Biometrika (1965), **52**, 3 *and* 4, *p.* 309
Printed in Great Britain

The goodness-of-fit statistic V_N: distribution and significance points†

BY M. A. STEPHENS
McGill University, Montreal

1. INTRODUCTION AND SUMMARY

1·1. Kuiper (1960) has proposed V_N, an adaptation of the Kolmogorov statistic, to test the null hypothesis that a random sample of size N comes from a population with given continuous distribution function $F(x)$. If the sample distribution function is $F_N(x)$, V_N is defined by

$$V_N = \sup_{-\infty<x<\infty} (F_N(x) - F(x)) - \inf_{-\infty<x<\infty} (F_N(x) - F(x)). \tag{1}$$

Kuiper showed that:

(*a*) the distribution of V_N, on the null hypothesis, is independent of $F(x)$;

(*b*) if the observations are points on a circle, the value of V_N obtained from (1) does not depend on the choice of origin for measuring x.

The Kolmogorov statistic K_N does not possess property (*b*). V_N is therefore very suitable for observations on a circle: another statistic designed for use in this situation, and similarly an adaptation of an older statistic, W_N^2, is U_N^2, introduced by Watson (1961, 1962). Both V_N and U_N^2 may also be used for observations on a line. The definitions of K_N, W_N^2 and U_N^2 will be found in § 5.

1·2. Throughout this paper, the distribution of V_N will refer to its distribution on the null hypothesis. Kuiper gave the distribution, for large N, by showing that

$$\Pr(\sqrt{N}V_N \geqslant z) = \sum_{m=1}^{\infty} 2(4m^2z^2 - 1)e^{-2m^2z^2} - \frac{8z}{3\sqrt{N}} \sum_{m=1}^{\infty} m^2(4m^2z^2 - 3)e^{-2m^2z^2} + O\left(\frac{1}{N}\right). \tag{2}$$

We give below the exact distribution of V_N, in both the upper and the lower tails. These results, together with (2), make it possible to calculate significance points to make the goodness-of-fit test available for a complete range of values of N. The test, with the tables, is described in § 2. The two theorems concerning the distribution, preceded by the relevant lemmas, are in §§ 3 and 4. In § 5 are collected together a number of interesting results, primarily concerning the relations between the asymptotic distributions of $\sqrt{N}V_N$, K_N, W_N^2 and U_N^2.

1·3. In practical applications, one will be interested in the relative performances of V_N and U_N^2 for circular observations. Tables for the test based on U_N^2 are in Stephens (1963, 1964). For observations on a line, they may also be compared, both with each other, and with K_N and W_N^2; for some alternatives, they might be expected to give greater power. For a preliminary study along these lines, see Pearson (1963).

2. THE GOODNESS-OF-FIT TEST BASED ON V_N

2·1. The test requires the steps set out below. A figure, with examples showing how V_N and the other test statistics are calculated, is in Pearson (1963). If the N given observations

† Research supported in part by the U.S. Office of Naval Research.

are observations on a circle, any point on the circumference may serve as origin in steps (2) and (3) below.

(*a*) Suppose the observations, in ascending order, are $x_1, x_2, \ldots, x_N$.

(*b*) Draw a figure, showing $F(x)$ and the sample distribution function $F_N(x)$, namely, the step function defined by

$$F_N(x) = 0, \quad x < x_1,$$
$$F_N(x) = i/N, \quad x_i \leqslant x < x_{i+1}, \quad 1 \leqslant i \leqslant N-1,$$
$$F_N(x) = 1, \quad x_N \leqslant x.$$

(*c*) If A is the maximum value of $(F_N(x) - F(x))$, and B the maximum value of

$$(F(x) - F_N(x)), \quad \text{then} \quad V_N = A + B.$$

Table 1. *Upper tail percentage points for* V_N *and (in parentheses) for* $\sqrt{N}\,V_N$

	Significance levels as percentages						
N	15·0	10·0	5·0	2·5	1·0	0·5	0·1
2	0·9250	0·9500	0·9750	0·9875	0·9950	0·9975	0·9995
3	·776	·817	·871	·909	·942	·959	·982
4	·683	·714	·768	·816	·864	·892	·937
5	·619	·652	·700	·740	·789	·822	·881
6	0·571	0·601	0·646	0·687	0·732	0·762	0·824
7	·532	·561	·604	·641	·686	·716	·775
8	·501	·528	·569	·605	·647	·676	·734
9	·475	·500	·539	·574	·614	·642	·699
10	·452†	·477†	·514	·547	·586	·613	·668
11	0·432	0·456	0·492	0·524	0·562	0·587	0·641
12	·415	·437	·471	·503	·540	·565	·617
14	·386	·408	·439	·469	·503	·527	·576
16	·363	·384	·414	·441	·473	·496	·542
18	·343	·363	·392	·417	·448	·470	—
20	0·326	0·346	0·372	0·397	0·427	0·447	—
	(1·460)	(0·546)	(1·665)	(1·776)	(1·908)	(1·998)	—
30	·269	·285	·307	·328	·352	·369	—
	(1·476)	(1·562)	(1·684)	(1·797)	(1·930)	(2·022)	—
40	·235	·248	·268	·286	·307	·322	—
	(1·484)	(1·571)	(1·695)	(1·808)	(1·941)	(2·034)	—
50	·211	·223	·241	·256	·276	·289	—
	(1·490)	(1·576)	(1·701)	(1·815)	(1·949)	(2·042)	—
60	0·193	0·204	0·220	0·235	0·252	0·264	—
	(1·494)	(1·582)	(1·705)	1·820	(1·955)	(2·047)	—
70	·179	·189	·204	·218	·234	·245	—
	(1·497)	(1·585)	(1·707)	(1·824)	(1·959)	(2·051)	—
80	·168	·178	·191	·204	·219	·230	—
	(1·500)	(1·588)	(1·711)	(1·826)	(1·962)	(2·055)	—
100	·151	·159	·172	·183	·197	·206	—
	1·505	(1·590)	(1·716)	(1·831)	(1·967)	(2·060)	—
∞	(1·537)	(1·620)	(1·747)	(1·862)	(2·001)	(2·098)	(2·303)

To assist interpretation for high values of N, percentage points for $\sqrt{N}\,V_N$ are given in parentheses. The horizontal line in each column is explained in §§ 2·4 and 4·13.

† These two percentage points have been found by making a special calculation of $C_r^*(z, d)$ for the stage $0{\cdot}4 < z \leqslant 0{\cdot}5$.

(*d*) Enter Table 1 at the appropriate row for N; if V_N exceeds an entry, the null hypothesis is rejected at the corresponding significance level α. The entries in parentheses are used for interpolation (see § 4·13).

2·2. In the above, we assume that the test is being used against the usual alternative of a poor fit. If it is necessary to test against too good a fit, Table 2 is used, the null hypothesis being rejected at significance level α if V_N is less than the corresponding entry. For an example of this situation, see Pearson (1963).

2·3. Steps (*b*) and (*c*) above may be replaced, if convenient, by the following:

(*b'*) Let $y_i = F(x_i)$ $(i = 1, 2, \ldots, N)$.

(*c'*) If A is the maximum value of $(i/N) - y_i$ for all i, and B the maximum value of $y_i - (i/N)$ for all i, then $V_N = A + B$.

Table 2. *Lower tail percentage points for V_N and (in parentheses) for $\sqrt{N}\,V_N$*

	Significance levels as percentages						
N	15·0	10·0	5·0	2·5	1·0	0·5	0·1
2	0·575	0·550	0·525	0·513	0·505	0·503	0·501
3	·491	·462	·425	·398	·374	·362	·346
4	·434	·411	·378	·351	·325	·309	·285
5	·388	·370	·343	·320	·296	·280	·254
6	0·356	0·337	0·314	0·295	0·273	0·260	0·234
7	·333	·315	·290	·273	·255	·243	·219
8	·313	·296	·274	·256	·239	·228	·207
9	·296	·281	·259	·243	·225	·215	·196
10	·282	·267	·247	·231	·214	·204	·187
11	0·270	0·256	0·237	0·221	0·205	0·195	0·178
12	·259	·245	·227	·213	·197	·188	·170
14	·241	·228	·211	·198	·184	·175	·159
16	·227	·215	·198	·186	·173	·165	·149
18	·215	·203	·188	·176	·163	·156	·141
20	0·204	0·193	0·179	0·168	0·156	0·148	0·135
	(0·913)	(0·864)	(0·798)	(0·749)	(0·696)	(0·662)	(0·601)
30	·169	·160	·147	·138	·128	·122	—
	(0·923)	(·874)	(·807)	(·757)	(·703)	(·669)	—
40	·147	·139	·129	·121	·112	·107	—
	(·929)	(·880)	(·813)	(·763)	(·709)	(·675)	—
50	·141	·125	·116	·108	·101	·095	—
	(·934)	(·885)	(·817)	(·766)	(·714)	(·681)	—
60	0·121	0·115	0·106	0·099	0·093	0·088	—
	(·937)	(·889)	(·821)	(·769)	(·717)	(·684)	—
70	·112	·107	·098	·092	·086	·082	—
	(·940)	(·891)	(·823)	(·772)	(·720)	(·687)	—
80	·105	·100	·092	·086	·081	·077	—
	(·942)	(·894)	(·826)	(·773)	(·722)	(·689)	—
100	·095	·090	·083	·078	·073	·069	—
	(·945)	(·897)	(·829)	(·777)	(·725)	(·692)	—
∞	(·973)	(·9275)	(·8613)	(·8095)	(·7550)	(·7212)	(·6590)

To assist interpolation for high values of N, percentage points for $\sqrt{N}\,V_N$ are given in parentheses. The horizontal line in each column is explained in §§ 2·4 and 4·13.

2·4. In the tables, in each column, the values below the horizontal line (except those for $N = \infty$) are estimates. The construction of the tables is described in § 4·13, together with comments on their use.

3. THREE LEMMAS

3·1. When N observations are in *ascending* order we say they are in *rank order*, or are *ranked*.

LEMMA 1. *Suppose $x_1, x_2, \ldots, x_r$ are r independent observations from the uniform distribution between 0 and 1, denoted by $U(0, 1)$. Let z, d be positive, such that $z + (r-1)\,d \leqslant 1$. The probability that*

(*a*) $0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_r \leqslant 1$, *and also that*

(*b*) $0 \leqslant x_i \leqslant z + (i-1)\,d$, *for* $i = 1, 2, \ldots, r$, *is given by*

$$A_r(z, d) = \frac{z}{r!}(z + rd)^{r-1}. \tag{3}$$

Throughout this paper, d will be constant, and we write $A_r(z, d) \equiv A_r(z)$. $A_0(z)$ is defined equal to unity. We note that

$$A_r(d) = \frac{d^r(r+1)^r}{(r+1)!}.$$

The result (3) is quoted, with further comment on its proof, in Birnbaum & Tingey (1951).

Lemma 1 gives the probability for the special case where the order of selection of the observations is the same as the rank order. With r independent observations, there will be $r!$ equi-probable original orders of selection which would give the same rank order. Thus we have the following:

COROLLARY. *The probability that, after being placed in ascending order, r independent observations from $U(0, 1)$ will satisfy the conditions of Lemma* 1, *is* $r!\,A_r(a, d)$.

3·2. *Introduction to Lemma* 2. Suppose $x_1, x_2, \ldots, x_r$ are as above and further that

(*a*) $0 \leqslant x_1 \leqslant x_2 \ldots \leqslant x_r \leqslant 1$, and

(*b*) $(i-1)\,d \leqslant x_i \leqslant z + (i-1)\,d$, where $z, d > 0$, such that $z + (r-1)\,d \leqslant 1$. We shall need the probability that both (*a*) and (*b*) are satisfied together; this will be called $C_r^*(z, d)$. In Lemma 2, an expression is derived for $C_r^*(z, d)$, for the case when, in addition,

(*c*) $z \geqslant (r-1)\,d$. When this expression is being used, the asterisk will be dropped, as in equation (4) below.

LEMMA 2. *For the random variables $x_1, x_2, \ldots, x_r$ discussed above, the probability, given* (*c*), *that* (*a*) *and* (*b*) *are jointly true, is*

$$C_r(z, d) = (z + d + rd)^{r-2}\,((z+d)^2 - rd^2)/r!. \tag{4}$$

As d will be a constant, we write $C_r(z, d) \equiv C_r(z)$, and define $C_0(z) \equiv 1$.

Proof. We imagine the following figure. Suppose a rectangle, length z, height d, lies on the x-axis, from $x = 0$ to $x = z$. On top of this rectangle lies a similar rectangle, moved a distance d to the right. This is repeated until r rectangles are in the pile. The length of each rectangle represents the permitted range of one of the x_i. The height d has been chosen only to give the type of figure which arises in the discussion of V_N. On the x-axis, at $x = z$, a vertical line A is drawn; x_1 must lie to the left of A while other x_i may lie to the left or the right. Because of the restriction (*c*) on z and d, no x_i lies wholly to the right of A. Let K_i be the event that

x_i lies to the left of A, and L_i be the event that it lies to the right. Then, denoting the probability of an event E by $P(E)$,

$$C_r(z) = P(K_1 L_2 L_3 \ldots L_r) + P(K_1 K_2 L_3 \ldots L_r) + \ldots + P(K_1 K_2 \ldots K_{r-1} L_r) + P(K_1 K_2 \ldots K_r).$$

Giving the probabilities term by term, this is easily seen from the figure to be

$$C_r(z) = A_1(z)\, A_{r-1}(d) + A_2(z-d)\, A_{r-2}(2d) + \ldots + A_{r-1}(z-(r-2)\,d)\, A_1((r-1)\,d) + A_r(z-(r-1)\,d)\, A_0(rd).$$

Thus

$$C_r(z) = \sum_{l=1}^{r} A_i(z-(i-1)\,d)\, A_{r-1}(id)$$

$$= \frac{1}{r} \sum_{l=1}^{r} \frac{(rd)^{r-i}\,(z+d)^{i-1}\,(z-(i-1)\,d)}{(i-1)!\,(r-i)!}.$$

The final bracket may be used to break the sum into two parts. This gives

$$C_r(z) = \frac{z}{r!} \sum_{i=1}^{r} \binom{r-1}{i-1} (rd)^{r-i}\,(z+d)^{i-1} - \frac{d}{r(r-2)!} \sum_{i=2}^{r} \binom{r-2}{i-2} (rd)^{r-i}\,(z+d)^{i-1}$$

$$= \frac{z}{r!}\,(z+d+rd)^{r-1} - \frac{d(z+d)}{r(r-2)!}\,(z+d+rd)^{r-2}$$

$$= \frac{1}{r!}\,(z+d+rd)^{r-2}\,((z+d)^2 - rd^2).$$

Corollary. *The probability that, after being placed in ascending order, r independent observations from $U(0,1)$ will satisfy conditions (a) and (b) of Lemma 2, is $r!\,C_r^*(z,d)$; if z satisfies condition (c), the probability is then $r!\,C_r(z,d)$.* The proof follows that for Lemma 1, Corollary.

3·3. Lemma 3. *Let*

$$D_r(z,d) \equiv \sum_{i=0}^{r} C_i(z-d)\, A_{r-i}(d) \quad (r \geqslant 0). \tag{5}$$

Then

$$D_r(z,d) = (y^r - 2rdy^{r-1} + r(r-1)\,d^2 y^{r-2})/r!, \tag{6}$$

where $y = z + (r+1)\,d$.

Proof. We start with an identity due to Abel, quoted by Birnbaum & Pyke (1958). This states that, for a, b real, and n an integer $\geqslant 0$,

$$(b-n) \sum_{i=0}^{n} \binom{n}{i} (a+1)^i\,(b-i)^{n-i-1} \equiv (a+b)^n.$$

Then

$$S_n(a,b) \equiv \sum_{i=0}^{n-1} \binom{n}{i} (a+i)^i\,(b-i)^{n-i-1} \equiv \{(a+b)^n - (a+n)^n\}/(b-n).$$

$S_n(a,b)$ from the left-hand sum, is continuous at $b = n$. Therefore

$$S_n(a,n) = \lim_{b \to n} ((a+b)^n - (a+n)^n)/(b-n)$$

that is,

$$S_n(a,n) = n(a+n)^{n-1}. \tag{7}$$

Using (3) and (4) in (5), we have

$$D_r(z,d) = \sum_{i=0}^{r} \frac{(z+id)^{i-2}\,(z^2 - id^2)\,d^{r-1}(1+r-i)^{r-i}}{i!\,(1+r-i)!}.$$

By the identity $$z^2 - id^2 \equiv (z+id)^2 - 2id(z+id) + i(i-1)\,d^2,$$

$D_r(z,d)$ is separated into three summations, $S_A + S_B + S_C$. The first of these is

$$S_A = \sum_{i=0}^{r} \frac{d^{r-1}(z+id)^i\,(1+r-i)^{r-i}}{i!\,(1+r-i)!}$$

$$= \frac{d^r}{(r+1)!} \sum_{i=0}^{r} \binom{r+1}{i} \left(\frac{z}{d}+i\right)^i (1+r-i)^{r-i} = \frac{d^r}{(r+1)!}\, S_{r+1}\left(\frac{z}{d}, r+1\right).$$

Using (7), we have $$S_A = \frac{d^r}{(r+1)!}\,(r+1)\left(\frac{z}{d}+r+1\right)^r = \frac{(z+(r+1)\,d)^r}{r!}.$$

The other two sums, after similar manipulation, become

$$S_B = \frac{-2d}{r!}\, r(z+(r+1)\,d)^{r-1}$$

and

$$S_C = \frac{d^2}{r!}\, r(r-1)\,(z+(r+1)\,d)^{r-2}.$$

The above forms are used to show that $S_B = 0$ when $r = 0$, and $S_C = 0$ when $r = 0$ or 1. The sum of these gives $D_r(z,d)$ in the form (6). We note that $D_0(z,d) \equiv 1$.

4. The distribution of V_N

4·1. *Assumptions.* We shall calculate the distribution of V_N, on the null hypothesis, by supposing that the N independent observations are from a uniform distribution on a circle of unit circumference. A specific observation, given by Lemma 4, will be chosen as the origin for x, and the positive direction will be clockwise. These assumptions, as stated earlier, do not affect the distribution of V_N.

4·2. The technique to be employed rests on the result of the following:

LEMMA 4. *If N points are given on a circle of circumference 1, it is possible to determine at least one point P_1 such that, if subsequent consecutive points clockwise are labelled $P_2, P_3, \ldots, P_N$, the arc lengths $P_1P_i \leqslant (i-1)/N$, for $2 \leqslant i \leqslant N$.*

Proof. Suppose that the points are consecutively labelled clockwise $B_1, B_2, \ldots, B_N$. *Assume H: the lemma is false.* A particle may then start at any point, B_i, and, for some k, jump k points to B_{i+k}, covering an arc distance greater than k/N. Imagine a succession of such jumps. Since N is finite, the particle eventually arrives at a previously occupied point, say B_r. Since last at B_r it has gone round the unit circle say C times, covering a distance C. In so doing it has jumped CN points and, by H, has covered an arc greater than

$$(1/N)\,(CN) = C.$$

Thus we have a contradiction; H is false, and the lemma true.

Further, it may easily be shown that, with probability 1, P_1 is unique. We therefore choose P_1 as the origin for x, and label the other observations, in order moving clockwise, $P_2, P_3, \ldots, P_N$. Let P_i have co-ordinate x_i. The population and sample distribution functions (D.F.) are now defined only for $0 \leqslant x \leqslant 1$. The population D.F. is $F(x) = x$, $0 \leqslant x \leqslant 1$; the sample D.F. is

$$F_N(x_1) = F_N(0) = \frac{1}{N};$$

$$F_N(x) = \frac{i}{N} \quad (x_i \leqslant x < x_{i+1}).$$

4·3. It will be helpful to have a figure, which the reader may draw as follows:

(*a*) Draw the usual rectangular x, y co-ordinate axes, and let the origin at O be labelled also P_1 and A_1. The observations are represented by points P_i ($\equiv (x_i, 0)$), on the x-axis. Let the point A_i have co-ordinates $((i-1)/N, 0)$, for $1 \leqslant i \leqslant N$. Suppose $D \equiv (1,0)$, $E \equiv (1,1)$ and $F \equiv (0,1)$. Draw the population D.F. (the line OE) and the sample D.F.

(*b*) Let d be $1/N$. Parallel to OE, draw dotted lines $y = x + nd$, $1 \leqslant n \leqslant N-1$. Draw also the solid line L, given by $y = x+z$, and let L cut the horizontal lines $y = id$, $1 \leqslant i \leqslant N$, in points M_i, within the rectangle $ODEF$. Let L cut the y-axis in M_0.

(*c*) When $1 - Kd < z \leqslant 1-(K-1)d$, we say that z and the line L are in stage K. For $1 \leqslant i \leqslant N$, let $y_i = F_N(x_i)$, and let Q_i be the point (x_i, y_i). When Q_i is above or on the line L, $V_N \geqslant Z$; we shall then say that Q_i exceeds L. The smallest value of V_N is d, so that values of z in stage N are not considered.

(*d*) *Event* E_s. We shall define an event E_s, for $1 \leqslant s \leqslant K$, as occurring when the sth Q, moving downwards from Q_N, is *the first* to exceed L, though lower Q's may also do so. More precisely, Q_{N+1-s} exceeds L while Q_i, for $i > N+1-s$, does not exceed L, and Q_j, for $j < N+1-s$, may or may not exceed L. Clearly s is restricted to $1 \leqslant s \leqslant K$ when L is in stage K.

(*e*) As illustration, suppose in the figure described above we have $N = 12$, and let z be just greater than $7d$. Suppose $F_N(x)$ is such that the first eight observations are so crowded together that Q_8 exceeds L and also Q_{10} exceeds L, but no other Q_i exceeds L. Since Q_{10} is the third Q moving downwards, the event E_3 is occurring.

4·4. *Probability notations.* Union of events A, B is denoted by $A \cup B$, and intersection by AB or $A \cap B$; the intersection of A and the complement of B, if B is a subset of A, is denoted by $A - B$. $P(E)$ is the probability of event E; $P(V_N \geqslant z)$ is called $P_N(z)$.

4·5. If G, H are two points not necessarily on the x-axis, the statement '$P_i \in GH$' or '$x_i \in GH$' will be used to mean that the point P_i, co-ordinates $(x_i, 0)$, lies in the closed interval $G'H'$ where G', H' are the projections of G, H on the x-axis.

4·6. *The distribution of V_N, upper tail. Introduction to Theorem* 1. We see that, for given z, $V_N \geqslant z$ whenever one of the mutually exclusive events E_s occurs. Thus for given z, so that K is known,

$$P_N(z) = \sum_{s=1}^{K} P(E_s) \tag{8}$$

and we now seek $P(E_s)$.

Probability of event E_s. We are given N observations independently chosen from a uniform distribution on a circle of circumference 1. Suppose these are divided into three groups as follows: one is chosen, at random, to be P_1; $s-1$ of the other observations are then picked at random to be a set called $S1$; the $N-s$ remaining observations form a set called $S2$. The point P_1 is then chosen as the origin of x. Let the observations in $S2$, in ascending order, be called x_2 to x_{N-s+1}, and let those in $S1$, in ascending order, be called x_{N-s+2} to x_N. The event E_s will then occur, provided the following conditions are met.

For set $S1$: $\quad S1a$: $x_j \in M_j A_j$, $\quad N-s+2 \leqslant j \leqslant N$;

and for set $S2$: $\quad S2a$: $x_j \in OM_{N-s+1}$, $\quad 2 \leqslant j \leqslant N-s+1$

and $\quad S2b$: $x_j \in OA_j$, $\quad 2 \leqslant j \leqslant N-s+1$.

These restrictions together state that the least in $S1$ must be greater than the greatest in $S2$: thus the way we have labelled the observations not only puts them in ascending order in

each set, but also ranks them within the entire group of N observations, i.e.

$$x_1 \leqslant x_2 \leqslant x_3 \ldots \leqslant x_N.$$

The observations may be assigned to the three groups in $N\binom{N-1}{s-1}$ ways. Once this is done, the event E_{sa}, (that the observations in $S1$ satisfy condition $S1a$), and the event E_{sb} (that those in $S2$ satisfy conditions $S2a$ and $S2b$) are clearly independent. Thus the total probability of event E_s is given by

$$P(E_s) = N\binom{N-1}{s-1} P(E_{sa}).P(E_{sb}) \tag{9}$$

and we now must find $P(E_{sa})$ and $P(E_{sb})$.

4·7. *Probability of event E_{sa}.* The conditions on the variables in $S1$ are those of Lemma 2, Corollary, with z now replaced by $z-d$. Thus

$$P(E_{sa}) = (s-1)!\, C^*_{s-1}(z-d, d). \tag{10}$$

4·8. *Probability of event E_{sb}.* For event E_{sb} we first define mutually exclusive events $G_1, G_2, G_3, \ldots$, etc., as follows:

$$G_1\colon \{x_j \in OM_{N-s+1}, \quad 2 \leqslant j \leqslant N-s+1\},$$
$$G_2\colon \{x_j \in A_2 M_{N-s+1}, \quad 2 \leqslant j \leqslant N-s+1\}$$

and, for $3 \leqslant m \leqslant K-s+1$,

$$G_m\colon \{x_i \in OA_1, \quad 2 \leqslant i \leqslant m-1, \quad x_j \in A_m M_{N-s+1}, \quad m \leqslant j \leqslant N-s+1\}.$$

The event $G_1 - G_2$ gives the event that all points in set $S2$ are in OM_{N-s+1}, with at least one in OA_2; $G_1 - G_2 - G_3$ gives the event that all points in $S2$ are in OM_{N-s+1}, with at least one in OA_2, and at least two in OA_3. Thus it may be seen that event E_{sb} is given by

$$E_{sb} = G_1 - G_2 - G_3 - \ldots - G_{K-s+1}, \quad \text{if} \quad 1 \leqslant s \leqslant K-1;$$

and by

$$E_{sb} = G_1, \quad \text{if} \quad s = K.$$

$$\left.\begin{aligned} &\text{So, when } 1 \leqslant s \leqslant K, && P(E_{sb}) = P(G_1) - \sum_{m=2}^{K-s+1} P(G_m), \\ &\text{and when } s = K, && P(E_{sb}) = P(G_1). \end{aligned}\right\} \tag{11}$$

We now need the probabilities $P(G_m)$ $(m = 1, 2, \ldots, K-s+1)$. For event G_1, we require that $N-s$ independent observations all fall into the interval OM_{N-s+1}, of length $1-z-(s-1)d$. Thus $P(G_1) = (1-z-(s-1)d)^{N-s}$. For event G_m, $m \geqslant 2$, we first must divide the $N-s$ members of set $S2$ into two subsets, $S21$ with $m-2$ members, and $S22$ with $N-s-m+2$ members. Subset $S21$, after being put in ascending order, will be $x_2, x_3, \ldots, x_{m-1}$ and will satisfy the conditions of Lemma 1, Corollary, with z equal to d. Subset $S22$ is all to fall in the interval $A_m M_{N-s+1}$, of length $1-z-(s+m-2)d$. The subsets $S21$, $S22$ are independent, so

$$P(G_m) = \binom{N-s}{m-2} (m-2)!\, A_{m-2}(d) \{1-z-(s+m-2)d\}^{N-s-m+2}.$$

Thus in (11) we have, for $1 \leqslant s \leqslant K$,

$$P(E_{sb}) = (1-z-(s-1)d)^{N-s} - \sum_{m=2}^{K-s+1} \frac{(N-s)!}{\{N-(s+m-2)\}!} A_{m-2}(d) \{1-z-(s+m-2)d\}^{N-(s+m-2)}. \tag{12}$$

We now have $P(E_{sa})$ and $P(E_{sb})$ and can substitute in (9). However, if we wish to evaluate $P(E_s)$ we must impose the condition of Lemma 2 to make it possible to calculate $C^*_{s-1}(z-d, d)$.

Thus we require $z-d \geqslant (s-1)\,d$ for all s from 1 to K, and K is connected with z by the relation $1-Kd < z \leqslant 1-(K-1)\,d$. These requirements lead to condition (a) in Theorem 1 below, restricting z to the upper tail. When this condition is met, we can say

$$P(E_{sa}) = (s-1)!\,C_{s-1}(z-d). \tag{13}$$

We now are in a position to prove the following theorem.

4·9. THEOREM 1. *Distribution of V_N, upper tail.*

(a) *Let z satisfy the inequality: $z \geqslant \frac{1}{2}$, if N is even, or $z \geqslant (N-1)/(2N)$, if N is odd;*

(b) *let $M = [N(1-z)]$, i.e. the greatest integer contained in $N(1-z)$;*

(c) *let $y = z+t/N$, and let*

$$T_t = y^{t-3}[y^3N - y^2t(3-2/N) + yt(t-1)\,(3-2/N)/N - t(t-1)\,(t-2)/N^2].$$

Then
$$P(V_n \geqslant z) = P_N(z) = \sum_{t=0}^{M} \binom{N}{t} (1-z-td)^{N-t-1}\,T_t. \tag{14}$$

Proof. When z satisfies (a), we can use equations (12) and (13) in equation (9) to obtain $P(E_s)$, and then get $P_N(z)$ from equation (8). The result is

$$P_N(z) = \sum_{s=1}^{K} P(E_s) = \text{sum } A - \text{sum } B,$$

where
$$\text{sum } A = \sum_{s=1}^{K} N\tbinom{N-1}{s-1}\,(s-1)!\,C_{s-1}(z-d)\{1-z-(s-1)\,d\}^{N-s}$$

and

$$\text{sum } B = \sum_{s=1}^{K-1} \sum_{m=2}^{K-s+1} N\tbinom{N-1}{s-1}\,(s-1)!\,C_{s-1}(z-d) \times \frac{(N-s)!}{\{N-(s+m-2)\}!}\,A_{m-2}(d)\{1-z-(s+m-2)\,d\}^{N-(s+m-2)}.$$

We first introduce g_t for $N(N-1)\ldots(N-t+1)$; then, in sum B, for given s change the variable m to t, given by $t = s+m-2$. Sum B becomes

$$\text{sum } B = \sum_{s=1}^{K-1} \sum_{t=s}^{K-1} g_t\,C_{s-1}(z-d)\,A_{t-s}(d)\,(1-z-td)^{N-t}.$$

Reversing the order of summation gives

$$\begin{aligned}\text{sum } B &= \sum_{t=1}^{K-1} g_t(1-z-td)^{N-t} \sum_{s=1}^{t} C_{s-1}(z-d)\,A_{t-s}(d) \\ &= \sum_{t=1}^{K-1} g_t(1-z-td)^{N-t}\,D_{t-1}(z,d), \quad \text{using (5).}\end{aligned}$$

In sum A we change index s to t given by $t = s-1$. Then

$$\begin{aligned}P_N(z) &= \sum_{t=0}^{K-1} g_{t+1}\,C_t(z-d)\,(1-z-td)^{N-t-1} - \sum_{t=1}^{K-1} g_t(1-z-td)^{N-t}\,D_{t-1}(z,d) \\ &= N(1-z)^{N-1} + \sum_{t=1}^{K-1} g_t(1-z-td)^{N-t-1}[(N-t)\,C_t(z-d) - (1-z-td)\,D_{t-1}(z,d)].\end{aligned}$$

With the results of Lemmas 2 and 3, the square bracket simplifies to $T_t/t!$, where T_t is defined in (c) above. Further, M has been defined in (b) so that $M = K-1$. Thus

$$P_N(z) = \sum_{t=0}^{M} \binom{N}{t} (1-z-td)^{N-t-1}\,T_t,$$

as given by (14).

4·10. *Distribution of V_N, lower tail.* Theorem 1 may be used to give the values of $P_N(z)$ only in the upper tail. Another method has been used to find the lower tail distribution and is illustrated in the proof of Theorem 2. We use the same notation as in §§ 4·1–4·9.

THEOREM 2 (*a*). *For* $1/N \leqslant z \leqslant 2/N$, $(N \geqslant 2)$,

$$P(V_N \leqslant z) \equiv 1 - P_N(z) = N!(z - 1/N)^{N-1}. \tag{15}$$

THEOREM 2 (*b*). *For* $2/N \leqslant z \leqslant 3/N$, $(N \geqslant 3)$,

$$P(V_N \leqslant z) \equiv 1 - P_N(z) = \frac{(N-1)!\{\beta^{N-1}(1-\alpha) - \alpha^{N-1}(1-\beta)\}}{N^{N-2}(\beta - \alpha)}, \tag{16}$$

where $t = \alpha, \beta$ *are the solutions of the quadratic equation*

$$t^2 - (Nz - 1)t + \tfrac{1}{2}(Nz - 2)^2 = 0. \tag{17}$$

Proof. THEOREM 2 (*a*). V_N takes its minimum value $1/N$, when every P_i is at A_i. For z between $1/N$ and $2/N$, and for $V_N \leqslant z$, each P_i may move to the left up to a distance $z - 1/N$; i.e. $id - z \leqslant x_i \leqslant (i-1)d$, $(2 \leqslant i \leqslant N)$. The probability of this event, counting all orderings, is easily seen to be $N!(z - 1/N)^{N-1}$.

THEOREM 2 (*b*). For $V_N \leqslant z$, $x_i \in M_i A_i$, for all $i \geqslant 3$. (Necessarily, $x_2 \in OA_2$.) $M_i A_i$ has length $z - d$. Suppose Q_i denotes the event that $x_i \in M_i A_{i-1}$, and R_i the event that $x_i \in A_{i-1} A_i$. Because of the ordering, if event Q_i occurs, the range of x_{i-1} is restricted. The probability of the compound event $Q_i R_{i-1}$, for $i \geqslant 3$, is

$$I = \int_0^{z-2d} (d - t)\, dt = \tfrac{1}{2}(6dz - z^2 - 8d^2).$$

We wish to describe the event E_s, in which the $N - s$ largest variables x_i, $(s+1 \leqslant i \leqslant N)$, are each in the appropriate interval $A_{i-1} A_i$ while the other variables, x_k, $(3 \leqslant k \leqslant s)$, still in ascending order, are in the intervals $M_j A_j$. Thus E_s is described by an intersection of events of the form $E_s \equiv R_N R_{N-1} \ldots R_{s+1} Z_s Z_{s-1} \ldots Z_3 R_2$, where Z_k may be the event Q_k or the event R_k, for $3 \leqslant k \leqslant s$. In such a sequence for E_s, a Q followed by R, as noted above, must be treated as a compound event with probability I; but all other letters in the sequence will represent independent events. Thus, e.g. $R_7 Q_6 Q_5 R_4 Q_3 R_2$ is the intersection of the events $R_7 \cap Q_6 \cap (Q_5 R_4) \cap (Q_3 R_2)$, with probability $d(z - 2d) I^2$. The event E_N gives all situations, for any one ordering, in which $V_N \leqslant z$. Thus we must find $P(E_N)$ as follows. We start with event E_3.

Event E_3. This is given by $E_{3u} \cup E_{3v}$ where E_{3u} is $R_N R_{N-1} \ldots R_4 R_3 R_2$, with probability $u_3 = d^{N-1}$ and E_{3v} is $R_N R_{N-1} \ldots R_4 Q_3 R_2$ with probability $v_3 = d^{N-3} I$. $P(E_3)$ is then $u_3 + v_3$.

Event E_4. Suppose we define the four mutually exclusive events following:

E_{411} is $R_N R_{N-1} \ldots R_5 R_4 R_3 R_2$, probability d^{N-1};
E_{412} if $R_N R_{N-1} \ldots R_5 R_4 Q_3 R_2$, probability $d^{N-3} I$;
E_{421} is $R_N R_{N-1} \ldots R_5 Q_4 R_3 R_2$, probability $d^{N-4} I d$;
E_{422} if $R_N R_{N-1} \ldots R_5 Q_4 Q_3 R_2$, probability $d^{N-4} y I$,

where $y = z - 2d$. Then $E_4 = E_{4u} \cup E_{4v}$, where

$$E_{4u} = E_{411} \cup E_{412}, \quad \text{with probability } u_4,$$

and

$$E_{4v} = E_{421} \cup E_{422}, \quad \text{with probability } v_4.$$

Then

$$u_4 = u_3 = v_3 \quad \text{and} \quad v_4 = I u_3/d^2 + y v_3/d.$$

Finally

$$P(E_4) = u_4 + v_4.$$

Event E_{s+1}. In general, E_{s+1} is obtained from E_s by

(*a*) keeping R_{s+1} as it is, producing events whose union we call $E_{s+1,u}$, with probability u_{s+1}, and by

(*b*) changing R_{s+1} to Q_{s+1}, producing events whose union we call $E_{s+1,v}$, with probability v_{s+1}.

With this procedure, a new combination $\ldots Q_{s+1} R_s \ldots$ replaces d^2 by I, and a new combination $\ldots Q_{s+1} Q_s \ldots$ replaces d by y. Thus

$$u_{s+1} = u_s + v_s \tag{18}$$

and

$$v_{s+1} = \frac{I}{d^2} u_s + \frac{y}{d} v_s. \tag{19}$$

Using (18) to eliminate v in (19) we have

$$u_{s+2} - (1 + y/d)\, u_{s+1} + (y/d - I/d^2)\, u_s = 0.$$

This difference equation is solved by standard techniques, using the known values for u_3, v_3 to solve for the arbitrary constants. The result for u_s is

$$u_s = d^{N-1}\{\beta^{s-2}(1-\alpha) - \alpha^{s-2}(1-\beta)\}/(\beta - \alpha),$$

where α, β are the solutions of

$$t^2 - (1 + y/d)\, t + y/d - I/d^2 = 0. \tag{20}$$

When the expressions for y, d and I are substituted in (20), the equation for t becomes (17). For the rank ordering, therefore, $P(E_N) = u_N + v_N$, which, by (18), equals u_{N+1}. Any of the $N!$ possible orderings of the observations might be the rank ordering, with equal probability. The total probability $P(V_N \leqslant z)$ is thus given by $N!\, u_{N+1}$, which gives (16).

4·11. *The mean of* V_N. At this point we add one isolated result. This is the mean of V_N, which may easily be deduced from equation (24) of Birnbaum & Pyke (1958). This gives the mean of $\sup(F_N(x) - F(x))$; the mean of $\inf(F_N(x) - F(x))$ is the negative of this, and from these results

$$E(V_N) = \frac{N!}{N^{N+1}} \sum_{i=0}^{N-1} \frac{N^i}{i!}.$$

4·12. *Extensions to Theorems* 1 *and* 2. Theorem 1 may clearly be extended if $C_r^*(z, d)$ can be evaluated to give $P(E_{sa})$ in (10). Theorem 2 may also be extended upwards, though at the next stage a quartic equation must be solved to give the solution of the finite difference equation which arises. In principle it would be gratifying to find the complete solution and a way of matching the two tails. This would perhaps make it possible also to obtain the complete asymptotic distribution, i.e. the extension of (2), by the method which Lauwerier (1963) has used to solve a similar problem. However, in practice, for the production of statistical tables the need is not great as will be seen below.

4·13. *Compilation of Tables* 1 *and* 2. Theorems 1 and 2 have been used to compute by inverse interpolation the exact significance points above the horizontal line in each column of Tables 1 and 2. The points for larger values of N have been obtained with the help of (2). This expression gives approximate points which are too low compared with the exact values in the upper tail, and are too high in the lower tail. The error in significance level which is given by using these approximate values is very small but, nevertheless, for higher values of N, better estimates of significance points may be obtained by interpolation in a graph of existing exact critical values of $\sqrt{N}\, V_N$ against $1/N$, including those for $N = \infty$. The remaining significance points have been obtained in this way, using the points given by (2) as a guide.

This interpolation may be continued for $N > 100$; to this end critical values of $\sqrt{N}V_N$ are included in the tables, placed in parentheses. Such interpolation will give better accuracy than that given by using the asymptotic points; for example, when $N = 100$, use of the asymptotic value of $\sqrt{N}V_N$, at the upper 5% level (1·747), gives very nearly a 4% test. However, it should be pointed out that for these high values of N inaccuracies in the measurement of x_i may affect the conclusion of the test more than slight errors in significance points.

5. Results on the asymptotic distributions

Some interesting relationships exist between the asymptotic distributions of the four test statistics, $\sqrt{N}V_N$, K_N, W_N^2 and U_N^2, the last three being defined by

$$K_N = \sqrt{N} \sup_{-\infty<x<\infty} |F_N(x) - F(x)|,$$

$$W_N^2 = N\int_{-\infty}^{\infty} (F_N(x) - F(x))^2 \, dF(x)$$

and

$$U_N^2 = N\int_{-\infty}^{\infty} \left(F_N(x) - F(x) - \int_{-\infty}^{\infty} (F_N(y) - F(y))\, dF(y)\right)^2 dF(x).$$

Using the notation K^2 for $\lim_{N\to\infty} K_N^2$ and $\phi(t; K^2)$ for the characteristic function of the null-hypothesis distribution of K^2, and similarly for the other statistics, the known characteristic functions are

$$\phi(t; K^2) = \prod_{j=1}^{\infty} \left(1 - \frac{it}{2j^2}\right)^{-1},$$

$$\phi(t; W^2) = \prod_{j=1}^{\infty} \left(1 - \frac{2it}{j^2\pi^2}\right)^{-\frac{1}{2}},$$

and

$$\phi(t; U^2) = \prod_{j=1}^{\infty} \left(1 - \frac{it}{2j^2\pi^2}\right)^{-1}.$$

To these we now add, defining $\lim_{N\to\infty} \sqrt{N}V_N$ as V_α,

$$\phi(t; V_\alpha^2) = \prod_{j=1}^{\infty} \left(1 - \frac{it}{2j^2}\right)^{-2}. \tag{21}$$

Watson (1961) had noticed the interesting fact that K^2/π^2 and U^2 have the same distribution, and Pearson & Stephens (1962) that the sth cumulants of W^2 and U^2, say κ_s and κ_s' respectively, are connected by the relation $\kappa_s' = 2^{1-2s}\kappa_s$. Equation (21) has been derived from the observation that the sth cumulant of V_α, say κ_s'', is connected with κ_s' by $\kappa_s'' = 2\pi^{2s}\kappa_s'$.

Thus if we consider 4 new statistics, S_1, S_2, S_3, S_4, derived from the above by the relations

$$S_1 = W^2/4, \quad S_2 = U^2, \quad S_3 = K^2/\pi^2, \quad S_4 = V_\alpha^2/\pi^2,$$

the sth cumulants, respectively κ_{1s}, κ_{2s}, κ_{3s}, κ_{4s}, of their distributions are easily shown to be connected by the simple relations

$$2\kappa_{1s} = \kappa_{2s} = \kappa_{3s} = \tfrac{1}{2}\kappa_{4s}.$$

The author acknowledges with thanks the help of the McGill University Computing Centre, where the tables were computed; also several helpful conversations with members of the University, notably Prof. I. G. Connell, and Messrs D. Sankoff, M. Angel and E. Rothman. The referee is also thanked for valuable suggestions to improve the form of the paper.

References

Birnbaum, Z. W. & Pyke, R. (1958). On some distributions related to the statistic D_N^+. *Ann. Math Statist.* **29**, 179–87.

Birnbaum, Z. W. & Tingey, Fred H. (1951). One-sided confidence contours for probability distribution functions. *Ann. Math. Statist.* **22**, 592–6.

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Proc. Koninkl. Nederl. Akad. Van Wettenschappen*, Series A, **63**, 38–47.

Lauwerier, H. A. (1963). The asymptotic expansion of the statistical distribution of N. V. Smirnov. *Z. Wahrscheinlichkeits theorie und Verw. Gebiete*, **2**, 61–8.

Pearson, E. S. (1963). Comparison of tests for randomness of points on a line. *Biometrika*, **50**, 315–25.

Pearson, E. S. & Stephens, M. A. (1962). The goodness-of-fit tests based on W_N^2 and U_N^2. *Biometrika*, **49**, 397–402.

Stephens, M. A. (1963). The distribution of the goodness-of-fit statistic U_N^2. I. *Biometrika*, **50**, 303–13.

Stephens, M. A. (1964). The distribution of the goodness-of-fit statistic U_N^2. II. *Biometrika*, **51**, 393–8.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. I. *Biometrika*, **48**, 109–14.

Watson, G. S. (1962). Goodness-of-fit tests on a circle. II. *Biometrika*, **49**, 57–63.