A COMPARISON OF THE JACKKNIFE AND BOOTSTRAP ESTIMATORS IN LINEAR MODELS WITH REFERENCE TO PRODUCTION MODELS USED BY SASOL

A THESIS SUBMITTED TO THE FACULTY OF SCIENCE, UNIVERSITY OF WITWATERSRAND, JOHANNESBURG, FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BY

STUART MAXWELL ANGUS

MAY 1988

I, THE UNDERSIGNED, HEREWITH DECLARE THAT THIS THESIS IS ENTIRELY MY OWN WORK AND HAS NOT BEEN PRESENTED FOR ANY DEGREE AT ANOTHER UNIVERSITY.

S. Angus

STUART MAXWELL ANGUS

ACKNOWLEDGEMENTS

I . . like to express my sincere appreciation for the support and assistance I received from friends, family and collegues who made this work possible.

I am particularly indebted to my supervisor, Dr C C Frangos, who introduced me to the jackknife and bootstrap techniques and whose patient guidance and encouragement assured the completion of this work.

I would like to thank my employer, Sasol Two, for its financial support during the four year period spent on completing this thesis.

The assistance received from all the technical staff of the Computer Department at Sasol Two is gratefully acknowledged.

Finally, I wish to thank Mrs N Hope for her extreme patience and devotion in typing the manuscript.

ABSTRACT

This thesis is based on research which has been carried out on two sub-sampling methods, namely the jackknife and the bootstrap. A typical problem in applied statistics involves the estimation of an unknown parameter θ. The two main questions asked are (1) What estimator $\hat{\theta}$ should be used? (2) Having chosen to use a particular $\hat{\theta}$, how accurate is it as an estimator of θ?. The jackknife and bootstrap are resampling methods for answering the second question. The jackknife is a simple but powerful method for bias reduction and distribution-free estimation of the variance. The bootstrap can be viewed as a closely related method of the jackknife and is used to generate sampling distributions of statistics and thereby to draw inferences about parameters.

Chapter 1 of this thesis is a brief survey of the research which has been carried out on the jackknife method and also under consideration are open questions suitable for further research. Similarly, in Chapter 2, a review of the bootstrap method is undertaken with future trends and possible new research topics discussed. In Chapters 3 to 5, three separate research areas are investigated mainly by Monte-Carlo simulation studies to evaluate the performance of the jackknife and bootstrap methods against the standard parametric methods. The areas under consideration are (1) component and system availability (2) non-linear regression models and (3) simple time-series models. These areas were chosen with particular reference to their applicability in industrial situations. Finally, in Chapter 6, the performance of the jackknife method is evaluated by considering several case studies which were undertaken at Sasol Two and Three between January 1984 and December 1987 using actual plant data.

C O N T E N T S

CHAPTER 1

OVERVIEW OF THE JACKNIFE METHOD IN STATISTICS

1.1 Derivation and description of the jackknife estimator

The jackknife is a member of the family of resampling techniques which are feasible today because of the availability of high speed computers. In particular, the jackknife is a method for bias reduction and distribution free interval estimation. The jackknife method was first introduced by Quenouille (1949), in relation to bias reduction of a serial correlation estimator in time series by splitting the sample into two half samples. In a subsequent paper, Quenouille (1956), generalised this idea into splitting the sample into g groups of size h each, such that n = gh, and formally defined the jackknife estimator as follows:

Let $Y_1 \ldots, Y_n$ be a sample of independent and identically distributed (i.i.d.) random variables, and let the real-valued parameter θ be associated with their distribution $F(x,\theta)$. Let $\hat{\theta}_n$ be an estimator of the parameter θ based on the sample of size n. The sample $Y_1 \ldots, Y_n$ is split into g groups, each of size h. Then, let $\hat{\theta}_{-i}$ be the corresponding estimator based on the sample of size (g-1)h, where the i^{th} group of size h has been deleted. The jackknife pseudovalues are defined as:

$$\tilde{\theta}_i = g\hat{\theta}_n - (g - 1)\hat{\theta}_{-i} \quad (i = 1,\ldots,g) \qquad (1.1.1)$$

and the jackknife estimator of θ is defined as follows:

$$\hat{\theta} = \frac{1}{g} \sum_{i=1}^{g} \tilde{\theta}_i \qquad (1.1.2)$$

If h = 1 and g = n, the jackknife estimator $\hat{\theta}$ has the form:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\theta}_i \qquad (1.1.3)$$

where $\tilde{\theta}_i = n\hat{\theta}_n - (n - 1)\hat{\theta}_{-i}$ (1.1.4)

The jackknife estimator $\hat{\theta}$ has the interesting property (Quenouille, 1956) that, if $\hat{\theta}_n$ is biased of order $1/n$, then $\hat{\theta}$ reduces the bias to order $1/n^2$.

i.e. if $E(\hat{\theta}_n) = \theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \ldots$ (1.1.5)

then,

$$E(\hat{\theta}) = n(\theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \ldots) - (n-1)(\theta + \frac{a(\theta)}{n-1} + \frac{b(\theta)}{(n-1)^2} + \ldots)$$

$$= \theta - \frac{b(\theta)}{n(n-1)} + \ldots. \qquad (1.1.6)$$

and hence $\hat{\theta}$ is biased to order $1/n^2$ only.

In general, the most popular version of the jackknife method involves one-at-a-time omissions, since this is considered to be more efficient than h-at-a-time omissions for $h \geq 2$ (Rao and Webster, 1966). However, with moderately large samples, and with the support of high speed computers, Mosteller and Tukey (1968), showed that the h-at-a-time omission scheme may possess valuable bias reduction properties and may also produce trustworthy confidence intervals for θ.

Tukey (1958) gave the name 'jackknife' to Quenouille's method, based on the idea that it would be a rough-and-ready statistical tool, which would be applicable in a variety of situations. Whereas Quenouille was concerned with the reduction of bias, Tukey's main objective was to obtain an estimator of variance. Hence, Tukey treated the pseudovalues $\widetilde{\theta}_i$ in (1.1.4) as if they were a sample of i.i.d. random variables with distribution function $F(\theta,x)$, and suggested the following estimator of the variance of the jackknife estimator $\hat{\theta}$:

$$\text{var}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\theta}_i - \hat{\theta})^2 \tag{1.1.7}$$

Tukey also conjectured that the statistic:

$$t = \frac{(\hat{\theta} - \theta)}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\theta}_i - \hat{\theta})^2}} \tag{1.1.8}$$

should have an approximate t distribution with (n - 1) degrees of freedom. Approximate confidence intervals and tests can then be based on the above proposal.

1.2 An example of a jackknife estimator

Let $X_i (i = 1,\ldots,n)$ be n i.i.d. random variables with cumulative distribution function $N(\mu,\sigma^2)$. An estimator of σ^2 is given by:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2, \tag{1.2.1}$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

It is clear that:

$$E\{\hat{\theta}_n\} = \sigma^2 - \frac{\sigma^2}{n}$$

The i^{th} partial estimator of σ^2 is given by:

$$\hat{\theta}_{-i} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j^2 - \left\{\frac{n\bar{X} - X_i}{n-1}\right\}^2$$

$$= \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j^2 - \frac{1}{(n-1)^2} (n^2\bar{X}^2 - 2nX_i\bar{X} + X_i^2)$$

Further, it can be shown that

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i} = \frac{\{(n-1)^2-1\}}{(n-1)^2 n} \sum_{i=1}^{n} X_i^2 + \frac{(2n-n^2)}{(n-1)^2} \bar{X}^2$$

$$= \frac{n(n-2)}{(n-1)^2} \hat{\theta}_n$$

Therefore,

$$E \left\{\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i}\right\} = \frac{n(n-2)}{(n-1)^2} E(\hat{\theta}_n) = \frac{(n-2)}{(n-1)} \sigma^2 = \sigma^2 - \frac{\sigma^2}{(n-1)}$$

The jackknife estimate of $\theta = \sigma^2$ is:

$$\hat{\theta} = n\hat{\theta}_n - \frac{(n-1)}{n} \sum_{i=1}^{n} \hat{\theta}_{-i} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 - \frac{n(n-2)}{(n-1)^2} \sum X_i^2 - \frac{n(n-2)}{(n-1)} \bar{X}^2$$

$$= \frac{(n^2-n-n^2+2n)}{(n-1)n} \sum_{i=1}^{n} X_i^2 - \frac{(n^2-n-n^2+2n)}{(n-1)} \bar{X}^2$$

$$= \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i-\bar{X})^2$$

Hence, the jackknife estimator of σ^2 is the minimum variance unbiased estimator of the variance in the case of the normal distribution.

1.3 Second order jackknife

The bias of a jackknife estimator $\hat{\theta}$, of order $1/n^2$, may be eliminated by jackknifing with weights n^2 the jackknife estimator. In Quenouille's paper (1956), the second-order jackknife estimator is defined as:

$$\hat{\theta}2 = \frac{n^2\hat{\theta}-(n-1)^2 \sum_{i=1}^{n} \tilde{\theta}_{-j}/n}{n^2-(n-1)^2}, \qquad (1.3.1).$$

where $\hat{\theta}$ is the jackknife estimator of θ from a sample of size n and $\tilde{\theta}_{-j}$ is the jackknife estimator of θ from a sample size (n - 1) with the j^{th} observation removed. Equation (1.3.1) can also be rearranged to express the second order jackknife estimator in terms of the original estimator $\hat{\theta}_n$ (see Miller, 1974a).

$$\text{i.e. } \hat{\theta}^{(2)} = (n-1)\{n^3\hat{\theta}_n - (2n^2-2n+1)(n-1)(\frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{-i}) +$$

$$(n-1)^2(n-2)\{\frac{2}{n(n-1)}\sum_{i<j}\hat{\theta}_{-ij}\}\}, \qquad (1.3.2)$$

where $\hat{\theta}_{-i}$ is the original estimator of θ from a sample of size (n-1) with the i^{th} observation removed and $\hat{\theta}_{-ij}$ is the original estimator of θ from a sample of size (n-2) with the i^{th} and j^{th} observations removed.

If $E(\hat{\theta}_n) = \theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2}$, then $E(\hat{\theta}^{(2)}) = \theta + \theta(1/n^3)$, but $\hat{\theta}^{(2)}$ is not unbiased. Schucany, Gray and Owen (1971), suggested modifying the weights to achieve complete unbiasedness when the bias has only first and second-order terms in 1/n. Their estimator, which has simpler weights then Quenouille's (1.3.1) is

$$\hat{\theta}^{(2)*} = \frac{1}{2}\{n^2\hat{\theta}_n - 2(n-1)^2(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}) + (n-2)^2\{\frac{2}{n(n-1)}\sum_{i<j}\hat{\theta}_{-ij}\}\} \qquad (1.3.3)$$

1.4 The generalised jackknife

In order to handle more general forms of bias, Schucany, Gray and Owen (1971), generalised the jackknife technique. Suppose there are two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ with expectations

$$E(\hat{\theta}_1) = \theta + f_1(n).b(\theta)$$

and $E(\hat{\theta}_2) = \theta + f_2(n).b(\theta)$

Then, the estimator

$$\hat{\theta}^* = \frac{\begin{vmatrix} \hat{\theta}_1 & \hat{\theta}_2 \\ f_1(n) & f_2(n) \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ f_1(n) & f_2(n) \end{vmatrix}} \qquad (1.4.1)$$

is referred to as the second order generalised jackknife

Let $R = \dfrac{f_1(n)}{f_2(n)} \neq 1$ and $f_2(n) \neq 0$

then,

$$E(\hat{\theta}^*) = \frac{E(\hat{\theta}_1) - R.E(\hat{\theta}_2)}{1 - R}$$

$$= \frac{\theta(1-R) + f_1(n) - Rf_2(n)}{1 - R}$$

$$= \theta$$

Hence, the estimator $\hat{\theta}^*$ is completely unbiased. The standard jackknife estimator, as given in (1.1.2), for g = n, can be considered as a partial case as follows:

i.e. put $\hat{\theta}_1 = \hat{\theta}_n$, $\hat{\theta}_2 = \sum \hat{\theta}_{-i}/n$, $f_1(n) = 1/n$ and $f_2(n) = 1/(n-1)$.

The same authors (Gray et el.-1972) also considered k separate terms in the bias, each of which factorized into distinct functions of n and θ. The expectation of each estimator is of the form:

$$E(\hat{\theta}_i) = \theta + \sum_{j=1}^{k} f_{ij}(n)\, b_j(\theta) \quad (i = 1,\ldots,k+1) \tag{1.4.2}$$

The k^{th} order generalised jackknife estimator of θ is defined as follows:

$$\hat{\theta}^* = \frac{\begin{vmatrix} \hat{\theta}_1 & \cdots & \hat{\theta}_{k+1} \\ f_{11}(n) & & f_{k+1,1}(n) \\ \vdots & & \vdots \\ f_{1k}(n) & \cdots & f_{k+1,k}(n) \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 \\ f_{11}(n) & & f_{k+1,1}(n) \\ \vdots & & \vdots \\ f_{1k}(n) & \cdots & f_{k+1,k}(n) \end{vmatrix}} \tag{1.4.3}$$

The effect of the first and second order jackknife estimators on a general bias term has been investigated by Adams, Gray and Watkins (1971). In a subsequent paper (1972), the same authors highlighted an interesting relationship between the jackknife technique and the e_1-transformation which is used in numerical analysis for increasing the speed of convergence of a series. Given a slowly converging series of numbers

$$S_n = \sum_{i=1}^{n} a_i \quad , \tag{1.4.4}$$

the transformation

$$e_1(S_n) = \frac{S_n - \rho(n).S_{n-1}}{1 - \rho(n)} \quad , \qquad (1.4.5)$$

for $\rho(n) = a_n/a_{n-1} \neq 1$, will increase the rate of convergence to the limit S_∞ in many cases. The analogy is $S_n \sim E(\hat{\theta}_n)$, $S_{n-1} \sim E(\sum \hat{\theta}_{-i}/n)$, $S_\infty \sim \theta$ and $\rho(n) \sim (n-1)/n$. The jackknife estimator is the linear extrapolation to $0 = 1/\infty$ from $\hat{\theta}$ plotted at $1/n$ and $\sum\hat{\theta}_{-i}/n$ at $1/(n-1)$.

1.5 Jackknifing the ratio estimator

Ratio estimation occupies an important place in sample surveys, and has become an area of application for the jackknife technique since the simple estimator $\bar{Y}/\bar{X}$ is biased. Given a sample (X_i, Y_i) $(i=1,\ldots,n)$ of paired random variables with $E(X_i) = \mu$ and $E(Y_i) = \eta$, the problem is to estimate $\theta = \eta/\mu$. In sample surveys the auxiliary population mean μ may be considered known, or at least estimated from a much larger sample. For this latter case, $\hat{\eta} = \hat{\theta}\mu$, where $\hat{\theta}$ is a ratio estimate based on (X_i, Y_i) $(i=1,\ldots,n)$, is often a more precise estimator of than the less sophisticated estimator $\bar{Y}$. Ratio estimation in scientific problems which have no connection with sample surveys also exists in many instances.

The application of the jackknife to ratio estimation was pioneered by Durbin (1959), in which the behaviour of (1.1.2) with g = 2 in the model

$$Y_i = \alpha + \beta X_i + e_i \qquad (1.5.1.)$$

was studied. The e_i's are i.i.d. with either a normal of gamma distribution. For the normal distribution, neglecting terms of $O(n^{-4})$, Durbin established that the jackknife estimator has both smaller bias and smaller variance than the simple estimator $\overline{Y}/\overline{X}$. In the case of gamma distributions with coefficient of variation less than 1/4, the jackknife reduces the bias, increases the variance, but reduces the mean squared error in comparison with $\overline{Y}/\overline{X}$. Rao (1965), proved that both the bias and variance of the jackknife estimator are in fact decreasing functions of g for the normal distribution, and therefore showed that g=n would be the optimum choice. Rao and Webster (1966), demonstrated through a combination of theoretical and numerical work that this also holds true for the gamma distribution. Finally, Tin (1965), Rao and Beegle (1967), Rao (1969) and Hutchison (1971) compared the jackknife ratio estimator with alternative competitors such as Mickey, Hartley and Ross, Tin and Beale estimators. Their findings favour the jackknife and the Tin estimators.

1.6 The validity of Tukey's proposal

This section describes general problems in which it has been proved that Tukey's proposal is indeed valid. Namely, the statistic (1.1.8) has an approximate t distribution or, for large n, an approximate normal distribution.

Consider the standard formulation in which the maximum likelihood estimate θ^* is a root of the equation

$$0 = \sum_{i=1}^{n} \frac{\partial \log f(X_i,\theta)}{\partial\theta} , \tag{1.6.1}$$

where $f(x;\theta)$ is the density function for the random variables X_i.

Brillinger (1964), jackknifed the maximum likelihood estimator θ^* of θ, by dividing the sample into g groups of size h each and investigated the case where g is held fixed and $h \to \infty$. In this case, Brillinger showed that the limiting distribution of the statistic:

$$\frac{g^{\frac{1}{2}}(\hat{\theta} - \theta)}{\left\{ \frac{1}{g-1} \sum_{i=1}^{g} (\hat{\theta}_i - \hat{\theta})^2 \right\}^{\frac{1}{2}}} \tag{1.6.2}$$

is a t distribution with n-1 degrees of frredom. Reeds (1978), considered the case where $g \to \infty$, with h=1 as $n \to \infty$, and showed the asymptotic normality of the jackknife version of the consistent root of the maximum likelihood equation. Reeds also showed that the jackknife estimator of the variance of the asymptotic distribution of the maximum likelihood estimator is consistent.

Miller (1964), investigated the case where the original estimator is a twice differentiable function of the sample mean. He showed that, if $X_1, X_2, \ldots, X_N$ are i.i.d. random variables with mean $\mu=0$ and variance $0 < \sigma^2 < \infty$ and the jackknife estimator of $\theta = f(\mu)$, where f is a real function, is $\hat{\theta}$, then the statistic

$$N^{\frac{1}{2}} (\hat{\theta} - \theta) \tag{1.6.3.}$$

is, as $N \to \infty$, asymptotically normally distributed with mean zero and variance $\sigma^2\{f'(\mu)\}^2$. Miller assumed that the first derivative f' of f was bounded.

The class of statistics which Miller considered is rather limited. A broadening of the class was realised by considering statistics of the form f(U) where the argument is a U-statistic. Any statistic of the form

$$U(X_1, X_2, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_c k(X_{i_1}, \ldots, X_{i_m}) , \qquad (1.6.4)$$

where the kernel function $k(X_1, \ldots, X_m)$ is symmetric in its m arguments and the summation is over all the combinations of m variables $X_{i_1}, \ldots, X_{i_m}$ out of the n variables $X_1, \ldots, X_n$, is called a U-statistic. Let $\mu = E\{k(X_1, \ldots, X_m)\}$ and the parameter of interest be $f(\mu)$, with corresponding jackknife estimator, $\hat{\theta} = f(U)$. Then, Arvesen (1969) proved that (1.6.2) with g=n, has a limiting unit normal distribution as $n \to \infty$, provided that $E\{k^2(X_1, \ldots, X_m)\}$ is finite, and f has a bounded second derivative near θ. Arvesen and Schmitz (1970) extended this result to the very general case of a real-valued function of several U-statistics $f(U_1, \ldots, U_r)$, where each U statistic U_i has a different kernel function k_i for the same set of basic i.i.d. variables $X_1, \ldots, X_n$ which can now be p-dimensional vectors. Examples of statistics falling into this framework, include ratios, the t-statistic, the Wilcoxon signed-rank statistic and the product-moment correlation coefficient.

Miller (1974b), widened the domain of applicability of the jackknife to the full linear model. He considered the model:

$$Y = X\beta + e \qquad (1.6.5)$$

where $Y = (Y_1, Y_2, \ldots, Y_n)'$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$, $e = (e_1, e_2, \ldots, e_n)'$, and X is an n x p matrix. The error variables e_i, i=1, 2,...,n are assumed to be i.i.d. with zero mean, variance σ^2, fourth moment μ_4, but not necessarily with a normal distribution. It is assumed that rank (X) = p.

Miller estimated $\theta = f(\beta)$ where f(.) is a smooth function of the regression parameters. The customary ad hoc estimator of θ would be $f(\hat{\beta})$, where $\hat{\beta}$ is the least squares estimator $(X'X)^{-1} X Y$. Applying the jackknife in the usual fashion by successively deleting each row of X and Y, produces $\hat{\theta}_{-i} = f(\hat{\beta}_{-i})$ (i=1,...,n) and also the corresponding pseudo-values. Under the condition $X'X/n \to \Sigma$, a positive definite matrix, as $n \to \infty$, Miller proved that the statistic (1.6.2) is asymptotically normally distributed, provided $\mu_4 < \infty$ and f(.) has bounded second derivatives in a neighbourhood of β. This result can be extended to non-linear regression problems and will be discussed in Chapter 4.

The validity of the jackknife has also been extended to include stochastic processes with stationary, independent increments. Gaver and Hoel (1970) jackknifed the reliability parameter $\theta = e^{-\lambda\tau}$ for fixed $\tau > 0$, where λ is the intensity parameter of a Poisson process $\{Y_t\}$. The standard estimator for λ is $\hat{\lambda} = Y_T/T$ over the interval $\{0,T\}$. The ad hoc estimator $\hat{\theta} = e^{-\hat{\lambda}\tau}$, of θ, is jackknifed by dividing the time interval $\{0,T\}$ into n equal-length sub-intervals. The estimator with the i^{th} sub-interval removed is

$\hat{\theta}_{-i} = e^{-\hat{\lambda}_{-i}\tau}$, where $\Delta Y_i = Y_{id} - Y_{(i-1)d}$, $d = T/n$ and

$\hat{\lambda}_{-i} = (Y_T - \Delta Y_i)/(T-d)$. As $n \to \infty$, the limit of the jackknife estimator $\hat{\theta}$ is defined as:

$$\lim_{n\to\infty} \hat{\theta} = e^{-\hat{\lambda}t} \{1 - Y_T (e^{\tau/T} - 1 - \tau/T)\} \tag{1.6.6}$$

In a subsequent paper, Gray, Watkins and Adams (1972), restricted the stochastic processes $\{Y_t\}$ to processes whose path functions are piecewise continuous and of bounded variation, the Wiener process component is eliminated, and $\{Y_t\}$ reduces essentially to a sum of independent Poisson processes with different jump sizes γ and intensity parameters λ. Let $\theta = f(\lambda)$, where $E(Y_t) = \lambda t$ and $\hat{\lambda} = Y_T/T$. The limit of the jackknife estimator $\hat{\theta}$, obtained by dividing the interval into n equal length sub-intervals, is defined as:

$$\lim_{n\to\infty} \hat{\theta} = f(\hat{\lambda}) - \sum_{\gamma} N_\gamma \{ f(\hat{\lambda} - \frac{\gamma}{T}) - f(\hat{\lambda}) + \frac{\gamma}{T} f'(\hat{\lambda})\}, \tag{1.6.7}$$

where N_γ is the number of jumps of size γ in $\{0,T\}$ and f' is the derivative of f. The estimator (1.6.7) is asymptotically normally distributed with mean θ and variance $\sigma^2\{f'(\lambda)\}^2/T$, as $n \to \infty$, under the set of conditions that $\Gamma = \{\gamma\}$ is a bounded set, f has a bounded second derivative near λ, and

$$\frac{1}{T}\sum_\gamma \gamma^2 N_\gamma \to \sigma^2 = \text{var}(Y_1) < \infty \qquad (1.6.8)$$

in probability as $T \to \infty$. The limit of the jackknife variance estimate $\hat{S}^2/n = \sum(\hat{\theta}_i - \hat{\theta})^2/\{n(n-1)\}$, as $n \to \infty$, is

$$\lim_{n\to\infty} \hat{S}^2/_n = \sum_\gamma N_\gamma \{f(\hat{\lambda} - \frac{\gamma}{T}) - f(\hat{\lambda})\}^2 \qquad (1.6.9)$$

As $T \to \infty$, (1.6.9) multiplied by T converges in probability to $\sigma^2\{f'(\lambda)\}^2$, under the conditions (1.6.8), Γ bounded and f' continuous near λ. Under these stated conditions, as $T \to \infty$, $T^{\frac{1}{2}}(\lim \hat{\theta}-\theta)/\lim \hat{S}^2/n)^{\frac{1}{2}}$ has a limiting unit normal distribution.

1.7 Examples of jackknifing failures

There are many counter examples where the jackknife does not work. A necessary 'ingredient' for the jackknife to work, is that the estimator $\hat{\theta}_n$ has to have a locally linear quality. Miller (1968), defines this linear quality such that, for an unmodified estimator $\hat{\theta}_n(X_1,\ldots,X_N)$, it can be expanded in a power series for each observation where

(i) the second and higher order terms are negligible and

(ii) the first order term is linear in the observation X_i or some simple function of X_i.

Given these properties for $\hat{\theta}_n$, then the jackknife estimate $\hat{\theta}$ can be expanded in a power series so that the large sample theory can be applied to establish asymptotic normality with the correct mean and variance.

Asymptotic normality is preserved through the linear quality of $\hat{\theta}_n$, and this normality is maintained under jackknifing.

In an earlier paper, Miller (1964), showed that the largest order statistic $\hat{\theta}_n = Y_{(n)}$ was non-normal, and therefore unsuitable for jackknifing. He defined $\hat{\theta}_n$ to be the $\max\{Y_1, Y_2 \ldots, Y_n\}$ such that

$$\begin{aligned} &\hat{\theta}_n = Y_{(n)} \\ \text{and } &\hat{\theta}_{-i} = Y_{(n)} \text{ if } i \neq (n), \text{ i.e. } (n-1) \text{ times} \\ &\quad\; = Y_{(n-1)} \text{ if } i = (n), \text{ i.e. once} \end{aligned} \tag{1.7.1}$$

where the order statistics $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ correspond to $(Y_1, Y_2 \ldots, Y_n)$.

The jackknife pseudovalues are defined in the normal manner:

$$\begin{aligned} \tilde{\theta}_i &= Y_{(n)} \quad \text{if } i \neq (n) \\ &= nY_{(n)} - (n-1)\, Y_{(n-1)} \text{ if } i = (n) \end{aligned} \tag{1.7.2}$$

The jackknife estimator and variance estimator are thus:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \tilde{\theta}_i}{n} = Y_{(n)} + \{(n-1)/n\}.(Y_{(n)} - Y_{(n-1)})\}$$

$$\text{and } \mathrm{Var}(\hat{\theta}) = (n-1)^{-1} \sum_i (\tilde{\theta}_i - \hat{\theta})^2 = \{(n-1)^2/n\} \; (Y_{(n)} - Y_{(n-1)})^2 \tag{1.7.3}$$

The t-statistic is:

$$t = \frac{n^{\frac{1}{2}}\,(\hat{\theta} - \theta)}{\mathrm{Var}\,(\hat{\theta})^{\frac{1}{2}}} = \frac{n}{n-1}\left\{\frac{Y_{(n)} + \{(n-1)/n\} \; \{Y_{(n)} - Y_{(n-1)} - \theta\}}{Y_{(n)} - Y_{(n-1)}}\right\}$$

$$\sim 1 - \frac{(\theta - Y_{(n)})}{Y_{(n)} - Y_{(n-1)}} \tag{1.7.4}$$

$$= 1 - A_n$$

where $A_n = \dfrac{\theta - Y_{(n)}}{Y_{(n)} - Y_{(n-1)}}$

The upper tail of the distribution of t from (1.7.4) is not identical with the Student t distribution, since $A_n \geq 0$. Hence, this is not a valid case for jackknifing, since the jackknife cannot create asymptotic normality.

Mosteller and Tukey (1968), in unpublished notes by L.E. Moses jackknifed the sample median with the interesting result that the jackknife estimator was the same as the original estimator. That is, if the sample is of size n=2h and the i^{th} observation (i=1, 2...,h) in the upper half of the sample is deleted, the median of the remaining sample will be the h^{th} highest random variable. Similarly, if the i^{th} observation (i=1, 2...,h) in the lower half of the sample is deleted, the median will be the $(h+1)^{th}$ order statistic in the original sample. Hence, the jackknife estimator is the sample median since the pseudovalues have two different values taking on each t times.

Wainer and Thissen (1975), applied the jackknife to estimating Fisher's Z transformation of the correlation coefficient between two variables in the case of non-normality, and found that the jackknife estimator increased the bias of the estimate.

For a truncation point problem with $\hat{\theta}_n = Y_{(n)}$, where the random variables are of a finite range, Robson and Whitlock (1984), had to modify the definition of the jackknife because of the particular bias expansion and derived $2Y_{(n)} - Y_{(n-1)}$ as the estimator instead. They also derived a corresponding confidence limit statement

$$\Pr\left\{Y_{(n)} + \left\{\frac{(1-\alpha)}{\alpha}\left(Y_{(n)} - Y_{(n-1)}\right)\right\} > \theta\right\} \simeq 1-\alpha \qquad (1.7.5)$$

Finally, an area where the jackknife has had very little application, is in time-series analysis. This is somewhat ironic since Quenouille (1949), originially proposed the jackknife concept for a time-series problem. However, unless the number of deleted groups, $g=2$, then the removal of data segments will violate the serially correlated sequence of observations. This particular application is discussed later in Chapter 5.

1.8 Application of the jackknife in multivariate analysis

Tests of hypotheses and confidence intervals concerning the correlation coefficient ρ, in bivariate normal populations, are commonly based on Fisher's-Z transformation of the sample correlation coefficient, $\tanh^{-1} r$. This statistic is approximately normally distributed with mean $\tanh^{-1}\rho$ and variance $1/(n-3)$. The normal theory test for $\rho=0$ based on the $\tanh^{-1} r = \frac{1}{2} \ln \frac{1+r}{1-r}$,
is asymptotically valid for any population having finite fourth moments. However, if $\rho \neq 0$, the asymptotic variance of $\tanh^{-1} r$ is not, in general, $1/(n-3)$, unless the underlying distribution is normal. Duncan and Layard (1973), used Monte-Carlo simulation to compare the small sample performance of the usual normal theory procedures for inference about correlation coefficients with that of two asymptotically robust procedures, one of which is based on a grouping of the observations and the other on the jackknife technique. For differing cases of normal and five non-normal distributions, jackknifing the statistic r, was shown to work well in terms of the nominal 95% confidence intervals for the correlation coefficient ρ.

Another area of multivariate analysis in which the jackknife has found application is discriminant analysis. The jackknife method has been employed by Lachenbruch and Mickey (1968), to estimate the discriminant coefficients and to assess their variability.

They have estimated the errors of misclassification probabilities as follows. First, the discriminant function has been computed from all the observations and, second, the i^{th} 'partial' discriminant functions have been found from all but the i^{th} random variable. This is the well-known U-method and has been found to produce less biased estimates of the error rates. Mosteller and Tukey (1968), have applied the U-method for solving the 'Authorship' problem. This is a discriminant analysis problem created by the Federalist papers and the dispute about their real author. Gray and Schucany (1972), have given a synopsis of the Mosteller-Tukey work.

1.9 An example of an area of jackknife application : Jackknifing in biomedical studies

In an interesting paper, Salsburg (1971), considered the jackknife to test dose-response effects when the responses are binomial variates and when the underlying p values are near 0 or 1. The evaluation of a drug for either toxicity or efficiency often involves a set of ex post facto data in the form of percentages of individuals affected at different dose levels. For example, 3 out of 50 might complain of dizziness at 25mg of drug per day, 2 out of 33 at 30mg per day, and 7 out of 84 at 50mg per day. The question to consider is whether this is a drug-related effect or a random somatic symptom. Similarly, consider the problem where all traces of a disease-causing organism are eliminated in 89% of patients given a low dose, in 100% of patients given a higher dose, and 96% given a still higher dose. Does this imply that a continuation of the therapy will improve the cure rate or that there is a small resistant sub-strain of organism that will not be affected by any course of drug?

Given a set of doses and percentage responses, a standard procedure is to fit a sigmoid curve to the log-dose. Unfortunately, for the type of problems described above, this approach will not be very successful.

The reasons are as follows:

(i) the hypothesis is being tested that the regression has 'zero slope', rather than attempting to estimate the regression. The test statistic used for fitting a sigmoid to percentage responses is a Student t, with two fewer degrees of freedom that the number of doses tested. In most experiments, the doses are seldom planned in terms of a good regression estimating experiment, and the number of doses tried is usually only between 3 and 5.

(ii) the observed percentages are usually near 0 or 1, thus making the curve fitting very difficult. Naylor (1964), showed that the usual sigmoid regressions did not produce good fits when this was the case, and therefore the validity of the t distribution to test the regression coefficient is somewhat dubious. Salsburg thus considered the jackknife as a method for a test of hypothesis. In particular if

$Y_{i,j} = \{ {C \atop 1} \}$ = response of i^{th} individual, j^{th} dose,

X_j = j^{th} dose

then, the least squares covariance estimator is:

$$\hat{\beta} = \sum(X_j - \bar{x})(Y_{ij} - \bar{y})/n \qquad (1.9.1)$$

The estimate $\hat{\beta}$ can be considered as a weighted sum of the variates Y_{ij} and fits Arvesen's (1969), criterion for valid jackknifing.

Salsburg considered a problem where an experimental drug was administered to a large group of patients infected with one of 3 organisms. The clinicians were allowed to choose any of 3 different dose levels for a given patient. The drug was known to be effective against 2 of the organisms and marginally effective against the third. The t-value for each organism was computed from the pseudovalues of the jackknifed covariance estimator, and the effect of the dose response for the different organisms was evaluated.

Salsburg extended the problem to consider how well the t-distribution fits for a finite number of observations N, and for what values of N the asymptotic results effectively hold. The results indicated that, for small sample sizes, the procedure is a conservative alpha(α)level test with the true size always less than nominal. Also, the lower α, the more conservative the test is.

A Monte Carlo procedure was carried out to evaluate the power of the test, under the alternative hypothesis of positive slope. The sample distribution functions were compared against the true distribution functions with Kolomogorov tests, all of them fitting with $\alpha > 0.30$. Although the jackknife produces a test with low power, there is no other easily available competitor.

In a subsequent paper by Frawley (1974), he reinvestigated the problem considered by Salsburg and found that the jackknife produced a hypothesis test, having better power than was previously indicated. For a large number of subjects, the jackknife test was shown to be quite useful. Also, Frawley suggested an alternative test for smaller groups of subjects, which was a more powerful procedure under the null hypothesis and various alternative hypotheses.

Heltshe and Forrester (1983), applied the jackknife to estimating the number of species in a community and also the variance of this number. The concept of diversity has been used as a method for characterising the structure of species abundance in a community. Although Zahl (1977), applied the jackknife technique to the estimation of species diversity indices, Heltshe and Forrester's paper considers the most fundamental concept of diversity, namely the number of species or the species richness in a community. Good (1953), and Engen (1978), used the number of species occurring with a frequency of one, to estimate the true number of species, based on a random sample of individuals. Their paper proposes an estimation procedure under the assumption that the sample of individuals is not necessarily random. The methodology is as follows:

Let $y^{\circ}=S$ be the number of species found in a pool of n quadrats; y^{-i}, i=1, 2...,n be the number of species found in a pool of (n-1) quadrats with the i^{th} quadrat removed. The jackknife pseudovalues y_i, i=1, 2...,n are defined as:

$$y_i = ny^{\circ} - (n-1)y^{-i} \qquad (1.9.2)$$

Let f_j be the number of quadrats containing j 'unique' species with

$$\sum_{j=o}^{s} f_j = n \text{ and } \sum_{j=o}^{s} jf_j = k. \qquad (1.9.3)$$

where k is the total number of unique species in a pool of n quadrats.

Since $y^{-i}=y^{\circ}-j$ if the i^{th} quadrat contains j unique species, the pseudovalues take on the following distinct values:

$$y_i = y^{\circ} + j(n-1) \qquad (1.9.4)$$

where j is the number of unique species in Quadrat i, Hence,

$$\begin{aligned} y_i &= y^{\circ} \text{ with frequency } f_o \\ &= y^{\circ} + (n-1) \text{ with frequency } f_1 \\ &\vdots \\ &= y^{\circ} + j(n-1) \text{ with frequency } f_j \end{aligned}$$

The jackknife estimator for the number of species is thus:

$$\hat{S} = \frac{1}{n} \sum y_i = \frac{1}{n} \sum_{o}^{s} f_j \{y^{\circ} + j(n-1)\}$$

$$= y^{\circ} + \frac{(n-1)}{n} k \qquad (1.9.5)$$

with variance estimate,

$$\mathrm{Var}(\hat{S}) = \frac{1}{n(n-1)} \sum_{1}^{n} (y_i - \bar{y})^2$$

$$= \frac{n-1}{n}\{\sum_{0}^{s} j^2 f_j - k^2/n\} \qquad (1.9.6)$$

Monte Carlo simulation techniques were used to evaluate the behaviour of the jackknife estimators of species richness. For different clumped populations, the percentage cover of the two-sided 95% confidence intervals was generally high in all cases. In terms of bias, the jackknife estimator was better than the original estimator until the sample size increased to above 80 to 100 quadrats. At this point, the bias of the jackknife estimator exceeded the bias of the original estimator but the magnitude of this bias was negligible for these sample sizes.

In a paper by Frangos and Stone (1984), they considered the jackknife as well as other non-parametric methods, to estimate a proportion with batches of different sizes. In this case, M randomly drawn batches $\{ (n_i, x_i);\ i=1,\ldots,M \}$ were considered, where n_i is the size of the i^{th} batch and x_i is the number of defectives in it.

Given the pure binomial estimator

$$\hat{u}_o = \sum_{i=1}^{M} x_i/N \qquad (1.9.7)$$

where $N = \sum n_i$. Gladen (1979), defined the jackknife estimator:

$$\hat{u}_{oj} = \sum_{i=1}^{M} w_i P_i\ , \qquad (1.9.8)$$

where $P_i = x_i/n_i$ and

$$w_i = \frac{n_i}{N} + \frac{M-1}{M}\{\frac{n_i}{N-n_i} - \frac{n_i}{N}\sum\{\frac{n_j}{N-n_j}\}\} \qquad (1.9.9)$$

Frangos and Stone (1984), as well as considering the alternative estimator to $\hat{u}_0$,

i.e. $\hat{u}_1 = \sum p_i/M = \bar{p}$, (1.9.10)

also considered other estimators such as cross-validatory estimators, jackknife and bootstrap cross-validatory estimators, a minimum jackknife risk estimator and a variation of the 'classical' estimator defined by Southward and van Ryzin (1972). The relationship between the jackknife and cross-validatory estimators is discussed in Section 1.14.

Monte Carlo simulation techniques were used to evaluate each of the estimators. Random samples of (n,x) were generated, where for a given value of n, x was binomial (n,p) with p from a beta (α,α) distribution. In terms of mean square error, Gladen's jackknife estimator, and the estimators with a jackknife 'element' did not show an improvement in results. Similarly, in terms of robust confidence limits, the jackknife estimator produced only modest results for small M. Much better results were obtained using the alternative estimator u_1, and Southwood and van Ryzin's (1972) amended classical estimator.

1.10 The infinitesimal jackknife

The concept of an infinitesimal jackknife estmator (I.J.E.) was introduced by L.B.Jaekel in an unpublished Bell Telephone Laboratories technical memorandum. Although it does not appear to be as practically useful as the ordinary jackknife estimator (O.J.E.), it does give a deeper insight into the nature of the jackknife procedure and, through the concept of the influence function, (Hampel,1974), establishes an important connection between the jackknife method and the theory of robust estimation (see Section 1.11).

In order to understand the connection between the I.J.E. and the theory of robust estimation, it is necessary to summarise some relevant aspects of the latter. Under regularity conditions, an estimator $\hat{\theta}_n = T(\hat{F})$, where $\hat{F}$ is the sample c.d.f., of $\theta = T(F)$ can be expressed in the form:

$$T(\hat{F}) = T(F) + \int T'(F,y)d(\hat{F}-F)(y) + 0_p(n^{-\frac{1}{2}}), \qquad (1.10.1)$$

where $T'(F,y)$ is a von Mises (1947) derivative defined by

$$\lim_{\varepsilon \to 0} \frac{T(F + \varepsilon G) - T(F)}{\varepsilon} = \int T'(F,y)dG(y) \qquad (1.10.2)$$

The term influence curve has been associated with $T'(F,y)$ by Hampel (1974), because it quantifies the degree a change in the mass at y will change the estimate.

If $T(cF) = T(F)$ for all F and $c > 0$, then

$$\int T'(F,y)dF(y) = 0,$$

so that

$$T(\hat{F}) = T(F) + \frac{1}{n}\sum_{i=1}^{n} T'(F,Y_i) + 0_p\ (n^{-\frac{1}{2}}) \qquad (1.10.3)$$

From the Central Limit theorem, the average of i.i.d. random variables in (1.10.3) is asymptotically normally distributed with mean zero and variance

$$\frac{1}{n} \int \{T'(F,y)\}^2 dF(y) \qquad (1.10.4)$$

If T is known, an empirical estimator of the asymptotic variance is

$$\frac{1}{n^2} \sum_{i=1}^{n} \{ T'(\hat{F},Y_i) \}^2 \qquad (1.10.5)$$

Jaeckel defined the I.J.E. as follows:

Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with distribution F. Let the estimator $\hat{\theta}_n$ be a function $T(Y;w)$ of the observations $Y = (Y_1,\ldots,Y_n)'$ and arbitrary weights $w = (w_1,\ldots,w_n)'$. If $w_i = 1/n$, then $\hat{\theta}_n = T(\hat{F})$, where $\hat{F}$ is the empirical distribution function. Suppose also that the function T is self-normalizing in the weights so that $T(Y;cw)$

$= T(Y;w)$ for all $c > 0$.

For the O.J.E., the deleted observation is given a weight zero.

i.e. $\hat{\theta}_{-i} = T(Y_1,\ldots,Y_n\ ;\ {}^1/n,\ldots,\ 0,\ldots,{}^1/n)$ (1.10.6)

In the case of the I.J.E., the deleted observation is given a weight slightly less than the others such that $\sum_{i=1}^{n} w_i$ does not necessarily equal 1.

i.e. $\hat{\theta}_{-i}(\varepsilon) = T(Y_1,\ldots,Y_n\ ;\ {}^1/n,\ldots,{}^1/n - \varepsilon,\ldots,{}^1/n)$ (1.10.7)

The I.J.E. for the asymptotic variance of $\hat{\theta}_n$, is defined to be:

$$\frac{\hat{S}^2(\varepsilon)}{n} = \frac{(1-\varepsilon)}{n^2\varepsilon^2} \sum_{i=1}^{n} \{\hat{\theta}_{-i}(\varepsilon) - \frac{1}{n} \sum \hat{\theta}_{-j}(\varepsilon)\}^2 \quad (1.10.8)$$

If $\varepsilon = {}^1/n$, then (1.10.4) equates with the O.J.E.

Assume that $T(y;w)$ is differentiable with respect to the weights.

Let $\hat{D}_i = \dfrac{\partial T(y;w)}{\partial w_i}$ and $\hat{D}_{ii} = \dfrac{\partial^2 T(y;w)}{\partial w_i^2}$ (1.10.9)

be the first and second derivatives, respectively, evaluated at $y = Y$ and $w = ({}^1/n,\ldots,{}^1/n)'$. Given the self-normalising condition on the weights in T, this implies that $\sum \hat{D}_i = 0$.

Consider the expansion

$$\hat{\theta}_{-i}(\varepsilon) = \hat{\theta}_n - \varepsilon\hat{D}_i + \frac{1}{2}\varepsilon^2 \hat{D}_{ii}\ldots, \quad (1.10.10)$$

then , it follows that

$$\frac{\hat{S}^2(0)}{n} = \lim_{\varepsilon \to 0} \frac{S^2(\varepsilon)}{n} = \frac{1}{n^2} \sum_{i=1}^{n} \hat{D}_i^2 \qquad (1.10.11)$$

However, since $\hat{D}_i$ equals $T'(\hat{F}, Y_i)$, then (1.10.11) equals (1.10.5) Therefore, the I.J.E. of the variance estimate (1.10.11) provides an estimate of the asymptotic variance (1.10.4)

For the O.J.E.,

$$\hat{\theta}_n - \hat{\theta} = (n-1)\ ({}^1/n \sum_{i=1}^{n} \hat{\theta}_{-i} - \hat{\theta}_n) \qquad (1.10.13)$$

estimates the bias of $\hat{\theta}_n$.

Also, if

$$\hat{b}(\varepsilon) = \frac{1-\varepsilon}{n\varepsilon^2} \ \{1/_n \textstyle\sum \hat{\theta}_{-i}(\varepsilon) - \hat{\theta}_n\} \qquad (1.10.14)$$

then, (1.10.15) equals (1.10.14) when $\varepsilon = {}^1/_n$. From the power series expansion (1.10.10), it follows that:

$$\hat{b}(0) = \lim_{\varepsilon \to 0} \hat{b}(\varepsilon) = \frac{1}{2n} \sum_{i=1}^{n} \hat{D}_{ii} \qquad (1.10.15)$$

where the I.J.E. is defined to be $\hat{\theta}(0) = \hat{\theta}_n - \hat{b}(0)$

In his paper, Jaekel proves, under general conditions, that for estimators which satisfy (1.10.3), $\hat{S}^2(0)$ and $n\hat{b}(0)$ converge to the correct asymptotic constants as $n \to \infty$. Jaekel also shows that under the same conditions, the O.J.E. behaves correctly asymptotically.

$$\text{Let } \hat{\theta}_n = T = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (1.10.16)$$

and denote $\bar{X} = M\,(X_1,\ldots\ldots, X_n;\ w_i,\ldots\ldots, w_n) = \dfrac{\sum w_i X_i}{w_i}$ (1.10.17)

Then,

$$M(\hat{F}) = \int x d\hat{F}(x) = \bar{X}$$

and $M(F) = \int x dF(x) = E(X)$ (1.10.18)

Moreover,

$$T(X_1,\ldots\ldots, X_n\ ;\ w_1,\ w_2,\ldots\ldots, w_n) = \frac{1}{\sum w_i}\ \{\sum w_i (X_i - M)^2\}$$

$$T(\hat{F}) = \int (x - M(\hat{F}))^2\ d\hat{F}(x) = \hat{\theta}_n \qquad (1.10.19)$$

$$T(F) = \int (x - M(F))^2\ dF(X) = \mathrm{var}(X)$$

Differentiating T with respect to the weights w_k, gives

$$\frac{\partial T}{\partial w_k} = \frac{1}{\sum w_i}\ \{\sum w_i . 2(X_i - M)(-\frac{\partial M}{\partial w_k}) + (X_k - M)^2\} - \frac{1}{(\sum w_i)^2} \sum w_i (X_i - M)^2$$

Since $\sum w_i X_i = M \sum w_i$, the differential equation simplifies to:

$$\frac{\partial T}{\partial w_k} = \frac{1}{\sum w_i} \{(X_k - M)^2 - T\}$$

Hence

$$\hat{Q}_k = \frac{\partial T}{\partial w_k} \Big|\ x_k = X_k,\ w_k = {}^1/n,\ j = 1,\ldots, n$$

$$= (X_k - \bar{X}) - \frac{1}{n} \sum_i (X_i - \bar{X})^2$$

The variance estimate of $\hat{\theta}_n$ is

$$\hat{V}(0) = \frac{1}{n^2} \sum_k (X_k - X)^2 \qquad (1.10.20)$$

Differentiating again leads to:

$$\frac{\partial^2 T}{\partial w_k^2} = \frac{1}{\sum w_i} \left\{2(X_k - M)\left(-\frac{\partial M}{\partial w_k}\right) - \frac{\partial T}{\partial w_k}\right\} - \frac{1}{(\sum w_i)^2}\left\{(X_k-M)^2 - T\right\}$$

Since $\frac{\partial M}{\partial w_k} = \frac{1}{\sum w_i}(X_k - M)$,

then

$$\hat{Q}_{kk} = \left.\frac{\partial^2 T}{\partial w_k^2}\right| x_k = X_k,\ w_k = \frac{1}{n},\ j = 1,\ldots,n$$

$$= -4(X_k - \bar{X})^2 + \frac{2}{n}\sum_i (X_i - \bar{X})^2$$

Replacing n by (n-1) in the denominator, gives

$$\hat{b}(0) = \frac{1}{2n(n-1)}\sum_k \hat{Q}_{kk} = -\frac{1}{n(n-1)}\sum_i (X_i - \bar{X})^2$$

Hence, the I.J.E. is defined to be

$$\hat{\theta}_n - \hat{b}(0) = \left(\frac{1}{n} + \frac{1}{n(n-1)}\right)\sum_i (X_i - \bar{X})^2 = \frac{1}{(n-1)}\sum_i (X_i - \bar{X})^2 \qquad (1.10.21)$$

where the estimate is known to be exactly unbiased.

1.11 The jackknife method and the theory of robust estimation

Let $x_1, x_2, \ldots, x_n$ be n i.i.d. random variables from the distribution $P(X_i < x) = F((x-\theta)/\sigma)$, where the functional form of F is not exactly known. Let T_n be an estimator of θ. According to Huber (1972), the estimator T_n is robust if it has one of the following properties:

(i) A high efficiency relative to the sample mean for all F.

(ii) A high efficiency over a strategically selected finite set $\{F_i\}$ of distribution functions (e.g. the normal, logistic, double exponential, Cauchy and rectangular distributions)

(iii) A small asymptotic variance over some neighbourhood of a distribution function, in particular of the normal

(iv) The distribution of the estimator should change little under arbitrary small variations of the underlying distribution F. (Hampel, 1971, 1974).

In a relatively narrow sense, 'robustness' can be interpreted as signifying insensitivity against small deviations from the assumptions. Distributional robustness means insensitivity of the estimator if the shape of the true underlying distribution deviates slightly from the assumed model (usually the normal distribution). Robust statistics are not non-parametric statistics because, in the theory of robustness, there exists an ideal parametric model which is evaluated, to make sure that the statistical methods work well in the model and in some neighbourhood of it. In an analogy with computers, robust estimators are a 'third generation' of statistics after parametric and non-parametric ones.

For a qualitative definition of robustness, the reader is referred to Hampel (1971). Huber (1972) has distinguished three kinds of robust estimators which he named M, L and R estimators respectively. If T_n is an M, L or R estimator of $\theta = T(F)$, then Huber (1972) found that $n^{\frac{1}{2}}(T_n - T(F))$ is asymptotically normal with mean zero and asymptotic variance:

$$\sigma^2(F) = \int \{ IC(x,F,T) \}^2 \, F(dx), \qquad (1.11.1)$$

where the function $IC(x,F,T)$ is the first order influence function of the functional T at the distribution F (Hampel, 1974). The influence function $IC(x,F,T)$ is an important characteristic of a robust estimate and a more detailed examination of it will be given in the next section.

1.12 Relationship between the jackknife statistic and the influence function

Hampel's influence function (1974), and its relation with the jackknife statistic provides an important bridge between the jackknife method and the theory of robust estimation. The influence function, which is a very useful heuristic tool in robust statistics, is defined as

$$IC_{T,F}(x) = \lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon}, \qquad (1.12.1)$$

where δ_x is the distribution which places point mass 1 at the value of an observation, and T(F) is the test statistic.

This function is a measure of the influence an additional observation of value x, has on the test statistic T(F).

If T is sufficiently regular, then it can be linearised near F in terms of the influence function. Using a Taylor series expansion:

$$T(G) = T(F) + \int IC_{T,F}(x)d(G-F)(x) + \quad \ldots \qquad (1.12.2)$$

where G is a distribution function.

Given $\int IC_{T,F}(x)dF = 0$, by substituting the empirical distribution F_n for G into (1.12.2), leads to

$$\sqrt{n}\,(T(F_n) - T(F)) = n \int IC_{T,F}(x)dF_n + \ldots$$

$$= \frac{1}{\sqrt{n}} \sum_i IC_{T,F}(x_i) + \ldots \qquad (1.12.3$$

Based on the central limit theorem, the leading term on the right hand side of (1.12.3) is asymptotically normal with mean 0, if the x_i are independent with common distribution F. Assuming the other terms are asymptotically negligible, then $\sqrt{n}\,(T(F_n) - T(F))$ is asymptotically normal with mean 0 and variance:

$$V(F,T) = \int \{IC_{F,T}(x)\}^2 \, dF(x) \qquad (1.12.4)$$

The influence function has two main uses. Firstly, it is a tool which can be used to assess the relative influence of individual observations towards the value of an estimator or test statistic.

Consider an observation which is an outlier and unbounded. **Hampel** denoted its maximum absolute value:

$$\gamma^* = \sup_x | IC_{T,F}(x) |$$

and termed the value ' gross error sensitivity'.

If M is a set of cumulative distribution functions, consider the 'gross error' model:

$$P_2(F_o) = \{ F | F = (1-\varepsilon)F_o + \varepsilon H, \ H \in M \}$$

Then, from Huber (1981),

$$T(F) - T(F_o) \approx \varepsilon \int IC_{T,F_o}(x) dH(x)$$

Hence,

$$b_1(\varepsilon) = \sup | T(F) - T(F_o) | \approx \varepsilon\gamma^* \qquad (1.12.5)$$

Secondly, since the influence function allows a guess of the explicit formula (1.12.4) for the asymptotic variance, then a simple and heuristic assessment of the asymptotic properties of an estimator can be made.

The 'sensitivity curve', (Tukey 1971), and the jackknife are two finite sample versions of the influence function. For the jackknife method, if F is replaced by F_n and ε by $^{-1}/(n-1)$ in (1.12.1), then

$$\hat{IC}_{T,F}(x_i) = \frac{T(\frac{n}{n-1} F_n - \frac{1}{n-1} \delta_{x_i}) - T(F_n)}{-\frac{1}{(n-1)}}$$

$$= (n-1) \{T_n(x_1, \ldots, x_n) - T_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\}$$

$$= \hat{\theta}_i - T_n , \qquad (1.12.6)$$

where $\hat{\theta}_i = nT_n - (n-1) T_{n-1}$, is the i^{th} jackknife pseudovalue. Then, the jackknife estimate of $\theta = T(F), \hat{\theta}$, is defined as:

$$\hat{\theta} = \frac{\sum_i \theta_i}{n} = \frac{\sum(\hat{IC}_{T,F}(x_i) + T_n)}{n}$$

$$= T_n + \frac{\sum_{i=1}^{n} \hat{IC}_{T,F}(x_i)}{n} , \qquad (1.12.7)$$

to the first order, where $\hat{\theta}$ is the jackknife estimate of $\theta = T(F)$ and $\hat{IC}_{T,F}(x_i)$ is the finite sample version of the influence function, which is given by (1.12.6).

Frangos (1984) calculated the influence function of the sample mean as follows :

Let $T(F) = \int x dF = \mu$ be the sample mean $T_n = \frac{1}{n} \sum_i x_i$ then the influence function: Following Frangos (1984), the influence function of the sample mean, is derived as follows :

$$IC_{T,F}(x) = \lim_{\varepsilon \to 0} \frac{x d\{(1-\varepsilon)F + \varepsilon\delta_x\} - \int x dF}{\varepsilon}$$

$$= \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\int x dF + \varepsilon x - \int x dF}{\varepsilon}$$

$$= x - \int x dF$$

$$= x - T(F)$$

$$= x - \mu$$

Finally, it should be noted that when the influence function does not depend smoothly on F, the jackknife produces poor results, with usually a very inaccurate variance estimate. An example of this phenomenon is the median, which is based on a small sample of ordered statistics.

1.13 Relationship between the jackknife and bootstrap methods

The bootstrap is another member of the family of resampling techniques, which provide estimators of bias and variance for an extremely wide class of statistics. The bootstrap method was developed by B.Efron in a series of papers (see Efron, 1979a, 1979b, 1981a, 1981b, 1982) and is described in detail in Chapter 2. Efron (1979a), attempted to explain the jackknife in terms of the bootstrap by showing that the jackknife can be thought of as a linear expansion method for approximating the bootstrap. Firstly, Efron considered the one-sample situation in which a random sample of size n is observed from a completely unspecified probability distribution F.

$$X_i = x_i \quad X_i \sim_{ind} F \quad i = 1,2,\ldots,n \tag{1.13.1}$$

where $X = (X_1, X_2,\ldots, X_n)$ and $x = (x_1, x_2,\ldots, x_n)$ denote the random sample and its observed realisation, respectively. The bootstrap method is described as follows :

Step 1 : Construct the sample probability distribution $\hat{F}$, putting mass $^1/n$ at each point $x_1, x_2,\ldots, x_n$

Step 2 : Draw a random sample of size n from $\hat{F}$, say

$X_i^* = x_i^*, X_i^* \sim_{ind} \hat{F} \quad i = 1, 2,\ldots,n$

Call this the bootstrap sample, $X^* = (X_1^*, X_2^*,\ldots, X_n^*)$, $x^* = (x_1^*, x_2^*,\ldots, x_n^*)$. The values of X^* are randomly selected with replacement from the set $\{x_1, x_2,\ldots, x_n\}$.

As a point of comparison, the jackknife can be thought of as drawing samples of size n-1 without replacement.

Step 3 : Approximate the sampling distribution of R(X,F) by the bootstrap distribution of

$$R^* = R(X^*, \hat{F}) \qquad (1.13.2)$$

where the distribution of R^* is induced by the random mechanism (1.13.1) with $\hat{F}$ held fixed at its observed value. The difficult part of the bootstrap procedure is to actually calculate the bootstrap distribution. One method of calculation is to use a Taylor series expansion to obtain the approximate mean and variance of the bootstrap distribution of R^*.

Let $N_i^* = \#\{X_i^* = x_i\}$ be the number of times x_i is selected in the bootstrap sampling procedure. Define $P_i^* = N_i^*/n$, such that

$$P^* = (P_1^*, P_2^*, \ldots, P_n^*) \qquad (1.13.3)$$

Using the properties of the multinomial distribution, P^* has mean vector and covariance matrix

$$E_* P^* = e/n \text{ , } Cov_* \, P^* = I/_{n}2 - e'e/n^3, \qquad (1.13.4)$$

under the bootstrap sampling procudere, where I is the identity matrix and e = (1, 1,...,1)

(Notations such as E_* and Cov_* indicate probability calculations relating to the bootstrap distribution of X^*, with x and $\hat{F}$ fixed)

The abbreviated notation

$$R(P^*) = R(X^*, \hat{F}) \qquad (1.13.5)$$

for the bootstrap realisation of R corresponding to P^*, can then be used. Hence, expanding $R(P^*)$ in a Taylor series about the value $P^* = e/n$ gives an approximation of the bootstrap distribution of $R(X^*, \hat{F})$

$$\text{i.e. } R(P^*) = R(e/n) + (P^* - e/n)U + \tfrac{1}{2}(P^* - e/n).V(P^* - e/n)', \qquad (1.13.6)$$

where

$$U = \begin{bmatrix} \vdots \\ \dfrac{\partial R(P^*)}{\partial P_i^*} \\ \vdots \end{bmatrix}_{P^*=e/n} \quad \text{and } V = \begin{bmatrix} & \vdots & \\ & \dfrac{\partial^2 R(P^*)}{\partial P_i^* \, \partial P_j^*} & \\ & \vdots & \end{bmatrix}_{P^*=e/n} \qquad (1.13.7)$$

The restriction $\sum P_i^* = 1$ has been ignored in (1.13.6) and (1.13.7). This computational convenience is justified by extending the definition of $R(P^*)$ to all vectors P^* having non-negative components, at least one positive, by the homogenous extension

$$R(P^*) = R\left(\frac{P^*}{\sum_i^n P_i^*}\right) \qquad (1.13.8)$$

The homogeneity of (1.13.8) leads to:

$$eU = 0, \quad eV = -nU' \text{ and } eVe' = 0 \qquad (1.13.9)$$

An approximation to the bootstrap expectation is obtained from (1.13.4) and (1.13.6)

$$\text{i.e. } E_* R(P^*) = R(e/n) + \tfrac{1}{2} \text{ trace } V\{I/n^2 - e'e/n^3\}$$
$$= R(e/n) + \frac{1}{2n}\bar{V} \qquad (1.13.10)$$

where

$$\bar{V} = \sum_{i=1}^{n} V_{ii}/n \qquad (1.13.11)$$

An approximation for the bootstrap variance is obtained by ignoring the last term in (1.13.6)

$$\text{i.e. } Var_* R(P^*) = U' \, (I/n^2 - ee/n^3) \, U = \sum_{i=1}^{n} U_i^2/n^2 \qquad (1.13.12)$$

It can be shown that the jackknife expressions for bias and variance are essentially the bootstrap results obtained in (1.13.10) and (1.13.12). Based on the usual jackknife theory, consider $R(X,F) = \theta(\hat{F}) - \theta(F)$, the difference between the non-parametric estimator of some parameter $\theta(F)$ and $\theta(F)$ itself. $R(e/n) = \theta(\hat{F}) - \theta(F) = 0$, since $R(X^*,F) = \theta(\hat{F}^*) - \theta(\hat{F})$, with F^* being the empirical distribution of the bootstrap sample. Then, (1.13.10) becomes $E_* \{\theta(\hat{F}^*) - \theta(\hat{F})\} = (1/2n)\bar{V}$, implying that $E_F \{\theta(\hat{F}) - \theta(F)\} = (1/2n)\bar{V}$. Similarly, (1.13.12) becomes $Var_* \{\theta(\hat{F}^*) - \theta(\hat{F})\} = \sum U_i^2/n^2$, implying $Var_F \theta(\hat{F}) = \sum U_i^2/n^2$. Also the approximations:

$$Bias_F \theta(\hat{F}) = \frac{1}{2n} \bar{V} \text{ and } Var_F \theta(\hat{F}) = \sum_{i=1}^{n} U_i^2/n^2 \qquad (1.13.13)$$

exactly agree with those given by Jaekel's infinitesimal jackknife as defined in Section 1.10 (see Jaekel, 1972 and Miller, 1974a).

Efron (1979b), demonstrated the conjecture that the jackknife is a linear approximation for the bootstrap, by a series of examples including the variance of the sample mean, error rates in a linear discriminant analysis, ratio estimation and estimating regression parameters.

1.14 Relationship between the jackknife and cross-validation methods

Stone (1974), applied a generalised form of a 'cross-validation' criterion to the choice and assessment of statistical prediction. Basically, this involves the partitioning of the data sample into two sub-samples, the choice of a statistical predictor, including any necessary estimation on one subsample and then, the assessment of its performance by measuring its prediction against the other sub-sample. The cross-validation criterion usually corresponds to the division of the sample (size n) into a 'construction' sub-sample (size n-1) and a 'validation' sub-sample (size 1) in all (n) possible ways. Stone developed the general framework and definitions to this resampling method as follows:

Consider a data sample S, of measurements (x,y) on each of n items, where x and y are quite general, such that

$$S = \{ (x_i, y_i) \mid i=1,\ldots.n \} \qquad (1.14.1)$$

Consider a new item for which only the x-value is known. It is then required to predict the y-value by $\hat{y}$, which is a function of x and s. The starting point is a 'prescription' (class of predictors)

$$\{\hat{y}(x, \alpha, S) \mid \alpha \in \mathcal{A}\} \qquad (1.14.2)$$

where the dependence of $\hat{y}(x,\alpha,S)$ on x and S is prescribed. The element of choice in (1.14.2) lies in allowing S to determine α. The method of cross-validatory choice of α and the method of cross-validatory assessment of this choice is developed in the following steps:

(i) A 'naive choice' of α is the value $\alpha^\circ(S) \in \mathcal{A}$ that minimises

$$L(\alpha) \overset{\text{def}}{=\!=} \frac{1}{n} \sum_{i=1}^{n} L\{y_i, \hat{y}(x_i; \alpha, S)\} \qquad (1.14.3)$$

where $L\{y,\hat{y}\}$ is a selected loss function of $\hat{y}$ as a predictor of the actual value y. For example, $L\{y,y\} = (y-\hat{y})^2$ would correspond to a least squares fitting procedure.

(ii) The 'naive assessment of this naive choice' would employ $L(\alpha^\circ(S))$ over the n items in S of $L\{y,\hat{y}(x;\,\alpha^\circ(S),S)\}$

(iii) The 'cross-validatory assessment of the naive choice' would employ

$$C^\circ \stackrel{\text{def}}{=\!=} \frac{1}{n}\sum_{i=1}^{n} L\{y_i,\hat{y}(x_i;\,\alpha^\circ(S_{\setminus i}),S_{\setminus i})\} \qquad (1.14.4)$$

where $S_{\setminus i}$ denotes the sample S with i^{th} item omitted and $\alpha^\circ(S_{\setminus i})$ is the naive choice of α which minimises

$$L_{\setminus i}(\alpha) \stackrel{\text{def}}{=\!=} \frac{1}{(n-1)} \sum_{\setminus i} L\{y_j,\,\hat{y}(x_j;\alpha,S_{\setminus i})\} \qquad (1.14.5)$$

where $\sum_{\setminus i}$ denotes the summation omitting the i^{th} item

(iv) The 'cross-validatory choice of α is the value $\alpha^+(S)\varepsilon\,\mathcal{A}$ that minimises

$$C(\alpha) \stackrel{\text{def}}{=\!=} \frac{1}{n}\sum_{i=1}^{n} L\{y_i,\hat{y}(x_i,\,\alpha,S_{\setminus i})\} \qquad (1.14.6)$$

(v) The 'cross validatory assessment of this cross-validatory choice' employs

$$C^+ \stackrel{\text{def}}{=\!=} \frac{1}{n}\sum_{i=1}^{n} L\{y_i,\hat{y}(x_i;\,\alpha^+(S_{\setminus i}),\,S_{\setminus i}\} \qquad (1.14.7)$$

where $\alpha^+(S_{\setminus i})$ is the cross-validatory choice of α which minimises

$$C_{\setminus i}(\alpha) \stackrel{def}{=\!=} \frac{1}{(n-1)} \sum_{\setminus i} L\,\{y_j, y(x_j, \alpha\,,\, S_{\setminus ij})\} \qquad (1.14.8)$$

where $S_{\setminus ij}$ denotes the sample with the i^{th} and j^{th} items omitted.

Two examples of a cross-validatory estimator are given by Frangos and Stone (1984), in estimating a proportion with batches of different sizes, as described in Section 1.9
i.e. consider the interval of values of the parameter of interest, μ, defined by $(\hat{\mu}_o, \hat{\mu}_1)$, where $\hat{\mu}_o$ and $\hat{\mu}_1$ are the pure-binomial and alternative estimator respectively. An estimator by cross-validatory choice of the index α is obtained in the prescription.

$$\{\, \hat{\mu}(\alpha) = \alpha\hat{\mu}_o + (1-\alpha)\,\hat{\mu}_1 \quad 0 \leqslant \alpha \leqslant 1\} \qquad (1.14.9)$$

Let $\hat{\mu}_{o\setminus i}$, $\hat{\mu}_{1\setminus i}$ denote the values taken by $\hat{\mu}_o$, $\hat{\mu}_1$ respectively when the i^{th} batch datum (n_i, x_i) is omitted. Two different choices of α, α^+ and α^+_w are considered that minimises the cross-validatory assessment criteria.

$$\sum\{P_i - \alpha\hat{\mu}_{o\setminus i} - (1-\alpha)\hat{\mu}_{1\setminus i}\}^2 \;,\; \sum n_i\{P_i - \alpha\hat{\mu}_{o\setminus i} - (1-\alpha)\hat{\mu}_{1\setminus i}\}^2 \qquad (1.14.10)$$

respectively. These choices give, when applied with the whole data set, the unweighted cross-validatory estimator.

$$\hat{\mu}^+ = \alpha^+\hat{\mu}_o + (1-\alpha^+)\hat{\mu}_1 \qquad (1.14.11)$$

and the weighted cross-validatory estimator

$$\hat{\mu}^+_w = \alpha^+_w\hat{\mu}_o + (1-\alpha^+_w)\hat{\mu}_1 \qquad (1.14.12)$$

Geisser (1971), described the cross-validatory approach as of 'predictive jackknife type'. The cross-validation and jackknife methods both employ the device of omission of items one or more at a time. However, the component of jackknifing that sharply distinguishes it from cross-validation is its manufacture of pseudovalues for the reduction of bias. Jackknifing the cross-validatory statistic is possible, although there is only one example of this, as yet, in the literature. This refers again to Frangos and Stone (1984), where the cross-validatory estimators (1.14.11) and (1.14.12) were jackknifed, in the 'proportion with batches of different sizes' example.

1.15 **Improvements in the jackknife method**

Hinkley (1977b), considered a modified Student t-approximation for the standardised jackknife estimator. If n=gh, where n is the sample size, g is the number of groups and h is the number of omissions, then Hinkley showed that, for small h, the jackknife variance estimate $V_{n,h}$ could behave quite unlike a multiple of X^2_{g-1}. In particular, the nominal g-1 degrees of freedom associated with $V_{n,h}$ might lead to inaccurate Student t approximations. Hinkely therefore considered a simple modification to the degrees of freedom, as follows: Firstly, approximate the distribution on $nV_{n,h}$ by $\sigma^2 X^2_f/f$. Then, estimate f from the data and use f as a replacement for g-1 as the degrees of freedom in the Student t approximation for jackknife confidence limits. In order to estimate f, Hinkley considered the function $gV_{n,h}$ where

$$V_{n,h} = \{g(g-1)\}^{-1} \sum_{j=1}^{g} (\hat{\theta}_j - \hat{\theta})^2 \qquad (1.15.1)$$

is the variance estimate of the jackknife estimator $\hat{\theta}$, to be the sample variance of g numbers $\hat{\theta}_j$, and used the formula for the jackknife variance of a sample variance to estimate the variance of $gV_{n,h}$. This leads to an estimate of $var(V_{n,h})$, where

$$K_{n,h} = \frac{\sum(\hat{\theta}_j-\hat{\theta})^4}{g(g-1)(g-2)^2} - \frac{gV^2_{n,h}}{(g-2)^2} \qquad (1.15.2)$$

Hinkley then derived the estimator for f,

$$f_{n,h} = 2V^2_{n,h}/K_{n,h} \qquad (1.15.3)$$

Although this approach is a rather crude use of double-jackknifing, the value $f_{n,s}$ is very easy to compute and should lend itself to the many Monte Carlo type simulations which can be used to evaluate the jackknife procedure. An evaluation of this approach with regard to jackknifing the 'availability'estimator is given in Chapter 3.

In a later paper, Hinkley and Wei (1984), considered improvements of jackknife confidence limits methods, by applying an Edgeworth expansion to the standard error, which was computed by a jackknife method.

Using the author's notation, they firstly define:

$$\hat{I}_j = (n-1)\ (T-T_{/j}) \qquad (j=1,\ldots\ldots,n) \qquad (1.15.4)$$

where $T_{/j}$ refers to the estimate of θ calculated from the sample with X_j omitted. The jackknife standard error $\hat{S}$ for an estimator T, is then defined by:

$$\hat{S} = \{\sum(\hat{I}_j-\hat{I})^2 / \{n(n-1)\}\}^{\frac{1}{2}} \qquad (1.15.5)$$

with $\hat{I} = n^{-1}\sum\hat{I}_j$

Then, using the influence function $I_T(x,F)$ of $T(F)$

$V(F) = \text{var}\{I_T(X,F)\}$, is estimated by $V(\hat{F}) = \sum\{\frac{I_T(X_j,\hat{F})}{n}\}^2$

Hence, the I.J.E. for the standard error of T is

$$\hat{S} = \{V(\hat{F})/\ n\}^{\frac{1}{2}} \qquad (1.15.6)$$