SOME BASIC THEOREMS FOR DEVELOPING TESTS OF FIT FOR THE CASE OF THE NON-PARAMETRIC PROBABILITY DISTRIBUTION FUNCTION, I

By Bradford F. Kimball

State Department of Public Service, New York, N. Y.

1. Summary. In developing tests of fit based upon a sample $O_n(x_i)$ in the case that the cumulative distribution function $F(X)$ of the universe of X's is not necessarily a function of a finite number of specific parameters—sometimes known as the non-parametric case—it has been pointed out by several writers that the "probability integral transformation" is a useful device (cf. [1]–[4]).

The author finds that a modification of this approach is more effective. This modification is to use a transformation of ordered sample values x_i from a random sample $O_n(x_i)$ based on successive *differences* of the cdf values $F(x_i)$.

A theorem is proved giving a simple formula for the expected values of the products of powers of these differences, where all differences from 1 to $n + 1$ are involved in a symmetrical manner.

The moment generating function of the test function defined as the sum of m squares of these successive differences is developed and the application of such a test function is briefly discussed.

2. Introduction. Let the sample values x_i be ordered so that

$$x_i \leqq x_{i+1}, \qquad (i = 1, 2, \cdots, n - 1). \tag{2.1}$$

Let F_r denote the value of the cdf $F(X)$ associated with the rth ordered sample value x_r. Thus

$$F_r = F(x_r). \tag{2.2}$$

Consider the following transformation of the ordered sample values x_i based upon the (hypothetically) known cumulative distribution function $F(X)$ which will be taken as a continuous function of X over its admissible range:

$$\begin{aligned} u_1 &= F_1, \\ u_r &= F_r - F_{r-1}, \qquad (r = 2, 3, \cdots, n) \\ u_{n+1} &= 1 - F_n. \end{aligned} \tag{2.3}$$

The restrictions on F_i are that

$$F_i \leqq F_{i+1}, \text{ and } 0 \leqq F_i \leqq 1. \tag{2.4}$$

The above transformation (2.3) translates these conditions into the *symmetrical* conditions

$$0 \leqq u_i, \text{ and } u_1 + u_2 + \cdots + u_n + u_{n+1} = 1. \tag{2.5}$$

A one-to-one correspondence between u_i and F_i exists if one of the u_i be omitted,—say u_β. With u_β omitted, the Jacobian of the transformation from F_i to u_i

has value unity. The probability density of the sample $O_n(x_i)$, with x_i ordered, is given by

$$P[O_n(x_i)]\, dO_n = n!\, dF_1\, dF_2 \cdots dF_n . \tag{2.6}$$

Hence with u_β omitted,

$$P[O_n(x_i)]\, dO_n = n!\, du_1\, du_2 \cdots du_{\beta-1}\, du_{\beta+1} \cdots du_{n+1} . \tag{2.7}$$

The sample space of the u_i with u_β omitted, is that portion of the $n + 1$ Euclidean space of all the u_i variables, bounded by the coordinate hyperplanes, which is on the projection of the hyperplane (2.5) upon the hyperplane $u_\beta = 0$. This is a region in the n-space of the u_i with u_β omitted, bounded by the coordinate hyperplanes and the hyperplane

$$u_1 + u_2 + \cdots + u_{\beta-1} + u_{\beta+1} + \cdots + u_n + u_{n+1} = 1. \tag{2.8}$$

Thus the formal integral of the pdf of the u_i over sample space is

$$n! \int n \int du_1 \cdots du_{\beta-1}\, du_{\beta+1} \cdots du_{n+1} = 1 \tag{2.9}$$

with $0 \leqq u_i$, and u_i bounded above by the hyperplane (2.8).

It is now clear that both the pdf and the sample space of the u_i (with u_β omitted) are symmetrical in the u_i. This fact leads to *complete* symmetry of the joint distribution function of any set of u_i, over $i = 1$ to $n + 1$ including u_β, relative to the u_i selected. Other interesting results are forthcoming.

3. Basic mathematical theorem. Using the techniques associated with the Beta function, the expectation of the products of powers u_i is found to be

$$\begin{aligned} &E[u_r^p \cdot u_s^q \cdot u_t^w \cdots] \\ &= \Gamma(n+1)\Gamma(p+1)\Gamma(q+1)\Gamma(w+1) \cdots / \Gamma(n+p+q+w+\cdots+1) \end{aligned} \tag{3.1}$$

where r, s, t, etc., are any set of different indices (for the present other than β) from the integers 1 to $n + 1$, and p, q, w, etc., are any real numbers greater than minus one. The relation (3.1) can further be generalized to the case where u_β may be included. This will be proved for the case $n = 2$, with p, q and w taken as integers. The generalization can be concluded from inspection. Thus with

$$u_3 = 1 - u_1 - u_2,$$

$$\begin{aligned} E[u_1^p \cdot u_2^q \cdot u_3^w] &= 2! \int_0^1 u_2^q\, du_2 \int_0^{1-u_2} u_1^p (1 - u_1 - u_2)^w\, du_1 \\ &= 2! \int_0^1 u_2^q (1 - u_2)^{p+w+1} \int_0^1 s^p (1 - s)^w\, ds \\ &= \frac{2!\,p!\,w!}{(p+w+1)!} \int_0^1 u_2^q (1 - u_2)^{p+w+1}\, du_2 = \frac{2!\,p!\,q!\,w!}{(p+q+w+2)!} . \end{aligned}$$

Hence the theorem:

THEOREM. *Given a random sample of n values of X from a universe with* cdf $F(X)$ *which is continuous over the range of X. With the sample values* x_i *ordered so that* $x_i \leqq x_{i+1}$ *define a set of* $n + 1$ *variables* u_i *as the successive differences of* $F(x_i)$ *by the relations* (2.3). *The expected value of the product of real powers greater than minus one of any or all of the* u_i, $(i = 1, 2, \cdots, n + 1)$, *is given by the relation* (3.1) *above (not subject to the omission of* u_β).

There are many interesting consequences of this theorem. Perhaps the most striking is the following:

COROLLARY 1. *Let a range* $\alpha(m, k)$ *for positive integer m be defined by*

$$\alpha(m, k) = F(x_{k+m}) - F(x_k) \tag{3.2}$$

with $k = 0, 1, 2, \cdots, n$, *and* $m \leqq n + 1 - k$

under the convention

$$F(x_0) = 0, \qquad F(x_{n+1}) = 1.$$

The probability distribution of $\alpha(m, k)$ *is independent of k and hence is the same as that of* $F(x_m)$.

Another interesting consequence (not new) is the following:

COROLLARY 2. *The correlation of* u_i *and* u_k, $i \neq k$, *is the same for all pairs* (i, k) *over the range of indices from* 1 *to* $n + 1$, *and has the value* $-1/n$.

Introducing the notation

$$[n + r]_r = (n + r)(n + r - 1) \cdots (n + 1), \tag{3.3}$$

the corollary follows from the relationships

$$E(u_i) = 1/(n + 1), \quad E(u_i^2) = 2/[n + 2]_2, \quad E(u_i u_k) = 1/[n + 2]_2 .$$

The fact that the correlation between any two frequency differences u_i and u_k is negative leads to the following more general relationship:

COROLLARY 3. *For any set of different indices i, j, k, etc., and for any positive numbers p, q, r, etc., the expectation of the product of the powers p, q, r,* $\cdots$ *of* u_i, u_j, $u_k \cdots$ *is less than the product of the expectations of the powers taken separately:*

$$E[u_i^p \cdot u_j^q \cdot u_k^r \cdots] < E(u_i^p) \cdot E(u_j^q) \cdot E(u_k^r) \cdots . \tag{3.4}$$

This follows from generalization of the relation

$$\frac{\Gamma(n + 1)\Gamma(p + 1)\Gamma(q + 1)\Gamma(r + 1)}{\Gamma(n + p + q + r + 1)} < \frac{[\Gamma(n + 1)]^3 \Gamma(p + 1)\Gamma(q + 1)\Gamma(r + 1)}{\Gamma(n + p + 1)\Gamma(n + q + 1)\Gamma(n + r + 1)} .$$

The above theorem suggests the possibility of test functions for fitted distributions, relative to a universe with a cdf which, since it is merely conditioned by a sufficient hypothesis for the theorem, may be of the non-parametric type.

A test function of the form

$$Y = \sum_m u_i^p, \qquad p \text{ real and positive} \tag{3.5}$$

might first come to mind. If $p = 1$, compensatory effects of deviations reduce the efficiency of the test function. One is thus led first to consider the test function (3.5) for the case $p = 2$.

4. The moments of the probability distribution of $y_m = \Sigma u_i^2$. We are first concerned with the problem of the determination of the moments of the function

$$y_m = \sum_m u_i^2 \tag{4.1}$$

where i ranges over any particular fixed set of m integers which for simplicity is usually taken as the first m.

One first recalls the fact that the result is independent of *which* m indices have been selected; and that the expected value of any combination of powers is independent of which specific subscripts of u_i are involved.

Since the u_i are correlated, principles of combinatory analysis are involved in determining the moments of y_m. One possible way of obtaining the moments is as follows:

Let v_r denote the rth moment of y_m about $y_m = 0$. Thus

$$E[(y_m)^r] = v_r = E[(\sum_m u_i^2)^r]. \tag{4.2}$$

Now in the expansion of $(\sum_m u_i^2)^r$, the sum of the power indices of each term is $2r$. Thus referring back to (3.1) and (3.3) it will be noted that the expected value of each such term will have the common factor

$$1/[n + 2r]_{2r}.$$

Consider a general term of the expansion of $(\sum_m u_i^2)^r$

$$C_{r_1 r_2 \cdots r_k} \cdot u_{i_1}^{2r_1} u_{i_2}^{r_2} \cdots u_{i_k}^{2r_k}, \qquad \text{with} \quad r_1 + r_2 + \cdots + r_k = r.$$

Clearly

$$E(u_{i_1}^{2r_1} u_{i_2}^{2r_2} \cdots u_{i_k}^{2r_k}) = 2r_1 \,!\, 2r_2 \,! \cdots 2r_k \,!/[n + 2r]_{2r}.$$

and the coefficient $C_{r_1 r_2 \cdots r_k}$ is the multinomial coefficient

$$C_{r_1 r_2 \cdots r_k} = \frac{r!}{r_1! r_2! \cdots r_k!}.$$

Now in the expansion of $(\sum_m u_i^2)^r$ group the terms which have the same set of k values of r_i, irrespective of which indices of u_i are involved. The number of such terms (since each involves k different indices) is $\binom{m}{k}$. If $r_1, r_2, \cdots, r_k$,

are all different each combination could be taken in $k!$ different ways. Thus with r's all different and fixed, the sum of all coefficients of terms with same combination of $2r_i$ powers (irrespective of variation of indices of the u_i) is

$$\binom{m}{k} k! \frac{r!}{r_1!r_2!\cdots r_k!}.$$

This would then constitute the total multiplier for

$$2r_1\,!\,2r_2\,!\cdots 2r_k\,!/[n+2r]_{2r}$$

for a given set of k r's which are all different.

If some of r's are repeated, let $k_1, k_2, \cdots, k_s$ denote the number of repetitions of each different r_i ($k_i \geqq 1$, and $k_1 + k_2 + \cdots + k_s = k$). Then each combination of the k r's corresponding to a set of k products could be taken in

$$k!/(k_1!\,k_2!\cdots k_s!)$$

different ways. Hence the lemma:

LEMMA 1. *Consider all admissible sets of k different subscripts of u_i and a fixed set of values of $r = r_1, r_2, \cdots, r_k$ where*

$$r_1 + r_2 + \cdots + r_k = r$$

such that s of these r's are different, and the number of repetitions in the set of r's is given by $k_1\,k_2\cdots k_s$ ($k_i \geqq 1$, and $k_1 + k_2 + \cdots + k_s = k$). The composite coefficient of the terms in v_r involving the factor

$$2r_1!\,2r_2!\cdots 2r_k!/[n+2r]_{2r}$$

is given by

$$\binom{m}{k} \frac{k!}{k_1!k_2!\cdots k_s!} \cdot \frac{r!}{r_1!r_2!\cdots r_k!}. \tag{4.3}$$

Examples of computation of v_r by means of the above lemma. The first order moment is given by

$$v_1 = E\Big(\sum_m u_i^2\Big) = m\,2!/[n+2]_2. \tag{4.4}$$

The second order moment is given by

$$v_2 = E\Big[\Big(\sum_m u_i^2\Big)^2\Big] = C_1E(u_i^4) + C_2E(u_i^2u_j^2),$$

and determining the values of C_i from Lemma 1,

$$v_2 = \left[m4! + \binom{m}{2}\binom{2!}{2!}\frac{2!}{1!1!}2!2!\right]\Big/[n+4]_4$$

or

$$v_2 = \left[m4! + 8\binom{m}{2}\right]\Big/[n+4]_4 = \left[m + \binom{m}{2}\frac{1}{3}\right]\Big/\binom{n+4}{4}. \tag{4.5}$$

Again for the third order moment,

$$v_3 = E[(\sum_m u_i^2)^3] = C_1E(u_i^6) + C_2E(u_i^2u_j^4) + C_3E(u_i^2u_j^2u_k^2),$$

and using Lemma 1,

$$= \left[m6! + \binom{m}{2}\frac{2!}{1!1!}\frac{3!}{1!2!}2!4! + \binom{m}{3}\frac{3!}{3!}\frac{3!}{1!1!1!}2!2!2!\right] \Big/ [n+6]_6$$

$$= \left[m6! + \binom{m}{2}2!3!4! + \binom{m}{3}2!2!2!3!\right] \Big/ [n+6]_6$$

or

$$v_3 = \left[m + \binom{m}{2}\frac{2}{5} + \binom{m}{3}\frac{1}{15}\right] \Big/ \binom{n+6}{6}. \tag{4.6}$$

Similarly writing the fourth moment in the form

$$v_4 = C_1E(u_i^8) + C_2E(u_i^6u_j^2) + C_3E(u_i^4u_j^4) + C_4E(u_i^2u_j^2u_k^4) + C_5E(u_i^2u_j^2u_k^2u_s^2)$$

and using Lemma 1 it reduces to

$$v_4 = \left[m + \binom{m}{2}\frac{2}{7} + \binom{m}{2}\frac{3}{35} + \binom{m}{3}\frac{3}{35} + \binom{m}{4}\frac{1}{105}\right] \Big/ \binom{n+8}{8}. \tag{4.7}$$

Higher order moments of the probability distribution function may be computed as desired.

An alternate method of computing the moments of the distribution of this test function is the following:

Consider a function $g_0(x)$ such that

$$\frac{d^r g_0(0)}{dx^r} = (2r)!, \qquad g_0(0) = 1. \tag{4.8}$$

Thus

$$E[u^{2r}] = [d^r g_0(0)/dx^r]/[n+2r]_{2r}. \tag{4.9}$$

From the principles of combinatory analysis of linear operators, it follows that[1]

$$E[(\sum_m u_i^2)^r] = \frac{d^r[g_0(x)]^m}{dx^r}\Big|_{x=0} /[n+2r]_{2r}. \tag{4.10}$$

Although this is an enlightening analytical form, actual computations seem to be simpler with the use of Lemma 1.

---

[1] One way of seeing this is to first think of the u_i as statistically independent. The numerators of the resulting terms would be the same as in (4.10). When the u_i are taken as dependent, by virtue of (3.1) the numerators will remain the same while all denominators will reduce to $[n+2r]_{2r}$.

Moment generating function. The moment generating function of the probability distribution of y_m can be written as

$$(4.11)\quad E(e^{ty}) = G_0(t, m) = 1 + \sum_{r=1}^{\infty} [d^r(g_0(x))^m/dx^r \mid_{x=0}]/[n + 2r]_{2r}\, t^r/r!$$

with

$$g_0(x) = 1 + 2!\,x + 4!\,x^2/2! + 6!\,x^3/3! + \cdots + (2r)!\,x^r/r! + \cdots$$

$$[n + 2r]_{2r} = (n + 2r)(n + 2r - 1)\cdots(n + 1).$$

Although $g_0(x)$ exists only as a formal power series, $G_0(t, m)$ is defined by (4.11) as a power series with positive coefficients, converging for all t.

5. Some comments on test function, $p = 2$. At the present time the study of the test function for $p = 2$ has not gone far enough to justify publication of results. One difficulty is that although its asymptotic distribution function appears to be normal, the convergence towards normalcy may be extremely slow in some cases.

Furthermore there are indications that the case $m = n + 1$ will give the most definitive results not only because the complete range of data is used, but also because errors of Type II would in general have a less erratic effect.

For the case $m = n + 1$ the mean, variance and third and fourth reduced moments (i.e. moments about the mean divided by corresponding power of σ) are:

Case $m = n + 1$.

$$E(y_{n+1}) = 2/(n + 2), \qquad \sigma^2 = 4n/[(n + 2)^2(n + 3)(n + 4)],$$

$$(5.1)\quad \alpha_3 = \mu_3/\sigma^3 = \frac{10n - 4}{(n + 5)(n + 6)}\sqrt{\frac{(n + 3)(n + 4)}{n}}$$

$$\alpha_4 = \left[\frac{n^5 + 101n^2 + 14n - 8}{(n + 5)(n + 6)(n + 7)(n + 8)}\right]\left[\frac{3(n + 3)(n + 4)}{n}\right]$$

$$\alpha_4 - 3 = \frac{6(41n^4 + 241n^3 + 118n^2 - 784n - 48)}{n(n + 5)(n + 6)(n + 7)(n + 8)}.$$

If data is not grouped the test may be applied as follows: Given a function $Q(X)$ which has been fitted to the cdf $F(X)$. From a random sample of size n with x_i ordered as in (2.1) compute the successive differences of $Q(x_i)$ to obtain the variables u_i^*. Then consider the sum of the squares

$$U^* = \sum_{n+1} u_i^{*2}.$$

If $Q(X)$ is a true representation of $F(X)$ the variation of U^* will follow that of y_{n+1}. Thus the expected value of U^*, its variance etc. will be *independent* of the fitted function $Q(X)$, which represents certain advantages over the χ^2 test.

The effect of Type II errors can be roughly analyzed as follows: In considering the effect of such errors the testing procedure must be criticized from the point of view that

$$Q(X) \neq F(X).$$

For $m = n + 1$ it still is true that

$$\Sigma u_i^* = 1$$

which tends to act as a control upon U^*. For example set

$$u_i^* = u_i + \chi_i .$$

Then from the above relation it follows that

$$\Sigma\chi_i = 0. \tag{5.2}$$

Write U^* as

$$\begin{aligned} U^* &= \Sigma u_i^2 + \Sigma\chi_i^2 + 2\Sigma u_i\chi_i \\ &= \Sigma u_i^2 + \Sigma\chi_i^2 + (2\Sigma\chi_i)/(n+1) + 2\Sigma\chi_i\delta(u_i) \end{aligned} \tag{5.3}$$

where $\delta(u_i)$ denotes the variation of the true frequency differences from their expected value $1/(n + 1)$.

The variation $\delta(u_i)$ will be to a considerable degree independent of χ_i. Thus the term $\Sigma\chi_i^2$ will in general tend to be larger than the last term on the right. The third term on the right will be zero by virtue of (5.2), and hence U^* will tend to be larger than y_{n+1}. A similar effect upon the sampling variance of U^* can be noted. Hence an interval of rejection

$$U^* \geqq A, \qquad P[y_{n+1} \leqq A] = \alpha = \text{confidence level},$$

is pointed to.

On the other hand if $m < n + 1$ the condition (5.2) no longer holds, the term $(2\ \Sigma\chi_i)/(n + 1)$ of (5.3) will not be zero and in many cases would dominate the other two error terms. Thus it is easily conceivable that one may have in the case $m < n + 1$

$$U_m^* < y_m$$

even when the discrepancies χ_i are large. Hence in the case $m < n + 1$ choice of confidence interval will require considerable care (see [1]).

Although the distribution of y_{n+1} for small n is decidedly non-normal, if the test function is replaced by

$$r_{n+1} = (\Sigma[u_i - 1/(n+1)]^2)^{\frac{1}{2}} \tag{5.4}$$

it will be found that the probability density function takes on the normal character quite rapidly with increasing n. Indeed the author has found that a computed approximation to the probability density function of r_{n+1} with $n = 4$ is decidedly normal in character.

REFERENCES

[1] J. NEYMAN, "Smooth test for goodness of fit," *Skand. Aktuar. Tidskn.* (1937) p. 149.
[2] E. S. PEARSON, "The probability integral transformation for testing goodness of fit and combining independent tests of significance," *Biometrika*, Vol. 30 (1938), pp. 134–148.
[3] E. J. GUMBEL, "Simple tests for given hypothesis," *Biometrika*, Vol. 32 (1942), pp. 317–333.
[4] H. SCHEFFÉ AND J. W. TUKEY, "Non-parametric estimation, I. Validation of order statistics," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187–192.