

Now, denoting by E and D^2 the expectation and the variance, respectively, in the general case, we have

$$E\Theta_k(X_1, \dots, X_k) = p_{(k)}.$$

Defining

$$(5.6) \quad \Phi_k'(t_1, \dots, t_k) = q_{(k)} \Theta_k(t_1, \dots, t_k),$$

$$(5.7) \quad \Psi_k'(t_1, \dots, t_k) = \Phi_k'(t_1, \dots, t_k) - 1,$$

we have

$$(5.8) \quad E\Phi_k'(X_1, \dots, X_k) = 1,$$

$$(5.9) \quad E\Psi_k'(X_1, \dots, X_k) = 0.$$

The equation (2.1) implies

$$(5.10) \quad M_k = p_{(k)} \binom{N}{k} (U_k' + 1),$$

where

$$(5.11) \quad U_k' = \binom{N}{k}^{-1} \sum \Psi_k'(X_{\alpha_1}, \dots, X_{\alpha_k}).$$

Combining (5.9), (5.10), (5.11),

$$(5.12) \quad E(U_k') = 0,$$

$$(5.13) \quad E(M_k) = p_{(k)} \binom{N}{k}.$$

If we define for k, c such that $k \geq 2$ and $1 \leq c \leq k$,

$$\Phi_k^{(c)}(t_1, \dots, t_c) = p_i^{k-c} q_{(k)}, \text{ if } \frac{i-1}{n} < t_1, \dots, t_c \leq \frac{i}{n}, \quad i = 1, \dots, n,$$

$$= 0, \text{ otherwise,}$$

$$(5.14) \quad \Psi_k^{(c)}(t_1, \dots, t_c) = \Phi_k^{(c)}(t_1, \dots, t_c) - 1,$$

then it follows that

$$(5.15) \quad E\Psi_k'(t_1, t_c, X_{c+1}, \dots, X_k) = \Psi_k^{(c)}(t_1, \dots, t_c),$$

where expectation are as before with respect to the X 's.

It is readily verified that

$$(5.16) \quad E\Psi_k^{(c)}(X_1, \dots, X_c) = 0,$$

$$E \left[\Psi_k^{(c)}(X_1, \dots, X_c) \right]^2 = q_{(k)}^2 p_{(2k-c)} - 1.$$

By the same method as in section 3, putting $c_1 = \max(1, 2k - N)$, we have

$$(5.17) \quad D^2(U_k') = \binom{N}{k}^{-1} \sum_{c=c_1}^k \binom{k}{c} \binom{N-k}{k-c} \left[q_{(k)}^2 p_{(2k-c)} - 1 \right].$$

Under the limiting condition (3.9) it follows from (5.3), (5.5) and (5.17) that

$$(5.18) \quad D^2(U_k') \sim \frac{1}{n} \left\{ \sum_{c=1}^k \frac{(k!)^2}{c! [(k-c)!]^2} \cdot \frac{f_{2k-c}}{r^c f_k^2} - \frac{k^2}{r} \right\} = \frac{1}{n} \sigma_k'^2 \text{ (say),}$$

and from (5.13), (5.10) that

$$(5.19) \quad E(M_k/n) \sim r^k f_k/k!,$$

$$D^2(M_k/n) \sim \frac{1}{n} r^{2k} \sum_{c=1}^k \frac{1}{c! [(k-c)!]^2} \cdot \frac{f_{2k-c}}{r^c} - \frac{r^{2k-1} f_k^2}{[(k-1)!]^2} \left\} = \frac{1}{n} \sigma_k^2 \text{ (say).}$$

6. Asymptotic distribution in the general case and the consistency of M_k -test.

Theorem 2. *If the population d.f. $F(x)$ has the continuous (except for a finite number of points) density function such as $\int_0^1 f^m(x) dx$ ($m = 2, 3, \dots, 2k-1$) exist, and if $n \rightarrow \infty$, $N/n \rightarrow r$ (const), then U_k' and M_k/n are asymptotically normally distributed $(0, n^{-1} \sigma_k'^2)$ and $(r^k f_k/k!, n^{-1} \sigma_k^2)$, respectively.*

As the theorem can be proved in parallel with Theorem 1, we shall omit the proof.

Theorem 3. M_k -test is consistent against every alternative hypothesis whose d.f. satisfies the condition stated in Theorem 2.

Proof. From Theorem 1 the asymptotic distribution of M_k/n under H_0 is normal $(r^k/k!, n^{-1}\sigma_{0k}^2)$ and from Theorem 2 it is normal $(r^k f_k/k!, n^{-1}\sigma_k^2)$ under H . As the difference of means is constant and both variances are $O(n^{-1})$, we have only to show that

$$(6.1) \quad f_k = \int_0^1 f^k(x) dx > 1,$$

provided that $f(x)$ is not identically 1. For this purpose we shall prove more general

Lemma. If $\varphi(t)$ is convex function of $t \geq 0$ and satisfies $\varphi(1) = 1$, and if $f(x)$ is a density function which is continuous almost everywhere in the interval $(0, 1)$, then

$$(6.2) \quad \int_0^1 \varphi[f(x)] dx \geq 1.$$

where the equality holds if $f(x)$ is equal to 1 almost everywhere.

Remark. (6.1) is a special case of (6.2), where $\varphi(t) = t^k$.

Proof. As $y = \varphi(t)$ is convex, the graph in t, y -plane is above its tangent at $(1, 1)$:

$$y = \varphi'(1)(t-1) + 1,$$

except the point $(1, 1)$ itself. Thus

$$\varphi(t) \geq \varphi'(1)(t-1) + 1,$$

where equality holds if and only if $t = 1$. Hence

$$\int_0^1 \varphi[f(x)] dx \geq \int_0^1 [\varphi'(1)(f(x)-1) + 1] dx = 1,$$

where equality holds if and only if $f(x) = 1$ almost everywhere.

7. Unbiasedness of M_k -test. The author has proved in a recent paper a theorem concerning the unbiasedness in the test of goodness of fit. We shall first give some notations.

Denote by N_i the number of X 's which fall in the interval $((i-1)/n, i/n]$ and by k_i the observed value of $N_i, i = 1, \dots, n$. Let W be the set of all (k_1, \dots, k_n) . The subset S of W will be called symmetric if it is invariant under all permutations of n coordinates. Finally S will be called to satisfy the condition O provided that, if S contains the point (k_1, \dots, k_n) with $k_i \leq k_k - 2$, then S contains also $(k_1, \dots, k_i + 1, \dots, k_i - 1, \dots, k_n)$.

Then the above-mentioned theorem runs as follows :

If the acceptance region of the test for H_0 is a symmetric subset of W and satisfies the condition O , then the test is unbiased against all alternatives.

Now, as one of its applications we have

Theorem 4. *M_k -test is unbiased against all alternatives.*

Proof. Putting $\binom{j}{k} = 0$, if $j < k$, we have

$$(7.1) \quad M_k = \sum_{i=1}^n \binom{N_i}{k}.$$

The acceptance region R of the M_k -test is determined by the inequality

$$M_k \leq M_k^0,$$

where M_k^0 is a constant depending only on the level of significance of the test.

It is obvious that R is symmetric.

The condition O means that if $\sum_{i=1}^n \binom{k_i}{k} \leq M_k^0$ and $k_i \leq k_j - 2$, then $\sum_{\alpha \neq i, j} \binom{k_\alpha}{k} + \binom{k_i+1}{k} + \binom{k_j+1}{k} \leq M_k^0$. In order to verify this, we have merely to show that, if $k_1 \leq k_2 - 2$, then

$$\binom{k_1+1}{k} + \binom{k_2-1}{k} \leq \binom{k_1}{k} + \binom{k_2}{k}.$$

This follows at once from the relations

$$\binom{k_1+1}{k} - \binom{k_1}{k} = \binom{k_1}{k-1} \leq \binom{k_2-1}{k-1} = \binom{k_2}{k} - \binom{k_2-1}{k}.$$

8. Relation between M_k -test and David's test. We have divided the interval $(0, 1]$ into n small intervals $((i-1)/n, i/n]$, $i = 1, \dots, n$. Denote by n_k the number of small intervals which contain exactly k X 's, $k = 0, 1, \dots, N$. David's test [2] for H_0 uses the statistic n_0 (the present author denoted it by v in [1]), i.e., we shall accept H_0 when and only when n_0 is sufficiently small.

Now n_0 has a certain relationship with M_k as follows. We have

$$\begin{aligned} n_0 + n_1 + n_2 + n_3 + \dots + n_N &= n, \\ n_1 + 2n_2 + 3n_3 + \dots + Nn_N &= N, \\ \binom{2}{2}n_2 + \binom{3}{2}n_3 + \dots + \binom{N}{2}n_N &= M_2, \\ \binom{3}{3}n_3 + \dots + \binom{N}{3}n_N &= M_3, \\ &\dots\dots\dots \\ \binom{N}{N}n_N &= M_N. \end{aligned}$$

Therefore, putting $M_0 = n$, $M_1 = N$, we obtain the general relation

$$M_k = \sum_{i=k}^N \binom{i}{k} n_i, \quad k = 0, 1, \dots, N.$$

Hence

$$n_j = \sum_{k=j}^N (-1)^{k-j} \binom{k}{j} M_k, \quad j = 0, 1, \dots, N,$$

in particular

$$(8.1) \quad n_0 = \sum_{k=0}^N (-1)^k M_k.$$

From (8.1) and (3.2) we obtain

$$(8.2) \quad E_0(n_0) = n \left(1 - \frac{1}{n}\right)^N.$$

Putting

$$(8.3) \quad n_0^* = \left[n_0 - E_0(n_0) \right] / n.$$

we have from (2.7)

$$(8.4) \quad n_0^* = \sum_{k=2}^N (-1)^k n^{-k} \binom{N}{k} U_k,$$

$$(8.5) \quad E_0(n_0^*) = 0.$$

It follows by the same method for obtaining the variance of U_k in section 3 that

$$(8.6) \quad E_0(U_k U_l) = \binom{N}{l}^{-1} \sum_{c=c_2}^{c_3} \binom{k}{c} \binom{N-k}{l-c} (n^{c-1} - 1),$$

where $c_2 = \max(2, k+l-N)$ and $c_3 = \min(k, l)$.

(8.4) and (8.6) imply

$$(8.7) \quad E_0(U_k n_0^*) = (1 - 1/n)^{N-k+1} - (1 - 1/n)^N,$$

and under the limiting condition (3.9),

$$(8.8) \quad E_0(U_k n_0^*) \sim n^{-1} (k-1) e^{-r}.$$

In particular

$$(8.9) \quad E_0(U_2 n_0^*) \sim n^{-1} e^{-r}.$$

It is proved in the author's paper [1] that

$$(8.10) \quad D_0^2(n_0^*) \sim n^{-1} e^{-2r} (e^r - 1 - r).$$

(This evaluation can be also obtained easily from (8.4) and (8.7)). Combining (3.13), (8.9) and (8.10), we have the correlation coefficient of U_2 and n_0^*

$$\rho(U_2, n_0^*) \sim \frac{r}{\sqrt{2(e^r - 1 - r)}} = \rho \text{ (say)}.$$

Since M_2 and n_0 are linear functions of U_2 and n_0^* , respectively, this is at the same time the correlation coefficient of M_2 and n_0 , that is,

$$\rho(M_2, n_0) \sim \rho.$$

When r is small, ρ is approximately equal to 1. This is actually what one would expect since when r is small $M_k, k \geq 3$, are negligible in comparison with M_2 and from (8.1) n_0 becomes almost linear in M_2 only.

9. Consideration of the other limiting condition. Thus far we have concerned ourselves with the limiting condition (3.9), while in this section the assumption $N/n \rightarrow r$ is substituted by $N \rightarrow \infty$. (The author does not know the consequence when n is fixed and N alone tends to infinity, except the special case $k = 2$.)

First, let the null hypothesis H_0 be true. From (3.7) we have

$$(9.1) \quad D_0^2(U_k) \sim \frac{1}{n} \sum_{c=2}^k c! \binom{k}{c}^2 \left(\frac{n}{N}\right)^c.$$

Under the limiting condition

$$(9.2) \quad n \rightarrow \infty \quad \text{and} \quad N/n \rightarrow \infty$$

we have

$$(9.3) \quad D_0^2(U_k) \sim 2 \binom{k}{2}^2 n N^{-2}.$$

(3.2), (9.3) and (2.7) imply

$$(9.4) \quad \begin{aligned} E_0(M_k/n) &\sim (N/n)^k/k!, \\ D_0^2(M_k/n) &\sim N^{2k-2} n^{-(2k-1)}/2 [(k-2)!]^2. \end{aligned}$$

Corresponding to Theorem 1, we obtain

Theorem 1'. *Under H_0 and the limiting condition (9.2), U_k and M_k/n are asymptotically normally distributed with the means and variances (3.1), (9.3), (9.4).*

Proof is omitted since it is almost similar to that of Theorem 1.

To the contrary, under the limiting condition

$$N \rightarrow \infty \quad \text{and} \quad N/n \rightarrow 0,$$

the asymptotic distribution of U_k and M_k/n are not necessarily normal.

Finally, under the alternative hypothesis whose d.f. satisfies the condition in Theorem 2, (5.17) implies

$$(9.5) \quad D^2(U_k) \sim \sum_{c=1}^k c! \binom{k}{c}^2 N^{-c} (n^{c-1} f_{2k-1}^{c-2} f_k^{-2} - 1).$$

If (9.2) holds, then

$$(9.6) \quad D^2(U_k') \sim k^2(f_{2k-1}f_k^{-2}-1)N^{-1}.$$

From (5.13), (9.6) and (5.10) it follows that

$$(9.7) \quad \begin{aligned} E(M_k/n) &\sim f_k(N/n)^k/k!, \\ D^2(M_k/n) &\sim (f_{2k-1}-f_k^2)n^{-2k}N^{2k-1}/[(k-1)!]^2. \end{aligned}$$

Consequently we obtain corresponding to Theorems 2 and 3,

Theorem 2'. *Under the limiting condition (9.2) and the alternative hypothesis whose d.f. satisfies the condition stated in Theorem 2, U_k' and M_k/n are asymptotically normally distributed with means and variances (5.12), (9.6) and (9.7).*

Theorem 3'. *Under the limiting condition (9.2) M_k -test for H_0 is consistent against every alternative whose d.f. satisfies the condition stated in Theorem 2.*

10. Relation between M_2 - and χ^2 -tests. The statistic used in the χ^2 -test in the case of equal probability is

$$\chi^2 = \sum_{i=1}^N \frac{(N_i - N/n)^2}{N/n} = \frac{n}{N} \left(\sum_{i=1}^N N_i^2 - \frac{N^2}{n} \right),$$

where $N_i, i = 1, \dots, n$, are defined in section 7.

On the other hand, as the special case of (7.1), it holds

$$M_2 = \sum_{i=1}^N \binom{N_i}{2} = \frac{1}{2} \left(\sum_{i=1}^n N_i^2 - N \right).$$

Combining these two equations, we have

$$\chi^2 = 2nM_2/N + n - N,$$

or, by (2.7) and (5.10),

$$\begin{aligned} \chi^2 &= (N-1)U_2 + n - 1, \\ \chi^2 &= np_{(2)}(N-1)U_2' + N(np_{(2)}-1) + n(1-p_{(2)}). \end{aligned}$$

Hence by (3.13) and (5.19)

$$\begin{aligned} E_0(\chi^2) &= n-1 \sim n, \\ D_0^2(\chi^2) &= 2(n-1)(N-1)/N \sim 2n, \\ E(\chi^2) &= N(np_{(2)}-1) + n(1-p_{(2)}) \sim N(f_2-1) + n, \\ D^2(\chi^2) &= 2n^2(N-1)N^{-1} [2(N-2)(p_{(3)}-p_{(2)}^2) + p_{(2)}(1-p_{(2)})] \\ &\sim 4N(f_3-f_2^2) + 2nf_2. \end{aligned}$$

Finally as the corollaries of Theorems 1, 1' ; 2, 2' ; 3, 3' we obtain the following

Corollary 1. *Under H_0 and the limiting condition (3.9) or (9.2) χ^2 is asymptotically normally distributed with mean n and variance $2n$.*

Corollary 2. *Under the alternative whose d.f. satisfies the condition stated in Theorem 2 and under the limiting condition (3.9) or (9.2), χ^2 is asymptotically normally distributed with mean $N(f_2-1)+n$ and variance $4N(f_3-f_2^2)+2nf_2$, where f_2 and f_3 are defined in (5.4).*

Corollary 3. *Under the limiting condition (3.9) or (9.2) the χ^2 -test for H_0 is consistent against every alternative whose d.f. satisfies the condition in Theorem 2.*

The facts in Corollaries 1 and 2 were already stated in the paper of H. B. Mann and A. Wald [5] but were not proved there.

(Received July 10, 1952)

References

- [1] M. Okamoto, On a non-parametric test, Osaka Math. J. 4 (1952), 77-85.
- [2] F. N. David, Two combinatorial tests of whether a sample has come from a given population, Biometrika, 37 (1950), 97-110.
- [3] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. of Math. Stat. 19 (1948), 293-325.
- [4] M. Okamoto, Unbiasedness in the test of goodness of fit, Osaka Math. J. (In this volume.)
- [5] H. B. Mann and A. Wald, On the choice of the number of class intervals in the application of the chi-square test, Ann. of Math. Stat. 13 (1942), 306-317.