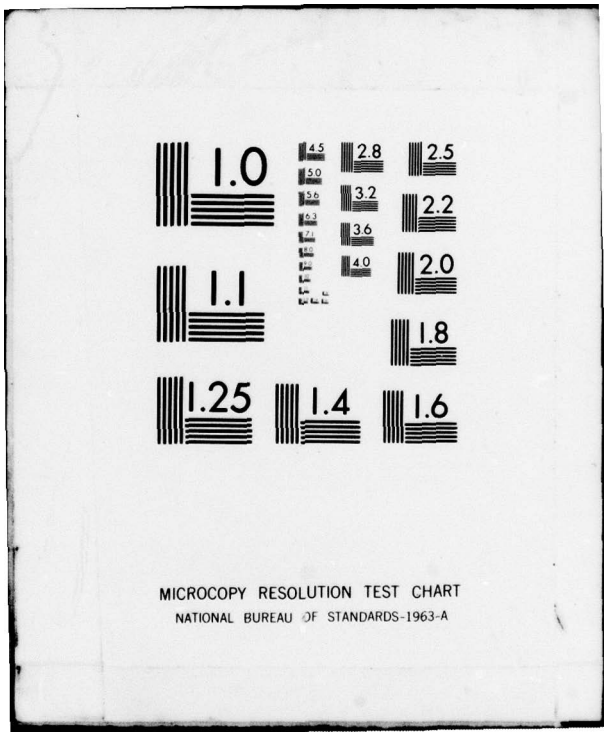
AD-A073 270 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
A DENSITY-QUANTILE FUNCTION APPROACH TO CHOOSING ORDER STATISTI--ETC(U)
JUL 79 R L EUBANK DAAG29-78-G-0180
UNCLASSIFIED TR-A-10 ARO-16228.10-M NL

END
DATE FILMED
10-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ARO 16228.10-M

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS

(18) ARO (19) 16228.10-M

(12)

(6) A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale Parameters /

LEVEL II

(10) Lester Randall L. Eubank
Institute of Statistics, Texas A&M University

(12) 65 p.

(9) Technical Report

(11) July 1979

(14) TR-A-18

AD A073270

DDC
RECEIVED
AUG 30 1979
C

DDC FILE COPY

Texas A & M Research Foundation
Project No. 3861

"Maximum Robust Likelihood Estimation and Non-parametric Statistical Data Modeling"

Sponsored by the U.S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

(15) DAAG29-78-G-0180

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE | READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. A-10
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale Parameters.
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Randall L. Eubank
8. CONTRACT OR GRANT NUMBER(s): DAAG29-78-G-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Research Office, Research Triable Park, NC 27709
12. REPORT DATE: July 1979
13. NUMBER OF PAGES: 129
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report): NA
18. SUPPLEMENTARY NOTES: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Order statistics, Data compression, Estimation, Location parameters, Scale parameters, Censored samples, Quantile estimation, Reproducing kernel Hilbert spaces.
20. ABSTRACT (Continue on reverse side if necessary and identify by block number): Parzen (1979) has shown that the estimation of location and scale parameters by linear systematic statistics may be formulated as a problem in regression analysis of a smoothed sample quantile process. In this dissertation a general approach to optimal spacings selection is presented that utilizes design techniques for continuous parameter time series regression, developed by Sacks and Ylvisaker (1966, 1968). This methodology is applied to several common distributions. The problems of optimal order statistic selection for estimation in censored samples, for quantile estimation and for the summarization of large data sets are also considered.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

347 380

79 08 28 005

ABSTRACT

A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale Parameters.

(August 1979)

Randall Lester Eubank, B.S., New Mexico State University; M.S., New Mexico State University; M.Stat, Texas A&M University; Chairman of Advisory Committee: Professor Emanuel Parzen

Parzen (1979) has shown that the estimation of location and scale parameters by linear systematic statistics may be formulated as a problem in regression analysis of a smoothed sample quantile process. In this dissertation, a general approach to optimal spacings selection is presented that utilizes design techniques for continuous parameter time series regression developed by Sacks and Ylvisaker (1966, 1968). This methodology is applied to several common distributions. The problems of optimal order statistic selection for estimation in censored samples, for quantile estimation and for the summarization of large data sets are also considered.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Emanuel Parzen for his interest and cooperation throughout this undertaking and for the invaluable experience of our association. Without his insight and guidance, this dissertation would not be possible.

My greatest personal indebtedness is to my wife, Elisa. She has provided constant support throughout my graduate education. Her assistance in the recopying and proofreading of this manuscript was invaluable.

I am grateful to Dr. W. B. Smith, Dr. L. J. Ringer, Dr. T. E. Wehrly, Dr. D. J. Hartfiel, and Dr. Dan Colunga for their willingness to serve on my advisory committee. A special word of thanks goes to Dr. W. B. Smith for his temporary service as my advisor. It has been both a pleasure and an honor to be associated with the faculty and staff at the Institute of Statistics. The interest and care they show in their students helps remove many of the obstacles and complexities associated with graduate work thereby making it a truly enjoyable experience.

Finally I would like to express my gratitude to Ms. Linda Bishop for the typing of the manuscript, to Mr. D. L. Hawkins for his programming assistance, and to Mr. Louis Gaston for his aid with the figures.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF TABLES

LIST OF TABLES (CONTINUED)

LIST OF TABLES (CONTINUED)

LIST OF FIGURES

1. INTRODUCTION

1.1 Preliminaries

A frequently occuring statistical model is the location and scale parameter model. In this model, it is assumed that the cumulative distribution function (c.d.f.) of independent identically distributed random variables, $X_1, X_2, \ldots, X_n$, is of the form

$$F(x) = F_o\left(\frac{x - \mu}{\sigma}\right) \qquad (1.1.1)$$

where F_o is a known distributional form and μ and σ are respectively unknown location and scale parameters. The maximum likelihood estimators of μ and σ are often difficult to compute. Thus practical considerations often dictate the usage of estimators that are inefficient when compared to the Cramér-Rao lower variance bound for unbiased parameter estimation.

A class of estimators of μ and σ that have good properties are those formed as linear functions of the sample order statistics, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, (the random sample $X_1, X_2, \ldots, X_n$ arranged in increasing order). These estimators have been called systematic statistics by Mosteller (1946). Systematic statistics will be of the form $\sum_{i=1}^{k} b_i X_{(n_i)}$ for $X_{(n_1)}, X_{(n_2)}, \ldots, X_{(n_k)}$ a subset of the n sample order statistics. For this reason the problem of

---

Citations follow the format of the Journal of the American Statistical Association.

estimation by systematic statistics consists of two parts: (a) the selection of a set of k order statistics and (b) the determination of the coefficients for the order statistics selected.

Definition 1.1.1. The quantile function, Q, corresponding to a distribution function F is defined to be

$$Q(u) = F^{-1}(u) = \inf(x : F(x) \geq u). \qquad (1.1.2)$$

The p^{th} quantile $(0 < p < 1)$ of the distribution is $Q(p)$.

Definition 1.1.2. Define the sample quantile function, $\tilde{Q}$, by

$$\tilde{Q}(u) = X_{(j)}, \quad \frac{j-1}{n} < u \leq \frac{j}{n}, \quad j = 1, \ldots, n. \qquad (1.1.3)$$

The p^{th} sample quantile $(0 < p < 1)$ is $\tilde{Q}(p)$.

It is often more convenient to consider linear systematic statistics as being linear functions of sample quantiles. By letting $u_i = n_i/n$, it follows that $\sum_{i=1}^{k} b_i X_{(n_i)} = \sum_{i=1}^{k} b_i \tilde{Q}(u_i)$ and hence the two formulations are equivalent.

A basic result which leads to the usefulness of systematic statistics is the following theorem due to Mosteller (1946).

Theorem 1.1.1. Let F be an absolutely continuous distribution with probability density function (p.d.f.) denoted by f. Let $0 < u_1 < u_2 < \ldots < u_k < 1$ be k real numbers and $Q(u_1), Q(u_2), \ldots, Q(u_k)$ the corresponding (population) quantiles. Further assume

that f is differentiable in the neighborhood of $x_i \equiv Q(u_i)$ and that $f(Q(u_i)) \equiv fQ(u_i) \neq 0$ for $i = 1, \ldots, k$. Then the joint distribution of the k sample quantiles, $\tilde{Q}(u_1)$, $\tilde{Q}(u_2)$, ..., $\tilde{Q}(u_k)$, tends to a k-variate normal distribution as n tends to infinity with

$$AE[\tilde{Q}(u_i)] = Q(u_i) \tag{1.1.4}$$

and

$$ACOV(\tilde{Q}(u_i),\tilde{Q}(u_j)) = \frac{1}{n} \frac{u_i(1 - u_j)}{fQ(u_i)fQ(u_j)} , \; u_i \leq u_j , \tag{1.1.5}$$

where AE and ACOV denote asymptotic expectation and covariance respectively.

When the location and scale parameter model (1.1.1) holds the p.d.f. and quantile function have the forms

$$f(x) = \frac{1}{\sigma} f_o\left(\frac{x - \mu}{\sigma}\right) ,$$

$$Q(u) = \mu + \sigma Q_o(u) \tag{1.1.6}$$

where f_o and Q_o are respectively the p.d.f. and the quantile function corresponding to the known c.d.f. F_o. A corollary to Theorem 1.1.1 concerning this model follows immediately using (1.1.6).

Corollary 1.1.1. In addition to the hypotheses of Theorem 1.1.1, assume that f and Q are of the form (1.1.6). Then the limiting

distribution of the k sample quantiles is k-variate normal with

$$AE[\tilde{Q}(u_i)] = \mu + \sigma Q_o(u_i) \tag{1.1.7}$$

$$ACOV(\tilde{Q}(u_i),\tilde{Q}(u_j)) = \frac{\sigma^2}{n} \frac{u_i(1 - u_j)}{f_oQ_o(u_i)f_oQ_o(u_j)} , \; u_i \leq u_j . \tag{1.1.8}$$

For the purposes of location and scale parameter estimation, Corollary 1.1.1 may be interpreted as stating that, asymptotically, the sample quantiles, $\tilde{Q}(u_1)$, $\tilde{Q}(u_2)$, ..., $\tilde{Q}(u_k)$, satisfy the conditions required for application of the Gauss-Markov Theorem. Thus asymptotically best linear unbiased estimators (ABLUE's) of μ and/or σ may be obtained through generalized least squares. Ogawa (1951) has given general formulae for these estimators and their asymptotic relative efficiencies (ARE's) when either one or both of the parameters are unknown.

Let $u_o \equiv 0$, $u_{k+1} \equiv 1$ and $f_oQ_o(u_o) = f_oQ_o(u_{k+1}) = 0$. Define

$$K_1 = \sum_{i=1}^{k+1} \frac{[f_oQ_o(u_i) - f_oQ_o(u_{i-1})]^2}{u_i - u_{i-1}} , \tag{1.1.9}$$

$$K_2 = \sum_{i=1}^{k+1} \frac{[Q_o(u_i)f_oQ_o(u_i) - Q_o(u_{i-1})f_oQ_o(u_{i-1})]^2}{u_i - u_{i-1}} , \tag{1.1.10}$$

$$K_3 = \sum_{i=1}^{k+1} \frac{[f_oQ_o(u_i) - f_oQ_o(u_{i-1})][Q_o(u_i)f_oQ_o(u_i) - Q_o(u_{i-1})f_oQ_o(u_{i-1})]}{u_i - u_{i-1}} ,$$

(1.1.11)

$$\Delta = K_1 K_2 - K_3^2 \quad , \tag{1.1.12}$$

$$Z = \sum_{i=1}^{k+1} \frac{[f_o Q_o(u_i) - f_o Q_o(u_{i-1})][f_o Q_o(u_i)\tilde{Q}(u_i) - f_o Q_o(u_{i-1})\tilde{Q}(u_{i-1})]}{u_i - u_{i-1}} \quad , \tag{1.1.13}$$

$$Y = \sum_{i=1}^{k+1} \frac{1}{u_i - u_{i-1}} \Big\{ [Q_o(u_i) f_o Q_o(u_i) - Q_o(u_{i-1}) f_o Q_o(u_{i-1})] \; [f_o Q_o(u_i)\tilde{Q}(u_i) - f_o Q_o(u_{i-1})\tilde{Q}(u_{i-1})] \Big\} \quad . \tag{1.1.14}$$

In this notation the ABLUE's and ARE's derived by Ogawa may be written as follows:

1. Assume σ is known. Then the ABLUE for μ is

$$\hat{\mu} = \frac{1}{K_1} Z - \sigma \frac{K_3}{K_1} \tag{1.1.15}$$

with asymptotic relative efficiency

$$ARE(\hat{\mu}) = \frac{K_1}{E\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]} \quad . \tag{1.1.16}$$

2. Assume μ is known. Then the ABLUE for σ is

$$\hat{\sigma} = \frac{1}{K_2} Y - \mu \frac{K_3}{K_2} \tag{1.1.17}$$

with asymptotic relative efficiency

$$ARE(\hat{\sigma}) \doteq \frac{K_2}{E\left[\left(\frac{X f'(X)}{f(X)}\right)^2\right] - 1} \tag{1.1.18}$$

($\doteq$ denoting approximate equality for large values of n).

3. Assume both μ and σ are unknown. ABLUE's for μ and σ are

$$\hat{\mu} = \frac{1}{\Delta} (K_2 Z - K_3 Y) \tag{1.1.19}$$

$$\hat{\sigma} = \frac{1}{\Delta} (K_1 Y - K_3 Z) \tag{1.1.20}$$

with asymptotic relative efficiency

$$ARE(\hat{\mu},\hat{\sigma}) = \frac{\Delta}{E\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]\left\{E\left[\left(X \frac{f'(X)}{f(X)}\right)^2\right] - 1\right\} - \left\{E\left[X\left(\frac{f'(X)}{f(X)}\right)^2\right]\right\}^2} \quad . \tag{1.1.21}$$

Examination of equations (1.1.15) - (1.1.21) reveals that the estimators and their asymptotic relative efficiencies are all functions of the spacings, $u_1, u_2, \ldots, u_k$. Therefore, through strategic placement of the spacings it is possible to further optimize asymptotic estimator efficiency. A set of spacings resulting in a maximum value for one of the efficiency expressions (1.1.16), (1.1.18) or (1.1.21) will be termed an optimal spacings set while the problem of finding such sets will be termed the optimal spacings problem.

1.2 Review of the Literature

1.2.1 Overview

Let

$$x_i = Q_o(u_i) , \quad i = 1, \ldots, k . \tag{1.2.22}$$

The classical approach to the optimal spacings problem has been to first show that there exist solutions, $x_1^*, \ldots, x_k^*$, to

$$\frac{\partial ARE(\hat{\mu})}{\partial x_i} = 0 \tag{1.2.23}$$

$$\frac{\partial ARE(\hat{\sigma})}{\partial x_i} = 0 \tag{1.2.24}$$

$$\frac{\partial ARE(\hat{\mu},\hat{\sigma})}{\partial x_i} = 0 \tag{1.2.25}$$

which satisfy the order restrictions

$$-\infty < x_1^* < x_2^* < \ldots < x_k^* < \infty \tag{1.2.26}$$

and maximize one of the expressions for asymptotic estimator efficiency given in Section 1.1. The next step is to find the optimal spacings that correspond to these solutions. Much of the literature on the optimal spacings problem is concerned with solutions obtained in this manner.

For certain distribution types, the asymptotic relative efficiency expressions become quite complicated. Consequently numerical methods have frequently been employed to find spacings sets resulting in optimal or near optimal efficiencies. The results obtained are usually expressed in the form of tables of optimal spacings and the corresponding coefficients for the ABLUE's for various values of k.

1.2.2 Normal Distribution

Ogawa (1951) has considered the optimal spacings problem for the normal distribution. In the event that σ is known, it was shown that there exists a unique set of optimal spacings for each value of k. These optimal spacings were shown to be symmetric, i.e. $u_i + u_{k-i+1} = 1$. Numerical techniques were employed to find the optimal spacings when k = 2(1)10.

For the case of a known location parameter, it was found that the function $ARE(\hat{\sigma})$ had many maxima. Although the greatest maximum was not found, of the spacing sets considered those resulting in the largest efficiency values were reported for k = 1(1)6.

Simultaneous estimation of both the location and scale parameter was considered only for an estimator based on two order statistics. This estimator was derived under the assumption of symmetric spacings.

Eisenberger and Posner (1965) have extended Ogawa's results. Assuming σ or μ to be known, optimal spacings have been calculated for the cases k = 2(2)20. Suboptimal spacings that minimize the

sum of the estimators' variances, $V(\hat{\mu}) + V(\hat{\sigma})$, were also found for $k = 2(2)20$.

1.2.3. Exponential Distribution

Sarhan and Greenberg (1958) have considered the estimation of the scale parameter of the exponential distribution under the assumption that the location parameter was known. Optimal spacings were obtained for linear systematic estimators based on $k = 1(1)15$ order statistics.

1.2.4 Pareto Distribution

For the Pareto distribution the location and scale parameter model takes the form

$$F(x) = 1 - \left(1 + \frac{x - \mu}{\sigma}\right)^{-\nu}, \quad x \geq \mu \qquad (1.2.27)$$

where $\nu > 0$ is a known shape parameter. Kulldorf and Vännman (1973) have considered the estimation of μ and σ by linear functions of optimally spaced order statistics.

For μ assumed known, optimal spacings for the estimation of σ were obtained when $\nu = .5(.5)5$ and $k = 1(1)10$. When $\nu = 1$, the optimal spacings were found to be the points $\frac{i}{k+1}$, $i = 1, \ldots, k$.

For both μ and σ unknown, it was shown that optimal spacings sets exist for the simultaneous estimation of μ and σ only if the sample is censored to the left. Thus a set of modified estimators

was suggested which use the optimal spacings for the estimation of σ (when μ is known) with $k - 1$ sample quantiles.

1.2.5 Cauchy Distribution

Bloch (1966) has considered the estimation of the location parameter of the Cauchy distribution by a linear function of five order statistics. Numerical techniques were utilized to obtain a set of spacings that was essentially the optimum. These spacings corresponded to an asymptotic relative efficiency of .95161. A linear systematic estimator based on these five spacings was found to be superior to the optimum trimmed mean and the sample median.

1.2.6 Logistic Distribution

Gupta and Gnanadesikan (1966) have considered the optimal spacings problem for the logistic distribution. In the case of a known scale parameter, the optimal spacings for location parameter estimation were found to be the points $\frac{i}{k+1}$, $i = 1, \ldots, k$. An explicit form for the estimator based on these optimal spacings was given as

$$\hat{\mu} = \frac{6}{k(k+1)(k+2)} \sum_{i=1}^{k} i(k+1-i)\bar{Q}\left(\frac{i}{k+1}\right) \qquad (1.2.28)$$

with

$$ARE(\hat{\mu}) = \frac{k(k+2)}{(k+1)^2} \quad . \qquad (1.2.29)$$

Optimal spacings for the estimation of σ, in the event that μ is known, were obtained for two or three symmetrically spaced order statistics.

Hassanein (1969b) has considered suboptimal simultaneous location and scale parameter estimation using spacings that minimize the sum of the variances of the estimators. The spacings for these suboptimal estimators were given for the cases k = 2(1)9.

1.2.7 Weibull Distribution

The location and scale parameter model for the Weibull distribution is of the form

$$F(x) = 1 - \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{\gamma}\right\}, \quad x > \mu, \qquad (1.2.30)$$

where γ is a known positive shape parameter. Hassanein (1971) has obtained optimal spacings for the simultaneous estimation of both the location and scale parameters of this distribution. The optimal spacings are given for estimators based on two, four, or six order statistics for the cases γ = 3(1)10(5)20.

1.2.8 Extreme Value Distribution

Hassanein (1968) has studied the estimation of location and scale parameters in the extreme value distribution. Optimum spacings for the estimation of the location parameter when the scale parameter is known were given for the cases k = 1(1)15. Under the

assumption that the location parameter is known, optimum spacings for scale parameter estimation were obtained for k = 1(1)4. An iterative scheme was proposed for simultaneous estimation of μ and σ by linear functions of two order statistics.

The problem of selecting optimal spacings for simultaneous location and scale parameter estimation has been considered by Hassanein (1969a, 1972). Spacings that minimize the sum of the variances of the estimators have been obtained for k = 2, 3, 4 and optimal spacings have been obtained for k = 1(1)10.

1.2.9 Gamma Distribution

The location and scale parameter model for the gamma distribution results in a p.d.f. of the form

$$f(x) = \frac{1}{\sigma\Gamma(p)}\left(\frac{x-\mu}{\sigma}\right)^{p-1} e^{-\left(\frac{x-\mu}{\sigma}\right)}, \quad x > \mu, \qquad (1.2.31)$$

where $p > 0$ is assumed known. Särndal (1964) has obtained nearly optimum spacings (in the sense that they result in near maximum ARE's) for the estimation of the scale parameter of the gamma distribution, when μ is assumed known for k = 1(1)10 with p = 2(1)5. The techniques utilized in the calculation of these spacings are a special case of a general approach to optimal spacings selection developed by Särndal (1962).

Hassanein (1977) has obtained suboptimal spacings for simultaneous location and scale parameter estimation that maximize the

sum of the efficiencies of the estimators. Spacings sets were given for estimators based on k = 2(1)5 order statistics with p = 3(1)10(5)20.

1.3 Objectives

The purpose of this research is threefold: (a) to formulate the problems of location and scale parameter estimation and spacing selection in a unified framework, based on regression analysis of the continuous parameter sample quantile function $\tilde{Q}(u)$, $0 \leq u \leq 1$, thereby developing a general computationally simple solution to the optimal spacings problem, (b) to apply this technique to several common distributions, and (c) to develop guidelines for the selection of order statistic subsets for the summarization of large data sets. The problems of optimal order statistic selection for estimation in censored samples and quantile estimation will also be considered.

2. SPACINGS FOR UNCENSORED SAMPLES

2.1 Preliminaries

As seen in Chapter 1, the classical approach to optimal location and/or scale parameter estimation by linear functions of order statistics has centered upon the efficiency of the estimators. The asymptotic relative efficiencies of these linear systematic statistics are functions of the spacings of the sample quantiles included in the estimators. Thus spacings that maximize expressions for asymptotic efficiency correspond to a best set of sample quantiles to be used for estimation purposes. Therefore, classically, optimal estimators were obtained by first finding spacings that resulted in maximum values for asymptotic efficiencies and then using the corresponding sample quantiles to estimate location and scale parameters by generalized least squares.

There are several problems associated with the classical approach. Two such difficulties are:

1. Finding spacings that result in maximum values for the asymptotic relative efficiencies is often computationally quite difficult.
2. There is no unified framework for solving the problem of optimal estimation.

In this chapter, a new approach to optimal location and scale parameter estimation is presented which alleviates many of the

problems inherent in the classical method. It will be seen in this and subsequent chapters that the computational aspects of this new procedure are quite simple. Also by using this approach, the problem of obtaining optimal linear systematic estimators of the location and/or scale parameters may be formulated in a unified regression framework.

The principal results of this chapter are contained in Section 2.5. Sections 2.2 through 2.4 provide the necessary background for the development of these results. Section 2.2 contains a few preliminary concepts and definitions. Section 2.3 treats the topic of regression design for a Brownian Bridge process. In Section 2.4, it is seen that the problem of location and scale parameter estimation can be formulated as one of continuous parameter *time series regression*. The results of Section 2.3 and 2.4 are combined in Section 2.5 to obtain estimators of μ and/or σ based on asymptotically optimal spacings. Finally in Section 2.6, the approach taken here is contrasted with those of Chernoff (1971) and Särndal (1962).

2.2 Definitions and Notation

This section contains definitions and notation that will be used in subsequent sections.

Definition 2.2.1. A reproducing kernel Hilbert space (RKHS), with reproducing kernel K, is a Hilbert space, H, of functions defined on a set T. The kernel K is a function on T × T with the following

properties for every u in T (where $K(\cdot,u)$ is the function defined on T whose value at s in T is $K(s,u)$):

$$K(\cdot,u) \in H \tag{2.2.1}$$

$$(g,K(\cdot,u)) = g(u) \tag{2.2.2}$$

for every g in H.

A kernel that will be seen to be of particular interest is the covariance kernel of a Brownian Bridge process.

Definition 2.2.2. A Brownian Bridge process $\{B(u),\ u \in [0,1]\}$ is *a zero mean normal process with covariance kernel*

$$K_B(u_1,u_2) = \min(u_1,u_2) - u_1u_2 \ . \tag{2.2.3}$$

The Hilbert function space, H_B, generated by K_B consists of L^2 differentiable functions satisfying $f(0) = f(1) = 0$ for every f in H_B. The inner product of two function f and g in H_B is

$$\langle f,g\rangle = \int_0^1 f'(u)g'(u)du \ . \tag{2.2.4}$$

By taking $g(u) = f(u)$, it is seen that for any f in H_B

$$\|f\|^2 = \int_0^1 (f'(u))^2du \ . \tag{2.2.5}$$

2.3 Regression Design for a Brownian Bridge Process

Let $\{B(u),\ u \in [0,1]\}$ be the Brownian Bridge process defined in Section 2.2. Consider the regression model

$$Y(u) = \sum_{i=1}^{m} \beta_i f_i(u) + \alpha B(u), \quad u \in [0,1], \tag{2.3.1}$$

$$\mathrm{COV}(Y(s), Y(t)) = \alpha^2 K_B(s,t)$$

where $f_1, f_2, \ldots, f_m$ are given regression functions, $\beta_1, \beta_2, \ldots, \beta_m$, α are $m+1$ unknown parameters and K_B is given by (2.2.3). For this model an infinite observation set is feasible, in which case parameter estimation may be accomplished using techniques developed by Parzen (1961a,b). It is often more convenient to take observations at only a finite number, k, of points on [0,1] as then parameter estimates may be obtained by generalized least squares. Since the point set to be selected is at the disposal of the experimenter, it should be selected in such a manner that the resultant estimators have optimal properties over all such estimators formed from the same number of observations.

The problem of selecting a "best" set of points (for estimation purposes) at which to sample from the process $\{Y(u),\ u \in [0,1]\}$ is one of design for continuous parameter time series regression. The point sets are called *designs* and the points themselves are called design points. Since sampling over time obviates the repetition of observations at a particular design point, it is necessary to define what is meant by a k-point design.

Definition 2.3.1. A k-point design for a Brownian Bridge process (and hence for $\{Y(u),\ u \in [0,1]\}$) is a k-tuple, $\{u_1, u_2, \ldots, u_k\}$, with $0 < u_1 < u_2 < \ldots < u_k < 1$. Denote by D_k the set of all such k point designs.

For $T \in D_k$, let $\hat{\beta}_T$ denote the BLUE of $\beta = (\beta_1, \beta_2, \ldots, \beta_m)'$ based on observations taken according to T. Let $\hat{\beta}$ denote the estimator of β obtained using observations over all of [0,1].

Optimal designs are those that minimize the variance of the estimator in the case $m = 1$ or minimize the generalized variance of the estimators in the case $m \geq 2$. As a result of the order restrictions satisfied by design points, D_k is not a compact set. Consequently, optimum designs frequently do not exist and are usually difficult to construct. This leads to the consideration of design sequences that are asymptotically optimal.

Definition 2.3.2. For the case $m = 1$ a design sequence $\{T_j\}_{j=1}^{\infty}$, $T_k \in D_k$, is asymptotically optimal for estimating $\beta_1 \equiv \beta$ if

$$\lim_{k \to \infty} \left\{ \frac{V(\hat{\beta}_{T_k}) - V(\hat{\beta})}{\inf_{T \in D_k} V(\hat{\beta}_T) - V(\hat{\beta})} \right\} = 1 \quad . \tag{2.3.2}$$

Definition 2.3.3. For the case $m \geq 2$ a design sequence $\{T_j\}_{j=1}^{\infty}$, $T_k \in D_k$ is asymptotically optimal for estimating β if

$$\lim_{k\to\infty}\left\{\frac{|V(\hat{\beta}_{T_k})^{-1}| - |V(\hat{\beta})^{-1}|}{\inf_{T\in D_k}|V(\hat{\beta}_T)^{-1}| - |V(\hat{\beta})^{-1}|}\right\} = 1 \quad . \tag{2.3.3}$$

Design sequences may be constructed through the use of density functions. Let h be a continuous non-negative density function on [0,1] with associated distribution function

$$H(u) = \int_0^u h(t)dt \quad . \tag{2.3.4}$$

The density, h, generates a design sequence $\{T_j\}_{j=1}^{\infty}$ where $T_k \in D_k$ and $T_k = \{H^{-1}(\frac{1}{k+1}), H^{-1}(\frac{2}{k+1}), \ldots, H^{-1}(\frac{k}{k+1})\}$. Finding an asymptotically optimal design sequence will be seen to be equivalent to finding an optimal density, i.e., finding a density that generates an asymptotically optimal density sequence.

The next theorem gives densities that generate asymptotically optimal design sequences. Its proof can be found as a straightforward application of results obtained by Sacks and Ylvisaker (1966, 1968).

Theorem 2.3.1. Let f_i be twice continuously differentiable on [0,1] and have the representation

$$f_i(u) = -\int_0^1 f_i''(t)K_B(u,t)dt \quad , \quad i = 1, \ldots, m \quad . \tag{2.3.5}$$

Then

1. For m = 1 the density

$$h^*(u) = \frac{[f_1''(u)]^{2/3}}{\int_0^1 [f_1''(t)]^{2/3}dt} \tag{2.3.6}$$

generates asymptotically optimal design sequences for estimating $\beta_1 \equiv \beta$.

2. For $m \geq 2$ let

$$\psi(u) = -(f_1''(u), f_2''(u), \ldots, f_m''(u))' \tag{2.3.7}$$

$$A = \{(f_i, f_j)\} \quad , \quad i, j = 1, \ldots, m \quad . \tag{2.3.8}$$

The density

$$h^*(u) = \frac{[\psi'(u)A^{-1}\psi(u)]^{1/3}}{\int_0^1 [\psi'(t)A^{-1}\psi(t)]^{1/3}dt} \tag{2.3.9}$$

generates asymptotically optimal designs for estimating $\beta = (\beta_1, \beta_2, \ldots, \beta_m)'$.

Remark 2.3.1. Let $T = \{u_1, u_2, \ldots, u_k\} \in D_k$ and m = 1. Denote by $P_T f_1$ the projection of f_1 onto the linear manifold generated by the functions $K_B(\cdot, u_i)$, $i = 1, \ldots, k$. It can be shown that

$$V(\hat{\beta}_T) = \|P_T f_1\|^{-2} = [\,\|f_1\|^2 - \|f_1 - P_T f_1\|^2\,]^{-1}. \quad (2.3.10)$$

Finding a sequence satisfying (2.3.2) is roughly (apart from convergence rate considerations) the same as finding a sequence $\{T_k\}_{k=1}^{\infty}$, such that

$$\lim_{k\to\infty} [\,\|f_1 - P_{T_k} f_1\|^2 - \inf_{T\subset D_k} \|f_1 - P_T f_1\|^2] = 0 \quad . \quad (2.3.11)$$

Remark 2.3.2. Let h be a density, and suppose $\{T_k\}_{k=1}^{\infty}$ is the design sequence generated by h. When m = 1, Sacks and Ylvisaker (1968) have shown that

$$\lim_{k\to\infty} k^2 \|f_1 - P_{T_k} f_1\|^2 = \frac{1}{12} \int_0^1 \left[\frac{f_1''(u)}{h(u)} \right]^2 du \quad (2.3.12)$$

provided h and f_1 satisfy any of the following conditions:

1. $\int_0^1 [h(u)]^{-2} du < \infty$ and f_1'' is continuous on [0,1].
2. $\frac{f_1''}{h}$ is continuous on [0,1].
3. There exist a constant C such that $(b-a)\int_a^b h^2(u)du \le C[\int_a^b h(u)du]^2$ for all $[a,b] \subset [0,1]$.

The authors have also shown that

$$\lim_{k\to\infty} k^2 \inf_{T\subset D_k} \|f_1 - P_T f_1\|^2 = \frac{1}{12} [\int_0^1 [f_1''(u)]^{2/3} du]^3 . \quad (2.3.13)$$

Thus in view of Remark 2.3.1 and equation (2.3.12) the task of finding an asymptotically optimal design sequence can be accomplished by finding a density function that attains the bound (2.3.13). Such a density is (2.3.6).

Remark 2.3.3. A multi-parameter version of Remark 2.3.2 may also be made. Sacks and Ylvisaker (1968) have shown that for $\{a_1, \ldots, a_m\}$ a set of positive numbers

$$\lim_{k\to\infty} \inf k^2 \sum_{i=1}^m a_i \|f_i - P_{T_k} f_i\|^2 \ge \frac{1}{12} \{\int_0^1 [\sum_{i=1}^m a_i (-f_i''(u))^2]^{1/3} du\}^3 \quad (2.3.14)$$

for any sequence of designs $\{T_k\}_{k=1}^{\infty}$. Finding an asymptotically optimal design is equivalent to finding a density h* that attains a lower bound of the form (2.3.14). Such an optimal density is given by (2.3.9).

Remark 2.3.4. The design sequences defined in Theorem 2.3.1 differ from those suggested by direct application of the formulae of Sacks and Ylvisaker (1966, 1968). These results may be reconciled by noting that no information is obtained from observations taken at 0 or 1 for regression functions in the RKHS generated by K_B. This is because (as noted in Section 2) such functions are necessarily zero at these points. By taking the $(k+2)^{th}$ element of the sequences suggested by these authors and disregarding design points at 0 and 1 Theorem 2.3.1 may be obtained.

Remark 2.3.5. To obtain designs for an interval [p,q] and regression functions of the form

$$f_i(u) = -\int_p^q f_i''(s)K_B(s,u)ds + C_1K_B(u,p) + C_2K_B(u,q), \quad i = 1, \ldots, m , \tag{2.3.15}$$

a modification of (2.3.6) and (2.3.9) is required. Although the form of the optimal densities are not altered, all limits of integration must be changed from 0 and 1 to p and q. The design points for the $(k+1)^{st}$ element in the sequence are then $H^{-1}(\frac{i}{k})$, $i = 0, \ldots, k$.

2.4 Location and Scale Parameter Estimation as a Continuous Parameter Time Series Regression Problem

Parzen (1979) has phrased the problem of linear estimation of location and scale parameters as a problem in regression analysis of a smoothed sample quantile process $\{f_oQ_o(u)\tilde{Q}(u),\ u \in [0,1]\}$. The formulation rests upon a theorem of Csorgo and Revesz (1978) regarding the deviation of $\tilde{Q}$ from the true quantile function Q. This theorem may be paraphrased as saying that, under suitable conditions on fQ, $\sqrt{n}\ fQ(u)[\tilde{Q}(u) - Q(u)]$, $0 \le u \le 1$, is asymptotically a Brownian Bridge process.

For the location and scale parameter model

$$F(x) = F_o\left(\frac{x-\mu}{\sigma}\right) \tag{2.4.1}$$

the true quantile function is

$$Q(u) = \mu + \sigma Q_o(u) \tag{2.4.2}$$

where Q_o is the quantile function corresponding to F_o. Thus the process

$$\{\frac{\sqrt{n}}{\sigma} f_oQ_o(u)[\tilde{Q}(u) - \mu - \sigma Q_o(u)] , \quad u \in [0,1]\} \tag{2.4.3}$$

may be considered as a Brownian Bridge process, $\{B(u),\ u \in [0,1]\}$, for large values of n.

Parzen (1979) has justified writing the expression

$$\frac{\sqrt{n}}{\sigma} f_oQ_o(u)[\tilde{Q}(u) - \mu - \sigma Q_o(u)] \approx B(u) \tag{2.4.4}$$

which holds asymptotically as $n \to \infty$. The estimation of μ and σ is then seen to be a problem in continuous parameter time series regression by writing (2.4.4) as

$$f_oQ_o(u)\tilde{Q}(u) = \mu f_oQ_o(u) + \sigma f_oQ_o(u)Q_o(u) + \sigma_B B(u) \tag{2.4.5}$$

where

$$\sigma_B = \sigma/\sqrt{n} . \tag{2.4.6}$$

The parameter σ_B is not constrained to be related to σ and is estimated as a free parameter. Therefore its estimate provides a diagnostic check on the goodness of fit of the model.

2.5 Selection of Optimal Spacings as a Regression Design Problem for the Quantile Process

The regression model for location and scale parameter estimation was seen in Section 2.4 to be

$$f_oQ_o(u)\tilde{Q}(u) = \mu f_oQ_o(u) + \sigma Q_o(u)f_oQ_o(u) + \sigma_B B(u) \qquad (2.5.1)$$

where

$$\sigma_B = \sigma/\sqrt{n} \qquad (2.5.2)$$

and $\{B(u),\ u \in [0,1]\}$ is a Brownian Bridge process. This model is seen to be a special case of the model (2.3.1) by making the identifications $f_1(u) = f_oQ_o(u)$, $f_2(u) = Q_o(u)f_oQ_o(u)$, $\beta_1 = \mu$, $\beta_2 = \sigma$ and $\alpha = \sigma_B$. Therefore, in an analogous manner to Section 2.3 optimal regression designs may be considered for what, in this case, would be location and scale parameter estimation. However, a comparison of the definition of a design for a Brownian Bridge process (Definition 2.3.1) with the properties of a set of spacings shows that such a design is nothing more than a set of spacings for sample quantiles. Thus for the model (2.5.1), selecting an optimum design for the estimation of location and scale parameters is equivalent to selecting an optimal set of spacings.

In the light of the previous discussion, it follows that densities which generate sequences of asymptotically optimal spacings

sets (designs) may be obtained by an application of Theorem 2.3.1. This statement is formalized in the next theorem.

Theorem 2.5.1. Let f_oQ_o and $Q_o \cdot f_oQ_o$ be twice continuously differentiable on $[0,1]$ and possess the representations

$$f_oQ_o(u) = -\int_0^1 [f_oQ_o(t)]''K_B(u,t)dt \qquad (2.5.3)$$

$$Q_o(u)f_oQ_o(u) = -\int_0^1 [Q_o(t)f_oQ_o(t)]''K_B(u,t)dt\ . \qquad (2.5.4)$$

The following conclusions hold:

1. The density

$$h^*(u) = \frac{[f_oQ_o''(u)]^{2/3}}{\int_0^1 [f_oQ_o''(t)]^{2/3}dt} \qquad (2.5.5)$$

generates a sequence of asymptotically optimal spacings sets for the estimation of μ when σ is known.

2. The density

$$h^*(u) = \frac{([Q_o(u)f_oQ_o(u)]'')^{2/3}}{\int_0^1 ([Q_o(t)f_oQ_o(t)]'')^{2/3}dt} \qquad (2.5.6)$$

generates a sequence of asymptotically optimal spacings sets for the estimation of σ when μ is known.

3. Let

$$\psi(u) = -([f_oQ_o(u)]'', [Q_o(u)f_oQ_o(u)]'')' \qquad (2.5.7)$$

and define the information matrix, A, by

$$A = \begin{bmatrix} \langle f_oQ_o, f_oQ_o\rangle & \langle f_oQ_o, Q_o f_oQ_o\rangle \\ \langle Q_o f_oQ_o, f_oQ_o\rangle & \langle Q_o f_oQ_o, Q_o f_oQ_o\rangle \end{bmatrix} . \qquad (2.5.8)$$

The density

$$h^*(u) = \frac{[\psi'(u)A^{-1}\psi(u)]^{1/3}}{\int_0^1 [\psi'(t)A^{-1}\psi(t)]^{1/3}dt} \qquad (2.5.9)$$

generates a sequence of asymptotically optimal spacings sets for the simultaneous estimation of μ and σ.

Remark 2.5.1. The asymptotic optimality of the spacings sets may be interpreted as meaning that as the number of spacings in a set, k, grows large the spacings in Theorem 2.5.1 give rise to estimators with approximately the same efficiency as estimators based on the optimal set of k spacings.

Given any density function, in particular those in Theorem 2.5.1, the form of the estimators may be deduced from those in Chapter 1. It will be useful to adopt a somewhat different notation than that employed there. Let h be a density function with associated distribution function H. Define for a specified number of spacings, k, the functions

$$K_1(h) = \sum_{i=1}^{k+1} \frac{[f_oQ_o(H^{-1}\left(\frac{i}{k+1}\right)) - f_oQ_o(H^{-1}\left(\frac{i-1}{k+1}\right))]^2}{H^{-1}\left(\frac{i}{k+1}\right) - H^{-1}\left(\frac{i-1}{k+1}\right)} , \qquad (2.5.10)$$

$$K_2(h) = \sum_{i=1}^{k+1} \left\{ \frac{1}{H^{-1}\left(\frac{i}{k+1}\right) - H^{-1}\left(\frac{i-1}{k+1}\right)} [f_oQ_o(H^{-1}\left(\frac{i}{k+1}\right))Q_o(H^{-1}\left(\frac{i}{k+1}\right)) \right.$$

$$\left. - f_oQ_o(H^{-1}\left(\frac{i-1}{k+1}\right))Q_o(H^{-1}\left(\frac{i-1}{k+1}\right))]^2 \right\} , \qquad (2.5.11)$$

$$K_3(h) = \sum_{i=1}^{k+1} \left\{ \frac{1}{H^{-1}\left(\frac{i}{k+1}\right) - H^{-1}\left(\frac{i-1}{k+1}\right)} [f_oQ_o(H^{-1}\left(\frac{i}{k+1}\right)) \right.$$

$$- f_oQ_o(H^{-1}\left(\frac{i-1}{k+1}\right))][f_oQ_o(H^{-1}\left(\frac{i}{k+1}\right))Q_o(H^{-1}\left(\frac{i}{k+1}\right))$$

$$\left. - f_oQ_o(H^{-1}\left(\frac{i-1}{k+1}\right))Q_o(H^{-1}\left(\frac{i-1}{k+1}\right))] . \right\} , \qquad (2.5.12)$$

$$\Delta(h) = K_1(h)K_2(h) - K_3(h)^2 . \qquad (2.5.13)$$

Also define the weight functions $W_\mu(i,h)$, $W_\sigma(i,h)$, $S_\mu(i,h)$, $S_\sigma(i,h)$ and correction factors $C_\mu(h)$, $C_\sigma(u)$ as follows:

$$W_\mu(i,h) = \frac{f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right))}{K_1(h)} \left[\frac{f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right)) - f_o Q_o(H^{-1}\left(\frac{i-1}{k+1}\right))}{H^{-1}\left(\frac{i}{k+1}\right) - H^{-1}\left(\frac{i-1}{k+1}\right)} \right.$$

$$\left. - \frac{f_o Q_o(H^{-1}\left(\frac{i+1}{k+1}\right)) - f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right))}{H^{-1}\left(\frac{i+1}{k+1}\right) - H^{-1}\left(\frac{i}{k+1}\right)} \right] , \qquad (2.5.14)$$

$$C_\mu(h) = \frac{K_3(h)}{K_1(h)} , \qquad (2.5.15)$$

$$W_\sigma(i,h) = \frac{f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right))}{K_2(h)} \left\{ \frac{1}{H^{-1}\left(\frac{i}{k+1}\right) - H^{-1}\left(\frac{i-1}{k+1}\right)} \left[f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right)) Q_o(H^{-1}\left(\frac{i}{k+1}\right)) \right.\right.$$

$$\left. - f_o Q_o(H^{-1}\left(\frac{i-1}{k+1}\right)) Q_o(H^{-1}\left(\frac{i-1}{k+1}\right)) \right] - \frac{1}{H^{-1}\left(\frac{i+1}{k+1}\right) - H^{-1}\left(\frac{i}{k+1}\right)}$$

$$\left. \cdot \left[f_o Q_o(H^{-1}\left(\frac{i+1}{k+1}\right)) Q_o(H^{-1}\left(\frac{i+1}{k+1}\right)) - f_o Q_o(H^{-1}\left(\frac{i}{k+1}\right)) Q_o(H^{-1}\left(\frac{i}{k+1}\right)) \right] \right\}$$

(2.5.16)

$$C_\sigma(h) = \frac{K_3(h)}{K_2(h)} , \qquad (2.5.17)$$

$$S_\mu(i,h) = \frac{1}{\Delta(h)} [K_2(h)W_\mu(i,h) - K_3(h)W_\sigma(i,h)] , \qquad (2.5.18)$$

$$S_\sigma(i,h) = \frac{1}{\Delta(h)} [K_2(h)W_\mu(i,h) - K_3(h)W_\sigma(i,h)] . \qquad (2.5.19)$$

Estimators of μ and/or σ based on sets of k asymptotically optimal spacings are given as follows:

1. Let h^* be defined as in (2.5.5). An estimator for μ when σ is known is

$$\mu^* = \sum_{i=1}^{k} W_\mu(i,h^*)\tilde{Q}(H^{*-1}\left(\frac{i}{k+1}\right)) - \sigma C_\mu(h^*) \qquad (2.5.20)$$

with

$$ARE(\mu^*) = \frac{K_1(h^*)}{\langle f_o Q_o, f_o Q_o \rangle} . \qquad (2.5.21)$$

2. Let h^* be defined as in (2.5.6). An estimator for σ when μ is known is

$$\sigma^* = \sum_{i=1}^{k} W_\sigma(i,h^*)\tilde{Q}(H^{*-1}\left(\frac{i}{k+1}\right)) - \mu C_\sigma(h^*) \qquad (2.5.22)$$

with

$$ARE(\sigma^*) = \frac{K_2(h^*)}{\langle f_o Q_o \cdot Q_o, f_o Q_o \cdot Q_o \rangle} \quad . \tag{2.5.23}$$

3. Let h^* be defined as in (2.5.9). Simultaneous estimators of μ and σ are

$$\mu^* = \sum_{i=1}^{k} S_\mu(i,h^*)\bar{Q}(H^{*-1}\left(\frac{i}{k+1}\right)) \tag{2.5.24}$$

$$\sigma^* = \sum_{i=1}^{k} S_\sigma(i,h^*)\bar{Q}(H^{*-1}\left(\frac{i}{k+1}\right)) \tag{2.5.25}$$

with

$$ARE(\mu^*,\sigma^*) = \frac{\Delta(h^*)}{|A|} \quad . \tag{2.5.26}$$

It should be noted that these formulae may be adapted to spacings generated by an arbitrary density, h, by substituting h for h^* in equations (2.5.20) through (2.5.26).

The remainder of this section will be devoted to considering special cases of the previous results. These serve to point out certain simplifications as well as some shortcomings of Theorem 2.5.1. First the case of a symmetric distribution (in particular the normal) will be considered. Secondly, the exponential distribution will be seen to pose certain problems in the application of Theorem 2.5.1.

For a symmetric distribution it can be shown that the off diagonal elements of the matrix A in (2.5.8) are zero (see Parzen (1979)). This forumla (2.5.9) may be simplified somewhat in this case.

As a specific example of a symmetric distribution, consider the normal distribution. The c.d.f. and p.d.f. of the normal distribution are often denoted by Φ and ϕ respectively. In this notation, the functions that have been considered so far are

$$Q_o(u) = \Phi^{-1}(u) \quad , \tag{2.5.27}$$

$$f_o Q_o(u) = (2\pi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}|\Phi^{-1}(u)|^2\} \quad . \tag{2.5.28}$$

It follows from (2.5.28) that

$$-f_o Q_o''(u) = \frac{1}{f_o Q_o(u)} \quad , \tag{2.5.29}$$

$$-[f_o Q_o(u) Q_o(u)]' = 2\frac{Q_o(u)}{f_o Q_o(u)} \tag{2.5.30}$$

and $A = \text{diag}(1,2)$.

The optimal density for simultaneous estimation of μ and σ is

$$h^*(u) = \frac{(1 + 2|\Phi^{-1}(u)|^2)^{\frac{1}{3}} \exp\{\tfrac{1}{3}|\Phi^{-1}(u)|^2\}}{\int_0^1 (1 + 2|\Phi^{-1}(t)|^2)^{\frac{1}{3}} \exp\{\tfrac{1}{3}|\Phi^{-1}(t)|^2\} dt} \quad . \tag{2.5.31}$$

The corresponding optimal c.d.f., H^*, must be tabulated by numeric integration (see Chapter 3). For a given k, the asymptotically

optimal spacings, $H^{*-1}\left(\frac{i}{k+1}\right)$, $i = 1, \ldots, k$, can then be found by interpolation.

For the exponential distribution

$$Q_o(u) = -\ln(1-u) \quad , \tag{2.5.32}$$

$$f_o Q_o(u) = 1 - u \quad . \tag{2.5.33}$$

The $f_o Q_o$ function in this case cannot be represented as in (2.5.3). In fact the $f_o Q_o$ function is not even in the RKHS of K_B as $f_o Q_o(0) \neq 0$.[1] Consequently, Theorem 2.5.1 does not pertain to the $f_o Q_o$ function in (2.5.33). This means that for exponential data, estimation of μ when σ is known or simultaneous estimation of μ and σ may not be accomplished using the theory developed in this section.

However, the function $f_o Q_o \cdot Q_o$ does possess the desired representation. Therefore, it is possible to obtain spacings for estimating σ when μ is known via equation 2.5.6. The optimal density is

$$h^*(u) = \frac{(u-1)^{-2/3}}{\int_0^1 (t-1)^{-2/3} dt} \quad . \tag{2.5.34}$$

[1]Other distributions (e.g. the Pareto) also have this problem. In such cases, there appears to be a correspondence between the distribution being non-regular and the $f_o Q_o$ functions not being a member of the RKHS of K_B. Whether this holds in general is a topic for further research.

The asymptotically optimal spacings generated by h^* are

$$H^{*-1}\left(\frac{i}{k+1}\right) = 1 - \left(1 - \frac{i}{k+1}\right)^3 , \; i = 1, \ldots, k \; . \tag{2.5.35}$$

2.6 Comparison with Other Approaches

In this section the results of the previous section are applied to solve problems considered by Chernoff (1971) and Sarndal (1962).

Chernoff (1971) considers the optimal spacings problem for the normal distribution. He assumes that the normalized quantiles, z_i, are selected according to some non-negative density function $g(z)$. This may be interpreted as meaning that for large k there should be approximately $kg(z)\Delta$ quantiles in a small interval $(z, z + \Delta)$.

Using the notation adopted in Section 2.5 for the normal distribution, the ARE of a linear systematic estimator of μ is proportional to

$$K_1 = \sum_{i=1}^{k+1} \frac{[\phi(z_i) - \phi(z_{i-1})]^2}{\phi(z_i) - \phi(z_{i-1})} \quad . \tag{2.6.1}$$

Chernoff writes K_1 as

$$K_1 = \sum_{i=1}^{k+1} \left[\frac{\phi(z_i) - \phi(z_{i-1})}{\phi(z_i) - \phi(z_{i-1})}\right]^2 [\phi(z_i) - \phi(z_{i-1})] \tag{2.6.2}$$

and notes that for large k

$$K_1 = \int_{-\infty}^{\infty} \left[\frac{d\phi(z)}{d\phi(z)}\right]^2 d\phi(z) = \int_{-\infty}^{\infty} z^2 \phi(z) dz = 1 \quad . \tag{2.6.3}$$

Thus K_1 is a discrete approximation to

$$\int_{-\infty}^{\infty} z^2 \phi(z) dz \quad . \tag{2.6.4}$$

Therefore the problem of selecting optimal spacings may be viewed as one of selecting a best set of points to discretely approximate (2.6.4).

By expanding $\phi(z)$ in its Taylor series about z_{i-1} the differences between the integral (2.6.4) and K_1 can be shown to be approximated for large k as follows:

$$\tfrac{1}{12} \sum_{i=1}^{k+1} (z_i - z_{i-1})^3 \phi(z_{i-1}) \approx \frac{1}{12k^2} \sum_{i=1}^{k+1} \frac{\phi(z_{i-1})}{g^2(z_{i-1})} (z_i - z_{i-1})$$

$$\approx \frac{1}{12k^2} \int_{-\infty}^{\infty} \frac{\phi(z)}{g^2(z)} dz \tag{2.6.5}$$

since $(z_i - z_{i-1}) \approx \frac{1}{g(z_{i-1})k}$. Thus minimizing the error in approximating (2.6.4) and selecting the density g^* that minimizes the right hand side of (2.6.5) are (asymptotically) equivalent problems. Chernoff solves this by variational methods to obtain

$$g^*(z) = \frac{1}{C} \phi^{1/3}(z) \tag{2.6.6}$$

where $C = \int_{-\infty}^{\infty} \phi^{1/3}(z) dz$. As $k \to \infty$ the z_i correspond to approximately the $\frac{i}{k+1}$ quantile of the g(z) distribution. Thus Chernoff's solution is to take the z_i such that they satisfy

$$\int_{-\infty}^{z_i} \phi^{1/3}(z) dz = \frac{i}{k+1} \int_{-\infty}^{\infty} \phi^{1/3}(z) dz \quad . \tag{2.6.7}$$

The error in approximating (2.6.4) is approximately

$$\frac{1}{k^2 12} \left[\int_{-\infty}^{\infty} \phi^{1/3}(z) dz\right]^3 \quad . \tag{2.6.8}$$

Equation (2.6.7) can be seen to be the same solution as suggested in Theorem 2.5.1 by making the change of variable $z = \phi^{-1}(u)$. This same procedure shows that the problem of finding a density which minimizes the error term (2.6.5) is identical to the problem of finding a density that attains the bound (2.3.13) (here $f_1 = \phi\phi^{-1}$). Thus not only the solutions but the problems themselves are the same.

Särndal (1962) treats the problem of selecting optimal spacings as one of selecting an optimal generating function. He defines a generating function to be a non-negative density function defined on an interval [a,b] of the real line. Let G be the c.d.f. that corresponds to a density function g. Denote by G^{-1} the inverse function of G. A set of k spacings, $\{u_1, u_2, \ldots, u_k\}$, taken according to g satisfy

$$u_i = F_o\left(G^{-1}\left(\frac{i}{k+1}\right)\right) \tag{2.6.9}$$

where F_o is the c.d.f. in (2.4.1).

Under the assumption that g and its first derivative are bounded and continuous with $g > 0$, the author shows that the loss in efficiency in estimation using linear estimates based on spacings chosen according to g is

$$L = \frac{1}{12(k+1)^2} \int_{-\infty}^{\infty} \frac{J^*(x) f_o(x)}{[g(x)]^2} + O(k^{-2}) \qquad (2.6.10)$$

where

$$J^*(x) = \frac{d^2 \log f_o(x)}{dx^2} = \nu_1(x) \qquad \text{for } \sigma \text{ known},$$

$$= \frac{d}{dx}\left[x \frac{d \log f_o(x)}{dx}\right] = \nu_2(x) \qquad \text{for } \mu \text{ known}, \qquad (2.6.11)$$

$$= [\nu_1(x), \nu_2(x)] A^{-1} \begin{bmatrix} \nu_1(x) \\ \nu_2(x) \end{bmatrix} \qquad \text{for both } \mu \text{ and } \sigma \text{ known}.$$

He then defines nearly optimal spacings to be those obtained according to a generating function that minimizes (2.6.10). A calculus of variations argument shows the optimal function to be

$$g^*(x) = \frac{1}{C} [J^*(x) f_o(x)]^{1/3} \qquad (2.6.12)$$

where

$$C = \int_{-\infty}^{\infty} [J^*(x) f_o(x)]^{1/3} dx \quad . \qquad (2.6.13)$$

Again by letting $x = Q_o(u)$ and substituting the functions $f_o Q_o(u)$ and $f_o Q_o(u) Q_o(u)$ for those in Remarks 2.3.2 and 2.3.3, the problems considered by Särndal and subsequent solutions are seen to be equivalent to those in Section 2.5.

3. APPLICATIONS

3.1 Preliminaries

The regression model for location and scale parameter estimation has been seen to be

$$f_o Q_o(u)\tilde{Q}(u) = \mu f_o Q_o(u) + \sigma Q_o(u) f_o Q_o(u) + \sigma_B B(u) \qquad (3.1.1)$$

where

$$\sigma_B = \sigma/\sqrt{n} \qquad (3.1.2)$$

and $\{B(u),\ u \in [0,1]\}$ *is a Brownian Bridge process.* From this model, it can be deduced that selecting optimal spacings is equivalent to selecting optimal regression designs.

In Chapter 2, design sequences that were asymptotically optimal were considered. Such sequences for the Brownian Bridge process were generated by density functions on [0,1]. For the model (3.1.1), the optimal densities are given by

$$h^*(u) = \frac{[f_o Q_o(u)'']^{2/3}}{\int_0^1 [f_o Q_o(t)'']^{2/3} dt} \qquad \text{if } \sigma \text{ is known}$$

$$= \frac{([Q_o(u) f_o Q_o(u)]'')^{2/3}}{\int_0^1 ([Q_o(t) f_o Q_o(t)]'')^{2/3} dt} \qquad \text{if } \mu \text{ is known} \qquad (3.1.3)$$

$$= \frac{[\psi'(u)A^{-1}\psi(u)]^{1/3}}{\int_0^1 [\psi'(t)A^{-1}\psi(t)]^{1/3} dt} \qquad \text{if } \mu \text{ and } \sigma \text{ are unknown,}$$

where

$$\psi'(u) = -([f_o Q_o(u)]'',\ [Q_o(u) f_o Q_o(u)]'') \qquad (3.1.4)$$

and A is the information matrix defined in Section 2.5.

In order to determine the asymptotically optimal set of spacings for a given number, k, of order statistics, it is first necessary to compute the optimal density, h*, and its corresponding c.d.f. H*. Then the required spacings are the points $H^{*-1}\left(\frac{i}{k+1}\right)$, $i = 1, \ldots, k$.

To utilize the preceding theory for data analysis, the researcher would require the H^{*-1} or H* functions for many of the common distributional forms. Such functions are derived in Section 3.2. Comparison of the spacings obtained using (3.1.3) with those obtained by other authors are also provided.

Once a spacings set has been decided upon, the estimators can then be constructed. The ABLUE's for a specified spacings set, $u_1, u_2, \ldots, u_k$, are given as follows:

1. When σ is known, the estimator for μ is

$$\hat{\mu} = \sum_{i=1}^{k} W_\mu(i)\tilde{Q}(u_i) - \sigma C_\mu . \qquad (3.1.5)$$

2. When μ is known, the estimator for σ is

$$\hat{\sigma} = \sum_{i=1}^{k} W_\sigma(i)\tilde{Q}(u_i) - \mu C_\sigma . \qquad (3.1.6)$$

3. Simultaneous estimators for μ and σ are

$$\hat{\mu} = \sum_{i=1}^{k} S_{\mu}(i)\tilde{Q}(u_i) \qquad (3.1.7)$$

and

$$\hat{\sigma} = \sum_{i=1}^{k} S_{\sigma}(i)\tilde{Q}(u_i) \quad . \qquad (3.1.8)$$

The exact formulas for these estimators and their efficiencies may be obtained by referring to Section 2.5. In subsequent sections, the coefficients W_{μ}, W_{σ}, S_{μ}, S_{σ}, and the correction factors C_{μ} and C_{σ} will be presented for certain spacings sets of interest. It should be noted that the correction factors are zero in the case of symmetric spacings for a symmetric distribution.

In Section 3.3, the problem of selecting order statistics for summarizing large data sets will be considered. It will be seen that the selection of a few strategically placed order statistics will provide sufficient information to construct efficient location and scale parameter estimators under a variety of distributional assumptions.

3.2 Spacings for Some Common Distributions

3.2.1. Normal Distribution

The c.d.f. and p.d.f. of the standard normal distribution are usually denoted by Φ and ϕ respectively. In this notation, the functions to be considered are:

$$f_o(x) = \phi(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}, \quad -\infty < x < \infty, \qquad (3.2.1)$$

$$F_o(x) = \Phi(x) = \int_{-\infty}^{x} \phi(t)dt \quad , \qquad (3.2.2)$$

$$Q_o(u) = \Phi^{-1}(u) \quad , \qquad (3.2.3)$$

and

$$f_oQ_o(u) = \phi\Phi^{-1}(u) = (2\pi)^{-1/2}\exp\{-|\Phi^{-1}(u)|^2/2\} \quad . \qquad (3.2.4)$$

The matrix A is found to be

$$A = \mathrm{diag}(1,2) \quad . \qquad (3.2.5)$$

When σ is known spacings taken according to the density

$$h^*(u) = \frac{[\phi\Phi^{-1}(u)]^{-2/3}}{\int_0^1 [\phi\Phi^{-1}(t)]^{-2/3}dt} \quad . \qquad (3.2.6)$$

will be asymptotically optimal for estimating the mean. Since h^* is symmetric, the spacings, u_i, will satisfy

$$u_{k-i+1} = 1 - u_i \quad , \quad i = 1, \ldots, k \quad . \qquad (3.2.7)$$

Using (3.2.6) it can be shown that

$$H^{*-1}(u) = \Phi(\sqrt{3}\,\Phi^{-1}(u)) \quad . \qquad (3.2.8)$$

The function in (3.2.8) is shown graphically in Figure 1 and is tabulated in Table 1 for points in the interval [0,.5]. To find

Figure 1. Normal Distribution, σ Known;
The Function $H^{*-1}(u)$

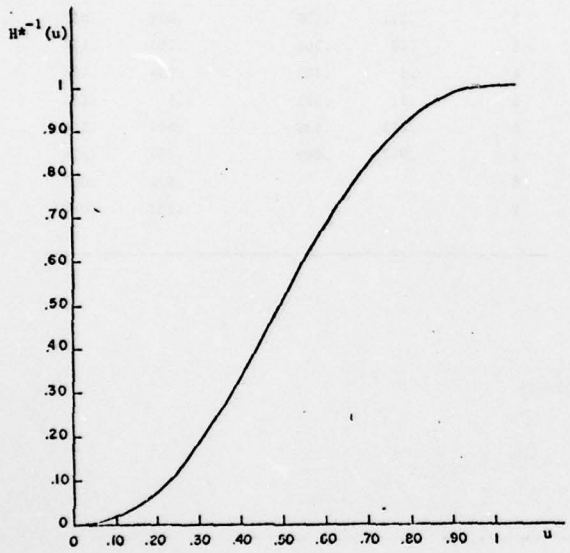


Table 1. Normal Distribution, σ Known;
The Function $H^{*-1}(u)$

u	$H^{*-1}(u)$	u	$H^{*-1}(u)$
.01	.00003	.26	.13136
.02	.00019	.27	.14457
.03	.00056	.28	.15865
.04	.00122	.29	.16853
.05	.00226	.3	.18141
.06	.00368	.31	.19489
.07	.00539	.32	.20897
.08	.00776	.33	.22363
.09	.01017	.34	.23885
.1	.01321	.35	.25143
.11	.01659	.36	.26763
.12	.02068	.37	.28434
.13	.025	.38	.29806
.14	.03074	.39	.31561
.15	.03593	.4	.32997
.16	.04272	.41	.34458
.17	.04947	.42	.36317
.18	.05705	.43	.3707
.19	.06426	.44	.39743
.2	.07353	.45	.41293
.21	.08226	.46	.4325
.22	.09176	.47	.44828
.23	.10027	.48	.46414
.24	.11123	.49	.48404
.25	.121	.5	.5

H^{*-1} values for $u > .5$ the relation (3.2.7) may be used. Table 2 contains spacings calculated using (3.2.8) and their corresponding coefficients for $k = 7,9$. The asymptotically optimal spacings for $k = 2,7,9$ are compared to the optimal sets obtained by Ogawa (1951) in Table 3.

To estimate σ when μ is known, the optimal density is

$$h^*(u) = \frac{[\phi^{-1}(u)/\phi\phi^{-1}(u)]^{2/3}}{\int_0^1 [\phi^{-1}(t)/\phi\phi^{-1}(t)]^{2/3}dt} \quad . \qquad (3.2.9)$$

A tabulation of the function $H^{*-1}(u)$ is given in Table 4 for u in the interval $[0,.5]$. For other values of u the relation (3.2.7) may be used. A graph of H^{*-1} is shown in Figure 2. Table 5 contains the asymptotically optimal spacings and corresponding coefficients and efficiencies for $k = 7,9$. Table 6 provides a comparison of the asymptotically optimal spacings with the optimal spacing given by Ogawa (1951) for $k = 2,6$.

For simultaneous estimation of μ and σ, spacings should be taken according to the density

$$h^*(u) = \frac{(1 + 2|\phi^{-1}(u)|^2)^{1/3}\exp\{1/3|\phi^{-1}(u)|^2\}}{\int_0^1 (1 + 2|\phi^{-1}(t)|^2)^{1/3}\exp\{1/3|\phi^{-1}(t)|^2\}dt} \quad . \qquad (3.2.10)$$

Since the spacings generated by h^* are symmetric, the function H^{*-1} has been tabulated only over the interval $[0,.5]$. This tabulation is given in Table 7. A graph of H^{*-1} appears in Figure 3. Table 8

Table 2. Normal Distribution, σ Known; Asymptotically Optimal Spacings and Coefficients for Seven or Nine Order Statistics

	$k = 7$		$k = 9$	
i	u_i	$W_\mu(i)$	u_i	$W_\mu(i)$
1	.023	.049	.013	.028
2	.121	.138	.074	.087
3	.29	.201	.184	.134
4	.5	.223	.334	.164
5	.71	.201	.5	.173
6	.879	.138	.666	.164
7	.977	.049	.816	.134
8			.926	.087
9			.987	.028

Table 3. Normal Distribution, σ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

Spacing	k = 2		k = 7		k = 9	
	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.227	.27	.023	.04	.013	.024
u_2	.773	.73	.121	.147	.074	.092
u_3			.29	.308	.184	.202
u_4			.5	.5	.334	.343
u_5			.71	.692	.5	.5
u_6			.879	.852	.666	.657
u_7			.977	.96	.816	.798
u_8					.926	.908
u_9					.987	.976
Efficiency	.8005	.8097	.9637	.9654	.976	.9771

Table 4. Normal Distribution, μ Known; The Function $H^{*-1}(u)$

u	$H^{*-1}(u)$	u	$H^{*-1}(u)$
.01	.0	.26	.04176
.02	.00002	.27	.04712
.03	.00006	.28	.05296
.04	.00013	.29	.05931
.05	.00026	.3	.0662
.06	.00045	.31	.07367
.07	.00072	.32	.08176
.08	.00106	.33	.09051
.09	.00156	.34	.09996
.1	.00215	.35	.11017
.11	.00288	.36	.1212
.12	.00376	.37	.1331
.13	.00481	.38	.14596
.14	.00604	.39	.15987
.15	.00747	.4	.17492
.16	.00912	.41	.19124
.17	.01101	.42	.209
.18	.01314	.43	.2284
.19	.01556	.44	.2497
.2	.01826	.45	.27328
.21	.02127	.46	.29968
.22	.02461	.47	.32982
.23	.02831	.48	.36537
.24	.03239	.49	.41045
.25	.03686	.5	.5

Figure 2. Normal Distribution, μ Known; The Function $H^{*-1}(u)$

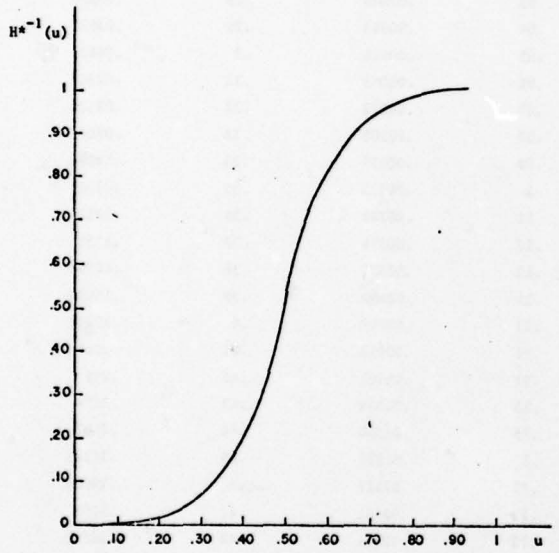


Table 5. Normal Distribution, μ Known; Asymptotically Optimal Spacings, Coefficients and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\mu(i)$	u_i	$W_\mu(i)$
1	.004	-.031	.002	-.015
2	.037	-.111	.018	-.062
3	.14	-.2	.066	-.115
4	.5	.0	.175	-.164
5	.86	.2	.5	.0
6	.963	.111	.825	.164
7	.996	.031	.934	.115
8			.982	.062
9			.998	.015
Efficiency	.8848		.9231	

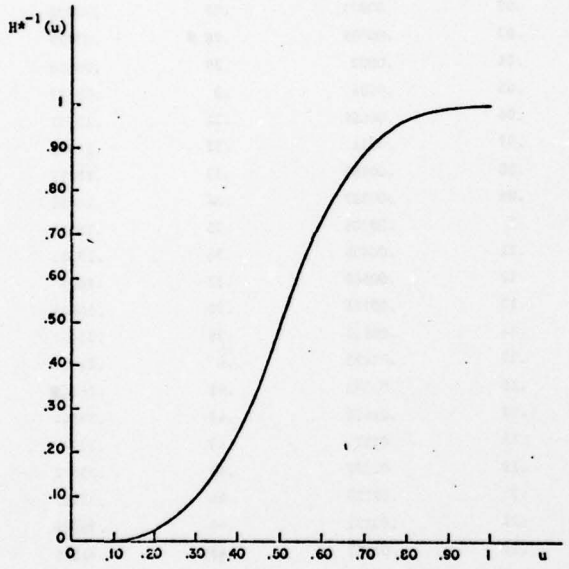
Table 6. Normal Distribution, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 6	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.091	.069	.006	.01
u_2	.909	.931	.056	.055
u_3			.228	.17
u_4			.772	.83
u_5			.944	.945
u_6			.994	.99
Efficiency	.6415	.6522	.8761	.8943

Table 7. Normal Distribution, Both μ and σ Unknown; The Function $H^{*-1}(u)$

u	$H^{*-1}(u)$	u	$H^{*-1}(u)$
.01	.0	.26	.06164
.02	.00003	.27	.06938
.03	.00009	.28	.07779
.04	.0002	.29	.08688
.05	.0004	.3	.09672
.06	.00069	.31	.10731
.07	.0011	.32	.11872
.08	.00166	.33	.13097
.09	.00237	.34	.14411
.1	.00326	.35	.15819
.11	.00436	.36	.17323
.12	.00569	.37	.1893
.13	.00726	.38	.20644
.14	.00911	.39	.22467
.15	.01125	.4	.24407
.16	.01372	.41	.26465
.17	.01652	.42	.28644
.18	.0197	.43	.30947
.19	.02327	.44	.33372
.2	.02727	.45	.35917
.21	.03171	.46	.38574
.22	.03663	.47	.41331
.23	.04205	.48	.44171
.24	.048	.49	.4707
.25	.05453	.5	.5

Figure 3. Normal Distribution, μ and σ Unknown; The Function $H^{*-1}(u)$



contains asymptotically optimal spacings and the corresponding coefficients and efficiencies for estimators based on seven or nine order statistics.

In the case that k = 2, Ogawa (1951) has shown that the spacings $u_1 = .134$ and $u_2 = .866$ with efficiency .4066 are optimal for the simultaneous estimation of μ and σ. The spacings obtained using (3.2.10) are $u_1^* = .132$ and $u_2^* = .868$. The efficiency for this latter spacings set is .4065.

3.2.2 Exponential Distribution

The exponential c.d.f. is

$$F_o(x) = 1 - \exp\{-x\}, \quad x > 0 . \tag{3.2.11}$$

The functions required in construction of the optimal density are:

$$f_o(x) = \exp\{-x\} , \tag{3.2.12}$$

$$Q_o(u) = -\log(1 - u) , \tag{3.2.13}$$

and

$$f_o Q_o(u) = 1 - u . \tag{3.2.14}$$

Thus

$$\langle f_o Q_o \cdot Q_o, f_o Q_o \cdot Q_o \rangle = 1 . \tag{3.2.15}$$

Table 8. Normal Distribution, Both μ and σ Unknown; Asymptotically Optimal Spacings, Coefficients, and Efficiencies for Seven or Nine Order Statistics

	k = 7			k = 9		
i	u_i	$S_\mu(i)$	$S_\sigma(i)$	u_i	$S_\mu(i)$	$S_\sigma(i)$
1	.006	.018	-.025	.003	.009	-.012
2	.005	.093	-.082	.027	.044	-.045
3	.197	.24	-.106	.096	.109	-.075
4	.5	.357	.0	.244	.214	-.075
5	.802	.24	.106	.5	.286	.0
6	.945	.093	.082	.756	.214	.075
7	.994	.018	.025	.903	.109	.075
8				.973	.044	.045
9				.997	.009	.012
Efficiency	.8382			.8903		

As was seen in Section 2.5, estimation of μ when σ is known or simultaneous estimation of both μ and σ could not be accomplished using the design techniques of Chapter 2. Consequently, only the estimation of σ when μ is known will be considered.

When the location parameter is known, the optimal density of scale parameter estimation is

$$h^*(u) = \frac{(u-1)^{-2/3}}{\int_0^1 (t-1)^{-2/3} dt} \quad . \tag{3.2.16}$$

Thus

$$H^{*-1}(u) = 1 - (1-u)^3 \quad . \tag{3.2.17}$$

The function (3.2.17) is shown graphically in Figure 4.

The optimal spacings for estimating σ when μ is known have been found by Sarhan and Greenberg (1958). A comparison of the performance of these optimal spacings with that of the spacings obtained using H^{*-1} is provided by Table 9 in the case of k = 2,7,9. The asymptotically optimal spacings, correction factors and coefficients are given in Table 10 for estimators based on either seven or nine order statistics.

3.2.3 Pareto Distribution

The distribution function for the Pareto distribution is

$$F_Q(x) = 1 - (1+x)^{-\nu} \quad , \quad x > 0 \quad , \tag{3.2.18}$$

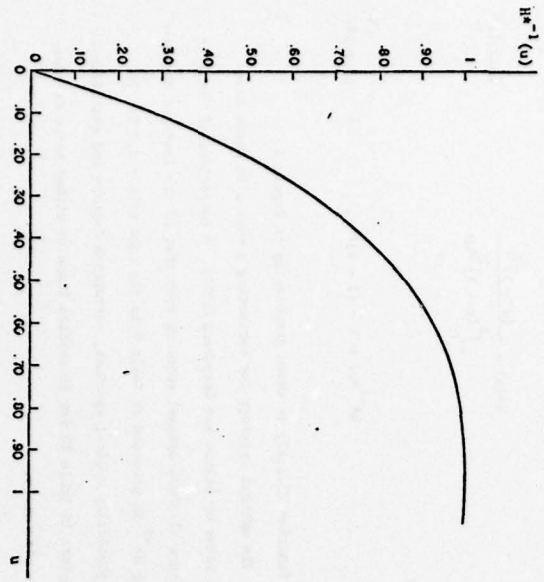


Figure 4. Exponential Distribution, μ Known; The Function $H^{*-1}(u)$

Table 9. Exponential Distribution, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.704	.637	.33	.312	.271	.258
u_2	.963	.927	.578	.551	.488	.468
u_3			.756	.728	.657	.634
u_4			.875	.851	.784	.761
u_5			.974	.93	.875	.855
u_6			.984	.975	.936	.921
u_7			.998	.995	.973	.963
u_8					.992	.987
u_9					.999	.997
Efficiency	.806	.8203	.958	.9693	.978	.980

Table 10. Exponential Distribution, μ Known; Asymptotically Optimal Spacings, Coefficients and Correction Factors for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.33	.297	.271	.248
2	.578	.219	.488	.196
3	.756	.152	.657	.15
4	.875	.129	.784	.11
5	.974	.033	.875	.076
6	.984	.016	.936	.049
7	.998	.005	.973	.027
8			.992	.012
9			.999	.001
C_σ		.849		.869

where $\nu > 0$ is a known shape parameter. The required functions are:

$$f_o(x) = \nu(1 + x)^{-(\nu+1)} , \tag{3.2.19}$$

$$Q_o(u) = (1 - u)^{-1/\nu} - 1 , \tag{3.2.20}$$

and

$$f_oQ_o(u) = \nu(1 - u)^{1 + 1/\nu} . \tag{3.2.21}$$

Since the f_oQ_o function does not vanish when evaluated at zero, it cannot be in the RKHS generated by the covariance kernel K_B (see Section 2.2 for this notation). Thus as in the case of the exponential distribution, optimum spacings for estimating μ when σ is known or for simultaneous parameter estimation cannot be obtained using the theory of Chapter 2. However, the $Q_o \cdot f_oQ_o$ function has the desired properties so the estimation of σ when μ is known using asymptotically optimal spacings can still be accomplished.

The optimal density for scale parameter estimation in the case that μ is known is

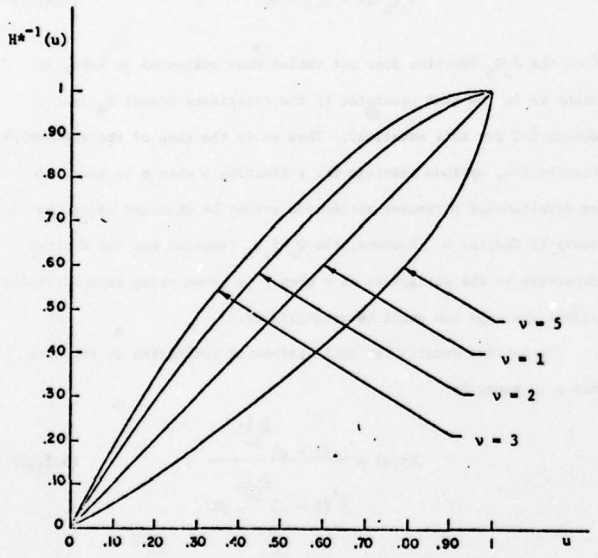
$$h^*(u) = \frac{(1 - u)^{\frac{2-2\nu}{3\nu}}}{\int_0^1 (1 - t)^{\frac{2-2\nu}{3\nu}} \, dt} . \tag{3.2.22}$$

Thus

$$H^{*-1}(u) = 1 - (1 - u)^{\frac{3\nu}{2+\nu}} . \tag{3.2.23}$$

The function H^{*-1} is shown graphically in Figure 5 for $\nu = .5, 1, 2, 3$.

Figure 5. Pareto Distribution, μ Known, $\nu = .5,1,2,3$; The Function $H^{*-1}(u)$



In the case that $\nu = 1$, the spacings obtained from (3.2.23) are the points $\frac{i}{k+1}$, $i = 1, \ldots, k$. This solution agrees with the optimal spacings given by Kulldorf and Vännman (1973). An explicit expression for the ABLUE is

$$\hat{\sigma} = \frac{6}{k(k+1)(k+2)} \sum_{i=1}^{k} (k - i + 1)^2 \tilde{Q}\left(\frac{i}{k+1}\right) - \frac{2k+1}{k+2}\mu. \tag{3.2.24}$$

For $\nu \neq 1$ the spacings generated by h^* are only asymptotically optimal. A comparison of the optimal spacings with those generated by (3.2.22) is presented in Tables 11-13 for the cases $\nu = .5,2,3$ respectively. Tables 14-16 contain the asymptotically optimal spacing and corresponding coefficients and correction factors for estimators based on either seven or nine order statistics for when $\nu = .5,2,3$ respectively.

3.2.4 Cauchy Distribution

For the Cauchy distribution, the required functions are:

$$F_o(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x), \quad -\infty < x < \infty, \tag{3.2.25}$$

$$f_o(x) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{3.2.26}$$

$$Q_o(u) = \tan[\pi(u - \tfrac{1}{2})] \tag{3.2.27}$$

$$f_o Q_o(u) = \pi^{-1}\sin^2(\pi u). \tag{3.2.28}$$

Table 11. Pareto Distribution, $\nu = .5$, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.216	.224	.077	.078	.061	.062
u_2	.483	.503	.159	.161	.125	.127
u_3			.246	.249	.193	.195
u_4			.34	.345	.264	.267
u_5			.445	.452	.34	.344
u_6			.565	.575	.423	.428
u_7			.713	.728	.514	.522
u_8					.619	.629
u_9					.749	.762
Efficiency	.9027	.9033	.9869	.987	.9917	.9917

Table 12. Pareto Distribution, $\nu = 2$, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.456	.44	.182	.179	.146	.144
u_2	.808	.786	.35	.345	.284	.281
u_3			.506	.499	.414	.409
u_4			.646	.638	.535	.529
u_5			.77	.761	.646	.639
u_6			.875	.866	.747	.74
u_7			.956	.949	.836	.828
u_8					.911	.904
u_9					.968	.963
Efficiency	.8701	.8711	.9808	.9809	.9877	.9877

Table 13. Pareto Distribution, $\nu = 3$, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.518	.491	.214	.208	.173	.169
u_2	.862	.832	.404	.395	.331	.324
u_3			.571	.559	.471	.465
u_4			.713	.7	.601	.591
u_5			.829	.816	.712	.702
u_6			.918	.906	.808	.797
u_7			.976	.969	.885	.876
u_8					.945	.931
u_9					.984	.979
Efficiency	.8581	.8606	.9785	.9787	.9861	.9862

Table 14. Pareto Distribution, $\nu = .5$, μ Known; Asymptotically Optimal Spacings, Coefficients and Correction Factors for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.077	.438	.061	.367
2	.159	.321	.125	.292
3	.246	.222	.193	.223
4	.34	.143	.264	.163
5	.445	.08	.34	.114
6	.565	.036	.423	.073
7	.713	.009	.514	.041
8			.619	.018
9			.749	.005
C_σ		1.308		1.295

Table 15. Pareto Distribution, $\nu = 2$, μ Known; Asymptotically Optimal Spacings, Coefficients and Correction Factors for Seven or Nine Order Statistics

i	k = 7		k = 9	
	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.182	.876	.146	.737
2	.35	.645	.284	.583
3	.506	.447	.414	.447
4	.646	.286	.535	.328
5	.77	.161	.646	.228
6	.875	.071	.747	.146
7	.956	.017	.836	.082
8			.911	.036
9			.968	.009
C_σ		2.503		2.595

Table 16. Pareto Distribution, $\nu = 3$, μ Known; Asymptotically Optimal Spacings, Coefficients and Correction Factors for Seven or Nine Order Statistics

i	k = 7		k = 9	
	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.214	1.171	.173	.985
2	.404	.861	.331	.769
3	.571	.598	.471	.598
4	.713	.382	.601	.441
5	.829	.215	.712	.305
6	.918	.094	.808	.194
7	.976	.022	.885	.11
8			.945	.048
9			.984	.011
C_σ		3.343		3.461

For the computation of efficiencies it should be noted that

$$\langle f_o Q_o, f_o Q_o \rangle = \frac{1}{2} \quad . \tag{3.2.29}$$

When the scale parameter is assumed known, spacings taken according to the density

$$h^*(u) = \frac{[2\pi\cos(2\pi u)]^{2/3}}{\int_0^1 [2\pi\cos(2\pi t)]^{2/3} dt} \tag{3.2.30}$$

will be asymptotically optimal for estimating the location parameter. A tabulation of the optimal c.d.f., H*, is given in Table 17 for the interval [0,.5]. Values for H^{*-1} may be calculated from the table by interpolation in combination with the use of (3.2.7). A graph of H^{*-1} is shown in Figure 6.

Bloch (1966) has considered the estimation of μ by a linear function of five order statistics. The optimal spacings set in this case was found to be {.13, .4, .5, .6, .87} yielding an asymptotic relative efficiency of .9516. Interpolation in Table 17 results in the spacings set {.125, .372, .5, .628, .871}. This latter set has .9481 as its efficiency. Asymptotically optimal spacings and their corresponding coefficients and efficiencies are given in Table 18 for either seven or nine order statistics.

3.2.5 Logistic Distribution

A frequent parameterization for the logistic distribution is

$$F(x) = \left[1 + \exp\left\{-\frac{\pi}{\sqrt{3}}\frac{(x-\mu)}{\sigma}\right\}\right]^{-1}, \quad -\infty < x < \infty \ , \tag{3.2.31}$$

Table 17. Cauchy Distribution, σ Known; The Function H*(u)

u	H*(u)	u	H*(u)
.01	.01402	.26	.25133
.02	.028	.27	.25422
.03	.04191	.28	.25829
.04	.0557	.29	.26336
.05	.06935	.3	.26935
.06	.08281	.31	.27616
.07	.09604	.32	.28373
.08	.10902	.33	.29202
.09	.12171	.34	.30095
.1	.13406	.35	.3105
.11	.14605	.36	.32061
.12	.15762	.37	.33125
.13	.16875	.38	.34238
.14	.17939	.39	.35395
.15	.1895	.4	.36594
.16	.19905	.41	.37829
.17	.20798	.42	.39098
.18	.21627	.43	.40395
.19	.22384	.44	.41719
.2	.23065	.45	.43065
.21	.23664	.46	.4443
.22	.24171	.47	.45809
.23	.24578	.48	.472
.24	.24867	.49	.48598
.25	.25	.5	.5

Figure 6. Cauchy Distribution, σ Known; The Function $H^{*-1}(u)$

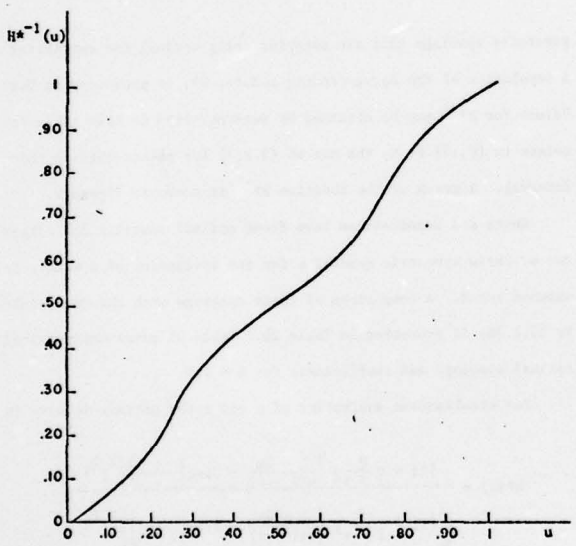


Table 18. Cauchy Distribution, σ Known; Asymptotically Optimal Spacings, Coefficients, and Efficiencies for Seven or Nine Order Statistics

i	k = 7		k = 9	
	u_i	$W_\mu(i)$	u_i	$W_\mu(i)$
1	.093	-.031	.073	-.014
2	.25	.0	.161	-.048
3	.407	.343	.339	.104
4	.5	.376	.427	.26
5	.593	.343	.5	.397
6	.75	.0	.573	.26
7	.907	-.031	.661	.104
8			.839	-.048
9			.927	-.014
Efficiency	.9579		.9743	

where μ and σ are respectively the mean and standard deviation of the distribution. The introduction of the factor $\pi/\sqrt{3}$ requires a slight modification of the model (3.1.1). In this case the model becomes

$$f_1(u)\bar{Q}(u) = \mu f_1(u) + \sigma f_2(u) + \sigma_B B(u) \qquad (3.2.32)$$

where

$$f_1(u) = \frac{\pi}{\sqrt{3}}\, u(1-u)\,, \qquad (3.2.33)$$

$$f_2(u) = u(1-u)\log\left(\frac{u}{1-u}\right)\,, \qquad (3.2.34)$$

$$\sigma_B = \sigma/\sqrt{n}\,, \qquad (3.2.35)$$

and $\{B(u),\ u \in [0,1]\}$ is a Brownian Bridge process. The work of Gupta and Gnanadesikan (1966) may be used to deduce that for the model (3.2.32)

$$A = \mathrm{diag}\left(\frac{\pi^2}{9}\,,\ \frac{3+\pi^2}{9}\right)\,. \qquad (3.2.36)$$

To estimate μ when σ is known, spacings should be taken according to the uniform distribution on [0,1], i.e.,

$$H^{*-1}(u) = u\,. \qquad (3.2.37)$$

For a given value of k, the spacings are the points $\frac{i}{k+1}$, $i = 1, \ldots, k$, which was seen to be the optimal solution in Section 1.2.6.[2]

[2]This is the same as the solution obtained for the Pareto distribution with $\gamma = 1$. Thus a plot of H^{*-1} is given in Figure 5 (p. 61).

Reference may be made to that section for an explicit form for the estimator and its efficiency.

When μ is known, the density function

$$h^*(u) = \frac{\left[\frac{1-2u}{u(1-u)} - 2\log\left(\frac{u}{1-u}\right)\right]^{2/3}}{\int_0^1 \left[\frac{1-2t}{t(1-t)} - 2\log\left(\frac{t}{1-t}\right)\right]^{2/3} dt} \qquad (3.2.38)$$

generates spacings that are asymptotically optimal for estimating σ. A tabulation of the corresponding c.d.f., H^*, is presented in Table 19. Values for H^{*-1} may be obtained by interpolation in this table for points in [0,.5] or by the use of (3.2.7) for points outside this interval. A graph of the function H^{*-1} is shown in Figure 7.

Gupta and Gnanadesikan have found optimal spacings for either two or three symmetric quantiles for the estimation of σ when μ is assumed known. A comparison of their spacings with the ones generated by (3.2.38) is presented in Table 20. Table 21 gives asymptotically optimal spacings and coefficients for k = 7,9.

For simultaneous estimation of μ and σ the optimal density is

$$h^*(u) = \frac{\left[12 + \frac{9}{3+\pi^2}\left(\frac{1-2u}{u(1-u)} - 2\log\left(\frac{u}{1-u}\right)\right)^2\right]^{1/3}}{\int_0^1 \left[12 + \frac{9}{3+\pi^2}\left(\frac{1-2t}{t(1-t)} - 2\log\left(\frac{t}{1-t}\right)\right)^2\right]^{1/3} dt}\,. \qquad (3.2.39)$$

A tabulation of the corresponding distribution function, H^*, is given in Table 22. Spacings may be obtained from this table through

Table 19. Logistic Distribution, μ Known; The Function $H^*(u)$

u	$H^*(u)$	u	$H^*(u)$
.01	.12276	.26	.42392
.02	.15956	.27	.42946
.03	.18612	.28	.43478
.04	.20774	.29	.4399
.05	.22632	.3	.4448
.06	.2428	.31	.44951
.07	.25771	.32	.45401
.08	.27139	.33	.45832
.09	.28408	.34	.46244
.1	.29593	.35	.46636
.11	.30706	.36	.47009
.12	.31758	.37	.47363
.13	.32755	.38	.47697
.14	.33703	.39	.48012
.15	.34606	.4	.48307
.16	.35469	.41	.48582
.17	.36295	.42	.48836
.18	.37086	.43	.4907
.19	.37845	.44	.49281
.20	.38574	.45	.4947
.21	.39274	.46	.49635
.22	.39947	.47	.49774
.23	.40595	.48	.49885
.24	.41217	.49	.49964
.25	.41816	.5	.5

Figure 7. Logistic Distribution, μ Known; The Function $H^{*-1}(u)$

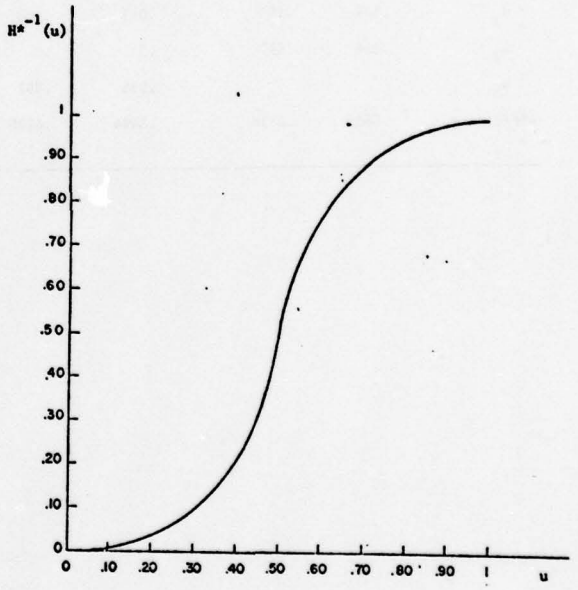


Table 20. Logistic Distribution, μ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 3	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.134	.103	.065	.103
u_2	.866	.897	.5	.5
u_3			.935	.897
Efficiency	.6686	.6838	.6494	.6838

Table 21. Logistic Distribution, μ Known; Asymptotically Optimal Spacings, Coefficients and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.011	-.035	.008	-.023
2	.065	-.136	.036	-.066
3	.185	-.26	.104	-.142
4	.5	.0	.221	-.214
5	.815	.26	.5	.0
6	.935	.136	.779	.214
7	.989	.035	.896	.142
8			.964	.066
9			.992	.023
Efficiency	.9016		.9364	

Table 22. Logistic Distribution, Both μ and σ Unknown; The Function H*(u)

u	H*(u)	u	H*(u)
.01	.10529	.26	.37907
.02	.1373	.27	.38499
.03	.16045	.28	.3908
.04	.17933	.29	.39649
.05	.1956	.3	.40207
.06	.21008	.31	.40755
.07	.22323	.32	.41294
.08	.23535	.33	.41825
.09	.24663	.34	.42347
.1	.25722	.35	.42861
.11	.26724	.36	.43368
.12	.27675	.37	.43869
.13	.28582	.38	.44364
.14	.2945	.39	.44853
.15	.30285	.4	.45337
.16	.31088	.41	.45816
.17	.31863	.42	.46291
.18	.32613	.43	.46763
.19	.3334	.44	.47231
.2	.34045	.45	.47697
.21	.34731	.46	.4816
.22	.35398	.47	.48622
.23	.36048	.48	.49082
.24	.36682	.49	.49541
.25	.37301	.5	.5

interpolation and the use of (3.2.7). The graph of H^{*-1} is shown in Figure 8.

Hassanein (1969b) has obtained spacings for simultaneous estimation of μ and σ that minimize the sum of the variances of the estimators. A comparison of these suboptimal spacings with spacings generated by (3.2.39) is provided by Table 23 for k = 2,7,9. Table 24 contains the asymptotically optimal spacings, coefficients and efficiencies for simultaneous estimation using either seven or nine order statistics.

3.2.6 Weibull Distribution

The Weibull c.d.f. is

$$F_o(x) = 1 - e^{-x^\gamma}, \quad x > 0, \tag{3.2.40}$$

where $\gamma > 0$ is a known shape parameter. The other functions that will be needed are:

$$f_o(x) = \gamma x^{\gamma-1} e^{-x^\gamma}, \tag{3.2.41}$$

$$Q(u) = [\log \frac{1}{1-u}]^{1/\gamma}, \tag{3.2.42}$$

and

$$f_o Q_o(u) = \gamma(1-u)[\log\left(\frac{1}{1-u}\right)]^{1-1/\gamma}. \tag{3.2.43}$$

The work of Harter and Moore (1967) may be used to deduce that

$$\langle f_o Q_o \cdot Q_o, f_o Q_o \cdot Q_o \rangle = \gamma^2. \tag{3.2.44}$$

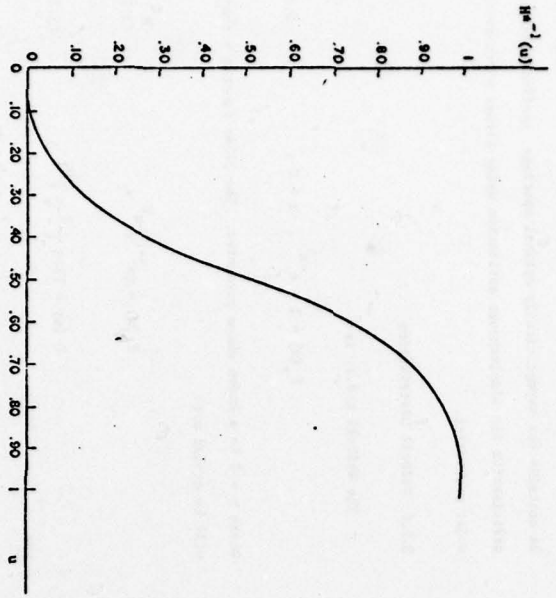


Figure 8. Logistic Distribution, Both μ and σ Unknown; The Function $H^{*-1}(u)$

Table 23. Logistic Distribution, Both μ and σ Unknown; A Comparison of Asymptotically Optimal and Suboptimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Suboptimal	Asymptotically Optimal	Suboptimal	Asymptotically Optimal	Suboptimal
u_1	.19	.238	.016	.039	.01	.023
u_2	.81	.762	.093	.152	.053	.096
u_3			.253	.316	.147	.211
u_4			.5	.5	.296	.351
u_5			.747	.684	.5	.5
u_6			.907	.848	.704	.649
u_7			.984	.961	.853	.789
u_8					.947	.094
u_9					.99	.977
Efficiency	.4422	.4142	.8651	.8605	.9095	.9051

Table 24. Logistic Distribution, μ and σ Unknown; Asymptotically Optimal Spacings and Coefficients for Seven or Nine Order Statistics

	k = 7			k = 9		
i	u_i	$S_\mu(i)$	$S_\sigma(i)$	u_i	$S_\mu(i)$	$S_\sigma(i)$
1	.016	.004	-.042	.001	.001	-.022
2	.093	.059	-.145	.053	.02	-.077
3	.253	.228	-.187	.147	.088	-.137
4	.5	.366	.0	.296	.212	-.131
5	.747	.228	.187	.5	.293	.0
6	.907	.059	.145	.704	.212	.131
7	.984	.004	.042	.853	.088	.137
8				.947	.02	.077
9				.99	.001	.022
Efficiency	.8651			.9095		

For estimating σ when μ is known, spacings should be taken according to the density

$$h^*(u) = \frac{1}{3}(1-u)^{-2/3} . \qquad (3.2.45)$$

Thus

$$H^{*-1}(u) = 1 - (1-u)^3 . \qquad (3.2.46)$$

As this is the same solution that was given for the exponential distribution, the properties of H^{*-1} (specifically its graph) can be found in Section 3.2.2. Asymptotically optimal spacings, coefficients, correction factors, and efficiencies for seven or nine order statistics are given in Tables 25-27 for $\gamma = 4/3, 2, 4$ respectively.

3.2.7 Extreme Value Distribution

The c.d.f. of the extreme value distribution is

$$F_o(x) = \exp\{-\exp\{-x\}\} , \quad -\infty < x < \infty , \qquad (3.2.47)$$

with corresponding p.d.f.

$$f_o(x) = \exp\{-(x + \exp\{-x\})\} . \qquad (3.2.48)$$

The other functions required for optimal density construction are:

$$Q_o(u) = -\log\log\left[\frac{1}{u}\right] , \qquad (3.2.49)$$

and

Table 25. Weibull Distribution, $\gamma = 1/3$, μ Known; Asymptotically Optimal Spacings, Coefficients, Correction Factors, and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.33	.238	.271	.186
2	.578	.211	.488	.177
3	.756	.165	.657	.152
4	.875	.151	.784	.122
5	.974	.045	.875	.091
6	.984	.023	.936	.063
7	.998	.008	.973	.037
8			.992	.019
9			.999	.002
C_σ		.841		.85
Efficiency	.958		.978	

Table 26. Weibull Distribution, $\gamma = 2$, μ Known; Asymptotically Optimal Spacings, Coefficients, Correction Factors and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.33	.189	.271	.139
2	.578	.204	.488	.16
3	.756	.18	.657	.155
4	.875	.181	.784	.136
5	.974	.062	.875	.11
6	.984	.032	.936	.081
7	.998	.012	.973	.051
8			.992	.028
9			.999	.004
C_σ		.861		.864
Efficiency	.958		.978	

Table 27. Weibull Distribution, $\gamma = 4$, μ Known; Asymptotically Optimal Spacings, Coefficients Correction Factors and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.33	.15	.271	.105
2	.578	.196	.488	.145
3	.756	.197	.657	.158
4	.875	.217	.784	.151
5	.974	.086	.875	.132
6	.984	.046	.936	.104
7	.998	.019	.973	.071
8			.992	.042
9			.999	.006
C_σ		.912		.913
Efficiency	.958		.978	

$$f_o Q_o(u) = -u \log(u) \quad . \tag{3.2.50}$$

It can be shown that

$$\langle f_o Q_o, f_o Q_o \rangle = 1 \quad . \tag{3.2.51}$$

When σ is known, the optimal density for use in the estimation of μ is

$$h^*(u) = \frac{1}{3} u^{-2/3} \quad . \tag{3.2.52}$$

An explicit expression for H^{*-1} is readily obtained and is found to be of the form

$$H^{*-1}(u) = u^3 \quad . \tag{3.2.53}$$

A graph of this function appears in Figure 9.

The performance of spacings generated by (3.2.52) is compared with that of the optimal spacings found by Hassanein (1968) in Table 28 for k = 2,7,9. The asymptotically optimal spacings, coefficients, and correction factors for seven or nine order statistics are given in Table 29.

3.2.8 Gamma Distribution

The density function for the gamma distribution is

$$f_o(x) = \frac{1}{\Gamma(p)} e^{-x} x^{p-1} \quad , \quad 0 < x < \infty \quad , \tag{3.2.54}$$

Figure 9. Extreme Value Distribution, σ Known; The Function $H^{*-1}(u)$

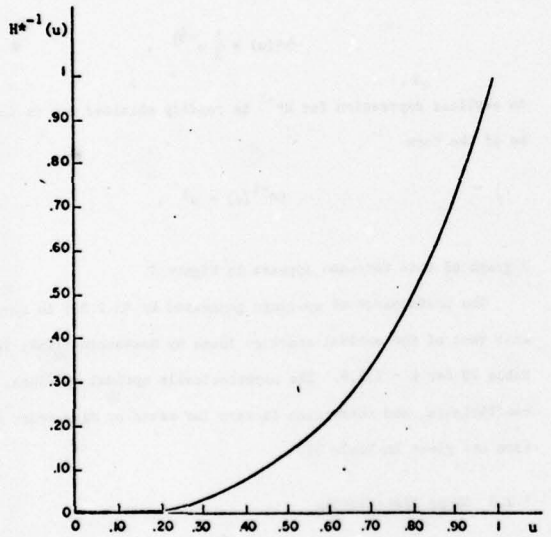


Table 28. Extreme Value Distribution, σ Known; A Comparison of Optimal and Asymptotically Optimal Spacings and Their Corresponding Efficiencies

	k = 2		k = 7		k = 9	
Spacing	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal	Asymptotically Optimal	Optimal
u_1	.037	.073	.002	.005	.001	.003
u_2	.296	.361	.016	.025	.008	.014
u_3			.053	.07	.027	.037
u_4			.125	.149	.064	.079
u_5			.244	.272	.125	.145
u_6			.422	.449	.216	.239
u_7			.67	.688	.343	.366
u_8					.512	.532
u_9					.729	.742
Efficiency	.8060	.8203	.968	.9693	.9791	.9798

Table 29. Extreme Value Distribution, σ Known; Asymptotically Optimal Spacings, Coefficients and Correction Factors for Seven or Nine Order Statistics

i	k = 7		k = 9	
	u_i	$W_\mu(i)$	u_i	$W_\mu(i)$
1	.002	.03	.001	.016
2	.016	.096	.008	.056
3	.053	.157	.027	.098
4	.125	.199	.064	.133
5	.244	.212	.125	.158
6	.422	.187	.216	.168
7	.67	.118	.343	.16
8			.512	.131
9			.729	.078
C_μ		-.444		-.437

where $p > 0$ is known. Denote by F_o the c.d.f. corresponding to f_o. As in the case of the normal distribution, it is not possible to derive an explicit formula for $Q_o(u) = F_o^{-1}(u)$ for all values of p. However, Q_o exists and hence its values and the values of the f_oQ_o function may be calculated through numerical procedure.

When μ is known, spacings taken according to the density

$$h^*(u) = \frac{Q_o(u)^{\frac{2(1-p)}{3}} \exp\{\tfrac{2}{3}Q_o(u)\}}{\int_0^1 Q_o(t)^{\frac{2(1-p)}{3}} \exp\{\tfrac{2}{3}Q(t)\}dt} \qquad (3.2.55)$$

will be asymptotically optimal for the estimation of σ. Spacings obtained using h^* have been computed by Särndal (1964) for k = 1(1)10 and p = 2,3,4,5. A graph of H^{*-1} for these same values of p is presented in Figure 10.

3.2.9 Lognormal Distribution

As in previous work, denote the standard normal p.d.f. and c.d.f. by ϕ and Φ respectively. For the lognormal distribution, the necessary functions for the construction of optimal densities are:

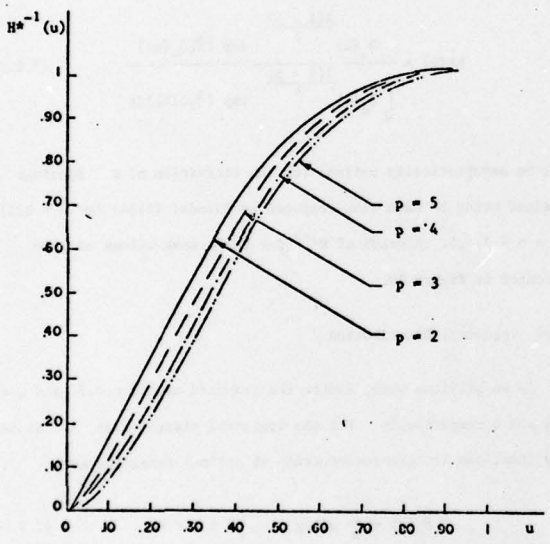
$$f_o(x) = \frac{1}{x}\phi(\log x), \quad 0 < x < \infty, \qquad (3.2.56)$$

$$Q_o(u) = \exp\{\Phi^{-1}(u)\}, \qquad (3.2.57)$$

and

$$f_oQ_o(u) = \phi\Phi^{-1}(u)\exp\{-\Phi^{-1}(u)\}. \qquad (3.2.58)$$

Figure 10. Gamma Distribution, μ Known; The Function $H^{*-1}(u)$



When μ is assumed known, the density

$$h^*(u) = \frac{[\phi\phi^{-1}(u)]^{-2/3}}{\int_0^1 [\phi\phi^{-1}(t)]^{-2/3}dt} \qquad (3.2.59)$$

generates spacings that are asymptotically optimal for the estimation of σ. This solution is identical to the one found for the estimation of the mean of the normal distribution when the standard deviation was known. Consequently,

$$H^{*-1}(u) = \phi(\sqrt{3}\,\phi^{-1}(u)) \quad . \qquad (3.2.60)$$

Asymptotically optimal spacings computed from (3.2.60) and the corresponding coefficients, correction factors and efficiencies for estimators based on seven or nine order statistics are given in Table 30.

3.2.10 Comparison of Solutions

The H^{*-1} functions provide one means of comparing different distributional forms. In Figure 11 the H^{*-1} functions for location parameter estimation are shown, when applicable, for the distributions considered in this chapter. A point of interest is that the logistic tends to behave more like the Cauchy than the normal in the case of spacings for location parameter estimation. A similar comparison for scale parameter estimation is shown in Figure 12.

Table 30. Lognormal Distribution, μ Known; Asymptotically Optimal Spacings, Coefficients, Correction Factors and Efficiencies for Seven or Nine Order Statistics

	k = 7		k = 9	
i	u_i	$W_\sigma(i)$	u_i	$W_\sigma(i)$
1	.023	.363	.013	.263
2	.121	.446	.074	.368
3	.29	.349	.184	.33
4	.5	.223	.334	.253
5	.71	.115	.5	.173
6	.879	.043	.666	.107
7	.977	.007	.816	.055
8			.926	.02
9			.987	.003
C_σ		1.546		1.571
Efficiency	.9637		.976	

Figure 11. The H^{*-1} Functions for Various Distributions in the Case that σ Is Known

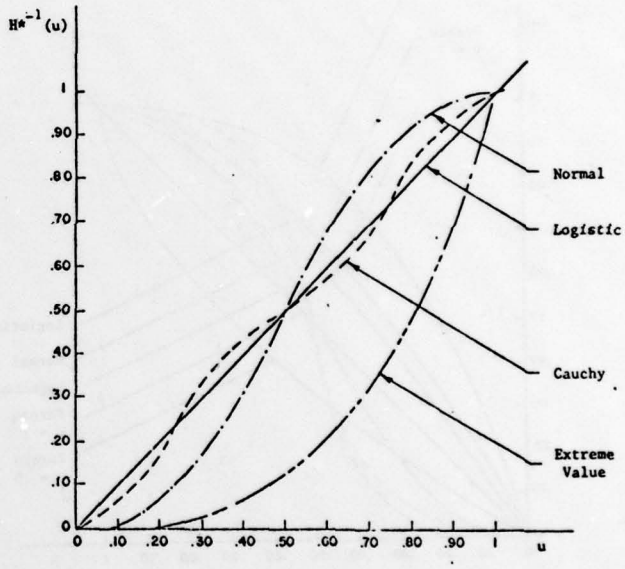
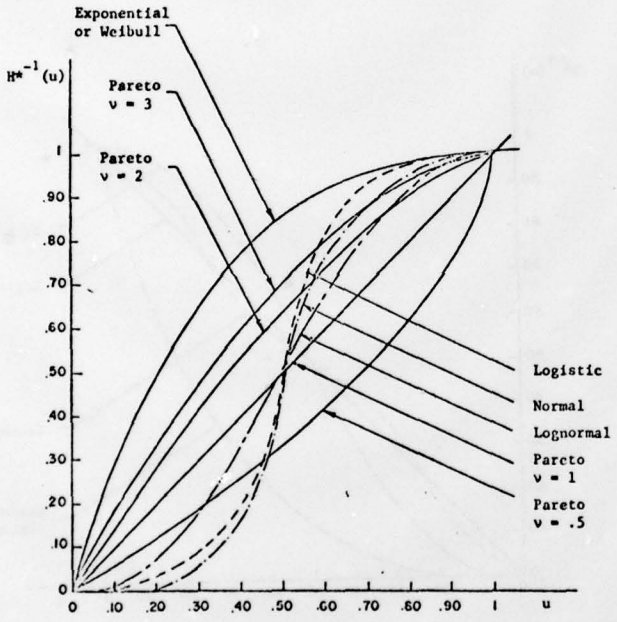


Figure 12. The H^{*-1} Functions for Various Distributions in the Case that μ Is Known



3.3 Data Summaries for Large Samples

An integral part of exploratory data analysis is the summarization of a data batch by a few select order statistics. These summaries are used to estimate location and scale parameters and/or to find re-expressions (transformations) of the data to other distributional forms (often normal). Several rules of thumb regarding the selection of order statistics for summary purposes have been proposed, such as using the median, quartiles and extremes, or taking the median, quartiles, $\frac{1}{8}$ percentiles and $\frac{1}{16}$ percentiles.

In Section 3.2 it was seen that the optimal placement of order statistics for location and scale parameter estimation depends heavily upon the assumptions made regarding the distributional type of the data. Thus for large data sets, where only a small portion of the sample is to be utilized for estimation purposes, the use of order statistics from nonparametric five or seven number data summaries may result in estimators with low efficiencies. A useful tool would be a summary technique that, perhaps after goodness of fit tests, could be adapted to the distributional form of the data. The objective of this section is to suggest such an adaptive technique for summarizing large data sets through the use of a few order statistics.

The specification of an adaptive data summary rule, in the present context, may be considered as consisting of two parts: (a) a rule of thumb regarding the selection of a set of order

statistics for a data summary and (b) the specification of optimal subsets of these order statistics for use in the estimation of μ or σ for various location and scale parameter models. In this section, adaptive rules based on 19 or 21 order statistics and considering subsets of size seven and nine respectively will be constructed using results from Section 3.2.

Upon review of the asymptotically optimal spacings for seven or nine order statistics in Section 3.2, it is seen that these spacings tend to cluster about certain points in the interval [0,1]. Since for $k = 7$, the spacings are the values $H^{*-1}\left(\frac{i}{8}\right)$, $i = 1, \ldots, 7$, it is no real surprise that the points where the spacings cluser are generally multiples of $\frac{1}{8}$. Similarly for $k = 9$, spacings tend to accumulate about multiples of $\frac{1}{10}$. The clustering behavior of the spacings may be used to justify the order statistics placement in the 19 and 21 number data summary rules that follow.

A 19 number adaptive data summary consists of the following sample quantiles:

1. The median, $\tilde{Q}(.5)$.
2. The $\frac{7}{16}$ percentile, $\tilde{Q}(.4375)$ and $\tilde{Q}(.5625)$.
3. The $\frac{3}{8}$ percentiles, $\tilde{Q}(.375)$ and $\tilde{Q}(.625)$.
4. The $\frac{5}{16}$ percentiles, $\tilde{Q}(.3125)$ and $\tilde{Q}(.6875)$.
5. The quartiles, $\tilde{Q}(.25)$ and $\tilde{Q}(.75)$.
6. The $\frac{3}{16}$ percentiles, $\tilde{Q}(.1875)$ and $\tilde{Q}(.8125)$.
7. The $\frac{1}{8}$ percentiles, $\tilde{Q}(.125)$ and $\tilde{Q}(.875)$.

8. The $\frac{1}{16}$ percentiles, $\tilde{Q}(.0625)$ and $\tilde{Q}(.9375)$.
9. The $\frac{2}{100}$ percentiles, $\tilde{Q}(.02)$ and $\tilde{Q}(.98)$.
10. The $\frac{1}{100}$ percentiles, $\tilde{Q}(.01)$ and $\tilde{Q}(.99)$.

For the distributions considered in Section 3.2, Tables 31-32 indicate which order statistics should be utilized in the estimation of μ or σ respectively by an estimator based on seven sample quantiles. The selection of an order statistic is indicated by a check mark across from its corresponding spacing. For example, for the exponential distribution with μ known Table 32 indicates that the spacings set {.3125, .5625, .75, .875, .9375, .98, .99} should be utilized for the estimation of σ.

For estimators composed of nine order statistics, a 21 number adaptive data summary may be defined as consisting of the following sample quantiles:

1. $\tilde{Q}(.5)$
2. $\tilde{Q}(.4)$, $\tilde{Q}(.6)$
3. $\tilde{Q}(.33)$, $\tilde{Q}(.66)$
4. $\tilde{Q}(.27)$, $\tilde{Q}(.73)$
5. $\tilde{Q}(.20)$, $\tilde{Q}(.80)$
6. $\tilde{Q}(.17)$, $\tilde{Q}(.83)$
7. $\tilde{Q}(.10)$, $\tilde{Q}(.90)$
8. $\tilde{Q}(.07)$, $\tilde{Q}(.93)$
9. $\tilde{Q}(.03)$, $\tilde{Q}(.97)$
10. $\tilde{Q}(.02)$, $\tilde{Q}(.98)$
11. $\tilde{Q}(.01)$, $\tilde{Q}(.99)$

Table 31. Order Statistic Selection for Location Parameter Estimation by Seven Order Statistics

Spacing	Distribution			
	Normal	Cauchy	Logistic	Extreme Value
.01				✓
.02	✓			✓
.0625				✓
.125	✓	✓	✓	✓
.875				
.25		✓	✓	✓
.3125	✓			
.375		✓	✓	
.4375				✓
.5	✓	✓	✓	
.5625				
.625		✓	✓	
.6875	✓			✓
.75		✓	✓	
.3125				
.875	✓	✓	✓	
.9375				
.98	✓			
.99				

Table 32. Order Statistic Selection for Scale Parameter Estimation by Seven Order Statistics

Spacing	Distribution						
	Exponential or Weibull	Pareto ν = .5	Pareto ν = 1	Pareto ν = 2	Pareto ν = 3	Logistic	Normal	Lognormal
.01						✓	✓	✓
.02							✓	
.0625		✓				✓		✓
.125			✓				✓	
.1875		✓		✓		✓		
.25		✓	✓		✓			✓
.3125	✓			✓				
.375		✓	✓		✓			
.4375				✓		✓	✓	✓
.5		✓	✓		✓			
.5625	✓			✓				✓
.625		✓	✓		✓			
.6875				✓		✓		✓
.75	✓		✓					
.8125				✓	✓	✓	✓	✓
.875	✓		✓	✓				
.9375	✓			✓	✓	✓	✓	✓
.98	✓						✓	
.99	✓						✓	

For the distributions considered in Section 3.2, Tables 33-34 indicate how order statistics should be selected for the estimation of μ or σ.

Table 35 contains the coefficients (denoted by W), correction factors (denoted by C) and efficiencies for the spacings sets suggested in Tables 31-34. To use this table, arrange the spacings under consideration in increasing order, $u_1 < u_2 < \ldots < u_k$, $k = 7,9$. Then the coefficient corresponding to u_i is W(i). For example, in the case of the exponential distribution with μ known, an explicit expression for the estimator of σ formed from 7 order statistics is

$$\hat{\sigma} = .2906\tilde{Q}(.3125) + .2252\tilde{Q}(.5625) + .1587\tilde{Q}(.75) + .0898\tilde{Q}(.875) + .055\tilde{Q}(.9375) + .0204\tilde{Q}(.98) + .0144\tilde{Q}(.99) - .854\mu \,. \qquad (3.3.1)$$

The efficiency of this estimator is .9653.

Table 33. Order Statistic Selection for Location Parameter Estimation by Nine Order Statistics

Spacing	Distribution			
	Normal	Cauchy	Logistic	Extreme Value
.01	✓			✓
.02				✓
.03				✓
.07	✓	✓		✓
.1			✓	✓
.17	✓	✓		
.2			✓	✓
.27				
.33	✓	✓	✓	✓
.4		✓	✓	
.5	✓	✓	✓	✓
.6		✓	✓	
.67	✓	✓	✓	
.73				✓
.8			✓	
.83	✓	✓		
.9			✓	
.93	✓	✓		
.97				
.98				
.99	✓			

Table 34. Order Statistic Selection for Scale Parameter Estimation by Nine Order Statistics

Spacing	Distribution							
	Exponential or Weibull	Pareto ν = .5	Pareto ν = 1	Pareto ν = 2	Pareto ν = 3	Logistic	Normal	Lognormal
.01						✓	✓	✓
.02							✓	
.03						✓		
.07		✓					✓	✓
.1		✓	✓			✓		
.17		✓		✓	✓		✓	✓
.2			✓			✓		
.27	✓	✓		✓				
.33		✓	✓		✓			✓
.4		✓	✓	✓				
.5	✓	✓	✓	✓	✓	✓	✓	✓
.6		✓	✓		✓			
.67	✓			✓				✓
.73		✓	✓	✓	✓			
.8	✓		✓		✓	✓		
.83				✓			✓	✓
.9	✓		✓	✓	✓	✓		
.93	✓				✓		✓	✓
.97	✓			✓		✓		
.98	✓				✓		✓	
.99	✓					✓	✓	✓

Table 35. Coefficients, Correction Factors and Efficiencies for the Summary Rule Spacings for Seven or Nine Order Statistics

k	W(1)	W(2)	W(3)	W(4)	W(5)	W(6)	W(7)	W(8)	W(9)	C	Efficiency
					a. Normal Distribution, σ Known						
7	.0464	.1517	.2026	.1986	.2026	.1517	.0464			.0	.9622
9	.0238	.0834	.1313	.1729	.1771	.1729	.1313	.0834	.0238	.0	.975
					b. Normal Distribution, μ Known						
7	-.0383	-.0691	-.2338	.0	.2338	.0691	.0383			.0	.8538
9	-.0358	-.0468	-.1131	-.1611	.0	.1611	.1131	.0468	.0358	.0	.9149
					c. Exponential Distribution, μ Known						
7	.2906	.2252	.1587	.0898	.055	.0204	.0144			.854	.9653
9	.2573	.2028	.1532	.1188	.0569	.0398	.0207	.0107	.0142	.8746	.9729
					d. Pareto Distribution, ν = .5, μ Known						
7	.5398	.3195	.1508	.1548	.0954	.0349	.0127			1.3079	.9838
9	.2873	.2448	.3018	.1752	.0987	.082	.0473	.0217	.0067	1.2656	.9903
					e. Pareto Distribution, ν = 1, μ Known						
7	.5833	.4286	.2976	.1905	.1071	.0477	.0119			1.6667	.9844
9	.4914	.4465	.2724	.1857	.1517	.1116	.0442	.0206	.0061	1.7303	.9889

Table 35. (continued)

k	W(1)	W(2)	W(3)	W(4)	W(5)	W(6)	W(7)	W(8)	W(9)	C	Efficiency
					f. Pareto Distribution, $\nu = 2$, μ Known						
7	.7698	.6708	.4717	.2893	.1947	.0684	.0264			2.4912	.9792
9	.6741	.5157	.4177	.4233	.2207	.1358	.0868	.0442	.0086	2.5269	.9852
					g. Pareto Distribution, $\nu = 3$, μ Known						
7	1.1211	.9445	.6031	.3995	.2934	.0779	.0154			3.4549	.9757
9	.9882	.8637	.5624	.4338	.2729	.208	.0848	.0515	.0163	3.4818	.984
					h. Cauchy Distribution, σ Known						
7	-.0518	.0	.3018	.5	.3018	.0	.0518			.0	.9496
9	-.0142	-.0483	.1038	.2601	.3972	.2601	.1038	.0483	.0142	.0	.9743
					i. Logistic Distribution, σ Known						
7	.0833	.1429	.1786	.1905	.1786	.1429	.8333			.0	.9844
9	.0546	.1116	.1341	.1238	.1517	.1238	.1341	.1116	.0542	.0	.9889
					j. Logistic Distribution, μ Known						
7	-.0334	-.1351	-.2644	.0	.2644	.1351	.0334			.0	.9009
9	-.0238	-.0557	-.1399	-.2138	.0	.2138	.1399	.0557	.0238	.0	.93

Table 35. (continued)

k	W(1)	W(2)	W(3)	W(4)	W(5)	W(6)	W(7)	W(8)	W(9)	C	Efficiency
					k. Weibull Distribution, $\gamma = 1/2$, Known						
7	.2274	.2147	.1722	.1078	.071	.0286	.021			.8427	.9653
9	.1927	.185	.1572	.1338	.0702	.0509	.0283	.0151	.0209	.8541	.9729
					l. Weibull Distribution, $\gamma = 2$, μ Known						
7	.1779	.2047	.1868	.1294	.0916	.0403	.0308			.8616	.9653
9	.1443	.1688	.1613	.1508	.0864	.0649	.0388	.0213	.0306	.8762	.9729
					m. Weibull Distribution, $\gamma = 4$, μ Known						
7	.1392	.1952	.2027	.1554	.1183	.0566	.0451			.9126	.9653
9	.1081	.154	.1655	.1698	.1065	.0829	.0531	.0299	.0448	.9147	.9729
					n. Extreme Value Distribution, σ Known						
7	.0661	.0796	.1526	.1866	.22	.1861	.1089			-.4314	.9653
9	.0656	.0421	.0726	.1059	.1311	.1913	.1699	.1405	.081	-.4251	.9729
					o. Lognormal Distribution, μ Known						
7	.3615	.4793	.3303	.1986	.1243	.048	.0059			1.548	.9622
9	.244	.3649	.341	.2684	.1771	.1113	.0506	.0191	.0023	1.5787	.975

4. SPACINGS FOR CENSORED SAMPLES AND QUANTILE ESTIMATION

In this chapter, techniques similar to those of Section 2.5 will be developed for spacing selection in censored samples. The selection of order statistics for the optimal estimation of population quantiles will also be considered.

4.1 Optimal Spacings for Censored Samples

Estimating location and scale parameters given a censored set of order statistics $X_{(np)}, \ldots, X_{(nq)}$ is most easily formulated as using the sample quantile function, $\tilde{Q}$, over the interval $[p,q] \subset [0,1]$. It can be shown (Parzen (1979)) that a model for location and scale parameter estimation in this case is

$$f_o Q_o(u)\tilde{Q}(u) = \mu f_o Q_o(u) + \sigma Q_o(u) f_o Q_o(u) + \sigma_B B(u) , \quad u \in [p,q] , \tag{4.1.1}$$

where $\{B(u), u \in [p,q]\}$ is a Brownian Bridge process on $[p,q]$ with covariance kernel

$$K_B(u_1,u_2) = \min(u_1,u_2) - u_1u_2 , \quad u_1,u_2 \in [p,q] , \tag{4.1.2}$$

and

$$\sigma_B = \frac{\sigma}{\sqrt{n}} . \tag{4.1.3}$$

The RKHS generated by K_B, H(K), consists of L^2 differentiable functions. For f and g in H(K) the inner product is

$$\langle f,g \rangle_{p,q} = \int_p^q f'(u)g'(u)du + \frac{f(p)g(p)}{p} + \frac{f(q)g(q)}{1-q} . \tag{4.1.4}$$

If $f \in H(K)$ is twice differentiable, the reproducing property and integration by parts can be used to show that f has the representation

$$f(u) = -\int_p^q f''(t)K_B(u,t)dt + K_B(u,p)[\frac{1}{p} f(p) - f'(p)] + K_B(u,q)[\frac{1}{1-q} f(q) + f'(q)]. \tag{4.1.5}$$

By making the identifications

$$C_1 = \frac{1}{p} f(p) - f'(p) ,$$
$$C_2 = \frac{1}{1-q} f(q) + f'(q) \tag{4.1.6}$$

Remark 2.3.5 of Section 2.3 is seen to be applicable. The next theorem which is the censored sample analogue of Theorem 2.5.1 follows immediately from this fact.

Theorem 4.1.1. Suppose the sample quantile function, $\tilde{Q}(u)$, is available over the interval $[p,q] \subset [0,1]$. Then the following results hold:

1. If $f_o Q_o$ has the representation (4.1.5) on $[p,q]$, define the density

$$h^*_{p,q}(u) = \frac{[f_o Q_o(u)'']^{2/3}}{\int_p^q [f_o Q_o(t)'']^{2/3} dt} \qquad (4.1.7)$$

with corresponding c.d.f. $H^*_{p,q}$. The spacings $H^{*-1}_{p,q}\left(\frac{i-1}{k-1}\right)$, $i = 1, \ldots, k$, are asymptotically optimal for estimating μ when σ is known.

2. If $Q_o \cdot f_o Q_o$ has the representation (4.1.5) on $[p,q]$, define the density

$$h^*_{p,q}(u) = \frac{([Q_o(u) f_o Q_o(u)]'')^{2/3}}{\int_p^q ([Q_o(t) f_o Q_o(t)]'')^{2/3} dt} \qquad (4.1.8)$$

with corresponding c.d.f. $H^*_{p,q}$. The spacings $H^{*-1}_{p,q}\left(\frac{i-1}{k-1}\right)$, $i = 1, \ldots, k$, are asymptotically optimal for estimating σ when μ is known.

3. If both $f_o Q_o$ and $Q_o \cdot f_o Q_o$ admit the representation (4.1.5) on $[p,q]$, define the density function

$$h^*_{p,q}(u) = \frac{[\psi'(u) A^{-1}_{p,q} \psi(u)]^{1/3}}{\int_p^q [\psi'(t) A^{-1}_{p,q} \psi(t)]^{1/3} dt} \qquad (4.1.9)$$

where

$$\psi'(u) = -([f_o Q_o(u)]'', [Q_o(u) f_o Q_o(u)]'') \qquad (4.1.10)$$

and

$$A_{p,q} = \begin{bmatrix} \langle f_o Q_o, f_o Q_o \rangle_{p,q} & \langle f_o Q_o, f_o Q_o \cdot Q_o \rangle_{p,q} \\ \langle f_o Q_o \cdot Q_o, f_o Q_o \rangle_{p,q} & \langle f_o Q_o \cdot Q_o, f_o Q_o \cdot Q_o \rangle_{p,q} \end{bmatrix}. \qquad (4.1.11)$$

Denote by $H^*_{p,q}$ the c.d.f. corresponding to $h^*_{p,q}$. The spacings $H^{*-1}_{p,q}\left(\frac{i-1}{k-1}\right)$, $i = 1, \ldots, k$, are asymptotically optimal for simultaneous estimation of μ and σ.

Theorem 4.1.1 provides a solution to the optimal spacings problem for censored samples. The corresponding formulas for the estimators of μ and σ based on asymptotically optimal spacings can be constructed by replacing h^* by $h^*_{p,q}$ and $H^{*-1}\left(\frac{i}{k+1}\right)$ by $H^{*-1}_{p,q}\left(\frac{i-1}{k-1}\right)$ in equations (2.5.20), (2.5.22), (2.5.24) and (2.5.25) of Section 2.5.

4.2 Optimal Spacings for Quantile Estimation

For the location and scale parameter model

$$F(x) = F_o\left(\frac{x-\mu}{\sigma}\right) \qquad (4.2.1)$$

the population quantile function, Q, has the form

$$Q(u) = \mu + \sigma Q_o(u) \qquad (4.2.2)$$

where $Q_o(u) = F_o^{-1}(u)$. This section will address the problem of how to optimally select order statistics for the estimation of the p^{th} population quantile, $Q(p)$.

First observe that since $Q_o(p)$ is known, $Q(p)$ is a linear combination of the unknown parameters μ and σ. Thus quantile estimation may be considered as a special case of the estimation of linear functions of the form $\ell_1\mu + \ell_2\sigma$.

For a given vector, $\ell' = (\ell_1,\ell_2)$, it is known from the theory of least squares (Graybill (1976)) that

$$\widehat{\ell_1\mu + \ell_2\sigma} = \ell_1\hat{\mu} + \ell_2\hat{\sigma} , \tag{4.2.3}$$

i.e., the ABLUE of a linear combination of the parameters is the same linear combination of the ABLUE's of μ and σ. Also note that

$$V(\ell_1\hat{\mu} + \ell_2\hat{\sigma}) = \frac{\sigma^2}{n}\,\ell'A^{-1}\ell \tag{4.2.4}$$

$$= \frac{\sigma^2}{n}\,\mathrm{tr}(A^{-1}\ell\ell') \tag{4.2.5}$$

where A is the information matrix of Section 5.2 and tr denotes the trace. Hence to minimize the variance of $\ell_1\hat{\mu} + \ell_2\hat{\sigma}$ it suffices to choose order statistics in such a manner that $\mathrm{tr}(A^{-1}\ell\ell')$ is a minimum. Sacks and Ylvisaker (1968) have derived an asymptotic solution to this problem that may be used to prove the next theorem.

Theorem 4.2.1. Let $\ell' = (\ell_1,\ell_2)$ be a known vector of constants and define $\psi(u)$ as in (4.1.9). Spacings generated by the density

$$h^*(u) = \frac{[\psi'(u)\ell\ell'\psi(u)]^{1/3}}{\int_0^1 [\psi'(t)\ell\ell'\psi(t)]^{1/3}dt} \tag{4.2.6}$$

will be asymptotically optimal for the estimation of $\ell_1\mu + \ell_2\sigma$.

The following corollary to Theorem 4.2.1 details an asymptotic solution to the problem of optimal spacing selection for quantile estimation.

Corollary 4.2.1. Let Q have the form (4.2.2) and let $p \in (0,1)$ be specified. Define the density function

$$h^*(u) = \frac{[\psi'(u)M\psi(u)]^{1/3}}{\int_0^1 [\psi'(t)M\psi(t)]^{1/3}dt} \tag{4.2.7}$$

with

$$M = \begin{bmatrix} 1 & Q_o(p) \\ Q_o(p) & Q_o^2(p) \end{bmatrix} \tag{4.2.8}$$

and $\psi(u)$ defined as in (4.1.9). The sequence of spacings sets generated by h^* is asymptotically optimal for the estimation of $Q(p)$.

5. CONCLUSION

5.1 Summary

A general approach to obtaining optimal spacings for linear systematic estimators of location and/or scale parameters has been formulated in this dissertation. By treating the problem of location and scale parameter estimation by linear functions of order statistics as one of regression analysis of a sample quantile process, it was found that the optimal spacings problem was equivalent to a regression design problem. This approach was seen to have advantages over classical techniques in that it provided a unified regression framework for optimal location and scale parameter estimation and led to computationally simple solutions to the optimal spacings problem.

The basic theory was developed in Chapter 2 where asymptotic results regarding designs for continuous parameter time series were employed to obtain spacings sets that were asymptotically optimal. This asymptotic optimality can be interpreted as meaning that the spacings sets result in nearly optimal efficiencies as the number of spacings included in the sets becomes large.

The theory developed in Chapter 2 was applied to several common distributional forms in Chapter 3. The asymptotically optimal spacings sets were seen to give nearly optimal efficiency for set size as small as seven or nine. Further, the propensity of these spacings

to cluster about certain values made it possible to propose some adaptive procedures for summarizing large data sets with a few order statistics.

In Chapter 4, an analogue of the asymptotic theory for optimal spacings selection in uncensored samples was developed for the case of censored samples. Asymptotically optimal spacings for population quantile estimation were also obtained.

5.2 Problems for Further Research

Several problems arise in the application of the theory developed in Chapter 2 due to the integral representation assumed for the f_oQ_o and $Q_o \cdot f_oQ_o$ functions. For this reason an approach to optimal spacings selection for functions that can only claim membership in the RKHS generated by K_B would be worthwhile.

There are several cases where either f_oQ_o'' or $(Q_o \cdot f_oQ_o)''$ behave so poorly at zero and/or one that they fail to be integrable on the closed interval [0,1]. An obvious procedure in this case would be to use an appropriate subinterval [p,q] of [0,1] and employ the results of Section 4.1 to obtain spacings. However, this approach seems to be quite sensitive to the choices for p and q. Thus techniques for selecting p and q in an optimal manner would be quite useful.

An extension of the results of Chapter 3 to other distributions and estimation situations is needed. Of particular interest is whether the placement of order statistics suggested in Section 3.3

will suffice for a still wider range of distribution types than those for which it was constructed.

REFERENCES

Bloch, D. (1966), "A Note on the Estimation of the Location Parameter of the Cauchy Distribution," *Journal of the American Statistical Association*, 61, 852-855.

Chernoff, H. (1971), "A Note on Optimal Spacings for Systematic Statistics," Technical Report No. 70, Department of Statistics, Stanford University.

Czorgo, M. and Revesz, P. (1978), "Strong Approximations of the Quantile Process," *Annals of Statistics*, 6, 882-894.

Eisenberger, I. and Posner, E. C. (1965), "Systematic Statistics Used for Data Compression in Space Telemetry," *Journal of the American Statistical Association*, 60, 97-133.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, North Scituate, Mass.: Duxbury Press.

Gupta, S. S. and Gnanadesikan, M. (1966), "Estimation of the Parameters of the Logistic Distribution," *Biometrika*, 53, 565-570.

Harter, H. L. and Moore, A. E. (1969), "Asymptotic Variances and Covariances of Maximum-Likelihood Estimators, from Censored Samples, of the Parameters of the Weibull and Gamma Populations," *Annals of Mathematical Statistics*, 38, 557-570.

Hassanein, K. M. (1968), "Analysis of Extreme Value Data by Sample Quantiles for Very Large Samples," *Journal of the American Statistical Association*, 63, 877-888.

_____ (1969a), "Estimation of the Parameters of the Extreme Value Distribution by Use of Two or Three Order Statistics," *Biometrika*, 56, 684-687.

_____ (1971), "Percentile Estimators for the Parameters of the Weibull Distribution," *Biometrika*, 58, 673-676.

_____ (1972), "Simultaneous Estimation of the Parameters of the Extreme Value Distribution by Sample Quantiles," *Technometrics*, 14, 63-70.

_____ (1977), "Simultaneous Estimation of the Location and Scale Parameter of the Gamma Distribution by Linear Functions of Order Statistics," *Scandinavian Actuarial Journal*, 60, 88-93.

Kulldorf, G. and Vännman, K. (1973), "Estimation of the Location and Scale Parameter of the Pareto Distribution by Linear Functions of Order Statistics," *Journal of the American Statistical Association*, 68, 218-227.

Mosteller, F. (1946), "On Some Useful Inefficient Statistics," *Annals of Mathematical Statistics*, 17, 175-213.

Ogawa, J. (1951), "Contributions to the Theory of Systematic Statistics, I," *Osaka Mathematical Journal*, 3, 131-142.

Parzen, E. (1961a), "An Approach to Time Series Analysis," *Annals of Mathematical Statistics*, 32, 951-989.

_____ (1961b), "Regression Analysis of Continuous Parameter Time Series," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, I, 469-489.

_____ (1979), "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association*, 74, 105-121.

Sacks, J. and Ylvisaker, D. (1966), "Designs for Regression Problems with Correlated Errors," *Annals of Mathematical Statistics*, 37, 66-89.

_____ (1968), "Designs for Regression Problems with Correlated Errors; Many Parameters," *Annals of Mathematical Statistics*, 39, 40-69.

Sarhan, A. E. and Greenberg, B. G. (1958), "Estimation Problems in Exponential Distribution Using Order Statistics," *Proceedings of the Statistical Techniques in Missile Evaluation Symposium*, Blacksburg, Va., 123-173.

Särndal, Carl-Erik (1962), *Information from Censored Samples*, Stockholm: Almqvist and Wiksell.

_____ (1964), "Estimation of the Parameters of the Gamma Distribution by Sample Quantiles," *Technometrics*, 6, 405-414.