Open access • Journal Article • DOI:10.1093/BIOMET/65.1.13

Improving the jackknife with special reference to correlation estimation — Source link

David Hinkley

Published on: 01 Apr 1978 - Biometrika (Oxford University Press)

Topics: Jackknife resampling, Estimator and Asymptotic theory (statistics)

IMPROVING THE JACKKNIFE
WITH SPECIAL REFERENCE TO CORRELATION ESTIMATION

by

David V. Hinkley*

Technical Report No. 294

July, 1977

*Research supported in part by National Science Foundation Grant MCS77-00959.

IMPROVING THE JACKKNIFE
WITH SPECIAL REFERENCE TO CORRELATION ESTIMATION

by

David V. Hinkley

Technical Report No. 294

July, 1977

Some Key Words: Jackknife; Influence Function; Robustness; Residual; Correlation

SUMMARY

Second-order asymptotic properties of the jackknife procedure are discussed, and the jackknifed estimator is shown to be a vulnerable estimator whose variation can be severely underestimated by the jackknife standard error. Simple robust alternatives to the average pseudovalue are discussed. Particular emphasis is placed on estimation of a correlation coefficient. Numerical examples are given.

1. INTRODUCTION

The jackknife is a method for distribution-free bias reduction and standard error estimation. For a wide class of problems it is known that the jackknife produces consistent results. An excellent review of applications and asymptotic theory is given by Miller (1974). Recently there have been several investigations of small-sample properties of the jackknife procedure (Hinkley, 1977a, 1977b), which show that some adjustments are necessary in order to obtain accurate confidence intervals using the jackknife. In an unpublished paper, Efron has shown that the jackknife gives a rough (linear) approximation to another sub-sampling method for getting confidence intervals.

In the present paper we examine two further aspects of the jackknife, namely the use of second-order asymptotics in assessing finite-sample properties, and the use of jackknife pseudovalues in obtaining estimates less sensitive to extreme data points. The discussion is illustrated throughout with results for the correlation estimate.

A brief summary of the results is as follows: jackknifed estimators can have very large random bias compared to the original estimators; the jackknife estimate of standard error can severely underestimate the standard error of the jackknifed estimator; and that use of the jackknife pseudovalues in residual and trimmed-mean analyses can give considerably improved estimators.

Section 2 summarizes the standard jackknife method and illustrates it on an artificial data set, where certain difficulties are apparent. Second-order properties of the jackknife are derived in Section 3, and numerical results are given for the correlation example. The same example is used in Section 4, where robust analysis via pseudovalues is discussed. Section 5 gives brief conclusions.

Throughout the paper we assume that the basic estimate is obtained from independent, identically distributed random variables. Moreover we assume that the estimate is a regular differentiable functional of the empirical distribution function, with at least two derivatives.

2. THE JACKKNIFE: AN EXAMPLE

In general discussion we shall assume that $T_n = t(Y_1,\dots,Y_n)$ is an estimate of the parameter of interest θ, and that the Y_i are identically and independently distributed. The jackknife procedure may be briefly defined as follows. Let $T_{n,-i}$ denote the estimate computed from $Y_1,\dots,Y_{i-1}$, $Y_{i+1},\dots,Y_n$ for $i=1,\dots,n$. Then define the pseudovalues

$$P_{n,i} = n\,T_n - (n-1)\,T_{n,-i} \quad (i=1,\dots,n) . \tag{2.1}$$

An adjusted form of T_n is the jackknifed estimate

$$T_n^* = n^{-1} \sum P_{n,i} = \bar{P}_n . \tag{2.2}$$

If T_n has systematic bias of order n^{-1} then T_n^* has bias of smaller order. A distribution-free estimate of the standard error of T_n is

$$S_n = \sqrt{n^{-1} V_n} , \tag{2.3}$$

where

$$V_n = (n-1)^{-1} \sum (P_{n,i} - \bar{P}_n)^2 \tag{2.4}$$

is the sample variance of the pseudovalues. The standard error of T_n^* is also estimated by S_n. Under mild regularity conditions (Miller, 1974) one can show that both

$$(T_n - \theta)/S_n \quad \text{and} \quad (T_n^* - \theta)/S_n$$

are asymptotically standard normal as $n \to \infty$, so that approximate confidence intervals for θ can be obtained using a normal approximation to either pivot.

Throughout this paper we shall be concerned with the extent to which these asymptotic results are reliable in small samples and with the sensitivity of the jackknife to deviant data values. For illustration we use the sample correlation estimate, or rather its z-transform. Thus, if the Y_i are data pairs, we take

$$T_n = \tfrac{1}{2} \log_e\{(1+R_n)/(1-R_n)\} ,$$

where R_n is the sample product-moment correlation. In practice the z-transform is preferable because it is usually more nearly normal than R_n and because inadmissible values are possible for the jackknifed correlation.

To illustrate the jackknife procedure and to motivate the following discussion we use five artificial data sets illustrated in Figure 2.1. The samples have a basic set of 19 bivariate normal data points in common, and are distinguished by the twentieth data pair $(v,-v)$ which has $v = 0$, 0.5, 1.0, 1.5 and 2.0 respectively in samples 1-5. Table 2.1 gives the data and pseudovalues in the form $P_{n,i}-T_n$. The values of T_n, T_n^* and V_n are given in Table 2.2.

Table 2.1 Five bivariate samples of size n=20 and the corresponding sample influence values $I_-(Y_i) = (n-1)(T_n - T_{n,-i})$, $I_+(Y_i) = (n+1)(T_{n,+i} - T_n)$ for T = z-transformed correlation.

		$I_-(Y_i) = P_{n,i}-T_n$					I_+				
X_1	X_2	v=0	0.5	1.0	1.5	2.0	v=0	0.5	1.0	1.5	2.0
0.774	0.693	0.62	0.67	0.67	0.67	0.67	0.58	0.62	0.63	0.63	0.63
-1.325	-0.650	-1.39	-0.60	0.02	0.27	0.39	-0.89	-0.38	0.07	0.27	0.37
0.148	0.547	-0.83	-0.53	-0.20	-0.05	0.03	-0.71	-0.46	-0.18	-0.04	0.28
-1.567	-0.915	-0.52	0.11	0.61	0.82	0.92	-0.14	0.25	0.60	0.75	0.82
-0.553	-0.256	-0.15	-0.12	-0.06	-0.03	-0.01	-0.14	-0.12	-0.06	-0.03	-0.01
1.017	0.973	1.17	1.17	1.16	1.16	1.17	1.05	1.05	1.04	1.04	1.04
0.092	0.192	0.05	0.03	0.03	0.03	0.03	0.05	0.03	0.03	0.03	0.03
-1.211	-1.142	0.88	1.01	1.06	1.06	1.06	0.81	0.92	0.95	0.96	0.95
-1.264	-1.350	0.44	1.01	1.26	1.34	1.36	0.54	0.93	1.12	1.18	1.19
1.013	0.960	1.15	1.15	1.15	1.15	1.15	1.03	1.03	1.03	1.03	1.03
-0.447	-0.320	0.06	0.05	0.04	0.04	0.04	0.06	0.05	0.04	0.04	0.04
-0.917	-0.764	0.47	0.48	0.47	0.46	0.45	0.45	0.45	0.45	0.44	0.43
-0.841	-0.778	0.34	0.42	0.44	0.44	0.43	0.33	0.40	0.42	0.42	0.41
0.428	0.486	0.29	0.27	0.27	0.27	0.28	0.29	0.27	0.26	0.27	0.27
0.042	-0.223	-1.03	-0.40	-0.12	-0.04	-0.01	-0.93	-0.38	-0.12	-0.04	-0.01
1.017	1.032	1.22	1.20	1.20	1.21	1.22	1.09	1.07	1.07	1.08	1.08
0.020	-0.516	-3.87	-1.58	-0.55	-0.22	-0.10	-2.76	-1.36	-0.52	-0.22	-0.10
0.423	0.516	0.28	0.26	0.26	0.27	0.28	0.27	0.25	0.26	0.27	0.27
-0.164	0.129	-0.33	-0.27	-0.17	-0.11	-0.09	-0.31	-0.26	-0.16	-0.11	-0.09
v	-v	-0.04	-5.88	-13.54	-19.65	-24.46	-0.04	-3.65	-5.49	-6.11	-6.38

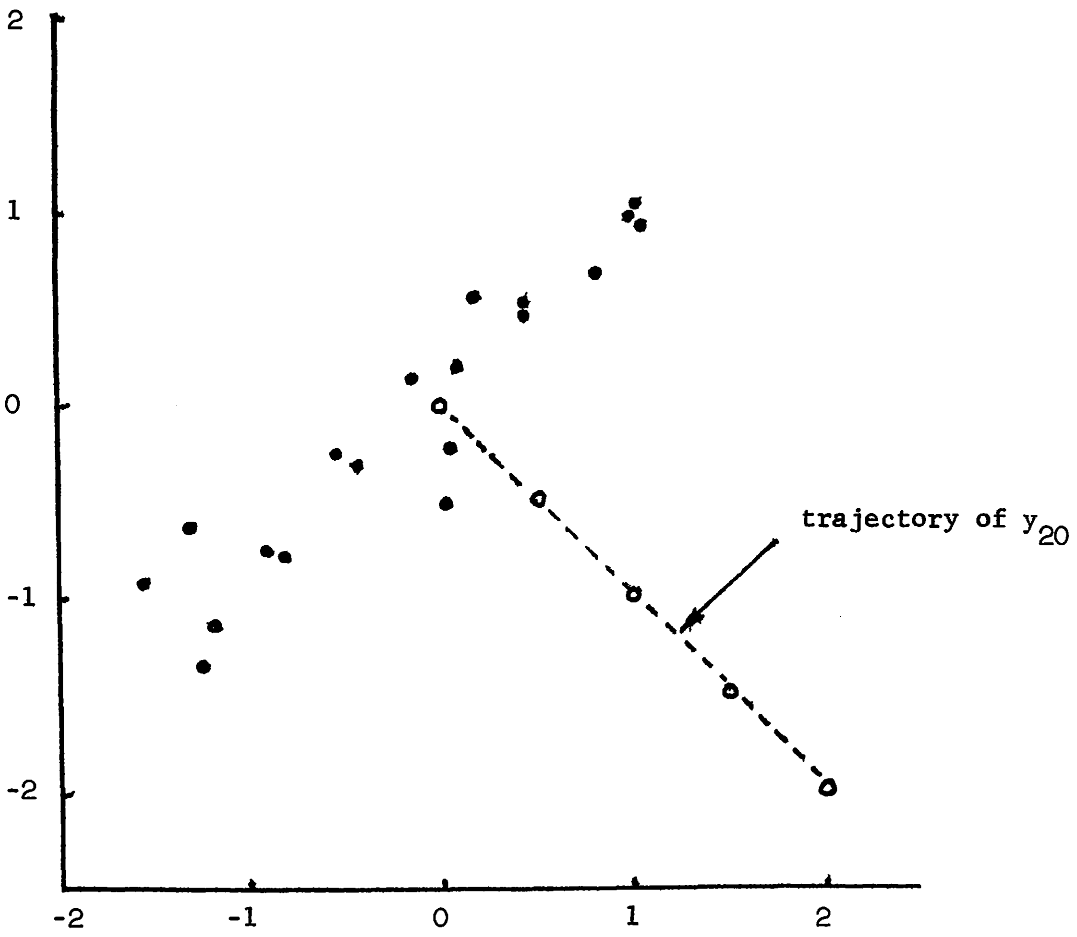

Figure 2.1. Five artificial bivariate samples of size $n=20$. All samples contain pairs represented by ● . Samples differentiated by values of y_{20} , represented by ○ . Nominal distribution of $y_1,\ldots,y_{19}$ is bivariate normal with correlation 0.95 and $N(0,1)$ marginals.

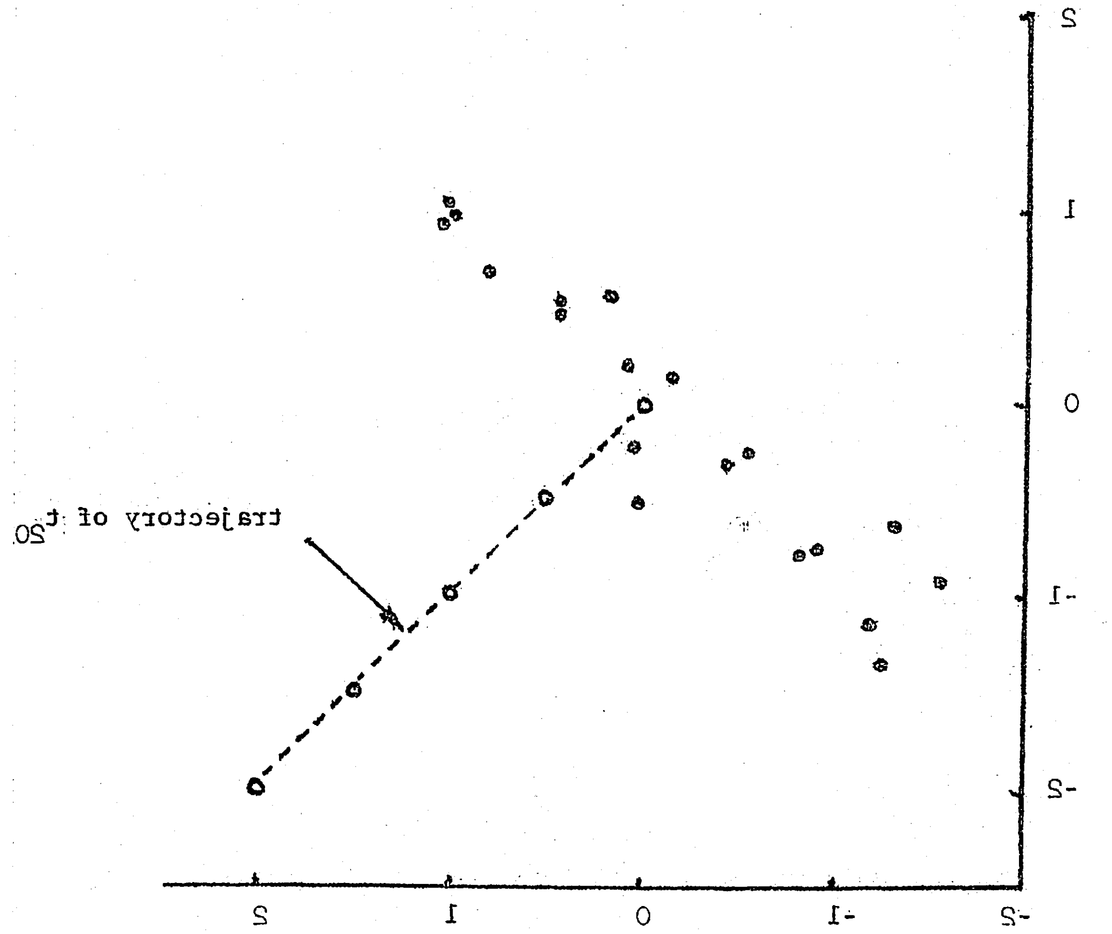

Figure 2.1. Five artificial bivariate samples of size n=20. All samples contain pairs represented by ● . Samples differentiated by values of y_{20} , represented by * . Nominal distribution of $y_1,\dots,y_{19}$ is bivariate normal with correlation 0.95 and N(0,1) marginals.

Table 2.2 Jackknife statistics for T = z-transformed correlation in the five samples given in Table 2.1.

	Twentieth data pair				
	(0,0)	(0.5,-0.5)	(1.0,-1.0)	(1.5,-1.5)	(2.0,-2.0)
T_n	1.76	1.45	1.05	0.73	0.48
T_n^*	1.70	1.38	0.75	0.18	-0.28
$T_n^*(.05)$ (see Section 4.2)	1.84	1.63	1.42	1.23	0.96
V_n	1.32	2.37	10.01	20.48	31.35
V_n^+ (see Section 3.2)	0.79	1.14	1.95	2.34	2.52

What is quite apparent from Table 2.2 is that T_n^* is more seriously affected by a deviant point y_{20} than is T_n. Both estimates are vulnerable to a very extreme value of y_{20}. The standard error estimate S_n defined by (2.3) is in close agreement with normal theory for $v=0$, but as v increases S_n becomes so large as to give fairly uninformative confidence intervals for θ via the pivot $(T_n^* - \theta)/S_n$. Similar behavior for the jackknifed correlation has been previously noted by Miller (1974) and Wainer and Thiessen (1975).

For long-tailed or contaminated data distributions the anomalies present in the above example would appear as random effects. In extreme cases the difficulty would be quite evident from suitable inspection of the data (Section 4), but in cases with weak contamination or non-normality this would not often be possible. It is therefore useful to see to what extent we can use theoretical arguments to appraise such small-sample behavior of T_n , T_n^* and V_n .

3. FURTHER PROPERTIES OF THE JACKKNIFE PROCEDURE

3.1 Theory

To some extent the inexactness of asymptotic properties can be understood and repaired by an examination of second-order properties. In the case of the jackknife, such an examination throws light on the observed finite-sample discrepancies between $\mathrm{var}(T_n)$, $\mathrm{var}(T_n^*)$ and their estimate $S_n^2 = V_n/n$. We make use of well-known expansion results for differentiable statistical functions; see, e.g., von Mises (1947).

Suppose that T_n is a regular differentiable statistical function of the form $T_n = t(\tilde{F}_n)$, where $\tilde{F}_n$ is the usual empirical distribution function. The corresponding parameter is $\theta = t(F)$ where F is the distribution function for Y . In fact, suppose that T_n admits the expansion

$$(3.1) \qquad T_n \sim \theta + n^{-1} \sum t_1(Y_j) + \tfrac{1}{2} n^{-2} \sum\sum t_2(Y_j,Y_k) \ .$$

Here $t_1(y)$ and $t_2(y,z)$ are equivalent to first and second Volterra, or von Mises, derivatives of the functional $t(F)$ and they may be defined by the identities

$$(3.2) \qquad \begin{aligned} \frac{d}{d\epsilon} t\ \{(1-\epsilon)F + \epsilon G\}_{\epsilon=0} &= \int t_1(y)dG(y) \ , \\ \frac{d^2}{d\epsilon^2} t\ \{(1-\epsilon)F + \epsilon G\}_{\epsilon=0} &= \iint t_2(y,z)dG(y)dG(z) \ , \end{aligned}$$

where G is an arbitrary probability distribution function. The first derivative $t_1(y)$ has the statistical name "influence function", whose role in statistics has been discussed in excellent papers by Hampel (1974), Jaeckel (unpublished) and Mallows (unpublished). Three important properties of the derivatives (3.2) are

$$(3.3) \qquad E_F\{t_1(Y)\} = 0 \ , \quad t_2(Y,Z) = t_2(Z,Y) \ , \quad E_F\{t_2(a,Y)\} = 0 \ \text{ for any fixed } a \ .$$

The jackknife pseudovalues $P_{n,i}$ defined in (2.1) give estimates of the influence function $t_1(y)$ at $y=Y_1,\ldots,Y_n$. Devlin et al (1975), following Mallows (unpublished), use the mnemonic notation

$$(3.4) \qquad I_-(Y_i) = (n-1)(T_n - T_{n,-i}) = P_{n,i} - T_n$$

for such estimates. A detailed analysis of these quantities is possible using all terms given in (3.1), which with (2.1) gives

$$(3.5) \qquad P_{n,i} \sim \theta + t_1(Y_i) + t_2^*(Y_i,Y) \ ,$$

where

$$(3.6) \qquad t_2^*(Y_i,Y) = \frac{2n \sum\limits_k t_2(Y_i,Y_k) - nt_2(Y_i,Y_i) - \sum\limits_{j}\sum\limits_{k} t_2(Y_j,Y_k)}{2n(n-1)} \ .$$

Before proceeding, we should note that the usual first-order properties of T_n, T_n^* and V_n follow from (3.1) and (3.5) by ignoring the terms involving $t_2(\ ,\ )$. Thus $n\ \mathrm{var}(T_n) \sim n\ \mathrm{var}(T_n^*) \sim V_n \sim \mathrm{var}\{t_1(Y)\}$ as $n \to \infty$. The final terms given in (3.1) and (3.5) are of order n^{-1}, and vanish only when T_n is a linear statistic.

The effect of the term $t_2^*(Y_i,Y)$ in (3.5) is to induce a "small" correlation between the pseudovalues, which in turn affects T_n, T_n^* and V_n in different ways. For simplicity we shall denote $t_1(Y_i)$ and

$t_2(Y_j, Y_k)$ by $D_{1,i}$ and $D_{2,jk}$ respectively. Then, with the definitions

(3.7) $\sigma_{11} = \mathrm{var}(D_{1,i})$, $\sigma_{12} = E(D_{1,i}D_{2,ii})$, $\sigma_{22} = \mathrm{var}(D_{2,jk})$, $j \neq k$,

routine calculation from (3.1), (3.5) and (3.6) leads to

$$\mathrm{var}(P_{n,i}) \sim \sigma_{11} + n^{-1}(\sigma_{12} + \sigma_{22})$$

$$\mathrm{cov}(P_{n,i}, P_{n,j}) \sim -n^{-2}(\sigma_{12} + \tfrac{1}{2}\sigma_{22})$$

(3.8a) $$n\ \mathrm{var}(T_n) \sim \sigma_{11} + n^{-1}(\sigma_{12} + \tfrac{1}{2}\sigma_{22})$$

(3.8b) $$n\ \mathrm{var}(T_n^*) \sim \sigma_{11} + \tfrac{1}{2}n^{-1}\sigma_{22}$$

(3.8c) $$E(V_n) \sim \sigma_{11} + n^{-1}(\sigma_{12} + \sigma_{22}) .$$

Whilst theoretical calculation of σ_{12} and σ_{22} is in principle straightforward, we have not followed this through in the correlation example. Rather we have pursued the more useful approach of estimating these quantities via repeated sample estimates. This has the advantage of simultaneously showing whether or not the theoretical differences in (3.8) can be corrected for in actual data analysis.

Fairly routine calculation from (2.1) shows that the following estimates are consistent:

(3.9a) $$\hat{D}_{1,i} = P_{n,i} - T_n$$

(3.9b) $$\hat{D}_{2,ii} = n\ \{(n+1)T_{n,+i} - 2n\ T_n + (n-1)\ T_{n,-i}\}$$

(3.9c) $$\hat{D}_{2,jk} = n\ \{n\ T_n - (n-1)(T_{n,-j} + T_{n,-k}) + (n-2)\ T_{n,-j-k}\}$$

for $j \neq k$. The notation is self-explanatory. Note that $\hat{D}_{2,ii}$ is expressible in terms of two estimates of $D_{1,i}$, namely $I_-(Y_i)$ and

(3.10) $\quad I_+(Y_i) = (n+1)(T_{n,+i}-T_n)$;

the latter has been suggested as an alternative to I_- by Mallows.

Estimates for σ_{12} and σ_{22} are formed from the derivative estimates by computing the corresponding sample moments. We have abbreviated the estimate of σ_{22} and worked with

$$\hat{\sigma}_{12} = \text{sample covariance of } D_{1,i} \text{ and } D_{2,ii}$$

$$\hat{\sigma}_{22} = \text{sample variance of } D_{2,i,i-1}$$

taking $D_{2,10} = D_{2,1n}$.

Note that the usual estimate V_n of σ_{11} could be replaced by the sample variance of $I_+(Y_i)$. Denoting this variance estimate by V_n^+ , it is straightforward to show that

(3.11) $\quad E(V_n^+) \sim \sigma_{11} + n^{-1}(\sigma_{22} + 3\sigma_{12})$.

Comparison with (3.8c) suggests that V_n is generally preferable to V_n^+ .

3.2 The Correlation Example

We return to the example of Section 2, taking T_n = z-transformed correlation. Table 2.1 gives both I_- and I_+ for each of the five samples. It should be apparent that the two estimates of D_1 agree well unless a data point is moderately deviant from the centre of the data. This indicates that for contaminated or long-tailed data distributions the values of $\hat{D}_{2,ii}$ will be large. The estimates for σ_{12} and σ_{22} in the first sample (v=0) are -5.8 and 3.4 respectively, which from (3.8) would indicate little difference between $\mathrm{var}(T_n)$ and $\mathrm{var}(T_n^*)$ for bivariate normal data. Even in this case V_n^+ will tend to underestimate both variances if (3.11) is accurate. As the pair y_{20} becomes more extreme, so do the estimates of σ_{12} and σ_{22}. Values of the estimates V_n and V_n^+ are given in Table 2.2, from which we might guess that the latter estimate is generally poor because it is too small.

To better assess the general conclusions that can be drawn from the second-order theory, we have obtained a few Monte Carlo results for the correlation case. Table 3.1 presents results for one case, where $n=20$, $\rho = 0.8$ ($\theta = 1.10$) and m data pairs are consistently multiplied by 3 ($m=0,1,2$). The obvious conclusions to be drawn from this and similar tables are: (i) the variance of T_n^* can be appreciably larger than that of T_n; (ii) V_n gives a good estimate of $\text{var}(T_n)$, whereas V_n^+ can give a severe underestimate of the variance; (iii) the differences are predictable in that σ_{12} is appreciably negative, the more so as data becomes more contaminated.

Table 3.1 Simulation results for jackknife statistics. Twenty bivariate normal pairs with $\rho = 0.8$ and $m = 0,1,2$ pairs multiplied by 3. T = z-transformed correlation. Results are from 1000 trials.

	m = 0		m = 1		m = 2	
statistic	mean	variance	mean	variance	mean	variance
T_n	1.124	0.0589	1.144	0.109	1.142	0.138
T_n^*	1.098	0.058	1.11	0.157	1.098	0.198
V_n	1.242	0.36	2.295	6.00	2.91	8.51
V_n^+	0.73	0.06	0.88	0.17	0.98	0.24
$\hat{\sigma}_{12}$	-6.00	24.9	-19.0	986	-26.5	1385
$\hat{\sigma}_{22}$	7.3	108	11.0	261	66.4	45255

The fine differences represented in (3.8) seem to be moderately good approximations. Thus the difference $\mathrm{var}(T_n^*) - \mathrm{var}(T_n) \sim n^{-2}\sigma_{12}$ holds up fairly well in Table 3.1. Also the approximation $E(V_n) - \mathrm{var}(T_n) \sim \frac{1}{2}n\, \sigma_{22}$ appears to be reasonably good. The approximations do not work as well for $n=10$, although they still give the right qualitative comparisons between $\mathrm{var}(T_n)$, $\mathrm{var}(T_n^*)$ and S_n^2 .

From a practical standpoint, we are interested in the ability to correct V_n and so obtain a more accurate standard error for T_n or T_n^* . It is apparent from Table 3.1 that the estimates of σ_{12} and σ_{22} can be quite imprecise, suggesting that such corrections cannot be made accurately. We have obtained a few Monte Carlo results which are somewhat more optimistic. We compared the basic standardized form of T_n^* , i.e., $(T_n^* - \theta)/S_n$, with the corrected form $(T_n^* - \theta)/\sqrt{\{S_n^2 - \frac{1}{2}n^{-2}(\hat{\sigma}_{22} + 2\,\hat{\sigma}_{12})\}}$ suggested by (3.8). Table 3.2 summarizes results of variance and coverage of nominal 90% confidence intervals derived from each form.

Table 3.2 Comparison of basic and adjusted standardized forms of T_n^* in the correlation case. Sample size n=20, bivariate normal data with m pairs multiplied by 3, $\rho = 0.8$. 1000 trials.

	variance of standardized form		coverage of nominal 90% confidence intervals	
	basic form	adjusted form	basic form	adjusted form
m=1	1.39	1.08	15%	10%
m=4	1.73	1.34	19%	13%
m=6	1.55	1.19	16%	12%

4. IMPROVING THE ESTIMATES USING PSEUDOVALUE ANALYSIS

4.1 General Theory

One aspect of the jackknife that was readily apparent from Table 2.2 is the sensitivity of T_n^* to outlying data pairs. This phenomenon is probably common if the large variance of T_n^* in Section 3 is a reliable indication. For a particular sample, as in the last two or three samples of Table 2.1, the poor value of T_n^* is well diagnosed by the extreme pseudovalue. Both Mallows and Devlin et al (1975) have alluded to the analogy between the influence estimates and residuals. From the results in Section 3 and the example of Section 2 we conclude that I_- is superior to I_+ as an estimate of t_1, and hence as a "residual". The analogy with linear-model residuals is not exact because of the non-zero second derivatives; indeed, this explains the difference between I_- and I_+.

As we have suggested, inspection of the pseudovalues, or I_- values, is a useful part of a jackknife analysis. If one or two extreme values are evident, then reanalysis with the corresponding data points omitted would be indicated. In certain cases, a probability plot of the pseudovalues will provide a useful way of determining whether or not T_n^* is a good estimate; for details see Devlin et al (1975). Here we consider another use of the pseudovalues that seems to be suitable when no extreme data points suggest themselves for deletion.

Our idea may be explained loosely as follows. The pseudovalue $P_{n,i}$ is like an observation on θ with "error" $t_1(Y_i)$, and the jackknifed estimate T_n^* is the arithmetic average of these observations. Now in certain situations $t_1(Y)$ has a symmetric distribution; see the correlation example in Section 4.2. Then there are consistent alternatives to $\bar{P}_n$ which are much less sensitive to deviation of the data distribution from that under which T_n is the preferred estimate. Actually a more accurate representation for $P_{n,i}$ is given by (3.1), but the above idea still bears fruit.

We have considered typical classes of robust alternatives to the average in conjunction with the "observations" $P_{n,i}$. One such class for which the theory is straightforward is the Huber M-estimates (Huber, 1972). Suppose that $T_n^*(c)$ is defined to be the unique solution to

$$n^{-1} \sum c(P_{n,i} - t) = 0 , \tag{4.1}$$

where $c(\ )$ is an odd monotone function. Then a straightforward calculation shows that symmetry of $t_1(Y)$ implies consistency of $T_n^*(c)$, so that Taylor expansion methods can be applied to (4.1) to show that

$$T_n^*(c) - \theta \sim \frac{n^{-1} \sum \{c(D_{1,i}) - t_2^*(Y_i, Y)c'(D_{1,i})\}}{n^{-1} \sum c'(D_{1,j})} . \tag{4.2}$$

Two things may be deduced from this relationship. Firstly, $T_n^*(c)$ has an asymptotic normal distribution by virtue of a Central Limit Theorem for the exchangeable random variables in the numerator summand; details are given in an as yet unpublished technical report by the author. Secondly, the influence function of $T_n^*(c)$ is

$$\frac{c(t_1(y)) - E\{t_2(y,Z)c'(Z)\}}{E\{c'(Z)\}} ,$$

which is unbounded in general. It is conjectured, but not proved, that other classes of robust estimates based on the pseudovalues have the same properties. In practice trimmed means of the pseudovalues are the easiest to use, and we consider these in the correlation case below.

From a purist's point of view estimates with unbounded influence are, by definition, not robust. However we take the view that estimates should only be insensitive to the type of departures from ideal conditions that cannot be detected by residual analysis. In any event, our purpose here is solely to show how to use the jackknife procedure to greater advantage.

It should be pointed out that the influence function itself may be used to derive estimates with bounded influence. One such instance is mentioned in Hinkley (1977b), and further research on this topic is being carried out by the author and H. Wang.

4.2 The Correlation Example

We return again to the example of the correlation estimate. For the five artificial samples in Table 2.1 the inspection of I_- values clearly indicates that y_{20} is an outlier when v exceeds 1.0; note that the I_+ values do not give such a clear indication. For all samples we have computed 5% trimmed means of the pseudovalues (i.e., averages of all but the largest and smallest $P_{n,i}$). The results are given in Table 2.2. These estimates seem to represent substantial improvement over T_n^* and T_n in cases other than the ideal $(v=0)$.

The trimmed mean pseudovalue has been evaluated further in Monte Carlo trials, some of which we describe here. First we should point out that the theory outlined in Section 4.1 applies here so long as the bivariate distribution has a density that depends only on the Normal exponent; otherwise trimmed means and M-estimates will not be consistent.

Table 4.1 gives Monte Carlo results for the 5% trimmed mean pseudovalue, T_n and T_n^* in sample size n=20 with $\rho = 0.6$. The results are very similar for other values of ρ. The first sampling situation is bivariate normal, and the second is bivariate normal with 10% of data pairs multiplied by 3. Both bias and variance of the estimates are given. Note that the trimmed mean sacrifices little in the ideal Normal case, but gives substantial improvement in the contaminated case. Other non-Normal distributions give similar qualitative results, but for very long-tailed distributions all three estimates are poor. In such cases extreme values of I_- occur. More detailed numerical results are given in the author's unpublished technical report.

Table 4.1 Monte Carlo bias and variance for T_n, T_n^* and the 5% trimmed mean pseudovalue in samples of size n=20 with $\rho = 0.6$. Results are based on 1000 trials.

		T_n	T_n^*	trimmed mean pseudovalue
bivariate normal	bias	0.009	-0.009	-0.006
	variance	0.056	0.055	0.057
bivariate normal with 10% multiplied by 3	bias	0.025	-0.001	-0.001
	variance	0.114	0.147	0.084

5. CONCLUSIONS

The jackknife method is a useful tool for reducing systematic bias and for calculating a distribution-free estimate of standard error. However, it appears that T_n^* may have a substantial random bias. Thus, the estimate T_n is to be preferred unless it is known that the random bias is small. For moderate sample sizes it may be possible to get a better estimate of $var(T_n^*)$ via estimates of the second derivatives as in Section 3.

When feasible, analysis of the influence estimates I_- is useful; an initial diagnostic might be the magnitude of $T_n - T_n^*$. An extreme value of I_- is probably grounds for removal of the corresponding data point and recalculation of the estimate T_n or T_n^*. The alternative influence estimate I_+ seems to be quite inferior both for residual analysis and for estimating standard errors. Routine deletion of smallest and largest pseudovalues before averaging produces a very definite improvement over T_n^* in the case of the correlation, and may do so in other cases. However, if a truly robust estimate is required, more complex methods, such as those devised by Devlin et al (1975) for the correlation, are needed.

ACKNOWLEDGEMENTS

This work was supported in part by a Public Health Service Grant at Stanford University in 1976, and in part by a National Science Foundation Grant. The author's work referred to in Section 4 is contained in Biostatistics Technical Report No. 19 at Stanford University. The author is grateful to Colin Mallows and Louis Jaeckel for the opportunity to see their important unpublished work on the influence function. Thanks are also due to Hai-Li Wang for useful discussions on some of the material.

REFERENCES

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika 62, 531-45.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. 69, 383-93.

Hinkley, D. V. (1977a). Jackknife confidence limits using Student t approximations. Biometrika 64, 21-8.

Hinkley, D. V. (1977b). Jackknifing in unbalanced situations. Technometrics 19, (to appear).

Huber, P. J. (1972) Robust statistics: a review. Ann. Math. Statist. 43, 1041-67.

Miller, R. G., Jr. (1974). The jackknife: a review. Biometrika 61, 1-15.

von Mises, R. (1947). On the asymptotic distributions of differentiable statistical functions. Ann. Math. Statist. 18, 309-48.

Wainer, H. and Thissen, D. (1975). When jackknifing fails (or does it?). Psychometrika 40, 113-4.