INFERENCCE PROCEDURES
BASED ON ORDER STATISTICS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Jesse C. Frey, M.S.

* * * * *

The Ohio State University

2005

Dissertation Committee:

Prof. H. N. Nagaraja, Adviser
Prof. Steven N. MacEachern
Prof. Ömer Öztürk
Prof. Douglas A. Wolfe

Approved by

__________________________
Adviser
Graduate Program in
Statistics
ABSTRACT

In this dissertation, we develop several new inference procedures that are based on order statistics. Each procedure is motivated by a particular statistical problem. The first problem we consider is that of computing the probability that a fully-specified collection of independent random variables has a particular ordering. We derive an equal conditionals condition under which such probabilities can be computed exactly, and we also derive extrapolation algorithms that allow approximation and computation of such probabilities in more general settings. Romberg integration is one idea that is used. The second problem we address is that of producing optimal distribution-free confidence bands for a cumulative distribution function. We treat this problem both in the case of simple random sampling and in the more general case in which the sample consists of independent order statistics from the distribution of interest. The latter case includes ranked-set sampling. We propose a family of optimality criteria motivated by the idea that good confidence bands are narrow, and we develop theory that makes the identification and computation of optimal bands possible. The Brunn-Minkowski Inequality from the theory of convex bodies plays a key role in this work. The third problem we investigate is that of how best to take advantage of auxiliary information when estimating a population mean. We develop a general procedure, intentionally representative sampling, that is unbiased in the nonparametric sense, yet offers great flexibility for incorporating auxiliary
information. The final problem we consider is that of modeling imperfect judgment rankings in ranked-set sampling. We develop a new class of models so large that essentially any judgment rankings model is a limit of models in this class, and we propose an algorithm for selecting one-parameter families from the class.
ACKNOWLEDGMENTS

I thank my adviser, Dr. H. N. Nagaraja, for his advice, his encouragement, and his consistently positive outlook. I also thank the other members of my committee, Dr. Steven N. MacEachern, Dr. Ömer Öztürk, and Dr. Douglas A. Wolfe, for their advice both on this dissertation and on other work during the past four years.
VITA

July 11, 1975 ............................... Born - Greenville, SC

1996 ........................................B.S. Mathematics and Physics,
B.A. History,
Presbyterian College, Clinton, SC

2000 ........................................M.S. Mathematics,
University of North Carolina,
Chapel Hill, NC

PUBLICATIONS


FIELDS OF STUDY

Major Field: Statistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>Computational Algorithms</td>
<td>5</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>The Exact Algorithm</td>
<td>7</td>
</tr>
<tr>
<td>2.3</td>
<td>Extending the Algorithm to More General Cases</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>A Product Trapezoid Rule Approach</td>
<td>23</td>
</tr>
<tr>
<td>2.5</td>
<td>Some Applications of the General Algorithm</td>
<td>29</td>
</tr>
<tr>
<td>3.</td>
<td>Optimal Distribution-free Confidence Bands for a CDF</td>
<td>34</td>
</tr>
<tr>
<td>3.1</td>
<td>An Overview of the Literature</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>A Class of Distribution-free Confidence Bands</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>A Class of Optimality Criteria</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Shape and Power Comparisons for the Confidence Band Procedures</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Optimal Confidence Bands in the Ranked-set Sampling Case</td>
<td>72</td>
</tr>
<tr>
<td>3.6</td>
<td>A Comparison of RSS-based Confidence Bands</td>
<td>86</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Simultaneous 92.4% nonparametric confidence intervals for the deciles of an unknown continuous distribution when the sample size is ( n = 40 )</td>
<td>13</td>
</tr>
<tr>
<td>2.2 The triangular array of approximations obtained via Romberg integration</td>
<td>18</td>
</tr>
<tr>
<td>2.3 Relative errors in computing the probability ( P(X_1 \leq X_2 \leq X_3) ), where ( X_i \sim \text{Beta}(i, 4 - i), i = 1, \ldots, 3 ). The true probability is ( \frac{64}{105} ).</td>
<td>22</td>
</tr>
<tr>
<td>2.4 Relative errors in computing the probability ( P(X_1 \leq \cdots \leq X_{10}) ), where ( X_i \sim \text{Beta}(i, 11 - i), i = 1, \ldots, 10 ). The true probability is approximately 0.00836.</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Distribution of the ranks ( R_{[1]} ), ( R_{[2]} ), and ( R_{[3]} ) when ( m = 3 ), ( n_1 = n_2 = n_3 = 1 ), and the rankings are perfect.</td>
<td>30</td>
</tr>
<tr>
<td>2.6 Exact critical values (in scientific notation) and levels for the exact test. In cases where the levels 0.05 and 0.10 are not achieved (to four decimal places), the two bracketing values are given.</td>
<td>33</td>
</tr>
<tr>
<td>3.1 The values ( {p(t, i)} ) when the algorithm of Section 2.2 is applied to the example in Section 3.2.</td>
<td>48</td>
</tr>
<tr>
<td>3.2 The values ( {s_t(i, j)} ) when the algorithm of Section 2.2 is applied to the example in Section 3.2.</td>
<td>49</td>
</tr>
<tr>
<td>3.3 Values of the optimality criterion ( \sum_i (b_i - a_i) ) for 95% confidence bands when the set size is 3 and 1 cycle of observations is measured.</td>
<td>87</td>
</tr>
<tr>
<td>3.4 Values of the optimality criterion ( \sum_i (b_i - a_i) ) for 95% confidence bands when the set size is 2 and 2 cycles of observations are measured.</td>
<td>90</td>
</tr>
</tbody>
</table>
4.1 Distribution of the sample to be measured. ............................ 95

4.2 The groups identified by the rank algorithm when \( n = 10 \) and \( k = 3 \). 100

4.3 Simulated efficiencies (relative to SRS) for the IRS and RSS mean estimators. ................................................................. 102

4.4 Simulated efficiencies (relative to SRS) for the IRS and RSS mean estimators when diameter is used as a covariate for height. The value \( r \) is the set size used for RSS. .......................... 103
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 The weights $\theta_{i,j}$ for the case in which $n = 2$ and $M = 8$.</td>
<td>26</td>
</tr>
<tr>
<td>3.1 Six different 95% confidence bands when $n = 10$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.</td>
<td>73</td>
</tr>
<tr>
<td>3.2 Six different 95% confidence bands when $n = 20$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.</td>
<td>74</td>
</tr>
<tr>
<td>3.3 Six different 95% confidence bands when $n = 40$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.</td>
<td>75</td>
</tr>
<tr>
<td>3.4 Power curves for the 95% confidence bands when $n = 10$. The alternative is the $Beta(\eta, \eta)$ distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.</td>
<td>76</td>
</tr>
</tbody>
</table>
3.5 Power curves for the 95% confidence bands when \( n = 20 \). The alternative is the \( \text{Beta}(\eta, \eta) \) distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes. . . . . . 77

3.6 Power curves for the 95% confidence bands when \( n = 40 \). The alternative is the \( \text{Beta}(\eta, \eta) \) distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes. . . . . . 78

3.7 Optimal conditional 95% confidence bands based on RSS when the set size is 3 and 1 cycle of observations is measured. Each band is plotted against data taking on the expected values of the respective order statistics given the observed sequence of order statistics. Individual plots are labeled by the observed sequence of order statistics. . . . . . 88

3.8 Optimal conditional 95% confidence bands based on RSS when the set size is 2 and 2 cycles of observations are measured. Each band is plotted against data taking on the expected values of the respective order statistics given the observed sequence of order statistics. Individual plots are labeled by the observed sequence of order statistics. . . . . . 89

4.1 Plot of the variable of interest height (in feet) against the covariate diameter at chest height (in centimeters) for the longleaf pine data. . 104

5.1 Densities for the judgment order statistics for various values of the power \( \rho \) when \( k = 3, r = 4 \) and \( h_1(u) = u \). The underlying distribution is uniform on \([0,1]\), and the plotted lines are \( f_{[1]} \) and \( f_{[1]} + f_{[2]} \). . . . . 133

5.2 Densities for the judgment order statistics for various values of the power \( \rho \) when \( k = 3, r = 4 \), and \( h_2(u) = (1 - u)^{-1} \). The underlying distribution is uniform on \([0,1]\), and the plotted lines are \( f_{[1]} \) and \( f_{[1]} + f_{[2]} \). 134

5.3 Densities for the judgment order statistics for various values of the power \( \rho \) when \( k = 3, r = 4 \), and \( h_3(u) = -u^{-1} \). The underlying distribution is uniform on \([0,1]\), and the plotted lines are \( f_{[1]} \) and \( f_{[1]} + f_{[2]} \). 135
CHAPTER 1

INTRODUCTION

The use of order statistics in statistical inference is almost as old as statistical inference itself. The sample median, for example, has long been used to estimate the center of a distribution, and the interquartile range has a similar history of use for estimating the spread of a distribution. Inference procedures based on order statistics have played an important role in the development of nonparametric statistics, and they are also prominent in areas such as robust estimation and outlier detection. In this dissertation, we develop several new inference procedures that are based on order statistics and that use ideas associated with the theory of order statistics.

In Chapter 2, we develop two computational algorithms that involve order statistics and that have applications in distribution-free statistical inference. The first of these algorithms, although applicable more generally, is most useful because it allows the computation of rectangle probabilities for standard uniform order statistics. That is, it allows the computation of probabilities of the form

$$P(u_1 \leq U_1 \leq v_1, \ldots, u_n \leq U_n \leq v_n),$$

where $(u_1, \ldots, u_n) \in \mathbb{R}^n$, $(v_1, \ldots, v_n) \in \mathbb{R}^n$, and $(U_1, \ldots, U_n)$ is a vector of standard uniform order statistics from a simple random sample. The second of the algorithms
allows the computation of certain probabilities involving independent order statistics from a continuous parent distribution. In particular, it allows the computation of probabilities of the form

\[ P(u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq U_{k_n:m_n} \leq v_n, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n}), \]

where \( U_{k_1:m_1}, \ldots, U_{k_n:m_n} \) are independent and \( U_{k_i:m_i} \) has a \( \text{Beta}(k_i, m_i + 1 - k_i) \) distribution. Note that the density for the \( \text{Beta}(k, m + 1 - k) \) distribution is given by

\[ f(x) = \begin{cases} \frac{m!}{(k-1)!((m-k))!} x^{k-1}(1-x)^{m-k}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \]

Both of the algorithms use the idea of creating a finite, nonhomogeneous, discrete Markov chain such that the probability of interest can be computed in terms of state space probabilities at the final time step. The second algorithm also uses the extrapolation technique known as Romberg integration. A collection of \( R \) functions that can be used to implement the computational algorithms developed in this chapter is given in the appendix.

In Chapter 3, we consider the problem of creating optimal distribution-free confidence bands for a continuous cumulative distribution function (CDF), and we use the computational algorithms developed in Chapter 2 in doing so. We treat this problem both for the case in which the sample is a simple random sample and for the case in which the sample consists of independent order statistics from the distribution of interest. We propose a family of optimality criteria that is motivated by the idea that good confidence bands are narrow, and we develop theory that makes the identification and computation of optimal bands possible. We also address uniqueness issues. The Brunn-Minkowski Inequality from the theory of convex bodies plays an
important role in our theoretical work in the simple random sampling case, and the more general Prékopa-Leindler Inequality plays a similar role in the independent order statistics case. We show that both in terms of narrowness and in terms of power, our optimal bands compare favorably to other distribution-free confidence bands that have been proposed in the literature.

In Chapter 4, we describe a novel sampling scheme, intentionally representative sampling (IRS), that is designed to improve on schemes like simple random sampling and ranked-set sampling by offering greater flexibility and by allowing for full use of auxiliary information. This scheme avoids the problem of drawing a bad, unrepresentative sample by creating and then using a space of potential samples consisting entirely of good samples. We describe a method of implementing IRS that is based on ranking units according to a concomitant, and we demonstrate the good performance of IRS both theoretically and through simulations based on a real biological data set.

In Chapter 5, we consider the problem of modeling imperfect rankings in ranked-set sampling (RSS). We develop a broad new class of models for imperfect rankings by thinking of the process of drawing a sample of size $k$ as a two-stage process in which one first draws a sample of size $r > k$, then subsamples $k$ units from the $r$. We show that this class of models is so large that essentially any reasonable imperfect rankings model is a limit of models in the class. We also present specific examples of useful one-parameter families that are contained in the class. These one-parameter families can be constructed to have properties such as left-skewness or right-skewness that are believed appropriate for a particular setting.
In Chapter 6, we summarize what was accomplished in Chapters 2 to 5. We also describe some future work that could be carried out in areas related to those discussed in this dissertation.

Only a handful of abbreviations are used in this dissertation. These are CDF (cumulative distribution function), EDF (empirical distribution function), EC (equal conditionals), SRS (simple random sampling), RSS (ranked-set sampling), and IRS (intentionally representative sampling).
CHAPTER 2

COMPUTATIONAL ALGORITHMS

2.1 Introduction

In this chapter, we develop algorithms for solving certain computational problems that arise in statistical inference. Each of these problems involves computing a probability of the form $P(X_1 \leq \cdots \leq X_n)$, where $X_1, \ldots, X_n$ are fully-specified independent random variables. We first derive an algorithm that allows exact, direct computation of $P(X_1 \leq \cdots \leq X_n)$ whenever the random variables $X_1, \ldots, X_n$ satisfy a certain equal conditionals (EC) condition. One important situation in which this EC condition is met is that in which the random variables $X_1, \ldots, X_n$ are interval-truncated versions of the same parent distribution. This algorithm can thus be applied to compute probabilities $P(u_1 \leq U(1) \leq v_1, \ldots, u_n \leq U(n) \leq v_n)$, where $(u_1, \ldots, u_n) \in \mathbb{R}^n$, $(v_1, \ldots, v_n) \in \mathbb{R}^n$, and $(U(1), \ldots, U(n))$ is a vector of standard uniform order statistics. We next derive an extrapolation algorithm that builds upon the exact algorithm to allow computation of $P(X_1 \leq \cdots \leq X_n)$ when $X_1, \ldots, X_n$ are independently distributed as order statistics from a common parent distribution. Since Beta distributions with integer parameters arise as the distributions for order statistics from the standard uniform distribution, this algorithm allows the computation of probabilities of the form $P(U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n})$, where $U_{k_1:m_1}, \ldots, U_{k_n:m_n}$
are independently distributed and $U_{k_i,m_i}$ has a $Beta(k_i, m_i + 1 - k_i)$ distribution. This algorithm can also be extended to allow computation of probabilities of the form $P(u_1 \leq U_{k_1,m_1} \leq v_1, \ldots, u_n \leq U_{k_n,m_n} \leq v_n, U_{k_1,m_1} \leq \cdots \leq U_{k_n,m_n})$. As part of a search for ways to improve the speed of this extrapolation algorithm, we examine a second extrapolation algorithm that is based upon a method for integrating over a simplex that was developed by Lyness and Puri (1973). We then combine the best aspects of these two extrapolation algorithms into a hybrid algorithm that offers both the speed of the Lyness and Puri (1973) algorithm and the flexibility of the first algorithm. Each of the algorithms that we develop proceeds by creating a finite, nonhomogeneous, discrete Markov chain such that the probability of interest can be expressed in terms of state space probabilities at the final time step. A similar strategy was used by Lou (1996) to compute distributions for certain run statistics that can be used to test for independence within a sequence of binary random variables. A collection of R functions for implementing the best of the algorithms from this chapter is given in the appendix.

The algorithms that we develop here have numerous potential applications in order statistics and nonparametrics. For example, probabilities of the form $P(X_1 \leq \cdots \leq X_n)$, where $X_1, \ldots, X_n$ are interval-truncated standard uniform distributions, arise in computing coverage probabilities for distribution-free confidence bands for a CDF. Probabilities of the form $P(X_1 \leq \cdots \leq X_n)$, where $X_1, \ldots, X_n$ are distributed like independent order statistics from the same parent distribution, arise as the probabilities for particular collections of ranks when doing ranked-set sampling (RSS),
nomination sampling, or any other type of sampling that leads to independent, non-
identically distributed observations. Being able to compute these probabilities is thus
useful for making nonparametric inference based on such samples.

In Section 2.2, we define the EC condition, and we show that when the condition
is satisfied, the probability \( P(X_1 \leq \cdots \leq X_n) \) can be computed exactly. In Section
2.3, we explore the extent to which the algorithm from Section 2.2 fails when the
EC condition does not hold, and we describe an extrapolation scheme that allows
approximation of \( P(X_1 \leq \cdots \leq X_n) \) to a high level of accuracy when \( X_1, \ldots, X_n \) are
distributed like order statistics from the same parent distribution. In Section 2.4, we
build on the work of Lyness and Puri (1973) to create an alternate, faster method for
computing these probabilities. We then combine the best aspects of this algorithm
with those of the algorithm from Section 2.3 to obtain a hybrid method that is both
fast and usable for computing probabilities of the form \( P(u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq
U_{k_n:m_n} \leq v_n, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n}) \), where \( U_{k_1:m_1}, \ldots, U_{k_n:m_n} \) are independent and
each \( U_{k_i:m_i} \) has a \( \text{Beta}(k_i, m_i + 1 - k_i) \) distribution. In Section 2.5, we discuss some
applications of the extrapolation algorithm.

2.2 The Exact Algorithm

In this section, we develop an algorithm that allows exact computation of proba-
bilities \( P(X_1 \leq \cdots \leq X_n) \) for appropriate choices of the random variables \( X_1, \ldots, X_n \).
We begin by presenting this algorithm in the greatest possible generality. We then
show how the algorithm can be applied to the computation of rectangle probabil-
ities for standard uniform order statistics. Finally, we mention some uses of these
probabilities in statistical inference.
Suppose that $X_1, \ldots, X_n$ are continuous, independently distributed random variables with CDFs $F_1, \ldots, F_n$, respectively. We can compute the probability $P(X_1 \leq \cdots \leq X_n)$ using an exact algorithm provided that there exists a sequence of real numbers $a_1 < \cdots < a_K$, called break points, such that the random variables $X_1, \ldots, X_n$ and the sequence $a_1 < \cdots < a_K$ satisfy the following condition.

**Definition 2.1.** Given the sequence of break points $a_1 < \cdots < a_K$, define disjoint intervals $I_0 = (-\infty, a_1], I_1 = (a_1, a_2], \ldots, I_{K-1} = (a_{K-1}, a_K]$, and $I_K = (a_K, \infty)$. We say that the random variables $X_1, \ldots, X_n$ and the sequence $a_1 < \cdots < a_K$ satisfy the equal conditionals (EC) condition if for each $j \in \{0, 1, \ldots, K\}$, the conditional distributions $[X_i | X_i \in I_j]$ are the same for all $i$ for which $P(X_i \in I_j) > 0$.

In the case where $X_1, \ldots, X_n$ are identically distributed, the EC condition is satisfied by $X_1, \ldots, X_n$ and every possible choice of the sequence $a_1 < \cdots < a_K$. If $X_1, \ldots, X_n$ are truncated versions of some parent distribution, then the EC condition is satisfied for every sequence of break points $a_1 < \cdots < a_K$ such that all of the truncation points are among the $a_i$. A condition equivalent to the EC condition is the condition that each of the random variables $X_1, \ldots, X_n$ is distributed as a mixture of a fixed collection of distributions $G_0, \ldots, G_K$, where each $G_i$ has support in $I_i$.

Suppose that $X_1, \ldots, X_n$ and $a_1 < \cdots < a_K$ satisfy the EC condition. To show that we can then compute $P(X_1 \leq \cdots \leq X_n)$ exactly, we develop an alternative representation for the probability $P(X_1 \leq \cdots \leq X_n)$. First, using the fact that a probability is the expectation of an indicator function, we can write that

$$P(X_1 \leq \cdots \leq X_n) = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} I(x_1 \leq \cdots \leq x_n) \, dF_1(x_1) \cdots dF_n(x_n).$$

(2.1)
Since the disjoint intervals \( I_0, \ldots, I_K \) have the real line as their union, we can also decompose the full integral (2.1) into a sum
\[
\sum_{i_1=0}^{K} \cdots \sum_{i_n=0}^{K} \int_{x_1 \in I_{i_1}} \cdots \int_{x_n \in I_{i_n}} I(x_1 \leq \cdots \leq x_n) \, dF_1(x_1) \cdots dF_n(x_n) \tag{2.2}
\]
of simpler integrals. It is immediately apparent, moreover, that because of the indicator function, the summands on the right side of (2.2) can be nonzero only if \( i_1 \leq i_2 \leq \cdots \leq i_n \). If we define \( S = \{(i_1, \ldots, i_n) \in \{0, \ldots, K\}^n | i_1 \leq \cdots \leq i_n \} \), we can write \( P(X_1 \leq \cdots \leq X_n) \) as
\[
\sum_{(i_1, \ldots, i_n) \in S} \int_{x_1 \in I_{i_1}} \cdots \int_{x_n \in I_{i_n}} I(x_1 \leq \cdots \leq x_n) \, dF_1(x_1) \cdots dF_n(x_n). \tag{2.3}
\]
We now consider the computation of individual terms in the sum (2.3).

Suppose that the vector \((i_1, \ldots, i_n) \in S\) satisfies the strict inequality \( i_1 < i_2 < \cdots < i_n \). Then the indicator function in the expression
\[
\int_{x_1 \in I_{i_1}} \cdots \int_{x_n \in I_{i_n}} I(x_1 \leq \cdots \leq x_n) \, dF_1(x_1) \cdots dF_n(x_n) \tag{2.4}
\]
is always 1, meaning that we can write (2.4) as
\[
\int_{x_1 \in I_{i_1}} \cdots \int_{x_n \in I_{i_n}} dF_1(x_1) \cdots dF_n(x_n) = \left( \int_{x_1 \in I_{i_1}} dF_1(x_1) \right) \cdots \left( \int_{x_n \in I_{i_n}} dF_n(x_n) \right) = \prod_{r=1}^{n} (F_r(a_{i_r+1}) - F_r(a_{i_r})) = \prod_{r=1}^{n} p(r, i_r), \tag{2.5}
\]
where \( p(i, j) \equiv P(X_i \in I_j) = F_i(a_{j+1}) - F_i(a_j) \).

Now suppose that the vector \((i_1, \ldots, i_n) \in S\) includes repeated values. To be specific, suppose that the vector \((i_1, \ldots, i_n)\) includes \( s_i \geq 0 \) copies of each value \( i \), \( i = 0, \ldots, K \). Then, noting that (2.4) is just
\[
P(X_1 \in I_{i_1}, X_2 \in I_{i_2}, \ldots, X_n \in I_{i_n}, \text{ and } X_1 \leq \cdots \leq X_n),
\]
we can use conditioning to rewrite (2.4) as a product

\[ P(X_1 \in I_{i_1}, X_2 \in I_{i_2}, \ldots, X_n \in I_{i_n}) P(X_1 \leq \cdot \cdot \cdot \leq X_n | X_1 \in I_{i_1}, \ldots, X_n \in I_{i_n}). \]  

(2.6)

Using the independence of \( X_1, \ldots, X_n \), we can write the first factor in (2.6) as

\[ \prod_{r=1}^{n} p(r, i_r). \]

Then, using the assumption that \( X_1, \ldots, X_n \) and \( a_1 < \cdot \cdot \cdot < a_K \) satisfy the EC condition, we may write the second term as \( \left( \prod_{i=0}^{K} s_i! \right)^{-1} \). This follows from noting that for any \( j, i, \) and \( l, \)

\[ P(X_i \leq \cdot \cdot \cdot \leq X_{i+l-1} | X_j, \ldots, X_{i+l-1} \in I_j) = \frac{1}{l!}. \]

Thus, (2.4) can be written as

\[ \left( \prod_{r=1}^{n} p(r, i_r) \right) \left( \prod_{i=0}^{K} s_i! \right)^{-1}, \]

(2.7)

and we can write \( P(X_1 \leq \cdot \cdot \cdot \leq X_n) \) as

\[ \sum_{(i_1, \ldots, i_n) \in S} \left\{ \left( \prod_{r=1}^{n} p(r, i_r) \right) \left( \prod_{i=0}^{K} s_i! \right)^{-1} \right\}, \]

(2.8)

where \( s_i \) is the number of the indices \( i_1, \ldots, i_n \) that are equal to \( i \).

The formula (2.7) suggests a Markov property that makes it possible to sequentially compute the probability (2.4) corresponding to any vector \( (i_1, \ldots, i_n) \in S \). Starting with a value of 1, we think of moving from left to right through the indices \( i_1, \ldots, i_n \). If the \( t \)th index is \( i_t \), then we multiply our current value by \( p(t, i_t) \), in this way accounting for the first factor in (2.7). If the \( t \)th index is appearing for the \( j \)th time, then we divide by \( j \), in this way accounting for the second factor in (2.7). At the \( t \)th stage, the only information needed to carry out the computation is the index
\( i_t \) and the number of times that \( i_t \) appeared among the previous indices \( i_1, \ldots, i_{t-1} \). This sequential computation method suggests an algorithm for computing the full probability (2.8).

To compute (2.8) efficiently, we define an appropriate set of probabilities and compute them using a recursive scheme. For \( t = 1, \ldots, n, i = 0, \ldots, K, \) and \( j = 1, \ldots, t \), we define \( s_t(i, j) \) to be the probability of the event \( \{ X_1 \leq \cdots \leq X_t, X_{t-j} \notin I_i, \text{ and } X_{t-j+1}, \ldots, X_t \in I_i \} \). Then, stated in words, \( s_t(i, j) \) is the probability that the first \( t \) random variables are in increasing order and that exactly the last \( j \) of the first \( t \) random variables fall in the interval \( I_i \). The values \( \{ s_t(i, j) \} \) for \( t = 1 \) are easily computed as

\[
s_1(i, j) = \begin{cases} 
  p(1, i), & j = 1, \\
  0, & \text{otherwise}, 
\end{cases}
\]  

(2.9)

and a recursive scheme for computing the values \( \{ s_{t+1}(i, j) \} \) from the values \( \{ s_t(i, j) \} \) is given by

\[
s_{t+1}(i, j) = \begin{cases} 
  p(t+1, i) \sum_{r<i} \sum_{l=1}^t s_t(r, l), & j = 1, \\
  p(t+1, i)s_t(i, j-1)/j, & j > 1. 
\end{cases}
\]  

(2.10)

The probability \( P(X_1 \leq \cdots \leq X_n) \) is then obtained as

\[
P(X_1 \leq \cdots \leq X_n) = \sum_{i=0}^K \sum_{j=1}^n s_n(i, j). 
\]  

(2.11)

One important special case of this algorithm comes in computing rectangle probabilities for uniform order statistics. Suppose that \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\) are vectors in \([0, 1]^n\) that satisfy \( u_i < v_i \) for all \( i \). Suppose further that we wish to compute

\[
P(u_1 \leq U(1) \leq v_1, \ldots, u_n \leq U(n) \leq v_n), 
\]  

(2.12)

where \((U(1), \ldots, U(n))\) is a vector of standard uniform order statistics. The relationship between the density of the order statistics and the density of a simple random sample
$U_1, \ldots, U_n$ allows us to rewrite (2.12) as

$$n!P(u_1 \leq U_1 \leq v_1, \ldots, u_n \leq U_n \leq v_n, U_1 \leq \cdots \leq U_n).$$

If we then define random variables $V_1, \ldots, V_n$ by setting $V_i = [U_i| U_i \in (u_i, v_i)]$ for $i = 1, \ldots, n$, we can write (2.12) as

$$n!(v_1 - u_1) \cdots (v_n - u_n) P(V_1 \leq \cdots \leq V_n).$$

Since the $\{V_i\}$ are just interval-truncated versions of the standard uniform distribution, the EC condition is met, and the algorithm given by equations (2.9) to (2.11) can be applied. An R function for computing these rectangle probabilities for uniform order statistics is given in the appendix.

Rectangle probabilities for uniform order statistics show up as coverage probabilities for distribution-free confidence bands for a continuous CDF. That topic will be discussed in Section 3.2. They also show up when one wishes to compute simultaneous nonparametric confidence intervals for a finite set of quantiles of a continuous distribution. The idea behind constructing such sets of simultaneous confidence intervals is discussed by Breth (1980), but no concrete examples are presented. Suppose that $q_1, \ldots, q_k$ are the $\alpha_1, \ldots, \alpha_k$ quantiles of some continuous distribution. If we draw a simple random sample $X_1, \ldots, X_n$, then any interval $(X_{(r)}, X_{(s)})$ whose endpoints are order statistics gives a nonparametric confidence interval for a specified quantile. Suppose that for each $i$, $(X_{(r_i)}, X_{(s_i)})$ is a nonparametric confidence interval for $q_i$. Then the simultaneous coverage probability for the $k$ intervals is the probability

$$P(X_{(r_1)} \leq q_1 \leq X_{(s_1)}, \ldots, X_{(r_k)} \leq q_k \leq X_{(s_k)}). \tag{2.13}$$

By the probability integral transformation, this probability can be rewritten as

$$P(U_{(r_1)} \leq \alpha_1 \leq U_{(s_1)}, \ldots, U_{(r_k)} \leq \alpha_k \leq U_{(s_k)}), \tag{2.14}$$
Table 2.1: Simultaneous 92.4% nonparametric confidence intervals for the deciles of an unknown continuous distribution when the sample size is $n = 40$.

where $(U_1, \ldots, U_n)$ are standard uniform order statistics. By manipulating the inequalities in (2.14), the probability (2.13) can then be written in the form (2.12), where each of the bounds $u_1, \ldots, u_n, v_1, \ldots, v_n$ is either 0, 1, or one of the $\alpha_i$. As a concrete example, Table 2.1 gives a set of simultaneous 92.4% nonparametric confidence intervals for the deciles of a continuous distribution when the sample size is $n = 40$. The individual intervals were chosen both to have roughly the same marginal coverage probabilities and to be “best possible” nonparametric confidence intervals in the sense of Zielinski and Zielinski (2005). The advantages of balancing the marginal coverage probabilities in a simultaneous confidence statement have been stressed by Beran (1990) and others.

### 2.3 Extending the Algorithm to More General Cases

In this section, we first describe how the algorithm of Section 2.2 can be crudely extended to situations in which the random variables $X_1, \ldots, X_n$ do not satisfy the
EC condition. We then discuss a particular case, namely that in which each $X_i$ is a Beta distribution with integer parameters, where the probability $P(X_1 \leq \cdots \leq X_n)$ can be computed using Romberg integration. Finally, we extend the Romberg integration algorithm to the case in which each $X_i$ is an interval-truncated version of a Beta density with integer parameters. Probabilities of this latter form arise when constructing distribution-free confidence bands for a continuous CDF based on a ranked-set sample or some other sample that consists of independent order statistics from the distribution of interest.

Suppose as in Section 2.2 that $X_1, \ldots, X_n$ are continuous, independently distributed random variables with cumulative distribution functions $F_1, \ldots, F_n$, respectively. Let $a_1 < \cdots < a_K$ be an arbitrary sequence of real numbers, and define disjoint intervals $I_0 = (-\infty, a_1], I_1 = (a_1, a_2], \ldots, I_{K-1} = (a_{K-1}, a_K], \text{ and } I_K = (a_K, \infty)$ as in Section 2.2. Because the intervals $I_0, \ldots, I_K$ have the real line as their union, the argument from Section 2.2 holds exactly all the way to (2.3). Moreover, if the sequence of indices $(i_1, \ldots, i_n) \in S$ satisfies the strict inequality $i_1 < \cdots < i_n$, then the argument leading up to (2.5) also holds exactly. It is only when we consider a sequence of indices with repeated values that the argument from Section 2.2 fails. This observation suggests one crude way of approximating the desired probability $P(X_1 \leq \cdots \leq X_n)$.

Let $S_t$ be defined by $S_t = \{(i_1, \ldots, i_n) \in \{0, \ldots, K\}^n | i_1 < \cdots < i_n \}$. Then $S_t$ is the proper subset of $S$ consisting of all those sets of indices for which we can compute the probability (2.4) exactly even without the EC condition. Let $S_c = S \cap S_t'$, the
complement of \( S_t \) in \( S \). Then we have that
\[
P((i_1, \ldots, i_n) \in S_c) = P(\text{at least two consecutive indices are the same})
\]
\[
\leq \sum_{r=1}^{n-1} P(i_r \text{ and } i_{r+1} \text{ are the same})
\]
\[
= \sum_{r=1}^{n-1} \sum_{i=0}^{K} p(r, i)p(r + 1, i)
\]
\[
\leq \sum_{r=1}^{n-1} \max_{t,i} \{p(t, i)\} \sum_{i=0}^{K} p(r + 1, i)
\]
\[
= (n - 1) \max_{t,i} \{p(t, i)\},
\]
where \( p(i, j) \equiv P(X_i \in I_j) = F_i(a_{j+1}) - F_i(a_j) \) as before. Thus, if we choose a list of sequences \( \{A_r\} \), where \( A_r = (a_1^{(r)}, \ldots, a_K^{(r)}) \), such that
\[
\lim_{r \to \infty} \max_{t,i} \{p(t, i); A_r\} = 0,
\]
and we compute
\[
\sum_{(i_1,\ldots,i_n) \in S_t} \left\{ \left( \prod_{r=1}^{n} p(r, i_r) \right) \left( \prod_{i=0}^{K} s_i! \right)^{-1} \right\}
\]
(2.15)
for each of the sequences, the limit will be \( P(X_1 \leq \cdots \leq X_n) \). Modifying the recursive computation procedure described at the end of Section 2.2 in an obvious way, we can define \( s_t(i) \) to be the probability that \( X_1 \leq \cdots \leq X_t \), that \( X_t \in I_i \), and that each of \( X_1, \ldots, X_t \) lies in a distinct interval \( I_j \). The values \( \{s_t(i)\} \) for \( t = 1 \) are then given by
\[
s_1(i) = p(1, i),
\]
and we can compute the values \( \{s_{t+1}(i)\} \) from the values \( \{s_t(i)\} \) by setting
\[
s_{t+1}(i) = p(t + 1, i) \sum_{r=0}^{i-1} s_t(r).
\]
(2.16)
We then have that
\[
\sum_{(i_1, \ldots, i_n) \in S_t} \left\{ \left( \prod_{r=1}^{n} p(r, i_r) \right) \left( \prod_{i=0}^{K} s_i! \right)^{-1} \right\} = \sum_{i=n-1}^{K} s_n(i). \tag{2.17}
\]

Approximating the probability \( P(X_1 \leq \cdots \leq X_n) \) by choosing a list of sequences \( \{A_r\} \) and computing the approximations (2.17) as just described is not the best available procedure, however, because we know that for any particular sequence \( A_r \), the approximation (2.17) understates the true probability. One way to improve this approximation is simply to act as though the EC condition holds even when it does not. The algorithm is then exactly as given in equations (2.9) to (2.11), the only difference being that the final value is approximate rather than exact. While the assumption that the EC condition holds may be quite unrealistic when the intervals between the values \( a_1^{(r)} < \cdots < a_K^{(r)} \) are wide, it becomes more and more reasonable as the intervals narrow. Thus, it is clear that applying the algorithm from Section 2.2 to a list of sequences \( \{A_r\} \) such that \( \lim_{r \to \infty} \max_{t,i} \{p(t, i); A_r\} = 0 \) yields a sequence of approximations that tends to converge to the correct answer faster than does the corresponding sequence obtained from (2.17). However, even this faster algorithm is unnecessarily wasteful for many cases of interest.

In what follows, we develop an alternate method for making this computation in the case in which each \( X_i \) has a Beta distribution with integer parameters. This method is based on the exact algorithm given by equations (2.9) to (2.11), and it uses the idea of Romberg integration, which is described in Atkinson (1989). The central notion behind Romberg integration is that if we know the structure of the error terms involved in some numerical approximation scheme, then we can take linear combinations of specially chosen approximations to obtain a higher-order approximation that converges at a much faster rate.
For example, suppose that when we use $M$ intervals, the error $E_M$ in the approximation has the form

$$E_M = \frac{c_2}{M^2} + \frac{c_4}{M^4} + \frac{c_6}{M^6} + \ldots,$$

(2.18)

where $c_2, c_4, \ldots$, are unknown real constants. If $I_{M}^{(0)} = I_{\text{true}} + E_M$ and $I_{2M}^{(0)} = I_{\text{true}} + E_{2M}$ are approximations of $I_{\text{true}}$ based on $M$ and $2M$ intervals, respectively, then the linear combination $I_{2M}^{(1)} = \frac{4}{3}I_{2M}^{(0)} - \frac{1}{3}I_{M}^{(0)}$ has an error in which the leading term goes like $\frac{1}{M^4}$ rather than $\frac{1}{M^2}$. If the value $I_{4M}^{(0)}$ is also available, then we may compute $I_{4M}^{(1)} = \frac{4}{3}I_{4M}^{(0)} - \frac{1}{3}I_{2M}^{(0)}$ and move to a higher order of accuracy by setting $I_{4M}^{(2)} = \frac{16}{15}I_{4M}^{(1)} - \frac{1}{15}I_{2M}^{(1)}$. This last approximation $I_{4M}^{(2)}$ has an error in which the leading term goes as $\frac{1}{M^6}$. In general, we can compute an entire triangular array of approximations by applying the recursion

$$I_{2k}^{(j+1)} = \frac{4^{j+1}}{4^{j+1} - 1}I_{2k}^{(j)} - \frac{1}{4^{j+1} - 1}I_{k}^{(j)}$$

(2.19)

whenever two approximations $I_{2k}^{(j)}$ and $I_{k}^{(j)}$ are available. This triangular array of approximations would be as shown in Table 2.2. The best available approximation at a given stage is the highest-order approximation on the row of the array that corresponds to the largest value of $M$. Moreover, by looking at the difference between the ratio of the two most recent best approximations and 1, we can obtain a typically conservative estimate of the size of the relative error in the best available approximation.

It turns out that if $X_1, \ldots, X_n$ are all Beta distributions with integer parameters, and if we use a sequence of break points consisting of the equally-spaced points $\frac{1}{M} < \frac{2}{M} < \cdots < \frac{M-1}{M}$ from the interval $(0,1)$, then the error of the approximation obtained using the algorithm given by equations (2.9) to (2.11) has exactly the structure given
Table 2.2: The triangular array of approximations obtained via Romberg integration in equation (2.18). Though we do not yet have a rigorous proof of this fact for general \( n \), we can show that the result holds for \( n = 2 \). The proof relies on the following result, which was first noticed by the pioneering probabilist Jacob Bernoulli. This and other results involving the Bernoulli numbers are summarized in Sebah and Gourdon (2002).

**Theorem 2.1 (Bernoulli).** Given positive integers \( p \) and \( n \), define

\[
s_p(n) = \sum_{k=1}^{n-1} k^p.
\]

Then there exists a fixed sequence of real numbers \( \{ B_k \}_{k=0}^\infty \) such that

\[
s_p(n) = \sum_{k=0}^{p} \frac{B_k}{k!} \frac{p!}{(p+1-k)!} n^{p+1-k}
\]

for all \( p \) and \( n \). Moreover, the numbers \( \{ B_k \} \), which are called the Bernoulli numbers, have the property that among the \( B_k \) with \( k \) odd, only \( B_1 = -\frac{1}{2} \) is nonzero.

Using Theorem 2.1, we can prove the following result.

**Theorem 2.2.** Suppose that \( X_1 \) and \( X_2 \) are continuous random variables with support on \([0,1]\). Let \( F_1 \) and \( F_2 \) be their respective cumulative distribution functions, and
assume that $F_1$ and $F_2$ are each polynomials on $[0, 1]$. Let the sequence of break points be the evenly spaced points $\frac{1}{M} < \cdots < \frac{M-1}{M}$. Then the algorithm given by equations (2.9) to (2.11) leads to an error of the form (2.18).

**Proof of Theorem 2.2.** We first note that since the algorithm given by (2.9) to (2.11) is linear in $F_1$ and $F_2$, it suffices to show the result for the case when both $F_1$ and $F_2$ are monomials. Since $F_1(0) = F_2(0) = 0$, neither $F_1$ nor $F_2$ can have a constant term. Thus, we may assume without loss of generality that there exist positive integers $r$ and $s$ such that $F_1(x) = x^r$ and $F_2(x) = x^s$ for $x \in [0, 1]$. The values $\{s_1(i, 1)\}$ are given by

$$s_1(i, 1) = \left(\frac{i+1}{M}\right)^r - \left(\frac{i}{M}\right)^r, \quad i = 0, \ldots, M - 1.$$  

Then, since $p(2, i) = \left(\frac{i+1}{M}\right)^s - \left(\frac{i}{M}\right)^s$ for $i = 0, \ldots, M - 1$, we have that

$$s_2(i, 1) = p(2, i) \sum_{k=0}^{i-1} s_1(i, 1) = \left[\left(\frac{i+1}{M}\right)^s - \left(\frac{i}{M}\right)^s\right] \left(\frac{i}{M}\right)^r$$

and

$$s_2(i, 2) = \frac{1}{2} \left[\left(\frac{i+1}{M}\right)^s - \left(\frac{i}{M}\right)^s\right] \left[\left(\frac{i+1}{M}\right)^r - \left(\frac{i}{M}\right)^r\right].$$

Summing over all possible states at time $t = 2$, we have that the approximation is given by

$$\sum_{i=0}^{M-1} \frac{1}{2} \left[\left(\frac{i+1}{M}\right)^s - \left(\frac{i}{M}\right)^s\right] \left[\left(\frac{i+1}{M}\right)^r + \left(\frac{i}{M}\right)^r\right]. \quad (2.20)$$

We will have proved the result if we can show that (2.20) is a polynomial in $\frac{1}{M^2}$. 

19
We can rewrite (2.20) as

\[
\frac{1}{2M^{s+r}} \sum_{i=0}^{M-1} ((i+1)^s - i^s) ((i+1)^r + i^r)
\]

\[
= \frac{1}{2M^{s+r}} \sum_{i=0}^{M-1} (i^r(i+1)^s + (i+1)^{r+s} - (i+1)^r i^s - i^{r+s})
\]

\[
= \frac{1}{2M^{s+r}} \left[ \sum_{i=0}^{M-1} (i+1)^{r+s} - \sum_{i=0}^{M-1} i^{r+s} + \sum_{i=0}^{M-1} i^r(i+1)^s - \sum_{i=0}^{M-1} (i+1)^r i^s \right].
\]

Changing the summation variable in two of the sums, we obtain the expression

\[
\frac{1}{2M^{s+r}} \left[ \sum_{j=1}^{M} j^{r+s} - \sum_{i=1}^{M-1} i^{r+s} + \sum_{i=0}^{M-1} i^r(i+1)^s - \sum_{j=0}^{M} j^r(j-1)^s \right]
\]

\[
= \frac{1}{2M^{s+r}} \left[ M^{r+s} + \sum_{i=0}^{M-1} i^r [(i+1)^s - (i-1)^s] - M^r(M-1)^s \right]
\]

\[
= \frac{1}{2} + \frac{1}{2M^{s+r}} \left\{ - \sum_{k=0}^{s} M^r \binom{s}{k} M^{s-k}(1)^k + \sum_{i=0}^{M-1} i^r \left[ \sum_{k=0}^{s} \binom{s}{k} i^{s-k} (1-(1)^k) \right] \right\}.
\]

Then, using the sums \(s_p(n)\) defined in Theorem 2.1, we can rewrite (2.21) as

\[
\frac{1}{2} + \frac{1}{2M^{s+r}} \left\{ - \sum_{k=0}^{s} \binom{s}{k} M^{r+s-k}(1)^k + \sum_{k=0}^{s} \binom{s}{k} \sum_{i=0}^{M-1} i^{r+s-k} (1-(1)^k) \right\}
\]

\[
= \frac{1}{2} + \frac{1}{2M^{s+r}} \left\{ - \sum_{k=0}^{s} \binom{s}{k} M^{r+s-k}(1)^k + \sum_{k=0}^{s} \binom{s}{k} s_{r+s-k}(M)(1-(1)^k) \right\}.
\]

(2.22)

We now consider the terms in this expression that involve odd powers of \(\frac{1}{M}\).

Because of the factor \(\frac{1}{2M^{s+r}}\) outside the brackets in (2.22), we get a term involving an odd power of \(\frac{1}{M}\) whenever the sums in the brackets lead to a term of the form \(CM^{s+r-w}\) for \(w\) odd. The first sum yields such a term for every odd \(w\) less than \(s\), and that term is \(\binom{s}{w} M^{s+r-w}\). The second sum has a summand that takes the value 0 whenever the summation index \(k\) is even. Thus, we need consider only the odd values
of $k$ in that sum. But if $k$ is odd, then the only term of the form $CM^{r+s-w}$ for $w$ odd that occurs in the polynomial $s_{r+s-k}(M)$ is the term $-\frac{1}{2}M^{r+s-k}$. This follows from the fact that the only nonzero Bernoulli number with an odd index is $B_1 = -\frac{1}{2}$. This leads, when $k = w$, to a summand of the form $-\frac{1}{2}M^{r+s-w} \binom{s}{w} (2) = -\binom{s}{w} M^{r+s-w}$. Since this term is exactly the negative of the term arising from the first sum, the two terms cancel, and there are no odd powers of $\frac{1}{M}$ in the final expression.

We may thus approximate the probability $P(X_1 \leq \cdots \leq X_n)$ as follows. For a sequence of choices of $M$ that starts with some value $M_0$ ($M_0 = 1$ is a reasonable choice) and doubles at each step, we use the sequence of break points $\frac{1}{M} < \cdots < \frac{M-1}{M}$ to obtain the basic approximation $I_M^{(0)}$. After each new basic approximation is obtained, we apply the recursion (2.19) to compute all of the available higher-order approximations. When the estimate of the relative error falls below some prespecified tolerance level, we stop the process and use the most recent best approximation as our answer.

As an example, suppose that we wish to compute the probability $P(X_1 \leq X_2 \leq X_3)$, where $X_1, X_2,$ and $X_3$ are independently distributed as $Beta(1, 3), Beta(2, 2),$ and $Beta(3, 1)$, respectively. Direct computation of the integral (2.1) shows that this probability is equal to $\frac{64}{105}$. The relative errors in the approximations based on selected numbers of intervals and levels of extrapolation are given in Table 2.3. Note that by relative error we mean the ratio of the error to the true probability. The results in Table 2.3 show that once the number of intervals is sufficiently large, the error in the basic approximations is reduced by a factor of roughly 4 with each doubling of the number of intervals. For the leftmost column of higher-order approximations, the error is reduced by a factor of roughly 16 with each doubling of the number of
<table>
<thead>
<tr>
<th>Number of intervals</th>
<th>Basic approximations</th>
<th>Higher-order approximations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-1.63·10⁻¹</td>
<td>-2.32·10⁻³</td>
</tr>
<tr>
<td>6</td>
<td>-4.27·10⁻²</td>
<td>-1.43·10⁻⁴ 2.01·10⁻⁶</td>
</tr>
<tr>
<td>12</td>
<td>-1.08·10⁻²</td>
<td>-8.91·10⁻⁶ 3.14·10⁻⁸ -7.29·10⁻¹⁶</td>
</tr>
<tr>
<td>24</td>
<td>-2.70·10⁻³</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Relative errors in computing the probability $P(X_1 \leq X_2 \leq X_3)$, where $X_i \sim Beta(i, 4 - i), i = 1, \ldots, 3$. The true probability is $\frac{64}{105}$.

intervals, while for the next column, the corresponding factor is 64. Using a maximum of just 24 intervals, an approximation accurate to 15 significant digits was obtained.

As a more difficult example along the same lines, suppose that we wish to compute the probability $P(X_1 \leq \cdots \leq X_{10})$, where $X_i \sim Beta(i, 11 - i), i = 1, \ldots, 10$. The relative errors obtained for this computation are given in Table 2.4. Here the errors displayed in the table were obtained by comparing the respective basic and higher-order approximations to the best available approximation based on 5,120 intervals. This approximate value was roughly 0.00836. Table 2.4 shows that an approximation accurate to 9 significant digits was obtained using a maximum of just 160 intervals.

To see that these results for Beta distributions extend to any situation in which $X_1, \ldots, X_n$ are distributed like order statistics from the same parent distribution, it suffices to apply the probability-integral transformation. That result tells us that if $F$ is the parent distribution, then the random variables $F(X_1), F(X_2), \ldots, F(X_n)$ each have Beta distributions with integer parameters. Specifically, if $X_i$ is distributed like the $k$th order statistic from a sample of size $m$, then $F(X_i)$ has a $Beta(k, m + 1 - k)$ distribution.
Table 2.4: Relative errors in computing the probability $P(X_1 \leq \cdots \leq X_{10})$, where $X_i \sim Beta(i, 11 - i), i = 1, \ldots, 10$. The true probability is approximately 0.00836.

<table>
<thead>
<tr>
<th>Number of intervals</th>
<th>Basic approximations</th>
<th>Higher-order approximations</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$-3.41 \cdot 10^{-1}$</td>
<td>$-2.12 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>20</td>
<td>$-1.01 \cdot 10^{-1}$</td>
<td>$-1.51 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>40</td>
<td>$-2.64 \cdot 10^{-2}$</td>
<td>$-9.76 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>80</td>
<td>$-6.67 \cdot 10^{-3}$</td>
<td>$-3.23 \cdot 10^{-6}$</td>
</tr>
<tr>
<td>160</td>
<td>$-1.67 \cdot 10^{-3}$</td>
<td>$-6.15 \cdot 10^{-6}$</td>
</tr>
</tbody>
</table>

To compute probabilities of the form $P(u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq U_{k_n:m_n} \leq v_n, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n})$, we proceed in the following fashion. We first apply the algorithm from Section 2.2 using a sequence of break points $a_1 < \cdots < a_K$ that includes all of the points $\{u_i, v_i\}$. To move from the current sequence of break points to the next sequence, we double the number of intervals by adding to the original sequence the midpoints of each of the intervals $(0, a_1), (a_1, a_2), \ldots, (a_K, 1)$. It is possible to show rigorously that if the Romberg integration algorithm works for the case in which $u_1 = \cdots = u_n = 0$ and $v_1 = \cdots = v_n = 1$, then it will also work for this more general situation.

### 2.4 A Product Trapezoid Rule Approach

In this section, we describe an algorithm for computing integrals of the form

$$I = \int_{S_n} \prod_{i=1}^{n} f_{k_i:m_i}(x_i) dx_n \ldots dx_1,$$

where $S_n = \{(x_1, \ldots, x_n) | 0 \leq x_1 \leq \cdots \leq x_n \leq 1\}$ is an $n$-dimensional simplex and $f_{k,m}$ is the density for a $Beta(k, m + 1 - k)$ distribution. The initial part of this development is due to Lyness and Puri (1973), and it applies to computing any
integral of the form

\[ I = \int_{S_n} f(x_1, \ldots, x_n) dx_n \ldots dx_1, \]

where \( f(x_1, \ldots, x_n) \) is a polynomial. The latter part, which is new, applies specifically to integration of separable polynomials like the product \( f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{k_i,m_i}(x_i) \).

Since Beta densities are polynomials, this algorithm can be used to compute the same sorts of probabilities that we computed in Section 2.3.

Suppose that \( f \) is a continuous function defined on the interval \([a,b]\), and let \( M \) be a positive integer. This integer \( M \) determines the width \( h = 1/M \) for the intervals used in the trapezoid rule. Following Lyness and Puri (1973), we may approximate

\[ I = \int_{x=a}^{b} f(x) dx \]

using a symmetric trapezoid rule

\[ R_x^M[a,b]f = \frac{1}{M} \sum_{i=-\infty}^{\infty} \theta_i f(i/M), \]

where the weights \( \{\theta_i\} \) are given by

\[ \theta_i = \begin{cases} 
1, & a < \frac{i}{M} < b, \\
1/2, & a = \frac{i}{M} < b \text{ or } a < \frac{i}{M} = b, \text{ and} \\
0, & a = \frac{i}{M} = b.
\end{cases} \]

For the case in which the interval \([a,b]\) is the interval \([0,1]\), this trapezoid rule simplifies to

\[ R_x^M[0,1]f = \frac{1}{M} \sum_{i=0}^{M} \theta_i f(i/M), \]

where

\[ \theta_i = \begin{cases} 
1, & 0 < i < M, \text{ and} \\
1/2, & i = 0 \text{ or } i = M.
\end{cases} \]

Suppose now that \( f \) is a continuous function defined on the 2-dimensional simplex \( S_2 = \{(x,y)|0 \leq x \leq y \leq 1\} \). One way to approximate the integral

\[ I = \int_{x=0}^{1} \int_{y=x}^{1} f(x,y) dy dx = \int_{S_2} f(x,y) dy dx \]
is to use a product trapezoid rule. We first apply the trapezoid rule $R^M_y[x, 1]$ in the $y$ direction, and we then apply the trapezoid rule $R^M_x[0, 1]$ in the $x$ direction. This gives an overall rule of the form

$$R^M_x[0, 1]R^M_y[x, 1]f = \frac{1}{M^2} \sum_{i=0}^{M} \sum_{j=1}^{M} \theta_{i,j} f(i/M, j/M),$$

where

$$\theta_{i,j} = \begin{cases} 
1, & 0 < i < j < M, \\
1/2, & 0 = i < j < M \text{ or } 0 < i = j < M \text{ or } 0 < i < j = M, \\
1/4, & 0 = i = j < M \text{ or } 0 = i < j = M, \text{ and } \\
0, & 0 < i = j = M.
\end{cases}$$

Figure 2.1 shows the grid points and weights used when $M = 8$.

Generalizing this product trapezoid rule to the $n$-dimensional simplex $S_n$ gives a rule of the form

$$R^M_{x_1}[0, 1] \cdots R^M_{x_n}[x_{n-1}, 1]f = \frac{1}{M^n} \sum_{i_1=0}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_{n-1}=1}^{M} \theta_{i_1, \ldots, i_n} f(i_1/M, \ldots, i_n/M),$$

(2.23)

where

$$\theta_{i_1, \ldots, i_n} = \begin{cases} 
(1/2)^{c(i_1, \ldots, i_n)}, & i_{n-1} \neq M, \\
0, & i_{n-1} = M,
\end{cases}$$

$c(i_1, \ldots, i_n) = I(0 = i_1) + I(i_1 = i_2) + \cdots + I(i_{n-1} = i_n) + I(i_n = M)$, and $I(A)$ is the indicator of the event $A$. Note that there is a lack of symmetry in this rule in that $\theta_{i_1, \ldots, i_n}$ is zero whenever $i_{n-1} = M$, but $\theta_{i_1, \ldots, i_n}$ need not be zero even if $i_2 = 0$.

Lyness and Puri (1973) show that if the function $f(x_1, \ldots, x_n)$ is a polynomial, then the approximations obtained using the rule (2.23) satisfy an Euler-Maclaurin expansion

$$R^M_{x_1}[0, 1] \cdots R^M_{x_n}[x_{n-1}, 1]f = I + \sum_{q=1}^{K} A_q/M^q,$$

(2.24)
Figure 2.1: The weights $\theta_{i,j}$ for the case in which $n = 2$ and $M = 8$. 
where $I$ is the integral of $f$ over $S_n$, $K$ is a finite integer, and $A_q = 0$ for $q$ odd. That is, the error in the approximation (2.23) is an even polynomial in $1/M$. This Euler-Maclaurin expansion allows us to use Romberg integration to combine a sequence of approximations into a new sequence with a higher order of accuracy. To see this, it suffices to note that the Euler-Maclaurin expansion tells us that the error in approximation has exactly the form (2.18). Thus, we may apply Romberg integration exactly as we did in Section 2.3.

These results of Lyness and Puri (1973) are interesting, but they are not computationally feasible except when $n$ is small if one must actually compute the value $f(i_1/M, \ldots, i_n/M)$ at each point in the sum (2.23). However, the joint densities that we wish to integrate have the form $f(x_1, \ldots, x_n) = f_{k_1:m_1}(x_1) \cdots f_{k_n:m_n}(x_n)$. This separable form for the function $f$ allows us to compute the sum on the right side of (2.23) in a recursive fashion.

Note first that the weight $\theta_{i_1, \ldots, i_n}$ can be written as a product involving simple indicator functions. Specifically, we can write $\theta_{i_1, \ldots, i_n}$ as

$$\left(1 - \frac{1}{2} I(i_1 = 0)\right) \left(1 - \frac{1}{2} I(i_2 = i_1)\right) \cdots \left(1 - \frac{1}{2} I(i_n = M)\right) \left(1 - I(i_{n-1} = M)\right).$$

Define values $\{V(t, i)| t = 1, \ldots, n; i = 0, \ldots, M\}$ by setting

$$V(t, i) = \sum_{i_1=0}^{M} \sum_{i_2=i_1}^{M} \cdots \sum_{i_{t-1}=i_{t-2}}^{M} \theta_{i_1, \ldots, i_{t-1}, i} \left(\prod_{j=1}^{t-1} f_{k_j;m_j}(i_j/M)\right) f_{k_t;m_t}(i/M),$$

where

$$\theta_{i_1, \ldots, i_t} = \left(1 - \frac{1}{2} I(i_1 = 0)\right) \left(1 - \frac{1}{2} I(i_2 = i_1)\right) \cdots \left(1 - \frac{1}{2} I(i_t = M)\right) \left(1 - I(i_{t-1} = M)\right).$$

Then the values $\{V(1, i)\}$ are given by

$$V(1, i) = \begin{cases} \frac{1}{2} f_{k_1;m_1}(0), & i = 0, \\ f_{k_1;m_1}(i/M), & 0 < i < M, \text{ and} \\ \frac{1}{2} f_{k_1;m_1}(1), & i = M, \end{cases}$$

27
and the values \{V(t + 1, i)\} may be computed from the values \{V(t, i)\} via the recursion

\[
V(t + 1, i) = \begin{cases} 
\left( \sum_{j=0}^{i-1} V(t, j) + \frac{1}{2} V(t, i) \right) f_{k_{i+1}; m_{i+1}}(i/M), & 0 \leq i \leq M - 1, \\
\left( \sum_{j=0}^{i-1} V(t, j) \right) f_{k_{i+1}; m_{i+1}}(1), & i = M.
\end{cases}
\]

The product trapezoid rule approximation (2.23) is then given by

\[
I_M^{(0)} = \frac{1}{M^n} \sum_{i=0}^{M} V(n, i).
\]

An \texttt{R} function for implementing this algorithm is given in the appendix.

Because this product trapezoid rule algorithm uses a smaller state space than the one used in the algorithm of Section 2.3, it is significantly faster than that algorithm. However, if we attempt to use the product trapezoid rule algorithm to compute probabilities of the form

\[
P(u_1 \leq U_{k_1; m_1} \leq v_1, \ldots, u_n \leq U_{k_n; m_n} \leq v_n, U_{k_1; m_1} \leq \cdots \leq U_{k_n; m_n}),
\]

then the expansion (2.24) is no longer a polynomial in \(1/M^2\). Instead, terms involving odd powers of \(1/M\) also appear in the expansion. As a result, the product trapezoid rule is not as effective in this more general situation as is the rule from Section 2.3. However, the product trapezoid rule suggests a way to improve the earlier rule. Specifically, instead of dividing by \(s_i!\) as in (2.7), we could divide by \(2^{s_i-1}\) when \(s_i > 1\). This factor would certainly be the wrong factor to use in an exact calculation, but it offers the advantage of reducing the dimension of the state space from two to one. In addition, we find that an expansion (2.24) holds for the approximation error under this new scheme.
This hybrid algorithm can be summarized as follows. We use states $s_t(i)$, where $t = 1, \ldots, n$ and $i = 0, \ldots, K$. For $t = 1$, the values $\{s_t(i)\}$ are given by

$$s_1(i) = p(1, i),$$

and a recursive scheme for computing the values $\{s_{t+1}(i)\}$ from the values $\{s_t(i)\}$ is given by

$$s_{t+1}(i) = p(t + 1, i) \left( \frac{1}{2} s_t(i) + \sum_{r < i} s_t(r) \right),$$

where the factor of 1/2 takes care of the necessary division by $2^{s_i-1}$. The final approximation is then given by

$$\sum_{i=0}^{K} s_n(i).$$

An R function for implementing this hybrid algorithm is given in the appendix.

### 2.5 Some Applications of the General Algorithm

As will be seen in Section 3.5, probabilities of the form

$$P \left( u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq U_{k_n:m_n} \leq v_n, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n} \right) \tag{2.25}$$

arise when one wants to construct optimal distribution-free confidence bands for a continuous CDF on the basis of a ranked-set sample or some other sample consisting of independent order statistics. Another setting in which these probabilities arise is in the construction of nonparametric tests. For example, probabilities of the form (2.25) can be used for testing whether the judgment rankings used in drawing a given ranked-set sample are perfect or not. This topic is the subject of a paper by Öztürk, Frey, and Deshpande (2005).

Suppose that $\{X_{ij}; i = 1, \ldots, m, j = 1, \ldots, n_i\}$ is a ranked-set sample from some distribution with continuous CDF $F$. That is, suppose that the values $\{X_{ij}\}$ are
mutually independent and that each $X_{ij}$ is a $j$th judgment order statistic from a set of $m$ observations drawn from $F$. Let $R_{ij}$ be the rank of the measured value $X_{ij}$ among the full set of $N = n_1 + \cdots + n_m$ observations. Since $F$ is a nondecreasing function, the ranking of the values $\{X_{ij}\}$ is the same as the ranking of the values $\{F(X_{ij})\}$. Under the hypothesis of perfect rankings, these latter values $\{F(X_{ij})\}$ are jointly distributed like independent order statistics from a standard uniform parent distribution, meaning that the distribution of the ranks $\{R_{ij}\}$ is distribution-free over the class of continuous $F$. Thus, we can construct a nonparametric test for perfect rankings simply by selecting a subset of the set of all rankings and taking that subset to be our rejection region.

In nonparametric tests based on simple random samples, all rankings are typically equally likely under the null hypothesis. In RSS, however, the observations in each judgment order statistic class have a different distribution, meaning that the probabilities of different rankings may be quite different. As an example, Table 2.5 gives the probability of each possible ranking when $m = 3$, $n_1 = n_2 = n_3 = 1$, and the rankings are perfect.

<table>
<thead>
<tr>
<th>$(R_{[1]1}, R_{[2]1}, R_{[3]1})$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 2, 3)</td>
<td>64/105 ≈ 0.6095</td>
</tr>
<tr>
<td>(1, 3, 2)</td>
<td>143/840 ≈ 0.1702</td>
</tr>
<tr>
<td>(2, 1, 3)</td>
<td>143/840 ≈ 0.1702</td>
</tr>
<tr>
<td>(2, 3, 1)</td>
<td>17/840 ≈ 0.0202</td>
</tr>
<tr>
<td>(3, 1, 2)</td>
<td>17/840 ≈ 0.0202</td>
</tr>
<tr>
<td>(3, 2, 1)</td>
<td>1/105 ≈ 0.0095</td>
</tr>
</tbody>
</table>

Table 2.5: Distribution of the ranks $R_{[1]1}$, $R_{[2]1}$, and $R_{[3]1}$ when $m = 3$, $n_1 = n_2 = n_3 = 1$, and the rankings are perfect.

30
Since each possible ranking \( \{R_{ij}\} \) has a (potentially) different probability under perfect rankings, one natural way to create a rejection region is to take the set of all rankings whose probabilities fall below a certain threshold value. Since these low-probability rankings are in some sense least consistent with perfect rankings, such a test has appeal even without reference to the potential alternatives. In particular, using a rejection region that consists of the data points that are least likely under the null hypothesis is an approach analogous to that used by Fisher in developing his tests of significance. As an example, we note that in the case presented in Table 2.5, a test with level exactly 0.05 can be obtained by rejecting when the observed ranking has probability 17/840 or less.

An alternate motivation for using the probability of the observed collection of ranks as a test statistic can be obtained by noting that under random rankings, each possible ranking is equally likely. Thus, the test that rejects when the probability of the observed ranking is too low is the same as the test that rejects when the ratio of the probability of the observed ranking under perfect rankings to the probability under random rankings is small. Since tests based on such likelihood ratios are known by the Neyman-Pearson Lemma (see, for example, Lehmann (1997), p. 74) to be most powerful tests of one simple hypothesis against another, such tests can be motivated as most powerful rank tests of perfect rankings against random rankings.

Because the distribution theory for the ranks \( \{R_{ij}\} \) is not tractable, critical values for these exact tests are not available analytically. They can, however, be approximated via simulation or determined (for sufficiently small sample sizes) via exhaustive enumeration. In either case, one of the keys to determining critical values is being able to compute the value of the test statistic, which is simply the probability of the
observed ranking under the null hypothesis of perfect rankings. This computation can be done using combinatorical arguments if the sample is very small, but computations for larger samples are more complicated. They can, however, be completed by writing those probabilities in the form (2.25) and applying the algorithms developed in this chapter. To set notation, suppose that for \( i = 1, \ldots, N \), the measured value with rank \( i \) came from the \( k_i \)th judgment order statistic class. That is, assume that the value with rank \( i \) was ranked as the \( k_i \)th smallest value in its set. Then the probability under the null hypothesis that the ranks \( \{R_{ij}\} \) take their observed values is the probability \( P(U_{k_1:m} \leq \ldots \leq U_{k_N:m}) \), where the random variables \( U_{k_1:m}, \ldots, U_{k_N:m} \) are distributed like independent order statistics from a standard uniform parent distribution. This probability can be computed using either the product trapezoid algorithm or the hybrid algorithm developed in Section 2.4.

As an example, we computed critical values for these tests for choices of the set size \( m \) and the number of cycles \( n \) for which the total number of possible rankings was manageable. These critical values are given in Table 2.6, and they are stated in scientific notation so that 5.467E-5, for example, means 0.00005467. One may note that since the number of distinct values of the test statistic is large even for small values of \( m \) and \( n \), tests with levels very close to their nominal level are readily available even for small sample sizes \( N \). This is not always the case when a test statistic can take on only a discrete set of values.
Table 2.6: Exact critical values (in scientific notation) and levels for the exact test. In cases where the levels 0.05 and 0.10 are not achieved (to four decimal places), the two bracketing values are given.
CHAPTER 3

OPTIMAL DISTRIBUTION-FREE CONFIDENCE BANDS FOR A CDF

3.1 An Overview of the Literature

A good starting point for a discussion of the literature on distribution-free confidence bands for a continuous cumulative distribution function (CDF) is the landmark paper of Kolmogorov (1933). This paper originally appeared in Italian, but an English translation credited to Meneghini is included in the collection edited by Kotz and Johnson (1992). Kolmogorov’s paper deals with certain issues related to the problem of estimating a CDF using the empirical distribution function (EDF).

To set notation, suppose that $X_1, \ldots, X_n$ is a collection of independent observations from a distribution with unknown continuous CDF $F$. Then the EDF is the real-valued function $F_n$ defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t),$$

where $I(A)$ is the indicator that the event $A$ is true. It is easily shown that for fixed $t$, $F_n(t)$ is an unbiased estimator of $F(t)$, and it can also be shown that $F_n(t)$ is a consistent estimator of $F(t)$ as $n \rightarrow \infty$. The issues that interested Kolmogorov, however, concerned the functional convergence of $F_n$ to $F$. He defined the statistic
$D_n$ to be

$$D_n = \sup_t |F_n(t) - F(t)|,$$

the supremum of the pointwise absolute deviations between the EDF and the CDF, and he asked whether $D_n$ goes to 0 in probability as $n \to \infty$. To prove that in fact it does, Kolmogorov derived the asymptotic distribution of $D_n$, showing that as $n \to \infty$, $P(D_n \leq \lambda \sqrt{n})$ tends uniformly in $\lambda$ to $G(\lambda)$, where

$$G(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 \lambda^2}.$$

In addition to deriving the asymptotic distribution of $D_n$, Kolmogorov noted that the statistic $D_n$ is distribution-free provided $F$ is continuous, and he derived recursions for computing finite-sample probabilities of the form $P(D_n \leq u/n)$ for positive integers $u$. Kolmogorov did not discuss in his paper the possibility of using $D_n$ in a goodness-of-fit test, and he also did not discuss the possibility of inverting such a test to produce a confidence band for $F$, but other researchers wasted little time in doing so.

One of these researchers was Smirnov, who made a variety of theoretical and practical contributions towards turning Kolmogorov’s results into usable statistical procedures. These contributions are summarized by Stephens (1992). Smirnov proposed one-sided versions

$$D_n^+ = \sup_t \{F_n(t) - F(t)\}$$

and

$$D_n^- = \sup_t \{F(t) - F_n(t)\}$$

of Kolmogorov’s original statistic $D_n$, and he also proposed two-sample versions of these statistics that can be used to test for equality of two unknown continuous
distributions. He published tables of the asymptotic distributions for the statistics $D_n$, $D_n^+$, and $D_n^-$, all of which are now referred to as Kolmogorov-Smirnov statistics.

The testing and inversion process for the Kolmogorov-Smirnov statistic $D_n$ is as follows. Let $F$ be an unknown continuous CDF from which we draw independent observations $X_1, \ldots, X_n$, and suppose that $F_0$ is a fully-specified continuous CDF whose goodness-of-fit we wish to test. Our null and alternative hypotheses are then

$$H_0 : F = F_0 \text{ and } H_1 : F \neq F_0.$$  

If we use $D_n = \sup_t |F_n(t) - F_0(t)|$ as our test statistic and seek to carry out a level $\alpha$ test, then we can determine an appropriate critical value $c_{\alpha}$ by solving the equation

$$P_{H_0}(D_n \leq c_{\alpha}) = 1 - \alpha. \quad (3.1)$$

The test that rejects $H_0$ exactly when $D_n$ exceeds $c_{\alpha}$ is then a level $\alpha$ test. Inverting this test gives a confidence set consisting of all those continuous CDFs $G$ that would not have been rejected had they been used as $F_0$, and this confidence set has a confidence coefficient of $100(1 - \alpha)$% by the usual argument. Because $D_n$ is just the supremum of the pointwise absolute deviations between $F_0$ and $F_n$, the confidence set obtained by inverting the test based on $D_n$ with critical value $c_{\alpha}$ is the set of all continuous CDFs $G$ that never differ from $F_n(t)$ by more than $c_{\alpha}$. It is thus a confidence band with upper bound $\min\{1, F_n(t) + c_{\alpha}\}$ and lower bound $\max\{0, F_n(t) - c_{\alpha}\}$.

One area of research that attracted many leading statisticians and probabilists in the decades following Kolmogorov’s paper was that of asymptotics. Connections were established between Kolmogorov’s statistic $D_n$ and the Brownian bridge on the interval $[0, 1]$, and a variety of theoretical results, including more elegant proofs of
Kolmogorov’s result on the asymptotic distribution of $D_n$, were produced. More relevant to our work, however, are two other areas of research that developed out of Kolmogorov’s paper, namely research into better methods for computing the finite-sample distribution of $D_n$ and other distribution-free statistics and research on distribution-free alternatives to the basic statistic $D_n$. We discuss first the literature on computational methods.

In his 1933 paper, Kolmogorov derived recursions for computing finite-sample probabilities of the form $P(D_n \leq u/n)$, where $u$ is a positive integer. A number of other early papers also presented methods for computing general probabilities of the form $P(D_n^+ \leq \lambda)$ or $P(D_n^- \leq \lambda)$ for the one-sided versions $D_n^+$ and $D_n^-$ of $D_n$. Massey (1949) derived a method for computing $P(D_n \leq \lambda)$ for any rational $\lambda$, but his method slows down considerably when the denominator of the fraction $\lambda$ is different from $n$. It does not seem to have been until almost 1970, in fact, that computational algorithms were developed that were fast enough and general enough to allow computation of values $P(D_n \leq \lambda)$ for arbitrary $\lambda$, thus making it possible to find quantiles of the distribution of $D_n$. These quantiles of the distribution of $D_n$, of course, are what the applied statistician needs to carry out tests and to find confidence bands in small-sample situations.

The determination of critical points for $D_n$ and the related statistics discussed here is essentially a root-finding problem. Given a particular critical value, we compute the corresponding coverage probability. Since the coverage probability is a nondecreasing continuous function of the critical value, we are guaranteed that there is a critical value that gives the desired coverage probability. A simple bisection scheme, starting with an interval known to contain the solution, gives an answer to any desired level of
accuracy in a fixed number of iterations, but more sophisticated root-finding schemes may be faster.

Steck (1971) derived a method for computing the probability that a collection of standard uniform order statistics lies in a specified multidimensional rectangle. To understand his approach, suppose that $0 \leq u_1 \leq \ldots \leq u_n \leq 1$ and $0 \leq v_1 \leq \ldots \leq v_n \leq 1$ are sequences of lower and upper bounds satisfying $u_i < v_i, i = 1, \ldots, n$. Set

$$P_n(u, v) = P(u_i \leq U(i) \leq v_i, i = 1, \ldots, n),$$

(3.2)

where $U(1), \ldots, U(n)$ are order statistics based on a simple random sample of size $n$ from the standard uniform distribution. Steck proved the following result, which expresses the probability (3.2) as the determinant of a matrix.

**Theorem 3.1 (Steck).** Let $M$ be the $n \times n$ matrix whose $(i, j)$ entry is

$$M_{ij} = \begin{cases} (v_i - u_j)^{j-i+1}/(j-i+1)!, & j-i+1 \geq 0, \\ 0, & j-i+1 < 0, \end{cases}$$

(3.3)

where $(x)_+ = \max\{x, 0\}$. Then

$$P_n(u, v) = n! \det[M].$$

Steck’s theorem allows the computation of critical values for $D_n$, the one-sided statistics $D_n^+$ and $D_n^-$, and many other distribution-free statistics. Because it gives an analytic expression for the coverage probability, it is also useful in deriving theoretical results. Unfortunately, because computation of determinants involves subtraction as well as addition and multiplication, Steck’s method becomes numerically unstable when the sample size $n$ becomes even moderately large. Steck reports, for example, that even double precision arithmetic gives inaccurate results at sample sizes as small as $n = 50$, and similar computational difficulties are reported by Calitz (1987).
A method for computing the probability (3.2) that avoids the numerical difficulties encountered by Steck’s determinant method was developed by Noé (1972). Noé’s method, which requires a set-up process similar to that of the algorithm discussed in Section 2.2, avoids the use of subtraction. It does, however, require computation of potentially large binomial coefficients. A cleverly chosen set of events involving standard uniform order statistics is defined, and the probabilities of these events are computed in a recursive scheme that leads in the end to the probability of interest. Like Steck’s determinant, Noé’s recursion can also be used for direct computations of power for specified alternative hypotheses. The computational time involved in implementing Noé’s recursion is comparable to the time required for the new algorithm developed in Section 2.2. One reason for preferring that new algorithm, however, is the fact that it can be adapted for use in computing probabilities similar to (3.2), but involving independent order statistics.

Looking back over the history leading up to the algorithms developed by Steck and Noé, it seems that the act of writing the coverage probability in the form (3.2) was actually a key step in developing a usable algorithm. It was well-known, of course, that because the distribution of $D_n$ was distribution-free, it was sufficient to do the calculations for the case of the standard uniform distribution, which has CDF $F(t) = t$ on the interval [0, 1]. Because of this observation, authors such as Massey (1949) formulated the problem of computing $P(D_n \leq \lambda)$ as that of computing

$$P(\max(t - \lambda, 0) \leq F_n(t) \leq \min(t + \lambda, 1), t \in [0, 1]).$$

This formulation expresses the probability $P(D_n \leq \lambda)$ in terms of $F_n(t)$, whose distribution is not known. Since the only values $F_n(t)$ can take are integer multiples of $1/n$, this formulation also led these authors to think in terms of discrete grids of
points. Only later, when the problem was framed in terms of standard uniform order statistics, were good general algorithms found.

We now discuss the literature on alternatives to the basic Kolmogorov-Smirnov statistic $D_n$. Even very soon after Kolmogorov’s paper, statisticians recognized that the test statistic $D_n$, while sensitive to certain types of alternatives, was perhaps not the ideal test statistic for other situations. In particular, it did not seem likely to perform well against heavy-tailed alternatives. Wald and Wolfowitz (1939) were among the first to point out that very general distribution-free confidence bands were possible, and they even proposed that principles similar to those applied in the Neyman-Pearson theory of hypothesis testing be used to find appropriate confidence band procedures to use in testing against specific alternatives. The first paper to propose a specific test statistic that stood the test of time, however, was Anderson and Darling (1952).

Anderson and Darling (1952) consider weighted versions of the basic statistic $D_n$. Specifically, they let $\psi(x)$ be a positive weight function defined on the interval $[0, 1]$. They then define the weighted Kolmogorov-Smirnov statistic $D_\psi$ as

$$D_\psi = \sup_t |F_n(t) - F(t)| \sqrt{\psi(F(t))}.$$  \hspace{1cm} (3.4)

Given almost any choice of the positive function $\psi$, a valid distribution-free test can be based on $D_\psi$. However, in order for that test to yield a confidence band for $F$ when inverted, the solution to the inequality

$$|F_n(t) - F(t)| \leq K \psi(F(t))^{-1/2}$$

must be an interval for each positive critical value $K$ and each possible value for $F_n(t)$. The constant weight function $\psi(x) = 1$ gives the usual Kolmogorov-Smirnov statistic.
However, the more interesting weight function now associated with the names Anderson and Darling is the function
\[ \psi(x) = (x(1-x))^{-1/2}. \]
The motivation for this choice of weights is the observation that since \( F_n(t) \sim \frac{1}{n} Bin(n, F(t)) \), the variance of \( F_n(t) \) is proportional to \( F(t)(1-F(t)) \). Letting \( \psi(x) \) be \((x(1-x))^{-1/2}\) thus has the effect of standardizing the pointwise deviations \( F_n(t) - F(t) \). Confidence bands based on this Anderson-Darling test statistic tend to have better power against heavy-tailed alternatives than do the confidence bands based on \( D_n \). They also, however, tend to be very wide.

A variety of other weighted versions of \( D_n \) have been explored, and some authors have proposed test statistics based on taking the maximum of two or more weighted test statistics. Mason and Schuenemeyer (1983), for example, show that certain test statistics arising from taking the maximum of test statistics that are separately sensitive to heavy-tailed or light-tailed distributions yield tests with nice asymptotic properties.

One approach to confidence bands that does not rely on the absolute deviation \(|F_n(t) - F(t)|\) was proposed by Owen (1995). Owen’s approach, which is motivated by the work of Berk and Jones (1979), consists of using a test statistic that is formed by taking the infimum of a collection of pointwise test statistics, one for each real number \( t \). For fixed \( t \), \( nF_n(t) \sim Bin(n, F(t)) \). A likelihood ratio type test of whether \( F(t) = F_0(t) \) can thus be based on the quantity
\[
L(F_0(t), F_n(t)) = \frac{F_0(t)^{nF_n(t)}(1-F_0(t))^{n-nF_n(t)}}{F_n(t)^{nF_n(t)}(1-F_n(t))^{n-nF_n(t)}},
\]
where a high value of \( L(F_0(t), F_n(t)) \) favors the null hypothesis. Letting the overall test statistic be \( R_n = \inf_t L(F_0(t), F_n(t)) \) and inverting the resulting hypothesis test gives what Owen has called the nonparametric likelihood confidence band for the
CDF. Owen appeals to asymptotic theory developed by Berk and Jones (1979) to argue that nonparametric likelihood confidence bands should have better asymptotic power properties than bands based on the Kolmogorov-Smirnov statistic $D_n$ or any of the weighted test statistics $D_\psi$.

Our focus thus far has been on confidence bands for an unknown continuous CDF $F$. However, associated with each CDF $F$ is a quantile function $Q$ defined by $Q(u) = F^{-1}(u) = \inf\{t : F(t) \geq u\}$, and interest often lies more in the properties of $Q$ than in the properties of $F$. What is very convenient about the theory of distribution-free confidence bands for a continuous CDF is that in producing a distribution-free confidence band for $F$, one automatically obtains a distribution-free confidence band for the quantile function $Q$. This is true for all continuous CDFs $F$, but the argument can be presented most clearly if we assume that $F$ is strictly increasing on the real line. Suppose that $C(X_1, \ldots, X_n) \subset \mathbb{R}^2$ is a confidence band for $F$ with coverage probability $100(1 - \alpha)%$. Define the set $C_{\text{inv}}(X_1, \ldots, X_n) \subset \mathbb{R}^2$ by

$$C_{\text{inv}}(X_1, \ldots, X_n) = \{(y, x) : (y, x) \in C(X_1, \ldots, X_n)\}.$$  \hspace{1cm} (3.5)

The continuous CDF $F$ is contained in $C(X_1, \ldots, X_n)$ exactly when all points $(x, y)$ of the graph of $F$ are contained in $C(X_1, \ldots, X_n)$. But a point $(x, y)$ is part of the graph of $F$ if and only if the point $(y, x)$ is part of the graph of $Q$. It thus follows by (3.5) that $F$ is in $C(X_1, \ldots, X_n)$ if and only if $Q$ is in $C_{\text{inv}}(X_1, \ldots, X_n)$. As a result, the set $C_{\text{inv}}(X_1, \ldots, X_n)$ is a $100(1 - \alpha)%$ confidence band for $Q$.

Although most of the literature on distribution-free confidence bands for a continuous CDF has treated the case in which the observations $X_1, \ldots, X_n$ are independent draws from the distribution with CDF $F$, there do exist more general situations in which distribution-free confidence bands can be constructed. One such situation is
that of ranked-set sampling (RSS), a sampling procedure introduced by McIntyre (1952). Stokes and Sager (1988) propose a method for finding a confidence band for $F$ based on a balanced ranked-set sample. To set notation, let $k$ be the set size, and let $n$ be the number of cycles. Then the corresponding balanced ranked-set sample consists of the $nk$ observations

$$X_{(1)1}, \ldots, X_{(1)n}$$
$$X_{(2)1}, \ldots, X_{(2)n}$$
$$\vdots$$
$$X_{(k)1}, \ldots, X_{(k)n},$$

where the $\{X_{(i)j}\}$ are mutually independent, and $X_{(i)j}$ is distributed like the $i$th order statistic from a random sample of size $k$ drawn from the distribution with CDF $F$.

Stokes and Sager (1988) showed that in this balanced RSS situation, the RSS EDF

$$F^*(t) = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} I(X_{(i)j} \leq t)$$

continues to be pointwise an unbiased estimator of $F$. They then proposed that one use the Kolmogorov-Smirnov type statistic

$$D^* = \sup_t |F^*(t) - F_0(t)|$$

to conduct distribution-free tests of $H_0$ versus $H_1$. This statistic leads to a test that for fixed level $\alpha$ has a smaller critical value than one would obtain when testing with a simple random sample of size $nk$. Consequently, inverting the test based on $D^*$ gives a narrower confidence band for $F$ than does inverting the test based on the usual Kolmogorov-Smirnov statistic $D_{nk}$. One objection to the use of this method in practice is the fact that when the judgment rankings used in drawing the ranked-set sample are not perfect, the procedure will not have the prescribed level. Another
objection, which applies even when the judgment ranking is done perfectly, is that
the test based on $D^*$, while exact in an unconditional sense, may have wildly different
conditional levels depending on the observed sequence of order statistics in the sample.

In Section 3.2, we describe the class of distribution-free confidence bands that we
will consider, and we explain the connection between coverage probabilities for these
bands and rectangle probabilities for standard uniform order statistics. In Section
3.3, we introduce a broad class of optimality criteria motivated by the idea that
the difference between the upper and lower bounds of the confidence band should
be small, and we derive the theoretical results needed for applying these optimality
criteria in practice. The key result that is used in much of this work is the Brunn-
Minkowski Inequality from the theory of convex bodies. In Section 3.4, we compare
the performance of the optimal confidence bands to that of other distribution-free
confidence bands both in terms of width and in terms of power against a variety
of alternative distributions. In Sections 3.5 and 3.6, we generalize the results of
Sections 3.2 to 3.4 to the situation in which the sample is a ranked-set sample or
some other sample consisting of independent order statistics from a continuous parent
distribution $F$.

### 3.2 A Class of Distribution-free Confidence Bands

Let $X_1, \ldots, X_n$ be a simple random sample from the distribution with continuous
CDF $F$. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics corresponding to $X_1, \ldots, X_n$, and
set $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. Let the vectors $a = (a_0, a_1, \ldots, a_n)$ and $b =
(b_0, b_1, \ldots, b_n)$ be vectors of values from the interval $[0, 1]$. We can then define a
confidence band $CB(a, b)$ as follows. We say that a continuous CDF $G$ is contained
in the confidence band $CB(a, b)$ if for each $i = 0, \ldots, n$, $a_i \leq G(t) \leq b_i$ for all $t$ in the interval $(X_{(i)}, X_{(i+1)})$. It is immediately clear that the confidence band must have coverage probability 0 if $a_0 > 0$ or $b_n < 1$. Moreover, if $a_{i+1} \leq a_i$, then the coverage probability of the band will be unchanged if we increase $a_{i+1}$ to $a_i$, and if $b_{i+1} \leq b_i$, then the coverage probability of the band will be unchanged if we decrease $b_i$ to $b_{i+1}$. Finally, given that the lower bounds $a_0, \ldots, a_n$ and the upper bounds $b_0, \ldots, b_n$ are ordered as $a_0 \leq \ldots \leq a_n$ and $b_0 \leq \ldots \leq b_n$, the coverage probability will necessarily be 0 unless $a_i \leq b_{i-1}$, $i = 1, \ldots, n$, so that the intervals $(a_{i-1}, b_{i-1})$ and $(a_i, b_i)$ always overlap. It thus suffices to consider only vectors of bounds $a$ and $b$ satisfying the conditions

\begin{align*}
0 &= a_0 \leq a_1 \leq \ldots \leq a_n \leq 1, \\
0 &\leq b_0 \leq \ldots \leq b_n = 1, \text{ and} \\
a_i &\leq b_{i-1}, i = 1, \ldots, n. \quad (3.6)
\end{align*}

Note that the randomness of the band $CB(a, b)$ arises from the randomness in the values $X_{(1)}, \ldots, X_{(n)}$ and not from any randomness in the vectors of bounds $a$ and $b$. All Kolmogorov-Smirnov bands are confidence bands $CB(a, b)$ for properly chosen $a$ and $b$, but many other possibilities also exist. In particular, all weighted Kolmogorov-Smirnov confidence bands and the nonparametric likelihood confidence bands discussed in Section 3.1 have this form.

One may wonder whether the class of confidence bands that we are considering includes all distribution-free confidence bands. It does not, as it excludes, for example, randomized confidence bands. There is evidence, however, that the class includes all distribution-free confidence bands that one would want to use in practice. For
example, the class contains all confidence bands whose upper and lower bounds are functions of the EDF $F_n(t)$, which is often called the nonparametric sufficient statistic. In addition, Lehmann (1997, pp. 326-328) showed that the class contains all confidence bands that are equivariant under a continuous, strictly increasing transformation of the data.

Given vectors of bounds $a$ and $b$ that satisfy (3.6), we may compute the coverage probability of $CB(a, b)$ as follows. First, using the fact that $F$ is a nondecreasing, continuous function, we may rewrite the probability $P_F(F \in CB(a, b))$ in a more convenient form. Consider the interval $(X_{(0)}, X_{(1)})$. A necessary and sufficient condition for $F$ not to pass above the upper bound $b_0$ on this interval is that $F(X_{(1)})$, the value of $F$ at the right-most point on the interval, be no larger than $b_0$. Similarly, a necessary and sufficient condition for $F$ not to pass below the lower bound $a_0 = 0$ on this interval is that $F(X_{(0)}) = 0$ be no smaller than $a_0$. The lower bound condition is thus trivially met on this particular interval, but not in general. Similar pairs of conditions are obtained for each of the intervals $(X_{(i)}, X_{(i+1)})$, $i = 1, \ldots, n$, and we may write that

$$P_F(F \in CB(a, b)) = P_F(a_1 \leq F(X_{(1)}) \leq b_0, \ldots, a_n \leq F(X_{(n)}) \leq b_{n-1}).$$

However, since $F$ is a continuous CDF, the vector $(F(X_{(1)}), \ldots, F(X_{(n)}))$ is distributed exactly like a vector of order statistics from the standard uniform distribution. As a result, we may write the coverage probability in the form

$$P_F(F \in CB(a, b)) = P(a_1 \leq U_{(1)} \leq b_0, \ldots, a_n \leq U_{(n)} \leq b_{n-1}),$$

which is exactly the sort of rectangle probability that can be computed using Steck’s determinant or Noé’s recursion. As a final simplification, we note that because of the
relationship between the distribution of a simple random sample \( U_1, \ldots, U_n \) and the
distribution of the order statistics \( U_{(1)}, \ldots, U_{(n)} \), we may write \( P_F(F \in CB(a, b)) \) as
\[
n!P(a_1 \leq U_1 \leq b_0, \ldots, a_n \leq U_n \leq b_{n-1}, U_1 \leq U_2 \leq \ldots \leq U_n).
\] (3.8)

To compute this probability (3.8), we apply the algorithm for computing rectangle
probabilities for standard uniform order statistics that we developed in Section
2.2. As a concrete example of the computations involved, we compute the coverage
probability for a particular Kolmogorov-Smirnov confidence band.

Example 3.1: Suppose that the sample size is \( n = 4 \), and suppose that the half-width
is 1/2. Then the vectors of lower and upper bounds for the confidence band are
\[
\left(0, 0, 0, \frac{1}{4}, \frac{1}{2}\right) \quad \text{and} \quad \left(\frac{1}{2}, \frac{3}{4}, 1, 1, 1\right),
\]
respectively, meaning that the coverage probability can be written as
\[
P \left(0 \leq U_{(1)} \leq \frac{1}{2}, 0 \leq U_{(2)} \leq \frac{3}{4}, \frac{1}{4} \leq U_{(3)} \leq 1, \frac{1}{2} \leq U_{(4)} \leq 1\right)
= 4!P \left(0 \leq U_1 \leq \frac{1}{2}, 0 \leq U_2 \leq \frac{3}{4}, \frac{1}{4} \leq U_3 \leq 1, \frac{1}{2} \leq U_4 \leq 1, U_1 \leq U_2 \leq U_3 \leq U_4\right).
\]

Letting \( V_1, \ldots, V_4 \) be independent standard uniform random variables truncated to
the intervals \((0, \frac{1}{4}), (0, \frac{3}{4}), (\frac{1}{4}, 1)\), and \((\frac{1}{2}, 1)\), respectively, we then have that the
coverage probability is
\[
4! \left(\frac{1}{2}\right)^2 \left(\frac{3}{4}\right)^2 P(V_1 \leq V_2 \leq V_3 \leq V_4).
\]

The sequence \((\frac{1}{4}, \frac{1}{2}, \frac{3}{4})\) and the random variables \( V_1, V_2, V_3, V_4 \) satisfy the EC con-
dition of Section 2.2, meaning that we can apply the algorithm given by (2.9) to
(2.11). The values \( \{p(t, i)\} \) are as given in Table 3.1. The values \( \{s_t(i, j)\} \) for times
\( t = 1, \ldots, 4 \) are then as given in Table 3.2. Summing the last column of Table
Table 3.1: The values \( \{p(t,i)\} \) when the algorithm of Section 2.2 is applied to the example in Section 3.2.

3.2 gives us that \( P(V_1 \leq \ldots \leq V_4) = 13/54 \), and the coverage probability is then 
\[ 4!(1/4)(9/16)(13/54) = 13/16, \text{ or exactly } 81.25\%. \] In practice, of course, the zeros that appear in Table 3.2 are known in advance (since they are determined by the values \( \{p(t,i)\} \)) and are used to reduce the computational burden.

The coverage probabilities we obtain from this procedure are correct provided that the underlying CDF \( F \) is continuous. When it fails to be continuous, however, the coverage probability depends on the specific distribution \( F \). If one specifies the distribution \( F \), then arguments described by Gleser (1985) can be used to write the coverage probability for any confidence band as a rectangle probability involving uniform order statistics. Since \( F \) is in practice unknown, however, it is more useful to know that the coverage probabilities are conservative when \( F \) fails to be continuous. This important fact can be shown using the following theorem, which was proved by Guilbaud (1986).

**Theorem 3.2 (Guilbaud).** Let \( \mathcal{F} \) be the family of all CDFs. For each \( F \in \mathcal{F} \), let 
\[ V_F = \{ \nu : F(x) = \nu \text{ for some } x \in \mathbb{R} \} \] 
be the range of \( F \), and let \( D_F \) be the random
<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>$s_1(i,j)$</th>
<th>$s_2(i,j)$</th>
<th>$s_3(i,j)$</th>
<th>$s_4(i,j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1/2</td>
<td>1/6</td>
<td>1/36</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/12</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/108</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1/3</td>
<td>1/9</td>
<td>7/216</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1/18</td>
<td>1/36</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/108</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2/9</td>
<td>25/216</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/18</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.2: The values $\{s_t(i,j)\}$ when the algorithm of Section 2.2 is applied to the example in Section 3.2.
variable defined by $D_F = \sup_t d(F_n(t), F(t))$, where $d$ is a real-valued function defined on $[0, 1]^2$. Then the distribution of $D_F$ depends only on $V_F$, and $D_{F_1}$ is stochastically smaller than $D_{F_2}$ if $V_{F_1} \subset V_{F_2}$.

The fact that these confidence bands are conservative when $F$ fails to be continuous then follows as a corollary.

**Corollary 3.1.** If $CB(a, b)$ has coverage probability $\alpha$ when $F$ is continuous, then $CB(a, b)$ has coverage probability at least $\alpha$ when $F$ fails to be continuous.

**Proof of Corollary 3.1.** Define a function $d : [0, 1]^2 \to \mathbb{R}$ by setting

$$
d(x, y) = \begin{cases} 
\frac{y-\frac{1}{n}(a_i+b_i)}{\frac{1}{n}(b_i-a_i)} & , \quad x = \frac{i}{n} \\
0, & \text{otherwise},
\end{cases}
$$

and let $D_F = \sup_t d(F_n(t), F(t))$. Then the coverage probability of the band $CB(a, b)$ is exactly the probability $P(D_F \leq 1)$. Let $F$ be a continuous CDF, and let $F_d$ be some discontinuous CDF. Then $V_F = [0, 1]$, while $V_{F_d} \subset [0, 1]$. Thus, we have that $D_{F_d}$ is stochastically smaller than $D_F$, meaning that $P(D_{F_d} \leq 1) \geq P(D_F \leq 1) = \alpha$. $\square$

### 3.3 A Class of Optimality Criteria

In most of the literature on distribution-free confidence bands for a continuous CDF, the emphasis is on finding confidence bands that have good power properties when they are used in the context of goodness-of-fit testing. However, if testing is really the goal, then much better tests are available than the ones that can be inverted to give confidence bands for $F$. For example, the familiar Neyman-Pearson Lemma allows one to find the most powerful test of the point null hypothesis $H_0 : F = F_0$ against any simple alternative. One might argue, instead, that statisticians producing
confidence bands for $F$ intend to use the band as an exploratory analysis tool or as a tool for simultaneous inference rather than merely as a tool for goodness-of-fit testing.

When one thinks about a confidence band for $F$ as an exploratory analysis tool or a tool for simultaneous inference on quantiles or values $F(t)$, it becomes clear that the quality of a given band is essentially a function of how narrow that band is. Since the confidence bands introduced in Section 3.2 are of fixed width between pairs of consecutive order statistics, the $n + 1$ differences $b_0 - a_0, \ldots, b_n - a_n$ summarize what there is to say about the narrowness of the band. One simple optimality criterion would thus be the sum $\sum_{i=0}^{n} (b_i - a_i)$, which just totals the differences between the upper and lower bounds of the confidence band on each interval $(X_{(i)}, X_{(i+1)})$. Alternatively, one might choose a vector of positive weights $(w_0, \ldots, w_n)$ and seek to minimize the sum $\sum_i w_i (b_i - a_i)$ over all bands $CB(a, b)$ with a specified coverage probability. If one chooses the weights on the extremes to be large, that indicates special interest in having the band be narrow at the extremes. Similarly, choosing the weights to be large in the middle would indicate special interest in having the band be narrow in the middle.

Consider the problem of minimizing the criterion $\sum_i w_i (b_i - a_i)$ over all bands $CB(a, b)$ with coverage probability $\alpha \in (0, 1)$. We will prove that a minimizer must exist, that any minimizer must satisfy a small set of necessary and sufficient conditions, and that the minimizer will be unique under certain conditions that are often met in practice. First, however, we need to prove some results about coverage probabilities. Note that since $a_0 = 0$ and $b_n = 1$ are required for a band to have a nonzero coverage probability, there are actually only $2n$ bounds that need to be chosen to specify a band. Thus, we will assume that $a_0 = 0$ and $b_n = 1$ in the remainder of this
chapter, shortening the vectors \( a \) and \( b \) accordingly. Let \( D \) be the set of all vectors 
\((a, b) = (a_1, \ldots, a_n, b_0, \ldots, b_{n-1}) \in [0, 1]^{2n}\) that satisfy the conditions (3.6). Then, 
since \( D \) is the intersection of a finite collection of closed half-spaces, it is convex and 
closed. Since \( D \) is contained in the bounded set \([0, 1]^{2n}\), it is also bounded and hence 
compact. We may define the coverage probability function \( H : D \rightarrow [0, 1] \) by 
\[
H(a_1, \ldots, a_n, b_0, \ldots, b_{n-1}) = P_F(F \in CB(a, b)).
\]

An extremely useful alternate representation of \( H \) is the integral representation 
\[
H(a_1, \ldots, a_n, b_0, \ldots, b_{n-1}) = n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_i = a_i}^{b_i} \cdots \int_{u_n = a_n}^{b_n-1} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1, \quad (3.9)
\]
which follows immediately from (3.8). We now derive some results dealing with 
derdifferentiability properties of \( H \). In what follows, we use the upper-case letters 
\( A_1, \ldots, A_n, B_0, \ldots, B_{n-1} \) to indicate the bounds of a confidence bound as variables 
and the corresponding lower-case letters to indicate particular values for the bounds.

We first consider differentiation of \( H \) with respect to a lower bound. Let \((a, b) \in D\), 
where \( a = (a_1, \ldots, a_n) \) and \( b = (b_0, \ldots, b_{n-1}) \), and suppose that \( a_i \) is a lower bound 
such that \( a_i < a_{i+1} \). Then if \( a_i \) is increased by a sufficiently small amount \( \epsilon > 0 \), the 
modified vector is still in \( D \). We have that 
\[
\Delta(\epsilon) \equiv H(a_1, \ldots, a_i + \epsilon, \ldots, b_{n-1}) - H(a_1, \ldots, a_i, \ldots, b_{n-1}) \\
= n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_i = a_i + \epsilon}^{b_i} \cdots \int_{u_n = a_n}^{b_n-1} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1 - \\
\quad n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_i = a_i}^{b_i} \cdots \int_{u_n = a_n}^{b_n-1} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1 \\
= -n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_i = a_i + \epsilon}^{b_i} \cdots \int_{u_n = a_n}^{b_n-1} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1.
\]
If we then take \( \lim_{\epsilon \to 0} \frac{\Delta(\epsilon)}{\epsilon} \), we find that the right-hand partial derivative of \( H \) with respect to \( A_i \) is just

\[
\frac{\partial H}{\partial A_i^+} \bigg|_{(A,B)=(a,b)} = -n! \int_{a_1}^{b_0} \int_{a_i-\epsilon}^{b_i} \int_{a_{i+1}}^{b_{i} \land a_i} \int_{a_{i+1}}^{b_{i+1}} \int_{a_{i-1}}^{b_{i-1}} \int_{a_n}^{b_n-1} I(u_1 \leq \ldots u_{i-1} \leq u_{i+1} \ldots \leq u_n) du_n \ldots du_{i+1} du_i du_1,
\]

where \( x \land y = \min\{x, y\} \). Note that this partial derivative is the sort of quantity that can be computed using the algorithm developed in Section 2.2.

Suppose now that \( a_i \) satisfies \( a_i > a_{i-1} \). Then if \( a_i \) is decreased by a sufficiently small amount \( \epsilon > 0 \), the modified vector is still in \( D \). We then have that

\[
\Delta(\epsilon) \equiv H(a_1, \ldots, a_i - \epsilon, \ldots, b_{n-1}) - H(a_1, \ldots, a_i, \ldots, b_{n-1})
\]

\[
= n! \int_{u_i=a_1}^{b_0} \int_{u_i=a_{i-\epsilon}}^{b_i} \int_{u_{i+1}=a_{i+1}}^{b_{i+1}} \int_{u_{i-1}=a_{i-1}}^{b_{i-1}} \int_{u_n=a_n}^{b_n-1} I(u_1 \leq \ldots u_{i-1} \leq u_{i+1} \ldots \leq u_n) du_n \ldots du_1 -
\]

\[
= n! \int_{u_i=a_1}^{b_0} \int_{u_i=a_i}^{b_i} \int_{u_{i+1}=a_{i+1}}^{b_{i+1}} \int_{u_{i-1}=a_{i-1}}^{b_{i-1}} \int_{u_n=a_n}^{b_n-1} I(u_1 \leq \ldots u_{i-1} \leq u_{i+1} \ldots \leq u_n) du_n \ldots du_1.
\]

If we then take \( \lim_{\epsilon \to 0} \frac{\Delta(\epsilon)}{\epsilon} \), we find that the left-hand partial derivative of \( H \) with respect to \( A_i \) is just

\[
\frac{\partial H}{\partial A_i^-} \bigg|_{(A,B)=(a,b)} = -n! \int_{a_1}^{b_0} \int_{a_i-\epsilon}^{b_i} \int_{a_{i+1}}^{b_{i+1}} \int_{a_{i-1}}^{b_{i-1}} \int_{a_n}^{b_n-1} I(u_1 \leq \ldots u_{i-1} \leq u_{i+1} \ldots \leq u_n) du_n \ldots du_{i+1} du_i du_1,
\]

which is the same as the right-hand partial derivative. We have thus shown that \( H \) has a regular partial derivative with respect to \( A_i \) whenever \( a_{i-1} < a_i < a_{i+1} \). Note, furthermore, that because the argument \( a_i \) appears continuously in the expressions
(3.10) and (3.11), the left-hand and right-hand partial derivatives are continuous in $A_i$ whenever they exist.

We now consider differentiation of $H$ with respect to an upper bound. Let $(a, b) \in D$ as before, and suppose that $b_j$ is an upper bound such that $b_j < b_{j+1}$. Then if $b_j$ is increased by a sufficiently small amount $\epsilon > 0$, the modified vector is still in $D$. We have that

$$\Delta(\epsilon) \equiv H(a_1, \ldots, b_j + \epsilon, \ldots, b_{n-1}) - H(a_1, \ldots, b_j, \ldots, b_{n-1}) = n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_j+1 = a_{j+1}}^{b_j+\epsilon} \cdots \int_{u_n = a_n}^{b_{n-1}} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1 \left[ \int_{u_1 = a_1}^{b_0} \cdots \int_{u_j = a_j}^{b_j} \cdots \int_{u_n = a_n}^{b_{n-1}} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1 \right].$$

If we then take $\lim_{\epsilon \to 0} \frac{\Delta(\epsilon)}{\epsilon}$, we find that the right-hand partial derivative of $H$ with respect to $B_j$ is just

$$\frac{\partial H}{\partial B_j^+}_{(A,B) = (a,b)} = n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_j = a_j}^{b_j+1} \cdots \int_{u_n = a_n \lor b_j}^{b_{n-1}} I(u_1 \leq \ldots u_j \leq u_{j+2} \ldots \leq u_n) du_n \ldots du_{j+2} du_j \ldots du_1,$$

where $x \lor y = \max\{x, y\}$.

(3.12)
Suppose, finally, that \( b_j \) satisfies \( b_j > b_{j-1} \). Then if \( b_j \) is decreased by a sufficiently small amount \( \epsilon > 0 \), the modified vector is still in \( D \). We then have that

\[
\Delta(\epsilon) \equiv H(a_1, \ldots, b_j - \epsilon, \ldots, b_{n-1}) - H(a_1, \ldots, b_j, \ldots, b_{n-1})
\]

\[
= n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_{j+1} = a_{j+1}}^{b_j - \epsilon} \cdots \int_{u_n = a_n}^{b_{n-1}} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1 -
\]

\[
= n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_{j+1} = a_{j+1}}^{b_j} \cdots \int_{u_n = a_n}^{b_{n-1}} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1
\]

\[
= -n! \int_{u_1 = a_1}^{b_0} \cdots \int_{u_{j+1} = b_{j-\epsilon}}^{b_j} \cdots \int_{u_n = a_n}^{b_{n-1}} I(u_1 \leq \ldots \leq u_n) du_n \ldots du_1.
\]

If we then take \( \lim_{\epsilon \to 0} \frac{\Delta(\epsilon)}{\epsilon} \), we find that the left-hand partial derivative of \( H \) with respect to \( B_j \) is just

\[
\left. \frac{\partial H}{\partial B_j} \right|_{(A,B) = (a,b)} = n! \int_{u_1 = a_1}^{b_0} \cdots \int_{a_j}^{b_{j-1}} \int_{a_{j+2} \vee b_j}^{b_{j+1}} \cdots \int_{a_n \vee b_j}^{b_{n-1}} I(u_1 \leq \ldots u_j \leq u_{j+2} \ldots \leq u_n) du_n \ldots du_{j+2} du_j \ldots du_1,
\]

which is the same as the right-hand partial derivative. We have thus shown that \( H \) has a regular partial derivative with respect to \( B_j \) whenever \( b_{j-1} < b_j < b_{j+1} \).

Note, furthermore, that just as in the case of taking derivatives with respect to a lower bound \( A_i \), the left-hand and right-hand partial derivatives are continuous in \( B_j \) whenever they exist. Putting these results on partial derivatives with respect to upper bounds together with the results on partial derivatives with respect to lower bounds, we have shown the following theorem.

**Theorem 3.3.** The coverage probability function \( H \) has one-sided partial derivatives with respect to each lower bound \( A_i \) and each upper bound \( B_j \) at every point where modifying the bound in question by an arbitrarily small amount in the direction of
interest does not take \((a, b)\) out of \(D\). The left-hand and right-hand partial derivatives coincide when they both exist. Moreover, whenever a left-hand or right-hand partial derivative exists, it is continuous in the coordinate corresponding to that bound.

Theorem 3.3 implies that \(H\) is continuous in \(D\). More importantly, because the one-sided partial derivatives are continuous when they exist, Theorem 3.3 guarantees that we can write the change in coverage probability resulting from multiple infinitesimal changes in the bounds in terms of the partial derivatives of \(H\) at the starting point.

**Theorem 3.4.** Let \((a, b) \in D\), and let \(\delta = (\delta_{a_1}, \ldots, \delta_{b_{n-1}})\) be a fixed vector of real numbers. Suppose that there exists \(r > 0\) small enough that adding \(\epsilon \delta\) to \((a, b)\) results in another point in \(D\) whenever \(\epsilon < r\). Then, if \(\frac{\partial H}{\partial A_1}\bigg|_{(A,B)=(a,b)}\) and \(\frac{\partial H}{\partial B_{n-1}}\bigg|_{(A,B)=(a,b)}\) are the partial derivatives of \(H\) with respect to the bounds \(A_i\) and \(B_j\) in the directions corresponding to \(\delta_{a_i}\) and \(\delta_{b_j}\), we have that

\[
H(a_1 + \epsilon \delta_{a_1}, \ldots, b_{n-1} + \epsilon \delta_{b_{n-1}}) = H(a_1, \ldots, b_{n-1}) + \epsilon \left( \delta_{a_1} \left. \frac{\partial H}{\partial A_1} \right|_{(A,B)=(a,b)} + \ldots + \delta_{b_{n-1}} \left. \frac{\partial H}{\partial B_{n-1}} \right|_{(A,B)=(a,b)} \right) + o(\epsilon).
\]

(3.14)

The following theorem uses the continuity of \(H\) to show that an optimal confidence band must exist.

**Theorem 3.5.** Let \((w_0, \ldots, w_n)\) be a vector of positive weights, and let \(\alpha \in (0, 1)\) be a specified coverage probability. Then there exists a confidence band that minimizes \(\sum_i w_i(b_i - a_i)\) over the set of all confidence bands \((a, b) \in D\) such that \(P_{F}(F \in CB(a, b)) = \alpha\).
Proof of Theorem 3.5. Since $H$ is continuous, and since coverage probability 0 and coverage probability 1 are each attainable in $D$, the set $\{(a,b) \in D : P_F(F \in CB(a,b)) = \alpha\}$ is nonempty. Also, since $H$ is continuous and $D$ is bounded, the set $\{(a,b) \in D : P_F(F \in CB(a,b)) = \alpha\}$ is compact. Since the function $(a,b) \rightarrow \sum_i w_i(b_i - a_i)$ is obviously continuous, it must achieve a minimum on the compact set $\{(a,b) \in D : P_F(F \in CB(a,b)) = \alpha\}$.

We now consider the problem of finding necessary and sufficient conditions for a given confidence band to be optimal. Suppose that $(a,b)$ is a vector that minimizes the criterion $\sum_i w_i(b_i - a_i)$ over all vectors $(a,b) \in D$ with coverage probability $\alpha$. The next theorem shows that for any optimal band, the lower bounds $a_1, \ldots, a_n$ and the upper bounds $b_0, \ldots, b_{n-1}$ are strictly increasing as opposed to merely nondecreasing.

The following lemma provides a key observation that is used in the proof of the theorem.

**Lemma 3.1.** If a confidence band $(a,b) \in D$ is optimal with respect to the weights $(w_0, \ldots, w_n)$, then no confidence band can have the same value as $(a,b)$ for the optimality criterion and yet have a higher coverage probability.

**Proof of Lemma 3.1.** We proceed by proving the contrapositive. Let $(a,b) \in D$ be a confidence band with coverage probability $\alpha \in (0,1)$, and suppose that $(c,d) \in D$ is another band such that $\sum_i w_i(b_i - a_i) = \sum_i w_i(d_i - c_i)$, but $H(c,d) > H(a,b) = \alpha$. Since $H(c,d) > 0$, the bound $d_0$ satisfies $d_0 > c_1$. If we lower $d_0$ towards $c_1$, then the modified confidence band remains in $D$. As we decrease $d_0$, we decrease the optimality criterion $\sum_i w_i(d_i - c_i)$, and we also decrease the coverage probability from $H(c,d) > \alpha$ at the start to 0 once $d_0 = c_1$. Since $H$ is a continuous function, there must exist
a real number \( d^* \in (c_1, d_0) \) such that \( H(c_1, \ldots, c_n, d^*, d_1, \ldots, d_{n-1}) = \alpha \). Since this new band has the same coverage probability \( \alpha \) as \((a, b)\), but a smaller value for the optimality criterion, the confidence band \((a, b)\) is not optimal with respect to the weights \((w_0, \ldots, w_n)\).

\[ \square \]

**Theorem 3.6.** Let \((a, b) \in D\) be a confidence band that is optimal at level \( \alpha \in (0, 1) \) with respect to the weights \((w_0, \ldots, w_n)\). Then \( a_1 < \ldots < a_n \), and \( b_0 < \ldots < b_{n-1} \).

**Proof of Theorem 3.6.** Suppose that there exists \( i \in \{1, 2, \ldots, n-1\} \) such that \( a_i = a_{i+1} \). Then there either exists an integer \( l \) such that \( a_i = \ldots = a_{i+l} < a_{i+l+1} < 1 \) or an integer \( l \) such that \( a_i = \ldots = a_{i+l} < 1 \) and \( i + l = n \), where the last inequality (involving 1) follows in each case from the fact that \( \alpha = H(a, b) > 0 \). By (3.10), the partial derivative \( \frac{\partial H}{\partial A_i} \big|_{(A, B) = (a, b)} \) is equal to 0, while by (3.11), the partial derivative \( \frac{\partial H}{\partial B_0} \big|_{(A, B) = (a, b)} \) is strictly positive. If we increase \( a_{i+l} \) by \( \epsilon / w_{i+l} \) and decrease \( b_0 \) by \( \epsilon / w_0 \) for small enough \( \epsilon \) that the band remains in \( D \), then the optimality criterion is unchanged. However, by the expansion (3.14), the coverage probability changes by \( \epsilon \frac{\partial H}{\partial B_0} \big|_{(A, B) = (a, b)} / w_0 + o(\epsilon) \), which is positive once \( \epsilon \) is sufficiently small. Thus the confidence band \((a, b)\) cannot be optimal. The proof that the upper bounds \( b_0, \ldots, b_{n-1} \) are strictly ordered is completely analogous. \[ \square \]

The next theorem shows that, up to the vector of weights \((w_0, \ldots, w_n)\), the partial derivatives of \( H \) at an optimal band must all be the same in magnitude unless the corresponding bounds are equal to 0 or 1.

**Theorem 3.7.** Suppose that \((a, b)\) is an optimal confidence band with respect to the optimality criterion with weights \((w_0, \ldots, w_n)\). Then there exists a constant \( C > 0 \)}
such that for each lower bound $a_i > 0$,

$$\frac{\partial H}{\partial A_i} \big|_{(A,B)=(a,b)} = -Cw_i,$$

and for each upper bound $b_j < 1$,

$$\frac{\partial H}{\partial B_j} \big|_{(A,B)=(a,b)} = Cw_j.$$

If $a_1 = 0$, then

$$\frac{\partial H}{\partial A_1^+} \big|_{(A,B)=(a,b)} \leq -Cw_1,$$

and if $b_{n-1} = 1$, then

$$\frac{\partial H}{\partial B_{n-1}} \big|_{(A,B)=(a,b)} \geq Cw_{n-1}.$$

**Proof of Theorem 3.7.** This proof uses the same type of argument used in the proof of Theorem 3.6. Let $a_i$ be a lower bound, $i = 2, \ldots, n$. Then, since the lower bounds are strictly ordered, we can change the value of $a_i$ slightly in either direction without leaving $D$. We can also change the value of $b_0$ slightly in either direction without leaving $D$. If we decrease $a_i$ by $\epsilon/w_i$ and decrease $b_0$ by $\epsilon/w_0$ for small $\epsilon$ (positive or negative), then the optimality criterion is unchanged, but by the expansion (3.14), the coverage probability changes by

$$\epsilon \left( -\frac{\partial H}{\partial A_i} \big|_{(A,B)=(a,b)} / w_i - \frac{\partial H}{\partial B_0} \big|_{(A,B)=(a,b)} / w_0 \right) + o(\epsilon). \quad (3.15)$$

Since $(a, b)$ is an optimal confidence band, the quantity (3.15) must always be nonpositive. But since we can choose $\epsilon$ to be positive or negative, (3.15) can only always be nonpositive if the quantity in parentheses, which doesn’t depend on $\epsilon$, is equal to 0. This proves that

$$\frac{\partial H}{\partial A_i} \big|_{(A,B)=(a,b)} = - \left( \frac{\partial H}{\partial B_0} \big|_{(A,B)=(a,b)} / w_0 \right) w_i \equiv -Cw_i.$$
The case of $b_j < 1$ follows by a similar argument.

Suppose that $a_1 = 0$. Then, since the lower bounds are strictly ordered, we can increase the value of $a_1$ slightly without leaving $D$. We can also increase the value of $b_0$ slightly without leaving $D$. If we increase $a_1$ by $\epsilon/w_1$ and increase $b_0$ by $\epsilon/w_0$, then the optimality criterion is unchanged, but the coverage probability changes by

$$
\epsilon \left( \frac{\partial H}{\partial A_1} \bigg|_{(A,B)=(a,b)} / w_1 + \frac{\partial H}{\partial B_0} \bigg|_{(A,B)=(a,b)} / w_0 \right) + o(\epsilon).
$$

(3.16)

Since $(a, b)$ is an optimal confidence band, the quantity (3.16) must always be non-positive. But since we can choose $\epsilon$ to be arbitrarily small, but positive, (3.16) can only always be nonpositive if the quantity in parentheses, which doesn’t depend on $\epsilon$, is nonpositive. This proves that

$$
\frac{\partial H}{\partial A_1} \bigg|_{(A,B)=(a,b)} \leq - \left( \frac{\partial H}{\partial B_0} \bigg|_{(A,B)=(a,b)} / w_0 \right) w_1 = -C w_1.
$$

The case of $b_{n-1} = 1$ follows similarly.

Given the necessary conditions that we found in Theorem 3.7, it is natural to ask whether these conditions are not only necessary, but perhaps sufficient for a band $(a, b)$ to be optimal. It turns out that the conditions given in Theorem 3.7 are indeed sufficient for a band in the class $D$ to be optimal. Before we can show this result, however, we need to state an important result, the Brunn-Minkowski Inequality, from the theory of convex bodies. Proofs of this result are given in Leichtweiss (1998) and many other readily-accessible references. Gardner (2002) gives a particularly fascinating and readable study of the Brunn-Minkowski Inequality, its history, and its relationship to other well-known inequalities in analysis and geometry. We note here that the convex combination $(1 - \lambda)K_1 + \lambda K_2$ of the sets $K_1$ and $K_2$ is defined
by

\[(1 - \lambda)K_1 + \lambda K_2 = \{(1 - \lambda)x + \lambda y : x \in K_1, y \in K_2\}\].

**Theorem 3.8 (Brunn-Minkowski).** Let \(K_1\) and \(K_2\) be convex bodies in \(\mathbb{R}^n\), and let \(\lambda\) be a real number with \(0 < \lambda < 1\). Then if \(V_n(\cdot)\) is the volume function in \(\mathbb{R}^n\), we have that

\[V_n((1 - \lambda)K_1 + \lambda K_2) \geq ((1 - \lambda)V_n(K_1)^{1/n} + \lambda V_n(K_2)^{1/n})^n.\]  \hspace{1cm} (3.17)

If both \(V_n(K_1)\) and \(V_n(K_2)\) are positive, then equality holds in (3.17) iff

\[K_2 = \rho K_1 + t\]  \hspace{1cm} (3.18)

for \(\rho > 0\) and \(t \in \mathbb{R}^n\).

Theorem 3.8 holds even under the weaker hypothesis that \(K_1\) and \(K_2\) are merely measurable rather than convex, provided that \(\lambda K_1 + (1 - \lambda)K_2\) is also measurable (see, for example, Gardner (2002)). For our results in this chapter, however, we need only the classical version of the inequality. The Brunn-Minkowski Inequality is relevant to our situation because any confidence band \(CB(a, b)\) may be thought of as a convex set in \(\mathbb{R}^n\). Specifically, the confidence band \(CB(a, b)\) may be thought of as the set of all points \((x_1, \ldots, x_n) \in \mathbb{R}^n\) such that \(a_1 \leq x_1 \leq b_0, \ldots, a_n \leq x_n \leq b_{n-1}\), and \(x_1 \leq x_2 \leq \ldots \leq x_n\). This set \(CB(a, b)\) is a convex set because it is the intersection of a finite collection of half-spaces, which are themselves convex sets. The coverage probability of the confidence band \(CB(a, b)\), meanwhile, is just \(n!\) times the volume of \(CB(a, b) \subset \mathbb{R}^n\). The next theorem shows that the necessary conditions for optimality given in Theorem 3.7 are also sufficient for optimality.
Theorem 3.9. Suppose that \((a, b) \in D\) is a confidence band that satisfies the four conditions given in Theorem 3.7. Then \((a, b)\) minimizes \(\sum_i w_i(b_i - a_i)\) within the set of all confidence bands with coverage probability \(\alpha = H(a, b)\).

Proof of Theorem 3.9. We prove the result by contradiction. Suppose that there exists a confidence band that has the same coverage probability as \((a, b)\), but a strictly smaller value for the optimality criterion. Then, there must exist a confidence band \((c, d)\) such that \(\sum_i w_i(b_i - a_i) = \sum_i w_i(d_i - c_i)\), but \(H(c, d) > H(a, b)\). Let \(\lambda \in (0, 1)\). Then, since \(D\) is a convex set, the convex combination \((1 - \lambda)CB(a, b) + \lambda CB(c, d)\) is another confidence band in \(D\). Specifically, it is the band \(CB((1 - \lambda)a + \lambda c, (1 - \lambda)b + \lambda d)\). For this band, the value of the optimality criterion is

\[
\sum_i w_i ((1 - \lambda)b_i + \lambda d_i - ((1 - \lambda)a_i + \lambda c_i)) \\
= (1 - \lambda) \sum_i w_i(b_i - a_i) + \lambda \sum_i w_i(d_i - c_i) \\
= \sum_i w_i(b_i - a_i),
\]

which is the same, for any \(\lambda\), as the values of the criterion for the confidence bands \((a, b)\) and \((c, d)\). Consider the function

\[
G(\lambda) \equiv H ((1 - \lambda)a + \lambda c, (1 - \lambda)b + \lambda d) = n!V_n ((1 - \lambda)CB(a, b) + \lambda CB(c, d)).
\]
This function is differentiable via the chain rule since $H$ is differentiable, and we have by the Brunn-Minkowski Inequality that

$$
\frac{dG}{d\lambda} \bigg|_{\lambda=0} = n! \frac{d}{d\lambda} V_n ((1 - \lambda)CB(a, b) + \lambda CB(c, d)) \bigg|_{\lambda=0}
\geq n! \frac{d}{d\lambda} ((1 - \lambda)H(a, b)^{1/n} + \lambda H(c, d)^{1/n}) \bigg|_{\lambda=0}
\geq n! n ((1 - \lambda)H(a, b)^{1/n} + \lambda H(c, d)^{1/n})^{n-1} (H(c, d)^{1/n} - H(a, b)^{1/n}) \bigg|_{\lambda=0}
= n! (n(H(a, b)^{1/n})^{n-1}(H(c, d)^{1/n} - H(a, b)^{1/n}))
> 0.
$$

But this derivative $\frac{dG}{d\lambda} \bigg|_{\lambda=0}$ could also be computed by noting that for sufficiently small $\lambda$, we can use (3.14) to write $G(\lambda)$ in terms of the partial derivatives of $H$ at $(a, b)$. Specifically, if the vector $\delta$ of small changes in the bounds is

$$
\delta = (c_1 - a_1, \ldots, c_n - a_n, d_0 - b_0, \ldots, d_{n-1} - b_{n-1}),
$$

then (3.14) tells us that if neither a bound $a_1 = 0$ or a bound $b_{n-1} = 1$ is changed, then changing $(a, b)$ by $\epsilon \delta$ changes the coverage probability by

$$
\epsilon(-c_1 - a_1)Cw_1 - \ldots - (c_n - a_n)Cw_n + (d_0 - b_0)Cw_0 + \ldots + (d_{n-1} - b_{n-1})Cw_{n-1}) + o(\epsilon),
$$

which is just

$$
\epsilon C \left( \sum_{i=0}^{n} w_i (d_i - c_i) - \sum_{i=0}^{n} w_i (b_i - a_i) \right) + o(\epsilon) = o(\epsilon).
$$

But then $\frac{dG}{d\lambda} = 0$, whereas the Brunn-Minkowski Inequality told us that $\frac{dG}{d\lambda}$ must be positive. If a bound $a_1 = 0$ or $b_{n-1} = 1$ is changed, then the derivative $\frac{dG}{d\lambda}$ might even be negative. In either case we have a contradiction, showing that $(a, b)$ is an optimal band with coverage probability $\alpha$ with respect to the weights $(w_0, \ldots, w_n)$. 

\[\square\]
Now that necessary and sufficient conditions for a band to be optimal with respect to a vector of weights \((w_0, \ldots, w_n)\) have been determined, one may wonder whether optimal bands are unique. The Brunn-Minkowski Inequality is again the key to obtaining useful results.

**Theorem 3.10.** Suppose that \((a, b) \in D\) and \((c, d) \in D\) are confidence bands that are each optimal at level \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\). Then \((a, b)\) and \((c, d)\) are translated versions of each other.

*Proof of Theorem 3.10.* Note first that if \((a, b)\) and \((c, d)\) are different in any way, then taking any convex combination of them yields a new confidence band with the same value for the optimality criterion. By the Brunn-Minkowski Inequality, this new band must have a coverage probability strictly larger than \(\alpha\) unless the equality condition (3.18) holds. Since both \((a, b)\) and \((c, d)\) are optimal, the new band cannot have a higher coverage probability. Thus, by the equality condition in the Brunn-Minkowski Inequality, the convex sets \(CB(a, b)\) and \(CB(c, d)\) are related as \(CB(c, d) = \rho CB(a, b) + t\) for \(\rho > 0\) and \(t \in \mathbb{R}^n\). Since \(CB(c, d)\) and \(CB(a, b)\) have the same coverage probability, they have the same volume. Hence \(\rho = 1\). Thus, the confidence bands \(CB(c, d)\) and \(CB(a, b)\) are translated versions of each other. \(\Box\)

The next few results use Theorem 3.10 to show that uniqueness holds in many cases.

**Theorem 3.11.** Suppose that \((a, b)\) is a confidence band that is optimal at level \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\). If \((a_i, b_{i-1})\) overlaps with \((a_{i+1}, b_i)\) for \(i = 1, \ldots, n - 1\), or in other words if

\[ a_{i+1} < b_{i-1}, i = 1, \ldots, n - 1, \]

(3.19)
then \((a,b)\) is the unique optimal band up to translation of every bound by the same real constant.

**Proof of Theorem 3.11.** Recall that \(CB(a,b)\) is the set of all values \((x_1, \ldots, x_n) \in \mathbb{R}^n\) satisfying \(a_1 \leq x_1 \leq b_0, \ldots, a_n \leq x_n \leq b_{n-1}\), and \(x_1 \leq x_2 \leq \ldots \leq x_n\). Suppose that we translate the vector \((x_1, \ldots, x_n)\) by \(t = (t_1, \ldots, t_n)\), obtaining a new vector of variables \(y_i = x_i + t_i, i = 1, \ldots, n\). Then, if the vector \((x_1, \ldots, x_n)\) satisfies \(a_i \leq x_i \leq b_{i-1}, i = 1, \ldots, n\), the vector \((y_1, \ldots, y_n)\) satisfies \(a_i + t_i \leq y_i \leq b_{i-1} + t_i, i = 1, \ldots, n\). However, even if \((x_1, \ldots, x_n)\) satisfies \(x_i \leq x_{i+1}, i = 1, \ldots, n-1\), the vector \((y_1, \ldots, y_n)\) only satisfies \(y_i - t_i \leq y_{i+1} - t_{i+1}, i = 1, \ldots, n-1\), or \(y_i \leq y_{i+1} + (t_i - t_{i+1}), i = 1, \ldots, n-1\). The constraints that \((y_1, \ldots, y_n)\) satisfies do not define a confidence band in \(D\) unless for each \(i = 1, \ldots, n-1\), either \(t_i = t_{i+1}\) or the condition \(x_i \leq x_{i+1}\) is redundant in that it is implied by the rectangular constraints on \(x_i\) and \(x_{i+1}\). If the interval \((a_i, b_{i-1})\) overlaps with the interval \((a_{i+1}, b_i)\) for each \(i = 1, \ldots, n-1\), then none of the conditions \(x_i \leq x_{i+1}, i = 1, \ldots, n-1\) are redundant, meaning that \(t_1 = \ldots = t_n\) must hold.

**Corollary 3.2.** Suppose that \((a,b)\) is a confidence band that is optimal at level \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\). If \(w_0\) and \(w_n\) are different and condition (3.19) is satisfied, then \((a,b)\) is the unique minimizer of \(\sum_i w_i (b_i - a_i)\) among confidence bands in \(D\) with coverage probability \(\alpha\).

**Proof of Corollary 3.2.** Since the condition (3.19) is satisfied, every other optimal band must arise from adding a constant value \(s\) to each of the bounds for \((a,b)\). But if a constant \(s\) is added to each of the bounds, then the value of the optimality
criterion becomes
\[
    w_0(b_0 + s) + \sum_{i=1}^{n-1} w_i(b_i + s - (a_i + s)) + w_n(1 - (a_n + s))
\]
\[
= \sum_i w_i(b_i - a_i) + s(w_0 - w_n)
\]
(3.20)
since the values \(a_0 = 0\) and \(b_n = 1\) are fixed. This value (3.20) can only be the same as \(\sum_i w_i(b_i - a_i)\) if \(w_0 = w_n\). The result thus follows.

**Corollary 3.3.** Suppose that \((a, b)\) is a confidence band that is optimal at level \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\), and suppose that condition (3.19) is satisfied. If \(a_1 = 0\) and \(b_{n-1} = 1\), then \((a, b)\) is the unique minimizer of \(\sum_i w_i(b_i - a_i)\) among confidence bands in \(D\) with coverage probability \(\alpha\).

**Proof of Corollary 3.3.** This result follows immediately by noticing that if any positive constant \(s\) is added to \(b_{n-1} = 1\) or subtracted from \(a_1 = 0\), the resulting set of bounds is not located in the interval \([0, 1]\). \qed

**Corollary 3.4.** Suppose that \((a, b)\) is a confidence band that is optimal at level \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\), and suppose that condition (3.19) is satisfied. If \((a, b)\) is a symmetric confidence band in the sense that \(a_i + b_{n-i} = 1, i = 1, \ldots, n\), then there is no other optimal confidence band that is symmetric.

**Proof of Corollary 3.4.** To prove this result, it suffices to note that translating all bounds by the same constant \(s\) would destroy the symmetry of the band \((a, b)\). Hence \((a, b)\) is the unique symmetric optimizer. \qed

The final theoretical result in this section, which is given next, tells us that if we use a vector of weights \((w_0, \ldots, w_n)\) that is symmetric in the sense that \(w_i = w_{n-i}\) for
\[ i = 0, \ldots, n, \text{ then we are guaranteed that there exists a symmetric optimal confidence} \]

band at any desired level.

**Theorem 3.12.** Suppose that \((w_0, \ldots, w_n)\) is a vector of positive weights satisfying

the symmetry condition \(w_i = w_{n-i}, i = 0, \ldots, n, \) and suppose that \(\alpha \in (0, 1)\) is the

desired coverage probability. Then there must exist a confidence band \((a, b) \in D\) that

is both symmetric in the sense that \(a_i + b_{n-i} = 1, i = 1, \ldots, n, \) and optimal at coverage

probability \(\alpha\) with respect to the weights \((w_0, \ldots, w_n)\).

**Proof of Theorem 3.12.** By Theorem 3.5, there must exist an optimal band \((c, d) \in D\). Consider the band \((e, f)\) with lower bounds \((1 - d_{n-1}, \ldots, 1 - d_0)\) and upper bounds

\((1 - c_n, \ldots, 1 - c_1)\). Since order statistics \(U_{(1)}, \ldots, U_{(n)}\) from the standard uniform distribution satisfy

\[
(U_{(1)}, \ldots, U_{(n)}) \overset{d}{=} (1 - U_{(n)}, \ldots, 1 - U_{(1)}),
\]

the coverage probability of the band \((e, f)\) is exactly the same as the coverage probability of the band \((c, d)\). Moreover, because of the symmetry in the weights \((w_0, \ldots, w_n)\), the optimality criterion for the new band is

\[
\sum_i w_i (1 - c_{n-i} - (1 - d_{n-i})) = \sum_i w_{n-i} (d_{n-i} - c_{n-i}) = \sum_i w_i (d_i - c_i),
\]

the same as for the band \((c, d)\). Thus, the band \((e, f)\) is also optimal. Now construct

a new band by taking the convex combination \(\frac{1}{2}CB(c, d) + \frac{1}{2}CB(e, f)\). This new band

is symmetric by construction, and the Brunn-Minkowski Inequality shows that it is

optimal. \(\square\)

Through the theoretical results obtained in this section, we know that optimal confidence bands exist, that they are unique under certain conditions, and that they
can be recognized as optimal based on the values of the partial derivatives of the coverage probability function $H$ with respect to the bounds $A_1, \ldots, A_n$ and $B_0, \ldots, B_{n-1}$.

What is still lacking, however, is a computational algorithm for finding optimal bands given $\alpha$ and the weights $(w_0, \ldots, w_n)$. Given below is a computational procedure which, though crude, works in practice. The motivation for this algorithm is the observation that one can realize the partial derivative conditions required by Theorem 3.7 if one relaxes bounds that lead to high values of the quotients $\left| \frac{\partial H}{\partial A_i} / w_i \right|$ or $\left| \frac{\partial H}{\partial B_j} / w_j \right|$ and tightens bounds that lead to low values of those quotients. Note that in this particular optimization problem, it is impossible to have a locally optimal band that is not also a globally optimal band, which removes one of the usual concerns in an optimization problem.

**Computational Algorithm:**

*Step 1:* Use bisection to find the Kolmogorov-Smirnov confidence band with the desired coverage probability $\alpha$. This Kolmogorov-Smirnov confidence band will serve as a starting point from which we will work towards a better band.

*Step 2:* Specify a small starting step size, say $h = (4n)^{-1}$, which gives the size of the moves that will be made initially.

*Step 3:* Compute the values at the current confidence band of the partial derivatives of $H$ with respect to the lower and upper bounds. Update these partial derivatives whenever the current confidence band is changed in any way. Computational time can be saved by noticing that changing one bound will not noticeably change derivatives other than those corresponding to bounds in the immediate vicinity of the bound that was changed.
Step 4: When the current coverage probability is less than or equal to $\alpha$, increase the coverage probability of the band by increasing (if it is an upper bound) or decreasing (if it is a lower bound) by $h$ the moveable bound with the highest value of the quotient $\left| \frac{\partial H}{\partial A_i} / w_i \right|$ or $\left| \frac{\partial H}{\partial B_j} / w_j \right|$. By a moveable bound we mean any upper or lower bound for which making a change of size $h$ in the specified direction would not take the confidence band out of the space $D$.

Step 5: When the current coverage probability is greater than $\alpha$, decrease the coverage probability by decreasing (if it is an upper bound) or increasing (if it is a lower bound) by $h$ the moveable bound with the lowest value of the quotient $\left| \frac{\partial H}{\partial A_i} / w_i \right|$ or $\left| \frac{\partial H}{\partial B_j} / w_j \right|$.

Step 6: If making the most recent move has returned the band to a band already visited at the current step size $h$, then decrease $h$ by a factor of 2 to avoid entering an infinite loop.

Step 7: Repeat Steps 4 to 6, keeping Step 3 in mind, until a prespecified stopping criterion is met. The stopping criterion that we use is simply the size of $h$. However, another reasonable stopping criterion could be based on the difference between the maximum and minimum values of $\left| \frac{\partial H}{\partial A_i} / w_i \right|$ or $\left| \frac{\partial H}{\partial B_j} / w_j \right|$ over the set of all moveable bounds.

For the case of symmetric weights $(w_0, \ldots, w_n)$, Theorem 3.12 guarantees that a symmetric optimal confidence band will exist. Thus, if one has symmetric weights and is seeking a symmetric band, it suffices to work with only the lower bounds in the algorithm, taking the upper bounds to be given by $b_i = 1 - a_{n-i}, i = 0, \ldots, n$. 

69
3.4 Shape and Power Comparisons for the Confidence Band Procedures

Though narrowness of the confidence band was the motivation for proposing the optimality criteria discussed in Section 3.3, it is of interest to compare the performance of the optimal bands to that of other confidence bands both in terms of narrowness and in terms of the power of the associated hypothesis test. The confidence bands chosen for comparison with the optimal bands were (a) the basic Kolmogorov-Smirnov band, (b) the Anderson-Darling band based on the weighted Kolmogorov-Smirnov statistic

\[ D_{AD} = \sup_t \frac{|F_n(t) - F_0(t)|}{\sqrt{F_0(t)(1 - F_0(t))}} \]

and (c) the nonparametric likelihood confidence bands proposed by Owen (1995). Optimal bands based on three different choices of weights were used in the comparison. The first choice of weights was uniform weights \((1, \ldots, 1)\). The second choice of weights, made to emphasize narrowness of the band in the extremes, set \(w_i = 100\) for the outermost \(n/5\) intervals on either extreme and \(w_i = 1\) for the intervals in the middle. The final choice of weights, made to de-emphasize narrowness of the band in the extremes, set \(w_i = 1\) for the outermost \(n/5\) intervals on either extreme and \(w_i = 100\) for the intervals in the middle.

Symmetric confidence bands with coverage probability 95% were produced for sample sizes \(n = 10, n = 20,\) and \(n = 40\). These confidence bands are plotted for the case of equispaced data in Figures 3.1 through 3.3. A number of important features of both the optimal bands and their competitors are evident from these plots. In looking at the plots for the three competitors, one sees that the Kolmogorov-Smirnov band is centered at the EDF and has a uniform half-width, that the Anderson-Darling band is
so wide throughout its range that one almost suspects that the coverage probability has been computed incorrectly, and that the nonparametric likelihood band has a roughly elliptical shape. In looking at the three optimal bands in each figure, one sees that different choices of weights can lead to substantially different optimal bands.

To compare the power properties of the optimal bands to those of the three competitor bands, a class of alternative distributions that includes both heavy-tailed and light-tailed distributions was used. The null hypothesis distribution was taken to be the standard uniform distribution, and the alternatives considered were the $Beta(\eta, \eta)$ distributions for $\eta \in (0, 2)$. When $\eta$ is small, the distribution $Beta(\eta, \eta)$ has heavy tails, while when $\eta$ is large, the tails are light. The case $\eta = 1$, of course, gives the standard uniform distribution. Plots of power versus the parameter $\eta$ for the 95% confidence bands plotted in Figures 3.1 through 3.3 are given in Figures 3.4 through 3.6. One fact immediately evident from these plots is that all of the confidence bands are associated with biased tests when the sample size is small. As the sample size increases, the nonparametric likelihood band seems to have the best overall power performance, with the performance of the optimal band with weights emphasizing the extremes being comparable. The Anderson-Darling band has the best power against heavy-tailed alternatives, but it accomplishes this at the cost of having almost no power against light-tailed alternatives. Overall, the optimal confidence bands seem to be comparable to the best of their competitors in terms of power, and the fact that the weights can be changed allows flexibility to produce a band tailored to a particular kind of alternative.

One strategy for choosing the weights $(w_0, \ldots, w_n)$ to give good power against a particular alternative might proceed as follows. One could identify the region of the
null hypothesis distribution $F_0$ where the absolute difference $|F_0(t) - F_1(t)|$ between $F_0$ and the alternative distribution $F_1$ is greatest. One could then try to make the band narrow in that region by assigning high weights to intervals involving order statistics likely to fall in that region. There are limits, however, to how successful such a strategy can be. These limits arise from the fact that no confidence band can be optimal for any choice of the weights $(w_0, \ldots, w_n)$ unless it satisfies the condition

$$\frac{\partial H}{\partial A_i} = \frac{\partial H}{\partial B_i} \text{ for all } i \text{ with } 0 < a_i < b_i < 1.$$ 

Thus, most powerful bands for testing $H_0 : F = F_0$ versus $H_1 : F = F_1$ will typically not fall in the class of optimal bands.

### 3.5 Optimal Confidence Bands in the Ranked-set Sampling Case

In this section, we first present an example that highlights the need to use conditional confidence bands when producing a confidence band based on a ranked-set sample. We then discuss how to generalize the optimal bands of Section 3.3 to ranked-set samples and other samples consisting of independent order statistics.

**Example 3.2:** Suppose that we draw a ranked-set sample of size 2 consisting of a value $X_{1:2}$ distributed as the minimum of a simple random sample of size 2 and an independent value $X_{2:2}$ distributed as the maximum of a simple random sample of size 2. The unconditional approach to producing a confidence band, which was proposed by Stokes and Sager (1988), consists of computing the RSS empirical distribution function (EDF) $F^*$ and using a uniform-width confidence band centered at the $F^*$. If the half-width for this confidence band is $1/2$, then the lower and upper bounds of the band are $(a_0, a_1, a_2) = (0, 0, 1/2)$ and $(b_0, b_1, b_2) = (1/2, 1, 1)$. The coverage
Figure 3.1: Six different 95% confidence bands when $n = 10$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
Figure 3.2: Six different 95% confidence bands when $n = 20$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
Figure 3.3: Six different 95% confidence bands when $n = 40$. The bands are (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
Figure 3.4: Power curves for the 95% confidence bands when $n = 10$. The alternative is the $Beta(\eta, \eta)$ distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
Figure 3.5: Power curves for the 95% confidence bands when $n = 20$. The alternative is the $Beta(\eta, \eta)$ distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
Figure 3.6: Power curves for the 95% confidence bands when \( n = 40 \). The alternative is the \( \text{Beta}(\eta, \eta) \) distribution. The plots are for (a) the Kolmogorov-Smirnov band, (b) the Anderson-Darling band, (c) Owen’s nonparametric likelihood band, (d) the uniform-weight optimal band, (e) an optimal band with weights emphasizing the extremes, and (f) an optimal band with weights de-emphasizing the extremes.
probability is

\[ P(0 \leq \min\{X_{1:2}, X_{2:2}\} \leq 1/2, 1/2 \leq \max\{X_{1:2}, X_{2:2}\} \leq 1) = 0.625. \]

Suppose, however, that we condition on the ordering of the independent order statistics \(X_{1:2}\) and \(X_{2:2}\). If \(X_{1:2} < X_{2:2}\), which occurs \(5/6\) of the time, then the conditional coverage probability is

\[
\frac{P(0 \leq X_{1:2} \leq 1/2, 1/2 \leq X_{2:2} \leq 1)}{P(X_{1:2} \leq X_{2:2})} = \frac{(3/4)^2}{5/6} = 0.675.
\]

If the opposite ordering \(X_{2:2} < X_{1:2}\) occurs, however, then the conditional coverage probability is

\[
\frac{P(0 \leq X_{2:2} \leq 1/2, 1/2 \leq X_{1:2} \leq 1)}{P(X_{2:2} \leq X_{1:2})} = \frac{(1/4)^2}{1/6} = 0.375.
\]

There is thus a difference of 0.3 between the conditional coverage probability given one observed sequence of order statistics and the conditional coverage probability given the other ordering. This suggests that it would be inappropriate not to condition on the information available in the observed sequence of order statistics when presenting a confidence band.

In order to compute coverage probabilities for confidence bands based on ranked-set samples or other collections of independent order statistics, we need a more general computational algorithm than the one that we applied in the case of a simple random sample. The needed algorithm is provided in Section 2.3, and we describe its relationship to coverage probabilities here. Suppose that independent order statistics \(X_{k_1:m_1}, \ldots, X_{k_n:m_n}\) are observed from a parent distribution with continuous CDF \(F\). Here the notation \(X_{k_i:m_i}\) indicates an observation distributed like the \(k_i\)th order
standard uniform parent distribution, we can rewrite (3.21) as

\[ P_F(F \in CB(a, b) | X_{k_1:m_1} \leq \ldots \leq X_{k_n:m_n}) \]

\[ = P_F(a_1 \leq F(X_{k_1:m_1}) \leq b_0, \ldots, a_n \leq F(X_{k_n:m_n}) \leq b_{n-1}, X_{k_1:m_1} \leq \ldots \leq X_{k_n:m_n}) \]

\[ = \frac{P_F(a_1 \leq F(X_{k_1:m_1}) \leq b_0, \ldots, a_n \leq F(X_{k_n:m_n}) \leq b_{n-1}, X_{k_1:m_1} \leq \ldots \leq X_{k_n:m_n})}{P_F(X_{k_1:m_1} \leq \ldots \leq X_{k_n:m_n})} \]

\[ (3.21) \]

Using the fact that \( F(X_{k_i:m_i}) \stackrel{d}{=} U_{k_i:m_i} \), where \( U_{k_i:m_i} \) is an order statistic from the standard uniform parent distribution, we can rewrite (3.21) as

\[ \frac{P(a_1 \leq U_{k_1:m_1} \leq b_0, \ldots, a_n \leq U_{k_n:m_n} \leq b_{n-1}, U_{k_1:m_1} \leq \ldots \leq U_{k_n:m_n})}{P(0 \leq U_{k_1:m_1} \leq 1, \ldots, 0 \leq U_{k_n:m_n} \leq 1, U_{k_1:m_1} \leq \ldots \leq U_{k_n:m_n})}. \]

\[ (3.22) \]

This last expression shows that the conditional coverage probability of the band \( CB(a, b) \) is distribution-free, and it suggests that all that is needed for computing coverage probabilities is an algorithm for computing probabilities of the form

\[ P(a_1 \leq U_{k_1:m_1} \leq b_0, \ldots, a_n \leq U_{k_n:m_n} \leq b_{n-1}, U_{k_1:m_1} \leq \ldots \leq U_{k_n:m_n}) \]

\[ (3.23) \]

given the bounds \((a, b) \in D\). Exactly such an algorithm was developed in Section 2.3.
Given that we can compute the probabilities (3.23) necessary for computing conditional coverage probabilities, we can again define optimality criteria \( \sum_i w_i (b_i - a_i) \) and seek to find the band that minimizes \( \sum_i w_i (b_i - a_i) \) among all confidence bands with coverage probability \( \alpha \) conditional on the observed sequence of order statistics. As in Section 3.3, the partial derivatives of the conditional coverage probability \( H \) with respect to each of the bounds are what allow us to recognize and compute optimal bands. Setting

\[
P_d = P(0 \leq U_{k_1:m_1} \leq 1, \ldots, 0 \leq U_{k_n:m_n} \leq 1, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n}),
\]

we obtain the following result.

**Theorem 3.13.** Whenever a small change in the appropriate bound does not take \((a, b)\) out of the space \( D \), the one-sided partial derivatives of the conditional coverage probability \( H \) with respect to each of the bounds \( \{a_i, b_j\} \) are given by

\[
\frac{\partial H}{\partial A^+_i} \bigg|_{(A,B)=(a,b)} = -\frac{f_{k_i;m_i}(a_i)}{P_d} \int_{a_1}^{b_0 \land a_i} \cdots \int_{a_{i-1}}^{b_{i-2} \land a_i} \int_{a_{i+1}}^{b_i} \cdots \int_{a_n}^{b_{n-1}} \left( \prod_{j \neq i} f_{k_j;m_j}(u_j) \right) \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) I(u_1 \leq \cdots u_{i-1} \leq u_{i+1} \leq \cdots \leq u_n) du_n \cdots du_{i+1} du_{i-1} \cdots du_1,
\]

\[
\frac{\partial H}{\partial A^-_i} \bigg|_{(A,B)=(a,b)} = -\frac{f_{k_i;m_i}(a_i)}{P_d} \int_{a_1}^{b_0 \land a_i} \cdots \int_{a_{i-1}}^{b_{i-2} \land a_i} \int_{a_{i+1}}^{b_i} \cdots \int_{a_n}^{b_{n-1}} \left( \prod_{j \neq i} f_{k_j;m_j}(u_j) \right) \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) I(u_1 \leq \cdots u_{i-1} \leq u_{i+1} \leq \cdots \leq u_n) du_n \cdots du_{i+1} du_{i-1} \cdots du_1,
\]

\[
\frac{\partial H}{\partial B^+_j} \bigg|_{(A,B)=(a,b)} = \frac{f_{k_j;m_j}(b_j)}{P_d} \int_{a_1}^{b_0} \cdots \int_{a_j}^{b_{j-1}} \int_{a_{j+2} \land b_j}^{b_{j+1}} \cdots \int_{a_n \land b_j}^{b_{n-1}} \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) I(u_1 \leq \cdots u_{j-1} \leq u_{j+1} \leq \cdots \leq u_n) du_n \cdots du_{j+1} du_{j-1} \cdots du_1,
\]

and

\[
\frac{\partial H}{\partial B^-_j} \bigg|_{(A,B)=(a,b)} = \frac{f_{k_j;m_j}(b_j)}{P_d} \int_{a_1}^{b_0} \cdots \int_{a_j}^{b_{j-1}} \int_{a_{j+2} \land b_j}^{b_{j+1}} \cdots \int_{a_n \land b_j}^{b_{n-1}} \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) \left( \prod_{i \neq j} f_{k_i;m_i}(u_i) \right) I(u_1 \leq \cdots u_{j-1} \leq u_{j+1} \leq \cdots \leq u_n) du_n \cdots du_{j+1} du_{j-1} \cdots du_1,
\]
where \( f_{k:m}(x) \) is the probability density function for a Beta\((k, m+1-k)\) distribution. Thus, whenever a bound \( a_i \) or \( b_j \) can be moved at least a small amount in either direction without taking \((a, b)\) out of \(D\), a regular partial derivative exists.

With the help of this result, which generalizes our calculations in Section 3.3, most of the theoretical results we obtained in Section 3.3, including Theorems 3.3, 3.7, and 3.9, may be found to hold in this more general case in which the sample consists of independent order statistics. However, because the probability (3.23) cannot be interpreted in terms of volume in \(\mathbb{R}^n\), the Brunn-Minkowski Inequality is no longer sufficient to give the results we want. A result that is sufficient for our purposes is the Prékopa-Leindler Inequality (see, for example, Gardner (2002)), which generalizes the Brunn-Minkowski Inequality. A similar result was discussed and proved by Rinott (1976).

**Theorem 3.14 (Prékopa-Leindler).** Let \( \lambda \in (0, 1) \), and let \( f, g, \) and \( h \) be nonnegative integrable functions on \(\mathbb{R}^n\) satisfying

\[
h((1 - \lambda)x + \lambda y) \geq f(x)^{1-\lambda}g(y)^\lambda
\]  

(3.24)

for all \( x, y \in \mathbb{R}^n \). Then

\[
\int_{\mathbb{R}^n} h(x)dx \geq \left( \int_{\mathbb{R}^n} f(x)dx \right)^{1-\lambda} \left( \int_{\mathbb{R}^n} g(x)dx \right)^\lambda.
\]

The following corollary then follows. Here \( \Omega \) is the support set \(\{(x_1, \ldots, x_n) : 0 \leq x_1 \leq \cdots \leq x_n \leq 1\}\) for the standard uniform order statistics.

**Corollary 3.5.** Let \( \lambda \in (0, 1) \), and let \( \mu \) be a probability measure on \(\Omega\) with a log concave density \( d \). Let \( K_1 \) and \( K_2 \) be convex subsets of \(\Omega\). Then

\[
\mu((1 - \lambda)K_1 + \lambda K_2) \geq \mu(K_1)^{1-\lambda} \mu(K_2)^\lambda.
\]
Proof of Corollary 3.5. Set \( h(x) = d(x)I(x \in ((1 - \lambda)K_1 + \lambda K_2)) \), \( f(x) = d(x)I(x \in K_1) \), and \( g(x) = d(x)I(x \in K_2) \). Consider the condition (3.24). If either \( x \notin K_1 \) or \( y \notin K_2 \), then the condition holds immediately since the right-hand side is 0. If \( x \in K_1 \) and \( y \in K_2 \), then \((1 - \lambda)x + \lambda y \in ((1 - \lambda)K_1 + \lambda K_2) \) by the definition of the set \((1 - \lambda)K_1 + \lambda K_2 \). Thus, we have that

\[
h((1 - \lambda)x + \lambda y) = d((1 - \lambda)x + \lambda y) \geq d(x)^{1-\lambda}d(y)^{\lambda} = f(x)^{1-\lambda}g(y)^{\lambda},
\]

where the inequality follows from the fact that \( d \) is log concave. Noting that

\[
\int_{\mathbb{R}^n} h(x)dx = \mu((1 - \lambda)K_1 + \lambda K_2),
\]

\[
\int_{\mathbb{R}^n} f(x)dx = \mu(K_1), \text{ and}
\]

\[
\int_{\mathbb{R}^n} g(x)dx = \mu(K_2)
\]

and applying the theorem then proves the corollary. \( \Box \)

To see the relevance of this corollary, note that the joint conditional density of \( U_{k_1:m_1}, \ldots, U_{k_n:m_n} \) given the ordering \( U_{k_1:m_1} \leq \ldots \leq U_{k_n:m_n} \) is given by

\[
f(u_{k_1:m_1}, \ldots, u_{k_n:m_n}) \propto \left( \prod_{i=1}^n u_{k_i:m_i}^{k_i-1}(1 - u_{k_i:m_i})^{m_i-k_i} \right) I(0 \leq u_{k_1:m_1} \leq \ldots \leq u_{k_n:m_n} \leq 1).
\]

Taking logarithms, we have that

\[
\log f(u_{k_1:m_1}, \ldots, u_{k_n:m_n}) = C + \sum_{i=1}^n (k_i \log(u_{k_i:m_i}) + (m_i - k_i) \log(1 - u_{k_i:m_i}))
\]

in the support set \( \Omega \). Since each of the functions \( \log(u_{k_i:m_i}), i = 1, \ldots, n \) and \( \log(1 - u_{k_i:m_i}), i = 1, \ldots, n \) is concave, it follows that \( \log f \) is concave.

To show how the corollary may be applied, we prove a generalization of Theorem 3.9. The proof of the analog to Theorem 3.7 holds exactly as before.
Theorem 3.15. Suppose that \((a,b) \in D\) is a conditional confidence band whose derivatives satisfy the four conditions given in Theorem 3.7. Then \((a,b)\) minimizes \(\sum_i w_i(b_i-a_i)\) within the set of all conditional confidence bands with coverage probability \(\alpha = H(a,b)\). The event that we condition on, of course, is the observed sequence of order statistics.

Proof of Theorem 3.15. We proceed by contradiction as in the proof of Theorem 3.9. Suppose that there exists a conditional confidence band that has the same coverage probability as \((a,b)\), but a strictly smaller value for the optimality criterion. Then there must exist a conditional confidence band \((c,d)\) such that \(\sum_i w_i(b_i-a_i) = \sum_i w_i(d_i-c_i)\), but \(H(c,d) > H(a,b)\). Let \(\lambda \in (0, 1)\). Then, since \(D\) is a convex set, the convex combination \((1-\lambda)CB(a,b) + \lambda CB(c,d)\) is another conditional confidence band in \(D\). Specifically, it is the band \(CB(((1-\lambda)a + \lambda c, (1-\lambda)b + \lambda d)\). For this band, the value of the optimality criterion is exactly the same as for the bands \(CB(a,b)\) and \(CB(c,d)\).

Consider the function

\[
G(\lambda) \equiv H((1-\lambda)a + \lambda c, (1-\lambda)b + \lambda d) = \frac{1}{P_d}\mu((1-\lambda)CB(a,b) + \lambda CB(c,d)),
\]

where \(\mu\) is the measure determined by the joint distribution of the observed order statistics, conditional on their sequence. As in the proof of Theorem 3.9, writing \(\frac{dG}{d\lambda}\big|_{\lambda=0}\) in terms of the partial derivatives of \(H\) allows us to show that \(\frac{dG}{d\lambda}\big|_{\lambda=0} \leq 0\). If we compute the \(\frac{dG}{d\lambda}\big|_{\lambda=0}\) via the chain rule, however, we get a different result.
Specifically, we have by the corollary that

\[
\frac{dG}{d\lambda} \bigg|_{\lambda=0} \propto \frac{d}{d\lambda} \mu((1 - \lambda)CB(a, b) + \lambda CB(c, d)) \bigg|_{\lambda=0} \\
\geq \frac{d}{d\lambda} \left\{ \left( \mu(CB(a, b)) \right)^{1-\lambda} \left( \mu(CB(c, d)) \right)^{\lambda} \right\} \bigg|_{\lambda=0} \\
\geq (\log \mu(CB(c, d)) - \log \mu(CB(a, b))) \cdot \\
\left\{ \left( \mu(CB(a, b)) \right)^{1-\lambda} \left( \mu(CB(c, d)) \right)^{\lambda} \right\} \bigg|_{\lambda=0} \\
> 0.
\]

This gives a contradiction, proving the result.

Since the conditions under which equality holds in Theorem 3.14 are not yet known, we are not in as strong a position to derive uniqueness results as we were in Section 3.3. It seems clear, however, that using a measure $\mu$ that does not correspond to volume in $\mathbb{R}^n$ (i.e., a nonuniform measure) can only make it more difficult for equality to hold in the Prékopa-Leindler Inequality than in the Brunn-Minkowski Inequality. Thus, we suspect that the analogues of Theorem 3.11, Corollary 3.2, and Corollary 3.3 hold for the conditional confidence bands developed in this section. We conclude this section with a simple example illustrating the point that uniqueness may be more easily obtained with a nonuniform measure.

**Example 3.3:** Suppose that we observe a sample consisting of a single observation that is the maximum from a sample of size two. Identifying a conditional confidence band then reduces to choosing the values $b_0, a_1 \in [0, 1]$, and the coverage probability can be written as $P(a_1 \leq U_{2:2} \leq b_0)$, where $U_{2:2}$ has a $Beta(2, 1)$ distribution. Simple integration shows that the coverage probability is given by $b_0^2 - a_1^2$, and the value for the optimality criterion is $w_0 b_0 + w_1 (1 - a_1)$.
Suppose that we seek a band with coverage probability \( \alpha \in (0, 1) \) and that we choose weights \( w_0, w_1 \) such that \( w_1/w_0 = r \). The fact that \( \alpha = b_0^2 - a_1^2 \) tells us that \( a_1 = \sqrt{b_0^2 - \alpha} \), and minimizing the optimality criterion becomes equivalent to minimizing the function \( G(b_0) = b_0 - r\sqrt{b_0^2 - \alpha} \) over the range \([\sqrt{\alpha}, 1]\) of possible values for \( b_0 \). Differentiation gives us that

\[
G'(b_0) = \frac{dG}{db_0} = 1 - rb_0(b_0^2 - \alpha)^{-1/2}
\]

on the interval \((\sqrt{\alpha}, 1]\). Suppose that \( r \geq \sqrt{1-\alpha} \). Then, since \( b_0/\sqrt{b_0^2 - \alpha} \) achieves its minimum value of \( 1/\sqrt{1-\alpha} \) when \( b_0 = 1 \), we have that

\[
G'(b_0) = 1 - \frac{rb_0}{\sqrt{b_0^2 - \alpha}} \leq 1 - \frac{1}{\sqrt{1-\alpha}} \frac{b_0}{\sqrt{b_0^2 - \alpha}} \leq 1 - \frac{1}{\sqrt{1-\alpha}} = 0,
\]

which implies that \( G \) is minimized when \( b_0 = 1 \). Suppose on the other hand that \( r < \sqrt{1-\alpha} \). The equation \( G'(b_0) = 0 \) has a single positive solution \( b_0 = \sqrt{\frac{r}{1-r}} \), with \( G' \) being negative for smaller values and positive for larger values. Thus, the minimum value of \( G \) occurs when \( b_0 = \sqrt{\frac{r}{1-r}} \).

We thus see that in this example, an optimal band is always unique. This may be contrasted with the simple random sampling situation, where when \( n = 1 \) and \( r = 1 \), every single band with level \( \alpha \) is optimal.

### 3.6 A Comparison of RSS-based Confidence Bands

To compare the performance of the optimal conditional confidence bands based on RSS to that of the confidence bands described by Stokes and Sager (1988), we considered two situations in which the number of possible observed sequences of order statistics is managably small. These situations are (i) when a single cycle of size three is drawn and (ii) when two cycles of size two are drawn. In each of these
two situations, the total number of possible observed sequences of order statistics is 6. Figure 3.7 shows optimal conditional 95% confidence bands based on uniform weights for each of the 6 possible orderings of the order statistics \(U_{1:3}, U_{2:3}, \text{ and } U_{3:3}\). Figure 3.8, meanwhile, shows the analogous optimal bands for the second situation. Since the entire justification for using conditional confidence bands is that, given the observed sequence of order statistics, the data will depart from evenly-spaced data in predictable ways, the data values used on the horizontal axis of each plot are the expected values of the ordered values \(U_{k_1:m_1} \leq \ldots \leq U_{k_n:m_n}\) given the observed sequence of order statistics and assuming that the parent distribution is the standard uniform distribution.

Figures 3.7 and 3.8 show that when the observed order statistics are in their natural sequence, the optimal uniform-weight band tends to be narrow on the extremes. When the observed order statistics do not appear in the natural ordering, in contrast, the optimal uniform-weight band tends to be relatively narrow in the middle.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Optimality Criterion for Unconditional K-S band</th>
<th>Conditional Coverage Probability</th>
<th>Optimality Criterion for Conditional Optimal Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2,3)</td>
<td>3.03</td>
<td>0.966</td>
<td>2.94</td>
</tr>
<tr>
<td>(1,3,2)</td>
<td>3.03</td>
<td>0.944</td>
<td>2.96</td>
</tr>
<tr>
<td>(2,1,3)</td>
<td>3.03</td>
<td>0.944</td>
<td>2.96</td>
</tr>
<tr>
<td>(2,3,1)</td>
<td>3.03</td>
<td>0.867</td>
<td>2.99</td>
</tr>
<tr>
<td>(3,1,2)</td>
<td>3.03</td>
<td>0.867</td>
<td>2.99</td>
</tr>
<tr>
<td>(3,2,1)</td>
<td>3.03</td>
<td>0.819</td>
<td>2.95</td>
</tr>
</tbody>
</table>

Table 3.3: Values of the optimality criterion \(\sum_i (b_i - a_i)\) for 95% confidence bands when the set size is 3 and 1 cycle of observations is measured
Figure 3.7: Optimal conditional 95% confidence bands based on RSS when the set size is 3 and 1 cycle of observations is measured. Each band is plotted against data taking on the expected values of the respective order statistics given the observed sequence of order statistics. Individual plots are labeled by the observed sequence of order statistics.
Figure 3.8: Optimal conditional 95% confidence bands based on RSS when the set size is 2 and 2 cycles of observations are measured. Each band is plotted against data taking on the expected values of the respective order statistics given the observed sequence of order statistics. Individual plots are labeled by the observed sequence of order statistics.
Optimality Criterion

Table 3.4: Values of the optimality criterion $\sum_i (b_i - a_i)$ for 95% confidence bands when the set size is 2 and 2 cycles of observations are measured.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Optimality Criterion for Unconditional K-S band</th>
<th>Conditional Coverage Probability</th>
<th>Optimality Criterion for Conditional Optimal Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,2,2)</td>
<td>3.74</td>
<td>0.964</td>
<td>3.64</td>
</tr>
<tr>
<td>(1,2,1,2)</td>
<td>3.74</td>
<td>0.943</td>
<td>3.68</td>
</tr>
<tr>
<td>(1,2,2,1)</td>
<td>3.74</td>
<td>0.925</td>
<td>3.69</td>
</tr>
<tr>
<td>(2,1,1,2)</td>
<td>3.74</td>
<td>0.925</td>
<td>3.69</td>
</tr>
<tr>
<td>(2,1,2,1)</td>
<td>3.74</td>
<td>0.903</td>
<td>3.72</td>
</tr>
<tr>
<td>(2,2,1,1)</td>
<td>3.74</td>
<td>0.876</td>
<td>3.69</td>
</tr>
</tbody>
</table>

In Tables 3.3 and 3.4, the conditional optimal 95% confidence bands of Figures 3.7 and 3.8 are compared to the unconditional Stokes and Sager (1988) bands in terms of the optimality criterion $\sum_i (b_i - a_i)$. For both situations and for every possible observed sequence of order statistics, the optimal conditional band has a lower value for the optimality criterion. This finding holds despite the fact that the conditional coverage probabilities for the unconditional bands, which are also given in Tables 3.3 and 3.4, exceed 95% in some cases. Thus, there seems little doubt that the optimal conditional confidence bands are to be preferred to the unconditional bands.
CHAPTER 4

INTENTIONALLY REPRESENTATIVE SAMPLING

4.1 Introduction

Suppose that there is some population whose mean we wish to estimate. If this population consists of only a few units, each of which can be accessed and measured at negligible cost, then it may be feasible to obtain a measurement for every unit in the population and determine the population mean through complete enumeration. This is typically not feasible, however, either because the population of interest is large or infinite or because the cost of accessing and measuring a unit is not negligible. As a result, measurements can be obtained for only a limited number of units, and the quality of the resulting estimate of the population mean depends on the particular sampling scheme chosen.

The basic sampling scheme that serves as the foundation for more complex sampling schemes is simple random sampling (SRS), in which every possible sample is equally likely to be chosen. If auxiliary information is available in the form of relevant covariates, then stratified random sampling, in which one draws separate simple random samples from each of several disjoint subsets of the population, tends to produce better estimates of the population mean than does SRS. In the event that it is inexpensive to rank or approximately rank units without actually measuring them, ranked-set
sampling (RSS) is another approach that can produce better estimates than does SRS. However, none of these three sampling schemes offers protection against obtaining a sample that is obviously unrepresentative of the population. We introduce a novel sampling scheme, intentionally representative sampling (IRS), that yields an unbiased estimator of the population mean while allowing the researcher to ensure that each potential sample is, as far as possible, representative of the population. It thus allows a researcher to take full advantage of whatever auxiliary information is available.

One reason for preferring IRS to other sampling schemes is that it allows interested parties to take a role in selecting the sample. Consider, for example, the scenario discussed by Willemain (1980) in which nursing home operators are to be reimbursed by the Federal government according to the average level of debility of patients under their care. Since determining each patient’s level of debility requires expensive individual assessments, a complete enumeration is not feasible. On the other hand, nursing home operators would be unhappy with a plan based on SRS because of the possibility that a simple random sample might fail to include any of the most expensive patients, thus seriously underestimating the true cost per patient. Using IRS, the potential samples could be hand-selected by the nursing home operators, thus allaying their fears, without introducing any bias into the estimation of the mean cost per patient.

In this chapter, we describe IRS, prove that it always yields an unbiased estimator of the population mean, and suggest a simple rank-based method for implementing IRS when a single covariate is available. We then show via a simulation study that when the covariate and the variable of interest have a bivariate normal distribution, the IRS estimator of the population mean has a much smaller variance than either the
SRS or RSS estimators. We also consider a real data set consisting of measurements of diameters and heights for longleaf pines in a Georgia forest. We show that if diameter is taken as a covariate, and the goal is estimating the less accessible variable height, then IRS again allows much more efficient estimation than either SRS or RSS.

In Section 4.2, we motivate IRS by thinking of RSS as a two-phase sampling scheme and imagining how we might improve it. In Section 4.3, we provide a full description of IRS, prove that it leads to an unbiased estimate of a population mean, and suggest a method for computing standard errors. In Section 4.4, we propose a simple rank-based method for implementing IRS. In Section 4.5, we report on two simulation studies that show the improved efficiency of IRS relative to SRS and RSS when ranking is done according to a covariate. In Section 4.6, we summarize our findings and discuss the advantages of using IRS.

4.2 Motivation

Ranked-set sampling (RSS), proposed by McIntyre (1952), is a sampling scheme appropriate for use when it is inexpensive to rank or approximately rank units without actually measuring them. To draw a ranked-set sample, one first identifies a set size $r$ (typically between two and five) small enough that sets of $r$ units can easily be ranked accurately. One then draws $r$ independent simple random samples of size $r$. The units within each of these $r$ samples are ranked from smallest to largest using some method that does not require any actual measurement, and the $i$th smallest unit from the $i$th sample is selected for measurement. The procedure just described constitutes a single cycle of RSS, and it gives a sample of $r$ measured values. Larger samples are obtained by doing multiple independent cycles.
It is well-known that the RSS sample mean is an unbiased estimator of the population mean, and one may appeal to the theory of order statistics to show that the RSS sample mean is more efficient than the corresponding sample mean based on SRS. However, an alternate explanation for the improved efficiency of the RSS mean estimator can be obtained by thinking of both SRS and RSS as two-phase sampling schemes and looking at the sorts of samples they produce. What we will argue, essentially, is that the samples that arise from RSS are more representative of the population in the sense that samples consisting either entirely of large values or entirely of small values are excluded or given lower probabilities of occurring. Suppose, for example, that we wish to estimate a population mean, but that we can afford to make only two measurements. A ranked-set sampling approach to the problem might begin by setting $r = 2$ and selecting a random sample of size four from the population. This sample would then be randomly split into two smaller groups of size two. The smaller of the two units from the first group would be selected for measurement, and the larger of the two units from the second group would be selected for measurement. A simple random sample of size two, meanwhile, could be drawn simply by randomly selecting two units from the sample of size four. If we write $y_1 < y_2 < y_3 < y_4$ for the ordered values in the sample of size four, then the distribution of the sample to be measured would be as given in Table 4.1.

One can see from Table 4.1 that the difference between the distribution of the sample obtained using RSS and the sample obtained using SRS is that the two samples $\{y_1, y_2\}$ and $\{y_3, y_4\}$, which are unrepresentative in the sense that they each consist entirely of values from one extreme, are impossible under RSS. The probability assigned by SRS to those unrepresentative samples is redistributed by RSS to the
Table 4.1: Distribution of the sample to be measured.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Probability (SRS)</th>
<th>Probability (RSS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>{y_1, y_2}</td>
<td>1/6</td>
<td>0</td>
</tr>
<tr>
<td>{y_1, y_3}</td>
<td>1/6</td>
<td>1/6</td>
</tr>
<tr>
<td>{y_1, y_4}</td>
<td>1/6</td>
<td>1/3</td>
</tr>
<tr>
<td>{y_2, y_3}</td>
<td>1/6</td>
<td>1/3</td>
</tr>
<tr>
<td>{y_2, y_4}</td>
<td>1/6</td>
<td>1/6</td>
</tr>
<tr>
<td>{y_3, y_4}</td>
<td>1/6</td>
<td>0</td>
</tr>
</tbody>
</table>

more representative samples \{y_1, y_4\} and \{y_2, y_3\}. Even so, under RSS the samples \{y_2, y_4\} and \{y_1, y_3\}, which are arguably less representative than the samples \{y_1, y_4\} or \{y_2, y_3\}, can still occur. One might suppose that a still better scheme for drawing a sample of size two from the set \{y_1, y_2, y_3, y_4\} could consist of taking the sample \{y_1, y_4\} with probability 1/2 and the sample \{y_2, y_3\} with probability 1/2. Under such a scheme, the mean of the two measurements would continue to be an unbiased estimator of the population mean, and the variance of the resulting estimator might be expected to be smaller than that of either the RSS or the SRS estimator. This proposed sampling scheme represents a particular case of IRS.

4.3 The Sampling Procedure

Assume that a total of \(k\) measurements are to be made, and suppose that \(n\) units are available for measurement. These \(n\) units may be the entire population, or they may simply be a simple random sample drawn from a larger, possibly infinite, population. Let \(m\) be a positive integer such that \(k\) divides \(nm\), and imagine that we have \(m\) identical markers for each of the \(n\) units available for measurement. We then have a total of \(nm\) markers, and we may split these markers into \(l = nm/k\) groups of
size \( k \). If we randomly select one of these \( l \) groups of markers, obtain a measurement for each of the units corresponding to the markers in the group selected, and compute the sample mean, we obtain the IRS estimate \( \hat{\mu}_{IRS} \) of the population mean \( \mu \). The following theorem shows that this estimate is always unbiased for \( \mu \).

**Theorem 4.1.** Let \( X_1, \ldots, X_n \) be a simple random sample from the population of interest. Suppose that we select \( k \) units from this random sample by IRS. Then the mean \( \hat{\mu}_{IRS} \) of the \( k \) sampled values is unbiased for the population mean \( \mu \).

**Proof of Theorem 4.1.** Since the SRS sample mean \( \bar{X} \) is unbiased for the population mean,

\[
E[\bar{X}] = E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \mu.
\]

Let \( m \) be the number of markers per unit, and let \( l = nm/k \) be the number of groups. Define \( \{Z_{ij}; i = 1, \ldots, n, j = 1, \ldots, m\} \) to be occupation variables such that \( Z_{ij} \) is 1 when the \( j \)th marker for \( X_i \) is in the group chosen and 0 otherwise. We can then write the IRS sample mean as

\[
\hat{\mu}_{IRS} = \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} X_i.
\]

Since there are a total of \( l \) groups, each random variable \( Z_{ij} \) is marginally distributed as a Bernoulli random variable with success probability \( 1/l \). Moreover, each individual \( Z_{ij} \) is independent of the associated \( X_i \). The theorem then follows by noting that

\[
E[\hat{\mu}_{IRS}] = E \left[ \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} X_i \right] = \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} E[Z_{ij}X_i]
\]

\[
= \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} E[Z_{ij}]E[X_i] = \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{l} E[X_i]
\]

\[
= \frac{m}{kl} \sum_{i=1}^{n} E[X_i] = \frac{nm}{kl} E[\bar{X}] = \mu.
\]
Since the unbiasedness of the IRS mean estimator does not depend on the way we split the \( nm \) markers into groups, we should seek to select these groups in such a way that each group has about the same sample mean as any other group. Ideally, the groups would be such that we could not tell, before actual measurement, that any one group had a higher sample mean than any other. Selecting the groups in this way will clearly tend to minimize the variance of the estimator, and we avoid the situation that sometimes arises in SRS of drawing a sample that we do not believe is representative of the population, but being unable to do anything about it. A systematic method for choosing the groups will be discussed in Section 4.4. More generally, however, a researcher with expert knowledge of the matter at hand may simply select the groups to be representative in all the ways believed important. Units that are expected to have high values of the variable of interest may be balanced with units expected to lead to low values, and all varieties of auxiliary information may be considered in forming the groups. As a result, the expert knowledge of the researcher can be put to full use.

When we estimate the population mean \( \mu \), we are also interested in knowing something about the accuracy of our estimate. Because the samples from which measurements are taken in the IRS procedure do not arise from SRS, standard error estimates based on looking at the in-sample variation are not appropriate. Instead, if a standard error is desired, one may draw multiple independent IRS samples. Provided that the groups used in drawing the multiple IRS samples are constructed in a consistent manner, the average of the sample means from these multiple samples can be used as an overall estimate of the population mean, and a standard error may
be computed by looking at the amount of variability in the individual IRS sample means.

4.4 A Rank-Based Implementation Method

IRS is ideally implemented by taking advantage of the researcher’s expert knowledge. However, in cases where this expert knowledge does not exist or where the number of judgments that would need to be made is prohibitively large, alternatives must be sought. If units can be ranked either according to a covariate or by judgment without making any actual measurements, systematic procedures can be used to construct representative groups to be the potential samples for measurement. As in Section 4.3, assume that \( k \) measurements are to be made, and suppose that there are \( n \) units available for measurement. We take the number of markers \( m \) for each unit to be \( k \), and we use the ordered vector of integers

\[
C = (1, 1, \ldots, 1, 2, 2, \ldots, 2, 3, \ldots, n - 1, n, n, \ldots, n),
\]

where each of \( 1, 2, \ldots, n \) appears exactly \( k \) times, to represent the \( nk \) markers that we will divide into \( n \) groups. If we imagine using the \( nk \) markers as the elements of an \( n \) by \( k \) matrix \( A \) whose \( n \) rows represent the groups, then one set of groups may be constructed using the following algorithm. The idea behind this algorithm is to construct the groups in such a way that the sum of the ranks of the units in each group is roughly constant across all groups.

**The Rank Algorithm:** Let \( r \) be the integer such that either \( k = 2^r \) or \( k = 2^r - 1 \). We fill in the first \( r \) columns of the matrix \( A \) by using the elements of the row vector \( C \) sequentially from left to right. We let the first \( n \) elements of \( C \) run down the first
column, and we then let the next $n$ elements of $C$ run up the second column. We continue this alternating procedure (i.e., down, up, down, up, etc.) until the first $r$ columns are filled. We then fill in the last $k - r$ columns of the matrix by using the remaining elements of the vector $C$ sequentially from right to left, starting with the $k$th column. We let the last $n$ elements of $C$ run down the $k$th column, and we let the next-to-last $n$ elements of $C$ run up the $(k - 1)$st column. We then alternate as before until we have filled in the entire matrix. Having found the matrix $A$, we obtain the groups by taking the $i$th group to consist of the units whose ranks are given on the $i$th row of the matrix.

As an example, consider the case in which $k = 2$ and $n = 4$. The vector $C$ will be

$$C = (1, 1, 2, 2, 3, 3, 4, 4),$$

and the matrix whose rows represent the groups will be given by

$$A = \begin{bmatrix}
1 & 4 \\
1 & 4 \\
2 & 3 \\
2 & 3
\end{bmatrix}.$$

Thus, the two potential samples are the group consisting of the units with ranks 2 and 3 and the group consisting of the units with ranks 1 and 4. Groups for the larger case in which $k = 3$ and $n = 10$ are given in Table 4.2.

When $k$ is even, the rank algorithm leads to groups such that the ranks of the units in each group sum to the same value. For the case of $k$ odd, the ranks of the units in each group do not sum to exactly the same value, but the group-to-group differences are not large. As can be seen in Table 4.2, the rank algorithm may lead to groups that contain repeated values. Since this seems undesirable, there may be an
Table 4.2: The groups identified by the rank algorithm when $n = 10$ and $k = 3$.

<table>
<thead>
<tr>
<th>Number</th>
<th>Ranks for the group elements</th>
<th>Number</th>
<th>Ranks for the group elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>${1, 7, 10}$</td>
<td>6</td>
<td>${2, 5, 9}$</td>
</tr>
<tr>
<td>2</td>
<td>${1, 7, 10}$</td>
<td>7</td>
<td>${3, 5, 8}$</td>
</tr>
<tr>
<td>3</td>
<td>${1, 6, 10}$</td>
<td>8</td>
<td>${3, 5, 8}$</td>
</tr>
<tr>
<td>4</td>
<td>${2, 6, 9}$</td>
<td>9</td>
<td>${3, 4, 8}$</td>
</tr>
<tr>
<td>5</td>
<td>${2, 6, 9}$</td>
<td>10</td>
<td>${4, 4, 7}$</td>
</tr>
</tbody>
</table>

advantage to slightly modifying the groups to eliminate repeated values. However, the comparisons we make in the remainder of this chapter are based on the most basic version of the rank algorithm.

Beyond making sure that no group contains repeated values, there are other changes one might make to the rank algorithm. Most obviously, one could make a different choice of which columns of $A$ to fill in from top to bottom and which columns to fill in from bottom to top. To a certain extent, this choice could even be made to counteract anticipated skewness in the underlying distribution. Another generalization of the rank algorithm could be obtained by considering expectations associated with each rank (e.g., normal scores) or even covariate values instead of ranks. The rank algorithm tries to make the groups representative in the sense that the ranks of the units in each group have roughly the same sum. We could instead attempt to make the groups representative in the sense that the sum of the expectations associated with the ranks of the units in each group or the sum of the covariate values corresponding to the units in each group is roughly the same across all groups.
4.5 Comparisons of Estimator Performance

To compare the performance of the IRS mean estimator to that of the SRS and RSS mean estimators, two simulation studies were conducted. The first simulation study assumed that the variable of interest and a more easily measured covariate were jointly distributed as bivariate normal random variables. The second simulation study used a real data set consisting of measurements of diameters and heights for longleaf pines in a Georgia forest.

In the first simulation study, the variable of interest was taken to be the value $X$, and it was assumed that there existed a covariate $Y$ such that $X$ and $Y$ were jointly distributed bivariate normal with some correlation $\rho$. There are no restrictions on the choices of $n$ and $k$ for IRS beyond the obvious constraint that $k \leq n$, but the choice $n = k^2$ was made to facilitate the comparison with RSS. The number of measurements $k$ was allowed to run from two to six, and the correlation $\rho$ ranged from 0.0 to 1.0. For each choice of $k$ and $\rho$, each sampling scheme was run 100,000 times. For each run, the SRS mean estimator $\bar{X}$ was obtained by making measurements on a randomly selected $k$ of the $n$ available units and averaging those measurements. The efficiencies of the IRS and RSS mean estimators relative to the SRS mean estimator were estimated using ratios of variances so that, for example, the efficiency of the IRS estimator relative to the SRS estimator was estimated as

$$
eff(IRS, SRS) = \frac{\hat{\text{Var}}(\bar{X})}{\hat{\text{Var}}(\hat{\mu}_{IRS})},$$

where $\hat{\text{Var}}(\bar{X})$ and $\hat{\text{Var}}(\hat{\mu}_{IRS})$ are the variance estimates from the simulation. The rank algorithm described in Section 4.4 was used to determine the groups for IRS, and all ranking was done according to the covariate $Y$. The RSS estimator was obtained
by drawing a single cycle of size \( k \) from the \( n = k^2 \) units available for measurement.

The results of the simulation study, which are given in Table 4.3, show that for all sample sizes and levels of correlation considered, the IRS estimator was more efficient than either the RSS estimator or the SRS estimator.

To demonstrate that the superior performance of IRS holds not just for simulated data, but also for the sorts of data encountered in practice, we considered a data set consisting of measurements of diameters at chest height and heights for 396 longleaf pines from a Georgia forest. This data set, developed by Platt, Evans, and Rathbun (1988), is given in Chen, Bai, and Sinha (2004). We assumed that height was the variable of interest, and we took diameter at chest height to be the covariate used for ranking. Figure 4.1, which gives a plot of height versus diameter at chest height, shows that diameter is in fact highly relevant for the estimation of height.

<table>
<thead>
<tr>
<th>( k )</th>
<th>Type</th>
<th>Correlation ( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>IRS</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>1.46</td>
</tr>
<tr>
<td>3</td>
<td>IRS</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>1.92</td>
</tr>
<tr>
<td>4</td>
<td>IRS</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>2.35</td>
</tr>
<tr>
<td>5</td>
<td>IRS</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>2.76</td>
</tr>
<tr>
<td>6</td>
<td>IRS</td>
<td>5.42</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>3.19</td>
</tr>
</tbody>
</table>

Table 4.3: Simulated efficiencies (relative to SRS) for the IRS and RSS mean estimators.
Table 4.4: Simulated efficiencies (relative to SRS) for the IRS and RSS mean estimators when diameter is used as a covariate for height. The value $r$ is the set size used for RSS.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$n$</th>
<th>$k$</th>
<th>RSS efficiency</th>
<th>IRS efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1.29</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>4</td>
<td>1.30</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>6</td>
<td>1.30</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>8</td>
<td>1.29</td>
<td>1.74</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>3</td>
<td>1.55</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>6</td>
<td>1.55</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>9</td>
<td>1.56</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>12</td>
<td>1.57</td>
<td>2.55</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>4</td>
<td>1.78</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>8</td>
<td>1.79</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>12</td>
<td>1.81</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>1.82</td>
<td>3.46</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>5</td>
<td>1.99</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>10</td>
<td>2.02</td>
<td>3.58</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>15</td>
<td>2.05</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>20</td>
<td>2.08</td>
<td>4.56</td>
</tr>
</tbody>
</table>
Figure 4.1: Plot of the variable of interest height (in feet) against the covariate diameter at chest height (in centimeters) for the longleaf pine data.
We let the RSS set size \( r \) run from two to five, and for each choice of \( r \), we let the ratio \( k/r \) run from one to four. The number of available units was taken to be \( n = rk \) since this is exactly the number of values needed for RSS. For each choice of the number of measured values \( k \) and the number of available units \( n \), each sampling procedure was run 100,000 times. The rank algorithm was used to choose the groups for IRS, and the efficiencies of the IRS and RSS estimators relative to the SRS mean estimator were estimated as before. The results, given in Table 4.4, show that the IRS estimator was more efficient than either the RSS or the SRS estimator for all choices of \( k \) and \( n \). Moreover, when \( r \) was kept fixed and the sample size \( k \) increased, the advantage of IRS over RSS and SRS increased. These results show clearly that when a relevant covariate like diameter at chest height is available, IRS is to be preferred to either RSS or SRS.

### 4.6 Discussion and Conclusions

IRS is a completely new sampling approach. It does, however, have connections to more established ideas in the sampling literature. For example, IRS is a *two-phase* sampling scheme in the sense that one begins by drawing a simple random sample of size \( n \), then uses auxiliary information about the units in this sample of size \( n \) to select the final sample of \( k \) units to be measured. It is also a *cluster* sampling scheme in the sense that it is groups of units rather than individual units that are sampled. However, unlike the clusters used in standard cluster-sampling schemes such as those described in Lohr (1999) and Murthy (1967), the clusters used in IRS may overlap. Another important difference is that while the clusters used in typical cluster-sampling schemes tend to consist of units which are similar to each other, the
clusters chosen in IRS are purposely selected to be representative in the sense that deviations from the population mean in one direction or another tend to cancel each other out. An intentionally representative sample is not a *stratified* sample in the usual sense. However, if stratification information is the only auxiliary information available, then the researcher would certainly ensure that all potential IRS samples included numbers of observations from each stratum proportional to the size of that stratum within the first-phase sample of size \( n \). IRS may not seem to be as well-positioned as stratified random sampling to take advantage of the gains in efficiency possible by sampling different strata at different rates when the strata have different levels of variability. However, if auxiliary information beyond the stratification information is available, one can imagine using IRS separately within each stratum as an alternative to stratified random sampling.

Our simulations in Section 4.5 showed that IRS outperforms both SRS and RSS when we rank the units according to a relevant covariate and choose the groups using the rank algorithm of Section 4.4. IRS is also more flexible than other sampling schemes because all kinds of auxiliary information can be used in creating the groups that are the potential samples. Another advantage for IRS is that, as described in Section 4.1, interested parties may play a role in the selection of the groups that are the potential samples. All of these facts suggest that IRS is a superior sampling method, but a certain amount of caution must still be used in applying it. One danger is that while unbiasedness is guaranteed no matter how the groups described in Section 4.3 are chosen, it is possible to choose the groups so poorly that IRS is less efficient than even SRS. This observation suggests that IRS is best suited for
application in those settings where auxiliary information is both abundant and well understood by the researcher.
CHAPTER 5

IMPERFECT RANKINGS MODELS FOR RANKED-SET SAMPLING

5.1 Introduction

Ranked-set sampling (RSS) is a sampling approach that leads to improved statistical inference in situations where the units to be sampled can be ranked relative to each other prior to formal measurement. A ranked-set sample is typically obtained in \( n \) independent cycles of size \( k \), giving a total sample size of \( nk \). A single cycle of the basic RSS procedure is carried out by first drawing \( k \) independent samples, or sets, of size \( k \) and ranking the units within each set. This ranking may be done either by subjective judgment or according to a concomitant variable, and it need not be completely accurate. The \( k \) units for measurement are then obtained by selecting, for \( r = 1, \ldots, k \), the unit in the \( r \)th set that has been given rank \( r \). If the rankings are done according to the variable of interest, meaning that the rankings are perfect, then the sampled values are independent order statistics from the underlying distribution. In the more general situation in which the rankings are not perfect, the sampled values are merely independent judgment order statistics.

When an inference procedure based on RSS is proposed, it is typically initially evaluated in terms of its performance when the rankings are perfect. The next step in
the evaluation process often consists of showing that regardless of the quality of the rankings, the procedure is at least as good as the corresponding simple random sampling procedure. However, only rarely is it possible to make more precise assessments of the performance of procedures based on RSS when the rankings are imperfect. One difficulty is that only a few plausible models for imperfect rankings are available in the literature. A second difficulty is that while there is only one way the rankings can be perfect, there seem to be almost unimaginably many ways in which the rankings could be imperfect. In this chapter, we develop a broad new class of imperfect ranking models, and we show that models from this class can be used to uniformly approximate essentially any reasonable imperfect rankings model. These models also have a simple structure that may sometimes allow efficiency computations to be done analytically rather than via simulation.

In Section 5.2, we describe this new class of models for imperfect rankings. In Section 5.3, we relate our class to a larger class that, we argue, contains essentially any reasonable imperfect rankings model. In Section 5.4, we show that any of the imperfect rankings models discussed in Section 5.3 can be approximated arbitrarily well by models in the class described in Section 5.2. In Section 5.5, we describe one specific method for selecting a one-parameter family of models from the class introduced in Section 5.2. We state our conclusions in Section 5.6.

5.2 A New Class of Models

In developing a class of models for imperfect rankings, there are several goals that should be kept in mind. Perhaps the most critical of these goals is that the model must be a valid model. Just as one cannot write down just any function $f(x, y)$ on $\mathbb{R}^2$ and
be assured that it is a valid covariance function, one cannot write down an arbitrary collection of cumulative distribution functions (CDFs) for judgment order statistics and be assured that they represent a valid imperfect rankings model. Instead, one must provide a mechanism by which the imperfect rankings specified in each model could actually occur. Another goal to keep in mind is that the class of models should be flexible enough to model a wide range of different types of judgment errors. Finally, the class should be such that it is easy to identify one-parameter families within the class that smoothly span the range from random rankings to perfect rankings. In this section, we describe a large class of models that achieves these three goals.

One model for imperfect rankings, explored by Bohn and Wolfe (1994), says that when the CDF for the underlying distribution is $F$, the cumulative distribution functions $F_{[1]}, \ldots, F_{[k]}$ for the judgment order statistics can be written as convex combinations of the CDFs $F_{(1)}, \ldots, F_{(k)}$ for the true order statistics. That is, the CDFs for the judgment order statistics take the form

$$ F_{[i]} = \sum_{j=1}^{k} p_{ij} F_{(j)}, \quad (5.1) $$

where the values $p_{ij} \in [0, 1]$ are the elements of a doubly stochastic $k \times k$ matrix $P$. The fact that any such model is a valid model is a consequence of Birkhoff’s Theorem, which states that any doubly stochastic matrix is a convex combination of permutation matrices. A proof of Birkhoff’s Theorem, which is nontrivial, is given in Bhatia (1997). To show that (5.1) yields a valid model for any choice of $P$, we let $P = (p_{ij})$ be an arbitrary $k \times k$ doubly stochastic matrix. Then $P$ can be written as a convex combination

$$ P = \sum_{(i_1, \ldots, i_k) \in S_k} \alpha(i_1, \ldots, i_k) Per(i_1, \ldots, i_k), $$

110
where $S_k$ is the set of all permutations of the integers 1 to $k$, $\text{Per}(i_1, \ldots, i_k)$ is the permutation matrix that has a 1 in the $(i_j, j)$th position for $j = 1, \ldots, k$ and zeros elsewhere, and $\{\alpha(i_1, \ldots, i_k)\}$ is a collection of nonnegative constants summing to 1. Consider the ranking scheme in which the order statistics $x_{1:k}, \ldots, x_{k:k}$ in any sample are given the ranks $i_1, \ldots, i_k$, respectively, with probability $\alpha(i_1, \ldots, i_k)$ independent of $x_{i:k}, \ldots, x_{k:k}$. One can easily see that under this ranking scheme, the CDFs for the judgment order statistics are exactly those given by (5.1).

In our description of a mechanism through which imperfect ranking models of the form (5.1) can arise, we have taken the probability that the order statistics $x_{1:k}, \ldots, x_{k:k}$ are given a particular list of ranks $i_1, \ldots, i_k$ to be independent of the actual values of these order statistics. Since one would expect values that are close together to be harder to order correctly than values that are far apart, this independence condition seems extremely unlikely to hold in practice, and it has been roundly criticized by, among others, Presnell and Bohn (1999). What must be pointed out, however, is that using a model of the form (5.1) need not signify that one is assuming that the independence condition holds. The independence condition simply provides a way to show that the model is valid, and models of the form (5.1) can arise even when one takes the actual values of $x_{1:k}, \ldots, x_{k:k}$ into account. To illustrate this phenomenon, we consider an example discussed by Presnell and Bohn (1999).

**Example 5.1:** Suppose that $k = 2$ and that the underlying distribution is uniform on $[0,1]$. Presnell and Bohn (1999) model the imperfect rankings by assuming that given $X_1$ and $X_2$, the true smallest order statistic $X_{(1)}$ is correctly declared to be the smallest with probability $\frac{1}{2} \left(1 + X_{(2)} - X_{(1)}\right)$. Thus, the probability of ordering
the observations correctly is an increasing function of the difference between \( X_{(1)} \) and \( X_{(2)} \). If we compute the CDFs \( F_1 \) and \( F_2 \) for the judgment order statistics \( X_{(1)} \) and \( X_{(2)} \), we find that \( F_1(t) = \frac{3}{2}t - \frac{1}{2}t^2 \) and \( F_2(t) = \frac{1}{2}t + \frac{1}{2}t^2 \) for \( t \in [0, 1] \). Since \( F_1(t) = 2t - t^2 \) and \( F_2(t) = t^2 \), we can write the CDFs for the judgment order statistics as

\[
F_1(t) = \frac{3}{4} F_1(t) + \frac{1}{4} F_2(t)
\]

and

\[
F_2(t) = \frac{1}{4} F_1(t) + \frac{3}{4} F_2(t)
\]

This shows that the judgment order statistics follow a model of the form (5.1), with the matrix \( P \) being given by

\[
P = \begin{pmatrix}
\frac{3}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{3}{4}
\end{pmatrix}.
\]

Example 5.1 shows that using a model of the form (5.1) does not require one to assume that the difficulty of ordering the values \( x_{1:k}, \ldots, x_{k:k} \) does not depend on the values of \( x_{1:k}, \ldots, x_{k:k} \). However, the space of models (5.1) may not be large enough to adequately model the imperfect rankings in a particular problem of interest. For example, in the case \( k = 2 \), the model (5.1) is determined by the choice of the single parameter \( p_{11} \). An obvious way to expand the space of valid models is to look at convex combinations of a larger class of functions than just the CDFs \( F_1, \ldots, F_k \) of the true order statistics. This larger class of functions could potentially be chosen.
in any number of ways. We need to be sure, however, that the models in the expanded space continue to be valid models. This suggests the constructive approach we describe next.

Suppose that we wish to carry out RSS with a set size of $k$. We may certainly draw our sets of size $k$ directly. However, another completely general way to obtain sets of size $k$ is to draw sets of size $r > k$ and then subsample $k$ units from the $r$. When we draw the $r$ units, the ordered values corresponding to those units are jointly distributed like order statistics $X_{1:r}, \ldots, X_{r:r}$ from the parent distribution. When we randomly draw a subsample of size $k$, we obtain each of the $\binom{r}{k}$ possible selections of $k$ out of the $r$ order statistics with equal probability. Suppose that having subsampled order statistics $x_{i_1:r}, \ldots, x_{i_k:r}$ ($i_1 < \cdots < i_k$) from the larger set of size $r$, we select a doubly stochastic matrix $A(i_1, \ldots, i_k)$. We may then take the distribution of $X_{[i]}$, conditional on the values $i_1, \ldots, i_k$, to be given by

$$F_{[i]}(t|i_1, \ldots, i_k) = \sum_{j=1}^{k} A(i_1, \ldots, i_k)_{ij} F_{ij:r}(t),$$

where $F_{ij:r}$ is the distribution of the true $i_j$th order statistic from a set of size $r$ and $A(i_1, \ldots, i_k)_{ij}$ is the $(i, j)$th entry of the matrix $A(i_1, \ldots, i_k)$. By our earlier discussion of the case in which a set of size $k$ is drawn directly, any choice of the doubly stochastic matrix $A(i_1, \ldots, i_k)$ yields a valid model, the argument being that the chance that $x_{i_j:r}$ is given the rank $i$ could be taken to be independent of the actual values of the order statistics $x_{i_1:r}, \ldots, x_{i_k:r}$.

Let $\Omega(i_1, \ldots, i_k)$ be the $k \times r$ matrix which has a 1 as the $(j, i_j)$th entry for $j = 1, \ldots, k$ and is otherwise zero. Then the matrix product

$$N(i_1, \ldots, i_k) \equiv A(i_1, \ldots, i_k)\Omega(i_1, \ldots, i_k)$$
is a $k \times r$ matrix such that the distribution of $X_{[i]}$, conditional on the values of $i_1, \ldots, i_k$, is given by

$$F_{[i]}(t|i_1, \ldots, i_k) = \sum_{j=1}^{r} N(i_1, \ldots, i_k)_{ij} F_{j;r}(t),$$

where $N(i_1, \ldots, i_k)_{ij}$ is the $(i,j)$th entry of $N(i_1, \ldots, i_k)$. It thus follows that once the contributions of all $\binom{r}{k}$ equally likely choices of the values $i_1, \ldots, i_k$ are taken into account, the CDF for the $i$th judgment order statistic $X_{[i]}$ is given by

$$F_{[i]}(t) = \sum_{j=1}^{r} p_{ij} F_{j;r}(t), \quad (5.2)$$

where $P = (p_{ij})$ is the $k \times r$ matrix average

$$P \equiv \binom{r}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq r} N(i_1, \ldots, i_k). \quad (5.3)$$

We note that whereas the matrix $P$ had to be doubly stochastic for the model (5.1), the matrix $P$ given by (5.3) must have row sums of 1 and column sums of $k/r$. We will see in Example 5.3, however, that simply having $P$ satisfy these row sum and column sum conditions is not sufficient to guarantee that (5.2) yields a valid imperfect rankings model.

**Example 5.2:** Suppose that $k = 2$ and $r = 3$. When we sample two values from the collection $\{x_{1:3}, x_{2:3}, x_{3:3}\}$, we obtain each of the pairs $\{x_{1:3}, x_{2:3}\}$, $\{x_{1:3}, x_{3:3}\}$, and $\{x_{2:3}, x_{3:3}\}$ with probability $1/3$. Suppose that the matrices $A(1,2)$, $A(1,3)$, and $A(2,3)$ corresponding to the three samples are

$$A(1,2) = \begin{pmatrix} a & 1-a \\ -a & a \end{pmatrix}, \quad A(1,3) = \begin{pmatrix} b & 1-b \\ -b & b \end{pmatrix}, \quad \text{and}$$

$$A(2,3) = \begin{pmatrix} c & 1-c \\ -c & c \end{pmatrix},$$

$$114$$
where $a, b, c \in [0, 1]$. The matrices $N(1, 2)$, $N(1, 3)$, and $N(2, 3)$ are then given by

$$N(1, 2) = \begin{pmatrix} a & 1-a & 0 \\ 1-a & a & 0 \end{pmatrix}, \quad N(1, 3) = \begin{pmatrix} b & 0 & 1-b \\ 1-b & 0 & b \end{pmatrix}, \quad \text{and}$$

$$N(2, 3) = \begin{pmatrix} 0 & c & 1-c \\ 0 & 1-c & c \end{pmatrix},$$

meaning that the $2 \times 3$ matrix $P$ is given by

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} a+b & 1-a+c & 2-b-c \\ 2-a-b & 1+a-c & b+c \end{pmatrix}.$$  

By construction, using $P$ in (5.2) gives a valid imperfect rankings model for any choice of the values $a, b,$ and $c,$ and this model is clearly more flexible than the model (5.1).

One may check that in this setting where $k = 2$ and $r = 3$, the constraints on the matrix $P$ reduce to the constraints that each entry be nonnegative, that each row sum to 1, and that each column sum to $2/3$.

\textit{Example 5.3:} Suppose now that $k = 2$ and $r = 4$. When we sample two values from the collection $\{x_{1:4}, x_{2:4}, x_{3:4}, x_{4:4}\}$, we obtain each of the six possible pairs with probability $1/6$. Suppose that we seek to have the judgment order statistic $X_{[1]}$ take on the smallest values possible. This would require that the smallest value in each pair be correctly identified, meaning that $x_{1:4}$ is identified as the smallest whenever it appears, that $x_{2:4}$ is identified as the smallest when it is sampled together with $x_{3:4}$ or $x_{4:4}$, and that $x_{3:4}$ is identified as the smallest when it is sampled together with $x_{4:4}$. This would tell us that the distribution of $F_{[1]}$ is given by

$$F_{[1]}(t) = \frac{1}{2}F_{1:4}(t) + \frac{1}{3}F_{2:4}(t) + \frac{1}{6}F_{3:4}(t),$$
which is stochastically larger than the distribution $\frac{1}{2}F_{1:4}(t) + \frac{1}{2}F_{2:4}(t)$. Thus, even though the latter distribution would correspond to a matrix

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

with row sums equal to 1 and column sums equal to $1/2 = k/r$, it is not a valid model for imperfect rankings. To find all the models (5.2) that are valid in this setting with $k = 2$ and $r = 4$, let $a, b, c, d, e, f \in [0, 1]$ be values that determine our choices of the doubly stochastic matrices $A(1, 2), A(1, 3), A(1, 4), A(2, 3), A(2, 4),$ and $A(3, 4)$, respectively. We can then write the resulting $2 \times 4$ matrix $P$ as

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \end{pmatrix} = \frac{1}{6} \begin{pmatrix} a + b + c & 1 - a + d + e & 2 - b - d + f & 3 - c - e - f \\ 3 - a - b - c & 2 + a - d - e & 1 + b + d - f & c + e + f \end{pmatrix}.$$

In implementing the model (5.2), we must choose a doubly stochastic matrix $A(i_1, \ldots, i_k)$ for each choice of the values $i_1 < \cdots < i_k$. Under the independence condition that we have used to show the validity of the model, choosing $A(i_1, \ldots, i_k)$ is equivalent to partitioning probability among the $k!$ possible orderings of the values $x_{i_1:r}, \ldots, x_{i_k:r}$. One way to assign probabilities to all of these orderings at once is to first assign probabilities to all orderings of $x_{1:r}, \ldots, x_{r:r}$. Probabilities for orderings of $x_{i_1:r}, \ldots, x_{i_k:r}$ are then obtained by summing over all orderings of the full set $x_{1:r}, \ldots, x_{r:r}$ that give the desired ordering for $x_{i_1:r}, \ldots, x_{i_k:r}$. To implement this projection in matrix form, suppose that $R$ is the doubly stochastic $r \times r$ matrix determined by assigning probabilities to all orderings of $x_{1:r}, \ldots, x_{r:r}$. That is, the $(i, j)$th
element of $R$ is the probability that the true $j$th value is given the rank $i$. We then have, under the independence condition, that

$$P(X_{[i:k]} = X_{j:r}) = \sum_{l=1}^{r} P(X_{[i:k]} = X_{[l:r]} \text{ and } X_{[l:r]} = X_{j:r})$$

$$= \sum_{l=1}^{r} \frac{(l-1) \left(\binom{r-l}{k-i}\right)}{\binom{r}{k}} \cdot R_{lj}, \quad (5.4)$$

where $R_{lj}$ is the $(l,j)$th entry of $R$. But (5.4) is precisely the expression one would obtain from a matrix multiplication of $C_{kr}$ and $R$, where $C_{kr}$ is the $k \times r$ matrix with $(i,l)$th entry $\left(\binom{l-1}{i-1} \left(\binom{r-l}{k-i}\right) / \binom{r}{k}\right)$. Thus, one may easily move from a model (5.1) for set size $r$ to a model (5.2) for any smaller set size. That these projection models do not, however, exhaust the possibilities for models of the form (5.2) is demonstrated in the next example.

**Example 5.4:** Suppose that $k = 3$ and $r = 4$. Suppose that the matrices $A(1,2,3)$, $A(1,2,4)$, $A(1,3,4)$, and $A(2,3,4)$ are given by

$$A(1,2,3) = A(1,2,4) = A(1,3,4) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A(2,3,4) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$ 

The matrix $P$ is then given by

$$P = \begin{pmatrix} 3/4 & 0 & 0 & 1/4 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/4 & 1/4 & 1/2 \end{pmatrix},$$

while the matrix $C_{34}$ may be computed to be

$$C_{34} = \begin{pmatrix} 3/4 & 1/4 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}.$$ 

Suppose that $R = (R_{ij})$ is the $4 \times 4$ matrix that, when projected onto dimension 3, leads to the matrix $P$. Then $P = C_{34}R$ implies that if we consider the entries
$P_{12}, P_{13}, P_{21},$ and $P_{24},$ which are all zero, we must have that

\[
\frac{3}{4} R_{12} + \frac{1}{4} R_{22} = 0, \\
\frac{3}{4} R_{13} + \frac{1}{4} R_{23} = 0, \\
\frac{1}{2} R_{21} + \frac{1}{2} R_{31} = 0, \text{ and} \\
\frac{1}{2} R_{24} + \frac{1}{2} R_{34} = 0.
\]

But then, since $R$ has only nonnegative entries, each of $R_{21}, R_{22}, R_{23},$ and $R_{24}$ are zero, which is impossible since $R$ is doubly stochastic.

### 5.3 Connection with More General Models

It is natural to wonder how broad the class of models introduced in Section 5.2 is. To answer this question, we need a way to specify a general imperfect rankings model. One completely general way to do this is by specifying the CDFs $F[1], \ldots, F[k]$ of the judgment order statistics. We saw in Example 5.3, however, that if $F[1], \ldots, F[k]$ are specified arbitrarily, there is no guarantee that a valid model will result. Consequently, it makes sense to consider a different way of specifying the model. One specification method that does guarantee a valid model consists of specifying a collection of measurable functions \(f_{pij}(x_1, \ldots, x_k)\) that determine a doubly stochastic $k \times k$ matrix $P(x_1, \ldots, x_k)$ at each point $(x_1, \ldots, x_k)$ in their domains. To see this, suppose that conditionally on observed values $x_1, \ldots, x_k$, the doubly stochastic matrix $P(x_1, \ldots, x_k) = (p_{ij})$ is such that $p_{ij}$ gives the probability that the true $j$th order statistic $x_{(j)}$ will be given the judgment rank $i$. Any model where the rankings are done according to a concomitant variable (see, for example, Dell and Clutter (1972)) can be specified by such a collection of functions \(\{p_{ij}\}\), as can any model based on
subjective judgment rankings that are consistent over time. Because of the mechanism arising out of Birkhoff’s Theorem, any specification of the functions \( \{p_{ij}\} \) is guaranteed to lead to a valid imperfect rankings model.

One simplification that we can make is to notice that if the matrix function \( P(x_1, \ldots, x_k) \) depends on the order of presentation of the values \( x_1, \ldots, x_k \), which it easily could (i.e., the order of presentation of the units could affect the ranking), we can obtain another matrix function \( P' \), depending only on the order statistics \( x_{(1)}, \ldots, x_{(k)} \), that leads to exactly the same judgment order statistic CDFs \( F_1, \ldots, F_k \). To see this, first note that, following Presnell and Bohn (1999), we can write that

\[
F_i(t) = P(X_i \leq t) = \sum_{j=1}^{k} P(X_{(j)} \leq t, X_i = X_{(j)})
\]

\[
= \sum_{j=1}^{k} \int_{\mathbb{R}^k} I(x_{(j)} \leq t) p_{ij}(x_1, \ldots, x_k) dF(x_k) \ldots dF(x_1).
\]

For every permutation \( \pi \in S_k \), there is a subset \( R_{\pi} \) of \( \mathbb{R}^k \) on which \( x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(k)} \). Since these subsets \( R_{\pi} \) exhaust \( \mathbb{R}^k \), but overlap only on a set of measure zero, we may write that

\[
F_i(t) = \sum_{j=1}^{k} \sum_{\pi \in S_k} \int_{R_{\pi}} I(x_{(j)} \leq t) p_{ij}(x_1, \ldots, x_k) dF(x_k) \ldots dF(x_1). \tag{5.5}
\]

For fixed \( \pi \in S_k \), consider the integral over \( R_{\pi} \) that appears in (5.5). Relabel the coordinates by defining new variables \( y_1 = x_{\pi(1)}, \ldots, y_k = x_{\pi(k)} \). Then, if \( \pi^{-1} \in S_k \) is the inverse of \( \pi \), we have that

\[
\int_{R_{\pi}} I(x_{(j)} \leq t) p_{ij}(x_1, \ldots, x_k) dF(x_k) \ldots dF(x_1)
\]

\[
= \int_R I(y_{(j)} \leq t) p_{ij}(y_{\pi^{-1}(1)}, \ldots, y_{\pi^{-1}(k)}) dF(y_k) \ldots dF(y_1),
\]

119
where $R = \{(y_1, \ldots, y_k) | y_1 \leq \cdots \leq y_k\}$.

This gives us that

$$F_{ij}(t) = \sum_{j=1}^{k} \sum_{\pi \in S_k} \int_{R} I(y_{(j)} \leq t) p_{ij}(y_{i_1}, \ldots, y_{i_k})dF(y_k) \cdots dF(y_1)$$

$$= \sum_{j=1}^{k} \int_{R} I(y_{(j)} \leq t) \sum_{\pi \in S_k} p_{ij}(y_{i_1}, \ldots, y_{i_k})dF(y_k) \cdots dF(y_1)$$

$$= \sum_{j=1}^{k} \int_{R} I(y_{(j)} \leq t) \left( \frac{1}{k!} \right) \sum_{\pi \in S_k} p_{ij}(y_{i_1}, \ldots, y_{i_k})(k! I(y_1 \leq \cdots \leq y_k))dF(y_k) \cdots dF(y_1),$$

where the last equality follows in part from the observation that the set of all $\pi \in S_k$ is exactly the same as the set of all $\pi^{-1} \in S_k$. If we then note that $k! I(y_1 \leq \cdots \leq y_k)dF(y_k) \cdots dF(y_1)$ is precisely the joint density of the order statistics $y_{(1)}, \ldots, y_{(k)}$, while $R$ is exactly the support of those order statistics, it is clear that setting

$$P'(x_{(1)}, \ldots, x_{(k)}) = \frac{1}{k!} \sum_{\pi \in S_k} P(x_{i_1}, \ldots, x_{i_k})$$

and integrating against the joint density of the order statistics would give us exactly the same judgment order statistic CDFs $F_{[1]}, \ldots, F_{[k]}$ that we obtained using the original functions $\{p_{ij}\}$. Thus, in the remainder of this chapter, we take each $p_{ij}$ to be a function solely of the order statistics in the sample.

We now examine the relationship between the class of models introduced in Section 5.2 and more general imperfect rankings models by looking at the functions $\{p_{ij}\}$. Let $k$ and $r$ be fixed, and suppose that we wish to find the function $p_{ij}(x_1, \ldots, x_k)$ corresponding to some choice of the matrices $\{A(i_1, \ldots, i_k), 1 \leq i_1 < \cdots < i_k \leq r\}$ discussed in Section 5.2. When the sample of $k$ measured values consists of the ordered values $x_1 < \cdots < x_k$, those values could have arisen as the order statistics
for any of the \( \binom{r}{k} \) choices of \( i_1, \ldots, i_k \). When they arise from a particular choice of \( i_1, \ldots, i_k \), the chance that \( x_{[i]} = x_{(j)} \) is just \( A(i_1, \ldots, i_k)_{ij} \). Thus, we have that

\[
p_{ij}(x_1, \ldots, x_k) = \frac{\binom{r}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq r} A(i_1, \ldots, i_k)_{ij} f_{(i_1, \ldots, i_k)}(x_1, \ldots, x_k)}{\binom{r}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq r} f_{(i_1, \ldots, i_k)}(x_1, \ldots, x_k)},
\]

(5.6)

where \( f_{(i_1, \ldots, i_k)}(x_1, \ldots, x_k) \) is the joint density of the order statistics \( X_{i_1:r}, \ldots, X_{i_k:r} \), evaluated at \( x_1, \ldots, x_k \). Using standard distribution theory given in, for example, David and Nagaraja (2003), the joint density for the order statistics \( X_{i_1:r}, \ldots, X_{i_j:r} \) can be written as

\[
F(x_1)^{i_1-1}(F(x_2) - F(x_1))^{i_2 - i_1 - 1} \cdots (F(x_k) - F(x_{k-1}))^{i_k - i_{k-1} - 1}(1 - F(x_k))^{r-i_k},
\]

where \( F \) and \( f \) are the CDF and the density, respectively, for the underlying distribution. Defining new indices \( s_1, \ldots, s_{k+1} \) by setting \( s_1 = i_1 - 1, s_2 = i_2 - i_1 - 1, \ldots, s_k = i_k - i_{k-1} - 1, \) and \( s_{k+1} = r - i_k \), we can rewrite this joint density as

\[
\frac{r!}{s_1!s_2! \cdots s_k!s_{k+1}!} f(x_1) \cdots f(x_k) \\
F(x_1)^{s_1}(F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k}(1 - F(x_k))^{s_{k+1}},
\]

which is the same as

\[
\frac{r!}{(r-k)!} \binom{r-k}{s_1 \cdots s_{k+1}} f(x_1) \cdots f(x_k) \\
F(x_1)^{s_1}(F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k}(1 - F(x_k))^{s_{k+1}}.
\]
Making the analogous change of indices in (5.6) and canceling the common factors of \( f(x_1) \cdots f(x_k) \) in the numerator and denominator, we obtain a denominator of

\[
\frac{k!(r-k)!}{r!} \left( \frac{r!}{(r-k)!} \right) \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} \left( \begin{array}{c} r-k \\ s_1 \cdots s_{k+1} \end{array} \right) .
\]

\[
F(x_1)^{s_1} (F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k} (1 - F(x_k))^{s_{k+1}} .
\]

\[
= k! \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} \left( \begin{array}{c} r-k \\ s_1 \cdots s_{k+1} \end{array} \right) .
\]

\[
F(x_1)^{s_1} (F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k} (1 - F(x_k))^{s_{k+1}} .
\]

Since the summands in this last expression are just the values of the probability mass function when the vector \((S_1,\ldots,S_{k+1})\) has a multinomial distribution with mass parameter \(r-k\) and probability vector \((F(x_1), F(x_2) - F(x_1), \ldots, 1 - F(x_k))\), this denominator is simply \(k!\). The full expression (5.6) is then given by

\[
p_{ij}(x_1, \ldots, x_k) = \frac{k!(r-k)!}{r!k!} \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} A(s_1, \ldots, s_k) \frac{r!}{(r-k)!} \left( \begin{array}{c} r-k \\ s_1 \cdots s_{k+1} \end{array} \right) .
\]

\[
F(x_1)^{s_1} (F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k} (1 - F(x_k))^{s_{k+1}} .
\]

\[
= \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} A(s_1, \ldots, s_k) \left( \begin{array}{c} r-k \\ s_1 \cdots s_{k+1} \end{array} \right) .
\]

\[
F(x_1)^{s_1} (F(x_2) - F(x_1))^{s_2} \cdots (F(x_k) - F(x_{k-1}))^{s_k} (1 - F(x_k))^{s_{k+1}} .
\]

For the important special case in which the underlying distribution is uniform on the interval \([0,1]\), we obtain the expression

\[
p_{ij}(x_1, \ldots, x_k) = \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} A(s_1, \ldots, s_k) \left( \begin{array}{c} r-k \\ s_1 \cdots s_{k+1} \end{array} \right) .
\]

\[
x_1^{s_1} (x_2 - x_1)^{s_2} \cdots (x_k - x_{k-1})^{s_k} (1 - x_k)^{s_{k+1}} . \quad (5.7)
\]

### 5.4 Uniform Approximation Results

We begin by proving a technical lemma about the uniform approximation of continuous functions on the compact domain \(D = \{(x_1, \ldots, x_k) | 0 \leq x_1 \leq \cdots \leq x_k \leq 1\}.\)
We then use this lemma to show that any imperfect rankings model specified by measurable functions \( \{p_{ij}\} \) as described in Section 5.3 can be uniformly approximated by models in the class that we introduced in Section 5.2.

Suppose that \( f(x_1, \ldots, x_k) \) is a continuous function defined on \( D \). Define the approximating function \( B_r f \) by setting

\[
B_r f(x_1, \ldots, x_k) = \sum_{s_i \geq 0 \forall i, \sum_{i=1}^{k+1} s_i = r-k} \left\{ f \left( \frac{s_1}{r-k}, \frac{s_1 + s_2}{r-k}, \ldots, \frac{s_1 + \ldots + s_k}{r-k} \right) \right\} 
\]

\[
\left[ \left( \begin{array}{c}
 r - k \\
 s_1 \cdots s_{k+1}
\end{array} \right) x_1^{s_1} \cdots (x_k - x_{k-1})^{s_k} (1 - x_k)^{s_{k+1}} \right].
\]

For the case \( k = 1 \) in which \( D \) is the interval \( [0, 1] \), the polynomials \( B_r f \) are known as Bernstein polynomials, and their properties are discussed in detail by Lorentz (1986). One well-known property of the one-dimensional Bernstein polynomials is that as \( r \to \infty \), \( B_r f \) converges to \( f \) uniformly on \( [0, 1] \). The more general result that we state as Lemma 5.1 does not seem to have been stated in the literature. However, both the result itself and the proof are obvious extensions of results presented in Chapter VII of Feller (1971).

**Lemma 5.1.** As \( r \to \infty \), \( B_r f \) converges to \( f \) uniformly on \( D \).

**Proof of Lemma 5.1.** We first note that since \( f \) is continuous and \( D \) is compact, \( f \) is uniformly continuous on \( D \). Also, since the continuous image of a compact set is bounded, there exists a real number \( m \) such that \( |f(x_1, \ldots, x_k)| < m \) for all \( (x_1, \ldots, x_k) \in D \). Let the value \( (x_1, \ldots, x_k) \) be fixed at some point of interest. We may then note that the bracketed expression on the right-hand side of (5.8) is just the probability mass function for the random vector \( (S_1, \ldots, S_{k+1}) \) when that vector has a multinomial distribution with mass parameter \( r - k \) and probability vector...
Suppose that \((S_1,\ldots,S_{k+1})\) has such a multinomial distribution, and let \(M(s_1,\ldots,s_{k+1})\) denote the corresponding probability mass function. Then, using familiar properties of the multinomial distribution, we have that \(S_1 \sim \text{Bin}(r-k,x_1), S_1+S_2 \sim \text{Bin}(r-k,x_2), \ldots, S_1+\cdots+S_{k-1} \sim \text{Bin}(r-k,x_{k-1}),\) and \(S_1+\cdots+S_k \sim \text{Bin}(r-k,x_k).\) Thus, the random variables \(T_1 = \frac{s_1}{r-k}, \ldots, T_k = \frac{s_1+\cdots+s_k}{r-k}\) have means \(x_1,\ldots,x_k\) and variances \(\frac{x_1(1-x_1)}{r-k}, \ldots, \frac{x_k(1-x_k)}{r-k},\) respectively. Since the function \(x(1-x)\) is bounded by \(\frac{1}{4}\) on the interval \([0,1]\), none of \(T_1,\ldots,T_k\) has a variance in excess of \(\frac{1}{4(r-k)}\), which in turn is less than \(\frac{1}{4r}\).

Pick any \(\epsilon > 0\). Since the function \(f\) is uniformly continuous, there exists a \(\delta > 0\) such that if \((y_1,\ldots,y_k) \in D\) satisfies \(|x_i - y_i| < \delta, i = 1,\ldots,k,\) then \(|p_{ij}(x_1,\ldots,x_k) - p_{ij}(y_1,\ldots,y_k)| < \epsilon,\) and this \(\delta\) can be chosen independently of \((x_1,\ldots,x_k).\) Let \(D_\delta(x_1,\ldots,x_k)\) denote the set of all values \((y_1,\ldots,y_k) \in D\) such that no coordinate \(y_i\) differs from \(x_i\) by more than \(\delta.\) We now show that the probability that \((T_1,\ldots,T_k)\) is in the set \(D_\delta\) goes uniformly to 1 as \(r \to \infty.\) Let \(\sigma(T_i)\) denote the standard deviation of \(T_i, i = 1,\ldots,k.\) Since \(T_i\) has mean \(x_i\) and variance less than \(\frac{1}{4r}\), we have by Chebyshev’s inequality that

\[
P (|T_i - x_i| > \delta) = P \left( \frac{|T_i - x_i|}{\sigma(T_i)} > \left( \frac{\delta}{\sigma(T_i)} \right) \right) \leq \left( \frac{\sigma(T_i)^2}{\delta^2} \right) < \frac{1}{4r\delta^2}.
\]

Thus, to make \(P (|T_i - x_i| > \delta)\) smaller than \(\epsilon,\) it suffices to choose \(r\) so that \(\frac{1}{4r\delta^2} < \epsilon,\) or equivalently so that \(r > \frac{1}{4\epsilon \delta^2}.\) Note that, like our previous choice of \(\delta,\) this choice of \(r\) can be made independently of the value of \((x_1,\ldots,x_k).\)
Having picked \( r \) so large that \( P(|T_i - x_i| > \delta) < \epsilon \) for each \( i \), we have by Bonferroni’s inequality that
\[
P(|T_i - x_i| > \delta \text{ for some } i) < k\epsilon.
\]
In other words, \( P((T_1, \ldots, T_k) \notin D_\delta) < k\epsilon. \) Thus,
\[
|f(x_1, \ldots, x_k) - B_r f(x_1, \ldots, x_k)| = \left| \int_D f(x_1, \ldots, x_k) dM(s_1, \ldots, s_{k+1}) \right.
- \left. \int_D f(T_1, \ldots, T_k) dM(s_1, \ldots, s_{k+1}) \right|
\leq \int_D |f(x_1, \ldots, x_k) - f(T_1, \ldots, T_k)| dM(s)
= \int_{D_\delta} |f(x_1, \ldots, x_k) - f(T_1, \ldots, T_k)| dM(s)
+ \int_{D_\delta'} |f(x_1, \ldots, x_k) - f(T_1, \ldots, T_k)| dM(s).
\]
On the set \( D_\delta \), the absolute difference \( |f(x_1, \ldots, x_k) - f(T_1, \ldots, T_k)| \) is bounded by \( \epsilon \) by construction. On the set \( D_\delta' \), the absolute difference \( |f(x_1, \ldots, x_k) - f(T_1, \ldots, T_k)| \) is bounded by \( 2m \) since \( f \) is a bounded function. Thus, we have that
\[
|f(x_1, \ldots, x_k) - B_r f(x_1, \ldots, x_k)| \leq \epsilon \cdot P((T_1, \ldots, T_k) \in D_\delta) +
2m \cdot P((T_1, \ldots, T_k) \notin D_\delta)
\leq \epsilon + 2mk\epsilon = (2mk + 1)\epsilon.
\]
This shows that when \( r \) is picked larger than \( \frac{1}{4\delta\epsilon} \), the difference between \( f \) and \( B_r f \) is bounded by \( (2mk + 1)\epsilon \). Hence, as \( r \to \infty \), \( f \) goes to \( B_r f \) uniformly on \( D \). \( \square \)

Using the lemma, we can prove the following result about the uniform approximation of imperfect ranking models when the underlying distribution is uniform on \([0,1]\). The key observation in this proof is that if one makes the choices \( A(i_1, \ldots, i_k)_{ij} = p_{ij} \left( \frac{i-1}{r-k}, \ldots, \frac{i-k}{r-k} \right) \) for each \( i, j \), and \((i_1, \ldots, i_k)\), then the resulting imperfect rankings
scheme has functions \( \{ p_{ij}^* \} \) given by \( \{ B_r p_{ij} \} \). This can be seen by comparing (5.7) and (5.8).

**Theorem 5.1.** Suppose that the underlying distribution is uniform on \([0,1]\). For each \( i \) and \( j \) in the set \( \{ 1, \ldots, k \} \), let \( p_{ij}(x_1, \ldots, x_k) \) be the function defined on \( D \) by 
\[
p_{ij}(x_1, \ldots, x_k) = P(x[i] = x_j).
\]
If the functions \( \{ p_{ij} \} \) are all measurable, then there exists a sequence of imperfect ranking models from the class developed in Section 5.2 such that the judgment order statistic CDFs \( F[1], \ldots, F[k] \) are uniformly approximated by the judgment order statistic CDFs for that sequence of models.

**Proof of Theorem 5.1.** Since each \( p_{ij} \) is measurable and satisfies \( |p_{ij}(x_1, \ldots, x_k)| < 1 \) for all \( (x_1, \ldots, x_k) \in D \), each \( p_{ij} \) is contained in the space \( L^1(D) \) of functions that are integrable with respect to Lebesgue measure on \( D \). Fix any \( \epsilon > 0 \). Since continuous functions are dense in \( L^1(D) \), there exists a collection of continuous functions \( \{ p^c_{ij}, i = 1, \ldots, k; j = 1, \ldots, k \} \) such that for each \( i \) and \( j \), \( \| p^c_{ij} - p_{ij} \|_{L^1(D)} < \epsilon \). We also have, since all of the \( p^c_{ij} \) are continuous functions, that there exists an \( r \) so large that for each \( i \) and \( j \), \( \max_{(x_1, \ldots, x_k) \in D} |p^c_{ij}(x_1, \ldots, x_k) - B_r p^c_{ij}(x_1, \ldots, x_k)| < \epsilon \). But since \( D \) has a finite volume \( 1/k! \), this means that for each \( i \) and \( j \), \( \| p^c_{ij} - B_r p^c_{ij} \|_{L^1(D)} < \frac{\epsilon}{k!} < \epsilon \).

Consider the CDF \( F[i] \) for the \( i \)th judgment order statistic. We can write that
\[
F[i](t) = P(X[i] \leq t)
= \sum_{j=1}^{k} P(X_j \leq t, X[i] = X_j)
= \sum_{j=1}^{k} \int_D I(x_j \leq t)p_{ij}k!d\mu,
\]  
where \( \mu \) is Lebesgue measure on \( D \). Let \( F[i]^* \) be the CDF of the \( i \)th judgment order statistic under the sampling scheme corresponding to \( B_r p^c_{ij} \). Then, using the same
argument that led to (5.9), we have that

$$F^*_i(t) = \sum_{j=1}^{k} \int_D I(x_j \leq t)B_r p_{ij}^*k!d\mu,$$

and the difference between the two CDFs is

$$|F^*_i(t) - F_i(t)| = \left| \sum_{j=1}^{k} \int_D I(x_j \leq t)(p_{ij} - B_r p_{ij}^*)k!d\mu \right|$$

$$\leq \sum_{j=1}^{k} \int_D I(x_j \leq t)|p_{ij} - B_r p_{ij}^*|k!d\mu$$

$$\leq \sum_{j=1}^{k} \int_D I(x_j \leq t)\left(|p_{ij} - p_{ij}^*| + |p_{ij}^* - B_r p_{ij}^*|\right)k!d\mu$$

$$\leq k! \sum_{j=1}^{k} \left( \|p_{ij} - p_{ij}^*\|_{L^1(D)} + \|p_{ij}^* - B_r p_{ij}^*\|_{L^1(D)} \right)$$

$$\leq (k!)k (\epsilon + \epsilon) = 2k(k!)\epsilon.$$

Since the bound here is independent of $i$ and $t$, this proves that the convergence is uniform.

We now extend this result to allow arbitrary continuous underlying distributions.

**Theorem 5.2.** Suppose that the underlying distribution has a continuous CDF $F$. Let $F_1, \ldots, F_k$ be the CDFs for the judgment order statistics resulting from some imperfect rankings model specified by measurable functions $\{p_{ij}\}$ as described in Section 5.3. Then there exists a sequence of imperfect rankings models in the class developed in Section 5.2 such that the CDFs $F_1, \ldots, F_k$ are uniformly approximated by the judgment order statistic CDFs for that sequence of models.

**Proof of Theorem 5.2.** Define the function $F^{-1}$ on $[0,1]$ by setting

$$F^{-1}(u) = \inf\{x|F(x) \geq u\}.$$
Then, since \( F \) is continuous, we have by standard results that
\[
F^{-1}(u) \leq x \iff u \leq F(x).
\]

We may use the CDFs \( F_1, \ldots, F_k \) to define CDFs on \([0,1]\) by setting
\[
G_i(t) = P(F(X_i) \leq t) = P(X_i \leq F^{-1}(t))
= F_i(F^{-1}(t)).
\]

The CDFs \( G_1, \ldots, G_k \) define an imperfect ranking scheme with an underlying distribution that is uniform on \([0,1]\), and they correspond to defining measurable functions \( \{p_{ij}'\} \) as
\[
p_{ij}'(u_1, \ldots, u_k) = p_{ij}(F^{-1}(u_1), \ldots, F^{-1}(u_k)).
\]

Thus, Theorem 5.1 applies, and we can find an imperfect ranking scheme from the class described in Section 5.2 such that \( |G_{ij}^*(t) - G_{ij}(t)| \leq \epsilon \) for all \( i \) and all \( t \). Then, defining the judgment order statistic CDFs \( F^*_1, \ldots, F^*_k \) by setting \( F^*_i(F^{-1}(t)) = G_{ij}^*(t) \) for \( i = 1, \ldots, k \), we have that \( F_{ij} \) and \( F^*_i \) are also uniformly close for each \( i \).

The theorems in this section show that essentially any imperfect rankings model can be uniformly approximated by models in the class developed in Section 5.2. The theorems even specify a sequence of models, namely those corresponding to the polynomials \( \{B_r p_{ij}\} \), that do the approximating. This sequence of models specified in the theorems is not in any sense the ideal set of models to use, however. In fact, one may show by example that the convergence of these models is not even monotone. A better way to proceed is simply to find a value of \( r \) large enough that some model in the class (5.2) approximates the desired model well. Our suspicion is that in most cases of interest, \( r \) need not be much larger than \( k \).
5.5 A Method for Selecting One-Parameter Families

In this section, we describe a simple method for selecting a one-parameter family from the class of imperfect ranking models developed in Section 5.2. We want each such family to smoothly span the range from random rankings to perfect rankings, and we would like the family to incorporate prior notions about properties that the imperfect rankings are likely to have in the situation of interest. If we use the framework developed in Section 5.2, specifying a family is reduced to choosing the $\binom{r}{k}$ doubly stochastic matrices $A(i_1, \ldots, i_k)$.

One way of specifying the matrix $A(i_1, \ldots, i_k)$ is to assign a probability to each of the $k!$ possible judgment orderings of the true order statistics $x_{i_1:r}, \ldots, x_{i_k:r}$. If we let $q(i_{\pi(1)}, \ldots, i_{\pi(k)})$ denote the probability corresponding to the ordering $x_{i_{\pi(1)}} < \cdots < x_{i_{\pi(k)}}$, then the stochastic matrix $A(i_1, \ldots, i_k)$ can be computed as

$$A(i_1, \ldots, i_k) = \frac{1}{k!} \sum_{\pi \in S_k} q(i_{\pi(1)}, \ldots, i_{\pi(k)}) \text{Per}(\pi(1), \ldots, \pi(k)),$$

where $\text{Per}(\pi(1), \ldots, \pi(k))$ is the permutation matrix whose nonzero entries are $(\pi(i), i)$ for $i = 1, \ldots, k$. The probabilities $q(i_{\pi(1)}, \ldots, i_{\pi(k)})$, in turn, may be assigned by associating a positive weight $w(\pi)$ with each $\pi \in S_k$ and then normalizing these weights so that they become probabilities.

We may ensure that a one-parameter family of rankings spans the range from random rankings to perfect rankings by ensuring that at one extreme of the parameter space, the weights $w(\pi)$ are the same, while at the other extreme, the weights $w(\pi)$ are negligible except when $\pi$ is the identity. This may be done by first assigning a score $h\left(\frac{i}{r+1}\right)$ to each order statistic $x_{i:r}$. The weight associated with a particular ordering $x_{i_{\pi(1)}}:r, \ldots, x_{i_{\pi(k)}}:r$ of the order statistics $x_{i_1:r}, \ldots, x_{i_k:r}$ might then be given
by
\[ w(\pi) = \exp \left\{ \rho \sum_{j=1}^{k} j \cdot h \left( \frac{i_{\pi(j)}}{r+1} \right) \right\}, \]
where \( \rho \in [0, \infty) \) is the parameter for the family. When \( \rho = 0 \), the weights are all equal to 1, meaning that the judgment rankings are completely random. When \( \rho \) approaches \( \infty \), on the other hand, the probability will be concentrated on the single permutation \( \pi \) which has the highest value for \( \sum_{j=1}^{k} j \cdot h \left( \frac{i_{\pi(j)}}{r+1} \right) \). One can guarantee that the permutation with the highest value is the true ranking by choosing \( h \) to be a strictly increasing function, such as the quantile function for some continuous distribution.

To illustrate this method, we present computational details for one choice of \( h \) when \( k = 2 \) and \( r = 3 \). We also present a trio of figures showing how one could create one-parameter families of imperfect ranking models that have the type of skewness desired.

**Example 5.5:** Suppose that \( k = 2 \) and \( r = 3 \), and let the function \( h(u) \) be given by \( h(u) = u^2 \). This choice of \( h \) is appropriate if we think that higher values may be easier to order accurately than lower values. If the sample is \( \{x_{1,3}, x_{2,3}\} \), then the scores associated with the two order statistics are \( h(1/4) = 1/16 \) and \( h(1/2) = 1/4 \). Thus, the weight for the permutation \( (1, 2) \in S_2 \) is given by \( e^{(1/16)+2(1/4)} = e^{(9/16)} \), while the weight for the permutation \( (2, 1) \) is given by \( e^{(2(1/16)+1(1/4))} = e^{(3/8)} \). The probability associated with the correct ordering is then simply the fraction
\[ \frac{e^{(9/16)} \rho}{e^{(9/16)} \rho + e^{(6/16)} \rho} = \frac{e^{(3/16)} \rho}{1 + e^{(3/16)} \rho}. \]
This tells us that the doubly stochastic matrix $A(1, 2)$, as a function of the power $\rho$, is simply

$$A(1, 2) = \frac{1}{1 + e^{(3/16)\rho}} \begin{pmatrix} e^{(3/16)\rho} & 1 \\ 1 & e^{(3/16)\rho} \end{pmatrix}.$$ 

Similar calculations show that the values for the matrices $A(1, 3)$ and $A(2, 3)$ are

$$A(1, 3) = \frac{1}{1 + e^{(1/2)\rho}} \begin{pmatrix} e^{(1/2)\rho} & 1 \\ 1 & e^{(1/2)\rho} \end{pmatrix}$$

and

$$A(2, 3) = \frac{1}{1 + e^{(5/16)\rho}} \begin{pmatrix} e^{(5/16)\rho} & 1 \\ 1 & e^{(5/16)\rho} \end{pmatrix}.$$ 

Computing and averaging the matrices $N(1, 2)$, $N(1, 3)$, and $N(2, 3)$ as described in Section 5.2 then gives the matrix $P$ as

$$P = \frac{1}{3} \left( \frac{e^{(3/16)\rho}}{1 + e^{(3/16)\rho}} + \frac{e^{(1/2)\rho}}{1 + e^{(1/2)\rho}} + \frac{e^{(5/16)\rho}}{1 + e^{(5/16)\rho}} \right) + \frac{1}{3} \left( \frac{e^{(1/2)\rho}}{1 + e^{(1/2)\rho}} + \frac{e^{(5/16)\rho}}{1 + e^{(5/16)\rho}} \right) + \frac{1}{3} \left( \frac{1}{1 + e^{(1/2)\rho}} + \frac{e^{(5/16)\rho}}{1 + e^{(5/16)\rho}} \right).$$

As a second illustration of this method, we considered three different choices of the function $h$. We considered a symmetric choice of $h$ given by $h_1(u) = u$, and we also considered right-skewed and left-skewed choices given by $h_2(u) = (1 - u)^{-1}$ and $h_3(u) = -u^{-1}$, respectively. The results of using these choices of $h$ together with $k = 3$, $r = 4$, and various values of the power $\rho$ are illustrated in Figures 5.1 to 5.3. In each of the three figures, the underlying distribution is taken to be uniform on $[0,1]$, and the plotted lines are the density $f[1]$ of the first judgment order statistic and the sum $f[1] + f[2]$ of the densities for the first and second judgment order statistic. Thus, each graph contains three areas, one corresponding to the density for each judgment order statistic. For each of the three figures, the quality of the judgment rankings...
ranges from random (when the power is $\rho = 0$) to nearly perfect (when the power is $\rho = 8$). From Figure 5.1, we see that symmetry in the function $h$ leads to symmetry in the imperfect rankings model. In Figures 5.2 and 5.3, meanwhile, we see that skewness in $h$ corresponds to skewness in the imperfect rankings. The one-parameter family of imperfect rankings models featured in Figure 5.2 might be appropriate if one believed that it was easy to order values corresponding to high quantiles, but difficult to order smaller values, while the one-parameter family featured in Figure 5.3 might be appropriate if one believed that it was easy to order values corresponding to low quantiles, but difficult to order larger values.

This single method for selecting a model is obviously not intended to exhaust the possibilities for imperfect ranking models. Instead, the goal in describing this method is simply to show that using the framework developed in Section 5.2, one-parameter families of models with desirable properties such as skewness in one particular direction can be constructed with minimal effort. It might seem that the construction of the $\binom{r}{k}$ matrices $A(i_1, \ldots, i_k)$ would be computationally intensive. This is not the case, however, since $k$ is generally chosen to be small. If $r$ is not much larger than $k$, then neither going through the $\binom{r}{k}$ choices of the indices $i_1, \ldots, i_k$ nor exploring the $k!$ possible orderings given each choice of the indices takes much computer time at all.

5.6 Conclusions

By thinking of drawing sets of size $k$ by first drawing sets of size $r$ and then subsampling, we developed a broad class of models for imperfect rankings. We then showed that essentially any reasonable imperfect rankings model is a limit of models in
Figure 5.1: Densities for the judgment order statistics for various values of the power $\rho$ when $k = 3$, $r = 4$ and $h_1(u) = u$. The underlying distribution is uniform on $[0,1]$, and the plotted lines are $f_{[1]}$ and $f_{[1]} + f_{[2]}$. 

133
Figure 5.2: Densities for the judgment order statistics for various values of the power $\rho$ when $k = 3$, $r = 4$, and $h_2(u) = (1 - u)^{-1}$. The underlying distribution is uniform on $[0,1]$, and the plotted lines are $f_{[1]}$ and $f_{[1]} + f_{[2]}$. 
Figure 5.3: Densities for the judgment order statistics for various values of the power $\rho$ when $k = 3$, $r = 4$, and $h_3(u) = -u^{-1}$. The underlying distribution is uniform on $[0,1]$, and the plotted lines are $f_{[1]}$ and $f_{[1]} + f_{[2]}$. 
the class. These results suggest that when evaluating the performance under imperfect rankings of inference procedures based on RSS, one may safely restrict attention to the models in the class. Being able to restrict attention to this class is perhaps not an advantage if results must be obtained by simulation. However, because of the simple form of the judgment order statistic CDFs for models in the class, we suspect that it may sometimes be possible to compute efficiencies analytically when a model in the class is assumed. We also presented a simple method that allows the selection of one-parameter families of models that both smoothly span the range from random rankings to perfect rankings and have properties such as left-skewness or right-skewness that are believed to hold for the imperfect rankings in a given problem.
CHAPTER 6

SUMMARY AND FUTURE WORK

In this chapter, we summarize what was shown in Chapters 2 to 5. In the process, we also describe some future work that could be carried out in areas related to those discussed in this dissertation.

In Chapter 2, we developed two computational algorithms for computing probabilities of the form $P(X_1 \leq \cdots \leq X_n)$, where $X_1, \ldots, X_n$ are independent, fully-specified random variables. The first of these algorithms was an exact algorithm that allows computation of $P(X_1 \leq \cdots \leq X_n)$ provided that the random variables $X_1, \ldots, X_n$ and some sequence of break points $a_1 < \cdots < a_K$ satisfy an EC condition. The second algorithm was an extrapolation algorithm that allows computation of probabilities of the form $P(u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq U_{k_n:m_n} \leq v_n, U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n})$, where $(u_1, \ldots, u_n) \in \mathbb{R}^n$, $(v_1, \ldots, v_n) \in \mathbb{R}^n$, the random variables $U_{k_1:m_1}, \ldots, U_{k_n:m_n}$ are independent, and $U_{k:m}$ has a $Beta(k, m + 1 - k)$ distribution. In developing the second algorithm, we considered two approaches, one based on extrapolating from the exact algorithm and one based on a product trapezoid rule for integrating over a simplex that was developed by Lyness and Puri (1973). Our proof in Section 2.2 that the exact algorithm works was completely rigorous, as was our development of the product trapezoid rule algorithm in Section 2.4. For the extrapolation rule described
in Section 2.3 and the hybrid rule developed in Section 2.4, however, we were able to provide a rigorous proof of the algorithm only in the case \( n = 2 \). It would be nice to have rigorous proof that these algorithms hold more generally. While it seems likely that such results would need to rely on properties of the Bernoulli numbers, we would not rule out the possibility that a more statistical proof could be found.

In Chapter 3, we showed how to find optimal distribution-free confidence bands for a continuous CDF based on either a simple random sample or a sample consisting of independent order statistics. In the SRS case, we were able, using the Brunn-Minkowski Inequality, to derive not only necessary and sufficient conditions for optimality, but also conditions that guarantee uniqueness. While we suspect that uniqueness is even easier to obtain in the more general independent order statistics case, conditions for equality in the Prékopa-Leindler Inequality are not yet available. It would be nice to derive rigorous uniqueness results for this setting. These results might be obtained by deriving necessary conditions for equality in the Prékopa-Leindler Inequality, but easier results may be available if we focus specifically on the case of confidence bands. Future work in the area of confidence bands might also address the problem, foreshadowed by Wald and Wolfowitz (1939), of finding confidence bands that are most powerful against specific types of alternatives.

In Chapter 4, we introduced intentionally representative sampling (IRS) and proved that it provides an unbiased estimate of a population mean. We also showed via simulation studies that it tends to outperform both SRS and RSS in settings where those schemes are typically used. By itself, however, an estimate of a population mean is of only limited interest. If that estimate is accompanied by a standard error, however, then it becomes possible to carry out hypothesis tests and produce
confidence intervals for the unknown mean. Thus, it would be of interest to derive one or more new variance estimators to go with the IRS mean estimator. As described in Section 4.3, variance estimates are clearly available if one draws multiple samples. However, it would be of greater interest to obtain variance estimates that are based on a single sample.

In Chapter 5, we developed a broad new class of models for imperfect rankings, and we showed that essentially any reasonable model for imperfect rankings is a limit of models in this class. We also suggested that for certain inference procedures, the new class of models may allow analytical computation of relative efficiencies. It would be good to make that argument more concrete by finding some commonly used statistical procedures for which efficiency computations can be made analytically. We could also make a more compelling argument for the value of the new class of models by showing that, for actual ranked-set sample data, models within the new class provide a good fit. Ideally, this good fit would occur for small values of the first-stage sample size parameter $r$. 
APPENDIX A

R CODE

In this section, we provide R functions for implementing the best of the algorithms developed in Chapter 2. These functions are also available from the author. The first function, `rect.unif`, implements the algorithm for computing rectangle probabilities for standard uniform order statistics that was developed in Section 2.2. The second function, `prodtrap.rom`, implements the product trapezoid rule algorithm developed in Section 2.4. The final function, `rect.rss`, implements the fast, flexible hybrid algorithm developed at the end of Section 2.4.

The function `rect.unif` can be used for computing probabilities of the form $P(u_1 \leq U(1) \leq v_1, \ldots, u_n \leq U(n) \leq v_n)$, where $(u_1, \ldots, u_n) \in \mathbb{R}^n$, $(v_1, \ldots, v_n) \in \mathbb{R}^n$, and $(U(1), \ldots, U(n))$ is a vector of standard uniform order statistics. The function applies the exact algorithm developed in Section 2.2. The input to the function consists of the pair of vectors $u$ and $v$, which need to be the same length, and the output is simply the probability sought.

```r
rect.unif<-function(u,v){
  #Here we make sure that the vectors u and v are nondecreasing.
  #We also make sure that the bounds are all in the interval [0,1].
  n<-length(u)
  ...
}
```

140
\begin{verbatim}

u <- ifelse(u > 0, u, 0)
v <- ifelse(v < 1, v, 1)
for (i in 1:(n-1)){
    u[i+1] <- max(u[i], u[i+1])}
for (i in seq(n-1, 1, -1)){
    v[i-1] <- min(v[i-1], v[i])}

value <- 0

# Here we check that the probability is nonzero.
if (min(v-u) > 0){

# The vector bp contains the sequence of break points.
# The vectors left, hi, and lo determine which values p(t,i) are zero
# and which are nonzero.

bp <- unique(sort(c(u, v)))
nint <- length(bp) - 1
mp <- (bp[1:nint] + bp[2:(nint+1)])/2
left <- c(rep(0, nint))
for (i in 1:nint){
    vec <- (1:n) * (u < mp[i] & mp[i] < v)
    vec <- ifelse(vec == 0, (n+1), vec)
    left[i] <- min(vec)
}
lo <- c(rep(0, n))
hi <- c(rep(0, n))
for (i in 1:n){
    vec <- 1:nint * (u[i] < mp & mp < v[i])
    lo[i] <- min(vec[vec != 0])
    hi[i] <- max(vec[vec != 0])
    p <- bp[2:(nint+1)] - bp[1:nint]

# The matrix s is the matrix of state probabilities.

s <- matrix(c(rep(0, n*nint)), nint, n)
newtot <- c(rep(0, nint))

# Here we find the state probabilities at time t=1.

t <- 1
for (i in lo[1]:hi[1]){  
    s[i, 1] <- p[i]
    newtot[i] <- -s[i, 1]
    totbelow <- c(rep(0, nint))

# Now we apply the recursion given in equation (2.10).

for (t in 2:n){
    if (lo[t-1] < lo[t]){  
        totbelow[lo[t]] <- sum(newtot[lo[t-1]: (lo[t]-1)])
    if ((lo[t]+1) <= hi[t]){

}

\end{verbatim}

for (j in (lo[t]+1):hi[t]){
    totbelow[j]<-totbelow[j-1]+newtot[j-1]}
newtot<-c(rep(0,nint))
for (i in seq(hi[t],lo[t],-1)){
    if ((t-left[i]+1) >= 2){
        for (j in seq((t-left[i]+1),2,-1)){
            s[i,j]<-s[i,j-1]*p[i]*t/j}
        s[i,1]<-totbelow[i]*p[i]*t
        newtot[i]<-sum(s[i,1:(t-left[i]+1)])}}}

#The final sum is computed here.
value<-sum(newtot)
list(prob=value)}

The function **prodtrap.rom** can be used for computing probabilities of the form
\[ P(U_{k_1:m_1} \leq \cdots \leq U_{k_n:m_n}) \], where the random variables \( U_{k_1:m_1}, \ldots, U_{k_n:m_n} \) are independent, the values \( \{k_i\} \) and \( \{m_i\} \) are all positive integers, and each \( U_{k_i:m_i} \) has a
\( \text{Beta}(k_i, m_i + 1 - k_i) \) distribution. The function applies the product trapezoid algorithm that was developed in Section 2.4. The input to the function consists of the
vector \( k \), the vector \( m \), and the tolerance for the relative error. The output consists
of an approximation to the probability sought, a conservative estimate of the relative error in that approximation, and the maximum number of intervals used. The
function **prodtrap** computes the basic product trapezoid rule approximations that
are used by **prodtrap.rom**.

prodtrap<-function(kvec,mvec,nint){
    ss<-length(kvec)
    #Here we find the state probabilities at time t=1.
    #The vector v is the vector of state probabilities.

    t<-1
    x<-(0:nint)/nint
    v<-dbeta(x,kvec[t],mvec[t]+1-kvec[t])
    v[1]<-v[1]/2
    v[nint+1]<-v[nint+1]/2
    tot<-c(rep(0,nint+1))
#Here we apply the recursion.

```r
for (t in 2:ss){
tot[1]<-0
for (i in 2:(nint+1)){
tot[i]<-sum(v[1:(i-1)])
tot<-tot+c(v[1:nint],0)/2
v<-tot*dbeta(x,kvec[t],mvec[t]+1-kvec[t])
}
v[nint+1]<-v[nint+1]/2
}
```

#Now we find the approximation to the integral.

```r
prob<-sum(v)/(nint**ss)
prob}
```

```r
prodtrap.rom<-function(kvec,mvec,tol){

#The value eest is the relative error estimate.
#The value expo keeps track of the number of intervals nint.
#The matrix est is the triangular array of approximations
# produced by Romberg integration.

eest<-1
expo<-1
est<-matrix(c(rep(0,144)),12,12)

#Here we check whether the error is less than the tolerance.

while ((eest > tol) && (expo < 13)){
nint<-2**(expo)
est[expo,1]<-prodtrap(kvec,mvec,nint)
}

#Here we find the higher-order Romberg approximations.

if (expo > 1){
  for (i in 2:expo){
    fac<-4**(i-1)
est[expo,i]<-(fac*est[expo,i-1]-est[expo-1,i-1])/(fac-1))
    best<-est[expo,expo]
    if (expo > 1){
      eest<-abs(est[expo-1,expo-1]/best-1)
      expo<-expo+1
    }
  out<-list(prob=best,rel.err=eest,num.int=nint)
  out}
}
```

The function `rect.rss` can be used for computing probabilities of the form $P(u_1 \leq U_{k_1:m_1} \leq v_1, \ldots, u_n \leq U_{k_n:m_n} \leq v_n, U_{k_{n+1}:m_{n+1}} \leq \cdots \leq U_{k_n:m_n})$. It applies the hybrid
algorithm developed at the end of Section 2.4. The input to the function consists of the vectors $u$, $v$, $k$, and $m$ and the tolerance for the relative error. The output consists of an approximation to the probability sought, a conservative estimate of the relative error in that approximation, and the maximum number of intervals used.

```r
rect.rss<-function(u,v,kvec,mvec,tol){
  #Here we make sure that the vectors u and v are nondecreasing.
  #We also make sure that the bounds are all in the interval [0,1].
  n<-length(u)
  u<-ifelse(u>0,u,0)
  v<-ifelse(v<1,v,1)
  for (i in 1:(n-1)){
    u[i+1]<-max(u[i],u[i+1])
  }
  for (i in seq(n-1,1,-1)){
    v[i-1]<-min(v[i-1],v[i])
  }
  best<-0
  #The vector bp contains the sequence of break points.
  #The vectors hi and lo determine which values $p(t,i)$ are zero
  # and which are nonzero.
  bp<-unique(sort(c(u,v)))
  nint<-length(bp)-1
  mp<-(bp[1:nint]+bp[2:(nint+1)])/2
  lo<-c(rep(0,n))
  hi<-c(rep(0,n))
  for (i in 1:n){
    vec<-1:nint*(u[i]<mp & mp<v[i])
    lo[i]<-min(vec[vec != 0])
    hi[i]<-max(vec[vec != 0])
  }
  #Here we check that the probability is nonzero.
  if (min(v-u) > 0){
    #The value eest is the relative error estimate.
    #The value expo keeps track of the number of intervals nint.
    #The matrix est is the triangular array of approximations
    # produced by Romberg integration.
    eest<-1
    expo<-1
    est<-matrix(c(rep(0,144)),12,12)
    #Here we check whether the error is less than the tolerance.
while ((eest > tol) && (expo < 13)){

#Here we find the state probabilities at time t=1.
#The vector s is the vector of state probabilities.

s<-c(rep(0,nint))
t<-1
pvec<-pbeta(bp,kvec[t],mvec[t]+1-kvec[t])
for (i in lo[1]:hi[1]){  
s[i]<-pvec[i+1]-pvec[i]}
totbelow<-c(rep(0,nint))

#Here we apply the recursion.

for (t in 2:n){
  pvec<-pbeta(bp,kvec[t],mvec[t]+1-kvec[t])
  if (lo[t-1] < lo[t]){  
    totbelow[lo[t]]<-sum(s[lo[t-1]:(lo[t]-1)])
  }  
  if ((lo[t]+1) <= hi[t]){  
    for (j in (lo[t]+1):hi[t]){  
      totbelow[j]<-totbelow[j-1]+s[j-1]}
  }  
  for (i in seq(hi[t],lo[t],-1)){  
    s[i]<-(totbelow[i]+s[i]/2)*(pvec[i+1]-pvec[i])}

value<-sum(s[lo[n]:hi[n]])
est[expo,1]<-value

#Here we find the higher-order Romberg approximations.

if (expo>1){
  for (i in 2:expo){  
    fac<-4**(i-1)
    est[expo,i]<-(fac*est[expo,i-1]-est[expo-1,i-1])/(fac-1)
  }  
  best<-est[expo,expo]
  if (expo> 1){  
    eest<-abs(est[expo-1,expo-1]/best-1)}
}

#Now we update the vectors bp, lo, and hi.

expo<-expo+1
for (i in seq(nint+1,2,-1)){  
  bp[2*i-1]<-bp[i]
  bp[2*i-2]<-(bp[i]+bp[i-1])/2}
for (i in 1:n){  
  lo[i]<-2*lo[i]-1  
  hi[i]<-2*hi[i]}
nint<-nint*2}
out<-list(prob=best,rel.err=eest,num.int=nint/2)
out}
BIBLIOGRAPHY


perfect judgment rankings, Under preparation.


[34] Sebah, P. and Gourdon, X. (2002). Introduction on Bernoulli’s numbers, Available at *numbers.computation.free.fr/Constants/constants.html*.

probability that the empirical distribution function lies between two distribution

and Johnson, N., eds., *Breakthroughs in Statistics, Volume II: Methodology and
Distribution*, New York: Springer-Verlag.

with application to estimating distribution functions, *Journal of the American
Statistical Association*, **83**, 374-381.


[40] Zieliński, R. and Zieliński, W. (2005). Best exact nonparametric confidence in-