THE DISTRIBUTION OF THE MAXIMUM DEVIATION BETWEEN TWO SAMPLE CUMULATIVE STEP FUNCTIONS

BY FRANK J. MASSEY, JR.[1]

University of Oregon

1. Summary. Let $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_m$ be the ordered results of two random samples from populations having continuous cumulative distribution functions $F(x)$ and $G(x)$ respectively. Let $S_n(x) = k/n$ when k is the number of observed values of X which are less than or equal to x, and similarly let $S'_m(y) = j/m$ where j is the number of observed values of Y which are less than or equal to y.

The statistic $d = \max_x | S_n(x) - S'_m(x) |$ can be used to test the hypothesis $F(x) \equiv G(x)$, where the hypothesis would be rejected if the observed d is significantly large. The limiting distribution of $d\sqrt{\frac{mn}{m+n}}$ has been derived [1] and [4], and tabled [5]. In this paper a method of obtaining the exact distribution of d for small samples is described, and a short table for equal size samples is included. The general technique is that used by the author for the single sample case [2]. There is a lower bound to the power of the test against any specified alternative, [3]. This lower bound approaches one as n and m approach infinity proving that the test is consistent.

2. Distribution of d. Denote by α_1 the number of observed values of Y which are less than x_1, by α_2 the number of values of Y which are between x_1 and x_2, $\cdots$, by α_{n+1} the number of values of Y which are greater than x_n. It is known that, if the hypothesis $F(x) \equiv G(x)$ is true, the probability of the occurrence of any set of $\alpha_1, \cdots, \alpha_{n+1}$ is $n!m!/(m+n)!$ Thus the probability that $d \leq a$ can be found by counting the number of sets of $\alpha_1, \cdots, \alpha_{n+1}$ which give values of $d \leq a$ and multiply this number by $n!m!/(m+n)!$ The method of counting is illustrated here for $n = m$, and some results are given in Table 1. If $n = m$ then $S_n(x)$ and $S'_n(y)$ can only differ by multiples of $1/n$. (If $n \neq m$ they can only differ by multiples of $1/mn$.) For integer k we count the number of sets of $\alpha_1, \cdots, \alpha_{n+1}$ such that $d \leq k/n$.

Denote by $U_i(j)$, $j = 1, 2, \cdots, n$, $i = 0, 1, 2, \cdots, 2k-1$, the number of sets of possible $\alpha_1, \alpha_2, \cdots, \alpha_j$ such that $S'_n(x_j) = (j + i - k)/n$ and such that $| S_n(x) - S'_n(x) |$ has been less than or equal to k/n for $x < x_j$. It is easily seen that these $X_i(j)$ satisfy the following difference equations.

$$\begin{aligned} U_0(j+1) &= U_0(j) + U_1(j), \\ U_1(j+1) &= U_0(j) + U_1(j) + U_2(j), \\ &\vdots \\ U_{2k-2}(j+1) &= U_0(j) + \cdots + U_{2k-1}(j), \\ U_{2k-1}(j+1) &= U_0(j) + \cdots + U_{2k-1}(j). \end{aligned}$$

[1] Research under contract N6-onr-218/IV with the Office of Naval Research.

TABLE 1

Probability of $d \leqq k/n$

$n = m$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
1	1.000000					
2	.666667	1.000000				
3	.400000	.900000	1.000000			
4	.228571	.771429	.971429	1.000000		
5	.126984	.642857	.920635	.992063	1.000000	
6	.069264	.525974	.857143	.974026	.997835	1.000000
7	.037296	.424825	.787879	.946970	.991841	.999417
8	.019891	.339860	.717327	.912976	.981352	.997514
9	.010537	.269889	.648293	.874126	.966434	.993706
10	.005542	.213070	.582476	.832179	.947552	.987659
11	.002903	.167412	.520850	.788524	.925339	.979261
12	.001515	.131018	.463902	.744225	.900453	.968564
13	.000788	.102194	.411804	.700080	.873512	.955728
14	.000408	.079484	.364515	.656680	.845065	.940970
15	.000211	.061669	.321862	.614453	.815584	.924536
16	.000109	.047744	.283588	.573707	.785465	.906674
17	.000056	.036893	.249393	.534647	.755041	.887623
18	.000029	.028460	.218952	.497410	.724582	.867606
19	.0^{4}148	.021922	.191938	.462071	.694311	.846827
20	.0^{5}761	.016863	.168030	.428664	.664409	.825467
21	.0^{5}390	.012956	.146921	.397187	.635020	.803688
22	.0^{5}199	.009943	.128321	.367614	.606260	.781632
23	.0^{5}102	.007623	.111963	.339899	.578218	.759422
24	.0^{6}52	.005839	.097600	.313983	.550963	.737166
25	.0^{6}27	.004468	.085007	.289796	.524546	.714958
26	.0^{6}14	.003417	.073980	.267263	.499005	.692877
27	.0^{7}69	.002611	.064338	.246303	.474362	.670992
28	.0^{7}35	.001994	.055914	.226833	.450633	.649362
29	.0^{7}18	.001522	.048563	.208772	.427823	.628036
30	.0^{8}91	.001161	.042154	.192037	.405929	.607055
31	.0^{8}46	.000885	.036570	.176546	.384946	.586455
32	.0^{8}23	.000674	.031710	.162223	.364861	.566264
33	.0^{8}12	.000513	.027483	.148989	.345657	.546505
34	.0^{9}60	.000391	.023808	.136773	.327316	.527198
35	.0^{9}31	.000297	.020616	.125505	.309816	.508355
36	.0^{9}16	.000226	.017845	.115120	.293133	.489989
37	.0^{10}79	.000172	.015440	.105553	.277243	.472107
38	.0^{10}40	.000131	.013355	.096747	.262121	.454713
39	.0^{10}20	.000099	.011547	.088645	.247738	.437811
40	.0^{10}10	.000075	.009981	.081195	.234069	.421400

TABLE 1—*Continued*

$n = m$	$k = 7$	$k = 8$	$k = 9$	$k = 10$	$k = 11$	$k = 12$
1						
2						
3						
4						
5						
6						
7	1.000000					
8	.999845	1.000000				
9	.999260	.999959	1.000000			
10	.997943	.999783	.999989	1.000000		
11	.995634	.999345	.999938	.999997	1.000000	
12	.992141	.998503	.999796	.999982	.999999	1.000000
13	.987351	.997125	.999500	.999938	.999995	1.000000
14	.981218	.995100	.998979	.999837	.999981	.999999
15	.973752	.992344	.998163	.999647	.999948	.999994
16	.965002	.988801	.996985	.999330	.999880	.999983
17	.955047	.984439	.995389	.998847	.999762	.999960
18	.943982	.979252	.993331	.998160	.999571	.999917
19	.931911	.973251	.990776	.997233	.999286	.999844
20	.918942	.966458	.987701	.996033	.998884	.999729
21	.905183	.958911	.984095	.99453	.99834	.99956
22	.890738	.950653	.979953	.99271	.99764	.99933
23	.875705	.941731	.975280	.99055	.99676	.99901
24	.860177	.932197	.970087	.98803	.99568	.99860
25	.844240	.922101	.964389	.98516	.99438	.99808
26	.827971	.911498	.958206	.98193	.99287	.99744
27	.811443	.900437	.951562	.97833	.99111	.99667
28	.794722	.888969	.944481	.97438	.98911	.99576
29	.777865	.877140	.936989	.97007	.98686	.99469
30	.760927	.864996	.929113	.96542	.98436	.99346
31	.743955	.852580	.920880	.96044	.98160	.9921
32	.726992	.839930	.912319	.95514	.97859	.9905
33	.710076	.827086	.903455	.94953	.97533	.9888
34	.693242	.814080	.894315	.94363	.97182	.9868
35	.676519	.800946	.884924	.93745	.96807	.9847
36	.659934	.787713	.875307	.93101	.96407	.9824
37	.643512	.774409	.865487	.92432	.95985	.9799
38	.627273	.761059	.855487	.91740	.95540	.9773
39	.611234	.747687	.845327	.91027	.95074	.9744
40	.595413	.734313	.835029	.90293	.94587	.9714

For small n these equations can be solved by iteration, which was done in constructing Table 1. Initial conditions an $U_k(0) = 1$, $U_i(0) = 0$ for $i \neq k$. It might be noted that the $U_i(j + 1)$ are subtotals of the $U_i(j)$ so that the iteration proceeds very rapidly on an adding machine. The probability that $d \leq k/n$ is $[U_0(n) + U_1(n) + U_2(n) \cdots + U_k(n)]n!n!/(2n)!$.

REFERENCES

[1] W. Feller, "On the Kolmogorov-Smirnov theorems," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 177–189.

[2] F. Massey, "A note on the estimation of a distribution function by confidence limits," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 116–119.

[3] F. Massey, "A note on the power of a non-parametric test," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 440–443.

[4] N. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bulletin Mathématique de l'Université de Moscou*, Vol. 2 (1939), fasc. 2.

[5] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 279–281.

A NOTE ON THE SURPRISE INDEX

By R. M. Redheffer

Harvard University

Let $p_m (m = 0, 1, 2, \cdots)$ be a set of probabilities of events E_m, and suppose that the event E_i, with probability p_i, actually occurred. Is the fact that E_i occurred to be regarded as surprising? In a recent article [1] this question is answered by introducing the surprise index S_i,

$$S_i = (\Sigma p_m^2)/p_i, \tag{1}$$

which gives a comparison between the probability expected and that actually observed.[1] The event is to be regarded as surprising when S_i is large.

The author remarks on the difficulty of computing (1) for the Poisson and binomial distribution. The problem consists in evaluating the numerator, which we shall express here in terms of tabulated functions. The Poisson case leads to Bessel functions, the binomial case to Legendre or hypergeometric functions, and the asymptotic behavior involves square roots only.

1. *The Poisson case.* For the Poisson case we have $p_m = \lambda^m e^{-\lambda}/m!$ so that the generating function is

$$e^{-\lambda}e^{\lambda x} = \Sigma p_m x^m. \tag{2}$$

Let $x = e^{i\theta}$, then $e^{-i\theta}$; multiply; integrate from 0 to 2π; and simplify slightly to obtain

$$\Sigma p_m^2 = (e^{-2\lambda}/\pi) \int_0^\pi e^{2\lambda \cos\theta}\, d\theta. \tag{3}$$

[1] Cf. also [6].