Using either method to calculate S, the approximate standard normal distribution of $Z = (T-\theta)/S$ provides approximate $(1-p)$ confidence limits $T + k_{\frac{1}{2}p}S$ for θ, where k_q denotes the q quantile of the standard normal distribution. In many applications, the effect of bias, skewness and so on can lead to the normal approximation not being very accurate. Hence, the use of Edgeworth expansions can be applied, primarily in the case of the I.J.E., to improve the situation.

Hinkley and Wei obtained an Edgeworth expansion for the distribution of $\hat{Z} = (T-\theta)/\hat{S}$ where $\hat{S}$ is the I.J.E. of the standard error given by (1.15.6). The results of the expansion are simply stated in this section rather than all the steps that were used in the paper.

Let α_{11} and α_{22} correspond respectively to the standardised first-order biases of T and $V(\hat{F})$; while α_{32} and α_{43} are measures of skewness and kurtosis of the linear approximation $n^{-1}\sum I_T(X_j,F)$ for $T-\theta$.

The studentized quantity $\hat{Z}$ can be expressed in the form:

$$\hat{Z} = \sqrt{n}w(\hat{F},F), \tag{1.15.7}$$

where $w(G,F) = \{T(G) - T(F)\} / \sqrt{V(G)}$,
and $\hat{Z}$ then has the Edgeworth expansion:

$$\text{pr } (Z \leqslant z) = \Phi(z) - \phi(z) \sum_{j=1}^{2} n^{-\frac{1}{2}j} \lambda_j(z) + o(n^{-1}), \tag{1.15.8}$$

where Φ and ϕ are the standard normal distribution and density functions respectively.

Now, if $\hat{\alpha}_{jk} = \alpha_{jk}(\hat{F})$ and $\hat{\alpha}_{jk} = \alpha_{jk} + O_p(n^{-\frac{1}{2}})$, then (1.15.9) becomes

$$\text{Pr } (\hat{Z} \leqslant z) = \Phi(z) - n^{-\frac{1}{2}} \{\hat{\alpha}_{11} + \frac{1}{6}\hat{\alpha}_{32}(z^2-1)\}\phi(z) + O_p(n-1)$$

By adjusting k_q to

$$k_q^* = k_q + n^{-\frac{1}{2}}\{ \hat{\alpha}_{11} + \frac{1}{6}\hat{\alpha}_{32}(k_q^2 - 1) \} , \qquad (1.15.10)$$

leads to $\Pr(\hat{Z} \leqslant k_q^*) = q + O(n^{-1})$

Hence, the pair of limits

$$T-k^*_{1-\frac{1}{2}p}\hat{S},\ T-k^*_{\frac{1}{2}p}\hat{S} \qquad (1.15.11)$$

ignoring terms of order n-1, give coverage frequency of 1-p for the parameter θ.

Applying Edgeworth expansions to the standard jackknife as far as the first-order correction is concerned, leads to a simple analogue of the corrected quantile (1.15.10).

Firstly, the standard jackknife quantities $\hat{I}_j$ defined in (1.15.4) can be expressed as $\hat{I}_j = \hat{I}_j - \frac{1}{2}n^{-1} Q_t(X_j, X_j, \hat{F}) + o_p(n^{-1})$ (see Hinkley and Wang; 1980).

The standard jackknife estimator and the I.J.E. for the standard error therefore satisfy:

$$\frac{n-1}{n}\hat{S}^2 = \hat{S}^2 - n^{-2}u(\hat{F}) + o_p(n^{-2})$$

where $u(F) = E\left\{ I_T(X,F)\, Q_T(X,X,F) \right\}$

If $r(F)$ is set to $u(F)/v(F)$, and since $r(\hat{F}) = r(F) + o_p(1)$, then,

$$\hat{Z} = \frac{T-\theta}{S} = \hat{Z}\left\{1 - \tfrac{1}{2}n^{-1}\left\{r(F)-1\right\}\right\}^{-1} + o_p(n^{-1})$$

As there is a constant factor in the square brackets above, the Edgeworth expansion (1.15.9) applies to Z with modification of the n^{-1} term.

Hence, the corrected confidence limits analogous to (1.15.11) can be obtained by using consistent jackknife replacements for $\hat{\alpha}_{11}$ and $\hat{\alpha}_{32}$ in (1.15.10).

Let $\hat{Q}_{jk} = n\{nT-(n-1)(T_{/j} + T_{/k}) + (n-2)T_{/(j,k)}$ $(j \neq k)$ (1.15.12)

estimate $Q_t(X_j, X_k, F)$

Hinkley and Wei then estimated the jackknife first-order bias of T and the skewness of the linear approximation $n^{-1}. \sum I_t(X_j,F)$ for $T-\theta$ as follows:

$$\hat{\alpha}_{11} = -\sum \hat{I}_j / \hat{V} - \frac{1}{2}(n^{-1}\sum \hat{I}_j^3 + 2n^{-2}\sum_{j \neq k} \hat{I}_j \hat{I}_k \hat{Q}_{jk})/\hat{V}^{3/2} \quad (1.15.13)$$

$$\hat{\alpha}_{32} = -(2n^{-1}\sum \hat{I}_j^3 + 3n^{-2}\sum_{j \neq k} \hat{I}_j \hat{I}_k \hat{Q}_{jk})/\hat{V}^{3/2} \quad (1.15.14)$$

where $\hat{V} = n\hat{S}^2$.

The corrected confidence limits will now be:

$$T - k^{**}_{1-\frac{1}{2}p}\hat{S}, \quad T - k^{**}_{\frac{1}{2}p}\hat{S}$$

where $k^{**}_q = k_q + n^{-\frac{1}{2}}\{\hat{\alpha}_{11} + \frac{1}{6}\hat{\alpha}_{32}(k_q^2 - 1)\}$

If T is replaced by $T + \hat{I}$ and the first term is removed from (1.15.13), the limits (1.15.15) can then be expressed in terms of the bias corrected estimate. The results of Hinkley and Wei (1984), are very interesting, although the application of these results through Monte Carlo simulation studies, based on large sample data for normal approximations, may however be limited, due to the considerable computation time required.

1.16 Open questions for further research

Where should research on the jackknife go from here? What worthwhile questions on the jackknife still remain unanswered? The following list includes some thoughts on the matter but they are by no means all-inclusive.

(i) A problem which arises in applying the jackknife method is the determination of the minimum number of random variables which are required in order for Tukey's asymptotic result to hold. Monte Carlo studies could be performed using different model distributions in order to throw some light into the solution of this problem.

(ii) Is there any connection between the jackknife method and the Bayesian theory? Suppose there is a sample of n i.i.d. random variables with distribution $F(\theta,x)$. After splitting the sample into n groups to find the jackknife estimator of θ, the n pseudovalues $\hat{\theta}_1, \hat{\theta}_2,\ldots,\hat{\theta}_n$ are derived. Tukey suggested that the pseudovalues are approximately distributed as a random sample from a normal distribution with mean θ and variance σ^2. Suppose that the prior distribution of θ, given σ^2, is G. An open question which requires additional research is to find in this case the jackknife estimator of θ.

(iii) Let $X_1, X_2,\ldots,X_n$ be n random variables which are <u>not</u> independent but there is a small correlation between them. Suppose they are distributed according to $F(\theta,x)$. How can the parameter θ be estimated by the jackknife method?

(iv) Can a lower bound for Tukey's variance of the jackknife be established?

(v) Suppose the jackknife pseudovalues are developed by adding and not subtracting a new observation at a time. Is the new estimator as efficient as the original jackknife and do these new 'pseudovalues' satisfy Tukey's conjecture?

(vi) In the application of the jackknife method, the reduction of the mean square error should play an important role besides the reduction of bias and variance.

(vii) The influence curve for the parameter β in the linear model $Y=\beta x+\alpha$ where $e \sim N(0,\sigma^2)$, behaves like a residual. Therefore it can be used to discover possible discrepancies in the sample.

(viii) The jackknife method could be used in the linear model where the errors are non-homogeneous. This could be done by theoretical as well as by Monte Carlo studies for various models.

(ix) Estimators based on single order statistics such as the median do not behave properly under jackknifing. Smooth functions of the order statistics such as $\hat{\theta} = \sum J(i/n)\, Y_{(i)}/n$, where J is a continuous function, could be considered.

(x) Most researchers of the jackknife technique would suggest using a variance stabilizing transformation on the estimator in conjunction with the jackknife. Transformations are needed to keep the jackknife on scale and thus prevent a distortion of the results. The connection between transformations and the jackknife is worth more exploration. Is there an optimal way to select a transformation for use with the jackknife?

(xi) The jackknife is not a device for correcting outliers. How do outliers perturb the jackknife estimator? Is the examination and correction of pseudovalues a good way of handling outliers?

(xii) How should the jackknife method be applied in a multi-sample problem. For example, consider a two-sample problem. One way to jackknife is to compute an estimate of the unknown parameter from the two samples and then jackknife by successively deleting each observation in the first sample with the second sample intact and then deleting observations in the second sample with the first intact. Alternatively, each sample could be jackknifed separately and the results combined. Both methods are valid asymptotically, but which one gives the better results?

(viii) The efficiency of the jackknife could be examined when the sample is split according to a prescribed structure of the data or when the groups are unequal.

CHAPTER 2

CRITICAL REVIEW OF THE BOOTSTRAP METHOD

2.1 Derivation and description of the bootstrap estimator

The bootstrap, (as mentioned in Section 1.13), is a member of the family of general resampling methods, which are feasible today because of the availability of high speed computers. This 'computer intensive' method which allows freedom from the assumption of normality, was developed by B.Efron in a series of papers (see Efron 1979a, 1979b, 1981a, 1981b, 1982). Efron gave the term 'bootstrap' to the method because the estimator is derived from the treatment of the original sample, which is analogous to boots which one straps in order to tie them.

The bootstrap method consists of approximating the distribution of the statistic of interest, by what Efron calls the bootstrap distribution of this statistic. This distribution is obtained by replacing the unknown distribution of the statistic by the empirical distribution of the observations, and then resampling the data to obtain a Monte Carlo distribution for the resulting random variable.

A detailed description of the method is as follows:
Consider a real-valued statistic $\hat{\theta}(X_1,X_2,\ldots,X_n)$ which is a function of n i.i.d. observations:

$$X_1,X_2,\ldots X_n \overset{\text{i.i.d.}}{\sim} F \tag{2.1.1}$$

where F is an unknown probability distribution.
The standard error of $\hat{\theta}$ is a function of F and the form of the statistic $\hat{\theta}$.

$$\text{i.e.} \quad \sigma(F,n,\hat{\theta}(.,.,\ldots,.))=\sigma(F) \tag{2.1.2}$$

The bootstrap estimate of the standard error, $\hat{\sigma}_B$, evaluated at $F=\hat{F}$, is simply

$$\hat{\sigma}_B=\sigma(\hat{F}), \qquad (2.1.3)$$

where $\hat{F}$ is the empirical probability distribution which places mass $1/n$ at each point $X_1,X_2,\ldots,X_n$.

The function $\sigma(F)$ cannot usually be expressed in a simple form, and $\hat{\sigma}_B$ must be calculated using a Monte Carlo algorithm. The algorithm consists of the following steps:

Step 1 : Construct $\hat{F}$ as defined above

Step 2 : Draw a random 'bootstrap sample' from $\hat{F}$, with replacement, and denote the new sample by $X_1^*,X_2^*,\ldots,X_n^*$

i.e. $X_1^*,X_2^*,\ldots,X_n^* \overset{i.i.d.}{\sim} \hat{F}$.

This process can be carried out as follows

(a) Set the observed values to be $x_{(1)},x_{(2)},\ldots,x_{(n)}$

(b) Draw a pseudo-random variable from the discrete uniform distribution in the range $\{1,n\}$, and let r be the number

(c) Put $X_1^*=x_{(r)}$

(d) Repeat steps a, b and c, n times to obtain the bootstrap sample

$X_1^*,X_2^*,\ldots,X_n^*$.

Then, after steps a, b, c and d, calculate the statistic

$$\hat{\theta}^*=\hat{\theta}(X_1^*,X_2^*,\ldots,X_n^*) \qquad (2.1.4)$$

Step 3 : Repeat step 2, independently, a large number B of times, to obtain the 'bootstrap replications' $\hat{\theta}^*(1),\hat{\theta}^*(2),\ldots, \hat{\theta}^*(B)$.

Then calculate

$$\hat{\sigma}_B=\left\{\sum_{b=1}^{B} \frac{(\hat{\theta}^*(b)-\hat{\theta}^*(.))^2}{B-1}\right\}^{\frac{1}{2}} \qquad (2.1.5)$$

$$\text{where } \hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B \tag{2.1.6}$$

As $B \to \infty$, $\hat{\sigma}_B$ converges to $\hat{\sigma}(F)$. In practice Efron has found B in the range 200 to 500 adequate for estimating standard errors.

2.2 **Double bootstrap and smoothed bootstrap methods :**

As in the case of the second order jackknife estimate, there is a double bootstrap estimator of the standard error of a statistic. Efron (1982) has shown that the bootstrap estimate, $\hat{\sigma}_B$, is usually an under-estimate of the true standard error. The double bootstrap method can be considered as 'bootstrapping the bootstrap' and consists of the following steps:

Step 1 : Construct $\hat{F}$ as before

Step 2 : Draw a 'bootstrap sample' from $\hat{F}$

$X_1^*, X_2^*, \ldots, X_n^* \overset{\text{i.i.d.}}{\sim} \hat{F}$.

and calculate the statistic $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$.

Step 3 : Repeat step 2, independently, B times obtaining the 'bootstrap replications' $\hat{\theta}^*(1), \ldots, \hat{\theta}^*(B)$.

Then, calculate

$$\hat{\sigma}_B = \left\{ \sum_{B=1}^{B} \frac{\{\hat{\theta}^*(b) - \hat{\theta}^*(.)\}^2}{B-1} \right\}^{\frac{1}{2}}$$

Step 4 : Each bootstrap sample $X_1^*, \ldots, X_n^*$ is considered as the original sample, and is then used to repeat, independently, steps 1, 2 and 3. Hence, for each bootstrap sample, the B 'double bootstrap' samples are calculated. From each set of 'double bootstrap' samples, the bootstrap standard error estimates $\hat{\sigma}^{**}(1), \ldots, \hat{\sigma}^{**}(B)$. are calculated. The double bootstrap standard error estimate is then

$$\hat{\sigma}_B(\text{Double}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\sigma}^{**}(b) \tag{2.2.1}$$

The smoothed bootstrap method introduces an amount of 'smoothness' to F. In drawing the bootstrap sample, instead of choosing randomly each X_i^* from the set $(x_1, x_2, \ldots, x_n)$, a continuous variable X_i^* is considered, such that

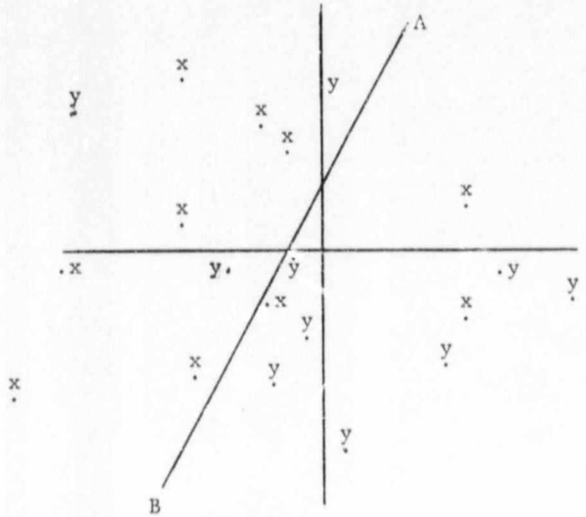
$$X_i^* = \bar{x} + C\,\{x_{I_i} - \bar{x} + \hat{\sigma} Z_i\} \tag{2.2.2}$$

where the I_i are chosen independently and randomly from the set $\{1, 2, \ldots, n\}$ and the Z_i is a random variable from a given distribution, having mean 0 and variance σ_z^2. The normal distribution is the most obvious choice for the Z_i. The quantities $\bar{x}$, $\hat{\sigma}$ and C are the sample mean, standard deviation and $(1+\sigma_z^2)^{\frac{1}{2}}$, respectively, so that X_i^* has mean $\bar{x}$ and variance $\hat{\sigma}^2$ under the bootstrap sampling procedure.

2.3 **An example of a bootstrap estimator :**

Efron (1979b) applied the bootstrap method in linear discriminant analysis. He considered the case of 20 random numbers, 10 of which were drawn from an x bivariate normal population with mean vector $(-\frac{1}{2}, 0)'$ and 10 from a y bivariate normal population with mean vector $(\frac{1}{2}, 0)'$, such that the common covariance matrix is equal to 1. The straight line AB is the linear discriminant boundary.

REGION A



REGION B

The linear discriminant function (L.D.F.) is defined as:

$$\left\{ w: (\bar{y}-\bar{x})' S^{-1} (w - \frac{\bar{x}+\bar{y}}{2}) = 0 \right\} \tag{2.3.1}$$

where

$$\bar{x} = \sum \frac{x_i}{10}, \quad \bar{y} = \sum \frac{y_i}{10}$$

and

$$S = \sum (x_i - \bar{x})(x_i - \bar{x})' + \sum (y_i - \bar{y})(y_i - \bar{y})'$$

The line AB divides the plane into two regions, A and B, such that an unlabelled future point w can be classified as being either an x or y, depending on whether it falls into A or B. The probability that a future x random number will be misclassified, is defined as:

$$err_x = \text{Prob}\{x \varepsilon B\} \tag{2.3.2}$$

An estimate of err_x is thus:

$$\widehat{err}_x = \frac{\text{number of } \{x_i \varepsilon B\}}{10} \tag{2.3.3}$$

Of interest would be to then estimate

$$bias_x = E\{err_x - \widehat{err}_x\} \tag{2.3.4}$$

The bootstrap estimate of $bias_x$ can be calculated as follows:

(i) Draw a random bootstrap sample, with replacement, from the original sample $x_1, x_2, \ldots, x_{10}$ and denote the new sample by $x_1^*, x_2^*, \ldots, x_{10}^*$. Similarly, create a bootstrap sample $y_1^*, y_2^*, \ldots, y_{10}^*$ from the 10 y points.

(ii) Using equation (2.3.1), substitute $\bar{x}^*$, $\bar{y}^*$ and $\bar{s}^*$, for x, y and S respectively. Construct the L.D.F. line and denote the bootstrap discriminant regions as A^* and B^*.

(iii) Let $$\text{bias}_x^* = \frac{\text{no. of } \{x_i \in B^*\}}{10} - \frac{\text{no. of } \{x_i^* \in B^*\}}{10} \qquad (2.3.5)$$

(iv) Repeat steps (i), (ii) and (iii), a large number B of times, to obtain independent values

$$\text{bias}_x^*(1),\ \text{bias}_x^*(2),\ldots,\ \text{bias}_x^*(B)$$

Then, the bootstrap estimate of bias is:

$$\text{bias}_x(\text{bootstrap}) = \frac{1}{B}\sum_{b=1}^{B} \text{bias}_x^*(b) \qquad (2.3.6)$$

2.4 **Bootstrapping regression models :**

The bootstrap method is valid (see Friedman,1981), for the regression model:

$$Y_i = f_i(\beta) + \varepsilon_i\ ,\ i=1,2,\ldots,n \qquad (2.4.1)$$

where f(.) is a known function of the unknown parameter vector β. The ε_i's are independent and identically distributed random variables with common distribution function F, such that $E_F(\varepsilon)=0$.

From the observed data X=x, the least squares estimate $\hat{\beta}$ is defined as:

$$\hat{\beta} : \min_\beta \sum_{i=1}^{n} \{x_i - f_i(\beta)\}^2 \qquad (2.4.2)$$

The bootstrap distribution of $\hat{\beta}$ is calculated as follows:

(i) Construct $\hat{F}$ such that

$$\hat{F}\text{: mass } \frac{1}{n} \text{ at } \hat{\varepsilon}_i = y_i - f_i(\hat{\beta}),\ i=1,2,\ldots,n$$

(ii) Draw with replacement from the observed residuals $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n$ and construct the bootstrap residuals $\varepsilon^* = (\varepsilon_1^*,\ldots,\varepsilon_n^*)$. Calculate the new random variable,

$$X_i^* = g_i(\hat{\beta}) + \varepsilon_i^* \quad , \quad \hat{\varepsilon}_i \overset{iid}{\sim} \hat{F} \ , \ i=1,2,\ldots,n \tag{2.4.3}$$

and consider the model

$$X_i^* = g_i(\hat{\beta}) + \varepsilon_i^{**} \quad , \ i=1,2,\ldots,n \tag{2.4.4}$$

From this model, calculate by the least squares method, the estimator $\hat{\beta}^*$ of β

(iii) Repeat independently the previous steps a large number B of times to obtain the bootstrap replications:

$\hat{\beta}^*(1),\ \hat{\beta}^*(2),\ldots,\hat{\beta}^*(B).$

Then calculate the bootstrap estimator of β, $\hat{\beta}^*(.)$, as follows:

$$\hat{\beta}^*(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^*(b) \tag{2.4.5}$$

with corresponding standard error estimator,

$$\hat{\sigma}_B = \left\{ \sum_{b=1}^{B} \frac{\{\hat{\beta}^*(b)-\hat{\beta}^*(.)\}^2}{B-1} \right\}^{\frac{1}{2}} \tag{2.4.6}$$

2.5 **Bootstrapping the mean :**

In a paper by Bickel and Freedman (1981), they considered the asymptotic validity of bootstrapping the mean.

Consider $X_1, X_2, \ldots, X_n$ to be independent random variables with common distribution function F. Assume that the mean μ and the variance σ^2 of F, are finite but unknown. The standard estimates for μ and σ^2 are calculated by the sample average μ_n, and the sample standard S_n, respectively.

$$\text{i.e. } \mu_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.5.1}$$

$$\text{and } S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_n)^2$$

Also, the distribution of the pivotal quantity

$$Q_n = \sqrt{n}(\mu_n - \mu)/S_n \tag{2.5.2}$$

tends weakly to N(0,1) from the Central Limit theorem. Although the asymptotics are known in this case, Bickel and Friedman investigated how the bootstrap would perform.

Let F_n be the empirical distribution of $X_1, \ldots, X_n$, putting mass $1/n$ on each X_i. Given $(X_1, \ldots, X_n)$, resample the data in the usual manner to obtain the bootstrap replications $X_1^*, \ldots, X_m^*$, where the resample size m may differ from the number n of data points, to e timate the distribution of the bootstrap pivotal quantity $Q_m^* = \sqrt{m}(\mu_m^* - \mu_n)/S_m^*$, where $\mu_m^* = (1/m).\ \sum_{i=1}^{m} X_i^*$ and $S_m^* = (1/m) \sum_{i=1}^{m} (X_i^* - \mu_m^*)^2$.

In the case of m=n, the distribution of Q could be computed from the data and used to approximate the unknown sampling distribution of Q_n. Alternatively, the bootstrap distribution of $\sqrt{n}(\mu_n^* - \mu_n)$ could be used to approximate the sampling distribution of $\sqrt{n}(\mu_n - \mu)$. Confidence intervals for μ could be derived from either approach, and the bootstrap method would therefore be useful if the Central Limit theorem were not available.

In the case of $m \neq n$, Bickel and Friedman used the following theorem: Suppose $X_1, X_2, \ldots$ are independent, identically distributed, and have finite positive variance σ^2. Along almost all sample sequences $X_1, X_2, \ldots$, given $(X_1, \ldots, X_n)$, as n and m tend to ∞:

(a) The conditional distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ converges weakly to $N(0, \sigma^2)$.

(b) $S_m^* \to \sigma$ in conditional probability: that is, for ε positive,

$$P\left\{ |S_m^* - \sigma| > \varepsilon | X_1, \ldots, X_n \right\} \to 0 \quad \text{a.s.} \tag{2.5.1}$$

Consider the situation where m and n are free to vary separately. The conditional law of $\sqrt{m}(\mu_m^* - \mu_n)$ tends to normal, with mean 0 and variance S_n^2, if m goes to infinity first. As n tends to infinity, this converges if and only if S_n^2 does.

For the simple case m=n, the parameter μ is replaced by μ_n when comparing the classical $\sqrt{n}(\mu_n - \mu)$ with the bootstrap $\sqrt{n}(\mu_n^* - \mu_n)$. This change represents an order of magnitude $1/\sqrt{n}$. Also, the X's have been replaced by X^*'s, resulting in a second error. Bickel and Freedman proved that the two errors almost cancel each other out, by showing that the distribution of the pivot does not change much when the empirical distribution F_n is replaced by the theoretical F. They showed that the bootstrap works for means as well as for pivotal quantities of the familiar 't-statistic' sort.

2.6 **Bayesian bootstrap** :

Rubin (1981), considers the Bayesian bootstrap (B.B.) as the Bayesian analogue of the bootstrap. The Bayesian bootstrap simulates the posterior distribution of the parameter, instead of simulating the sampling distribution of a statistic estimating a parameter. Rubin illustrates the B.B. in two simple examples, namely dichotomous X and bivariate X, as described by Efron (1979b).

2.7 **Bootstrap selection procedures :**

In an interesting paper by Swanepoel (1983), a new bootstrap procedure based on robust estimators is formulated which, in the small sample case, is shown through Monte Carlo experiments, to improve on the existing methods.

Many practical situations involve the problem of comparing several competing populations, treatments or processes. Gupta (1956), outlined one of the basic formulations of the selection problem, namely the subset selection formulation, which selects a random sized non-empty, (small), subset of the populations, at the same time controlling the probability of including the 'best' population in the selection subset (the term 'best' population is defined by Swanepoel in his paper). Prior knowledge of the parameter values is, however, usually not available.

Other basic goals such as the problem of selecting a subset to contain only δ-optimal populations or all populations better than a control have been well documented by Fabian (1962), Desu (1970), Carroll, Gupta and Huang (1975), Santer (1976), Panchapakesan and Santer (1977), Chen (1980), amongst others. Most of the research is based on parametric methods which assume, mainly either normal, binomial or multinomial distribution forms of the underlying observations. Bootstrap methods are free from this constraint and Swanepoel used these methods to obtain distribution-free selection procedures based on given robust estimators.

2.8 **Bootstrap and censored data :**

In Efron's earlier paper (1979b), a series of examples were considered, to show the applicability of the bootstrap method under a variety of situations. In a later paper (Efron, 1981a), a further example is considered involving right-censored data, which is used to derive the Kaplan-Meier (product limit) estimated survival curve (Kaplan and Meier, 1958). In particular, the bootstrap method is used to calculate the standard error of the Kaplan-Meier curve and to derive the standard error of a location estimate such as the trimmed mean with the appropriate confidence limits. Computer simulations to investigate the validity of the bootstrap method, are based on real data, namely the 'Channing House' data, which considered the survival curve of ninety seven men who lived in Channing House, a Palo Alto retirement centre, from its opening in 1964 to the data collection day, July 1, 1975.

Consider right-censored data of the form $\{(x_1,d_1),(x_2,d_2),\ldots,(x_n,d_n)\}$ where x_j is the j^{th} observation, censored or not, and

$$d_j = \begin{cases} 1 \text{ if } x_j \text{ is uncensored} \\ 0 \text{ if } x_j \text{ is censored} \end{cases} \tag{2.8.1}$$

under the assumption $x_1 < x_2 \ldots < x_n$.

The bootstrap estimate of accuracy, $\hat{\sigma}_{BOOT} = \sigma(\hat{F})$, where $\hat{F}$ is the empirical d.f., is calculated as follows:
Draw a bootstrap sample $(X_1^*,D_1^*),(X_2^*,D_2^*),\ldots,(X_n^*,D_n^*)$ by independent sampling n times with replacement from $\hat{F}$, the distribution putting mass $^1/_n$ at each points (x_j,d_j). Then, letting data* represent this artificial data set, calculate $\hat{\theta}^*=\theta(\text{data}^*)$, where $\hat{\theta}$ is an estimated functional such as the median. Next, repeat the previous steps N times, obtaining $\hat{\theta}^{*1},\hat{\theta}^{*2},\ldots,\hat{\theta}^{*N}$ and calculate $\hat{\sigma}_{BOOT}$ by the sample standard deviation formula:

$$\hat{\sigma}_{BOOT} = \sqrt{\frac{\sum_{j=1}^{N} (\hat{\theta}^*j)^2 - (\sum_{j=1}^{N} \hat{\theta}^*j)^2/N}{N-1}} \tag{2.8.2}$$

Efron considers statistics of the form $\hat{\theta}=\theta(\hat{S}_o)$, where $\hat{S}_o(t)$ is the Kaplan-Meier curve. It is possible to write $S_o=\Phi(\hat{F})$, where Φ is a certain mapping, from distributions on $R^1 \times \{0,1\}$ to distributions on R^1, as described in Peterson (1977).

Consider the random censorship model (Efron 1967), Gilbert (1962))

$$X_i = \min \{X_i^\circ, W_i\} \qquad (2.8.3)$$

where X_i° is the variable of interest and W_i is some independent censoring variable. The observed quantity is the pair (X_i, D_i) such that D_i equals 1 or 0 as X_i equals X_i° or W_i, respectively.

The Kaplan-Meier curve estimate $\hat{S}^\circ(t)$ of the true survival curve for X°, say $S^\circ(t)=\text{prob}\{X^\circ > t\}$, is nearly unbiased, and is given by the formula:

$$\hat{S}^\circ(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1}\right)^{d_j}$$

where k_t is the value of k such that $t \varepsilon [x_k, x_{k+1})$.

Consider drawing a single bootstrap sample $(X_1^*, D_1^*), (X_2^*, D_2^*), \ldots, (X_n^*, D_n^*)$ and define

$$m_j^* = \# \text{ of times } (x_j, d_j) \qquad (2.8.4)$$

appears in the bootstrap sample. Then $m^*=(m_1^*, m_2^*, \ldots, m_n^*)$ is an n-category multinomial, n draws, probability $1/n$ for each category:

$$m^* \sim \text{mult}(n, 1/n).$$

Also define

$$M_j^* = \sum_{i=j}^{n} m_i^* \text{ , } j=1,2,\ldots,n \qquad (2.8.5)$$

such that $M_1^*=n$, $M_2^*=n-m_1^*$ and so forth.

The Kaplan-Meier curve based on the bootstrap data $\{(x_1^*, d_1^*), (x_2^*, d_2^*), \ldots, (x_n^*, d_n^*)\}$ is

$$\hat{S}^{o*}(t) = \prod_{j=1}^{k_t} \left(1 - \frac{m_j^*}{M_j^*}\right)^{d_j} \tag{2.8.6}$$

Hence, the bootstrap estimate of standard deviation for $\hat{S}^o(t)$, t fixed, is

$$\hat{\sigma}_{BOOT} = \sqrt{\text{var}^* \; \hat{S}^{o*}(t)} \tag{2.8.7}$$

where 'var*' denotes the variance of (2.8.6) with the observed data $\{(x_1,d_1), (x_2,d_2),\ldots,(x_n,d_n)\}$ fixed and the vector m^* varying according to the multinomial distribution mult$(n, {}^1/n)$.

The usual formula for the standard deviation of $\hat{S}^o(t)$, is given by the "Greenwood formula".

$$\hat{\sigma}_{GREEN} = \hat{S}^o(t) \sqrt{\sum_{j=1}^{k_t} \frac{d_j}{(n-j)(n-j+1)}} \tag{2.8.8}$$

(see Kaplan and Meier (1958)).
Using the Channing House data, a comparison between $\hat{\sigma}_{GREEN}$ and $\hat{\sigma}_{BOOT}$ was made at nine different values of t: $x_{10}, x_{20}, \ldots, x_{90}$. The values of $\hat{\sigma}_{BOOT}$ were derived by Monte Carlo simulation, with N=400. The results indicated an excellent agreement between $\hat{\sigma}_{GREEN}$ and $\hat{\sigma}_{BOOT}$.

Using the bootstrap method, seven location estimates for the Channing House data were calculated. These included, for the estimated biases and standard deviations, the median, the mean, and various trimmed and Winsorized means. The estimated bias and variance both decreased as the amount of trimming or Winsorizing decreased, the worst case being the median and best being the mean. It is interesting to note that, although the median is often favoured as a location estimate in censored data problems, in this example, the superiority of the sample median as a point estimator, was not demonstrated.

A simple method for constructing confidence intervals based on the bootstrap distribution, known as the percentile method (see Section 2.9), was used to calculate 90% central confidence intervals for the mean, in the Channing House data. Also, using the percentile method, several Monte-Carlo experiments were run to calculate the actual coverage probabilities for the median. In these experiments, both X_i^o and W_i had exponential distributions, prob $\{X_i^o > t\} = e^{-t}$ and prob $\{W_i > t\} = e^{-t/c}$, $t \geq 0$. The probability of $X_i = \min\{X_i^o, W_i\}$ being uncensored, that is, $D_i = 1$, equals $c/(c+1)$. In the eight Monte Carlo experiments carried out, there was generally good agreement between the nominal and actual coverage probabilities for the median.

2.9 Improved bootstrap confidence limits :

In an interesting paper by Efron (1981b), the bootstrap is used to determine non-parametric standard errors to a real-valued statistic, in the case of small sample situations. Under the assumption of approximate normality, estimated standard errors provide rough confidence intervals: $\hat{\theta} + z_\alpha \hat{\sigma}$, where z_α is the α-point of a standard normal distribution. However, for small sample situations, confidence intervals are frequently highly asymmetric about the best point estimate $\hat{\theta}$. The correction made by replacing normal quantiles with Student t quantiles (replacing $\hat{\theta} \pm z_\alpha \hat{\sigma}$ by $\hat{\theta} \pm t_\alpha \hat{\sigma}$, with t_α the α-point of an appropriate t-distribution) is of magnitude $O(\frac{1}{n})$, whereas the actual magnitude of the asymmetry is $O(\frac{1}{\sqrt{n}})$. Efron considered two non-parametric methods, the percentile and the bias corrected percentile method, which attempts to capture the correct degree of asymmetry when assigning confidence intervals.

In Efron's paper (1981b), a simple correlation example is considered consisting of n=15 pairs of points. Having calculated the Pearson correlation coefficient $\hat{\rho}$ in the normal manner, a histogram of B = 1000 bootstrap replications $\hat{\rho}^*$, was constructed with the abscissa plotted in terms of $\hat{\rho}^* - \hat{\rho}$. Also plotted was the normal-theory density curve $f_{\hat{\rho}}(\hat{\rho}^*)$ for the given value of $\hat{\rho}$, against $\hat{\rho}^* - \hat{\rho}$.

There was a close similarity between the histogram and $f_{\hat{\rho}}(\hat{\rho}^*)$, suggesting that a simple method for constructing non-parametric confidence intervals was possible.

Let $\widehat{CDF}(t)$ denote the cumulative bootstrap distribution for some real-valued functional statistic $\hat{\theta}=\hat{\theta}(F)$:

$$\widehat{CDF}(t)=\text{Prob}_*\{\hat{\theta}^* < t\} = \frac{\#\{\hat{\theta}^*(b) < t\}}{B} \qquad (2.9.1)$$

where 'Prob$_*$' relates to the bootstrap probability, calculated by obtaining bootstrap replications in the normal manner.

$$\text{Define } \hat{\theta}(\alpha) = \widehat{CDF}^{-1}(\alpha) \qquad (2.9.2)$$

for a given value of α between 0 and 1, as a putative $(1-2\alpha)$ central confidence interval for the parameter $\theta=\theta(F)$. Efron called this approach the percentile method and used the method in the correlation example. With $\alpha = 0.16$, 1000 bootstrap replications gave $\rho\varepsilon\{0.654,0.908\} = \{\hat{\rho}-0.12,\hat{\rho} + 0.13\}$ as a central 68% interval for ρ, compared to the standard normal-theory interval $\{\hat{\rho}- 0.16, \hat{\rho}+ 0.09\}$, obtained by inverting

$$\phi \sim N(\phi + \frac{\rho}{2(n-1)}, \frac{1}{(n-3)})$$

The bias-corrected percentile method considers

$$\theta \,\varepsilon\, \{\widehat{CDF}^{-1}(\Phi(2z_0+z_\alpha)), \widehat{CDF}^{-1}(\Phi(2z_0+z_{1-\alpha}))\} \qquad (2.9.3)$$

where

$$z_0=\Phi^{-1}\widehat{CDF}(\hat{\theta}) \text{ and } \Phi(z)= (1/\sqrt{2\pi}) \int_{-\infty}^{z} e^{-s^2/2}\, ds \qquad (2.9.4)$$

For the correlation example, the bias-corrected central 68% interval was

$$\rho \varepsilon \{\widehat{CDF}^{-1}\Phi(-1.34),\ \widehat{CDF}^{-1}\Phi(0.66)\} = \{\hat{\rho} - 0.17,\ \hat{\rho} + 0.10\}$$

which is almost the same as the normal theory interval.

Efron also investigated the results from 10 Monte-Carlo trials with $X_1, X_2, \ldots, X_{15}$ bivariate normal, true correlation $\rho = 0.5$, using normal theory, the percentile method and the bias-corrected percentile method. The bias-corrected intervals gave an excellent representation of the correct left-right asymmetry in the normal theory intervals although the percentile method did not do this on a consistent basis.

In a recent paper by Efron (1987), the problem of setting approximate confidence intervals for a single parameter θ in a multi-parameter family, was further investigated using bootstrap intervals.

Approximate confidence intervals for θ are usually based on the maximum likelihood estimate

$$\theta \varepsilon \hat{\theta} + \hat{\sigma} z^{(\alpha)} \tag{2.9.5}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, $\hat{\sigma}$ an estimate of its standard deviation, usually based on the Fisher information matrix, and $z^{(\alpha)}$ the α point of a standard normal variate. Improvement in convergence to normality and constancy of σ can, in certain cases, be improved by considering a monotone transformation $\hat{\phi} = g(\hat{\theta})$ and $\phi = g(\theta)$, instead of $\hat{\theta}$ and θ. Efron's bias corrected percentile method assumes that normality and constant standard error can be achieved by some transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$, say

$$\frac{\hat{\phi} - \phi}{\tau} \sim N(-z_0, 1) \tag{2.9.6}$$

where τ is the constant standard error of $\hat{\phi}$ and z_0 is the bias constant.

In this latter paper, Efron introduces the improved bootstrap method, termed BC_a, which assumes that for some monotone transformation g, some bias constant z_0 and some 'acceleration constant' a, the transformation $\hat{\phi}=g(\hat{\theta})$, $\phi=g(\theta)$, leads to

$$\frac{\hat{\phi}-\phi}{\tau} \sim N(-z_0\sigma_\phi,\sigma_\phi^2) \quad , \quad \sigma_\phi=1+a\phi \tag{2.9.7}$$

When a=0, (2.9.7) reduces to the bias corrected percentile method.

Firstly, consider the simple case $\hat{\theta}\sim f_\theta$, where $f_\theta(\hat{\theta})$ is a one-parameter family of densities for the real valued statistic θ. The parametric bootstrap distribution is

$$\hat{\theta}^* \sim f_{\hat{\theta}} \tag{2.9.8}$$

which is the distribution of the statistic of interest when the unknown parameter θ is equated to the observed point estimate $\hat{\theta}$. The cumulative distribution function (c.d.f) of the bootstrap distribution is defined as

$$\hat{G}(S) = \int_{-\infty}^{s} f_{\hat{\theta}}(\hat{\theta}^*)d\hat{\theta}^* = \text{Prob}_{\hat{\theta}}\left\{\hat{\theta}^* < s\right\} \tag{2.9.9}$$

Assuming there exists a monotone increasing transformation g and constants z_0 and a such that

$$\hat{\phi}=g(\hat{\theta}),\ \hat{\phi}=g(\theta) \tag{2.9.10}$$

satisfying

$$\hat{\phi}=\phi+\sigma_\phi(Z-z_0) \qquad (Z \sim N(0,1)) \tag{2.9.11}$$

and $\tau=1$

with

$\sigma_\phi = 1+a\phi$

then, under these conditions, Efron showed that the correct central confidence interval of level $1-2\alpha$ for θ is

$$\theta \in \{ \hat{G}^{-1}(\phi(z\{\alpha\})),\ \hat{G}^{-1}(\phi(z\{1-\alpha\})) \} \qquad (2.9.12)$$

where

$$z\{\alpha\} = z_o + \frac{(z_o+z^{(\alpha)})}{1-a(z_o+z^{(\alpha)})} \qquad (2.9.13)$$

and likewise for $z\{1-\alpha\}$

If z_o and a equal 0 , then (2.9.13) refers to the percentile method discussed previously.

To calculate the 'acceleration constant' a, Efron considers the empirical influence function $U_i, i=1,\ldots,n$ of $\hat{\theta}=t(\hat{F})$, defined as

$$U_i = \lim_{\Delta \to 0} \frac{t((1+\Delta)\hat{F} + \Delta\delta_i) - t(\hat{F})}{\Delta} \qquad (1=1,2,\ldots,n) \qquad (2.9.14)$$

Here δ_i is a point mass of x_i such that U_i is the derivative of the estimate $\hat{\theta}$ with respect to the mass on point x_i. Efron gives the following approximation for the constant a.

$$a \simeq \frac{1}{6} \frac{\sum_{i=1}^{n} U_i^3}{(\sum_{i=1}^{n} U_i^2)^{3/2}} \qquad (2.9.15)$$

where the U_i's can be evaluated using finite differences in (2.9.14).

As an example, consider the linear model

$Y=X\beta+e$ (2.9.16)

where $\hat{\beta}$ is the least squares estimate of β, and U_i is the influence function for β at the point x_i. Then a finite sample version of U_i at the point x_i is

$$U_i = (n-1)(\hat{\beta}-\hat{\beta}_{-i}) \qquad (2.9.17)$$

where $\hat{\beta}_{-i}$ is the least squares estimate of β when the point (x_i, y_i) is omitted from the sample.

Using the 'correlation' example from Efron's earlier paper (1981b) 100 000 bootstrap replications resulted in the central 90% non-parametric BC_a internal $\{.43, .92\}$, where $\hat{\rho}=0.776$.

Finally, it should be noted that (2.9.15) is invariant under monotone changes of the parameter of interest, thus the BC_a intervals have the correct transformation properties.

2.10 **Future trends in bootstrapping methods :**

Efron (1987), as described in the previous section, introduces a new procedure for improving bootstrap confidence intervals using an 'acceleration' constant. Two approaches for further improvements of the coverage of the bootstrap confidence intervals are firstly, Beran's prepivoting bootstrap method (see Beran, 1987) and secondly, the use of studentized pivotal quantities (see Frangos and Schucany, 1987). A description of each method is as follows:

(i) the method of prepivoting is an iterated bootstrap method which relies entirely on nested bootstrapping. Suppose that θ is real valued and the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$ is normal, where $\hat{\theta}$ is an estimate of θ. Let $H_n(\cdot,\theta)$ denote the c.d.f. of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$. The estimated c.d.f. $H_n(\cdot,\hat{\theta})$ is called the parametric bootstrap estimate of $H_n(\cdot,\theta)$. Then, $H_n(x,\hat{\theta})$ converges in probability, uniformly in x, to the same normal c.d.f. as does $H_n(x,\theta)$. Moreover, the limiting distribution

$$Y = X\beta + e \tag{2.9.16}$$

where $\hat{\beta}$ is the least squares estimate of β, and U_i is the influence function for β at the point x_i. Then a finite sample version of U_i at the point x_i is

$$U_i = (n-1)(\hat{\beta} - \hat{\beta}_{-i}) \tag{2.9.17}$$

where $\hat{\beta}_{-i}$ is the least squares estimate of β when the point (x_i, y_i) is omitted from the sample.

Using the 'correlation' example from Efron's earlier paper (1981b) 100 000 bootstrap replications resulted in the central 90% non-parametric BC_a internal $\{.43, .92\}$, where $\hat{\rho}=0.776$.

Finally, it should be noted that (2.9.15) is invariant under monotone changes of the parameter of interest, thus the BC_a intervals have the correct transformation properties.

2.10 <u>**Future trends in bootstrapping methods**</u> :

Efron (1987), as described in the previous section, introduces a new procedure for improving bootstrap confidence intervals using an 'acceleration' constant. Two approaches for further improvements of the coverage of the bootstrap confidence intervals are firstly, Beran's prepivoting bootstrap method (see Beran, 1987) and secondly, the use of studentized pivotal quantities (see Frangos and Schucany, 1987). A description of each method is as follows:

(i) the method of prepivoting is an iterated bootstrap method which relies entirely on nested bootstrapping. Suppose that θ is real valued and the asymptotic distribution of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$ is normal, where $\hat{\theta}$ is an estimate of θ. Let $H_n(\cdot,\theta)$ denote the c.d.f. of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$. The estimated c.d.f. $H_n(\cdot,\hat{\theta})$ is called the parametric bootstrap estimate of $H_n(\cdot,\theta)$. Then, $H_n(x,\hat{\theta})$ converges in probability, uniformly in x, to the same normal c.d.f. as does $H_n(x,\theta)$. Moreover, the limiting distribution of $H_n\{n^{\frac{1}{2}}(\hat{\theta}-\theta),\hat{\theta}\}$ is considered uniform on $(0,1)$, thus motivating the bootstrap confidence interval

$$C_n = \{\theta : \alpha < H_n\{n^{\frac{1}{2}}(\hat{\theta}-\theta),\hat{\theta}\} < 1-\alpha\} \tag{2.10.1}$$

which has asymptotic coverage probability $1-2\alpha$ for θ.

Reduction in the coverage probability error of C_n due to the assumption that

$$R_{n,1}(\hat{\theta}_n,\theta) = H_n\{n^{\frac{1}{2}}(\hat{\theta}-\theta),\hat{\theta}\} \tag{2.10.2}$$

is exactly uniform $(0,1)$, can be achieved by starting with $R_{n,1}(\hat{\theta},\theta)$ in place of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$. Consequently, let $H_{n,1}(\cdot,\theta)$ denote the c.d.f. of $R_{n,1}(\hat{\theta},\theta)$ and define the new bootstrap confidence interval

$$\begin{aligned} C_{n,1} &= \{\theta : \alpha < H_{n,1}\{R_{n,1}(\hat{\theta},\theta),\hat{\theta}\} < 1-\alpha\} \\ &= \{\theta : \hat{\theta}-c_n(1-\alpha) < \theta < \hat{\theta}-c_n(\alpha)\} \end{aligned} \tag{2.10.3}$$

where $c_n(t)=n^{-\frac{1}{2}}H_n^{-1}\{H_{n,1}^{-1}(t,\hat{\theta}),\hat{\theta}\}$. For some models, the calculation of $C_{n,1}$ may require Monte Carlo approximations to $H_n(\cdot,\hat{\theta})$ and $H_{n,1}(\cdot,\hat{\theta})$. For the second of these cdf's, the natural Monte Carlo algorithm involves nested double parametric bootstrapping. 'Prepivoting' is the term used to describe the step involving the transformation of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$ by its estimated cdf. in the definition of $C_{n,1}$. The distribution of the original $n^{\frac{1}{2}}(\hat{\theta}-\theta)$ is often more dependent of θ than is the distribution of the transformed quantity $R_{n,1}(\hat{\theta},\theta)$. Consequently, the prepivoted bootstrap confidence set $C_{n,1}$ often has smaller coverage probability error than does C_n.

Another property of prepivoting is that it may be repeated to achieve higher order corrections automatically. For $i \geq 1$, let

$$R_{n,i+1}(\hat{\theta},\theta)=H_{n,i}\{R_{n,1}(\hat{\theta},\theta),\hat{\theta}_n\} \tag{2.10.4}$$

where $H_{n,i}(\cdot,\theta)$ is the cdf. of $R_{n,i}(\hat{\theta},\theta)$. Then, the i-times prepivoted bootstrap confidence interval for θ is defined to be

$$C_{n,i}=\{\theta:\alpha< H_{n,i}\{R_{n,i}(\hat{\theta},\theta),\hat{\theta}\}<1-\alpha\} \tag{2.10.5}$$

To consider an example of prepivoting, let the estimate $\hat{\theta}=(1+n^{-1})\bar{x}_n$, where $\bar{x}_n$ is the mean of a random sample of size n drawn from the N(0,1) model. The coverage probability error of the once prepivoted bootstrap confidence set $C_{n,1}$ is $O(n^{-2})$. The twice prepivoted bootstrap confidence interval $C_{n,2}$ is exact, as it coincides with the classical confidence interval based on $n^{\frac{1}{2}}(\bar{x}_n-\theta)$.

(ii) the second approach is to use studentized pivotal quantites to produce second-order accurate bootstrap confidence intervals. The use of Edgeworth expansions to approximate the distribution of pivotal statistics of the form $T=(\hat{\theta}-\theta)/\sigma(\theta)$ was considered by Abramovitch and Singh (1985). They showed that bootstrapping statistics of the form T improves the normal approximation. Frangos and Schucany (1987), considered two different ways of refining second order correct confidence intervals by estimating higher-order terms of an Edgeworth expansion of the distribution of a statistic and also by estimating higher order terms of the von Mises expansion for the statistic itself. They examined the performance of confidence intervals of the form:

$$\{\theta-H^{-1}(1-\alpha)S.E.(\hat{\theta}),\theta-H^{-1}(\alpha)S.E.(\hat{\theta})\} \tag{2.10.6}$$

where H(s) is the empirical distribution $T=(\hat{\theta}-\theta)/S.E.(\hat{\theta})$. The distribution of T is approximated by the bootstrap distribution of

$$T^* = \frac{(\hat{\theta}-\theta)}{S.E.(\hat{\theta})} \tag{2.10.7}$$

i.e. $H(s) = \#\{T^* \leq s\}/B$. The standard error of $\hat{\theta}$ is estimated by

$$S.E.(\hat{\theta}) = \sum_i I_i^2/n^2 \qquad (2.10.8)$$

where $I_i = (n-1)(\hat{\theta}-\hat{\theta}_{-i})$

$\hat{\theta} = t(X_1,\ldots,X_n)$

and $\hat{\theta}_{-i} = t(X_1,\ldots,X_{i+1},\ldots,X_n)$

The confidence intervals (2.10.6) have been called bootstrap-t (BST) confidence intervals, and the simulation studies carried out by Frangos and Schucany, for the variance and the correlation coefficient with various distributions supported the general improvement afforded to studentization.

2.11 Open questions for further research :

What further research should be carried out on the bootstrap method? The following list includes some thoughts which are by no means exhaustive on the matter.

(i) an important question which one can investigate is the application of the prepivoting method of Beran (1987), in areas where the conventional bootstrap methods of Efron do not produce good results.

(ii) Another question which concerns the importance of the bootstrap techniques is the incorporation of pivotal studentized quantities for non-parametric confidence intervals based on the paper of Abramovitz and Singh (1985).

(iii) The bootstrap method is established as a worthy non-parametric tool in statistics although further enhancement of the method could be achieved with more theoretical results. For example, one could investigate whether the Bayesian bootstrap methods (see Rubin, 1981) provide robust non-parametric confidence intervals for various parameters using different probability distributions.

(iv) a very important task could be to undertake a Monte Carlo comparison of all the different versions of the bootstrap procedures using different probability distributions and in different areas of statistics. This simulation study could be modelled on the Princeton Robustness study (1972).

(v) the employment of Edgeworth expansions in approximating the bootstrap distribution for certain statistics.

(vi) in the case of asymptotic expansions, one could not underestimate the von Mises expansions using influence functions for important bootstrap non-parametric confidence intervals. In this thesis, the construction of non-parametric confidence intervals has been investigated using conventional bootstrap methods. One could determine the influence function of, for example, the availability ratio (see Chapter 3), and estimate by jackknifing the variance of the statistic and then compare this estimator with the corresponding bootstrap estimator of the variance. Certainly, the importance of non-parametric bootstrap confidence intervals will necessitate the use of pivotal studentized quantites. In this respect, one must investigate the coverage and length of these non-parametric confidence intervals for the parameter θ, and also the ratio $Pr(\theta<L)/Pr(\theta>H)$, where L and H are the lower and higher confidence limits respectively.

(vii) one could also determine the best transformation, logarithmic or otherwise, for transforming a statistic which has a non-normal probability distribution into a statistic which has a normal probability distribution. The Johnson system of transformation (see Johnson, 1978) could be used for this purpose.

CHAPTER 3

NON-PARAMETRIC ESTIMATION OF COMPONENT AND SYSTEM AVAILABILITY

3.1 Introduction :

In the process of determining the worth of equipment to perform a given task, there are usually three quantities of primary concern, namely, reliability, maintainability and availability. These quantities form the cornerstones of reliability theory which grew out of the demands of modern technology. Khintchine (1932), and Palm (1947), investigated the area of machine maintenance which was one of the first areas of reliability to be approached with any mathematical sophistication. Parametric families of distributions that could be used as lifetime models in engineering contexts were also developed. Some of these such as the extreme value distributions (Gumbel, 1935) and the Weibull distribution (Weibull, 1939), appeared first in connection with the strength or fatigue life of materials. The attraction of the Weibull distribution was that it could describe increasing as well as decreasing failure rates.

Altman and Goor (1946), introduced non-parametric methods to the statistical analyses of lifetime data. There also coincided with the general increase of attention to reliability problems by engineers an increased interest in the statistical treatment of reliability data. This work marked the beginning of the widespread assumption of the exponential distribution in life-testing research. Davis (1952), presented the results of several goodness-of-fit tests for various competing failure distributions using some failure data. This data seemed to give a distinct edge to the exponential distribution and with a further paper by Epstein and Sobel (1953), the exponential distribution acquired a unique position in life testing. Reliability theory had now established itself as a major area in engineering. The exponential was well established as the distribution on which much statistical analysis and modelling of lifetimes was based. Although other distributions, such as the Weibull, were starting to gain use, it was only the exponential distribution where statistical methods were reasonably well developed. Two other developments also took place. Firstly, the estimation of system reliability was considered. Buehler (1957), and Steck (1957), for example, estimated the reliability for a series system where each component either fails or does not fail over some period of time, based on test data on component reliabilities. Secondly, more sophisticated models for systems were considered and concepts like availability and maintainability (e.g. see Shooman, 1968) were introduced. The reliability growth of systems under development was also studied and modelled.

3.2 **Statistical developments in reliability theory :**

Numerous advances in statistical theory and methods relevant to reliability have been made over the last 25 years. A brief summary of six of the main developments are as follows:

(i) the parametric inference for univariate life distributions: one of the major areas of activity was in the provision of estimation and hypothesis test procedures for various parametric life distribution families. Previously, a considerable amount of work has been done on inference procedures for the exponential distribution with censored and uncensored data (e.g. Epstein and Sobel, 1953, 1954; Bartholomew, 1957) and for the normal and log normal distributions. Some of the impetus for the great amount of activity in this area came from the realisation that whereas the exponential distribution was easily handled statistically, it gave non-robust methods (e.g. Zelen and Dannemiller, 1961) and was not an appropriate model in many situations. Other life distributions such as the Weibull were becoming increasingly popular and the Weibull distribution emerged later as the most widely used life distribution. However, it should also be said that useful work on other models, such as the gamma (e.g., Engelhardt and Bain, 1978), the generalised gamma (e.g. Farewell and Prentice, 1977; Lawless, 1982), the lognormal (e.g. Nelson and Schmee, 1979), and the inverse Gaussian (e.g. Chhikara and Folks, 1977) has been done in recent years, as has work on goodness-of-fit tests for various models. In addition, the understanding of general large sample maximum likelihood methods for use with censored data has improved considerably. Various papers by Blight (1970), Cox (1975), Aalen (1978) and Kalbfleisch and Prentice (1980), some of them motivated by biomedical lifetime problems, have established that standard maximum likelihood large sample methods can be used with many types of censored data.

(ii) non-parametric and graphical procedures with censored data: Kaplan and Meier (1958), developed the product-limit estimate of the survivor function $S(t)=Pr(T\geq t)$, for a lifetime random variable T. The product limit estimate is defined as follows: suppose that $t_1<t_2\ldots<t_k$ are the observed lifetimes in a sample of k lifetimes and n-k censoring times. Let n(t) denote the number of individuals known to be still alive (i.e. alive and uncensored) just prior to time t. Then the product-limit estimate of S(t) is :

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n(t_i)-1}{n(t_i)}\right) \tag{3.2.1}$$

with the proviso that if there happens to be a censoring time t^* larger than t_k, then $\hat{S}(t)$ is undefined past t^*. Nelson (1972a), developed a nonparametric estimate of the cumulative hazard function $H(t)=-\log S(t)$ as follows:

$$\hat{H}(t) = \sum_{i:t_i<t} \frac{1}{n(t_i)}\ , \tag{3.2.2}$$

which yields another estimate $\hat{S}(t)=\exp.\{-\hat{H}(t)\}$ of S(t). Besides giving nonparametric estimates of a life distribution's survivor function from censored data, equations (3.2.1) and (3.2.2) are extremely useful for providing plots from which one can carry out model assessment or rough parameter estimation for parametric families of models.

In another direction, nonparametric tests for the quality of two or more distributions were extended to handle censoring. The best known examples are perhaps the extensions of the Wilcoxon test (e.g., Efron, 1967; Prentice, 1978) and the exponential ordered scores test (e.g. Peto and Peto 1972; Prentice 1978).

(iii) regression analysis of lifetime data: maximum likelihood large sample theory or linear estimation methods are typically used to handle regression problems where lifetime, or some transform thereof, is the response variable. Such methods have been developed and applied in numerous papers; Nelson and Hahn (1972), Nelson (1972b), Lawless and Singhal (1980), and Zelen (1959), provide examples in the life testing area.

The regression models used most with lifetime data fall into two categories. Firstly, most of the common parametric models are of the form:

$$Y=\mu(x;\beta)+\sigma Z, \qquad -\infty < Z < \infty \tag{3.2.3}$$

where Y=logT represents log lifetime, $\mu(x)$ is a function of a vector of regressor variables $x=(x_1,\ldots,x_p)$, σ is a positive scale parameter, and Z has a known distribution. The second category of models, much used in the regression analysis of biomedical lifetime data, is the so-called proportional hazards family. Here, the effect of regressor variables is assumed to be multiplicative on the hazard function for T, which is given by $h(t)=f(t)/S(t)$, where $f(t)$ is the p.d.f. of a life distribution. The hazard function of T given x is assumed to be of the form:

$$h(t/x)=h_o(t;\theta)g(x;\beta) \tag{3.2.4}$$

Here, $h_o(t;\theta)$ can be thought of as the baseline hazard function for an item with $g(x;\beta)=1$.

Regression models based on binary response models, are frequently used when one considers only the failure or non-failure of equipment over a specified time period. One can of course adapt the models (3.2.3) or (3.2.4) to this situation, although, in practice, most of the models used are adaptions of (3.2.3). In particular, (3.2.3) gives, for a specified time, to the probability of failure by time t_o as

$$\Pr(T \leqslant t_o) = \Pr\left(\frac{Y-\mu(x;\beta)}{\sigma} \leqslant \frac{\log t_o - \mu(x;\beta)}{\sigma}\right)$$

$$= F\left(\frac{\log t_o - \mu(x;\beta)}{\sigma}\right) \qquad (3.2.5)$$

where F is the cumulative distribution function (cdf) of Z. Frequently, $\mu(x;\beta)$ is taken to be a linear function ; in the case of a single regressor variable this would be $\mu(x;\beta)=\beta_o+\beta_1 x$, and (3.2.5) would become:

$$\Pr(\text{failure by } t_o) = F(\alpha+\beta x) \qquad (3.2.6)$$

where $\alpha = (\log t_o - \beta_o)/\sigma$ and $\beta=-\beta_{1/\sigma}$. The most frequently used models are those for which F is the cdf of a standard logistic, standard normal, or standard extreme value distribution.

(iv) multiple failure mode problems: many problems dealing with equipment involve more than one type of failure. The main approach to this problem has been through the classical theory of competing risks (e.g. David and Moeschberger, 1978; Nelson, 1982). In this formulation, one supposes that there is associated with each of k possible failure causes $(i=1\ldots,k)$ a random variable T_i that represents time to failure from that cause. The time to failure for the item is then given by $T=\min(T_1,\ldots,T_k)$ and $C=\{i:T_i=\min(T_1,\ldots,T_k)\}$ is the observed failure cause. Most competing risk analysis assumes that $T_1,\ldots,T_k$ are independent. In this case the survivor function for T is $S(t)=S_1(t)\ldots S_k(t)$, where

$S_i(t)$ is the survivor function for T_i.

A main problem with the classical approach is that the concept of T_i, time to failure from cause i, is often without any physical basis. In addition, even when the T_i's are physically meaningful, they are often non independent. An alternative formulation of multiple failure mode problems is given by Altschuler (1970), Prentice et al, (1978), and others. Briefly, this models the joint distribution of (T,C) directly, in terms of cause-specific hazard functions:

$$h_j(t)=\lim_{\Delta t \to 0} \frac{\Pr(t<T<t+\Delta t, C=j/T>t)}{\Delta t} \tag{3.2.7}$$

or subsurvivor functions $S_j(t)=\Pr(T>t,C=j)$, $j=1,\ldots,k$. With this approach, both parametric and non-parametric inference procedures are easily developed. Also, in cases where the classical competing-risks framework is physically meaningful and $T_1,\ldots,T_k$ are independent, the formulation just given is equivalent to the competing-risks framework and so yields the same statistical procedures.

(v) Methods for repeated failure analysis: an important reliability problem is the treatment of repeated failures on the same piece of equipment or system. For example, consider the repeated failure of a piece of equipment over time, with repair time for simplicity being ignored. One way to model this is in terms of a point process with intensity function:

$$\lambda(t;H(t),x)=\lim_{\Delta t \to 0} \frac{\Pr\{\text{failure in } (t,t+\Delta t)\,|\,H(t),x\}}{\Delta t}, \tag{3.2.8}$$

where H(t) represents the past history of failures for the equipment, and x represents concomitant variables that might be related to failure. The key problem is to develop models and statistical methods for (3.2.8).

Two models that are widely used in reliability are the time dependent Poisson process for a homogeneous population, where $\lambda(t;H(t),x)=\lambda(t)$, and the renewal process were $\lambda(t;H(t),x)=\lambda(t-v(t))$, with $v(t)$ the time of the most recent failure. However, models of the form (3.2.8) generally require a quite different point of view than do models for the life distribution (or time to a single failure) of a piece of equipment. Generally, there needs to be more done in developing these models and in discriminating among alternative models.

(vi) estimation of system reliability: the determination of system reliability from reliability information on components of the system has been well documented (e.g. Barlow and Proschan, 1975), but statistical methods have played a fairly minor role. A typical statistical problem is the following: suppose that a system has components $C_1,\ldots,C_k$ connected in some configuration. Given data from independent tests on the various components, estimate the reliability of the system. A lot of work has gone into developing estimation procedures for systems in which the components $C_1,\ldots,C_k$ operate independently of one another. Also, the methods which are available, are essentially for exponentially failure time distributions, or binary (fail/no-fail) treatments of components and systems.

The assumption of component independence is of course critical to this work, and this limits its applicability. Consequently, besides the extension of existing approaches to wider ranges of life distributions, there is a need for models that capture more of the physical realities of system reliability problems.

3.3 **Definition of reliability concepts :**

In considering the various types of reliability problems discussed in the previous section, it is usually desired to analyze and calculate certain quantities of interest, designated in the literature by a variety of labels: reliability, availability, interval availability, efficiency, effectiveness, etc. It is therefore necessary to define clearly and concisely the basic quantities arising in reliability theory.

Consider a system whose state at time t is described by $X(t)=(X_1(t),\ldots,X_n(t))$, a vector-valued random variable. For example, $X(t)$ may be a vector of system parameter values, with each component $X_i(t)$ ranging over an interval of real numbers. $X(t)$, being a random variable, will be defined by a distribution function,

$$F(x_1,\ldots,x_n;t);$$

i.e. $F(x_1,\ldots,x_n;t)$ equals the probability that

$$X_1(t) \leqslant x_1,\ldots,X_n(t)\leqslant x_n$$

Now, for any state $x=(x_1,\ldots,x_n)$, there is a gain or payoff, $g(x)$. For a two state system where $x=0$ or 1, the gain accruing from being in the functioning state ($x=1$) might be one unit of value, so that $g(1)=1$, and the gain from being in the failed state ($x=0$) might be 0, so that $g(0)=0$. The expected gain $G(t)$ at time t may be calculated from

$$G(t) = E\{g(X_1(t))\} = \int\ldots\int g(x_1,\ldots,x_n)dF(x_1,\ldots,x_n;t) \qquad (3.3.1)$$

Also, the expected gain $G(t)$ may be averaged over some interval of time, $a\leqslant t\leqslant b$, with respect to some weight function $W(t)$ to obtain

$$H(a,b)= \int_a^b G(t).dW(t) \qquad (3.3.2)$$

Equations (3.3.1) and (3.3.2) can then be used to define the various reliability quantities as follows:

(i) 'Reliability' is the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered.

Let $X(u)=1$ if the device is performing adequately at time u, 0 otherwise ; then

$G(t)= E\{g(X(t))\} = P\{X(t)=1\}$ = probability that the device performs adequately over $\{0,t\}$

Thus, $G(t)$ is the reliability of the device as defined above.

(ii) 'Pointwise availability' is the probability that the system will be able to operate within the tolerances at a given instant of time.

As before, let $X(t)=1$ if the system is operating within tolerances at time t, 0 otherwise. Also, as before $g(1)=1$, $g(0)=0$. Given that the possibility of repair or replacement before time t, is not excluded, then

$G(t)=E\{g(X(t))\} = P\{X(t)=1\}$ is the probability that the system is operating within tolerances at time t.

Thus, $G(t)$ now defines the pointwise availability at the time point t.

(iii) 'Interval availability' is the expected fraction of a given interval of time that the system will be able to operate within the tolerances, whereby repair and/or replacement is permitted.

For a given time interval $\{a,b\}$, where $W(t)=(t-a)/(b-a)$, then $H(a,b)$, as defined in (3.3.2), can be computed as

$$H(a,b)= \frac{1}{(b-a)} \int_a^b G(t).dt = \frac{1}{(b-a)} \int_a^b E\{g(X(t))\}\, dt$$

so that under suitable regularity conditions

$$H(a,b)= E\, \frac{\int_a^b g(X(t))dt}{(b-a)}$$

which is the expected fraction of the time interval $\{a,b\}$ that the system is operating within tolerances. Thus $H(a,b)$ is the interval availability for the interval $\{a,b\}$.

(iv) 'Limiting interval availability' is the expected fraction of time in the long run that the system operates satisfactorily. The term $\lim_{T\to\infty} H(0,T)$, as defined previously, is simply computed, to obtain the limiting interval availability.

(v) 'Interval reliability' is the probability that at a specified time, the system is operating and will continue to operate for an interval of duration, say 'x'. Repair and/or replacement is permitted and the continued operation during the interval is to be achieved without the benefit of repair or replacement.

Let $X(t)=1$ if the system is operating at time t, 0 otherwise. Then the interval reliability $R(x,T)$ for an interval of duration x starting at time T is given by

$$R(x,T)=P\{X(t)=1, T\leqslant t\leqslant T+x\}$$

'Limiting interval reliability', often known as 'strategic reliability', is simply the limit of $R(x,T)$ as $T\to\infty$

3.4 **Parametric Methods used in setting confidence limits for system availability :**

A paper by Thompson (1966), describes techniques for determining a lower confidence limit on system availabilty A, and for testing the null hypothesis that $A=A_0$ against the alternative hypothesis that $A<A_0$ when the time-to-failure and time-to-repair are independent, exponentially distributed variables. The pointwise availability estimate is the statistic calculated for this purpose, and is expressed mathematically as

$$A = \frac{\theta}{\theta+\phi} = \frac{1}{1+(\phi/\theta)} \qquad (3.4.1)$$

where θ = system mean-time-to-failure

ϕ = system mean-time-to-repair

A random sample of n_1 times-to-failure and n_2 times-to-repair are drawn from exponential distributions with sample means $\hat{\theta}$ and $\hat{\phi}$ respectively. Given that $2n_1\hat{\theta}/\theta$ and $2n_2\hat{\phi}/\phi$ will be chi-square distributed variables with $2n_1$ and $2n_2$ degrees of freedom, since they are independent due to the independence of the variables t_1 and t_2, it is therefore possible to define two new variables

$$Z_1 = \left(\frac{2n_1\hat{\theta}/\theta}{2n_1}\right)/\left(\frac{2n_2\hat{\phi}/\phi}{2n_2}\right) = \frac{\hat{\theta}\phi}{\hat{\phi}\theta} \tag{3.4.2}$$

which is F-distributed with $2n_1$, $2n_2$ degrees of freedom, and

$$Z_2 = \frac{1}{Z_1} = \frac{\hat{\phi}\theta}{\hat{\theta}\phi} \tag{3.4.3}$$

which is F-distributed with $2n_2$, $2n_1$ degrees of freedom.

Using the variable Z_1 , a lower confidence limit for A can be derived as follows:

$$\Pr\left\{\frac{\hat{\theta}\phi}{\hat{\phi}\theta} < F_{1-\alpha}\ (2n_1, 2n_2)\right. = 1-\alpha$$

which simplif[illegible] to:

$$\Pr\left\{\frac{\hat{\theta}}{\hat{\theta}+\hat{\phi}F_{1-\alpha}(2n,2n)} < A\right\} = 1-\alpha \tag{3.4.4}$$

Given the special case when $n_1=n_2=n$, then equation (3.4.4) becomes:

$$\Pr\left\{\frac{\hat{\theta}}{\hat{\theta}+\hat{\phi}\,F_{1-\alpha}(2n,2n)} < A\right\} = 1-\alpha$$

and the $(1-\alpha)$ lower confidence limit is found from

$$LCL = \frac{\hat{\theta}}{\hat{\theta}+\hat{\phi}F_{1-\alpha}(2n,2n)} \tag{3.4.5}$$

A two-sided $(1-\alpha)$ confidence interval, derived in a similar manner, is given by:

$$LCL = \frac{\hat{0}}{\hat{0}+\phi F_{1-\alpha/2})2n,2n)} \qquad (3.4.6)$$

$$UCL = \frac{\hat{0}F_{1-\alpha/2}(2n,2n)}{\hat{0}F_{1-\alpha/2}(2n,2n)+\hat{\phi}} \quad , \qquad (3.4.7)$$

where LCL and UCL are the lower and upper confidence levels respectively.

A test of the null hypothesis $H_o:A=A_o$ against the alternative hypothesis $H_1:A<A_o$ at the α level of significance can be derived using Z_2 as the decision statistic and $F_{1-\alpha}(2n_2,2n_1)$ as the decision criteria. To perform a test of H_o vs H_1, it is necessary to find some k such that:

$$Pr\left\{\hat{A} < k \mid H_o\right\} = \alpha,$$

which is equivalent to

$$Pr\left\{\frac{1}{1+\hat{\phi}/\hat{\theta}} < k \mid H_o\right\} = \alpha$$

This simplifies to:

$$Pr\left\{Z_2 > F_{1-\alpha}(2n_2,2n_1)\right\} = \alpha \qquad (3.4.8)$$

The power function of H_o vs. H_1 is as follows:

$$Pr\left\{\frac{\hat{\phi}\theta_o}{\hat{\theta}\phi_o} > F_{1-\alpha}(2n_2,2n_1) \;/\; H_1 : A=A_1 < A_o\right\} = 1-\beta \qquad (3.4.9)$$

where β is the consumer's risk of a Type II error.

A paper by Gray and Lewis (1967), investigated a confidence interval for system availability under slightly differing assumptions to Thompson, namely that the time between failures and the time to repair are independent, exponential and lognormal random variables respectively.

The following notation was used to define system availability:

$$\boxed{A = \frac{u_y}{u_y + u_x} = \frac{1}{1 + \frac{u_x}{u_y}},} \tag{3.4.10}$$

where u_y = mean time between failures of the equipment

and u_x = mean time to repair the equipment

The problem was to establish a confidence interval for the ratio of the mean of a lognormal distribution to the mean of an exponential distribution. Gray and Lewis's paper proposed a method by which, under the assumption of known variance of the lognormal distribution, an exact confidence interval for u_x/u_y can be derived. A description of their method is as follows:

Given β^2, the variance of a lognormal distribution X is known, denote X by $\Lambda(\alpha, \beta^2)$. Then for a random sample of size n,

$$\frac{\left(\prod_{i=1}^{n} x_i\right)^{1/n}}{e^{\alpha}} = \frac{\bar{x}_G}{e^{\alpha}} \sim \Lambda(0, \beta^2/n),$$

where $\bar{x}_G$ is the geometric mean.

Let $U \sim \chi^2(k)$ and $V \sim \Lambda(0, \beta^2/n)$

Then, if U and V are independent

$$f(u,v) = \frac{u^{k_2 - 1}\, e^{-u/2} \sqrt{n}\, e^{-n/2(\mathrm{Ln}\, v/\beta)^2}}{\Gamma(k/2) 2^{k/2}\, \beta \sqrt{2\pi} V}, \qquad \begin{array}{l} 0<u<\infty \\ 0<v<\infty \end{array} \tag{3.4.11}$$

$= 0$ elsewhere

By letting W=U/V and Z=V and proceeding by standard methods, one obtains