NON-PARAMETRIC ANOVA IN SMALL SAMPLES; A MONTE CARLO STUDY OF THE ADEQUACY OF THE ASYMPTOTIC APPROXIMATION.

K. R. Gabriel
Princeton University and
Hebrew University, Jerusalem

and

P. A. Lachenbruch
University of North Carolina

Institute of Statistics Mimeo Series No. 569

NON-PARAMETRIC ANOVA IN SMALL SAMPLES: A MONTE CARLO STUDY OF THE ADEQUACY OF THE ASYMPTOTIC APPROXIMATION.

K. R. GABRIEL
Princeton University and
Hebrew University, Jerusalem

and

P. A. LACHENBRUCH
University of North Carolina

1. Introduction

Two k sample generalizations of the Wilcoxon-Mann-Whitney test are the Kruskal-Wallis (KW) [1952] and the Dwass-Steel (DS) [1960, 1961] tests. The former uses a sum of squares "between" mean ranks and is applicable for samples of any sizes. The latter uses the maximum of the mean rank differences of all pairs of samples (using separate rankings for each pair) applicable only if the sample sizes are equal. Asymptotically, the distribution of the KW statistic is chi-square with k-1 degrees of freedom whilst the DS statistic has the distribution of a range of k Normal variables. These distributions follow from the fact that the Wilcoxon-Mann-Whitney statistics are asymptotically Normal.

Little is known about the small sample distribution of the two statistics. For k=3 and various sample sizes up to 5, Kruskal and Wallis [1952] have computed exact probabilities for their statistic. They found the chi-square approximation to be conservative, i.e., the true probability of acceptance was greater than the $1-\alpha$ obtained from the chi-square distribution. For sample sizes above 5, Kruskal and Wallis recommended that the KW test be used with the asymptotic chi-square distribution.

For any two samples the ranks sum to $n(2n+1)$, a constant, so the absolute rank mean difference is a decreasing function of the smaller rank sum. Hence the maximum rank mean difference is inversely related to the minimum rank sum (MRS). Steel [1961] has tabulated percentage points for the MRS, anything not exceeding these points to be declared significant. These tables use the expectation and variance of rank sums, $E(RS)=n(2n+1)/2$ and $1/2\ \mathrm{Var}(RS)=n^2(2n+1)/24$, and the range distribution W_k of k Normals (the 1/2 factor enters

because the rank sums play the role of differences of sums.) Now, the value of MRS cannot be less than $n(n+1)/2$ - this occurs if all the observations of one sample are smaller than all those of another - but the "range" critical value

$$n(2n+1)/2 - W_{k,1-\alpha}n\sqrt{(2n+1)/24}$$

may, for some values of k,n and α, be less than that. In those cases the above critical values are clearly conservative, never allowing rejection, and Steel has therefore omitted them, providing no critical values for these cases. One wonders whether the range approximation might not also be conservative in other cases. Steel has checked the approximation used in his tables against exact distributions for $k=3$ and sample sizes $n=3(1)$ 6 and against Monte Carlo samples for a number of values of k (3, 4, 5, and 6) and n. He found a few deviations for $\alpha=0.10$ with exaggerated probabilities of rejection. For $\alpha=0.05$ he found no deviations, but for $\alpha=0.01$ he noted the tabluated values to be conservative.

2. The Monte Carlo Study.

Additional information on the true percentage points of the two statistics was obtained in the present study by Monte Carlo sampling. At any given combination of k(number of samples) and n(size of each sample), the integers 1, 2, ..., kn were randomly allocated into k sets of n and the two statistics computed for each random allocation. The principal purpose of the sampling was to determine the proportion of the allocations whose statistics are acceptable at the α level of their asymptotic distributions. The adequacy of these asymptotic approximations is checked by seeing how closely that observed proportion is to the given $1-\alpha$.

Random allocations were generated on computers by means of programs written and run independently by the two authors. Table 1 shows the number of allocations generated by the first author on the IBM 7040 of the Hebrew University, Jerusalem and the number generated by the second author on the IBM 360/75 of the University of North Carolina. The use of independent programs and computers by the two authors provided a check on the methods employed by each author.

Cross checking of the two authors' results could be carried out at each (k,n) combination at which both generated random allocations. The check was made by comparing the proportions of statistics acceptable with respect to the asymptotic 95% point. For the KW statistic 38 comparisons were possible: among these one showed no difference between the two authors' proportions, 16 showed a higher proportion for one author, 21 for the other. For the DS statistic there were only 28 comparisons: among these in 18 one author had higher proportions, and in 10 the other. Thus one could not say that the proportions obtained by either author were consistently or significantly higher than those of the other. Despite the fact that there was a rather larger frequency of extreme comparisons than expected, (For the KW statistic 9 out of the 38 comparisons were 5% significant, and 6 out of 28 were 5% significant for the DS statistic) the lack of systematic differences suggested that both sets of random allocations could be used to estimate the same percentages. Hence all proprotions acceptable were pooled and the rest of the data and discussion refers entirely to the pooled proportions.

The sampled proportions acceptable for KW statistics are shown in Figures 1, 2, and 3 for asymptotic 90%, 95%, and 99% points, respectively. Similarly, Figures 4, 5, and 6 show the respective proportions for the DS statistic -

entering asterisks wherever the statistic cannot attain the asymptotic point (as noted in section 1, above).

To obtain a general picture of the behavior of the proportion acceptable as a function of k and n, for any given $1-\alpha$, quadratic regression surfaces were fitted on k and n. The arc-sine-root transformation was used and proportions weighted according to the number of Monte Carlo allocations on which they were based.

For the KW statistic all sampled (k,n) points were entered in the regression computation. For the DS statistic, the computation included only points in which the true proportion acceptable was not 100%.

The fit of the quadratic surface was apparently close, though the tests for lack of fit were significant. The regression surfaces should therefore be regarded as fair approximations rather than as estimates of true quadratic surfaces. An idea of the shape of the regression surfaces and of their fit can be obtained from the contours plotted in the Figures for various proportions acceptable. Extrapolation of these contours and surfaces outside the range investigated here must be done most cautiously because such a procedure is well known to be most susceptible to random errors.

3. Results for the Kruskal-Wallis test.

The sample proportions acceptable for $\alpha=0.10$, 0.05, and 0.01 as given in Figures 1, 2, and 3, respectively, somewhat exceed $1-\alpha$ for almost the entire domain investigated. For $\alpha=0.10$ and $\alpha=0.05$ this excess is generally not above 2% and indeed not above 1% for most of that domain. The excess is most pronounced for low values of n and disappears for large n as expected from the asymptotic theory. A slight increase in the proportions acceptable

for n=20, 25 is suggested by the contours for α=0.05 and 0.01. Since there are only few sampled points with such values of n, this rise may well be due to random variability or to the inadequacy of the quadratic surface, and not indicate a real increase for large n. The effect of k, if any, is less obvious, except perhaps for α=0.10 where the excess seems to increase with increasing number of samples k.

The excesses at $1-\alpha$=0.95 are relatively more severe than those at $1-\alpha$=0.90, both being of the order of 2% at n=4. For $1-\alpha$=0.99 the excesses are relatively even greater, amounting to some 0.9% at n=4. The general pattern of excesses for $1-\alpha$=0.99 is otherwise similar to that for $1-\alpha$=0.95, clearly decreasing with increasing n and apparently independent of k.

These findings agree well with those of Kruskal and Wallis (1952) in confirming that the test is conservative for small n. The extent of this conservative bias is most serious for α=0.01 where the true α may be closer to 0.005 even with n as large as 8, whereas for α=0.05 and α=0.10 sample sizes of n=8 already ensure that the bias be no more than 1%.

4. Results for the Dwass-Steel test.

The sample proportions acceptable for asymptotic α=0.10, 0.05, and 0.01, as shown in Figures 4, 5, and 6, respectively, exceed $1-\alpha$ more seriously for the DS statistic than for the KW statistic. However, here the excesses depend both on k and on n. For any given k the bias decreases as n increases, as expected from asymptotic theory. But, for given n, the biases increase with k, so that they are greatest for large k and small n. For k=10, n=10, the biases are of the order of 3%, 2 1/2%, and 0.9% for α=10%, 5%, and 1%, respectively (compared with biases of 1%, <1%, and 0.7%, respectively, with the KW statistic). For other values of k and n, the biases are also considerable. This suggests that the use of the range distribution for this test is most dubious.

The reason for the increasing inadequacy of the asymptotic distribution as k increases may be that the range statistic is sensitive to extreme deviations. It is tabulated for the most extreme difference among several Normal variables, each of which may be arbitrarily large (albeit with small probability). Its present application, on the other hand, is to differences between mean ranks whose variation is bounded. For pairs of such variables, as in the Wilcoxon-Mann-Whitney test, the difference is apparently small and the Normal approximation good. But as the number of variables increases, the chance increases that some Normal pair will deviate much more than a pair of rank means can. Hence the range approximation to the DS statistic becomes increasingly poorer.

There is need for an alternative and less biased method of obtaining the significance level of a minimum rank sum (MRS) when the range approximation is inadequate. It is proposed that the probability α of randomly falling short of a given MRS be evaluated by solving the regression equation

$$2 \sin.^{-1}\sqrt{p} = .70397 + .04957K - .15828N + .20243m$$
$$-.00137K^2 - .00323N^2 - .00327m^2$$
$$-.00246KN + .00181Km$$

where m=MRS/10. The accuracy of this estimate may be gauged by the standard error of estimate which is 0.0366 (The multiple R^2 is 0.8250). This regression equation is based on the Monte Carlo allocations of the present study and presents an empirical approximation to a most intractable probability integral formula. (A check of Steel's table of critical values by means of this approximation confirms that the true significance level is almost always below the nominal one stated in that table, with the true α being on the average about one half of the nominal α. The only cases where the regression

estimate of α exceeds the nominal α are for k=3, 4 with the lowest n value tabulated by Steel, i.e., n=4, α=0.10; n=5, α=0.05; and n=7, α=0.01).

5. Conclusions

The Monte Carlo study has confirmed that the asymptotic chi-square approximation to the distribution of the KW statistic is slightly conservative. For α=0.10 and 0.05 the conservatism is slight unless $n \leq 6$. For α=0.01, the conservatism is serious if $n \leq 8$. Apart from these cases one would feel justified in using the chi-square approximation for this statistic. It should be stressed, however, that the validity of the present conclusions has been established only for equal sample rises.

The asymptotic range approximation to the distribution of the DS statistic has been found to be overly conservative. The extent of this conservatism is such that nominal significance levels computed from the range distribution may be much higher than true significance levels. This difference is so large as to make it doubtful whether the range distribution may be safely used in significance testing with level α=0.10 or 0.05 unless there are few samples and these are of size 15 or more. At level α=0.01, the range distribution should not be used even for samples of size 25.

The use of the DS statistic with the range distribution for multiple comparisons, as proposed by Steel [1961], is seriously impaired by the behavior of the statistic with increasing number of samples k. The more samples there are to be compared, the more seriously does the DS statistic's distribution appear to deviate from the range distribution. It is doubtful whether one may use this technique even approximately except for relatively few quite large samples.

Instead of the asymptotic range distribution one may use an empirical approximation formula to estimate the significance level of a DS statistic.

After this paper was completed, the work of Dunn-Rankin and Wilcoxon [1966] was brought to our attention. They provide exact values for the DS statistic for some values of k and n.

Acknowledgements:

This work was supported in part by the National Institute of Health, Grant Number GM-12868-03 and in part by the Office of Naval Research under Contract Nonr-1858(OS). The authors gratefully acknowledge computer time made available by the Computer Centers of the Hebrew University, Jerusalem and the University of North Carolina, Chapel Hill.

References

Dunn-Rankin, P. and Wilcoxon, F. [1966] "The true distribution of the range of rank totals in the two-way classification." Psychometrilia, pp. 573-580.

Dwass, Meyer [1960] "Some k-Sample rank-order tests", in Contributions to Probability and Statistics (Ed. I. Olkin et al), p. 198-202. Stanford University Press, Stanford.

Kruskal, William H. and Wallis, W. Allen [1952] "Use of ranks in one-criterion analysis of variance", Journal of the American Statistical Association, 47, 583-621.

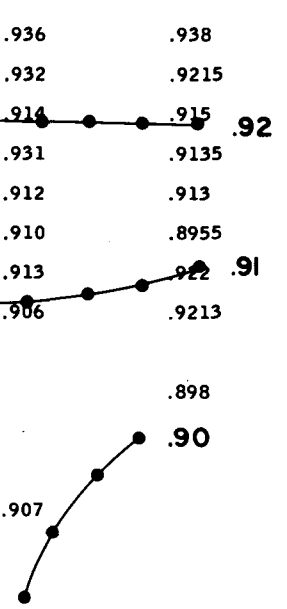
Steel, Robert G. D. [1961] "Some rank sum multiple comparisons tests", Biometrics, 17, 539-52.

Table 1: Number of Random Allocations Generated at Each Number of Samples - k - and size of Sample - n. Upper Number - generated in Jerusalem, lower number - generated in North Carolina

n \ k	3	4	5	6	7	8	9	10	12	15
3	1600	1600	1600	1600	1600	--	--	1000	--	1000
	--	--	--	--	--	--	--	--	--	--
4	1600	1600	1600	1600	1600	--	--	1000	--	1000
	--	--	--	--	--	--	--	--	--	--
5	1600	1600	1600	1000	1000	--	--	1000	--	1000
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6	1600	1000	1000	1600	1600	--	--	1000	--	1000
	1000	1000	1000	1000	1000	2000	1000	1000	1000	1000
7	--	--	--	--	--	--	--	--	--	--
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
8	3200	1000	1000	1000	1000	--	--	1000	--	1000
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
9	--	--	--	--	--	--	--	--	--	--
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
10	1600	1000	1000	--	500	--	--	500	--	500
	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
12	1600	1000	1000	--	500	--	--	500	--	500
	1000	1000	1000	1000	1000	1000	1000	1000	--	--
15	--	--	--	--	--	--	--	--	--	--
	--	--	--	--	--	--	--	--	1000	--
16	1600	1000	1000	500	500	--	--	500	--	--
	2000	2000	2000	2000	2000	2000	1000	--	--	--
20	1000	1000	500	500	--	--	--	--	--	--
	--	--	--	--	--	--	--	--	--	--
25	1600	500	500	--	--	--	--	--	--	--
	--	--	--	--	--	--	--	--	--	--

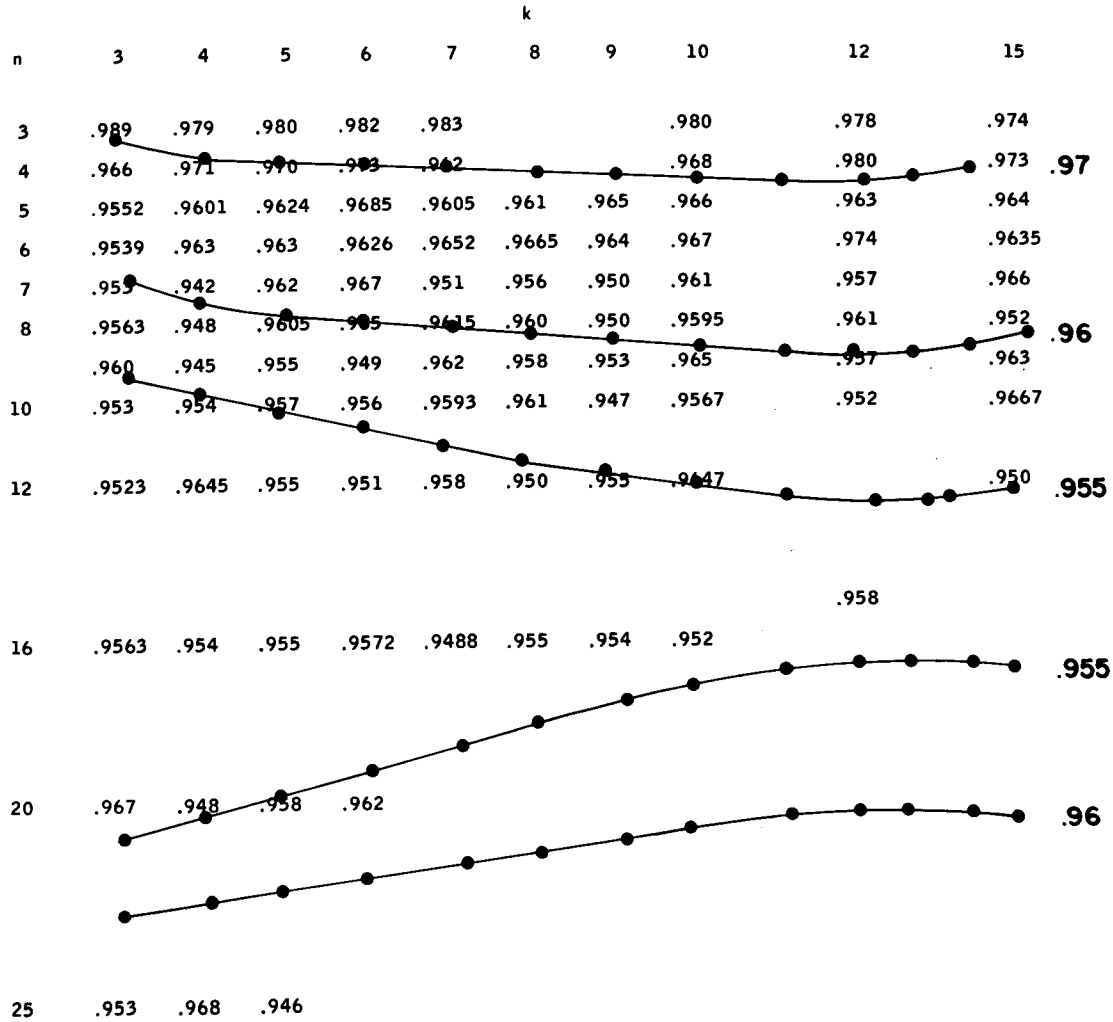
Figure 1

n \ k	3	4	5	6	7	8	9	10	12	15
3	.890	.918	.931	.932	.931			.935	.936	.938
4	.899	.918	.919	.918	.908			.920	.932	.9215
5	.9045	.9089	.9058	.9165	.9085	.914	.923	.915	.914	.915
6	.9020	.9075	.909	.9073	.9139	.911	.903	.9215	.931	.9135
	.908	.890	.904	.908	.890	.912	.900	.906	.912	.913
8	.9031	.8935	.910	.9105	.914	.907	.890	.913	.910	.8955
	.904	.8970	.898	.890	.908	.918	.904	.913	.913	.922
10	.9056	.8900	.902	.890	.916	.910	.904	.916	.906	.9213
12	.9109	.909	.910	.908	.8993	.892	.913	.920		.898
16	.9021	.9043	.9037	.906	.900	.9045	.906	.900		
20	.919	.891	.922	.930						
25	.899	.900	.904							



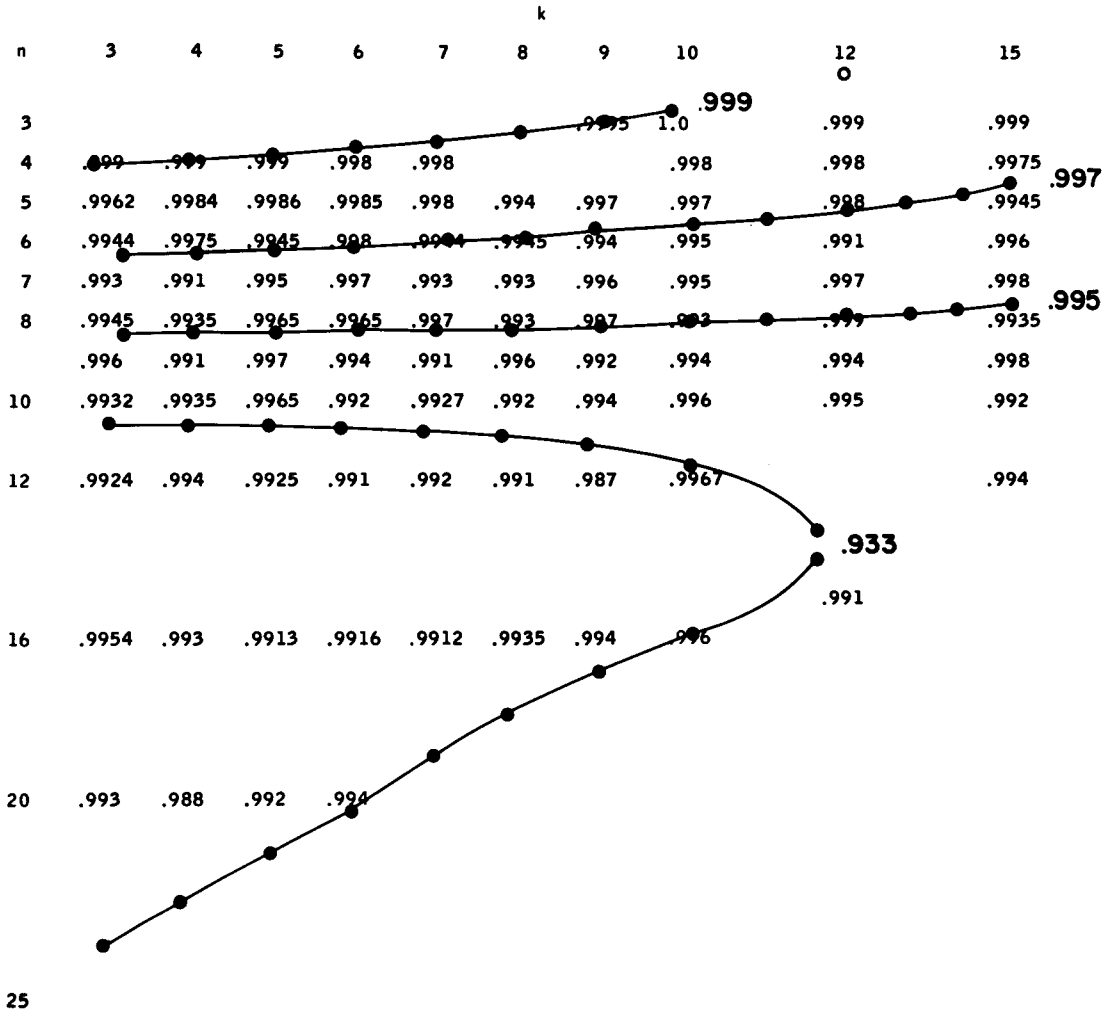
Proportions of KW statistics acceptable at $1-\alpha=0.90$ point of the chi-square distribution, with fitted quadratic contours.

Figure 2



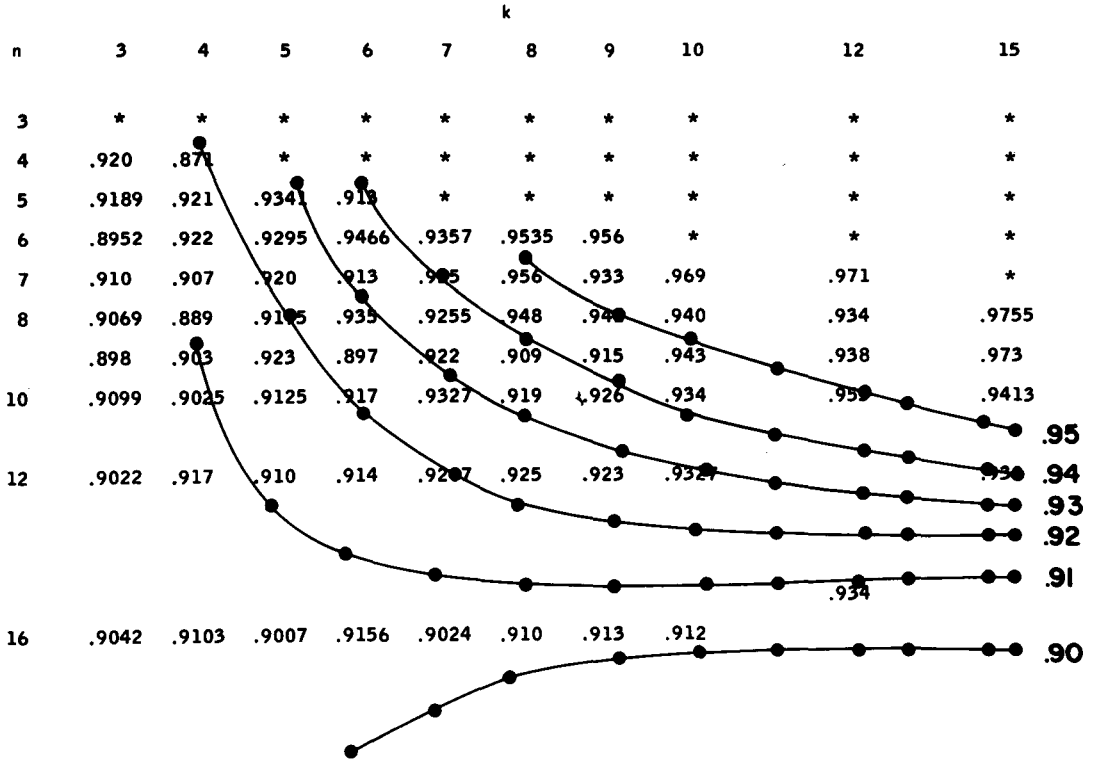
Proportions of KW statistics acceptable at $1-\alpha=0.95$ point of the chi-square distribution, with fitted quadratic contours.

Figure 3



Proportions of KW statistics acceptable at $1-\alpha=0.99$ point of the chi-square distribution, with fitted quadratic contours.

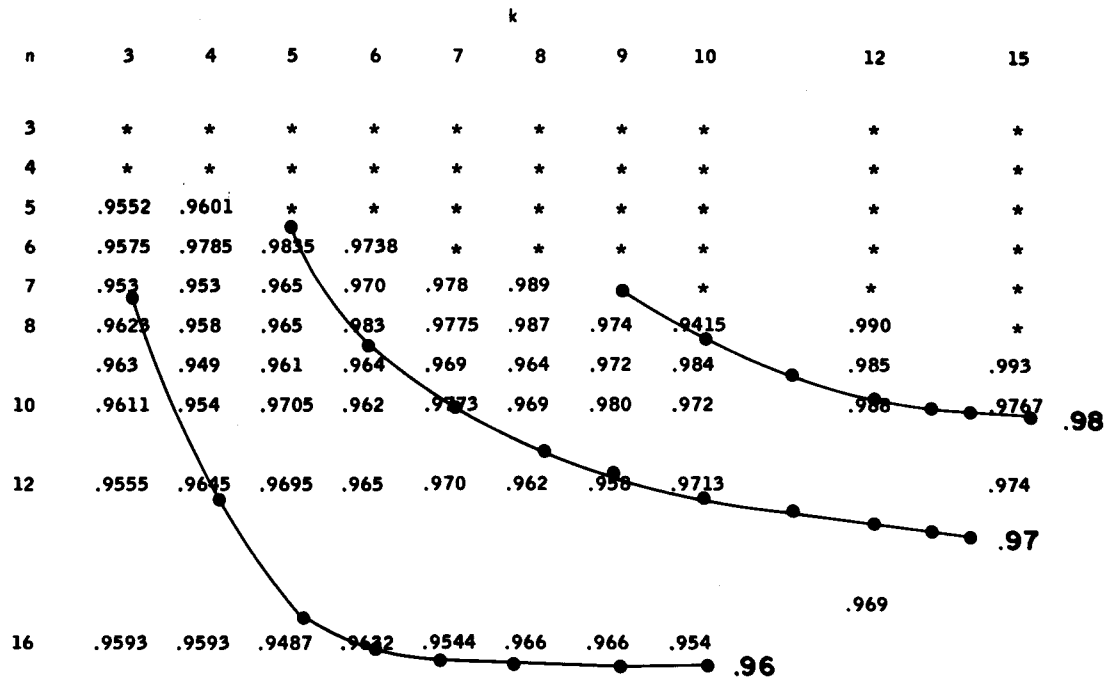
Figure 4



Proportions of DS statistics acceptable at $1-\alpha=0.90$ point of the range distribution, with fitted quadratic contours.

* No rejections possible at such k, n and α.

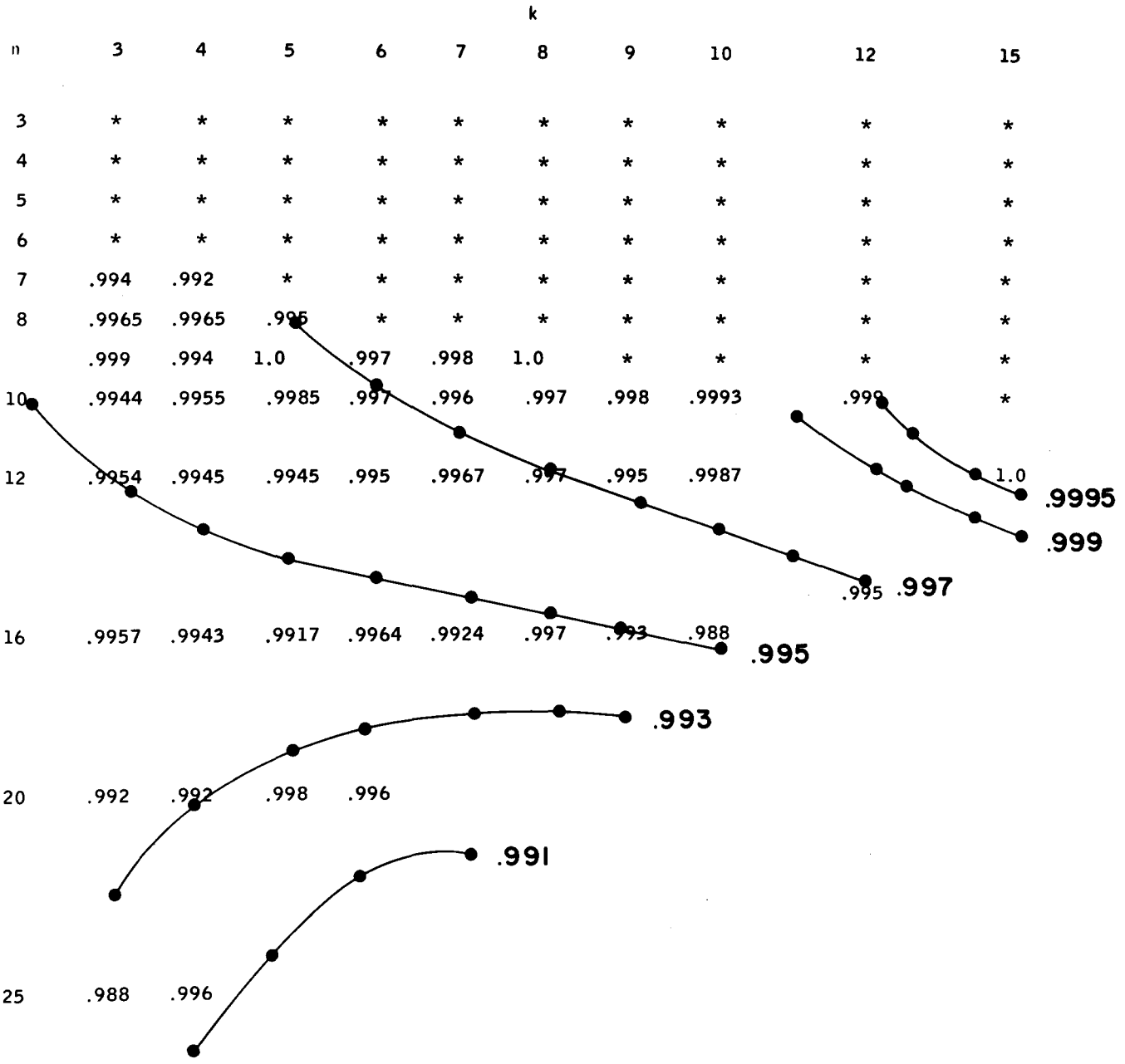
Figure 5



Proportions of DS statistics acceptable at 1- =0.95 point of the range distribution, with fitted quadratic contours.

* No rejections possible at such k, n and α.

Figure 6



Proportions of DS statistics acceptable at 1- =0.99 point of the range distribution, with fitted quadratic contours.

* No rejections possible at such k, n and α.