Stochastic Musings

Perspectives from the Pioneers of the Late 20th Century

edited by John Panaretos

STOCHASTIC MUSINGS: PERSPECTIVES FROM THE PIONEERS OF THE LATE 20th CENTURY

This page intentionally left blank

STOCHASTIC MUSINGS: PERSPECTIVES FROM THE PIONEERS OF THE LATE 20th CENTURY

Edited by

John Panaretos
Athens University of Economics and Business

(A Volume in Celebration of the 13 Years of the Department of Statistics of the Athens University of Economics & Business in Honor of Professors C. Kevork & P. Tzortzopoulos)

Psychology Press
Taylor & Francis Group
NEW YORK AND HOVE

Camera ready copy for this book was provided by the author.

First published by
Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

This edition published 2013 by Psychology Press

711 Third Avenue	27 Church Road
New York	Hove
NY 10017	East Sussex, BN3 2FA

Psychology Press is an imprint of the Taylor & Francis Group, an informa business

First issued in paperback 2013

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Stochastic musings : perspectives from the pioneers of the late 20th century : a volume in celebration of the 13 years of the Department of Statistics of the Athens University of Economics & Business in honour of Professors C. Kevork & P. Tzortzopoulos / [compiled by] John Panaretos.

p. cm.

Includes bibliographical references and index.

1. Statistics. I. Kevork, Konst. El., 1928– . II. Tzortzopoulos, P. Th. III. Panaretos, John.

QA276.16 .S849 2003
310—dc21

2002040845
CIP

ISBN 978-0-415-65197-4

Contents

List of Contributors

Vic Barnett, Department of Mathematics, University of Nottingham, University Park, Nottingham NG7 2RD, England.
e-mail: vic.barnett@ntu.ac.uk

David Bartholomew, The Old Manse Stoke Ash, Suffolk, IP23 7EN, England.
e-mail: DJBartholomew@compuserve.com

David, R. Cox (Sir), Department of Statistics, Nuffield College, Oxford, OX1 1NF, United Kindom.
e-mail: david.cox@nuffield.oxford.ac.uk

Bradley Efron, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA.
e-mail: brad@stat.stanford.edu

Katharina Emrich, Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

David Freedman, Department of Statistics, University of California, Berkeley, Berkeley, CA 94720-4735, USA.
e-mail: freedman.census@stat.Berkley.EDU

Joe Gani, School of Mathematical Sciences, Australian National University, Canberra ACT 0200, Australia.
e-mail: gani@wintermute.anu.edu.au

Norman, L. Johnson, Department of Statistics, University of North Carolina, Phillips Hall, Chapel Hill, NC, 27599-3260, USA.
e-mail: btrice@stat.unc.edu

Dimitris Karlis, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: karlis@aueb.gr

Leslie Kish, The University of Michigan, USA. †

† Leslie Kish passed away on October 7, 2000.

Samuel Kotz, Department of Engineering Management and System Analysis, The George Washington University, 619 Kenbrook drive, Silver Spring, Maryland 20902, USA.
e-mail: kotz@seas.gwu.edu

Irini Moustaki, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: moustaki@aueb.gr

Saralees Nadarajah, Department of Mathematics, University of South Florida, Tampa, Florida 33620, USA.
e-mail: snadaraj@chumal.cas.usf.edu

John Panaretos, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: jpan@aueb.gr

Stelios Psarakis, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: psarakis@aueb.gr

Bhaskara M. Rao, Department of Statistics, North Dakota State University, 1301 North University, Fargo, North Dakota 58105, USA.
e-mail: MB_rao@ndsu.nodak.edu

Radhakrishna C. Rao, Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA, USA 16802-2111, USA.
e-mail: crr1@psu.edu

Damodar, N. Shanbhag, Statistics Division, Department of Mathematical Sciences, University of Sheffield, Sheffield S3 7RH, England.
e-mail: d.shanbhag@sheffield.ac.uk

Jef Teugels, Department of Mathematics, Katholieke Universiteit Leuven, Cerestijnenlaan 200B, B-3030 Leuven, Belgium.
e-mail: Jef.Teugels@wis.kuleuven.ac.be

Zoi Tsourti, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: tsourti@aueb.gr

Wolfgang Urfer, Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany.
e-mail: urfer@omega.statistik.uni-dortmund.de

Evdokia Xekalaki, Department of Statistics, Athens University of Economics & Business, 76 Patision St. 104 34, Athens, Greece.
e-mail: exek@aueb.gr

Vladimir M. Zolotarev, Steklov Mathematical Institute, Russian Academy of Sciences, Ulitza Vavilova 42, Moscow 333, Russia.
e-mail: zolot@orc.ru

This page intentionally left blank

PREFACE

This volume is published in celebration of the 13 years of existence of the Department of Statistics of the Athens University of Economics and Business (www.stat-athens.aueb.gr). The Department was –and still is– the only Department exclusively devoted to Statistics in Greece. The Department was set up in 1989, when the Athens School of Economics and Business was renamed as the Athens University of Economics and Business. Until then, Statistics was part of the Department of Statistics and Informatics. In its 13 years of existence the Department has grown to a center of Statistics in Greece, both applied and theoretical, with many international links. As part of the 13th anniversary celebration, it was decided to put together a volume with contributions from scientists of international calibre as well as from faculty members of the Department.

The goal of this volume is to bring together contributions by some of the leading scientists in probability and statistics of the latter part of the 20th century who are the pioneers in the respective fields. (David Cox writes on "Statistics and Econometrics", C. R. Rao (with M. B. Rao & D. N. Shanbhag) on "Convex Sets of Multivariate Distributions and Their Extreme Points", Bradley Efron on "the Future of Statistics", David Freedman on "Regression Association and Causation", Vic Barnett on "Sample Ordering for Effective Statistical Inference with Particular Reference to Environmental Issues", David Bartholomew on "A Unified Statistical Approach to Some Measurement Problems in the Social Sciences", Joe Gani on "Scanning a Lattice for a Particular Pattern", Leslie Kish on "New Paradigms (Models) for Probability Sampling" (his last paper), Samuel Kotz & Norman L. Johnson on "Limit Distributions of Uncorrelated but Dependent Distributions on the Unit Square", Jef Teugels on "The Lifespan of a Renewal", Wolfgang Urfer (with Katharina Emrich) on "Maximum Likelihood Estimates of Genetic Effects", and Vladimir M. Zolotarev on "Convergence Rate Estimates in Functional Limit Theorems". The volume also contains the contributions of faculty members of the Department. All the papers in this volume appear for the first time in the present form and have been refereed.

Academic and Professional Statisticians, Probabilists and students can benefit from reading this volume because they can find in it not only new developments in the area but also the reflections on the future directions of the discipline by some of the pioneers of the late 20th century. Scientists and students in other scientific areas related to Probability and Statistics, such as Biometry, Economics, Physics and Mathematics could also benefit for the same reason.

The volume is dedicated to professors Constantinos Kevork and Panagiotis Tzorzopoulos who were the first two professors of Statistics of the former Athens School of Economics and Business who joined the newly established Department in 1989. Professor Tzortzopoulos has also served as Rector of the University.

What relates the Department to this volume is that the international contributors, all of them renowned academics, are connected to the Department, one way or another. Some of them (e.g. L. Kish, D. R. Cox, C. R. Rao) have been awarded honorary doctorate degrees by the Department. They, as well as the rest of the contributors, have taught as distinguished visiting professors in the international graduate program of the Department.

I am indebted to all the authors, especially those from abroad, for kindly contributing to this volume but also for the help they have provided to the Department. Finally, I would like to thank Lawrence Erlbaum Publishers for kindly accepting to publish the volume and to make it as widely available as its reputation guarantees.

John Panaretos
Chairman of the Department
Athens, Greece

STOCHASTIC MUSINGS: PERSPECTIVES FROM THE PIONEERS OF THE LATE 20th CENTURY

This page intentionally left blank

1
SAMPLE ORDERING FOR EFFECTIVE STATISTICAL INFERENCE, WITH PARTICULAR REFERENCE TO ENVIRONMENTAL ISSUES

Vic Barnett
Department of Computing and Mathematics
Nottingham Trent University, UK

1. Introduction

The random sample is the fundamental basis of statistical inference. The idea of ordering the sample values and taking account both of value and order for any observation has a long tradition. While it might seem strange that this should add to our knowledge, the effects of ordering can be impressive in terms of what aspects of sample behaviour can be usefully employed and in terms of the effectiveness and efficiency of resulting inferences.

Thus, for any random sample $x_1, x_2, \ldots x_n$ of a random variable X, we have the *maximum* $x_{(n)}$ or *minimum* $x_{(1)}$ (the highest sea waves or heaviest frost), the *range* $x_{(n)} - x_{(1)}$ (how widespread are the temperatures that a bridge must withstand) or the *median* (as a robust measure of location) as examples using the ordered sample. The concept of an *outlier* as a representation of extreme, possibly anomalous, sample behaviour or of contamination, also depends on ordering the sample and has played an important role since the earliest days of statistical enquiry. Then again, linear combinations of all ordered sample values have been shown to provide efficient estimators, particularly of location parameters.

An interesting recent development has further enhanced the importance and value of sample ordering. With particularly wide application in environmental studies, it consists of setting up a sampling procedure specifically designed *to choose potential ordered sample values* at the outset– rather than taking a random sample and subsequently ordering it. An example of such an approach is *ranked set sampling* which has been shown to yield high efficiency inferences relative to random sampling. The basic approach is able to be readily and profitably extended beyond the earlier forms of ranked set sampling. We shall review the use of ordered data

- as natural expressions of sample information
- to reflect external influences
- to reflect atypical observations or contamination
- to estimate parameters in models

with some new thoughts on distribution-free outlier behavior, and a new estimator (the *memedian*) for the mean of a symmetric distribution.

2. Inference from the Ordered Sample

We start with the random sample $x_1, x_2 \ldots x_n$ of n observations of a random variable X describing some quantity of, say, environmental interest. If we arrange the sample in increasing order of value as $x_{(1)}, x_{(2)} \ldots x_{(n)}$ then these are observations of the *order statistics* $X_{(1)}, X_{(2)} \ldots X_{(n)}$ from a potential random sample of size n. Whereas the x_i ($i = 1, 2 \ldots n$) are *independent* observations, the order statistics $X_{(i)}, X_{(j)}$, ($i \neq j$) are *correlated*. This often makes them more difficult to handle in terms of distributional behaviour when we seek to draw inferences about X from the ordered sample. (See David, 1981, for a general treatment of ordering and order statistics).

At the descriptive level, the extremes $x_{(1)}$ and $x_{(n)}$, the range $x_{(n)} - x_{(1)}$, the mid-range$(x_{(1)} + x_{(n)})/2$ and the median m (that is, $x_{([n+1]/2)}$ if n is odd, or $(x_{(n/2)} + x_{([n+1]/2)})/2$ if n is even) have obvious appeal and interpretation. In particular the extremes and the median are frequently employed as basic descriptors in exploratory data analysis, and modified order-based constructs such as the *box and whisker plot* utilize the ordered sample as a succinct summary of a set of data (see Tukey, 1977, for discussion of such a non-model-based approach).

More formally, much effort has gone into examining the distributional behavior of the ordered sample values (again David, 1981, gives comprehensive cover). As an example, we have an exact form for the probability density function (pdf) of the range r as

$$g(r) = n(n\text{-}1) \int_{-\infty}^{\infty} \{F(x + r) - F(x)\}^{n-2} f(x + r) dF(x)$$

where $f(x)$ is the pdf of X (see Stuart and Ord , 1994, p.494).

But perhaps the most important and intriguing body of work on extremes is to be found in their *limit laws*. Rather like the *Central Limit Theorem* for a sample mean, which ensures convergence to normality from almost any distributional starting point, so we find that *whatever the distribution of X* (essentially), the quantities $x_{(1)}$ and $x_{(n)}$ tend as n increases to approach in distribution one of *only three possible forms*. The starting point for this work is long ago and is attributed by Lieblein (1954) to W. S. Chaplin in about 1860. David (1981, Chapter 9) gives a clear overview of developments and research continues apace to the present time (see, for example, Anderson, 1984; Gomes, 1994).

The three limits laws, are known as, and have distribution functions (df's) in the forms:

A: (Gumbel) $F_A(x) = \exp\{-\exp[-(x-\lambda)/\delta]\}$ $-\infty < x < \infty$ $(\delta > 0)$

B : (Frechet) $F_B(x) = \exp\{-[(x-\lambda)/\delta]^{-\alpha}\}$ $x > \lambda$ $(\delta > 0)$

C: (Weibull) $F_C(x) = \exp\{-[-(x-\lambda)/\delta]^{-\alpha}\}$ $x < \lambda$ $(\delta > 0)$

Which of these is approached by $X_{(n)}$ (and $X_{(1)}$ which is simply dual to $X_{(n)}$ on a change of sign) is determined by the notion of *zones of attraction*, although it is also affected by whether X is bounded below or above, or unbounded.

A key area of research is the rate of convergence to the limit laws as n increases – the question of the so-called *penultimate distributions*. How rapidly, and on what possible modelled basis, $X_{(n)}$ approaches a limit law L is of much potential interest. What, in particular, can we say of how the distributions of $X_{(n)}$ stand in relation to each other as n progresses from 40 to 100, 250 or 1000, say? Little, in fact, is known but such knowledge is worth seeking! We shall consider one example of why this is so in Section 3.

Consider the following random sample of 12 daily maximum wind speeds (in knots) from the data of a particular meteorological station in the UK a few years ago:

19, 14, 25, 10, 11, 22, 19, 17, 49, 23, 31, 18

We have $x_{(1)} = 10$, $x_{(n)} = x_{(12)} = 49$.

Not only is $x_{(n)}$ (obviously) the largest value - the *upper extreme* - but it seems extremely extreme! This is the stimulus behind the study of *outliers*: which are thought of as extreme observations which *by the extent of their extremeness* lead us to question whether they really have arisen from the same distribution as the rest of the data (i.e., from that of X). The alternative prospect, of course, is that the sample is *contaminated* by observations from some other source. An introductory study of the links between *extremes, outliers,* and *contaminants* is given by Barnett (1983) – Barnett and Lewis (1994) provide an encyclopaedic coverage of outlier concepts and methods, demonstrating the great breadth of interest and research the topic now engenders.

Contamination can, of course, take many forms. It may be just a reading or recording error – in which case *rejection* might be the only possibility (supported by a *test of discordancy*). Alternatively, it might reflect low-incidence mixing of X with another random variable Y whose source and manifestation are uninteresting. If so, a robust inference approach which draws inferences about the distribution of X while *accommodating* Y in an uninfluential way might be what is needed. Then again, the contaminants may reflect an exciting unanticipated prospect and we would be anxious to *identify* its origin and probabilistic characteristics if at all possible. *Accommodation, identification,* and *rejection* are three of the approaches to outlier study, which must be set in terms of (and made conditional on) some model F for the distribution of X. This is so whether we are examining univariate data, time

series, generalized linear model outcomes, multivariate observations, or whatever the base of our outlier interest within the rich field of methods now available.

But what of our extreme daily wind speed of 49 in the above data? We might expect the wind speeds to be reasonably modelled by an extreme value distribution – perhaps of type *B* (Frechet) or *A* (Gumbel), since they are themselves maxima over a 24-hour period. Barnett and Lewis (1994, Section 6.4.4) describe various statistics for examining an upper outlier in a sample from a Gumbel distribution. One particular test statistic takes the form of a Dixon statistic,

$$(x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(1)}).$$

For our wind-speed data with n=12, this takes the value $\frac{18}{39} = 0.46$ which according to Table XXV on page 507 of Barnett and Lewis (1994) is not significant. (The 5% point is 0.53, so notice how critical is the value of $t_{(n-1)}$ i.e., $t_{(11)}$. If instead of 31 it were 28, then x = 49 would have been a *discordant outlier* at the 5% level. This illustrates dramatically how some outlier tests are prone to 'masking': Barnett & Lewis, 1994, pp. 122-124.) Thus we conclude that although 49 seems highly extreme it is not extreme enough to suggest contamination (e.g., as a mis-reading or a mis-recording or due to freak circumstances).

A fourth use of ordered data is in regard to basic estimation of the parameters of the distribution F followed by X. Suppose X has df which takes the form $F[(x - \mu/\sigma]$ where μ reflects location and σ scale or variation. If X is symmetric, μ and σ are its mean and standard deviation. Nearly 50 years ago, Lloyd (1952) showed how to construct the BLUE or *best linear unbiased estimator* of μ and of σ based on the order statistics, by use of the Gauss-Markov theorem.

Suppose we write $U_{(i)} = (X_{(i)} - \mu)/\sigma$ $(i = 1, 2. . . .n)$ as the *reduced* (*standardised*) *order statistics* and let α and V denote the mean vector and variance covariance matrix of $\boldsymbol{U}$. Note that V is *not* diagonal since the $U_{(i)}$ and $U_{(j)}$ (for $i \neq j$) are correlated. Using the principle of extended least squares we obtain the BLUE θ^* of θ - where $\theta' = (\mu, \sigma)$ - in the form

$$\theta^* = (\mathbf{A}' \mathbf{V}^{-1} \mathbf{A})^{-1} \mathbf{A}' \mathbf{V}^{-1} \mathbf{x}$$

with variance/covariance matrix

$$\text{var}(\theta^*) = \sigma^2 (\mathbf{A}' \mathbf{V}^{-1} \mathbf{A})^{-1}$$

where $\mathbf{A} = (\mathbf{1}, \alpha)$ with $\alpha' = \{\alpha_i\} = \{E(U_{(i)})\} = \{E[(X_{(i)} - \mu)/\sigma]\}$ and $\mathbf{V} = \{\upsilon_{ij}\}$ is the variance/covariance matrix of the reduced order statistics $U_{(i)} = (X_{(i)} - \mu)/\sigma$.

This can be readily separated to yield the individual BLUE's, μ^* and σ^*. (See David, 1981, for broader discussion of optimal and sub-optimal estimators

based on order statistics and of how they compare with estimators based on the unordered sample.)

This approach is central to the more modern environmentally important principles of *ranked set sampling*, which we consider briefly in Section 3.

3. Possible New Routes for Outliers and for Order-Based Samples

Some of the principles reviewed suggest possible developments in outlier methodology on the one hand and in order-based estimation on the other.

3.1 A Distribution-Free Approach to Outliers

It is clear from the above outline, that the methodology of outliers depends crucially on the form of the null (no-contamination) model. Thus, for example, even a discordancy test of a single upper outlier $x_{(n)}$ based on the statistic $t = (x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(1)})$ is constrained in this way – since the null distribution of t (and its critical values) depends *vitally* on the form of F. The distribution of t and its percentage points will obviously be different if F is normal, exponential, Gumbel, etc. Yet we may not have any sound basis for assuming a particular form of F, especially if the only evidence is the single random sample in which we have observed an outlier. *This is the fundamental problem of outlier study.*

In practice, this dilemma is well-recognized and is usually resolved by a judicious mix of historical precedent, broad principle, association and wishful thinking (as in all areas of model-based statistics).

Thus it may be that a previous related study, and general scientific or structural features of the practical problem, link with formal statistical considerations (e.g., the Central Limit Law, characteristics of extremal processes) to support a form for F, such as a normal distribution or an exponential distribution. We then relate our inferences to appropriate null (no-outlier) distributions for that particular F.

But we are concerned, of course, in studying outliers which as *extremes* must have the *distributional behaviour of extremes*, which we have just seen to be essentially distributionally independent of the family F from which the sample has been chosen – in view of the limit laws. So in principle it seems that we might essentially *ignore F and examine outlier behavior in terms of properties of the extreme value distribution which is being approached by* $x_{(1)}$ *or* $x_{(n)}$ (or by some appropriate outlier function of them). This is an attractive prospect: a *distribution-free* and hence *highly robust alternative to the usual model-based methods*.

So what is the difficulty? Precisely the following. Although $x_{(n)}$ approaches *A, B* or *C*, we have to deal with *finite samples* and not enough is known in detail about *how quickly and in what manner* the forms *A* or *B* or *C* are approached as n progresses from, say 40 to 100 to 250, etc. The study of convergence to the limit laws and of 'penultimate distributions' is not yet

sufficiently refined for our purpose (See Gomes, 1994, for some of the latest developments).

To consolidate this point consider Table 1.1 which shows samples of maximum daily, 3-daily, weekly, fortnightly, and monthly wind speeds (in knots) at a specific location in the UK.

Daily																			
	36	46	13	18	34	19	23	15	18	14	28	10	31	28	22	40	20	23	28
3-Daily																			
	33	31	21	19	22	28	29	25	36	41	16	24	43	20	38	51	34	20	31
Weekly																			
	40	36	47	21	41	27	34	32	45	42	54	19	30	31	24	31	33	34	36
Fortnightly																			
	35	32	45	37	39	31	34	28	47	58	31	33	51	42	50	47	40	41	52
Monthly																			
	40	44	39	32	48	36	51	40	38	52	62	51	39	50	42	56	29	36	45

Table 1.1: Samples of maximum windspeeds (in knots) at a single UK location over days, 3-days, weeks, fortnights, and months.

Assuming (reasonably) that these approach the limit law A (they are all maxima) we would expect to find that plots of $\ln \ln[(n + 1)/i]$ against $x_{(i)}$ in each case, will yield approximately linear relationships. It is interesting to confirm from the data that this is indeed so. We will further see that we obtain the natural temporal ordering we would expect (reflected, particularly, in the implied differences in the values of, particularly, λ) in the approximating extreme value distribution in each case. Essentially the plots are parallel with intercepts *increasing with the lengths of the periods over which the maximum is taken.*

Davies (1998) also carried out an empirical study of limiting distributions of wind speeds (again from a single UK site) and fitted Weibull distributions to maxima over days, weeks, fortnights, months, and 2-month periods. Figure 1 (from Davies, 1998) shows the fitted distributions in which the time periods over which the maxima are taken increase monotonically as we move from the left-hand distribution to the right-hand one.

We need to know much more about how the distributions change with change in the maximizing period. It might be hoped that we can obtain a clearer understanding of how the limit distribution of an extreme is approached in any specific case as a function of sample size n and that such knowledge might eventually lead to an essentially new (largely) distribution-free outlier methodology.

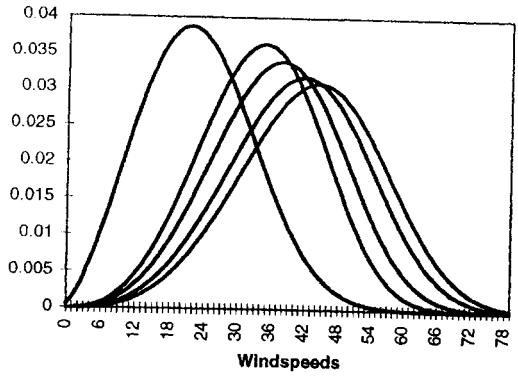

Figure 1.1: Fitted Weibull distributions to maximum windspeeds over different periods of time from 1 day to 2 months.

3.2 The Median and the Memedian

Ranked set sampling has become a valuable method particularly in environmental study. Barnett (2000) remarked:

"A method which is becoming widely used for sampling in the context of measuring expensive environmental risk factors is that of *ranked set sampling*. It can be used for the estimation of a mean, a measure of dispersion, quantiles or even in fitting regression models. The gains can be dramatic: efficiencies relative to simple random sampling may reach 300%.

The aim is to employ concomitant (and cheaply and readily available, sometimes subjective) information to seek to 'spread out' the sample values over their possible range. This can result in a dramatic increase in efficiency over simple random sampling. The method has been around for nearly 50 years since it was first mooted in an agricultural/environmental context (McIntyre, 1952). Further modifications continue apace to improve efficiency and applicability for different distributional forms of the underlying random variable and of the type of inference needed."

The method works as follows (Barnett, 2000):

"Conceptual random samples, of observations of the random variable X, take the form

$$\begin{array}{ccccc} x_{11} & \cdots & x_{21} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{22} & \cdots & x_{2n} \\ & & & & \\ x_{n1} & \cdots & x_{n2} & \cdots & x_{nn} \end{array}$$

From each subsample we take one measured observation $x_{i(i)}$: the ith

ordered value in the ith sample ($i = 1, 2, \ldots, n$). The ranked-set sample is then defined as $x_{1(1)}, x_{2(2)}, \ldots, x_{n(n)}$. In early applications, the mean μ of the underlying distribution was estimated by

$$\bar{\bar{x}} = \Sigma x_{i(i)}/n \tag{1}$$

which is the *ranked set sample mean*, which compares favorably with $\bar{x}$ (the mean of a random sample of size n ; not of size n^2, because measurement is assumed to be of overriding effort compared with ordering). We find that $\bar{\bar{x}}$ is unbiased and that (for $n > 2$) typically

$$\text{var}(\bar{\bar{x}}) < \text{var}(\bar{x})$$

often markedly so, for different sample sizes and distributions, if we have correctly ordered the potential observations in each conceptual subsample."

It will prove interesting to extend (1) to a more general form: that of an arbitrary linear combinations of the $x_{i(i)}$ terms. We consider estimators of the form

$$\mu^* = \Sigma \gamma_i x_{i(i)} \tag{2}$$

In the general case where X has df $F[(x - \mu)/\sigma]$, we just noted how to determine the BLUE of μ and σ from the ordered sample. For the ranked set sample $x_{1(1)}, x_{2(2)} \ldots, x_{n(n)}$ we have a simplification in that the variance covariance matrix is now *diagonal* (since $X_{i(i)}$, $X_{j(j)}$, are independent if $i \neq j$) and V in the development of Section 2 can be replaced with $W = \text{diag}(\upsilon_{ii}) = \text{diag}(\upsilon_i)$. So if we write the optimal estimators as

$$\mu^* = \sum_{j=1}^{n} \gamma_i x_{i(i)}, \qquad \sigma^* = \sum_{j=1}^{n} \eta_i x_{i(i)}$$

we have

$$\gamma_i = \frac{(1/\upsilon_i)\left[\sum_{j=1}^{n}(\alpha_j^{\,2}/\upsilon_j) - \alpha_i \sum_{j=1}^{n}(\alpha_j/\upsilon_j)\right]}{\Delta} \tag{3}$$

$$\eta_i = \frac{(1/\upsilon_i)\left[\alpha_i \sum_{j=1}^{n}(1/\upsilon_j) - \sum_{j=1}^{n}(\alpha_j/\upsilon_j)\right]}{\Delta} \tag{4}$$

where

$$\Delta = \sum\left(\frac{\alpha_j^2}{\upsilon_j}\right)\sum\left(\frac{1}{\upsilon_j}\right) - \left[\sum\left(\frac{\alpha_j}{\upsilon_j}\right)\right]^2$$

with

$$\left.\begin{aligned} \mathrm{var}(\mu^*) &= \sigma^2 \frac{\sum_{i=1}^{n}\left(\alpha_i^2/\upsilon_i\right)}{\Delta} \\ \mathrm{var}(\sigma^*) &= \sigma^2 \frac{\sum_{i=1}^{n}\left(1/\upsilon_i\right)}{\Delta} \end{aligned}\right\} \qquad (5)$$

The properties of these, and related estimators, are discussed by Barnett and Moore (1997), Sinha *et al.* (1996), Stokes (1995) and Barnett (2000). In particular μ^* is highly efficient in comparison with the random sample mean (from an unordered sample) and more efficient than the ranked set sample mean.

Modified schemes in which we take different numbers of observations of different $x_{i(i)}$ have been discussed in terms of sampling design and effect by Kaur *et al.* (1997) and Barnett (1999).

Suppose we consider an extreme version of such a differential choice of the $x_{i(i)}$: namely that from each conceptual sample we chose only the median m_i. So our sample is now the set of n values m_i ($i = 1, 2, \ldots n$). Rather than spreading out the sample – the original aim of ranked set sampling – we have now concentrated it into all the medians. Could this be sensible for estimating μ in a symmetric distribution where X has df $F[(x - \mu/\alpha)]$?

Suppose the median m has variance $v_{(m)}\sigma^2$. Then, if we define the *memedian* M to be the mean value of the medians:

$$M = \frac{1}{n}\sum_{1}^{n} m_i$$

its sampling variance will be $v_M = v_{(m)}\sigma^2 / n$ where $v_{(m)}\sigma^2$ is the variance of an individual sample median, obtained from the diagonal variance covariance matrix W. In comparison, we know that the ranked set sample mean, $\bar{\bar{x}}$, has variance $v_{\bar{\bar{x}}} = \sum_{1}^{n} v_{ii}\sigma^2 / n^2$ so that the relative efficiency of M and $\bar{\bar{x}}$ is $e_1 = \sum v_{ii} / (n v_{(m)})$. Clearly M will be more efficient than $\bar{\bar{x}}$ if $\bar{v}_{ii} \geq v_{(m)}$ which will be true if $v_{(m)} = min\{v_{ii}\}$. Can this happen? *We will see that it can.*

For illustrative purposes, we show in Table 1.2 the variances of standardized order statistics from samples of size 5 for four symmetric distributions which in standardized forms have pdf's as follows:

- *Normal* $\exp(-x^2/2)$
- *Uniform* 1
- *Triangular* $4x+2 \quad (-1/2<x<0)$ $\quad 2-4x \quad (0<x<1/2)$
- *Double exponential* $exp\{-|x|\}$

We see, not unsurprisingly, that the variances of the standardized order statistics are symmetric about that of the median, but *what is perhaps surprising* is that sometime the median has largest variance, sometimes the smallest. (Results for the triangular and double exponential distributions come from Sarhan, 1954).

Distribution	v_{ii}				
Normal	.4475	.3115	.2868	.3115	.4475
Uniform	.01984	.03175	.03571	.03175	.01984
Triangular	.1524	.1407	.1333	.1407	.1524
Double exponential	1.4703	.5025	.3512	.5025	1.4703

Table 1.2: Variances of standardized order statistics for samples of size 5 from symmetric distributions

So for the normal, triangular and double exponential distributions the *memedian* is more efficient than the ranked set sample mean. That this is not universally true is seen from the results for the uniform distribution. The values of the relative efficiency e_1 in the four cases are:

$$1.67, 0.78, 1.080, 2.45$$

showing major efficiency gains for the normal and double exponential distributions.

Is it even possible that M is more efficient than the ranked set BLUE, μ^*? The relative efficiency of M and μ^* is now $e_2 = n/[v_{(m)}\sum_{1}^{n}(1/v_{ii})]$.

Again, we can consider this for the four distributions in Table 1.2 illustrated for sample size $n = 5$. The values of e_2 are now

$$1.21, 0.73, 1.077, 1.74$$

so that again (for the same three cases) we conclude *rather surprisingly* that M can indeed be more efficient than μ^*. Further, it is much easier to calculate and we recall that it is also (in appropriate distributional circumstances) *a fortiori*