The Annals of Statistics
1977, Vol. 5, No. 1, 224–228

A NEW FORMULA FOR k-STATISTICS[1]

By I. J. Good

Virginia Polytechnic Institute and State University

A new formula for cumulants was given by Good (1975). A short proof is now given and the result is used to obtain a new formula for k-statistics. This formula can be used both for deriving the expressions of k-statistics in terms of power sums of the observations, and for checking and locating errors in formulae that are already in the literature.

1. Introduction. A new formula for cumulants was given by Good (1975) where, however, it was overlooked that the result could be used to prove a similar formula for k-statistics. We now rectify this and indicate how the result provides a new method of calculation of the k-statistics. [The method led to the detection of an incorrect sign in a formula of Zia ud-Din (1954).] In addition we give an elegant proof of the formula for cumulants since the proof given by Good (1975) was difficult, though the lemma on permanents is of independent interest.

As usual, cumulants

$$\kappa_{\mathbf{r}} = \kappa_{r_1, r_2, \cdots, r_n}$$

of an n-dimensional random vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)'$, are defined by the identity

$$\sum_{\mathbf{r}} \kappa_{\mathbf{r}} \mathbf{x}^{\mathbf{r}}/\mathbf{r}! = \log E \exp(\theta_1 x_1 + \cdots + \theta_n x_n)$$

where E denotes expectation, $x_1, x_2, \cdots, x_n$ are purely imaginary variables,

$$\mathbf{x}^r = x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}, \qquad \mathbf{r}! = r_1! \cdots r_n!$$

and $r_1, r_2, \cdots, r_n$ run through all nonnegative integers. The identity is valid when the characteristic function is analytic in the neighborhood of the origin. We assume this throughout.

2. The theorem for cumulants. Theorem 1 of Good (1975) expressed the cumulant $\kappa_{\mathbf{r}}$ as a *moment* of another random vector, in fact

$$\kappa_{\mathbf{r}} = R^{-1} E \prod_{\nu=1}^{n} [\omega\theta_\nu^{(1)} + \omega^2\theta_\nu^{(2)} + \cdots + \omega^R \theta_\nu^{(R)}]^{r_\nu} \tag{2.1}$$

where $R = |\mathbf{r}| = r_1 + r_2 + \cdots + r_n$, ω is any primitive Rth root of unity, $\theta_\nu^{(\rho)}$ is the νth component of the vector $\boldsymbol{\theta}^{(\rho)}$ ($\nu = 1, 2, \cdots, n$; $\rho = 1, 2, \cdots, R$), and $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(R)}$ are i.i.d. random vectors each with the same distribution as $\boldsymbol{\theta}$.

As mentioned in Good (1975), (2.1) can be used for the Monte Carlo evaluation

Received April 1975; revised May 1976.

[1] This work was supported in part by the Grant No. R01 GM 18770 of the Department of Health, Education and Welfare, National Institutes of Health.

AMS 1970 *subject classifications.* Primary 62D05; Secondary 62H10.

Key words and phrases. k-statistics, cumulants, permutations, partitions, permanents, roots of unity.

of cumulants, and for the expression of cumulants in terms of ordinary moments by means of programmed algebra.

We now give the new proof of (2.1).

As is well known, for any random vector $\mathbf{Y}$ (having requisite moments), the moment $\mu_{\mathbf{r}}'(\mathbf{Y})$ can be expressed as a polynomial in the cumulants of $\mathbf{Y}$ of which one term is $\kappa_{\mathbf{r}}(\mathbf{Y})$ and the other terms involve cumulants $\kappa_{\mathbf{q}}(\mathbf{Y})$ where $|\mathbf{q}| < |\mathbf{r}|$. Therefore

$$(2.2)\qquad \mu_{\mathbf{r}}'(\mathbf{Y}) = \kappa_{\mathbf{r}}(\mathbf{Y}) \qquad \text{if} \quad \kappa_{\mathbf{q}}(\mathbf{Y}) = 0 \quad \text{whenever} \quad |\mathbf{q}| < |\mathbf{r}| .$$

Let

$$\boldsymbol{\theta}^* = \textstyle\sum_{\rho=1}^{R} \omega^\rho \boldsymbol{\theta}^{(\rho)} ,$$

where the $\boldsymbol{\theta}^{(\rho)}$ each have the same cumulants $\kappa_{\mathbf{r}}$. Then, by the additive property of cumulants, we have

$$\begin{aligned}
\kappa_{\mathbf{q}}(\boldsymbol{\theta}^*) &= \textstyle\sum_{\rho=1}^{R} \kappa_{\mathbf{q}}(\omega^\rho \boldsymbol{\theta}^{(\rho)}) \\
&= \textstyle\sum_{\rho=1}^{R} \omega^{\rho|\mathbf{q}|}\kappa_{\mathbf{q}}(\boldsymbol{\theta}^{(\rho)}) \\
&= R\kappa_{\mathbf{q}} \qquad \text{if} \quad |\mathbf{q}| \quad \text{is a multiple of} \quad R \\
&= 0 \qquad \text{if} \quad |\mathbf{q}| \quad \text{is not a multiple of} \quad R .
\end{aligned}$$

Therefore, by applying (2.2) with $\mathbf{Y} = \boldsymbol{\theta}^*$, we see that $\kappa_{\mathbf{r}}(\boldsymbol{\theta}^*) = \mu_{\mathbf{r}}'(\boldsymbol{\theta}^*)$, and that

$$|\mathbf{r}|\kappa_{\mathbf{r}} = \kappa_{\mathbf{r}}(\boldsymbol{\theta}^*) = \mu_{\mathbf{r}}'(\boldsymbol{\theta}^*) ,$$

so that (2.1) is established.

It is curious that the characteristic function of $\boldsymbol{\theta}^*$ is formally $\prod_{\rho=1}^{R} \phi(\omega^\rho \mathbf{t})$ where $\phi(\mathbf{t})$ is the characteristic function of $\boldsymbol{\theta}$; and therefore

$$(2.3)\qquad \kappa_{\mathbf{r}} = \frac{\mathbf{r}!}{R} \mathscr{C}(\mathbf{u}^{\mathbf{r}}) \textstyle\prod_{\rho=1}^{R} \phi(-i\omega^\rho \mathbf{u})$$

where $\mathscr{C}(\mathbf{u}^{\mathbf{r}})\{\cdots\}$ denotes the coefficient of $\mathbf{u}^{\mathbf{r}}$ in $\{\cdots\}$.

3. A formula for k-statistics. Let $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(N)}$ be N independent observations of a vector random variable $\boldsymbol{\theta}$, where $N \geqq R$. [The same notation is used for these observations as for the random variables.] The k-statistic $k_{\mathbf{r}}$ is that symmetric function of these vectors, and which is a polynomial in their components, whose expectation (when the vectors are regarded as random variables) is $\kappa_{\mathbf{r}}$: in other words $k_{\mathbf{r}}$ is an unbiased estimate of $\kappa_{\mathbf{r}}$. It is known that $k_{\mathbf{r}}$ is unique (for each $\mathbf{r}$); see, for example, Kendall and Stuart (1963), page 278. Consider the average of the right side of (2.1), without the expectation sign, averaged over all possible ordered sequences of R of the observations, of which there are $N^{[R]} = N(N-1)\cdots(N-R+1)$. This statistic has $\kappa_{\mathbf{r}}$ as its expectation, and is a symmetric function of the observations. Therefore

$$(3.1)\qquad k_{\mathbf{r}} = \frac{1}{RN^{[R]}} \textstyle\sum \prod_{\nu=1}^{n} \{\omega\theta_\nu^{(j_1)} + \cdots + \omega^R\theta_\nu^{(j_R)}\}^{r_\nu}$$

is an unbiased estimate of $\kappa_{\mathbf{r}}$, where the summation is over all $N^{[R]}$ possible sequences $(j_1, j_2, \cdots, j_R)$ where $j_1, j_2, \cdots, j_R$ are R distinct numbers selected from the set $1, 2, \cdots, N$. Thus (3.1) is a formula for Fisher's k-statistic. Since a *cyclic* permutation of $(j_1, j_2, \cdots, j_R)$ leaves the product unchanged, we can sum over only those sequences in which j_1 is the smallest of the R numbers $j_1, \cdots, j_R$, provided that the factor R in the denominator is omitted.

4. Discussion. The univariate form of (3.1) is

$$k_r = \frac{1}{rN^{[r]}} \sum \{\omega\theta^{(j_1)} + \cdots + \omega^r\theta^{(j_r)}\}^r \tag{4.1}$$

and this confirms that we can subtract any constant from each of $\theta^{(1)}, \cdots, \theta^{(N)}$ without affecting k_r (if $r > 1$). In particular we can subtract the mean $\bar{\theta}$ from all the observations. When k_r is expressed as a polynomial in the symmetric power sums $s_1 = \theta^{(1)} + \cdots + \theta^{(N)}$, $s_2 = (\theta^{(1)})^2 + \cdots + (\theta^{(N)})^2$, etc., it is advantageous to force $s_1 = 0$, for this greatly simplifies the standard formulae for the k-statistics, the calculations become better conditioned, and the number of terms in the standard formula for k_r is reduced from $p(r)$ to $q(r)$, where $p(r)$ is the number of partitions of r and $q(r)$ is the number of partitions when no part is 1. We have $q(r) = p(r) - p(r-1)$ because the generating function for $q(r)$ is

$$[(1-x^2)(1-x^3)(1-x^4)\cdots]^{-1}$$

which is $1 - x$ times that for $p(r)$. [Tait (1882/85) computes the q's less simply.] From Euler's identity (Hardy and Wright (1938), page 282), or as the case $a = -x$ of a series for $\prod (1 + ax^n)$ due to Sylvester (1882, page 282), we have

$$\begin{aligned}(1-x^2)(1-x^3)(1-x^4)\cdots \\ = \textstyle\sum_{m=0}^{\infty} (-1)^m x^{\frac{1}{2}(3m^2+m)}(1 + x + x^2 + \cdots + x^{2m}),\end{aligned} \tag{4.2}$$

a fact that can be used to verify a table of values of $q(r)$. Similarly, in say the trivariate case, if $p(\mathbf{r})$ denotes the number of tripartite partitions of $\mathbf{r}$, and $q(\mathbf{r})$ the number having no part $\boldsymbol{\rho}$ with $|\boldsymbol{\rho}| = 1$, we have

$$\sum q(\mathbf{r})\mathbf{x}^{\mathbf{r}} = (1-x_1)(1-x_2)(1-x_3) \sum p(\mathbf{r})\mathbf{x}^{\mathbf{r}}, \tag{4.3}$$

from which $q(\mathbf{r})$ can be expressed as a linear combination of eight p's.

When N is large, it is impracticable to use (3.1) directly for numerical calculation. But it is known that $N^{[r]}k_r$ is a polynomial in $s_2, s_3, \cdots$ with integral coefficients, after forcing $s_1 = 0$; for example,

$$N^{[7]}k_7 = a(000001)s_7 + a(1001)s_2s_5 + a(011)s_3s_4 + a(21)s_2^2s_3$$

where $a(m_2, m_3, \cdots)$ denotes the coefficient of $s_2^{m_2}s_3^{m_3}\cdots$ and is a polynomial in N, divisible by N and of degree $\sum_{j=2}^{\infty} \min(0, m_j - 1)$. By fixing N and giving $x_1, x_2, \cdots, x_N$ special values (with $s_1 = 0$) in four different ways, we could obtain four linear equations from which the coefficients $a(000001)$, etc. could be found numerically, when the equations for the coefficients are linearly independent. Knowing that the coefficients are integers makes this process easier. If the coefficients were thereby obtained for several values of N, they could be expressed as explicit polynomials in N, again by solving sets of simultaneous linear equations.

Instead of carrying out these calculations completely, we may be content to verify or question published formulae for k_r by taking $N = r$ and then taking special values for $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(r)}$. For example, by taking $\theta^{(1)} = -(r-1)$, $\theta^{(2)} = \theta^{(3)} = \cdots = \theta^{(r)} = 1$, we find easily that

$$k_r = -(-r)^{r-1}. \tag{4.4}$$

I have used this special formula to check all the general formulae for k_r given, for example, by Kendall and Stuart ((1963), pages 280–281), for $r = 1(1)8$, and by Zia ud-Din (1954) for $r = 9$. These are good checks when $s_1 = 0$. This check revealed that when $r = 10$ the sign of $65N^3$ in the coefficient of $-37800 s_3^2 s_2^2$ in Zia ud-Din (1954) was printed as a plus when it should be a minus. Although this sign is only one of thousands of symbols the misprint multiples the answer by about -5000. The present check says nothing about the (less important) terms that involve s_1.

As already mentioned, the calculation of (4.1) when $N = r$, for arbitrary values of $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(r)}$, requires only $(r-1)!$ permutations of the r "observations." In fact only $(r-1)!/2$ permutations need be used, because a "reflection" ($\theta^{(1)}, \cdots, \theta^{(r-1)}$ replaced by $\theta^{(r-1)}, \cdots, \theta^{(1)}$ respectively) replaces $\sum \omega^s \theta^{(s)}$ by its complex conjugate. The permutations can be elegantly generated to take advantage of this complex conjugacy property, in the following manner. [See also Ord–Smith (1971).] We give a recipe for generating the mth permutation ($m = 0, 1, 2, \cdots$). First write m in the form $\sum_{s=1}^{t-1} a_s s!$ ($a_s = 0, 1, \cdots, s$). Then operate on the string $[A_1 A_2 A_3 \cdots]$ with the product of cyclic permutations $(s+1-a_s, s+2-a_s, \cdots, s, s+1)(s = 1, 2, \cdots, t-1)$, where the cycles are applied in the order $s = 1, 2, \cdots, t-1$. The cycles refer to the *positions* of the objects in the string, which is more convenient than when they refer to the *names* of the objects. For example, if $m = 11$ we have $a_4 = 0$, $a_3 = 1$, $a_2 = 2$, $a_1 = 1$, and $(3, 4)(1, 2, 3)(1, 2)[A_1 A_2 A_3 A_4] = (3, 4)(1, 2, 3)[A_2 A_1 A_3 A_4] = (3, 4)[A_3 A_2 A_1 A_4] = [A_3 A_2 A_4 A_1]$. This method of generating permutations is convenient for a computer program and happens to give a neat pattern of permutations that is easy to write out by hand.

A program to carry out this calculation was written by Mr. Byron Lewis. It was run (for $N = r$) with $\theta^{(1)} = 0$, $\theta^{(2)} = 1$, $\theta^{(3)} = 2, \cdots, \theta^{(r-1)} = r-2$, $\theta^{(r)} = -(r-1)(r-2)/2$; $r = 4(1)8$; and obtained further checks of the published formulae for k_r with $s_1 = 0$.

For multivariate k-statistics there are again cases that can be usefully worked out by hand. For example, in the bivariate case, $n = 2$, we can take $N = R$, and "observations"

$$(c-R, c), (c, c-R), (c, c), (c, c), \cdots, (c, c)$$

for which

$$s_{pq} = (c-R)^p c^q + c^p (c-R)^q + (R-2)c^{p+q}, \tag{4.5}$$

where s_{pq} denotes as usual the sum of $a^p b^q$ over all N observations (a, b). Then formula (3.1) shows that

$$k_{\mathbf{r}} = (-R)^{R-1}/(R - 1)(r_2 \neq 0, r_2 \neq R) . \tag{4.6}$$

Using (4.5) and (4.6) we can check any formula that expresses $k_{\mathbf{r}}$ in terms of the s_{pq}, and this check was applied to the bivariate formulae for $|\mathbf{r}| \leqq 4$, given, for example, by Kendall and Stuart (1963), page 308.

REFERENCES

FISHER, R. A. (1928). Moments and product moments of sampling distributions. *Proc. London Math. Soc. Ser. 2* **30** 199–238. (Also in Fisher's *Contributions to Mathematical Statistics*, paper no. 20. Wiley, New York.)

GOOD, I. J. (1975). A new formula for cumulants. *Math. Proc. Cambridge Philos. Soc.* **77** 333–337.

HARDY, G. H. and WRIGHT, E. M. (1938). *An Introduction to the Theory of Numbers.* Clarendon Press, Oxford.

KENDALL, M. G. and STUART, ALAN (1963). *The Advanced Theory of Statistics* **1**, 2nd ed. Griffin, London.

ORD-SMITH, R. J. (1970); (1971). Generation of permutation sequences. *Comput. J.* **13** 152–155; **14** 136–139.

SYLVESTER, J. J. (1882). A constructive theory of partitions, arranged in three acts, an interact and an exodian. *Amer. J. Math.* **5** 251–330.

TAIT, P. G. (1882/85). Note on a problem in partitions. Appendix on knots. *Edin. Roy. Soc. Trans.* **32** 340–342.

ZIA UD-DIN, M. (1954). Expression of the k-statistics k_9 and k_{10} in terms of power sums and sample moments. *Ann. Math. Statist.* **25** 800–803.

DEPARTMENT OF STATISTICS
VIRGINIA POLYTECHNIC INSTITUTE AND
STATE UNIVERSITY
BLACKSBURGH, VIRGINIA 24061