**HERBERT A. DAVID**

# ORDER STATISTICS

Second Edition

WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

# ORDER STATISTICS

SECOND EDITION

H. A. DAVID

*Iowa State University*

John & Sons, Inc.

New York · London · Sydney · Toronto

***Library of Congress Cataloging in Publication Data*:**

David, Herbert Aron.
Order statistics.

(Wiley series in probability and mathematical statistics) (A Wiley publication in applied statistics)
Bibliography: p.
Includes index.
1. Order statistics. I. Title.

QA278.7.D38 1980 519.5 80-16928
ISBN 0-471-02723-5

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Preface to the Second Edition

In the ten years since the first edition of this book there has been much activity relevant to the study of order statistics. This is reflected by an appreciable increase in the size of this volume. Nevertheless it has been possible to retain the outlook and the essential structure of the earlier account. The principal changes are as follows.

Sections have been added on order statistics for independent nonidentically distributed variates, on linear functions of order statistics (in finite samples), on concomitants of order statistics, and on testing for outliers from a regression model. In view of major developments the section on robust estimation has been greatly expanded. Important progress in the asymptotic theory of order statistics has resulted in the complete rewriting, with the help of Malay Ghosh, of the sections on the asymptotic joint distribution of quantiles and on the asymptotic distribution of linear functions of order statistics.

Many other changes and additions have also been made. Thus the number of references has risen from 700 to 1000, in spite of some deletions of entries in the first edition. Many possible references were deemed either insufficiently central to our presentation or adequately covered in other books. By way of comparison it may be noted that the first (and so far only) published volume of Harter's (1978b) annotated bibliography on order statistics contains 937 entries covering work prior to 1950.

I am indebted to P. G. Hall, P. C. Joshi, Gordon Simons, and especially Richard Savage for pointing out errors in the first edition. The present treatment of asymptotic theory has benefited from contributions by Ishay Weissman as well as Malay Ghosh. All the new material in this book has been read critically and constructively by H. N. Nagaraja. It is a pleasure to thank also Janice Peters for cheerfully given extensive secretarial help. In addition, I am grateful to Beatrice Shube of Wiley for her unfailing helpfulness and to the U.S. Army Research Office for longstanding support.

H. A. DAVID

*Ames, Iowa*
*July 1980*

# Preface

Order statistics make their appearance in many areas of statistical theory and practice. Recent years have seen a particularly rapid growth, as attested by the references at the end of this book. There is a growing recognition that the large body of theory, techniques, and applications involving order statistics deserves study on its own, rather than as a mere appendage to other fields, such as nonparametric methods. Some may decry this increased specialization, and indeed it is entirely appropriate that the most basic aspects of the subject be incorporated in general textbooks and courses, both theoretical and applied. On the other hand, there has been a clear trend in many universities toward the establishment of courses of lectures dealing more extensively with order statistics. I first gave a short course in 1955 at the University of Melbourne and have since then periodically offered longer courses at the Virginia Polytechnic Institute and especially at the University of North Carolina, where much of the present material has been tried out.

In this book an attempt is made to present the subject of order statistics in a manner combining features of a textbook and of a guide through the research literature. The writing is at an intermediate level, presupposing on the reader's part the usual basic background in statistical theory and applications. Some portions of the book, are, however, quite elementary, whereas others, particularly in Chapters 4 and 9, are rather more advanced. Exercises supplement the text and, in the manner of M. G. Kendall's books, usually lead the reader to the original sources.

A special word is needed to explain the relation of this book to the only other existing general account, also prepared in the Department of Biostatistics, University of North Carolina, namely, the multiauthored *Contributions to Order Statistics*, edited by A. E. Sarhan and B. G. Greenberg, which appeared in this Wiley Series in 1962. The present monograph is not meant to replace that earlier one, which is almost twice as long. In particular, the extensive sets of tables in *Contributions* will long retain their usefulness. The present work contains only a few tables needed to clarify the text but provides, as an appendix, an annotated guide to the massive output of tables scattered over numerous journals and books; such tables are essential for the ready use of many of the methods described. *Contributions* was not designed as a textbook and is, of course, no longer quite up to date. However, on a number of topics well developed by 1962 more extensive coverage will be found there than here. Duplication of all but the most fundamental material has been kept to a minimum.

In other respects also the size of this book has been kept down by deferring wherever feasible to available specialized monographs. Thus plans for the treatment of the role of order statistics in simultaneous statistical inference have largely been abandoned in view of R. G. Miller's very readable account in 1966.

The large number of references may strike some readers as too much of a good thing. Nevertheless the list is far from complete and is confined to direct, if often brief, citations. For articles dealing with such central topics as distribution theory and estimation I have aimed at reasonable completeness, after elimination of superseded work. Elsewhere the coverage is less comprehensive, especially where reference to more specialized bibliographies is possible. In adopting this procedure I have been aided by knowledge of H. L. Harter's plans for the publication of an extensive annotated bibliography of articles on order statistics.

It is a pleasure to acknowledge my long-standing indebtedness to H. O. Hartley, who first introduced me to the subject of order statistics with his characteristic enthusiasm and insight. I am also grateful to E. S. Pearson for his encouragement over the years. In writing this book I have had the warm support of B. G. Greenberg. My special thanks go to P. C. Joshi, who carefully read the entire manuscript and made many suggestions. Helpful comments were also provided by R. A. Bradley, J. L. Gastwirth, and P. K. Sen. Expert typing assistance and secretarial help were rendered by Mrs. Delores Gold and Mrs. Jean Scovil. The writing was supported throughout by the Army Research Office, Durham, North Carolina.

H. A. DAVID

*Chapel Hill, North Carolina*
*December 1969*

# Contents

# CHAPTER 1

# Introduction

## 1.1. THE SUBJECT OF ORDER STATISTICS

If the random variables $X_1, X_2, \ldots, X_n$ are arranged in ascending order of magnitude and then written as

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)},$$

we call $X_{(i)}$ the $i$th order statistic ($i = 1, 2, \ldots, n$). In much of this book the (unordered) $X_i$ are assumed to be statistically independent and identically distributed. The $X_{(i)}$ are then necessarily dependent because of the inequality relations among them. At times we shall relax the assumptions and consider nonidentically distributed $X_i$ as well as various patterns of dependence.

The subject of order statistics deals with the properties and applications of these ordered random variables and of functions involving them. Examples are the *extremes* $X_{(n)}$ and $X_{(1)}$, the *range* $W = X_{(n)} - X_{(1)}$, the *extreme deviate* (from the sample mean) $X_{(n)} - \bar{X}$, and, for a random sample from a normal $N(\mu, \sigma^2)$ distribution, the *studentized range* $W/S_\nu$, where $S_\nu$ is a root-mean-square estimator of $\sigma$ based on $\nu$ degrees of freedom. All these statistics have important applications. The extremes arise in the statistical study of floods and droughts, as well as in problems of breaking strength and fatigue failure. The range is well known to provide a quick estimator of $\sigma$ and has found particularly wide acceptance in the field of quality control. The extreme deviate is a basic tool in the detection of outliers, large values of $(X_{(n)} - \bar{X})/\sigma$ indicating the presence of an excessively large observation. In the same context, the studentized range is useful when outliers are not confined to one direction. However, it also supplies the basis of many quick tests in small samples and is of key importance for ranking "treatment" means in analysis of variance situations.

With the help of the Gauss-Markov theorem of least squares it is possible to use linear functions of order statistics quite systematically for the estimation of parameters of location and/or scale. This application is particularly useful when some of the observations in the sample have been "censored," since in that case standard methods of estimation tend to become laborious or otherwise unsatisfactory. Life tests provide an ideal illustration of the advantages of order statistics in censored data. Since such experiments may take a long time to complete, it is often desirable to stop after failure of the first $r$ out of $n$ (similar) items under test. The observations are the $r$ times to failure, which here, unlike in most situations, arrive already ordered for us by the method of experimentation; from them we can estimate the necessary parameters, such as the true mean life.

Other occurrences arise in the study of the reliability of systems. A system of $n$ components is called a *k-out-of-n system* if it functions if and only if at least $k$ components function. For components with independent life time distributions $P_1, P_2, \ldots, P_n$ the time to failure of the system is seen to be the $(n-k+1)$th order statistic from the set of underlying heterogeneous distributions $\{P_1, P_2, \ldots, P_n\}$. The special cases $k = n$ and $k = 1$ correspond respectively to series and parallel systems.

Computers have provided a major impetus for the study of order statistics. One reason is they have made it feasible to look at the same data in many different ways, thus calling for a body of versatile, often rather informal techniques commonly referred to as *data analysis* (cf. Tukey, 1962; Mosteller and Tukey, 1977). Are the data really in accord with ($a$) the assumed distribution and ($b$) the assumed model? Clues to ($a$) may be obtained from a plot of the ordered observations against some simple function of their ranks, preferably on probability paper appropriate for the distribution assumed. A straight-line fit in such a *probability plot* indicates that all is more or less well, whereas serious departures from a straight line may reveal the presence of outliers or other failures in the distributional assumption. Similarly, in answer to ($b$), one can in simple cases usefully plot the ordered *residuals* from the fitted model. Somewhat in the same spirit is the search for statistics and tests which, although not optimal under ideal (say normal-theory) conditions, perform well under a variety of circumstances likely to occur in practice. An elementary example of these *robust methods* is the use, in samples from symmetrical populations, of the *trimmed mean*, which is the average of the observations remaining after the most extreme $k$ ($k/n < \frac{1}{2}$) at each end have been removed. Loss of efficiency in the normal case may, for suitable choice of $k$, be compensated by lack of sensitivity to outliers or to other departures from an assumed distribution.

Finally, we may point to a rather narrower but truly space-age application. In large samples (e.g., of particle counts taken on a space craft) there are interesting possibilities for data compression (Eisenberger and Posner, 1965), since the sample may be replaced (on the space craft's computer) by enough order statistics to allow (on the ground) both satisfactory estimation of parameters and a test of the assumed underlying distributional form.

## 1.2. THE SCOPE AND LIMITS OF THIS BOOK

Although we shall be concerned with all of the topics sketched in the preceding section, and with many others as well, the field of order statistics impinges on so many different areas of statistics that some limitations in coverage have to be imposed. To start with, unlike Wilks (1948), we use "order statistics" in the narrower sense now widely accepted: we shall *not* deal with *rank-order statistics*, as exemplified by the Wilcoxon two-sample statistic, although these also require an ordering of the observations. The point of difference is that rank-order statistics involve the ranks of the ordered observations only, not their actual values, and consequently lead to nonparametric or distribution-free methods—at any rate for continuous random variables. On the other hand, the great majority of procedures based on order statistics depend on the form of the underlying population. The theory of order statistics is, however, useful in many nonparametric problems and also in an assessment of the non-null properties of rank tests, for example, by the power function.

Other restrictions in this book have more of an *ad hoc* character. Order statistics play an important supporting role in multiple comparisons and multiple decision procedures such as the ranking of treatment means. In view of the useful books by Miller (1966), updated in an annotated bibliography (Miller, 1977), and Gupta and Panchapakesan (1979), there seems little point in developing here the inference aspects of the subject, although the needed order-statistics theory is either given explicitly or obtainable by only minor extensions. However, some multiple decision procedures are considered in Chapter 8 on the treatment of outliers.

Much more could be said about asymptotic methods than we do in Chapter 9. In fact, the asymptotic theory of extremes and of related statistics has been developed at length in a book by Galambos (1978); on the more applied side Gumbel's (1958) account continues to be valuable. The asymptotic theory of central order statistics and of linear functions of order statistics has also been a very active research area in more recent years and would clearly justify an advanced monograph of its own. We have thought it best to confine ourselves to a detailed treatment of some of the most important results and to a summary of other developments.

The effective application of order-statistics techniques requires a great many tables. Inclusion even of only the most useful would greatly increase the bulk of this book. We therefore limit ourselves to a few tables needed for illustration; for the rest, we refer the reader to general books of tables, such as the two volumes by Pearson and Hartley (1970, 1972) and the collection of tables in Sarhan and Greenberg (1962). Extensive tables of many functions of interest have been prepared by Harter (1970a, b) in two large volumes devoted entirely to order statistics. Many references to tables in original papers are given throughout our text, and a guided commentary to available tables is provided in the Appendix.

## 1.3. NOTATION

Although this section may serve for reference, the reader is urged to look through it before proceeding further.

As far as is feasible, random variables, or *variates* for short, will be designated by uppercase letters, and their realizations (observations) by the corresponding lowercase letters. By order statistics we will mean either ordered variates or ordered observations. Thus:

| | | |
|---|---|---|
| $X_1, X_2, \ldots, X_n$ | unordered variates | |
| $x_1, x_2, \ldots, x_n$ | unordered observations | |
| $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ | ordered variates | order statistics |
| $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ | ordered observations | order statistics |
| $X_{1:n} \leqslant X_{2:n} \leqslant \cdots \leqslant X_{n:n}$ | ordered variates—extensive form | |

When the sample size $n$ needs to be emphasized, we use the more extensive form of notation, switching rather freely from the extensive to the brief form.

| | |
|---|---|
| $P(x) = \Pr\{X \leqslant x\}$ | cumulative distribution function (cdf) of $X$ |
| $P_n(x)$ | empirical distribution function |
| $p(x)$ | probability density function (pdf) for a continuous variate; probability function (pf) for a discrete variate |

| | |
|---|---|
| $F_r(x), F_{r:n}(x)$ | cdf of $X_{(r)}, X_{r:n}$ $\quad r = 1, 2, \ldots, n$ |
| $f_r(x), f_{r:n}(x)$ | pdf or pf of $X_{(r)}, X_{r:n}$ |
| $F_{rs}(x, y) = \Pr\{X_{(r)} \leqslant x, X_{(s)} \leqslant y\}$ | joint cdf of $X_{(r)}$ and $X_{(s)}$ $1 \leqslant r < s \leqslant n$ |
| $f_{rs}(x, y)$ | joint pdf or pf of $X_{(r)}$ and $X_{(s)}$ |
| $\xi_p$ | population quantile of order $p$, given by $P(\xi_p) = p$ or equivalently by $\xi_p = P^{-1}(p) = Q(p), 0 < p < 1$ |
| $\xi_{\frac{1}{2}}$ | population median |
| $X_{([np]+1)}$ | sample quantile of order $p$, where $[np]$ denotes the largest integer $\leqslant np$ |
| $X_{([n\lambda_i]+1)}$ | sample quantile of order $\lambda_i$, $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < 1$ |
| But the sample median is | |
| $X_{(\frac{1}{2}\overline{n+1})}$ | $n$ odd |
| $\frac{1}{2}(X_{(\frac{1}{2}n)} + X_{(\frac{1}{2}n+1)})$ | $n$ even |
| $W, W_n = X_{(n)} - X_{(1)}$ | (sample) range |
| $W_{(i)} = X_{(n+1-i)} - X_{(i)}$ | $i$th quasi-range ($W_{(1)} = W$) |
| $W_{rs} = X_{(s)} - X_{(r)}$ | |
| $\overline{W}, \overline{W}_{n,k}$ | mean of $k$ ranges of $n$ |
| ${}_jW$ | $W$ for $j$th sample |
| $Y_{[r]}, Y_{[r:n]}$ | concomitant of $X_{(r)}, X_{r:n}$ |
| $\mu = \mathcal{E}X, \sigma^2 = \text{var}\, X$ | mean, variance of $X$ |
| $\mu_X = \mathcal{E}X, \mu_Y = \mathcal{E}Y$ | means of $X$, $Y$ (bivariate case) |
| $\sigma_X^2 = \text{var}\, X, \sigma_Y^2 = \text{var}\, Y$ | variances of $X$, $Y$ |
| $\sigma_{XY} = \text{cov}(X, Y), \rho = \sigma_{XY}/\sigma_X\sigma_Y$ | covariance, correlation coefficient of $X$, $Y$ |
| $\mu_{r:n} = \mathcal{E}X_{r:n}$ | mean of $X_{r:n}$ |
| $\mu_{r:n}^{(k)} = \mathcal{E}(X_{r:n}^k)$ | $k$th raw moment of $X_{r:n}$ |
| $\mu_{rs:n} = \mathcal{E}(X_{r:n}X_{s:n})$ | |
| $\sigma_{r:n}^2 = \text{var}\, X_{r:n}$ | |
| $\sigma_{rs:n} = \text{cov}(X_{r:n}, X_{s:n})$ | |
| $Q(x) = P^{-1}(x)$ | inverse cdf |
| $p_r = r/(n+1), q_r = 1 - p_r$ | |
| $Q_r = Q(p_r), f_r = p(Q_r)$ | |
| $Q_r = dQ(p_r)/dp_r = 1/f_r$ | |

| | |
|---|---|
| $S_\nu$ | estimator of $\sigma$ based on $\nu$ DF; for a $N(\mu, \sigma^2)$ distribution $\nu S_\nu^2/\sigma^2 \frown \chi_\nu^2$. |
| $Q_{n,\nu} = W_n/S_\nu$ | studentized range ($W_n$, $S_\nu$ independent) |
| $S = [\sum(X_i - \bar{X})^2/(n-1)]^{\frac{1}{2}}$ | (internal) estimator of $\sigma$ |
| $S^{(P)} = \{[(n-1)S^2 + \nu S_\nu^2]/(n-1+\nu)\}^{1/2}$ | pooled estimator of $\sigma$ |
| ${}_jS$ | $S$ for $j$th sample |
| $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ $a>0, b>0$ | beta function |
| $I_p(a,b) = \int_0^p t^{a-1}(1-t)^{b-1}\,dt/B(a,b)$ | incomplete beta function (1.3.1) |
| $\beta(a,b)$ | beta variate $X$ having cdf $\Pr\{X \leqslant x\} = I_x(a,b)$ (1.3.2) |
| $\chi_\nu^2$ | chi-square variate with $\nu$ DF |
| $\phi(x) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2}$ $-\infty < x < \infty$ | unit normal pdf |
| $\Phi(x) = \int_{-\infty}^x \phi(t)\,dt$ | unit normal cdf |
| $N(\mu, \sigma^2)$ | normal variate, mean $\mu$, variance $\sigma^2$ |
| $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | multinormal variate, mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ |
| $n^{(k)} = n(n-1)\ldots(n-k+1)$ $k = 1, 2, \ldots, n$ | |
| $[x]$ | integral part of $x$, but $\mu_{[k]} = \mathcal{E}X^{(k)}$ |
| rv | random variable |
| pdf | probability density function |
| cdf | cumulative distribution function |
| iid | independent, identically distributed |
| c.f. | characteristic function |
| a.s. | almost surely |
| DF | degrees of freedom |
| ML | maximum likelihood |
| LS | least squares |
| BLUE | best linear unbiased estimator |
| UMVU | uniformly minimum variance unbiased |
| UMP | uniformly most powerful |
| H1, H2 | Harter (1970a, b)— *Order Statistics . . . 1, 2* |

| | |
|---|---|
| PH1, PH2 | Pearson and Hartley (1970, 1972)—*Biometrika Tables 1, 2* |
| SG | Sarhan and Greenberg (1962)—*Contributions to Order Statistics* |
| Ex. | exercise ("example" is written in full) |
| D | decimal (e.g., to 3D = to 3 decimal places) |
| S | significant (e.g., to 4S = to 4 significant figures) |
| A5.3 | appendix listing of tables relating to Section 5.3 |

## EXERCISES

1.1. Let $x_{r:n}(x_1, \ldots, x_n)$ denote the $r$th largest among the real numbers $x_1, \ldots, x_n$.
(*a*) Show that

$$x_{1:2}(x_1, x_2) = \min(x_1, x_2) = \tfrac{1}{2}(x_1 + x_2) - \tfrac{1}{2}|x_1 - x_2|,$$
$$x_{2:2}(x_1, x_2) = \max(x_1, x_2) = \tfrac{1}{2}(x_1 + x_2) + \tfrac{1}{2}|x_1 - x_2|.$$

(*b*) Show also that

$$x_{1:n}(x_1, \ldots, x_n) = x_{1:2}(x_{1:n-1}(x_1, \ldots, x_{n-1}), x_{1:n-1}(x_2, \ldots, x_n)),$$
$$x_{n:n}(x_1, \ldots, x_n) = x_{2:2}(x_{n-1:n-1}(x_1, \ldots, x_{n-1}), x_{n-1:n-1}(x_2, \ldots, x_n))$$

and that for $r = 2, 3, \ldots, n-1$

$$x_{r:n}(x_1, \ldots, x_n) = x_{1:n}(x_{r:n-1}(x_1, x_2, \ldots, x_{n-1}), \ldots, x_{r:n-1}(x_n, x_1, \ldots, x_{n-2})).$$

(Meyer, 1969)

1.2. For the real numbers $x_1, \ldots, x_n$ let $\max^{(\nu)}(x_1, \ldots, x_n)$ denote the $\nu$th largest ($\nu = 1, \ldots, n$). Also define

$$x_{n,\nu} = \max{}^{(\nu)}\left(x_1, x_1 + x_2, \ldots, \sum_{i=1}^{n} x_i\right)$$

and

$$x^*_{m,\nu} = \max{}^{(\nu)}\left(x_2, x_2 + x_3, \ldots, \sum_{i=2}^{m+1} x_i\right) \qquad m = 1, 2, \ldots; \nu = 1, \ldots, m.$$

Show that

(*a*) $x_{n,\nu} = x_1 + \max^{(\nu)}(0, x_2, \ldots, \sum_{i=2}^{n} x_i)$
(b) $x_{n,\nu} = x_1 + \max^{(\nu)}(0, x^*_{n-1,1}, \ldots, x^*_{n-1,\nu}) \qquad \nu = 1, \ldots, n-1$
(*c*) $x_{n,\nu} = \max^{(2)}(x_1, x_1 + x^*_{n-1,\nu-1}, x_1 + x^*_{n-1,\nu}) \qquad \nu = 2, \ldots, n-1; n \geqslant 3.$

(Pollaczek, 1975)

CHAPTER 2

# Basic Distribution Theory

## 2.1. DISTRIBUTION OF A SINGLE ORDER STATISTIC

We suppose that $X_1, X_2, \ldots, X_n$ are $n$ independent variates, each with cumulative distribution function (cdf) $P(x)$. Let $F_r(x)$ $(r = 1, 2, \ldots, n)$ denote the cdf of the $r$th order statistic $X_{(r)}$. Then the cdf of the largest order statistic $X_{(n)}$ is given by

$$\begin{aligned} F_n(x) &= \Pr\{X_{(n)} \leqslant x\} \\ &= \Pr\{\text{all } X_i \leqslant x\} = P^n(x). \end{aligned} \tag{2.1.1}$$

Likewise we have

$$\begin{aligned} F_1(x) &= \Pr\{X_{(1)} \leqslant x\} = 1 - \Pr\{X_{(1)} > x\} \\ &= 1 - \Pr\{\text{all } X_i > x\} = 1 - [1 - P(x)]^n. \end{aligned} \tag{2.1.2}$$

These are important special cases of the general result for $F_r(x)$:

$$\begin{aligned} F_r(x) &= \Pr\{X_{(r)} \leqslant x\} \\ &= \Pr\{\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x\} \\ &= \sum_{i=r}^{n} \binom{n}{i} P^i(x)[1 - P(x)]^{n-i} \end{aligned} \tag{2.1.3}$$

since the term in the summand is the binomial probability that *exactly* $i$ of $X_1, X_2, \ldots, X_n$ are less than or equal to $x$. We write (2.1.3) as

$$F_r(x) = E_{P(x)}(n, r) \tag{2.1.4}$$

and note that the $E$ function has been tabled extensively (e.g., Harvard Computation Laboratory, 1955, where the notation $E(n, r, P(x))$ is used). Alternatively, from the well-known relation between binomial sums and the incomplete beta function we have

$$F_r(x) = I_{P(x)}(r, n - r + 1), \tag{2.1.5}$$