DISTRIBUTIONS OF A M. KAC STATISTIC[1]

By J. L. Allen and J. A. Beekman

Ball State University

0. Summary. In 1949, M. Kac defined a statistic which appears useful for statistical problems arising in insurance, biology, and telephone engineering [2]. In those fields, the natural observation period is a fixed time period during which a random number of observations would be obtained. The distribution of the number of observations is given by a Poisson distribution. Distributions of this statistic can be used to determine upper and lower confidence contours for an unknown distribution, or in testing a distribution hypothesis. The purpose of this note is to extend the authors' earlier results, [1], to the two-sided Kac statistic.

1. Introduction. Let $N, X_1, X_2, \cdots$ be independent random variables, N having a Poisson distribution with mean λ and each X_i having the same continuous distribution function $F(y)$. Let $\psi_y(x)$ be 0 or 1 according as $x > y$ or $x \leqq y$. A modified empirical distribution function was defined by M. Kac [2] as

$$F_\lambda^*(y) = \lambda^{-1} \textstyle\sum_{j=1}^{N} \psi_y(X_j), \qquad -\infty < y < \infty, \tag{1.1}$$

where the sum is taken to be zero if $N = 0$. Notice that it is possible for $F_\lambda^*(y)$ to exceed one. The statistic analogous to the two-sided Kolmogorov statistic [3] is $\text{lub}_{-\infty<y<\infty} |F(y) - F_\lambda^*(y)|$ and will be referred to as the two-sided Kac statistic. It is noted [2] that as long as $F(y)$ is continuous, the distribution of the statistic is independent of $F(y)$. Hence we will confine our attention to the case $F(x) = x$, $0 \leqq x \leqq 1$.

A random sample will determine a confidence band for the unknown distribution $F(y)$:

$$F_\lambda^*(y) - k/\lambda < F(y) < F_\lambda^*(y) + k/\lambda, \ k \text{ a positive integer} < \lambda. \tag{1.2}$$

Remark. Recently L. Takács, [6], has derived the exact distributions for statistics based on (1.1) and also the empirical distribution function $N^{-1} \sum_{j=1}^{N} \psi_y(X_j)$, $-\infty < y < \infty$.

2. The distribution of the two-sided Kac statistic.

Theorem 1. *Assume that $N, X_1, X_2, \cdots$ satisfy the hypotheses of Section 1. Let $J(\lambda)$ be an integer such that $P[N > J] \leqq \delta$, for appropriately small δ. Let*

$$P_\lambda(k) = P\{\text{lub}_{-\infty<y<\infty} |F(y) - F_\lambda^*(y)| < k/\lambda\}, \tag{2.1}$$

Received 22 December 1966.

[1] This research was partially supported by the National Science Foundation through grant NSF GP-4570.

$k = 1, 2, \cdots, \lambda$, *where* λ *is a positive integer. Then*

$$P_\lambda(k) \doteqdot \sum_{n=\lambda-k+1}^{\min(J,\lambda+k-1)} U_{n,n-\lambda+k}(\lambda)e^{-\lambda} \tag{2.2}$$

where $U_{n,j}(m + 1)$, $j = 1, 2, \cdots, 2k - 1$, $m = 0, 1, \cdots, \lambda - 1$, *satisfy the equations*

$$U_{n,j}(m + 1) = \sum_{h=1}^{j+1} U_{n,h}(m)/(j + 1 - h)! \tag{2.3}$$

with

$$U_{n,h}(m) = 0 \quad \text{if} \quad h \geqq 2k, \quad \text{or if} \quad h + m > n + k. \tag{2.4}$$

Furthermore, the $U_{n,i}(t)$ *satisfy the boundary conditions*

$$\begin{aligned} U_{n,i}(0) &= 0 \quad \text{for } i \neq k, \\ U_{n,k}(0) &= 1 \quad \text{for} \quad i = k, \\ U_{n,i}(t) &= 0 \quad \text{for} \quad i + t > n + k. \end{aligned} \tag{2.5}$$

The error in approximation (2.2) *is at most* δ. *For* $J \geqq \lambda + k - 1$, *the error is zero.*

Proof. Using the distribution-free property of the statistic, the independence of N, X_1, X_2, $\cdots$, the definition of $J(\lambda)$, and the value of $\sum_{i=1}^n \psi_u(X_i)$ at $u = 1$,

$$\begin{aligned} P_\lambda(k) &= P\{\max_{0 \leqq u \leqq 1} |u - \lambda^{-1} \textstyle\sum_{i=1}^N \psi_u(X_i)| < k/\lambda\} \\ &= \textstyle\sum_{n=0}^\infty P\{\max_{0 \leqq u \leqq 1} |u - \lambda^{-1} \sum_{i=1}^n \psi_u(X_i)| < k/\lambda\} P[N = n] \\ &\doteqdot \textstyle\sum_{n=\lambda-k+1}^{\min(J,\lambda+k-1)} P\{\max_{0 \leqq u \leqq 1} |u - \lambda^{-1} \sum_{i=1}^n \psi_u(X_i)| < k/\lambda\} \\ &\quad \cdot e^{-\lambda}\lambda^n/n! \\ &= \textstyle\sum_{n=\lambda-k+1}^{\min(J,\lambda+k-1)} U_{n,n-\lambda+k}(\lambda)e^{-\lambda} \text{ by the following lemma.} \end{aligned} \tag{2.6}$$

Lemma. *For n such that* $\lambda - k + 1 \leqq n \leqq \lambda + k - 1$ *and* λ *an integer,*

$$\begin{aligned} P_n(k/\lambda) &= P\{\max_{0 \leqq u \leqq 1} |u - \lambda^{-1} \textstyle\sum_{i=1}^n \psi_u(X_i)| < k/\lambda\} \\ &= (n!/\lambda^n) U_{n,n-\lambda+k}(\lambda), \qquad k = 1, 2, \cdots, \lambda, \end{aligned} \tag{2.7}$$

where $U_{n,j}(m + 1)$, $j = 1, 2, \cdots, 2k - 1$, $m = 0, 1, \cdots, \lambda - 1$, *satisfy the system of equations* (2.3), (2.4), *and* (2.5). *We also have*

$$P_n(k/\lambda) = 0 \quad \text{for} \quad n < \lambda - k + 1 \quad \text{or} \quad n > \lambda + k - 1. \tag{2.8}$$

Remark. The condition on n also implies that $n - \lambda + k \geqq 1$, so that $U_{n,n-\lambda+k}(\lambda)$ is well defined.

Proof. The proof of this lemma involves only minor changes in Massey's proof on page 117, [4]. We let α_i be the number of observations falling in the interval $[(i - 1)/\lambda, i/\lambda]$, $i = 1, 2, \cdots, \lambda$. Then $\sum_{i=1}^\lambda \alpha_i = n$. Letting $U_{n,n-\lambda+k}(\lambda)$ be the sum of the terms $(\alpha_1 ! \cdots \alpha_\lambda !)^{-1}$, $\sum_{i=1}^\lambda \alpha_i = n$, such that $\lambda^{-1} \sum_{i=1}^n \psi_u(X_i)$,

TABLE 1

$P_\lambda(k/\lambda) = \beta,\ k \leqq \lambda$

κ	λ = 1	2	3	4	5	6
1	.36788	.13534	.04979	.01832	.00674	.00248
2		.69923	.53106	.40447	.30845	.23534
3			.84887	.75285	.66479	.58646
4				.91866	.86655	.81194
5					.95350	.92542
6						.97218

κ	λ = 7	8	9	10	15	20
1	.00091	.00034	.00012	.00005	.00000	.00000
2	.17960	.13708	.10462	.07985	.02068	.00536
3	.51733	.45642	.40276	.35545	.19048	.10213
4	.75818	.70676	.65825	.61282	.42805	.29909
5	.89313	.85873	.82360	.78861	.62806	.49792
6	.95680	.93805	.91684	.89398	.77250	.65893
7	.98281	.97415	.96324	.95038	.86857	.77841
8		.98915	.98410	.97768	.92798	.86205
9			.99305	.99002	.96224	.91756
10				.99550	.98081	.95253
11					.99042	.97351
12					.99525	.98557
13					.99765	.99227
14					.99886	.99590
15					.99945	.99784
16						.99888
17						.99942
18						.99971
19						.99986
20						.99993

κ	λ = 25	30	35	40	45	50
1	.00000	.00000	.00000	.00000	.00000	.00000
2	.00139	.00036	.00009	.00002	.00001	.00000
3	.05476	.02936	.01574	.00844	.00453	.00243
4	.20905	.14613	.10215	.07141	.04992	.03490
5	.39462	.31278	.24793	.19654	.15581	.12352
6	.56032	.47614	.40456	.34375	.29209	.24820
7	.69269	.61485	.54527	.48343	.42857	.37993
8	.79258	.72525	.66220	.60403	.55072	.50202
9	.86470	.80957	.75528	.70328	.65417	.60816
10	.91460	.87170	.82701	.78241	.73895	.69719
11	.94775	.91592	.88064	.84382	.80668	.77000
12	.96891	.94633	.91958	.89025	.85953	.82827
13	.98194	.96657	.94704	.92446	.89983	.87395
14	.98970	.97963	.96587	.94905	.92988	.90904
15	.99422	.98782	.97844	.96629	.95179	.93544
16	.99679	.99283	.98662	.97809	.96742	.95492
17	.99823	.99583	.99182	.98599	.97834	.96901
18	.99903	.99760	.99507	.99118	.98582	.97901
19	.99948	.99863	.99706	.99451	.99084	.98598
20	.99972	.99923	.99826	.99663	.99416	.99076
21	.99985	.99957	.99891	.99795	.99632	.99398
22	.99992	.99976	.99941	.99876	.99770	.99612

$0 \leqq u \leqq 1$, reaches $(1, n/\lambda)$ by a route within a k/λ band of u, $0 \leqq u \leqq 1$, we obtain (2.7).

The slightly different boundary conditions on U are easily verified. Finally, (2.8) applies since such n's provide too few or too many points to be within the band at $u = 1$.

The asymptotic distribution for the statistic has been derived by Kac in [2], and is now stated.

THEOREM 2. *For N, X_1, X_2, $\cdots$ subject to the previous conditions, and $\alpha > 0$,*

$$(2.9) \quad \lim_{\lambda\to\infty} P\{\mathrm{lub}_{-\infty<y<\infty} |F(y) - F_\lambda^*(y)| < \alpha/\lambda^{\frac{1}{2}}\} = (4/\pi) \sum_{k=0}^{\infty} [(-1)^k/(2k+1)] \exp[-(2k+1)^2\pi^2/8\alpha^2].$$

REMARK. As in [1], it is easy to show that the two-sided Kac statistic produces a consistent test. One first obtains a lower bound on the power of the test, using an idea of Massey, [5]. One completes the proof as in the proof of Theorem 3, [1].

TABLE 2

$P_\lambda(\epsilon/\lambda^{\frac{1}{2}}) = \beta$

ϵ	λ								
	15	20	25	30	35	40	45	50	$\lim_{\lambda\to\infty}$
.4	.011	.004	.001	.006	.006	.004	.003	.002	.0006
.5	.019	.028	.028	.022	.015	.019	.021	.020	.0092
.6	.076	.071	.055	.063	.063	.058	.053	.056	.0414
.7	.141	.128	.132	.127	.123	.125	.124	.119	.1027
.8	.214	.216	.209	.210	.209	.205	.206	.205	.1852
.9	.306	.304	.302	.301	.299	.298	.297	.296	.2776
1.0	.398	.393	.395	.391	.391	.389	.389	.389	.3708
1.1	.480	.482	.477	.480	.476	.477	.475	.475	.4593
1.2	.558	.557	.560	.556	.557	.555	.556	.554	.5404
1.3	.633	.629	.627	.628	.626	.626	.625	.625	.6130
1.4	.689	.690	.693	.689	.688	.689	.687	.688	.6770
1.5	.745	.744	.743	.743	.744	.742	.743	.741	.7328
1.6	.791	.791	.793	.790	.789	.790	.789	.788	.7808
1.7	.829	.829	.829	.829	.830	.829	.828	.829	.8217
1.8	.866	.865	.865	.863	.862	.862	.863	.862	.8563
1.9	.890	.890	.890	.890	.890	.891	.890	.889	.8851
2.0	.913	.914	.915	.914	.913	.912	.912	.913	.9090
2.1	.933	.931	.931	.931	.931	.931	.932	.931	.9285
2.2	.946	.947	.948	.947	.947	.947	.946	.946	.9444
2.3	.959	.959	.958	.958	.958	.958	.958	.959	.9571
2.4	.968	.968	.969	.968	.968	.968	.969	.969	.9672
2.5	.977	.976	.975	.976	.976	.976	.976	.976	.9752
2.6	.981	.981	.982	.982	.982	.982	.982	.982	.9814
2.7	.985	.986	.986	.986	.986	.986	.986	.986	.9861
2.8	.989	.989	.990	.990	.990	.990	.990	.990	.9898
2.9	.992	.992	.992	.992	.992	.992	.992	.992	.9925
3.0	.993	.994	.994	.994	.994	.994	.994	.994	.9946

3. Distribution tables. Part of Table 1 was computed on the Ball State University IBM 1620 digital computer. The most time consuming calculations of Table 1 were computed at the Research Computer Center, Indiana University, on a Control Data 3400/3600 computer. The authors wish to especially acknowledge the help of that center. Twenty-five places were kept in the calculations. The results were then rounded to five places. The δ involved in (2.2) was chosen equal to .0001. Hence, the total error estimate is $\leqq 1.5 \times 10^{-4}$.

Table 2 was derived by linear interpolation from Table 1, and indicates the convergence of the true distribution to the asymptotic distribution. The mild oscillation was caused by the interpolation process.

REFERENCES

[1] Allen, J. L., and Beekman, J. A. (1966). A statistical test involving a random number of random variables. *Ann. Math. Statist.* **37** 1305-1311.

[2] Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proc. Nat. Acad. Sci. U. S. A.* **35** 252-257.

[3] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.* **4** 1-11.

[4] Massey, F. J. (1950). A note on the estimation of a distribution function by confidence limits. *Ann. Math. Statist.* **21** 116-119.

[5] Massey F. J. (1950). A note on the power of a non-parametric test. *Ann. Math. Statist.* **21** 440-443.

[6] Takács, Lajos. (1965). Applications of a ballot theorem in physics and in order statistics. *J. Roy. Statist. Soc. Ser. B* **27** 130-137.