On Large-Sample Properties of Certain Non-Parametric Procedures*

by

Herman Rubin

Purdue University

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 238

September, 1970

*Research was supported in part by the Office of Naval Research Contract N00014-67-A-226-0008, project number NR042-216 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

On Large-Sample Properties

of Certain Non-Parametric Procedures*

by

Herman Rubin

Purdue University

Summary. Efficiencies of one-sided and two-sided procedures are considered from the standpoint of risk. It is shown that the two-sided Kolmogorov-Smirnov (K-S) and Kuiper procedures, which were shown in [4] to be asymptotically equiefficient with the median for translation alternatives for symmetric unimodal distributions have efficiencies for sample sizes in a wide range in the general vicinity of that of the median, but even if certain standard asymptotic approximations can be made, the efficiencies are not too close to that of the median, and in many cases the dominant asymptotic correction term does not even yield the sign of the deviation for samples of size 10^{20}.

A procedure briefly discussed in [1], for which the Pitman efficiency is zero, has good Bayes risk efficiency for translation alternatives for any distribution and merits further work for two-sided testing.

*Research was supported in part by the Office of Naval Research Contract N00014-67-A-226-0014, project number NR042-216 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

In the one-sided case, the one-sided K-S procedure appears to be somewhat worse to much worse than a procedure introduced by the author in [3]. Also, the K-S procedure involves a choice of significance level which is highly distribution dependent.

Introduction. We shall consider the "moderately large sample" efficiencies of certain well-known and not sufficiently well known non-parametric procedures from a decision-theoretic standpoint. By "moderatley large sample" we shall mean that central limit type theorems yield adequate approximations to the distributions involved, but that the further asymptotic approximations of the type in [4] are not necessarily very good. We shall also assume that the samples under consideration are sufficiently large that the large sample form of the risk can be used.

That is, we shall carry out our computations as if the observations can be considered as a stochastic process on [0,1] such that

$$X(t) = \theta\, h(t) + Y(t)\ , \tag{1}$$

where Y is a separable Gaussian process with mean 0 and covariance function

$$\Sigma(t,u) = 4(\min\,(t,u) - t\,u), \tag{2}$$

and $h(t)$ is a multiple of $fF^{-1}(t)$, chosen so that $h(.5) = 1$. The choice of these normalizing factors is for computational convenience; the median corresponds to $-X(.5)$ in standard units. For simulation purposes, we have chosen five distributions:

normal, with

$$h_N(t) = \exp(-.5\ \{[N(0,1)(t)]^{-1}\}^2)\ ; \tag{3}$$

logistic, with

$$h_L(t) = 4t(1-t)\ ; \tag{4}$$

double-exponential, with

$$h_D(t) = \begin{cases} 2t & t \leq .5\ , \\ 2(1-t) & t > .5\ ; \end{cases} \tag{5}$$

Cauchy, with

$$h_C(t) = \sin^2 \pi\, t\ ; \tag{6}$$

and a distribution with density $C/(1 + \tau|x - \theta|)^{10/9}$, in which case

$$h_T(t) = \begin{cases} (2t)^{10} & t \leq .5\ , \\ (2(1-t))^{10} & t > .5\ . \end{cases} \tag{7}$$

The loss structure was taken to be [2] $2|\theta|d\theta$ for a wrong decision in the one-sided testing problem -- it can be strongly argued that for "reasonably large" samples no other loss function is reasonable for this problem. For the two-sided problem the weight function was taken, as in [4], to be 1 if a type 1 error is made, and $|\theta|^k d\theta/\sqrt{2\pi}\,\mu_k$ for a type 2 error, where μ_k is the k-th absolute moment of the normal distribution. The choice of multiplicative constants was chosen so that if Z is $N(\theta,\sigma^2)$, then it will never pay to accept the null hypothesis if $\sigma > 1$, but for Z sufficiently small it will pay if $\sigma < 1$.

In the two-sided case these normalizations correspond to establishing a base for the sample size. In the one-sided case, if the value of the translation parameter at which there is indifference is θ^*, the risk is $E(\theta^{*2})$.

The procedures we have evaluated by Monte Carlo are, for the two-sided problem, Kolmogorov-Smirnov and Kuiper, and we have compared them to the median, for which it is known [4] that they are asymptotically equi-efficient. For the one-sided case, the Kolmogorov-Smirnov statistic has been compared with a symmetrized version introduced by the author in [3].

<u>Two-sided tests - asymptotic treatment</u>. The procedures that we shall consider are the median, Kolmogorov-Smirnov, and Kuiper. We shall also consider a test, suggested in [2], for which the Pitman efficiency is 0, but whose Bayes risk efficiency is that of the best order statistic. For the median (in our approximation: $X(.5) + \theta\, h(.5)$) the probability of exceeding C under the null hypothesis is

$$(8) \qquad P_M = \frac{2}{\sqrt{2\pi}} \int_c^\infty e^{-\frac{1}{2}t^2}\, dt \sim \frac{2}{\sqrt{2\pi}\, c}\, e^{-\frac{1}{2}c^2} .$$

For the K-S statistic ($\sup|X(t) + \theta h(t)|$) the corresponding probability is

$$(9) \qquad P_{KS} = 2\,\Sigma\,(-1)^{n-1}\,e^{-c^2 n^2/2} \sim 2e^{-c^2/2}\,,$$

and for the Kuiper statistic, the probability is

$$(10) \qquad P_K = 2\,\Sigma\,(n^2 c^2 - 1)e^{-c^2 n^2/2} \sim 2c^2\,e^{-c^2/2}\,.$$

(Note that there is a scale factor of 2 in the expressions for P_K and P_{KS}).

Now let us examine what happens under the alternative. Let $X^+(\theta) = \sup\,(X(t) + \theta\,h(t))$. For θ reasonably large, if $h(t) \sim 1 - \lambda\,|t - \frac{1}{2}|^\gamma$, $\gamma > \frac{1}{2}$, $X^+(\theta)$ is approximately $\theta + \lambda^{-1/\gamma}\,\theta^{-1/\gamma}\,Y_\gamma - Z$, where Z is normal $(0, 1)$ and Y_γ is a positive random variable whose distribution is not known except for $\gamma = 1$. Hence that θ for which $X^+(\theta) = c$ is approximately

$$(11) \qquad \theta_c = c + Z - \lambda^{-1/\gamma}\,c^{-1/\gamma}\,Y_\gamma.$$

Therefore

$$(12) \qquad E_{KS}(\theta_c{}^k) \sim c^k + \binom{k}{2}\,c^{k-2} - k\,K_\gamma\,c^{k-1-1/\gamma}\,.$$

For the Kuiper statistic we also need $X^-(\theta) = \inf\,(X(t) + \theta\,h(t))$. Here if θ is reasonably large and $h(t) \sim pt^\beta$, $\beta > \frac{1}{2}$, $X^-(\theta) \sim -p^{-1/\beta}\,\theta^{-1/\beta}\,W_\beta$, and

(13) $$E_K(\theta_c^{\,k}) \sim c^k + \binom{k}{2} c^{k-2} - kK_\gamma\, c^{k-1-1/\gamma} - kH_\beta\, c^{k-1-1/\beta} \,.$$

For the distributions we are considering, the values of γ and β are

	γ	β
normal	2	1^{+}
logistic	2	1
double-exponential	1	1
Cauchy	2	2
long-tailed	1	10 .

(For the normal, the tail behavior is slightly more complicated, but since for the Kuiper statistic the larger of β and γ is what counts, this is not a problem.)

Incidentally, in the case of the median,

(14) $$E_M(\theta_c^{\,k}) \sim c^k + \binom{k}{2} c^{k-2}.$$

Note that for the K-S and the Kuiper statistic, $E(\theta_c^{\,k})$ is smaller than for the median. However, the c required to obtain a given type I error is somewhat larger.

Now let us investigate what happens for m-th power loss for samples of size n if the cut-off point is c. We obtain for the type II risk

(15) $$R_2 = 2E(\theta_c^{\,m+1}/m+1)/\sqrt{2\pi}\,\mu_m n^{(m+1)/2}$$

Hence our combined risk is

$$(16)\quad R = P_I\ (c) + \left(\frac{c^{m+1}}{m+1} + \frac{m}{2}\ c^{m-1} - R\ c^{m-1/\gamma} + \ldots \right) B\ n^{-(m+1)/2}\ ,$$

where $P_I(c) \sim A\ c + e^{-c^2/2}$. Now a lengthy calculation shows that the dominant correction term to the asymptotic expression

$$(17)\qquad R \sim B(m+1)^{(m-1)/2} \left(\frac{\log n}{n} \right)^{(m+1)/2}$$

has the relative value

$$(18)\quad C_1 = \tfrac{1}{2}\ (q+1-m)\ \log\log n/((m+1)\ \log n)^{\frac{1}{2}},$$

which of course increases with q. Thus, for extremely large n, the median is better than the K-S test, which is better than the Kuiper test.

However, extremely large depends on log log n. Since $\log\log 10^{20} < 4$, for practical purposes the next term (which depends on A) comes into effect, and the $-R\ c^{m-1/\gamma}$ term may actually be dominant.

<u>Two-sided tests.</u> <u>Moderately large sample and empirical results.</u> A computation based on the likelihood ratio shows that for small n the K-S and Kuiper statistics are approximately equivalent to the best procedure. (This requires the probability of type I error to be nearly 1.) Apparently this efficiency drops off rapidly. Let us look at the results of Monte Carlo computations.

Efficiency (%)

		Kolmogorov-Smirnov			Kuiper		
		constant loss	absolute error	squared error	constant loss	absolute error	squared error
Normal	1	127	123	123	---	---	---
	2	129	126	119	70	72	72
	5	122	118	119	65	68	71
	10	119	119	118	68	70	73
	10^2	115	116	114	74	76	80
	10^3	117	114	113	75	81	85
	10^5	114	112	110	80	87	91
	10^{10}	112	109	107	88	93	96
	10^{20}	108	106	105	93	97	98
Logistic	1	111	114	116	---	---	---
	2	117	117	111	73	75	76
	5	114	112	113	69	72	74
	10	111	113	113	72	73	75
	10^2	110	112	111	75	78	81
	10^3	111	111	110	77	82	86
	10^5	110	109	108	82	88	91
	10^{10}	109	107	106	89	93	96
	10^{20}	107	105	104	94	97	98
Double-exp	1	---	---	---	---	---	---
	2	79	79	80	64	63	67
	5	78	80	82	60	61	62
	10	79	82	83	61	63	66
	10^2	82	86	87	64	68	72
	10^3	85	88	90	67	72	76
	10^5	88	91	93	72	78	82
	10^{10}	92	94	96	79	85	88
	10^{20}	94	96	97	86	90	92
Cauchy	1	---	---	---	---	---	---
	2	87	90	92	85	87	89
	5	88	91	91	80	84	87
	10	89	93	94	81	85	87
	10^2	93	97	98	84	89	92
	10^3	97	99	100	89	93	95
	10^5	98	101	101	92	96	98
	10^{10}	103	102	102	97	99	101
	10^{20}	102	102	102	99	101	102
Long-tailed	1	---	---	---	100+	105	112
	2	70	73	77	90	96	102
	5	66	65	65	78	86	91
	10	57	61	64	81	86	94
	10^2	61	68	73	81	93	100
	10^3	66	74	79	88	98	104
	10^5	74	82	86	94	103	107
	10^{10}	83	89	92	103	107	109
	10^{20}	93	94	95	106	108	108

The values are independent for the different distributions, but dependent within any one distribution. The standard deviations (estimated from 1000 sample processes) of these efficiencies are 1-2% for samples of size 10 and, with very few exceptions, .1 - .2% for samples of size 10^{20}. Thus, while individual figures for small sample sizes are not too reliable, the general picture is clear: for the Kolmogorov-Smirnov test, the flatness at the median determines the efficiency, and for samples of size 10^{20} relative to the base, the dominant asymptotic error has yet to make its presence felt.

The results are also similar for the Kuiper statistic. Several cases also clearly show the dip for small samples in the efficiencies. These results also agree with the exact calculations for K-S with 0-th power loss for the double-exponential in [3]. The optimal significance levels also are not much affected by the test.

A test occasionally considered (see, e.g. [1]) is to use $T_n = \sqrt{n} \sup \left| \frac{F_n - F}{\sqrt{F(1-F)}} \right|$. The statistic T_n is more sensitive to deviations in the tails than the K-S statistic. Now examination of

$$(19) \qquad T_n(x) = \sqrt{n} \; \frac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}}$$

by the usual methods shows that

$$(20) \qquad T_n \sim \sqrt{2 \log \log n}.$$

A further examination shows that the statistic cannot be very sensitive to Pitman alternatives since that x for which $T_n = |T_n(x)|$ is likely to be near 0 or 1. Of course,

$\sqrt{2 \log \log 10^{20}} < 2.15 \sqrt{2 \log \log 10}$, so that even this argument may not be too serious for reasonable sample sizes. But we note that for k-th power loss, for the K-S test the critical deviation is approximately $\frac{1}{2}\sqrt{(k+1) \log n}$, which grows much more rapidly. However, if we break the ordered observations below the median into groups of size 1, 2, 4, 8, ..., and examine the distribution of $T_n(x)$ in the corresponding intervals, we find that

$$(21) \qquad P(T_n > c) < K \log (n+1) e^{-\frac{1}{2}c^2} .$$

This shows that from the Bayes risk standpoint, this statistic bears much the same relationship to the best order statistic as the K-S or Kuiper statistic does to the median! This test consequently merits investigation.

<u>One-sided tests</u>. If the weight function is $2|\theta|d\theta$ for a wrong decision, and if the structural model is such that the observation $Y=\varphi(\theta,X)$, and for $\theta < \hat{\theta}(x)$ the decision is made that $\theta < 0$, the risk is $E[(\hat{\theta}(x))^2]$. This calculation can be applied in our model to the one-sided K-S tests and also to the symmetric test given by the author in [3]. The symmetric test has similar properties to the median for all symmetric unimodal distributions; its reciprocal efficiency relative to the median is between $2-\pi^2/6 = .355$ and a number bounded by $\left(\frac{\sqrt{\pi}}{3} + 1\right)^2 \sim 7.9$.

The one-sided K-S test does not fare so well. From equation (1), note that if θ is large the maximum of $X(t)$ will be large with large probability since $X(\frac{1}{2}) = \theta + Y(\frac{1}{2})$, but the minimum may still be quite negative if θ is not very large, especially if h is small some distance away from 0. This is indeed borne out by the empirical results. Unfortunately, it was not anticipated just how bad things would get, and hence it is necessary to crudely estimate some numbers. The results are as follows, with standard errors in parentheses:

Variance of Indifference Point

	Normal	Logistic	D-E	Cauchy	Long-tail
Symmetric	.735(.033)	.796(.034)	1.405(.061)	1.386(.062)	6.13(.19)
K-S, one-sided	.851(.027)	.942(.029)	1.737(.057)	1.902(.073)	~ 11(~3)
K-S, one-sided 50%	.852(.017)	.950(.029)	1.931(.07)	~ 4.5	~ 160
Optimal level	.49	.47	.39	.28	.008

Again, the values for the one-sided K-S at optimal level for the double exponential and the optimal level agree very well with the theoretical values of $\sqrt{14} - 2 = 1.74166$ and $e^{-\sqrt{.875}} \approx .39244$ respectively. Note that for not too bad distributions, the one-sided K-S test is fairly good if used at the optimal level, which varies considerably with the distribution. If the 50% level were used, as is optimal for any symmetric test statistic, the tail of the Cauchy is already bad enough to cause problems. It was not anticipated in the empirical procedure that when θ was chosen to make the maximum of $X(\theta)$ greater than 12.8 (or the minimum less than -12.8), which is beyond the 10^{-33} level for the Kuiper statistic that there would be any significant problems with the minimum (maximum). A few values for the Cauchy distribution were far enough out to give questionable accuracy at the 50th percentile; for the long-tailed distribution the figures given are probably slightly conservative.

Acknowledgments

The author wishes to express his thanks to Glennis Cohen for her assistance in producing the programs for the simulation and processing of the numerical results, and to Arthur Rubin for his invaluable help in the debugging of these programs.

References

[1] T. W. Anderson and D. A. Darling, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, Ann. Math. Statist Vol. 23 (1952) pp. 193-212.

[2] H. Chernoff, Sequential tests for the mean of a normal distribution, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1961, Vol. I, pp. 79 - 91.

[3] H. Rubin, Decision-theoretic evaluation of some non-parametric methods, Mimeo Series 193, Dept. of Statistics, Purdue University, Lafayette, Indiana. (To appear in Proceedings of the first Nonparametric Symposium).

[4] H. Rubin and J. Sethuraman, Bayes Risk efficiency, Sankhyā, A, Vol. 27 (1965) pp. 347 - 356.

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Nonparametric tests						
Kolmogorov-Smirnov test						
Kuiper test						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.