ARO Report 92-2

PROCEEDINGS OF THE THIRTY-SEVENTH
CONFERENCE ON THE DESIGN OF
EXPERIMENTS IN ARMY RESEARCH
DEVELOPMENT AND TESTING

Sponsored by
The Army Mathematics Steering Committee
on Behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT AND
ACQUISITION

U.S. Army Research Office

Report No. 92-2

May 1992

PROCEEDINGS OF THE THIRTY-SEVENTH CONFERENCE

ON THE DESIGN OF EXPERIMENTS,

Sponsored by the Army Mathematics Steering Committee

"Conference on the Design of Experiments
in Army Research Development and
Testing.

HOST

U.S. ARMY ENGINEERING WATERWAYS EXPERIMENT STATION
VICKSBURG, MISSISSIPPI

23-25 OCTOBER 1991

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The host for the Thirty-Seventh Army Design of Experiments Conference (DOE) was the Waterways Experiment Station (WES) at Vicksburg, Mississippi. Dr. Billy Johnson did an outstanding job in managing all the local arrangements. Dr. Robert Whalen, Technical Director of WES, kindly provided a 5:00 to 7:30 PM tour of the WES laboratories on the second day of the conference.

This conference is one of the two Army conferences that the Mathematical Sciences Division, of the Army Research Office, holds for Army scientists. Although the title of the DOE Conference suggests that it is concerned only with design of experiments, it is actually a statistic conference and covers a wide range of topics in statistics.

An occasional talk in high level mathematical statistics or probability theory intrudes on the scene, but generally topics such as these, that are heavily mathematical, are given in the Army Conference on Applied Mathematics and Computer Sciences. The DOE Conference is very much a teaching conference having clinical papers with a panel of discussants, analytic papers (application of known theory), the usual technical talks, and invited papers.

The keynote address was given by Professor G. Watson. He and the other invited speakers are listed below.

Speaker and Affiliation	Title of Address
Dr. William H. Du Mouchel BBN Software Products	Bayesian Meta Analysis
Professor Ian McKeague Florida State University	Identification of Nonlinear Time Series From First Order Cumulative Characteristics
Professor Emanuel Parzan Texas A&M University	Change Analysis
Professor Isabella Verdenelle Carnegie-Mellon University	A Bayesian Look at Experimental Design
Professor Geoffry Watson Princeton University	The Use of Simulation in Statistical Inference
Professor Edward Wegman George Mason University	The Straight Scoop on Wavelets and Nonparametric Function Estimation

James R. Thompson, Professor of Statistics at Rice University, was named recipient of the tenth U.S. Army Wilks Award for Contributions to Statistical Methodologies in Army Research, Development, and Testing. Professor Thompson's interest in confronting important Army problems and his willingness to interact with Army researchers is well established. His work, with Dr. M.S. Taylor, on data based nonparametric density estimation arising from modeling of multivariate ballistic data and his faithful support of the Design of Experiments Conference including presentation of the tutorial: Density Estimation, Modeling and Simulation: Studies in Empirical Model Building at the Thirty-Second Conference on the Design of Experiments Conference are signal contributions.

Each year a two-day tutorial precedes these conferences. This year the topic was recent developments in "Time Series Analysis", and was given by Professor Joseph Newton from Texas A&M University. He gave an excellent, information packed short course. His lecture notes, entitled "Applied Time Series Analysis" are reproduced in these proceedings.

The host for these conferences is the Army Mathematics Steering Committee. The members of this committee are duly aware of all the effort that goes into making these conferences such memorable events. Their thanks go to all those in attendance. The speakers in particular, need recognition for the time they spent in preparing and delivering their scientific papers.

Program Committee

Carl Bates	Robert Burge	Francis Dressel
Eugene Dutoit	Malcolm Taylor	Carl Russell
Douglas Tang	Henry Tingey	Jerry Thomas
Barry Bodt	Jock O. Grynovicki	

TABLE OF CONTENTS*

       *This table of Contents contains only the papers
       that are published in this manual.

vi

THIRTY-SEVENTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT, TESTING

23-25 OCTOBER 1991

HOST

U.S. ARMY ENGINEERING WATERWAYS EXPERIMENT STATION
VICKSBURG, MISSISSIPPI

Wednesday, 23 October 1991

0800 - 1400   REGISTRATION

0845 - 0900   **CALL TO ORDER:** Billy H. Johnson, U.S. Army Engineering
Waterways Experiment Station

OPENING REMARKS

0900 - 1100   GENERAL SESSION I

Chairperson: Jerry Thomas, U.S. Army Ballistic Research
Laboratory

**KEYNOTE ADDRESS: THE USE OF SIMULATION IN
STATISTICAL INFERENCE**
Geoffrey S. Watson, Princeton University

**THE STRAIGHT SCOOP ON WAVELETS AND
NONPARAMETRIC FUNCTION ESTIMATION**
Edward J. Wegman, George Mason University

1100 - 1130   BREAK

1130 - 1210   CLINICAL SESSION

Chairperson: Douglas S. Tang, Walter Reed Army Institute of
Research

Panelists: Emanuel Parzen, Texas A&M University
Jayaram Sethuraman, Florida State University
Nozer Singpurwalla, George Washington University

SENSITIVITY OF COMPONENT RELIABILITY FROM FATIGUE LIFE COMPUTATIONS
Donald M. Neal, William T. Matthews, Trevor D. Rudalevige, and Mark D. Vangel, U.S. Army Materials Technology Laboratory

1210 - 1330   **LUNCH**

1330 - 1530   **TECHNICAL SESSION* I**

     Chairperson: Linda L.C. Moss, U.S. Army Ballistic Research Laboratory

     ANALYSIS OF BALLISTIC DATA RANDOMIZATION
     Malcolm S. Taylor, U.S. Army Ballistic Research Laboratory

     A RANDOMIZATION TEST FOR COMPARING KINETIC ENERGY PENETRATORS
     Barry Bodt, U.S. Army Ballistic Research Laboratory

     ONE-SIDED TOLERANCE LIMITS FOR LOG-CONVEX AND IFR DISTRIBUTIONS
     Mark G. Vangel, U.S. Army Materials Technology Laboratory

     INFERENCE AND PREDICTIONS FROM THE POISSON-WEIBULL PROCESS BASED ON EXPERT OPINION AND COUNT DATA
     Nozer Singpurwalla, George Washington University

     ROBUST STATISTICAL DECISIONS (AN EMPIRICAL INVESTIGATION)
     Eugene Dutoit, U.S. Army Infantry School

1530 - 1600   **BREAK**

1600 - 1700   **GENERAL SESSION II**

     Chairperson: William Jackson, U.S. Army Tank-Automotive Command

     A BAYESIAN LOOK AT EXPERIMENTAL DESIGN
     Isabella Verdinelli, Carnegie Mellon University

* All but one of the technical sessions has at least one Applied paper in it.

0800 - 1200   **REGISTRATION**

0800 - 1000   **TECHNICAL SESSION 2**

Chairperson: Ann E.M. Brodeen, U.S. Army Ballistic Research
Laboratory

**SMALL SAMPLE DESIGN ALLOWABLES FROM PAIRED DATA
SETS**
Donald M. Neal, Trevor D. Rudalevige, and Mark G. Vangel,
U.S. Army Materials Technology Laboratory

**MINIMAX DISTANCE DESIGNS IN THE CONTEXT OF TWO-LEVEL
FACTORIAL EXPERIMENTS**
P.W.M. John, University of Texas
M. E. Johnson, University of South Florida
L. M. Moore, Los Alamos National Laboratory
D. Ylvisaker, University of California, Los Angeles

**A COMPARISON OF SAMPLING FREQUENCY OPTIONS FOR
SKIP-LOT SAMPLING PLANS**
C. J. Hunter, Air Command Headquarters, Manitoba, Canada

**TOXIC FUMES REDUCTION MODIFICATIONS TO THE M2A2
BRADLEY FIGHTING VEHICLE**
Linda L.C. Moss and William P. Johnson, U.S. Army Ballistic Research
Laboratory

**SOME LIMITATIONS OF THE RANK TRANSFORMATION TEST
FOR INTERACTION**
W.J. Conover, Texas Tech University

1000 - 1030   **BREAK**

1030 - 1200   **TECHNICAL SESSION 3**

Chairperson: Carl B. Bates, U.S. Army Concepts Analysis
Agency

**A PARALLELIZED SIMULATION BASED ALGORITHM FOR
PARAMETER ESTIMATION**
Martin Lawera and James R. Thompson, Rice University

APPLICATIONS OF SIMEST TO BIRTH AND DEATH PROCESSES
Kathy Ensor, Rice University

ON A NEW SYSTEM OF MULTIVARIANT DISTRIBUTIONS
Kevin M. Beam, The RAND Corporation

EVALUATION OF AFATDS COMMUNICATIONS THROUGHPUT AND DELAY
Virginia A.T. Kaste, Ann E.M. Brodeen and Barbara D. Broome,
U.S. Army Ballistic Research Laboratory

1030 - 1330   **LUNCH**

1330 -1430   **TECHNICAL SESSION 4**

  Chairperson: Virginia Kaste, U.S. Army Ballistic Research
             Laboratory

  VISUALIZATION AND ANIMATION OF STATISTICAL FUNCTIONS
  David Scott, Rice University

  CIRCULARITY: VALIDATION AND ASSOCIATED DIAGNOSTICS
  Devan V. Mehortra, G.D. Searle Company
  John W. Green, DuPont Company
  Henry B. Tingey, University of Delaware

1430 - 1500   **BREAK**

1500 - 1700   **GENERAL SESSION III**

  Chairperson: Terry Cronin, U.S. Army CECOM Center for
             Signals Warfare

  CHANGE ANALYSIS
  Emanuel Parzen, Texas A&M University

  IDENTIFICATION OF NONLINEAR TIME SERIES FROM FIRST ORDER CUMULATIVE CHARACTERISTICS
  Ian McKeague, Florida State University

  Friday, 25 October 1991

0800 - 1200   **REGISTRATION**

0800 - 1000   **TECHNICAL SESSION 5**

  Chairperson: Paul Broome, U.S. Army Ballistic Research
             Laboratory

x

AN ANALYSIS OF THE EFFECT OF THE DEGRADED-STATES METHODOLOGY FOR VULNERABILITY ASSESSMENT
William E. Baker, U.S. Army Ballistic Research Laboratory

VARIANCE COMPONENTS IN TANK GUN ACCURACY RESEARCH
David W. Webb, U.S. Army Ballistic Research Laboratory

IDENTIFY EXPERIMENTAL VARIABILITY AND ITS INFLUENCE ON THE RESEARCH, DEVELOPMENT, TESTING, AND EVALUATION PROCESS
Jock O. Grynovicki, U.S. Army Human Engineering Laboratory

MODELING GUNFIRE: A PROBLEM IN MAXIMUM LIKELIHOOD ESTIMATION
Henry B. Tingey, University of Delaware

EXPERIMENTAL DESIGN FOR VALUE ADDED ANALYSIS (VAA)
William Mann and Andrew Loerch, U.S. Army Concepts Analysis Agency

1000 - 1030   **BREAK**

1030 - 1200   **GENERAL SESSION IV**

Chairperson: Barry Bodt, U.S. Army Ballistic Research Laboratory, and Chairman of the AMSC Subcommittee on Probability and Statistics

OPEN MEETING OF THE PROBABILITY AND STATISTICS SUBCOMMITTEE OF THE ARMY MATHEMATICS STEERING COMMITTEE

BAYESIAN META ANALYSIS
William H. DuMouchel, BBN Software Products

ADJOURN

PAPER BY TITLE: (It came too late to be on the Agenda). **THE VARIABILITY OF THE LOWER TOLERANCE LIMITS FOR NORMAL SAMPLES,**
Bernard Harris, University of Wisconsin

Program Committee

Carl Bates	Robert Burge	Francis Dressel
Eugene Dutoit	Malcolm Taylor	Carl Russell
Douglas Tang	Henry Tingey	Jerry Thomas
Barry Bodt		Jock O. Grynovicki

xi

Simulation In Statistical Inference

Geoffrey Watson*, Princeton University

Abstract

In this paper I will not include one topic discussed in my actual talk since the paper is already too long. The bulk of the paper describes our mathematical and computational studies of the parametric bootstrap. This is largely expository - our examples are chosen to increase our intuition about how to proceed in analogous situations. There are many uses for the parametric bootstrap but the literature gives little practical guidance when there are many parameters. We hope that this paper will also be a useful introduction to the nonparametric bootstrap which was the initial problem of bootstrap studies and is still their central interest. I believe this wonderful tool is more subtle than many think it is. The last section is original and shows how, by converting a testing problem into one of estimation, simulation leads to the solution of an important problem in paleomagnetism; bootstrapping is part of this solution.

1 Introduction

Simulation has a long history. "Student" checked his formula for the t-distribution by manual simulation - drawing numbers on slips of paper. The early volumes of Biometrika are always fascinating to read; my undergraduate "senior thesis" was a summary of studies of the effects of non-normality which I mostly found there. Many involved manual simulation. In the mid-30's Pitman explored Fisher's ideas of randomization & permutation distributions. He found their early moments by algebraic methods. However everyone was clear that one could approximate the wanted distributions numerically. But it was then very slow work.

Even when computing gradually became easier and cheaper, few of us "changed heads". Machines were just used to do computations (e.g. numerical linear algebra) that were just a bit bigger than those we were already doing.

Efron changed everything with his "bootstrap" about 13 years ago. Perhaps I should have called this talk the "Stimulation in Stat'l Inference !!!!!!!!"

The emphasis in the bootstrap literature was, and still is, on NON_PARAMETRIC methods i.e. methods that will be 100% effective whatever the situation when the sample size tends to infinity. However an earlier paper by Efron pointing out that the method of getting accuracies of maximum likelihood estimators works better with sample likelihoods than expectations is a parametric bootstrap argument!

There is now a large, rapidly growing and difficult literature - the following monographs give the key ideas and facts: Efron(1982), Beran (1991), Hall (1991). But it is very hard, even for an academic, to "catch up" with current technology. I hope that this paper will help applied statisticians to understand and perhaps use some of the basic ideas. As the "keynote" speaker for this conference, I felt it was more essential for me to address a topic vital to everyone than to talk about something on which I have the most expertize!

I drifted into this area because I had to deal with complicated parametric estimation problems where classical theory was not much help so I tried "to simulate my way out of trouble - ever the innocent optimist! I was soon told that I was just rediscovering old bootstrap results. I had not troubled to study the bootstrap before. This was partly from laziness and partly because I was put off by what I thought was its uncritical embrace for small samples - I could see its asymptotic justifications. And I was not so keen on getting non-parametric methods. I soon discovered it was much more subtle than I thought! Further I think its use for parametric problems is not only very useful but a good introduction to the use of the nonparametric bootstrap.

Section 2 will introduce bootstrap ideas via parametric problems which are so simple that one can see what bootstrapping will do without having to simulate at all! Further we work with the m.l. estimators. Of course in more complex problems one must simulate. Section 3 shows several experiments with problems where the number of parameters goes from medium to large.- these simulations were carried out by *Javier Cabrera* . In Section 4 we will show how simulation may be used to solve an important problem in palaeomagnetism. Though it is logically different from the above, bootstrapping is an essential part of the trick! The computations shown were done by *Michel Debiche* .

2 Simple examples

Example 1   *Original "naive" bootstrap.*

     Consider a sample of n from

$$f(x,\theta) = \theta^{-1} \exp(-x/\theta)$$

with arithmetic mean m , the (unbiased) mle for ø. The "Naive" bootstrap way of getting a conf. interval "$\theta < c$" of size β is , I believe , to draw N samples of size n from $f(x,m)$, find the mean m^* of each , find c_β so that $N\beta$ of the m^*'s are less than c_β; we then have the "naive" or "percentile" interval "$\theta < c_\beta$"   (1)

     This is based on the plausible belief that the distribution of m, in sampling from $f(x,\theta)$,will be APPROX the same as that of m^* when sampling from $f(x,m)$, fixed m. Now if N is large, our simulation will give the same answer that we know , in this example, from distribution theory ---- given m, m^*/m distributed as $X_{2n}^2/2n$, because we know that

     m/θ distributed as $X_{2n}^2/2n$.   (2)

Write Prob $\{ X_{2n}^2 < c_{2n}(\beta) \} = \int$. Hence the "naive" interval (1) is, when N is large, close to

$$\theta < m\, c_{2n}(\beta)/2n .$$

But directly from (2) we get the correct (if N large) interval

     Prob $\{ m/\theta > c_{2n}(1-\beta)/2n \} = \beta$,   or

     Prob $\{ \theta < m \cdot 2n/ c_{2n}(1-\beta) \} = \beta$.   (3)

For n=5 , β=.95

     $c_{2n}(\beta)/2n = 1.83$, $2n/ c_{2n}(1-\beta)=2.53$

So the naive interval is much too small- its true coverage is much less than the nominal .95. Hence the use of "NAIVE"!

     What is wrong with the "naive" argument ? -

     The distribution of m^* , given m , is NOT the same as (only like) the distribution of m , given θ .

3

It is the distributions of m^*/m, given m and m/θ, given θ, that are the same! Notationally

$$L\, m^*/m \text{ , at } m = L\, m/\theta \text{ at } \theta$$

This is true - see (2) - because the distribution of $m/\emptyset$ does not depend on θ i.e.

$$m/\theta \text{ is a PIVOTAL FUNCTION .}$$

Pretend now that we don't know that m/θ has the distribution given by (2). We can here approximate it as closely as we like by taking N large enough in a simulation study of m^*/m. Thus we can get a an almost exact confidence interval by drawing N samples of size n from the exponential distribution with mean m and and finding k so that proportion of m^*/m's $>k \approx \beta$, to good approximation. Then we may assert that it should be accurate to say that

Prob$\{ m/\theta > k \} = \beta$, or equivalently that "$\theta < m/k$" is a confidence interval of size β. When N tends to infinity we must get the exact confidence interval this way.

The moral of this is that one should use pivotal , or more practically, asymptotically pivotal, statistics when bootstrapping.

Example 2 *Iteration of the bootstrap ,to reduce bias. Prepivoting*

Consider a sample of n from $U[0.\emptyset]$, the Uniform distribution on $[0,\theta]$. Then $\hat{\theta}_0 = M = \max \{ x_1 , ..., x_n\}$ is the biased {its expectation is $[n(n+1)]\, \theta$} mle for θ with SD θ/n , approx. Can we improve it e.g. get a less biased version, by using the bootstrap? I am pretending we don't know all about this estimator but I will use this knowledge in place of large simulations.

Consider the following idea:

1/ Draw N samples of size n from $U[0,M]$, each time finding its maximum M^*_j , j = 1, ... , N , Compute their average Ave M^*

Method (a). A *corrected-once* estimator is

$\hat{\theta}_1 = M -$ Bias estimate, Ave M^*-M

$= 2M -$ Ave M^*,

since we'd guess that Ave M^*-M would be close to the true bias EM-θ. We call this the additive or linear correction method

Method (b). *Another corrected-once* estimator is

$$\hat{\theta}_1 = M^2/\text{Ave } M^*$$

This comes from multiplying M by the bias correcting factor M/Ave M*. We call this the multiplicative correction method.

In each case the subscript "1" is used because one can repeat the process.

2/ Draw N samples from $U[0,\hat{\theta}_1]$, find their maxima M** and their average.

The twice-corrected estimators could plausibly be chosen to be

$$\hat{\theta}_2 = \hat{\theta}_1 - [\text{Ave } M^{**} - \hat{\theta}_1] = 2\hat{\theta}_1 - \text{Ave } M^{**}$$

$$\hat{\theta}_2 = \hat{\theta}_1 \, \hat{\theta}_1 / \text{Ave } M^{**} = \hat{\theta}_i^2/\text{Ave } M^{**}$$

However analysis shows this intuition is not right and that we should argue as follows.

Additive - The idea that gave $\hat{\theta}_1$ was to assert

$L[\hat{\theta}_0 - \theta]$ at $\theta = L \, [M^* - \hat{\theta}_0]$ at $\hat{\theta}_0$, suggesting

$$\hat{\theta}_0 - \hat{\theta}_1 = \text{Ave} \, [M^* - \hat{\theta}_0]$$

or $\hat{\theta}_1 = \hat{\theta}_0 - [\text{Ave } M^* - \hat{\theta}_0]$.

If $\hat{\theta}_1$ is closer to θ, similar use of

$L \, [\hat{\theta}_0 - \theta]$ at $\theta = L \, [M^{**} - \hat{\theta}_1]$ at $\hat{\theta}_1$

should lead to the better

$$\hat{\theta}_2 = \hat{\theta}_0 + [\hat{\theta}_1 - \text{Ave } M^{**}], \text{ and}$$

$$\hat{\theta}_3 = \hat{\theta}_0 + [\hat{\theta}_2 - \text{Ave } M^{***}], \text{ etc.}$$

Multiplicative - idea was to assert

$L \, [\hat{\theta}_0 / \theta]$ at $\theta = L \, [M^* / \hat{\theta}_0]$ at θ_0,

or better

$L \, [\hat{\theta}_0 / \varnothing]$ at $\theta = L \, [M^{**}/\hat{\theta}_1]$ at $\hat{\theta}_1$,

so use the latter to get new $\hat{\theta}_2$ as

$$\hat{\theta}_2 = \hat{\theta}_0 \, [\hat{\theta}_1 / \text{Ave } M^{**}], \text{ and}$$

$\hat{\theta}_3 = \hat{\theta}_0 [\hat{\theta}_2 / \text{Ave } M^{***}]$, etc.

We now show some properties of these sequences of estimator of θ.

Analysis If N is large, we can avoid arithmetic by recalling that

$EM = \theta n/(n+1)$, var $M = \theta^2/n^2$ approx

Method (a)

$$E\hat{\theta}_1 = 2\theta[n/(n+1)] - \theta[n/(n+1)]^2$$

$$= \theta\{1 - (n+1)^{-2}\}$$

$$\text{var } \hat{\theta}_1 = \text{varM}\{1 + N^{-1}\theta^2(1+1/n^2)\}$$

so that bias is now slightly down to $O(1/n^2)$ while variance is essentially that of $M = \hat{\theta}_0$. The bias in $\hat{\theta}_1$ is now smaller than its SD.

$$E\hat{\theta}_2 = E\hat{\theta}_0 + [E\hat{\theta}_1 - \text{EAve } M^{**}]$$

$$= E\hat{\theta}_0 + E\hat{\theta}_1[1 - n/(n+1)\}^{\wedge}$$

$$E\hat{\theta}_2 = \theta\{1 - 1/(n+1)^3\},$$

an improvement over $E\hat{\theta}1 = \{1 - 1/(n+1)^2\}$ but not worth having since the SDs of $\hat{\theta}_1$ and $\hat{\theta}_2$ are $O(1/n)$.

Method (b)

$$E\hat{\theta}_1 = EM \, E\{M^2/\text{Ave } M^* | M\}$$

$$= E_M M^2 / EM^*, \text{ for N large},$$

$$= E \, M^2/[n/(n+1)]M$$

$$= [(n+1)/n] \, EM$$

$$= \theta$$

so the BIAS is Zero

$$\text{var } \hat{\theta}_1 = \theta^2/n(n+2), \text{ approx}.$$

$$E\theta_2 = E\hat{\theta}_0 \hat{\theta}_1 / \text{Ave } M^{**}$$

$$= E \hat{\theta}_0 \hat{\theta}_1 / E \, M^{**}, \text{ N large}$$

$$= E \hat{\theta}_0 \hat{\theta}_1 / [n/(n+1)] \, \emptyset 1,$$

$$= [(n+1)/n] \, E\hat{\theta}_0$$

$$= \theta, \text{ as for } \hat{\theta}_1$$

6

Thus the multiplicative correction method is better than the additive but both reduce the bias. Why ? Because it is based on a correct assertion - and this is so because here $M/\emptyset$ is a pivot.

Now we look at CONFIDENCE INTERVALS FOR $\emptyset$ If we use the multiplicative approach , we expect here to get the right answer because $M/\emptyset$ is a pivot i.e.

$$L (M / \theta) \text{ at } \theta = L (M^* / M \text{ at } M. \tag{4}$$

The cdf of both random variables is k^n , $0 < k < 1$.

Let $q\beta(M^*)$ be estimated β-quantile of M^* , found from the N samples of size n drawn from $U(0,M)$ i.e. $N\beta$(approx) of the M^*'s are less than or equal to this q. Belief in (4) leads us to assert that

$$\text{Prob} \{ M/\theta < q\beta(M^*)/M \} = \beta \tag{5}$$

Now if we let N tend to infinity , for any fixed M, $q\beta(M^*)/M$ tends to k where $k^n = \beta$. So that $M/\emptyset < q\beta(M^*)/M$ or $\emptyset > M^2/ q\beta(M^*)$ gives a confidence interval with exactly the coverage desired, β!

However if, in the same problem, we use the linear method and the first round of simulations only ,we would not expect to get an exact confidence interval. In fact we would give as the one-sided β-interval (100β%)

$$M - q\beta(M^* - M) < \emptyset ,$$

or

$$\emptyset > 2M - q\beta(M^*) .$$

Now as N tends to infinity , M being fixed, we have just seen that $q\beta(M^*)$ tends to k M. Thus the interval becomes

$$\theta/M > 2-k , k^n = \beta, 0 < k < 1 .$$

The actual coverage of this interval is easily seen to be

$$(2-k)^{-n} = (2- \beta^{1/n})^{-n} ,$$

which is tabled below

	$\beta = .9$	$\beta = .95$	$\beta = .99$
n=4	.90244	.95062	.99002
8	.90123	.95031	.99001
10	.90099	.95025	.99001
16	.90062	.95016	.99000
20	.90050	.95012	.99000

Although we have made intervals from a nonpivital function, the coverages are very accurate. We may not be so lucky in other cases.

7

Iterative bias correction is worthwhile here because the variance is so small!

Remark 1

This is an example of non-regular m.l. Thus the variance of the estimator cannot be obtained from the likelihood and the limiting distribution is non-normal. In fact the studentized form

$t = [\phi-M]/[M/n]$ is exactly pivotal and has a standard exponential distribution when n tends to infinity. But to do this one needs to know the formula for the SD of M. In the regular case one could use the negative derivative of the likelihood to get the correct divisor.

To proceed this way here one would have to use a double bootstrap - in effect to find the variance formula.

Remark 2

Had we studied the estimation of μ from a sample of n from $N(\mu,1)$, of course the additive method would be best since sample mean - μ = m-μ is then pivotal. If we have $N(\mu,\beta^2)$ its distribution depends on β, and Student's t is a pivot now. See Example 3.

Remark 3

For FINITE Simulation size N, we keep adding a little variance and so mess things up a bit

Remark 4

CAN WE FIX UP A NONPIVOTAL FUNCTION? e.g. θ-M in EX 2?
Beran (1988) suggests the use of an idea he calls PREPIVOTING.

He calls the function of estimator and parameter that we use (e.g. θ-M , M/θ) a "root". In general call it $r(\hat{\theta}, \theta)$.

Recall that if rv X has cdf (continuous) F(x), F(X) is uniformly distributed on [0,1] so Prob { $F(X)<\alpha$} = α.

Suppose in some general problem we knew the cdf of the root $r(\hat{\theta}, \theta)$ to be H (r, θ). Then

H [$r(\hat{\theta}, \theta), \theta$] is U[0,1]

so that the set { θ: H [$r(\hat{\theta}, \theta), \theta$] $< \alpha$} has probability α . (6)

The trick then is to find or approximate the cdf H. Then (6) should give a confidence interval of the right coverage; note that theoretically there are many intervals with the right coverage, some more sensible than others.

8

One needs two levels of sampling i.e. a double (not iterated) bootstrap.

First try it in the Uniform of Example 2. Use sampling from U $(0, M)$ to find approximately the cdf of the root $M-M^*$. Call it $H(k,M)$. One might hope that

$$H(\theta-M,M) \text{ approx } U(0,1) \text{ so that}$$

$$\{\theta: H(\theta-M,M)<\alpha\} = \text{Conf Int. approx size } \alpha .$$

If N is very large , H will be what we get from distribution theory: $1-(1-k/M)^n$. Then Beran's interval would be

$$\{\theta: \ 1-[1-(\theta-M)/M]^n < \alpha \}$$

or

$$\{\theta: \ \theta/M < 2-(1-\alpha)^{1/n} .$$

But the probability of this last statement being true is was tabulated above - it was surprisingly close to α.

Thus here Prepivoting has converted a nonpivotal function $\theta-M$ into a pivotal one M/θ and lead to almost exact confidence intervals. With finite samples we should approximate this happy state.

More Generally

Draw N [1] samples from $f(x, \hat{\theta})$, obtaining N [1] estimates which we will suppose when ordered from least to largest to be

$$\theta^*1, ... , \theta^* N [1] .$$

Now draw N [2] samples from the density $f(x, \theta^*_i)$, computing each time the estimator θ^{**} and build their empirical distribution. Call it $H(., \theta^*_i)$. With this done for all N[1] values of i, we compute the N[1] values of $H(\hat{\theta}, \theta^*i)$.

If $H(\hat{\theta}, \theta^*_i) < \alpha$, accept θ^*_i ; if not, reject It. If the simulation sizes are large enough, these values should give a simple conf. int. for θ, obtained by a *double layer of sampling* !!

Notice how the arithmetic can get out of hand here- we need $O(N^2)$ samples

Example 3

Sample of n from $N(\mu,\beta^2)$, mean m, variance s^2. To find an interval for μ. *Suppose we don't know how to studentize and we start with the root*

9

$$n^{1/2}[m-\mu] .$$

By extensive sampling from $N(m,s^2)$ we will find that the cdf of the root $n^{1/2}[m^*-m]$ is $\tilde{I}(x/s)$ where $\tilde{I}$ is the cdf of the standard normal. Hence Beran's method says: use confidence set

$$\{\mu: \tilde{I}(\,n^{1/2}[m-\mu]\,/s) < \alpha \}$$

Thus his method has again studentized (made pivotal) the initial root. Moreover the actual coverage of the interval is only slightly wrong - since it uses the standard normal rather than the t-distribution.

Abstracting the ideas so far --

We have a statistic $T = \hat{\o}o$, computed from a sample of n from $f(x , \o)$ and want a point estimator of $\o$ with little bias and also perhaps a confidence interval for $\o$. In general we won't know ET as a function of $\o$ but we assume T is a reasonable estimator so

$$ET = \o+b(\o) \text{ or } \o(1+B(\o)) .$$

(1) Draw N samples of n from $f(x ,\hat{\o}o)$, get N estimates $\o^*$ and their average $\text{Ave}\o^*$.

Do your best to find a function $g(\hat{\o},\o)$ which is pivotal or whose distribution changes slowly with $\o$. Then assume

$$L\, g(\hat{\o},\o) = L\, g(\o^*,\hat{\o})$$

Solve $g(\hat{\o},\o) = \text{Ave}\, g(\o^*,\hat{\o})$ for $\o$ and call the solution $\hat{\o}_1$. Above we used $g(x,y) = x-y$ and x/y.

(2) Repeat with the same plan , solving

$$g(\hat{\o},\o) = \text{Ave}\, g(\o^{**},\hat{\o}_1)$$

to get $\hat{\o}_2$, and so on .

This gives a sequence of point estimators whose bias should go down. It is only worth continuing if the bias reduction at least size of SD!!!!

To get confidence intervals when we are happy with the additive assertions

$$L\,[\,\hat{\o}o - \o] = L\,[\,\o^* - \hat{\o}o\,] , \text{ or maybe}$$

$$L\,[\hat{\o}o - \o] = L\,[\,\o^{**} - \hat{\o}1\,],$$ we could e.g. find the 97.5% quantile of the N values of $\o^* - \hat{\o}o$. Call it $q^*(.975)$.

Then we have the approximate statement

$\text{Prob}\{\,\hat{\o}o - \o < q^*(.975)\,\} = .975$ so a 95% confidence interval would be $\{\,\hat{\o}o - q^*(.975),\ \hat{\o}o - q^*(.025)\}$. The naive method would give

$$\{q^*(.025) , q^*(.975)\} .$$

The multiplicative case goes similarly .

We can use prepivoting to get a more reliable confidence interval for ø, use the step (1) results to get the cdf of g(ø*,^ ø), H(., ø) say. Then propose the set {ø: H[g(θ, θ)]< α} as a confidence set for θ of size α. Or we could use a double bootstrap to adjust empirically the intervals to get the correct coverage.

Remark 4. There might be occasions when you'd want to estimate a known function of θ. Also when you might want to use Median θ* instead of Ave θ* to get median unbiasedness. We won't pursue these directions here - they pose no new problems.

Remark 5 Rather a lot of computation is required, especially in double bootstrapping. There is a literature on methods for reducing it by monte carlo methods e.g. importance sampling. Redoing whole calculations many times, to verify a method, is sound are particularly arduous.

3 Distributions with several parameters

The number of strategies for any problem increase with the dimension of the parameter. We will here only show some experiments on iterative simulation to reduce bias. In almost all high dimensional problems, we can at most find asymptotic pivots. These are usually a consequence of calling upon the central limit theorem. I think there may be cases where this strategy leading to nonparametric results may be inferior in small samples to using some statistic related to a parental distribution, when that assumption is not wildly wrong. This is a matter for future research. This work was done with Javier Cabrera .
Example 4

Efron's papers (and others) often refer to the problem of estimating the correlation coefficient ® from a sample of n from a bivariate normal. We try to see if iteratve simulation decreases the bias.

We used 1000 samples of 5 from a bivariate normal, both means zero, unit variances and ® = 0.7. Each pair of lines in Table 1 refers to a corrected set of estimates so we have the zero-th to fifth correction. The first line each time is the mean of all 1000 estimates. The second line the standard deviation of these 1000 estimates. Thus e.g. for the means, the SD should be 1/51/2 = .447, about the value in Table 1. Evaluating the square root of the large sample variance var r =

11

(1-®2)2/n, we find 0.228, a little less than the 0.33 in our table. The divisor n-3 here gives .361.

Following down the columns we see that the mean estimates don't change as we would expect, that the standard deviations get a little closer to unity and that the average r's approach .7.

But motion stops after the first iteration.

Remarks Better results would come if we used the Fisher transformation of r,

$$z = (1/2) \log(1+r)/(1-r)$$

$$\Omega = (1/2) \log(1+\circledR)/(1-\circledR),$$

since as n $\tilde{u}$ Ç , z-Ω becomes standard normal and so is then pivotal.

Efron and others have shown that a double bootstrap gives a confidence interval with good coverage.

Here we should probably not have bothered to iterate for the means and variances. Notice that there are here 5 parameters only one of which ® was of interest.

Example 5

We now turn to one of the two problems that lead us to this way of thinking.

Suppose we have samples of n_i from Fisher distributions $F(\mu_i, \dot{}_i)$, i = 1,2 , on the unit sphere $\|x\| = 1$ in three dimensions. So $\mu_i\| = 1$ where μ_i is the mean direction and axis of rotational symmetry of the distribution and $\dot{}_i$ controls its concentration. With $x.\mu = \cos\theta$,

Fisher prob density = $\{1/4 \pi \sinh \dot{}\} \exp \{ \chi\, x.\mu \}$

The estimates are m_i , the directions of the sum R_i of the observed unit vectors &

$$k_i = (n_i-1)/(n-\| R_i \|) .$$

These χ-estimators have distributions skewed to the right and are not unbiased. We really want to estimate the angle θ between the means

$$\mu_1 \text{ and } \mu_2 .$$

This is naturally estimated from

$$m_1.m_2 = \cos\theta_0$$

but it is clear that this will lead to an overestimate which can be very severely biased. So we want to take the bias out of the estimates of θ and χ_1 and χ_2 to the extent that this is possible. To use our method we must draw N samples of n_1 and n_2 from Fishers with concentrations k_1 and k_2 and whose mean vectors are

separated by angle $\hat{\theta}_0$. We did separate runs using all additive and all multiplicative methods but saw no obvious differences. We don't know any simple but more pivotal quantities that we could have used. We used many combinations of $\chi = \chi_1 = \chi_2$ and $n = n_1 = n_2$ and θ to study point estimation. But only in a few did we calculate the actual coverage of our 95% nominal intervals. See Table 2 for some results.

Morals

Over the range of sample sizes and χ's tried so far, the methods are not good
* correction dwarfed by sd
* poor coverages from percentile intervals so need to use better method (e.g. double sampling) to get CI's.
Competitive methods - see e.g. Debiche & Watson, (1991) - also have trouble. The problem is inherently hard.

Example 5

Our other motivation was a set of functional relationship problems. Three problems have come up from my geophysical contacts.
(1) the estimation of a linear transformation with positive determinant (Gleser and Watson, 1973) from the initial and final positions of n points, measured with errors with known covariance matrices
(2) radioactive dating methods lead to fitting a linear relation between three variables subject to errors with known, possibly unequal, covariance matrices (Kent , Watson , Onstott,1990). The method given there works in any number of dimensions.
(3) The motion of sea ice deduced from the initial and final positions of n radio beacons on the ice whose positions are measured with planar errors by satellite. One could possibly assume one knew the covariance matrix of the errors. Here then we are estimating a rigid motion - displacement and rotation.
In Problem 2 above, the actual errors were small and our mathematical method lead to standard errors that agreed very well with those obtained by John Kent who kindly used our current method at our request on these data sets. Problems 1 &3 are yet to be attacked.

Instead we tried the simple but classical problem of fitting a straight line with both variables subject to normal errors. Here if there are n points with the same ß's

13

in each coordinate and the true points fall on $\eta = \alpha + \beta \zeta$, we have n (the ζ's) plus 2 (α & β) plus 1 (ß), i.e. n+3 parameters from 2n observations.

In our experiments we also included the method (denoted in our table by " mle") recommended e.g. in Fuller (1987).

In problems 1&3 the mle type methods are complicated so one might want to use simpler least squares type methods. So we tried here to start from a least squares type method, re-estimating positions of the points by their orthogonal projections onto the currently fitted line. refitting the line, getting a new ^ß and points, & so on.

Our results (Table 3) are confusing , probably because we have so many parameters to estimate in each round and again we have no pivots.

Doubt now whether one could EXTRAPOLATE from what was learnt here!

The one clear result was that in many practical situations the mle's will not behave as advertised. That this is so is seen from Figures 1 and 2. We are now trying to see if we can "fix up" the mle's!!!

This, I suspect, will in fact be the major practical use of parametric bootstrapping.

4 A simulation solution to the "Fold Test Problem".

A simplified version of a classical problem on paleomagnetism is the following. Imagine that an eruption spreads hot lava across a plane. When the iron-rich lava cools below its Curie point, it will acquire magnetization parallel to the local earth's field at that time. If the slab is not too large all parts of the slab are parallely magnetized. We imagine that no subsequent events alter this magnetization. Suppose that it is subsequently folded. Then if we go to different sites on this folded slab and measure this frozen or remnant magnetization, we will not get parallel vectors. However the angle between the normal to the bedding (old horizontal) plane and the magnetization should be preserved. If however the magnetization of a folded formation was acquired somehow after folding, the magnetization in the folded formation would be parallel at all sites. The other alternative is that the magnetization was acquired at some time during the folding process. These three alternatives are called **pre, pos, and syn. This may be clarified in Figure 1.**

For some 40 years people have tried to sort this out with significance tests. It has, since 1953, been standard to describe the scatter of paleomagnetic direction measurements by the Fisher distribution mentioned in the last section. Nowadays bootstrap methods are being applied to this problem - see eg. Fisher & Hall

(1991). It is possible to produce tests that a set of mean vectors come from distributions with parallel mean directions (null hypothesis: Post). And it is possible to test that a set of angles is constant (null hypothesis: Pre). But one can get into a logical muddle. Many references could be given - but see e.g. Tauxe, Klystra, and Constable (1991).

We prefer to think about this differently - as an attempt to estimate the relative timing of magnetization and folding. Let us assume - as is very common here - that the axis of folding is known. If it is known only with error this may be coped with by further layer of simulation. Then the current folding can be unfolded by any amount until we get back to the original which is 100% unfolding. The present is 0% unfolding. *Thus we seek the % unfolding when the rock was magnetized. Then the measured unit vectors will be more parallel than at other %'s.*

A more complicated approached should be used in practice but the following will give our idea. The basic data will be one sample of n_i unit vectors (the directions of magnetization of specimens) at site i with mean direction m_i and concentration k_i, with $i = 1, ..., s$. Now the mean directions could be unfolded (i.e. rotated about the fold axis, and at each percent a "kappa" estimate based on these rotated means could be found and plotted against the %. If the curve had a maximum at say, 75%, we would want to conclude that the magnetization occurred at that %. But what is the variability of that % estimate? One suggestion, which we will illustrate below, is to assume that the data at the i-th site are distributed as $F(m_i, k_i)$, draw a sample of n_i from this and compute a new mean direction m^*_i say, for $i = i, ..., s$ and so a new curve. Then go on till one has N curves. Then one can get the statistics of the maxima - position, height, etc. The percentiles of the positions gives us a confidence interval for the position of the maxima. If the interval overlaps !00% (0%) one cannot rule out the "pre" ("post") hypothesis. In our worked example, the data used was supplied by K Kodama. We show the first 20 curves Fig. 2. From the statistics of the results (see Figures 3, 4, 5), we found that 95% of all positions of the maximum lay in the interval (65%, 80%). The practical inference in this case is clear. No doubt in other cases there might be so much noise that no clear - cut assertion could be made. These computations were done for me by Michel Debiche.

If the fold axis is measured with some known error it would be easy to add another layer of simulation to reflect it - because one will get different m_i at the same % unfolding. This will broaden a peak such as we gave in Fig 2 - and so broaden the confidence intervals.

Here we got our resamples from Fisher distributions. Instead we could draw bootstrap samples from the points in each original sample, find their mean directions, rotate each by each % unfolding, calculating Kappa each time. This too gives a family of graphs which would be treated as before.

Finally I would prefer to use not kappa but a statistic I have devised (Watson, 1983) to check the null hypothesis that a number of populations have the same mean direction. The use of kappa here goes back to early but incorrect paleomagnetic practice - McElhinney (1964).

These two methods are I suppose the applications of the parametric and nonparametric bootstrap to this problem! I'm sure that they will get refined and extended when they are used in practice.

References

Beran, R., Ducharme, G. R. (1991) *Asymptotic theory for Bootstrap methods in Statistics,* Centre De Reserches Mathematiques, Universite de Montreal (extensive list of references).

Efron , B. (1979) Bootstrap methods: another look at the jackknife, *Ann . Statist. ,* **7,** 1-26.

Efron, B. (1982) *The jackknife, bootstrap and other resampling plans,* SIAM (Regional Conference Series in Appplied Mathematics, 38) Philadelphia , 92pp.

Fisher, N. L., Hall, P. J. (1991) A general Statistical test for the effect of folding????????

Fuller, W. (1987) *Measurement error models,* John Wiley & Sons, New York, 440pp

Gleser L., Watson, G.S. (1973) The estimation of a linear transformation, *Biometrika ,* **60, #3 ,** 525-534.

Hall, P. J. (1991) *The Bootstrap and Edgeworth Expansion ,* Springer, to appear. (extensive list of references)

Kent, J. H., Watson, G. S., Onstott, T. (1990) Fitting straight lines and planes with an application to radiometric dating, *Earth & Planetary Sci . Lett. ,* **97,** 1-17.

McElhinney, M. W. (1964) Statistical significance of the fold test in palaeomagnetism, *Geophys. J. R. Astron. Soc. ,* 338-340

Tauxe, L., Kylstra, N., Constable, C. (1991) Bootstrap statistics for Palaeomagnetic data , *J. G. R. ,* Vol **90, 11,** 723-11740.

Watson, G.S. (1983) *Statistics on the Sphere ,* John Wiley & Sons, New York, pp 238

TABLE 1

BIVARIATE NORMAL

n = 5 10000 simulations 1000 replications 5 iterations

	mu1=0	mu2=0	sigma1=1	sigma2=1	ro=0.7
Mean	-0.021000	-0.019741	0.929537	0.925653	0.655325
SD	0.448826	0.438744	0.352741	0.345771	0.330464
Mean	-0.020863	-0.019644	0.988923	0.984850	0.689007
SD	0.448893	0.438763	0.375278	0.367848	0.340013
Mean	-0.021003	-0.019646	0.989047	0.984776	0.686186
SD	0.449051	0.438793	0.375146	0.367873	0.337448
Mean	-0.020925	-0.019757	0.988979	0.984864	0.687068
SD	0.448873	0.438851	0.375570	0.368098	0.337839
Mean	-0.020738	-0.019612	0.988741	0.984708	0.686768
SD	0.448984	0.439047	0.375188	0.367905	0.337750
Mean	-0.021048	-0.019722	0.988989	0.984861	0.686793
SD	0.449041	0.438764	0.375357	0.368018	0.337718

TABLE 2

USING MULTIPLICATIVE CORRECTION FOR THE ANGLE

Sample size = 5, True Parameter = 20, Kappa = 20

	Angle	Std.Dev.	Kappa1	Std.Dev.	Kappa2	Std.Dev.
Uncorrected Estimates	21.92	8.116	26.09	17.52	26.94	18.29
Corrected Estimates	20.32	8.664	17.41	11.56	17.94	12.28
Corrected Twice	19.11	9.116	20.28	13.69	20.99	14.37
Linear Correction	16.36	10.654	19.51	13.08	20.17	13.79

Confidence Intervals

	2.5%	5%	50%	95%	97.5
Corrected Twice	1.688	3.025	19.34	33.86	36.6
Linear Correction	0.000	0.000	16.77	33.43	36.3

Sample size = 20, True Parameter = 5, Kappa = 5

	Angle	Std.Dev.	Kappa1	Std.Dev.	Kappa2	Std.Dev.
Uncorrected Estimates	11.770	6.017	5.310	1.279	5.291	1.226
Corrected Estimates	9.139	6.348	5.012	1.214	4.993	1.163
Corrected Twice	7.947	6.597	5.024	1.218	5.006	1.169
Linear Correction	5.784	7.256	4.961	1.868	5.004	1.265

Confidence Intervals

	2.5%	5%	50%	95%	97.5
Corrected Twice	0.08285	0.2718	6.376	20.50	22.92
Linear Correction	0.00000	0.0000	1.773	19.74	22.84

Sample size = 20, True Parameter = 20, Kappa = 20

	Angle	Std.Dev.	Kappa1	Std.Dev.	Kappa2	Std.Dev.
Uncorrected Estimates	20.19	4.097	21.23	5.119	21.15	4.904
Corrected Estimates	19.77	4.195	20.06	4.837	19.98	4.635
Corrected Twice	19.71	4.223	20.11	4.851	20.04	4.654
Linear Correction	19.52	4.338	20.16	6.314	20.00	4.821

Confidence Intervals

	2.5%	5%	50%	95%	97.5
Corrected Twice	10.96	12.64	19.80	2..8	27.87
Linear Correction	10.80	12.32	19.67	26.33	27.68

Sample size = 5, True Parameter = 20, Kappa = 5

	Angle	Std.Dev.	Kappa1	Std.Dev.	Kappa2	Std.Dev.
Uncorrected Estimates	29.22	14.82	6.523	4.381	6.740	4.581
Corrected Estimates	24.62	15.95	4.342	2.902	4.476	3.083
Corrected Twice	20.71	16.32	5.018	3.464	5.202	3.639
Linear Correction	18.49	16.16	4.842	3.302	5.011	3.482

Confidence Intervals

	2.5%	5%	50%	95%	97.5
Corrected Twice	0.1962	0.6265866	17.90	50.36	57.43
Linear Correction	0.0000	0.0004011	14.66	48.61	56.17

Sample size = 20, True Parameter = 20, Kappa = 5

	Angle	Std.Dev.	Kappa1	Std.Dev.	Kappa2	Std.Dev.
Uncorrected Estimates	21.54	8.373	5.310	1.279	5.291	1.226
Corrected Estimates	19.65	8.969	5.012	1.214	4.993	1.163
Corrected Twice	18.97	9.389	5.024	1.218	5.006	1.169
Linear Correction	17.81	10.299	4.961	1.868	5.004	1.265

Confidence Intervals

	2.5%	5%	50%	95%	97.5
Corrected Twice	1.112	2.876	19.30	34.45	37.12
Linear Correction	0.000	0.000	18.52	34.25	37.12
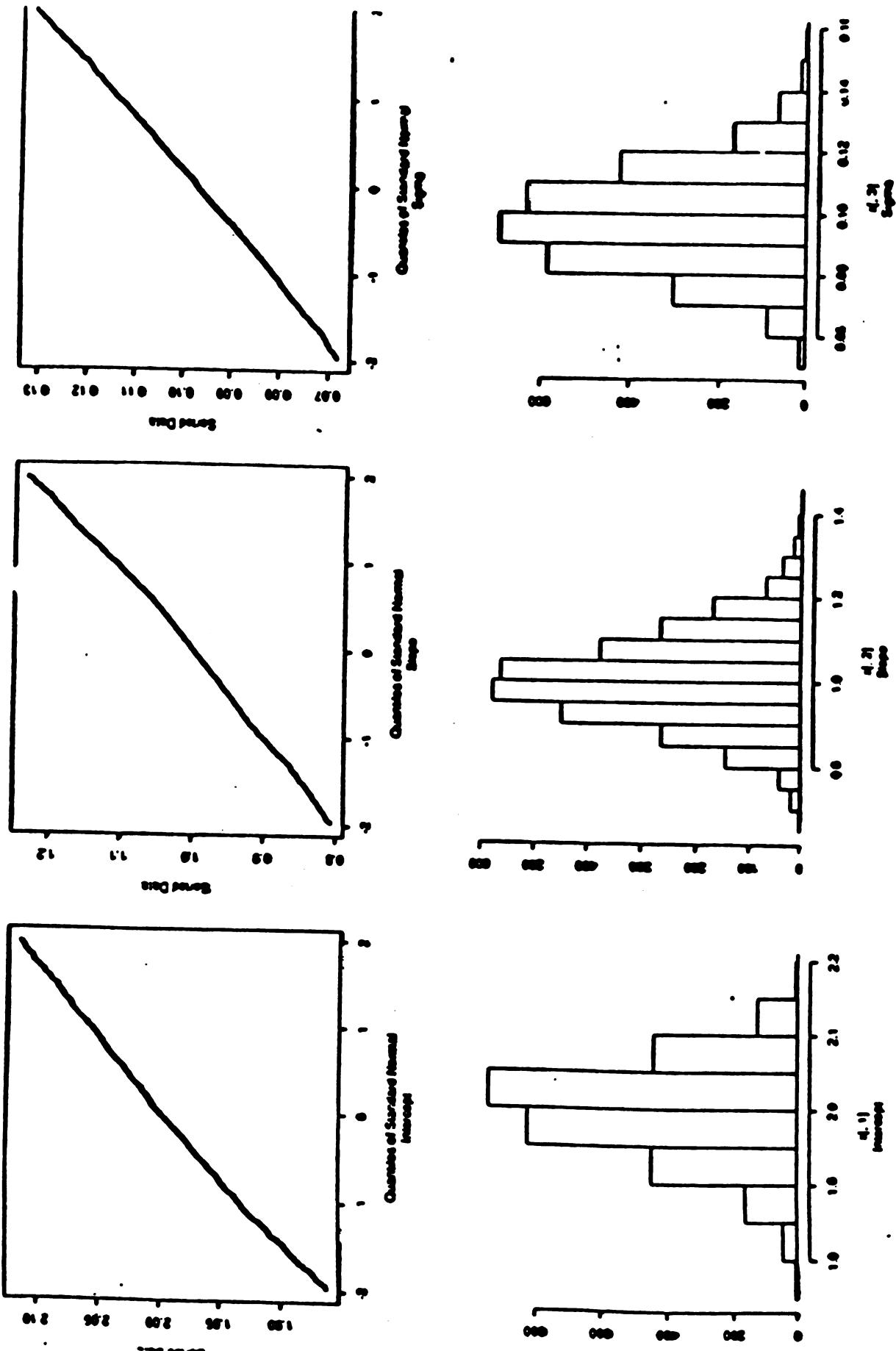
		Sample size = 5			Sample size = 20		
		a	b	Std.Dev.	a	b	Std.Dev.
σ=0.1	MLE	1.98998	1.02032	0.09189	1.99547	1.00855	0.09175
	Std.Dev.	0.11458	0.18878	0.03863	0.06259	0.10904	0.0.688
	Uncorrected	2.01377	0.97090	0.09334	2.04155	0.91647	0.10109
	Std.Dev.	0.10783	0.17564	0.04019	0.05566	0.09510	0.01799
	Corrected Once	1.99854	1.00120	0.10004	2.00542	0.98865	0.10061
	Std.Dev.	0.11137	0.18221	0.04242	0.06027	0.10419	0.01748
	Corrected Twice	1.99426	1.00979	0.09958	1.99960	1.00029	0.10025
	Std.Dev.	0.11281	0.18521	0.04188	0.06170	0.10716	0.01724
	Linear Correction	1.99083	1.01455	0.09973	1.99831	1.00282	0.10044
	Std.Dev.	0.11547	0.19039	0.04202	0.06217	0.10815	0.01732
σ=0.3	MLE	1.80990	1.32309	0.26485	1.93024	1.13252	0.29255
	Std.Dev.	3.18758	5.11863	0.10950	0.63185	1.06141	0.04966
	Uncorrected	2.13501	0.72845	0.29520	2.23287	0.53413	0.32352
	Std.Dev.	0.28726	0.45909	0.13687	0.12044	0.18987	0.06142
	Corrected Once	2.06939	0.85938	0.32011	2.14881	0.70196	0.32183
	Std.Dev.	0.33261	0.54902	0.15849	0.14471	0.24019	0.06334
	Corrected Twice	2.02968	0.93864	0.31434	2.11198	0.77549	0.31594
	Std.Dev.	0.37227	0.62756	0.15627	0.15884	0.26922	0.06196
	Linear Correction	1.94891	1.06672	0.31732	2.07814	0.84229	0.32037
	Std.Dev.	1.61161	5.84560	0.15687	0.17893	0.31108	0.06361
σ=0.5	MLE	1.4838	1.9459	0.40760	2.03963	0.8823	0.47362
	Std.Dev.	18.4917	32.5957	0.16660	6.57566	14.5629	0.07734
	Uncorrected	2.2811	0.4378	0.47946	2.35761	0.2854	0.53149
	Std.Dev.	0.4025	0.6011	0.21989	0.16119	0.2263	0.09648
	Corrected Once	2.2299	0.5417	0.53285	2.30134	0.3978	0.53822
	Std.Dev.	0.4800	0.7577	0.26366	0.19732	0.3121	0.10159
	Corrected Twice	2.1963	0.6102	0.52486	2.27339	0.4535	0.53221
	Std.Dev.	0.5440	0.8838	0.26271	0.21830	0.3582	0.10164
	Linear Correction	1.9099	0.8702	0.52921	2.23999	0.5182	0.53648
	Std.Dev.	5.7236	2.7234	0.26260	0.30512	0.4264	0.10122
σ=1.0	MLE	2.3755	1.4762	0.73618	2.23275	-0.0011	0.89460
	Std.Dev.	19.9005	55.9057	0.30012	11.81564	36.7087	0.14062
	Uncorrected	2.4190	0.1374	0.87690	2.46052	0.0823	1.00262
	Std.Dev.	0.6312	0.6792	0.39050	0.25924	0.2448	0.17349
	Corrected Once	2.4017	0.1678	0.98398	2.44374	0.1166	1.02264
	Std.Dev.	0.7367	0.8874	0.47236	0.29379	0.3529	0.18170
	Corrected Twice	2.3911	0.1867	0.97236	2.43486	0.1341	1.01557
	Std.Dev.	0.8322	1.0663	0.47299	0.31635	0.4126	0.18200
	Linear Correction	2.3126	0.0173	0.97908	2.42273	0.1597	1.02004
	Std.Dev.	6.3744	11.6193	0.47204	0.47741	0.5131	0.18116
σ=2.0	MLE	5.500	-1.586	1.39434	3.12274	-0.9583	1.73817
	Std.Dev.	181.505	20.599	0.57535	27.06611	73.2686	0.27413
	Uncorrected	2.462	0.042	1.66592	2.49416	0.0168	1.94763
	Std.Dev.	1.135	0.690	0.74930	0.47249	0.2464	0.33696
	Corrected Once	2.458	0.051	1.87017	2.48995	0.0236	1.98775
	Std.Dev.	1.280	0.896	0.90436	0.50090	0.3565	0.35256
	Corrected Twice	2.457	0.057	1.84784	2.48762	0.0270	1.97479
	Std.Dev.	1.413	1.067	0.90460	0.52055	0.4173	0.35310
	Linear Correction	1.959	0.633	1.86108	2.48089	0.0344	1.98286
	Std.Dev.	48.055	23.013	0.90312	0.76009	0.5600	0.35160

FIGURE 1

Samp= 20, C.Sim= 5000, Simsize= 3000, Sigma = 0.1

FIGURE 2

Samp= 20, C.Sim= 5000, Simsize= 3000, Sigma = 0.5
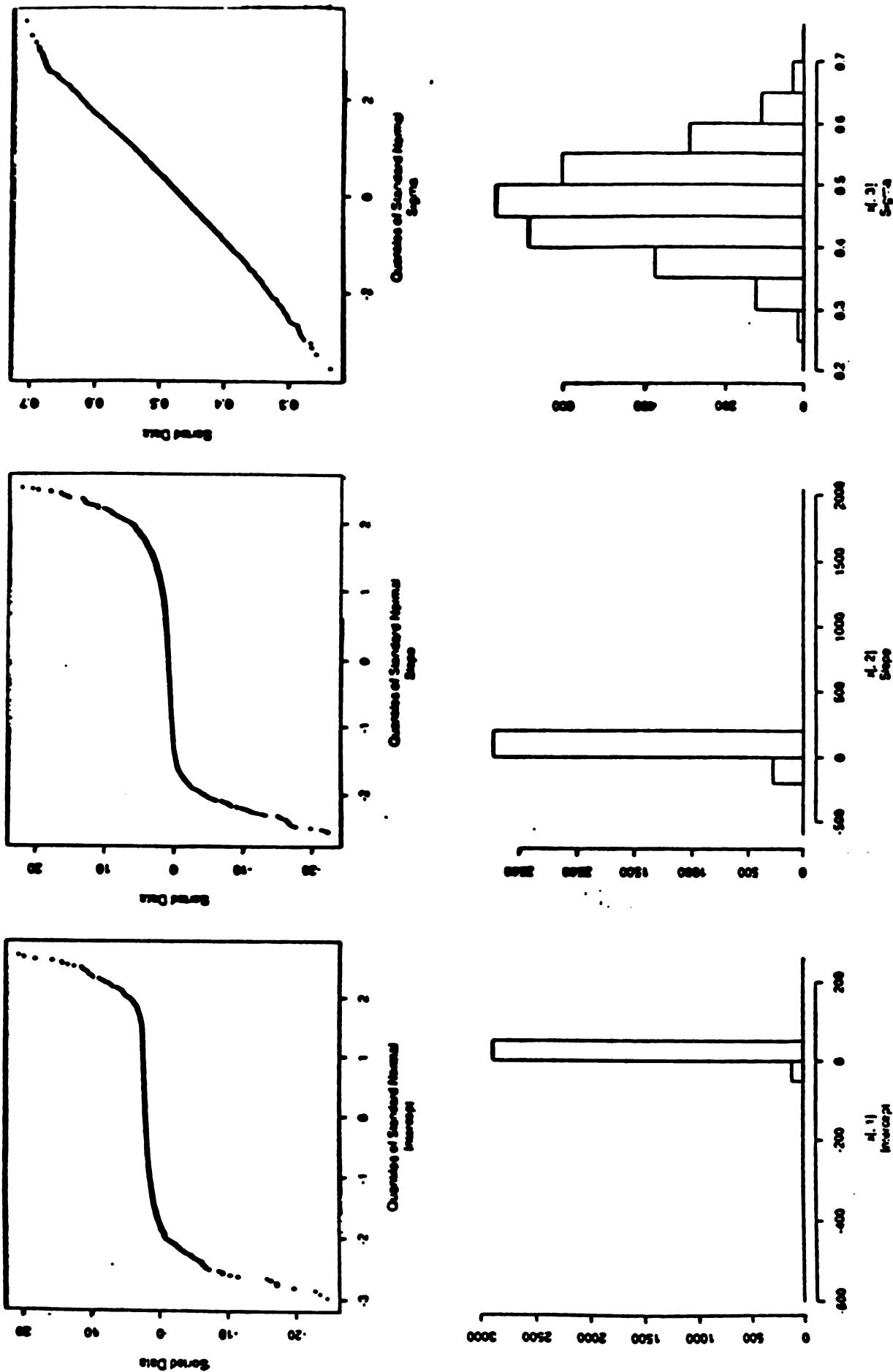
FIGURE 3

**Simulation: k: First 20 Expts**

Estimated kappa

Per cent rotation
no. sites = 8; no. limbs = 2

23
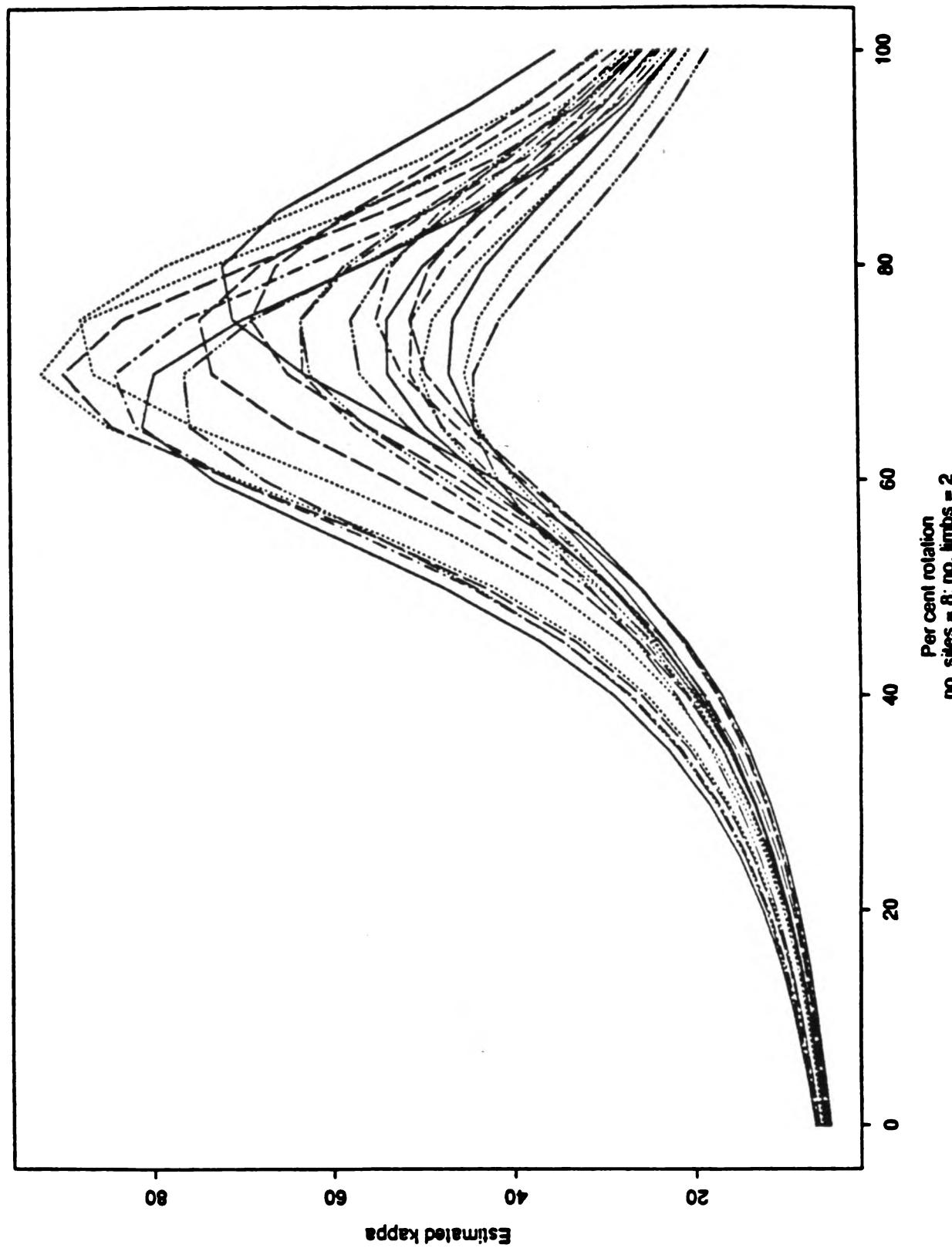
FIGURE 4
Histogram of maximum k values

kmax
First simulation

24
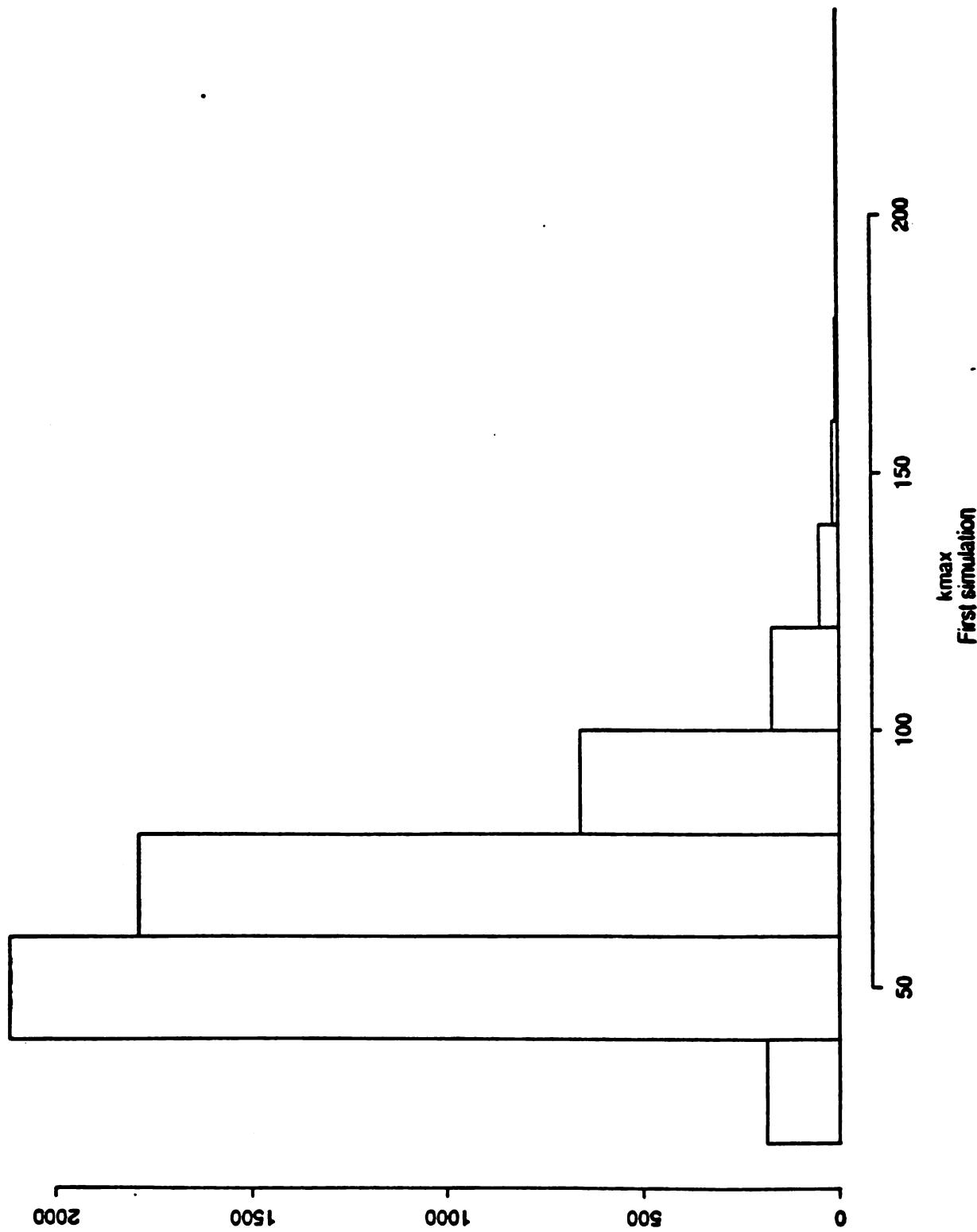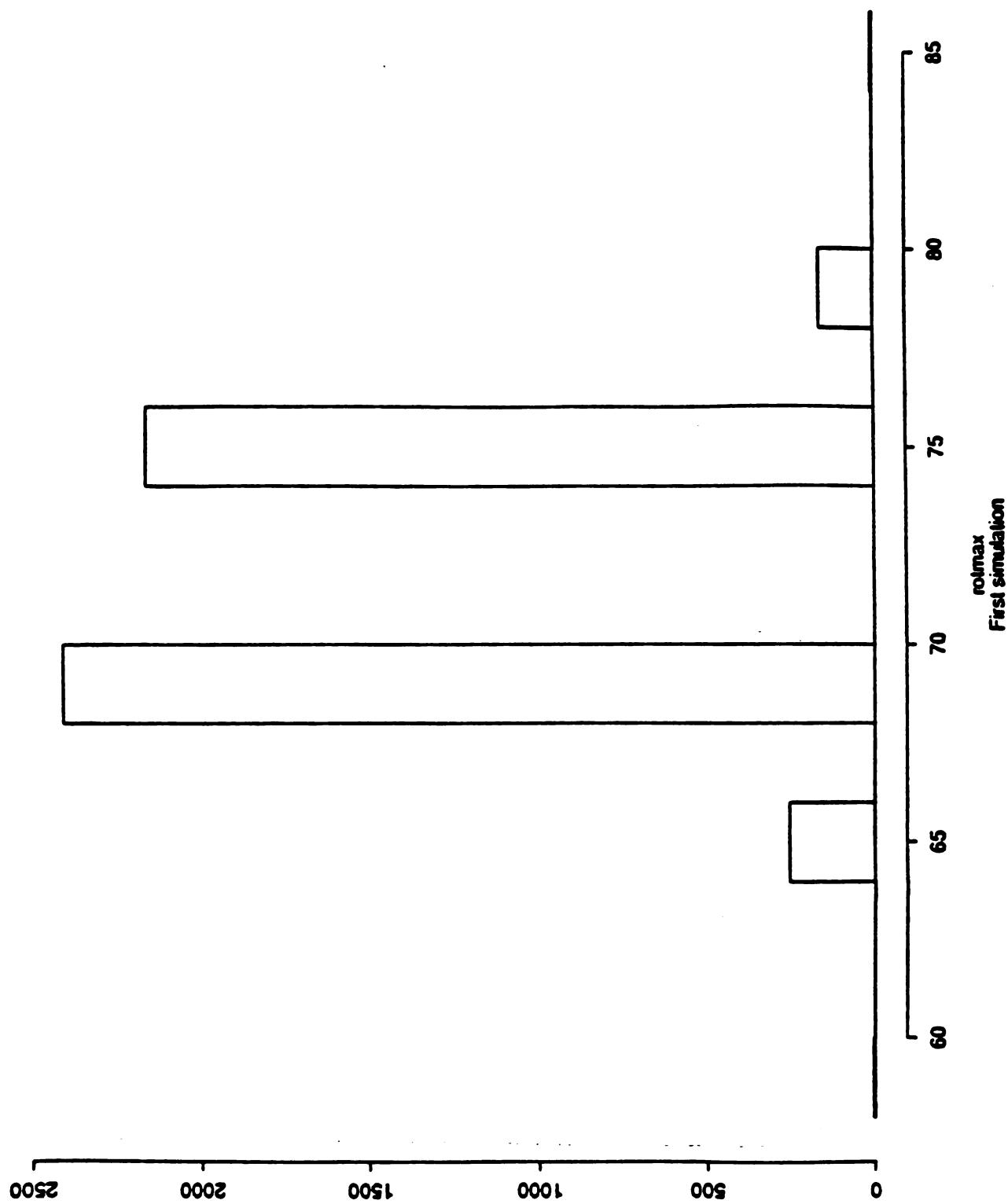
FIGURE 5

Histogram of % rotation for maximum k

rotmax
First simulation

Sensitivity of Component Reliability from Fatigue Life Computations

Donald M. Neal   William T. Matthews   Trevor D. Rudalevige   Mark G. Vangel

U.S. Army Materials Technology Laboratory
SLCMT-MRS-MM Arsenal St.
Watertown, Massachusetts 02172-0001

Abstract

Structures which fail due to cyclic loading are said to fail in fatigue. Damage accumulation due to fatigue is the primary factor which limits the useful life of aircraft.

Designers of new aircraft obtain information with which to estimate fatigue life by extensive testing of small specimens, in addition to *very limited* testing of actual structural components. Based on this data, an estimate is made of the lifetime for which very high (.999999) reliability is assured. This high reliability is currently a requirement in the construction of army helicopters and fixed wing aircraft.

There is little agreement among designers on how fatigue life should be determined, as well as insufficient understanding of the uncertainties involved in high reliability computations.

This presentation reviews the fatigue life determination procedures for several manufacturers and points out some ways in which these methods are deficient in obtaining high reliability.

The purpose in introducing this clinical paper is to obtain statistical procedures that will provide highly reliable fatigue loaded structures such as the Army helicopter.

INTRODUCTION

Methodology to substantiate helicopter fatigue life has received considerable attention during the last decade. This interest was stimulated by the substantial variability in the results from the study on the American Helicopter Society pitch link problem.[1] Recently, further interest has resulted from the U.S. Army's introduction of a structural fatigue reliability criterion for rotorcraft. This criterion has been interpreted[2] as a requirement for a component lifetime estimate to have a reliability of 0.999999.

Helicopter safe life reliability methodology has recently been the subject of several papers[3-6] and an American Helicopter Society subcommittee round robin.[7]

The authors[8] have investigated the sensitivity of high reliability estimates from simple stress-strength statistical model computations. Results showed substantial variability in reliability estimates even for almost undetectable differences in the assumed probability density functions (PDFs) representing the stress and strength data.

In this report, the uncertainties in determining high reliability for helicopter component safe life design are studied by introducing a simulation process to identify the effects of a small amount of variability in the design variables for determining the lifetime estimate. The reliability values are determined for a generic uniaxial steel structure loaded in tension similar to a helicopter pitch link component by applying Miner's Linear Damage Rule.[9] The six component fatigue test values were obtained from Arden[1] where the maximum applied stress (S) on the component is tabulated with respect to cycles to failure (N). In order to obtain an SN curve to represent the component fatigue test results, a separate regression analysis was applied to a larger set of coupon tests of a steel for which the results are tabulated in Bury.[10] The assumed spectrum load used in determining the lifetime estimate was obtained from Berens.[11] Note that only the six component fatigue test values are from Reference 1 and the remaining test values are from References 10 and 11.

THE COUPON TEST SN CURVE

This section describes the procedure for determining an SN regression curve to represent coupon fatigue test data,[10] as shown in Figure 1. The assumed functional representation[12] of the data is

$$S = S_\infty + (S_u - S_\infty)e^{-\beta (\log_{10} N)^\gamma} \qquad (1)$$

where S is the maximum applied stress and N is the number of cycles required for the coupon to fail. S_∞ is the coupon endurance limit representing the case when $N \rightarrow \infty$ and S_u represents the static strength of the coupon; i.e., the strength for $N = 1$. The shape of the SN curve is determined by β and γ. S_∞, S_u, β, and γ were determined from application of an IMSL computer code[13] for solving nonlinear regression problems. The resultant SN curve is shown in Figure 1 (solid line) with the individual coupon fatigue test values.
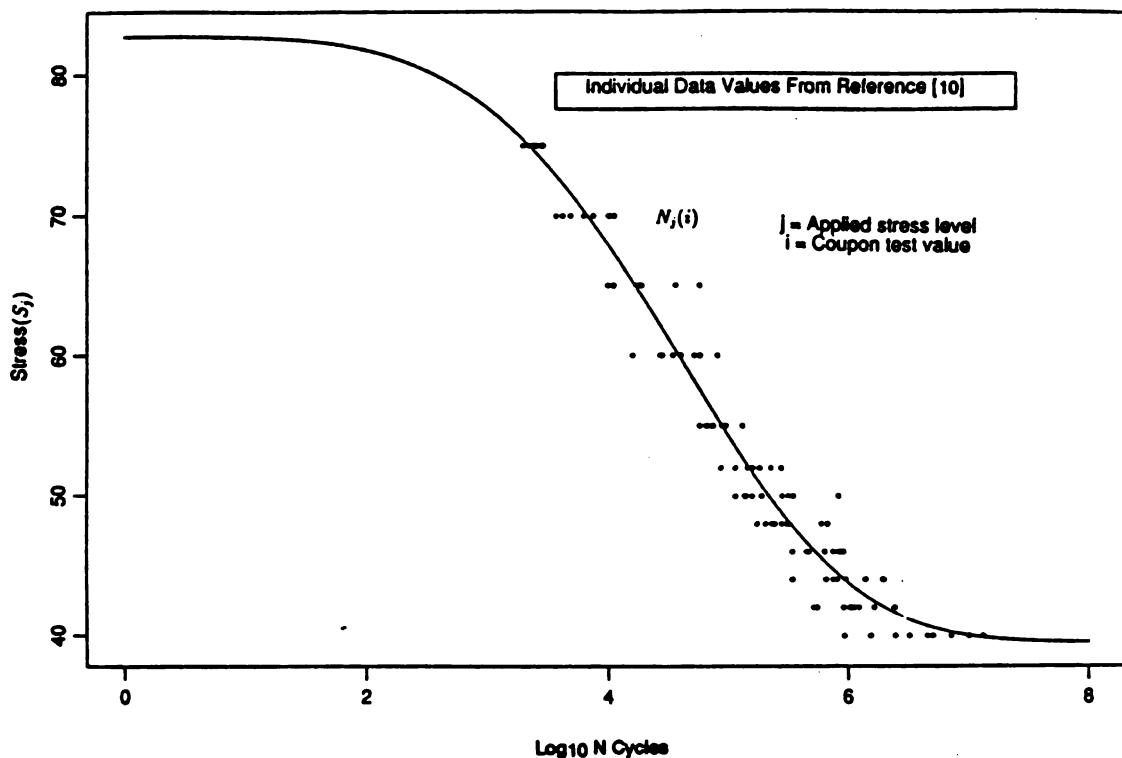
Figure 1. Mean coupon SN curve/fatigue data.

A review of the literature on the determination of component fatigue life showed that various functional representations similar to Equation 1 have been applied where N is the independent variable and S is the dependent (response) variable. This is counter to the conventional functional representation of test data where S would be the independent variable in the analysis since a *fixed* cyclic load (stress) value is applied and a *resultant* (dependent) number of cycles to failure is recorded. In order to obtain N as the dependent variable, Equation 1 can be inverted resulting in the following:

$$\log_{10} N = e^{\{\log[-\log((S - S_\infty)/(S_u - S_\infty))] - \log \beta\}/\gamma}.$$

(2)

Although Equation 2 is recommended in determining the functional representation of the data, Equation 1 was applied in this study since it is commonly used in engineering fatigue analysis, and the qualitative measure of the relative uncertainties in determining the reliability at a specified lifetime are not affected by the SN curve assumption.

In order to simplify the analysis, the fatigue data from Reference 10 was normalized with respect to the estimated S_∞ value determined from the initial application of regression analysis. Another SN curve was then obtained from the normalized data, where β, γ, and S_u were obtained for a known S_∞ of 1. The resultant SN(N) curve is shown in Figure 2. The figure also shows the regression results SN(S) from the application of Equation 2.

29
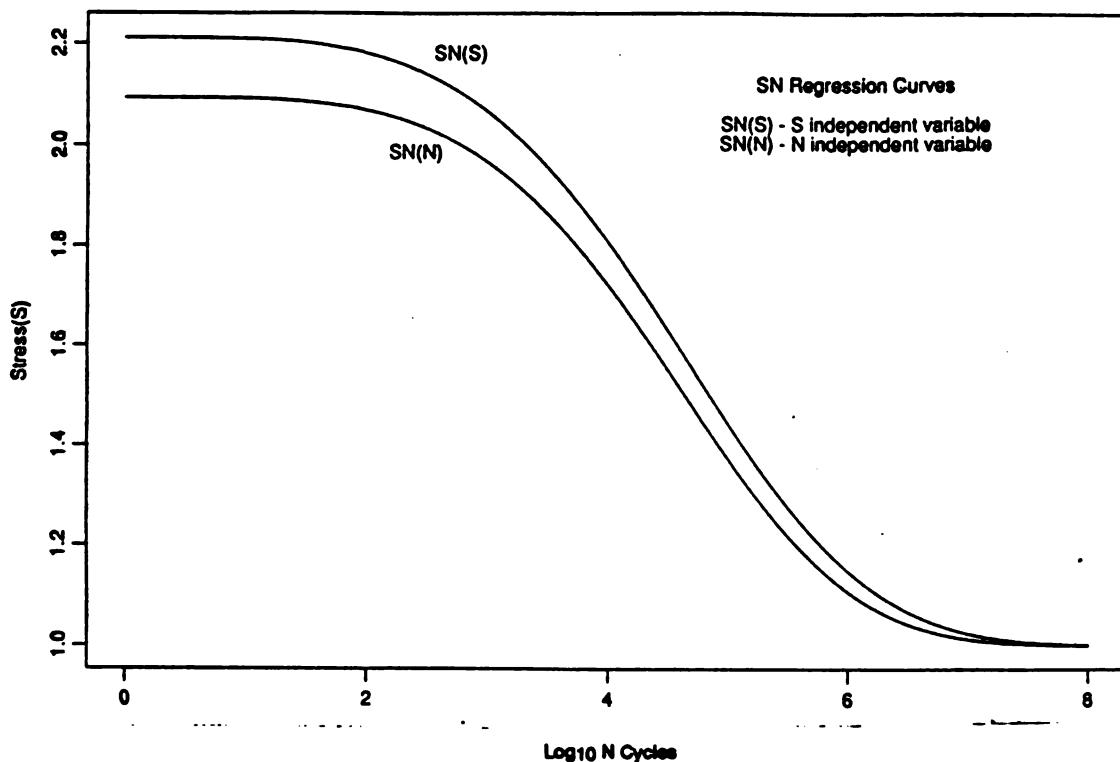
Figure 2. Regression curves from coupon test results.

THE COMPONENT SN CURVE

Usually the shape of the component SN curve is obtained from a prior coupon SN(N) curve, as shown in Figure 2. The location (ordinate position) of the curve is determined from extrapolating the individual component values, as shown in Figure 3, to $N = 10^8$ cycles. The original component values in Reference 1 have been rescaled so that they have scales similar to the S values in Figure 2. The extrapolation process involves vertically positioning the coupon SN curve (see Figure 3) to agree with the individual component values and then extending the curves to $N = 10^8$ cycles. S_i values are obtained for $N = 10^8$ and the component curves mean stress position at N is

$$S_m = \sum S_i / n,$$

(3)

where n is the number of component test results. The solid line in Figure 4 shows the representative component SN_c curve and component test data. Since there are usually only six component test results available, because of the costs in component testing, the above procedure is often applied. Using the more extensive, less expensive coupon test results to determine the shape of the SN curve assumes similar material, test, and environment for both coupon and component.

30
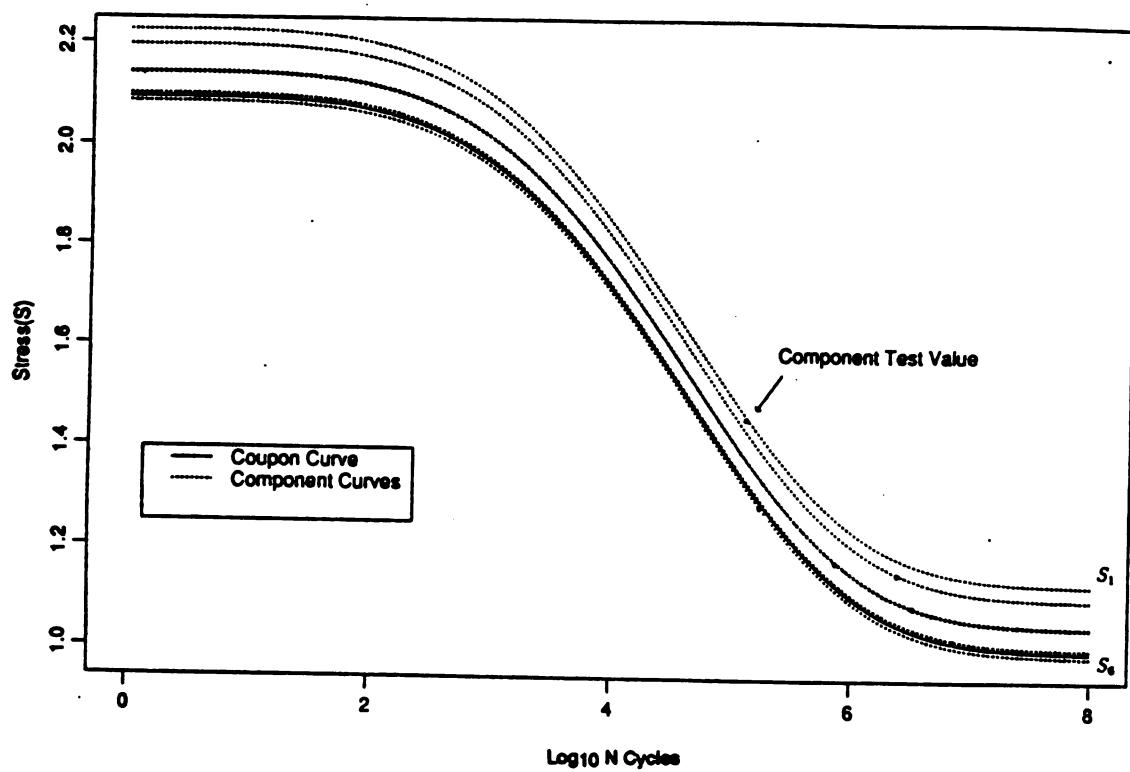
Figure 3. Extrapolation of component data.
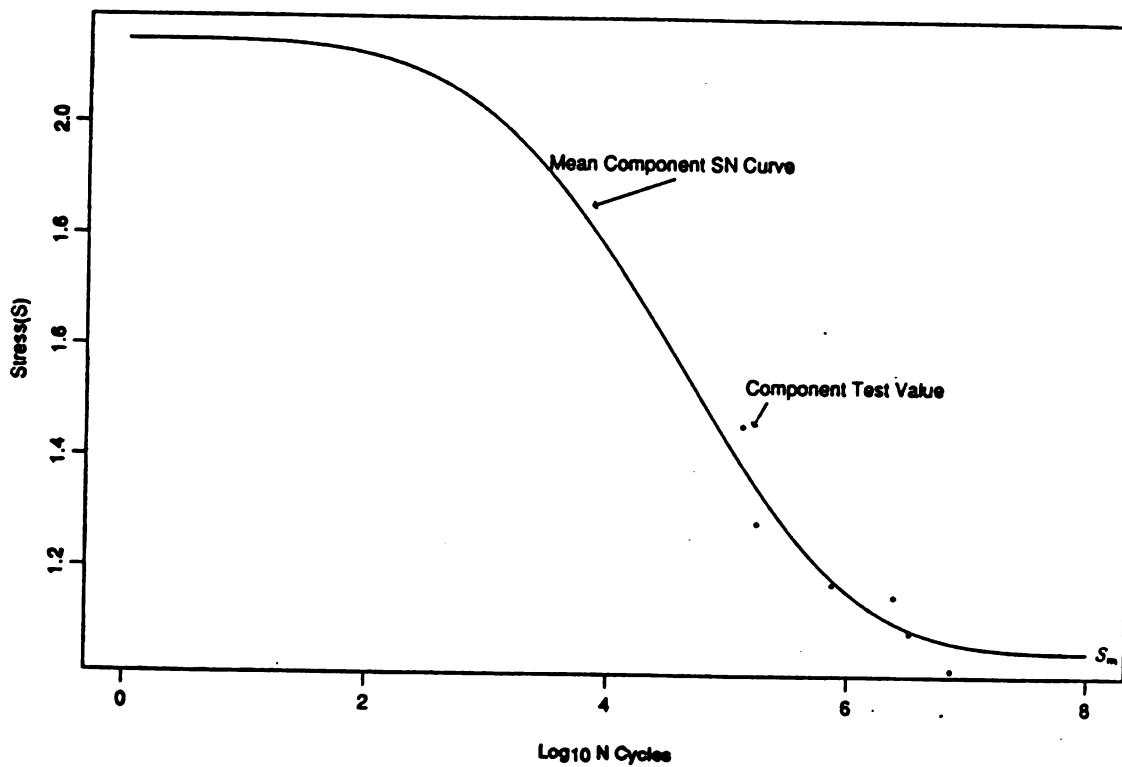
Figure 4. Mean SN curve for component test values.

31

SPECTRUM LOAD

The normalized spectrum loading used in the fatigue life analysis is shown in Figure 5a. The loading was obtained from a rainflow count of a modified combat history described in Reference 11. The spectrum was determined by the number of loads within discrete range increments. The spectrum is simplified to five loads $\{L_i\}_1^5$ by expanding the size of the range increments and including the appropriate cycle count $\{n_i\}_1^5$ within each expanded range. The normalization procedure involved dividing each L_i by the smallest damaging load $S_\bullet$ (endurance limit). This simplification was adequate for identifying the spectrum effects in this study.

MINER'S RULE

In order to obtain the lifetime estimate from the simplified fatigue load (L) and the normalized material strength (S) data shown in Figures 5a and 5b, the following linear damage rule[9] is applied where

$$DF = \sum_{i=1}^{5} \frac{n(i)}{N(i)} \tag{4}$$

is the damage fraction for each pass or repetition of the spectrum. This representation of operation hours is described in Reference 10. The n(i)s are the number of cycles corresponding to the applied load L(i), as shown in Figure 5a. The N(i) values are obtained from the SN curve, as shown in Figure 5b, where the corresponding S_i values are identified in the figure by the L(i) values obtained from the spectrum loads in Figure 5a. In addition, the rule requires that

$$N_P \cdot DF = 1 \tag{5}$$

in order to determine the maximum number of passes (N_P) that can occur prior to the component failure.
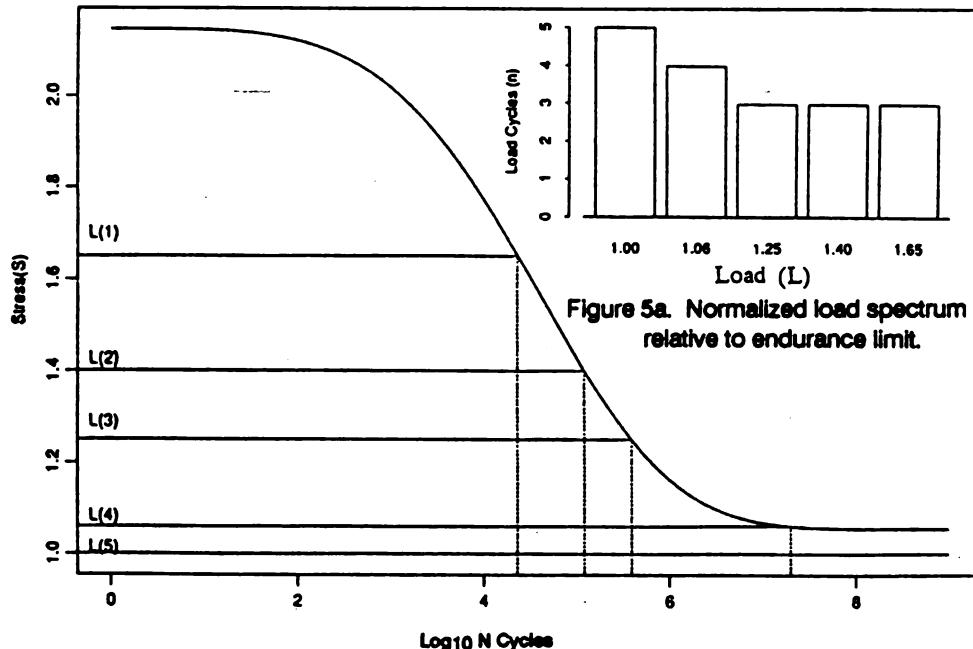
Figure 5a. Normalized load spectrum relative to endurance limit.
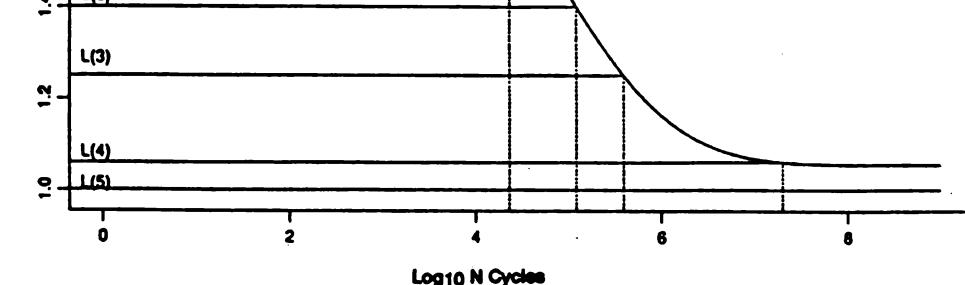
Figure 5b. Component mean SN curve.

Bootstrap Method Applied to Coupon SN Curve Computation

The Bootstrap Method,[14] a simulation process, was introduced in the fatigue life reliability analysis in order to examine the effects of uncertainties used in determining the coupon SN curve and the resultant component reliability. Only one reliability estimate can be obtained from a single set of data; however, even with all conditions the same, one would expect to determine a different reliability estimate from another set of data. The Bootstrap Method provides a technique for estimating the variability among random sets of data generated under equivalent conditions using data from only a single random sample. The idea is to create arbitrarily many *new* datasets by sampling *with replacement* from the original data. If there are n values in the original data, then a new dataset is created by selecting n values from among these observed data, allowing data values to be selected more than once. The probability distribution of the reliability calculated from these datasets, which are created by taking random samples from the single observed dataset, provides an estimate of the actual probability distribution of reliability which could, in principle, be determined from *future* datasets.

The material fatigue testing involves obtaining the number of cycles to failure for a specified applied load (S) shown as the individual data points in Figure 1. The Bootstrap Method involves selecting a random set of 9 values independently with replacement from the set of cycles to failure values $\{N_j(i)\}_{i=1}^9$ for each jth applied stress from $\{S_j\}_1^{12}$, as shown in Figure 1 and obtained from Reference 12. The result is a new set $\{N_j^*(i)\}_1^9$ for each of the S_j values.. The new set is called the *Bootstrap* sample where some values can be repeated once, twice, or more times. The new set is then used in the regression procedures described in the Coupon Test SN Curve Section in order to obtain a new SN curve (S in Equation 1).

In Figure A1 (see Appendix), the results of applying the Bootstrap show a 90% confidence band on the original SN(N) curve. Results in Figure A2 show the individual SN(N) curves obtained for the Bootstrap samples. The results from Figures A1 and A2 indicate that there is more variability for large or small N values than for the central region of the curve which is consistent with determining confidence bands on regression curves.

For calculating the effects of coupon SN curve uncertainties, a damage fraction (DF*) value is computed by applying the linear damage rule. The above procedure is repeated M_B times, so that a set of $\{DF_k^*(i)\}_1^{M_B}$ are obtained. The component reliability R can then be obtained by counting the number (N_B) times $N_P \cdot DF^* < 1$, $k = 1,2,...,M_B$, where N_P, the number of passes, is specified. The computed component reliability R including uncertainties in the coupon testing procedure is written as

$$R = N_B / M_B, \tag{6}$$

where M_B is the number of repeated applications of the Bootstrap procedure.

Reliability Estimates from SN Component Curve Simulations

The following simulation procedure was applied in order to identify the effects of uncertainties in the location of the component SN_c curves on the reliability estimates.

*Represents simulation results.

The uncertainties are assumed because of the potential differences in loading, material, surface conditions, and geometry between the coupon and component. Also contributing to these uncertainties are: the extrapolation of the component fatigue data in determining the S_i's, as shown in Figure 3, and the availability of only six values in computing S_m (mean of the curve), as shown in Figure 4. Examination of potential inaccuracies in the reliability computations due to assuming that the component and coupon SN curve shapes are similar was not included in the simulation process. Introducing variability in the curve's location was sufficient for showing sensitivity in the reliability computation. In the simulation process, a random set of M S_m^* values were obtained. These values are normally distributed about the S_m value in Figure 4 from the following:

$$S_m^*(i) = S_m(1 + V_S \cdot Z_i), \quad i = 1,2,...,M, \tag{7}$$

where the Z_i's are values randomly selected from a standard normal distribution with a mean of 0 and a variance of 1. The V_S value is the coefficient of variation (CV) and the mean is S_m. In Figure A3, a representative normally distributed set of S_m^* values are shown for $V_S = 0.01$ and $V_S = 0.02$. The newly obtained mean values (S_m^*) are now used in vertical positioning of the component SN curve, as shown in Figure 4, so that M SN curves can be obtained from Equation 1 by the following:

$$S_i^* = S(S_\infty, S_u, \beta, \gamma) + \Delta P_i, \quad i = 1, 2, ..., M, \tag{8}$$

where $\Delta P_i = S_m^*(i) - S_m$. M damage fraction values (DF_i^*) are obtained from applying the procedures described in the Miner's Rule Section and the schematics shown in Figures 5a and 5b using the newly available S_i^* values.

From Miner's Rule, compute $N_P \cdot DF_i^*$, $i = 1, 2, ..., M$ and record the number (N_S) of times $N_P \cdot DF_i^* < 1$ for a given N_P value, where N_P represents the specified number of passes. The component reliability R can be written as

$$R = N_S / M. \tag{9}$$

Note that in order to obtain 0.999999 reliability, $M = 1 \times 10^6$ simulations would be required.

Load Uncertainties Effect On Reliability Computations

A simulation procedure similar to that described in the previous section was applied in order to identify the sensitivity in computing component reliability by introducing uncertainties in the assumed spectrum loads (see Figure 5a). There exist potential errors involved in assuming a specific load spectrum.[15] They are the results of: an inaccurate measuring device, the location of the device, and assuming load patterns determined from short periods of data recording which differ from the actual loads the component would be subject to during its operational lifetime.

Application of the simulation process involved only modeling uncertainties in the L values, with n(i)s remaining constant for a given load. Introducing the same amount of variability in each $\{L(i)\}_1^5$ values was sufficient to show the sensitivity of the reliability estimates from uncertainties in the loading.

34

Initially, the simulation involves obtaining M1 sets where the jth set $\{L_j^*(i)\}_{i=1}^5$ is determined from the following:

$$L_j^*(i) = L(i)(1 + V_L \cdot Z_j), i = 1,2,...,5 \tag{10}$$

where $j = 1,2,...,M1$ and Z_j is a random value from a standard normal $N(0,1)$ distribution. V_L is the coefficient of variation representing an assumed variability in load $L(i)$.

For the jth simulation, the original five loads $\{L(i)\}_1^5$, as shown in Figure 5a, are modified resulting in a new set $\{L_j^*(i)\}_i^5$ from Equation 10. The distribution of $L_j^*(1)$ for all j values, for example, would be similar to that for S_m, as shown in Figure A3.

In the simulation process, the jth modified set L_j^*, and its associated N_j^*, determines a damage fraction value DF_j^*, as described in the Miner's Rule Section and Figures 5a and 5b. In order to obtain component reliability values from the load variability, Miner's Rule is then applied by recording the number (N_L) of times $N_P \cdot DF_j^* < 1$ for $j = 1,2,...M1$. The component reliability R is then written as

$$R = N_L / M1 . \tag{11}$$

Reliability Sensitivity from Uncertainties in Miner's Rule

A simulation procedure similar to those in the previous two sections is applied to the Miner's Rule relationship in Equation 5. This was done in order to examine the effects of a possible error in assuming the component will fail when $N_P \cdot DF = 1$ (see Equation 5). In order to identify the effects of this uncertainty in computing component reliability R, the following simulation process was performed:

Initially, the value 1 in Equation 5 is replaced by a set of random numbers $\{CR_i\}_1^{M2}$ resulting in $N_P \cdot DF^* < CR_i$, where

$$CR_i = 1 + V_M \cdot Z_i, i = 1,2,...,M2 \tag{12}$$

and V_M and Z_i are the assumed coefficient of variation and standard normal as previously defined in the above two sections.

The reliability R is determined from recording the number (N_Z) of times that

$$N_P \cdot DF^* < CR_i, \tag{13}$$

and then defining

$$R = N_Z / M2 \tag{14}$$

where M2 is the number of simulations.

WORKING SN CURVE

The adjustment of the mean component SN_c curve from a limited amount of component test data results in a certain amount of variability in estimating the location of the curve. In order to account for this variability, and in some instances other uncertainties in the fatigue analysis process, a component SN_c curve reduction factor is often introduced which results in a new working SN_w curve, as shown in Figure A4. There is no standard method for obtaining a working curve in the helicopter industry.[16] The working curve in Figure A4 was obtained by a uniform reduction in all S_c values. This approach maintains the same curve shape as in the original SN_c curve; i.e., the coupon SN curve shape. This approach is consistent with the use of the coupon curve shape in the extrapolation process for each component data value (see Figure 3) by which the original component curve S_m value (see Figure 4) is obtained. In Figure 3, a schematic of this uniformity is shown where for $N = 1$ and $N = 10^8$ show an equal amount of assumed dispersion in the S_i values.

Reduction Factors for Working Curves

Some of the reduction factors commonly used by the helicopter manufacturers are discussed in References 15 and 16. In some cases a multiplication factor is used to obtain working curve values, S_w; i.e.,

$$S_w = S_c - P \cdot S_m \tag{15}$$

where S_c represents the strength values from the component curve, SN_c for various P values. S_m was previously defined in Equation 3.

Another reduction procedure involves defining

$$S_w = S_c - 3 \cdot SD \tag{16}$$

where the standard deviation (SD) is often determined from an assumed standard coefficient of variation for a particular material to represent the S_i values shown in Figure 3 and in Equation 3. A typical value for the coefficient of variation for steel is 7%. The SD value is then written as $SD = 0.07 \cdot S_m$. One other method involves determining SD from the actual S_i values; i.e., $SD = \sqrt{(\sum (S_i - S_m)^2 / (n - 1)}$, and substituting the SD value in Equation 16.

The working curve was introduced in this report in order to evaluate its capability to include the possible variability in the reliability estimates from the simulation results.

RESULTS AND DISCUSSIONS

In this section, results from the simulation procedures are shown in both tabulated and graphical form. Variability is introduced in combination, as well as individually, for all of the following four factors: the spectrum load, the mean SN Curve, Miner's Rule, and the Bootstrap process.

In Table 1 all four factors were varied for a range of CV values (% variability) from 1% to 5%, except for the Bootstrap simulation where the variability is obtained from coupon test results. The component reliability results are tabulated as a function of the corresponding CV values assumed in the simulation procedures. The results were obtained by systematically randomly selecting values from each of the four factors so that 1×10^6 distinct factor combinations are obtained for computing the damage fraction (DF) in the Miner's Rule Section. The reliability (R) is then obtained from the sum of all the times $N_p \cdot DF^* < 1$ divided by 1×10^6.

Table 1. RELIABILITY VERSUS FACTOR VARIABILITY: LIFETIME = 3425

% Variability*	Reliability
1.0	0.999999
2.0	0.989676
3.0	0.937250
4.0	0.872101
5.0	0.816061

*Simultaneous variability assumed for the following: spectrum load, mean curve, Miner's Rule = factor (1) and the Bootstrap process on defining mean curve are applied.

In order to apply the simulation procedures, a 1% variability was introduced for each of the factors and the number of passes ($N_p = 3425$) was selected in order to obtain a baseline reliability value of 0.999999. This value was selected because of the helicopter industry's interest in obtaining high component reliability of 0.999999.

The results in Table 1 show a substantial instability when comparing the reliability estimate of 0.999999 versus 0.989676 for the respective 1% and 2% variabilities. The implication of these results is that in one case one in a million failures could occur compared to 10324 failures in a million in the other. This substantial difference for such a small increase in the inherent variability in the assumed fatigue life models shows a severe sensitivity in computing high reliability when there is a small degree of uncertainty in determining spectrum loads, SN curves, and assuming a failure requirement from Miner's Rule. The results from increasing the variability from 3% to 5% show a corresponding reduction in reliability values. The R = 0.816061 for 5% variability is a very large reduction from the original 0.999999 for 1% variability. The CV values shown in Table 1 represent a range of potential parameter uncertainties in the fatigue life model.

In Table 2, reliability values are tabulated as a function of the combined and individual variability of the four factors. This was done in order to examine the effects of the individual factor variability on computing component reliability. The 1% variability was applied to all factors resulting in R = 0.999999 when N_p is equal to 3425 (as in Table 1 at 1%). The 2% variability was applied to each factor individually with 1% variability for the other two factors. The Bootstrap process was applied in all of the cases. Introducing a 2% CV in the spectrum load (SPL) shows a substantial reduction in the reliability estimate from 0.999999 to 0.996404. The 2% variability in the component SN_c curve (MSN) shows a smaller reduction of 0.999999 to 0.999440 indicating that, based upon the particular spectrum considered, the *spectrum load uncertainties* could result in greater instability in the reliability values. Small variations in the Miner's Rule assumption (see Equation 13) do not appear to be as critical in the reliability computations. Increasing the variability from 3% to 5% shows a continued decrease in reliability estimates except for the case of Miner's Rule variability which has a very small reduction. The 5% variability on the spectrum load shows a value R = 0.862469 which is only 5.7% greater than the case where all factors were varied simultaneously, as shown in Table 1 for 5% variability.

Table 2. RELIABILITY VERSUS INDIVIDUAL FACTOR VARIABILITY: LIFETIME = 3425

% Variability (P) on Individual Factors*	Reliability		
	SPL	MSN	MR
1.0	0.999999	0.999999	0.999999
2.0	0.996404	0.999440	0.999999
3.0	0.967356	0.992375	——
4.0	0.912587	0.972164	0.999997
5.0	0.862469	0.941979	0.999994

*1% variability is applied to all factors except for individual increase in factor variability (P) in first column. Bootstrap process also included.

In Table 3, reliabilities are obtained for the individual factors, spectrum load (SPL), and location component SN curve (MSN). In order to obtain the R = 0.999999 value for 1% variability on each of the factors, the number of passes (Np) was 3700 for SPL and 4425 for MSN. The lower Np value for SPL is consistent with the results in Table 2 since the R values for SPL were lower than those for MSN when Np was 3425. In addition, it is obvious that a lower number of cycles of operation would usually increase the reliability value. The Bootstrap Method application resulted in a value of R = 0.999977 when combined with a 1% variability in MSN. This indicates that the method is not introducing any substantial variability compared to the SPL and MSN contribution in determining R. This is expected because of the small amount of variability in the SN curves, as shown in Figures A1 and A2. In addition, the range of cycle values contributing the most in determining the damage fraction has the least amount of variability.

Table 3. RELIABILITY VERSUS INDIVIDUAL FACTOR VARIABILITY / LIFETIME

% Variability	Reliability (R)	
	SPL*	MSN†
1.0	0.999999	0.999999
2.5	0.969376	0.965875
5.0	0.828010	0.818789

*3700 Lifetime value
†4425 Lifetime value
NOTE: Application of Bootstrap process simulation resulted in R = 0.999977 with 1% variability for MSN.

Table 4 shows the reliability results from reducing the S_m value, shown in Figure 4 and Equation 3, by the tabulated percentage in order to examine the possible material mean strength loss from environmental effects such as corrosion. New values equal $(1 - p/100)S_m$ where p is the tabulated percent reduction factor. In the case where p = 0, R = 0.999999 was obtained varying the SN_c curve by 1% with Np = 4425 which is in agreement with the result in Table 3. This variability in the SN_c curve (MSN) was maintained for each of the reduced S_m values. When p = 1, then $0.99S_m$ was used in the simulation process to obtain a reliability value equal to 0.999852 compared to 0.999999 for no reduction in S_m. This result is not as substantial a reduction in R as the case where the S_n value is reduced by 5% and R = 0.324206. The overall results indicate that loads which previously did not increase the damage fraction are now significant contributors in reducing the component reliability. If there is a potential for material strength loss due to corrosion, for example, then high reliability estimates are substantially reduced by small mean strength reduction.

**Table 4. RELIABILITY VERSUS PERCENT REDUCTION MSN:
LIFETIME = 4425**

% Reduction	Reliability
0.0	0.999999
1.0	0.999852
2.0	0.995542
3.0	0.946600
4.0	0.720650
5.0	0.324206

NOTE: 1% variability on MSN

Table 5A shows the deterministic fatigue lifetime values obtained from the application of various working curves described in the Working SN Curve Section. This computation was introduced to evaluate the curves relative effectiveness in accounting for the uncertainties in estimating the component SN_c curve. This evaluation involves comparing results from Tables 5A and 5B. In Table 5A, P = 0.5 which is the reduction from S_c in Equation 15. This shows a lifetime of 0.325, which is a very conservative estimate compared to the 6150 passes obtained from using the original component curve without a reduction. The least conservative lifetime estimate is 2000 which was obtained from reducing the component curves by three SDs. SD was obtained by using the S_i values in Figure 3 and Equation 3. This estimate was less conservative than the 1225 lifetime value obtained using an assumed CV = 0.07. The extrapolation process shown in Figure 3 may account for the relatively low SD estimate for the case when the life value is 2000. The other reduction factors result in a predictable decrease in the life estimate with an increase in the reduction percent P.

**Table 5A. LIFETIME VALUES FROM APPLICATION OF
WORKING CURVES**

Working Curve (Adjustment on S)	Lifetime Value
0.50*	0.325
0.44	48
0.30	500
0.25	850
0.20	1355
S - 3(sd)†	1225
S - 3(sd)	2000
NA ◊	6150

*Percent reduction of (P) on S: where (1-P)S is location of working curve
and S is mean component strength at endurance limit.
†Standard deviation determined from assuming 7% coefficient of
variation for S.
◊ NA: No adjustment of SN curve.

Table 5B. LIFETIME VALUES WITH 0.999999 RELIABILITY VERSUS VARIABILITY ON MSN AND SPL

% Variability	Lifetime Value
1.0	3425
2.0	1850
3.0	875
4.0	350
5.0	50

In Table 5B, simultaneous variability on the component curve (MSN) and the spectrum load (SPL) for 0.999999 reliability shows a reduction in the lifetime value with increasing variability, which is consistent with prior results. By comparing results from Tables 5A and 5B the effectiveness of the working curve in obtaining 0.999999 reliability can be identified. That is, for example, a 1% variability shows 3425 indicating that any of the working curves could provide the required reliability although the curve obtained from the three SD reductions would be the least conservative acceptable method. Introducing 2% variability shows a life estimate of 1850 which, in this case, requires using the three SD reduction procedure where SD is obtained from assuming a 0.07 CV value. If the variability is assumed to be 5%, then a lifetime value of 50 is obtained which would require a working curve reduction factor of 0.44 in order to provide the 0.999999 reliability. If a 5% variability in the loading and SN curve can exist, then most of the working curve procedure would be an undesirable method for obtaining high reliability.

Using Equation 7, the results of introducing a 1% uncertainty in the positioning of the component curve is shown in Figure 6 as a probability density function for the lifetime estimate (N_P = 1/DF) determined from Equation 4. A 7.3% coefficient of variation was obtained with a mean life of 6194. The inner range, $N_P \pm 3 \cdot$ SD, is 4964 to 7689 when the function is assumed to be log-normal; this is a substantial variability in the life estimate for a very small amount of variability in the location of the SN curve.

In Figure 7, a density function for the life estimate was obtained from an assumed 5% variability using the same procedures as described above. In this case, the CV was 37.5% with a mean equal to 6621. The inner three SD range is 2065 to 18587 for the lifetime value estimates. This exceptionally large dispersion in the life estimates for a moderate amount of variability (5%) in the location of the mean curve indicates instability in estimating lifetime values. It is noted that by taking the log of the data, a normal function was obtained indicating that the fatigue estimate can be represented by a log-normal distribution.

In Figure 8, a computation similar to that described in Figure 6 was performed in order to determine the difference in life values between 1% and 0.0001% points corresponding to reliabilities of 0.99 and 0.999999, respectively. A 1% variability in the spectrum was assumed in the computation of N_P. A CV of 10.8% was obtained with a mean of 6203. Results show a life of 4795 for the lower reliability of 0.99 and 3689 for the higher reliability of 0.999999 showing a 23% decrease in the lifetime estimate.
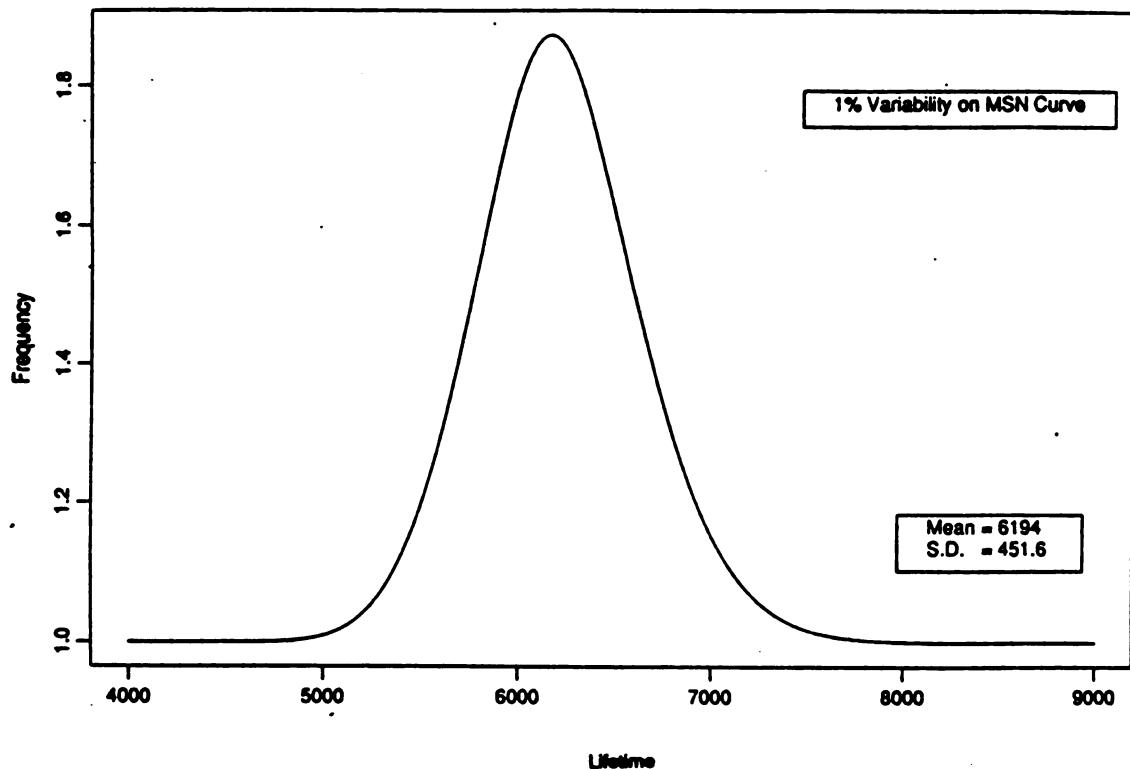
Figure 6. Component fatigue life probability density function.
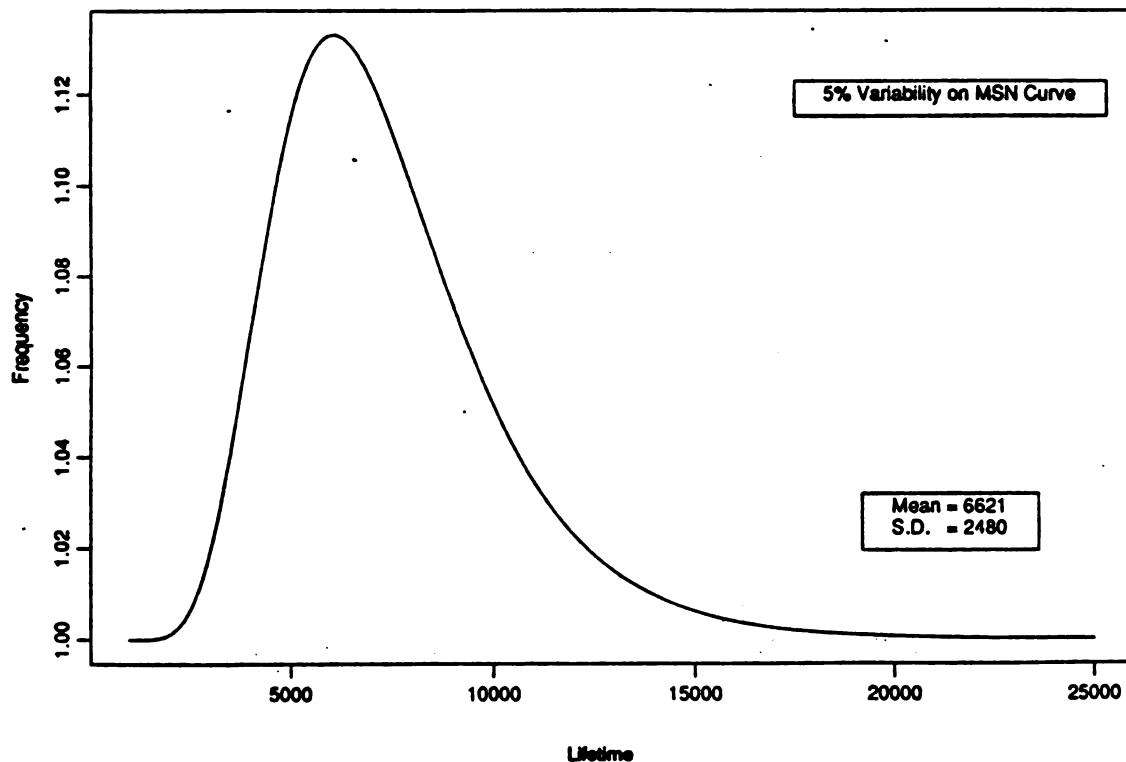
Figure 7. Component fatigue life prability density function.

41
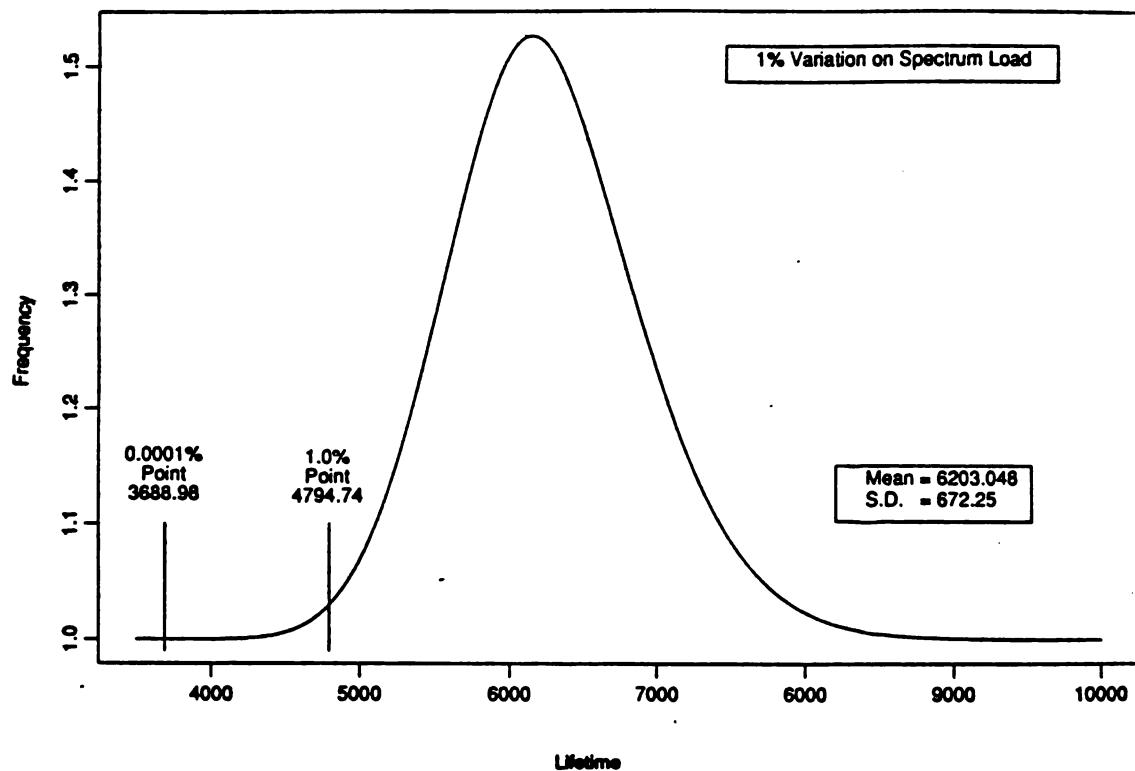
Figure 8. Component fatigue life probability density function and reliability.

Figure 9, where a 5% variability in the spectrum load was introduced, shows a log-normal distribution of lifetime values similar to that in Figure 7 for the SN_c curve variability. The inner range of 1075 to 31956 again shows the substantial variability in the life estimate indicating a serious instability in the fatigue life computation approach when even small uncertainties exist in assuming a specified spectrum load. Load spectrum and fatigue strength CVs in the range of 7% to 13% are being considered by the helicopter industry.[17] A comparison of the reliabilities of 0.99 and 0.999999 for the respective lifetimes showed 1702 and 448 passes which is a 74% decrease in lifetime. This is a much greater percent decrease than that of the 1% variability case in Figure 8. This assumed variability is probably more realistic than that of 1% which was previously assumed.

Comparison of these figures show uncertainties in safe life fatigue design in terms of changes in design lifetime for a fixed reliability, whereas the results in Tables 1 through 4 show variability in terms of changes in reliability for fixed lifetimes.

Although only a simple case has been considered, the modeling and simulation processes are capable of dealing with more complex safe life fatigue designs. Such designs could include more complex load spectra and additional parameters in the fatigue life model. The value of any high reliability based analyses, whether simple or complex, appears to be in question in view of the very substantial sensitivity of the reliability and lifetime results from this study.

Digitized by Google
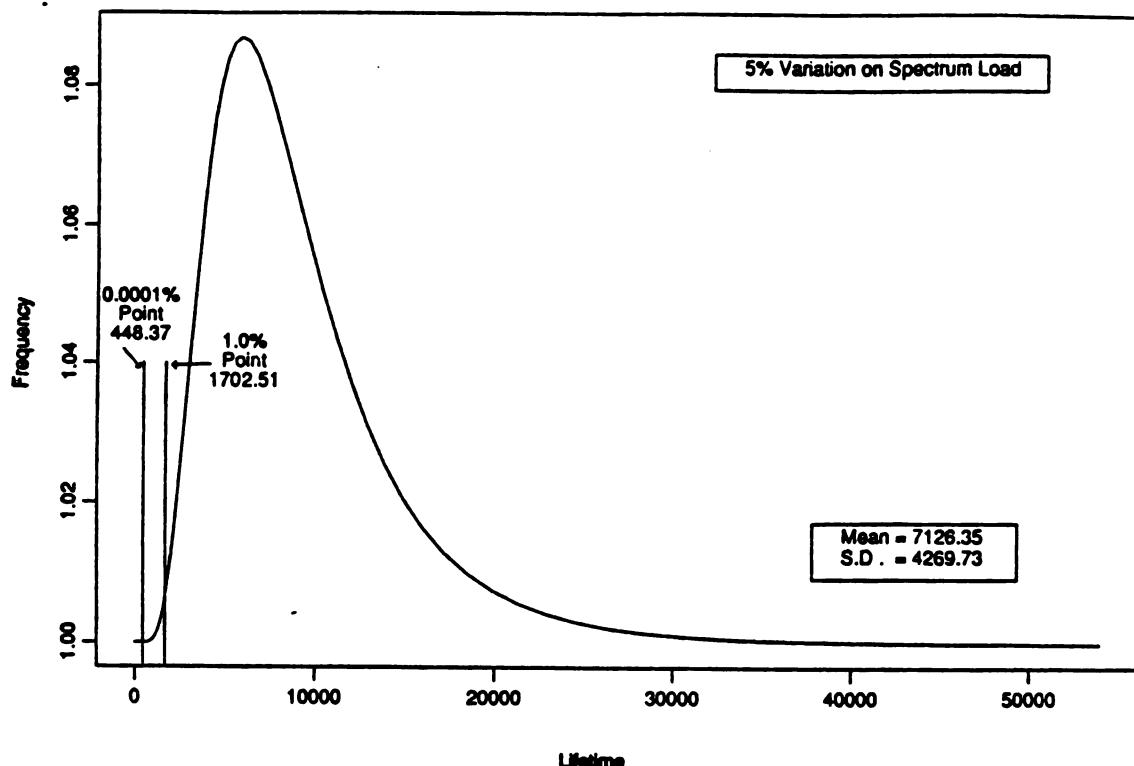
Figure 9. Component fatigue life probability density function and reliability.

CONCLUSIONS

A small amount of variability (uncertainty) in load or strength in the safe life fatigue model can result in a substantial reduction in high reliability values for a specified lifetime of a component. These uncertainties can also result in very unstable lifetime estimates for a given reliability. In contrast, the small variations assumed in the Miner's Rule criterion, and the variability in the SN coupon curve determination, caused a minimal amount of change in the reliability estimates.

A small percent reduction in the strength values in the component SN curve; for example, corrosion effects, can result in a large decrease in the reliability values.

Introducing working curves in the fatigue life computation is only effective when there is a small amount of variability in the SN component curve or when the reduction factor was very large.

In view of the sensitivity of the safe life reliability criterion of 0.999999 to the modest variability considered in this analysis, it appears that the 0.999999 reliability is ineffective as a criterion to ensure safety for a specified service life. **In summary, this report has identified a potential problem associated with obtaining a meaningful quantitative measure of reliability for a fatigue loaded component.**

43

REFERENCES

1. ARDEN, R. W. *Hypothetical Fatigue Life Problems*. Proceedings of the American Helicopter Society, Midwest Region, Helicopter Fatigue Methodology Specialists' Meeting, March 1980.

2. ARDEN, R. W., and IMMEN, F. H. *U.S. Army Requirements for Fatigue Integrity*. Proceedings of the American Helicopter Society National Technical Specialists' Meeting on Advanced Rotorcraft Structures, Williamsburg, VA, October 1988.

3. AMER, K. B. *A New Philosophy of Structural Reliability, Fail Safe Versus Safe Life*. Journal of the American Helicopter Society, v. 34, no. 1, January 1989.

4. KRASNOWSKI, B. R. *Designing Rotorcraft Dynamics Components to Reliability Requirements*. Presented at the National Fatigue Specialists' Meeting, American Helicopter Society, Scottsdale, AZ, October 1989.

5. SPIGEL, B. S. *Safe Life Design for Rotorcraft*. Journal of the American Helicopter Society, v. 36, no. 1, January 1991.

6. THOMPSON, A. E., and ADAMS, D. O. *A Computational Method for the Determination of Structural Reliability of Helicopter Dynamic Components*. Presented at the American Helicopter Society Annual Forum, Washington DC, May 1990.

7. EVERETT, R. A., BARTLETT, F. D., and ELBER, W. *Probabalistic Fatigue Methodology for Six Nines Reliability*. AVSCOM Technical Report 90-B-009, NASA Technical Memorandum 102757, December 1990.

8. NEAL, D. M., MATHEWS, W. T., and VANGEL, M. G. *Model Sensitivity in Stress-Strength Reliability Computations*. U.S. Army Materials Technology Laboratory, MTL TR 91-3, January 1991.

9. MINER, M. A. *Cumulative Damage in Fatigue*. Journal of Applied Mechanics, v. 12, 1945, p. A150-A164.

10. BURY, K. V. *Statistical Models in Applied Science*. John Wiley & Sons, 1975, p. 598.

11. BERENS, A. P. *Helicopter Fatigue Methodology*. University of Dayton Report, AVSCOM TR-87-D-13A, v. 1, December 1987.

12. WEIBULL, W. *Fatigue Testing and Analysis of Results*. Pergamon Press, New York, NY, 1961.

13. "RNLIN," IMSL Stat/Library: FORTRAN Subroutines for Statistical Analysis, v. 1, ch. 2, April 1987, p. 239-247.

14. EFRON, B. *Bootstrap Methods: Another Look at the Jackknife*. Annals of Statistics, v. 7, 1979, p. 1-26.

15. GUNSALLUS, C. T., HARDERSEN, C. P., and STENNETT, P. G. *Investigation of Fatigue Methodology*. US-AAVSCOM TR 87-D-17, U.S. Army Aviation Applied Technology Directorate, Ft. Eustis, VA, May 1988.

1.6 NOBACK, R. *State of the Art and Statistical Aspects of Helicopter Fatigue Substantiation Procedures*. Helicopter Fatigue Life Assessment, AGARD CP No. 297, North Atlantic Treaty Organization, March 1981.

17. SCHNEIDER, G., and GUNSALLUS, C. *Continuation of the AHS Round Robin on Fatigue Reliability and Damage Tolerance*. Presented at the American Helicopter Society 47th Annual Forum, Phoenix, AZ, May 1991.
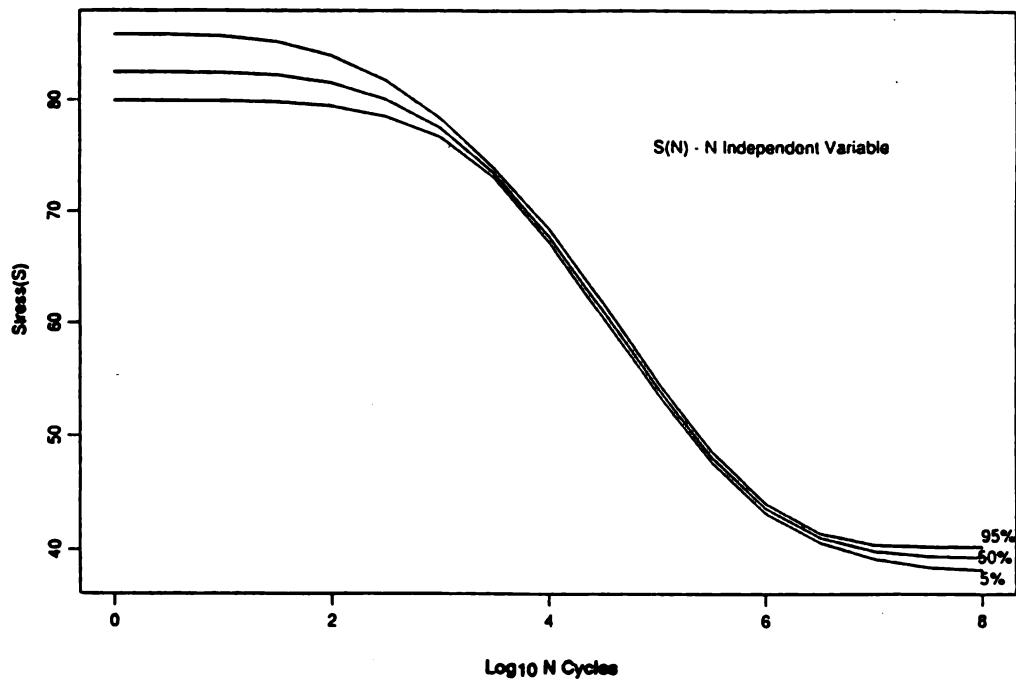
Figure A1. Confidence interval on mean coupon SN curve obtained from fatigue data.
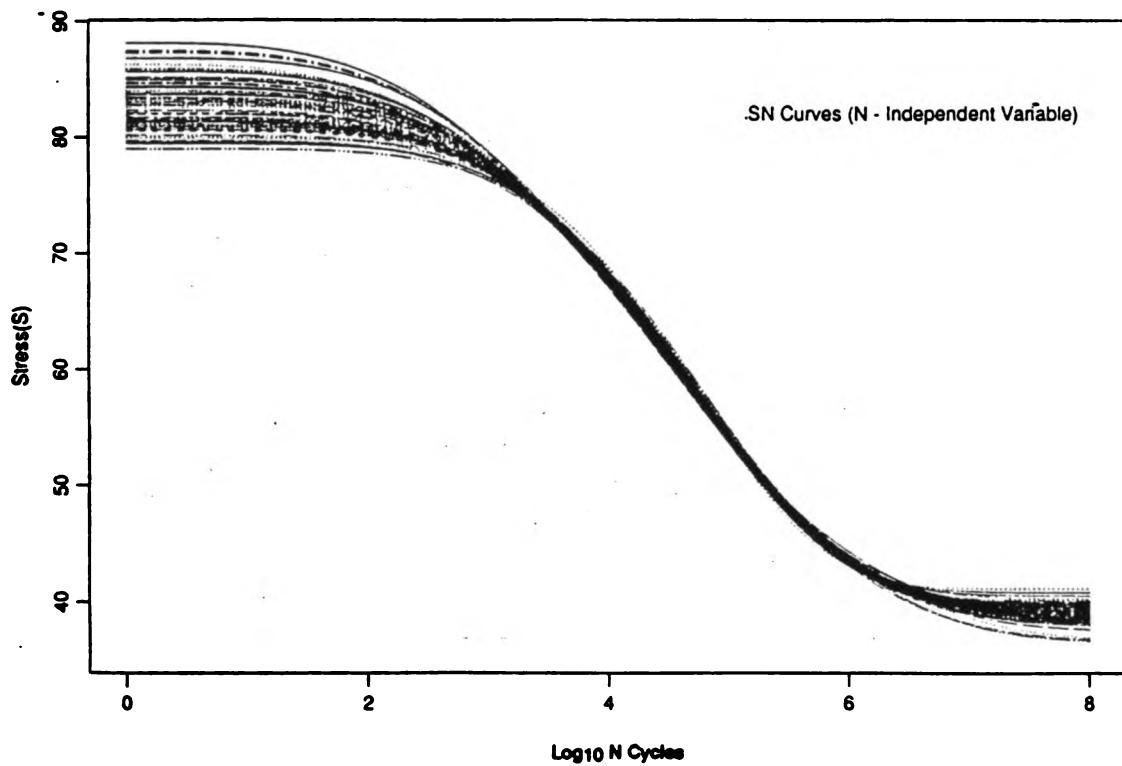
Figure A2. Regression SN curves from Bootstrapping (N-independent variable).

$$S_m^* = S_m(1 + V_S \cdot Z_i)$$
S_m = Deterministic Result from the Analysis
V_S = Coefficient of Variation
Z_i = Standard Normal N(0,1)
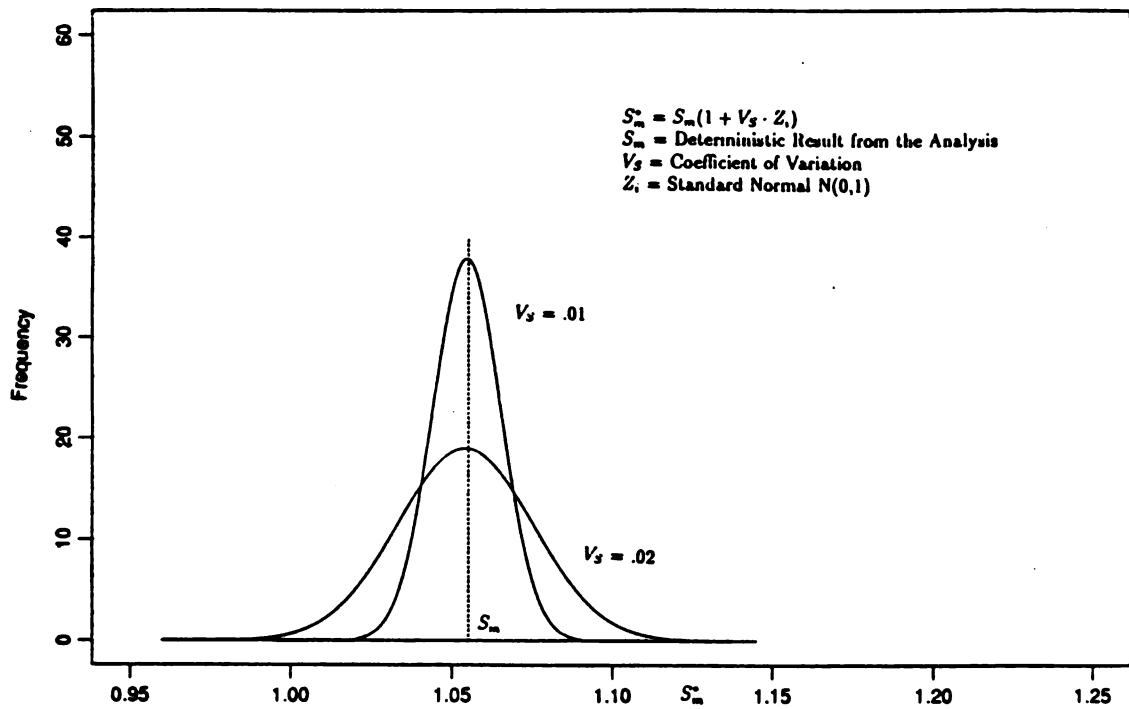
Figure A3. PDF component curve mean value.
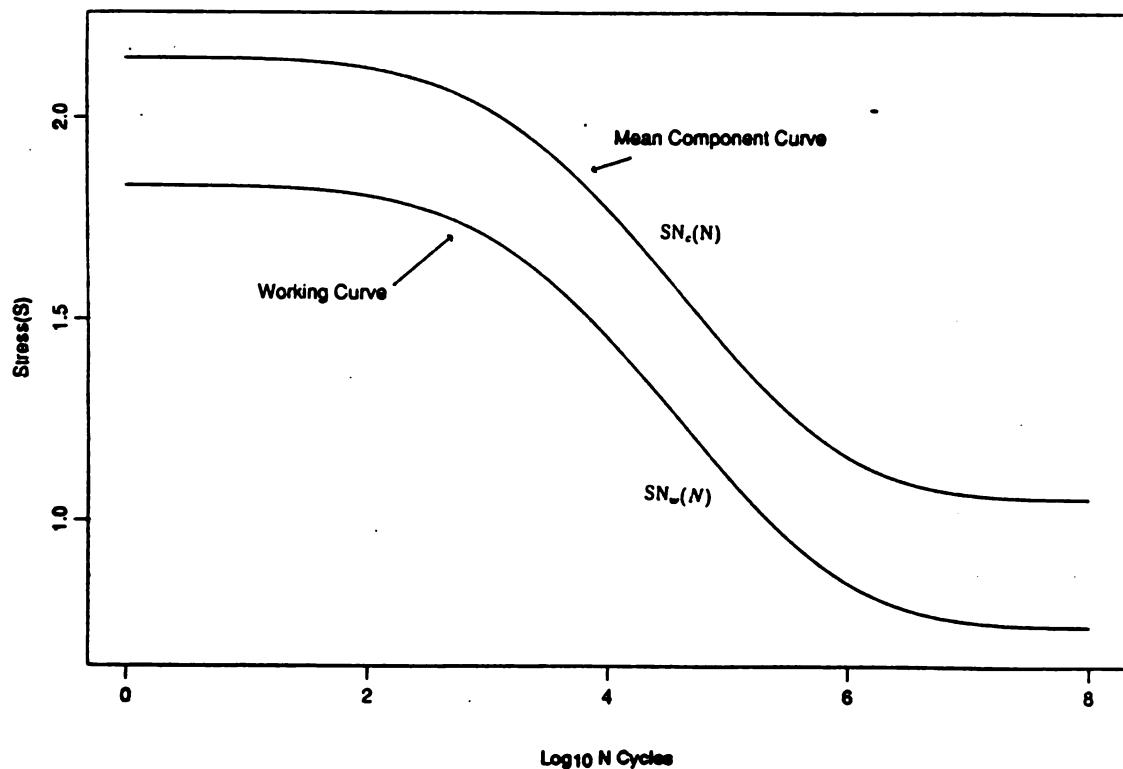
Figure A4. Mean and working component SN curves.

46

A Randomization Approach
to an Analysis of Ballistic Data

Malcolm S. Taylor
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5066

Abstract

Randomization procedures offer a viable approach to the analysis of ballistic data over a wide class of problems. Distribution assumptions may be avoided and, of even greater importance, random samples of data are not required. Small sample sizes, while never welcome, may be accommodated as well.

Introduction

This is an applications paper that details a problem that is representative in many respects of those engendered by ballistic data. Sample sizes are woefully small due to the cost of data collection and/or scarcity of testing materiel. The samples themselves are usually nonrandom and distribution assumptions are tentative. Historical data, when available, cannot be easily amalgamated to assist in inference.

An approach known generically as randomization, suggested by Fisher [2] and extended to nonrandom samples by Pitman [5] holds particular appeal, since distribution assumptions and random sample requirements may be relaxed. Edgington [1] asserts that "Few experiments in biology, education, medicine, psychology, or any other field use randomly selected subjects, and those that do usually concern populations so specific as to be of little interest. ... The population of interest to the experimenter is likely to be one that cannot be sampled randomly." Edgington's words ring true in the example to follow.

The problem: Stability of a kinetic energy penetrator

Kinetic energy penetrator technology has undergone a metamorphosis from the days when solid balls were launched from cannons or catapults against sailing ships and forts. The most obvious change has been in the overall configuration of the projectile. The ratio of the projectile's length to its diameter has gone from one to over twenty, as illustrated in Figure 1. This change has taken place largely in response to the changing targets which kinetic energy penetrators must confront.

47
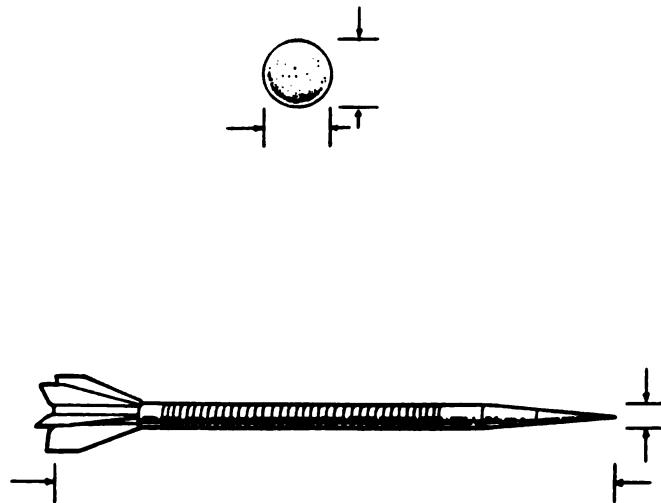
Figure 1. KE Penetrator Evolvement

The changes in tank ammunition design have followed the trend depicted in Figure 2. The armor piercing, discarding sabot ammunition (Fig. 2(b)) has a penetrator whose outer diameter is less than the inner bore diameter of the gun tube. The difference is compensated for by a sabot, which carries the penetrator down the gun tube and is then discarded. The long rod penetrator (Fig. 2(c)) is essentially a long rod of exceedingly dense material, typically tungsten alloy or depleted uranium, over twice as dense as steel. In addition to a discarding sabot, the penetrator has fins which increase the stability of the rod in flight.

(b)

(c)
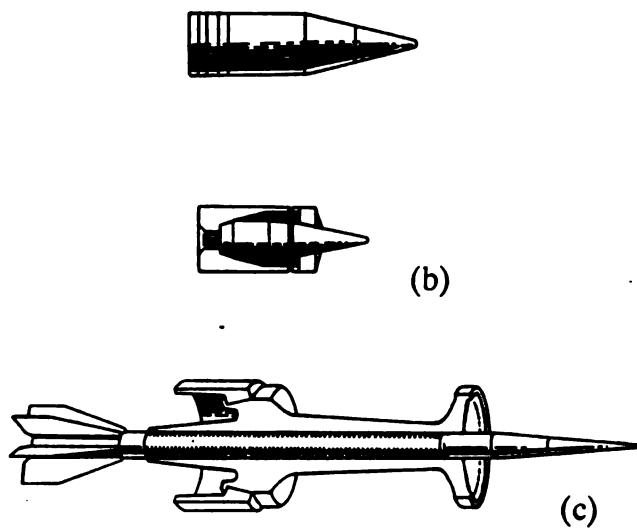
Figure 2. Armor Piercing Ammunition

Table 1 contains measurements of spin rates of long rod penetrators taken by Rapacki [6]. The natural frequency of the penetrators is about 120 cycles per second (hz). Spin rates close to this value amplify the initial manufacturing imperfections and increase in-flight bending.

To avoid this, the fins are reconfigured to reduce the spin rate to an appropriate level below 120 hz.

Table 1. Comparison of Two Fin Redesigns with a Control

design-0	redesign-1	redesign-2
163.6	97.5	78.1
109.0	122.2	76.7
218.7	108.2	88.5
143.2		
169.5		

If the spin rate is too high, as in design-0 (or control), the penetrator may become bowed in flight – sometimes to the point of breaking – and become unstable. Conversely, if the spin rate is too low, the penetrator may again become unstable. An optimal spin rate cannot be determined analytically, and resources are not adequate for extensive empirical study.

An engineering consideration with important implications for analysis of these data is the following: As the penetrator becomes more stable, the variance of the measured spin rates will decrease. For analysis of the data in Table 1, this establishes a multi-sample situation with possible heterogeneity of variance between samples, and where variance stabilizing transformations are inappropriate since both difference in location and dispersion is relevant to inference about the penetrator design. This effectively removes from consideration classical analysis of variance procedures for analysis of these data.
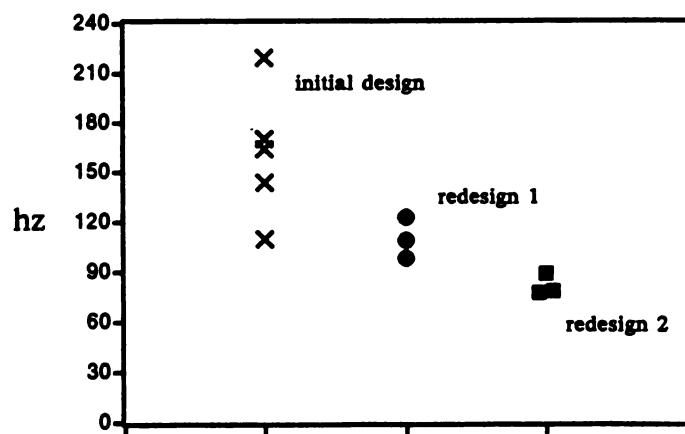
Figure 3. Spin Rate Data

A randomization approach to the data analysis

Figure 3 suggests that the fin reconfigurations had the intended effect – reducing the spin rate and increasing stability. To pursue quantitative support for this observation we will appeal to a randomization argument, choosing as a null hypothesis that the two fin redesigns are ineffectual and provide no improvement over the initial design. If the null hypothesis is true, then the categorical labels: design-0, redesign-1, and redesign-2, are completely arbitrary, and the eleven observations could be randomly assigned to the columns of Table 1 (retaining the same number of observations per column) without any attendant statistical consequences.[1]

We will consider restricted null hypotheses in which redesign-1 is compared to control and then redesign-2 is compared to control, rather than an omnibus test. This focuses attention on the comparisons of interest while easing the overall computational burden. Figure 4 represents the $_8C_3 = 56$ data configurations that are produced by systematic reassignment of datum values within columns one and two of Table 1. For each resultant configuration, the difference in location between control and redesign-1, $\bar{x}_0 - \bar{x}_1$, is plotted on the x-axis and the variance ratio control/redesign-1, s_0/s_1, is plotted on the y-axis.
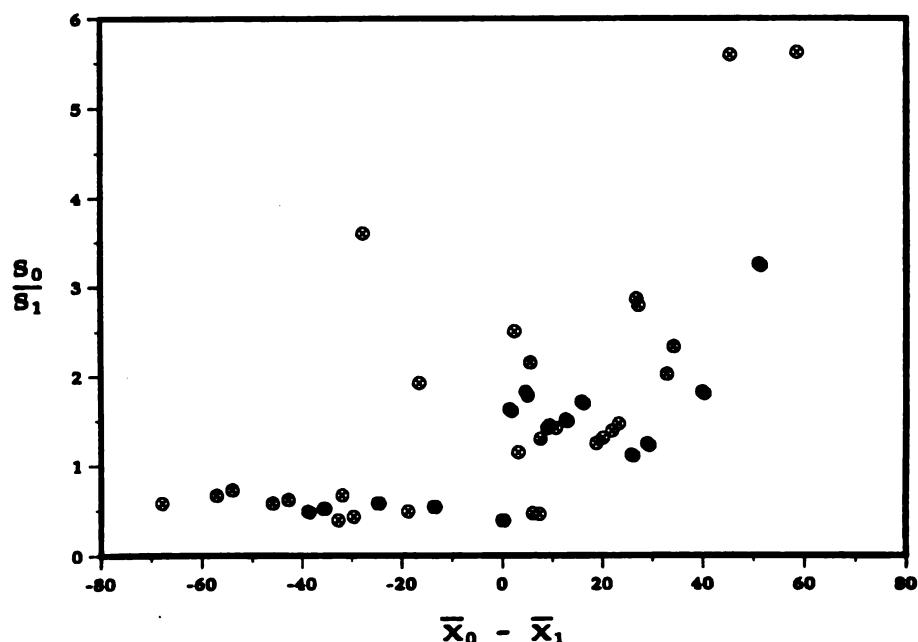
Figure 4. Data Display after Randomization

---

1. Some authors assume random assignment of homogeneous experimental units to control and treatment groups. We are necessarily in violation of this assumption, and arguably are detailing a permutation test rather than a randomization test. In either case, the procedure remains invariant.

To determine an observed significance level of the data in Table 1 relative to the data sets generated by reassignment, the point $(\overline{x_0} - \overline{x_1}, s_0/s_1)$ calculated from the data in columns one and two of Table 1 can be ranked against the remaining fifty-five points. We will specify a naive procedure for ranking the ordered pairs (x_i, y_i) which will suffice for these data, and which retains the structure of nonparametric rank tests (Lehmann [3]). We first rank the x-coordinates, assigning to the largest value the rank 1, the second largest rank 2, and so on. We rank the y-coordinates in the same way. Finally, we sum the ranks assigned to the x- and y-coordinates. In case of ties, the ranks are averaged.

Using this procedure, the observed data $(51.50, 3.24) \equiv (2, 4)$ having combined rank six, is tied with two other pairs and is assigned an overall rank of three. The observed significance level is then $3/56 = .054$

We knew beforehand the two restricted hypotheses of interest, and as such might invoke a planned comparisons argument. But, we are testing two of three possible comparisons and so a multiple comparisons procedure is a more conservative approach. Experimentwise error rate (Miller [4]) introduced through multiple comparisons will be controlled with the aid of Fisher's modified least significant difference procedure (Winer [7]) which has the desirable properties of being both nonparametric and applicable to unequal sample sizes.

Suppose we specify an experimentwise error rate of $\alpha' = .05$ for comparison of the two fin redesigns with the control. Adopting the obvious notation c, d1, d2 for control and redesign, we are interested in the comparisons c-d1 and c-d2. The observed significance level is determined for each of the pairwise comparisons following the randomization procedure outlined above. Each p-value is then multiplied by two (the number of comparisons) in accordance with Fisher's procedure to obtain an adjusted p-value. The p-values and adjusted p-values for comparison of c-d1 and c-d2 are given in Table 2.

Table 2. Multiple Comparison of Control and Two Treatments

comparison	c-d1	c-d2
p-value	.054	.018
adjusted p-value	.107	.036

The adjusted p-value, .036, corresponding to comparison of control and redesign-2, falls well below the $\alpha' = .05$ value chosen for experimentwise error rate, and reflects a statistically significant difference between the two penetrator designs. Comparison of control and redesign-1, with an adjusted p-value of .107, exceeds $\alpha' = .05$, and does not substantiate a claim of difference. These conclusions, now quantified, remain consistent with the display in Figure 1.

Conclusion

Randomization procedures offer a viable approach to the analysis of ballistic data over a wide class of problems. Distribution assumptions may be avoided and, of even greater

importance, random samples of data are not required. Small sample sizes, while never welcome, may be accommodated as well.

In statistics, as elsewhere, there is no free lunch. The price paid for randomization is increased computation, since every problem requires a tailored solution, reflected through the enumerative process required to determine the p-values. However, use of the normal theory statistics – t-test, F-test, chi-square test, etc. – may only be valid to the extent that they approximate the p-values obtained from randomization.

Acknowledgement

References

[1] E.S. Edgington, *Randomization Tests, 2nd. Ed.* (Marcel Dekker, New York, 1987).

[2] R.A. Fisher, *Design of Experiments* (Oliver & Boyd, Edinburgh, 1935).

[3] E.L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks* (Holden-Day, San Francisco, 1975) 226-232.

[4] R.G. Miller, Jr., *Simultaneous Statistical Inference, 2nd. Ed.* (Springer-Verlag, New York, 1981) 250-251.

[5] E.J.G. Pitman, Significance tests which may be applied to samples from any populations,*J. R. Statist. Soc., B.*4 (1937) 119-130.

[6] E.J. Rapacki, Private communication, unpublished analysis presented at XM907 Interim Program Review, U.S. Army Ballistic Research Laboratory, (1986).

[7] B.J. Winer, *Statistical Principles in Experimental Design* (McGraw-Hill, New York, 1971) 199-200.

A Randomization Test for Comparing 1/4-Scale Kinetic Energy Penetrators

Barry A. Bodt
Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Abstract The material properties of kinetic energy penetrators are studied in a 1/4-scale test environment at the Ballistic Research Laboratory. Metallurgists fire penetrators of various material compositions into semi-infinite steel blocks and record depths of penetration. Depth of penetration behaves approximately as a linear function of velocity, d(v), over the range of the four-velocity design routinely employed. Under a common slopes assumption, a difference in performance between penetrators k and l is computed as $d_k(v) - d_l(v)$. This difference is determined graphically, occasionally with the benefit of a least-squares fit to each performance. Statements of significance are not made at present. In this paper, a randomization test is examined as a means for providing analytical support for inference.

1. Introduction

Material properties of kinetic energy penetrators are compared at the Ballistic Research Laboratory in a 1/4-scale test environment. Metallurgists fire penetrators of various material compositions into semi-infinite steel blocks and record depths of penetration. Depth of penetration behaves approximately as a linear function of velocity, d(v), over the range of the four-velocity design routinely employed. Under a common slopes assumption, a difference in performance between penetrators k and l is computed as $d_k(v) - d_l(v)$. This difference is determined graphically, occasionally with the benefit of a least-squares fit to each performance. Statements of significance are not made at present. In this paper, a randomization test is presented as a means for providing analytical support for inference.

Inferences drawn from such experimentation may be considered the result of meta-analysis. Meta-analysis is loosely described as the "integration of independent studies" in a book by Hedges and Olkin [1985]. This area has received much recent attention in the social and biological sciences, but in the physical and engineering sciences it has received little notice with the exception of a few historical papers (e.g., Tippet [1931] and Fisher [1932]) that have been classified in retrospect as meta-analyses. The independent-studies quality of the aforementioned problem stems from the combination of data sets gathered at different times (often different years) and by different experimenters. This fact, practically speaking, invalidates a necessary assumption for normal theory analyses, namely the belief that the subjects for the combined data set are the result of a random sample.[1] Taylor and Bodt [1991] recommend surmounting this problem through the use of randomization tests and demonstrate applicability of this methodology to significance testing with ballistic data.

---

[1] In an ideal situation one would design a multiyear experiment where random sampling did occur, but the the obstacles are so formidable in this testing environment that it is not done.

The purpose of this paper is to introduce a randomization test for comparing 1/4-scale kinetic energy penetrators. A description of the data collection is followed by the discussion of a linear model through which significance testing of relevant contrasts can be made. It is then demonstrated how a reference distribution for determining significance can be achieved through randomization. Application of the procedure and discussion of the results follow.

2. The Data Collection

The measured response, d_{ij}, is the depth of penetration permitted by a semi-infinite steel block subjected to a hemi-nose penetrator of material j, fired at velocity i. Semi-infinite describes the independence of the penetration action to influences from side and rear free surfaces (i.e., the block is for practical purposes infinite with respect to width and depth). Hemi-nose refers to the hemispherical configuration of the projectile nose. Figure 1 shows the cut-away profile of a semi-infinite block, where the cut is made along the shot line. Depth of penetration is taken to be the maximum normal distance between the original entry-point surface and the bottom surface of the hole.

Depth of penetration from penetrators of several different material compositions are gathered over several velocities. The design structure suggests that the experimental units are the semi-infinite steel blocks. It is these that are exposed to the two treatments, velocity and penetrator material. Velocity is included as a test condition because it will affect penetration depth. Penetrator material is the only treatment of interest – materials are to be compared for relative effectiveness. Confidence in the assessment of relative performance is ensured through comparison over a range of velocities meaningful to the Army application (i.e., over a typical ordnance velocity range). A template for the experiment is to fire each penetrator (material) once at each of the following four nominal velocities: 1100 m/s, 1300 m/s, 1500 m/s, and 1700 m/s. Actual velocities will vary. A design matrix overlaid on a combined data set including different materials might appear as Figure 2.

Other facets of data collection influence the analysis. Penetrators are tested in separate experiments, quite possibly over as many as ten years if the purpose is to compare new materials to an historical control. Small sample sizes with no replication prevail if one adheres to the template for testing materials. There is no random sampling from a population of semi-infinite blocks – indeed, at the time of the first experiment, blocks used in later firings may have not yet been manufactured. Even if the sample were random, there is no guarantee that the population is normal, nor is it likely that the comfort of approximate normality can be afforded by the Central Limit Theorem with the sample sizes and replication considered.

3. The Linear Model

A linear models framework is presented in this section to support inference for this problem. Great detail is not given. For a comprehensive, but introductory, treatment, it is suggested the reader turn to Neter and Wasserman [1974]. The problem is first described in the context of a two-factor factorial design, followed by a refinement in the form of an analysis of covariance model. A convenient regression form of this model is then used to construct meaningful contrasts, and assumptions required for traditional significance testing of those contrasts are discussed.
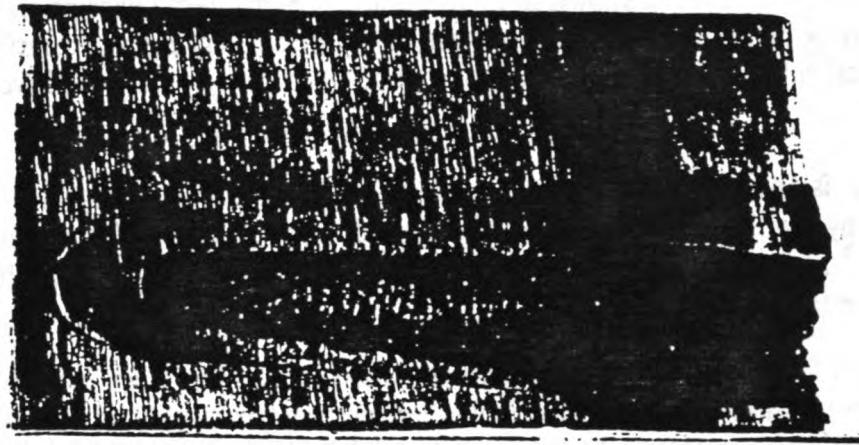
Figure 1. Cut-away Profile of a Semi-infinite Steel Block After Penetration.

Figure 2. Template for Data Collection.

3.1 Factorial Design

The design matrix shown in Figure 2 and the problem description suggest that a factorial design may be appropriate, with penetrator material serving as the principle treatment under study and velocity serving as an additional design variable. The additive model is expressed

$$d_{ij} = \mu + V_i + M_j + e_{ij}, \tag{1}$$

where μ is the common mean response, V_i and M_j are the effects (shifts from that mean) caused by the ith velocity and the jth material, respectively, and e_{ij} is the error associated with the $(ij)^{th}$ response. A Model-I stance is assumed, indicating that both material and velocity be treated as fixed effects.

Two facts render this approach less than ideal. The first, stated in the Introduction, is that experimenters know that velocity behaves approximately linearly with penetration depth. Even further, experience has shown that $d_k(v)$ and $d_l(v)$ are virtually parallel over the 1100 m/s to 1700 m/s velocity regime, hence the additivity assumption above. Beyond this regime the assumptions of linearity and parallel lines do not hold. The second is that although four nominal velocities are intended, the actual velocities tested often number as many as the number of 1/4-scale rods fired. Because firing velocity cannot be completely controlled, each nominal velocity actually encompasses a range of velocities close to the nominal. Figure 3 illustrates both linearity and firing velocity noise in replication of some tungsten alloy firings at the four nominal velocities.

This additional information impacts the method of analysis. Taking advantage of linearity would save the experimenter degrees of freedom to apply in the estimation of error – more efficiency in the model is possible. Left unconsidered, firing velocity noise would increase the estimate of response variability. In the next section the analysis of covariance model is suggested, having the advantage that both linearity and firing velocity variation can be incorporated.

3.2 Analysis of Covariance

3.2.1 Traditional Model

The linear relationship between velocity and depth of penetration can be made part of the linear model as follows. First, rewrite Equation 1 in terms of marginal means as

$$d_{ij} = \mu + (\mu_{i\cdot} - \mu) + (\mu_{\cdot j} - \mu) + (d_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu), \tag{2}$$

where the dot subscript means to pool over that index (i.e., to average based on the sum in the margin). Introduce in the model the term $\mu_{d/v}$ to represent the simple linear relationship between velocity and the mean response. Adding and subtracting $\mu_{d/v}$ from the right side of Equation 2 and rearranging terms leaves

$$d_{ij} = \mu_{d/v} + (\mu_{\cdot j} - \mu) + (d_{ij} - \mu_{d/v} - \mu_{\cdot j} + \mu). \tag{3}$$

Let v_{ij} represent the velocity of the $(ij)^{th}$ penetrator. The simple linear model which regresses penetration depth on velocity can then be expressed as $\mu + \gamma(v_{ij} - \bar{v}_{\cdot\cdot})$, where γ is the slope of

56
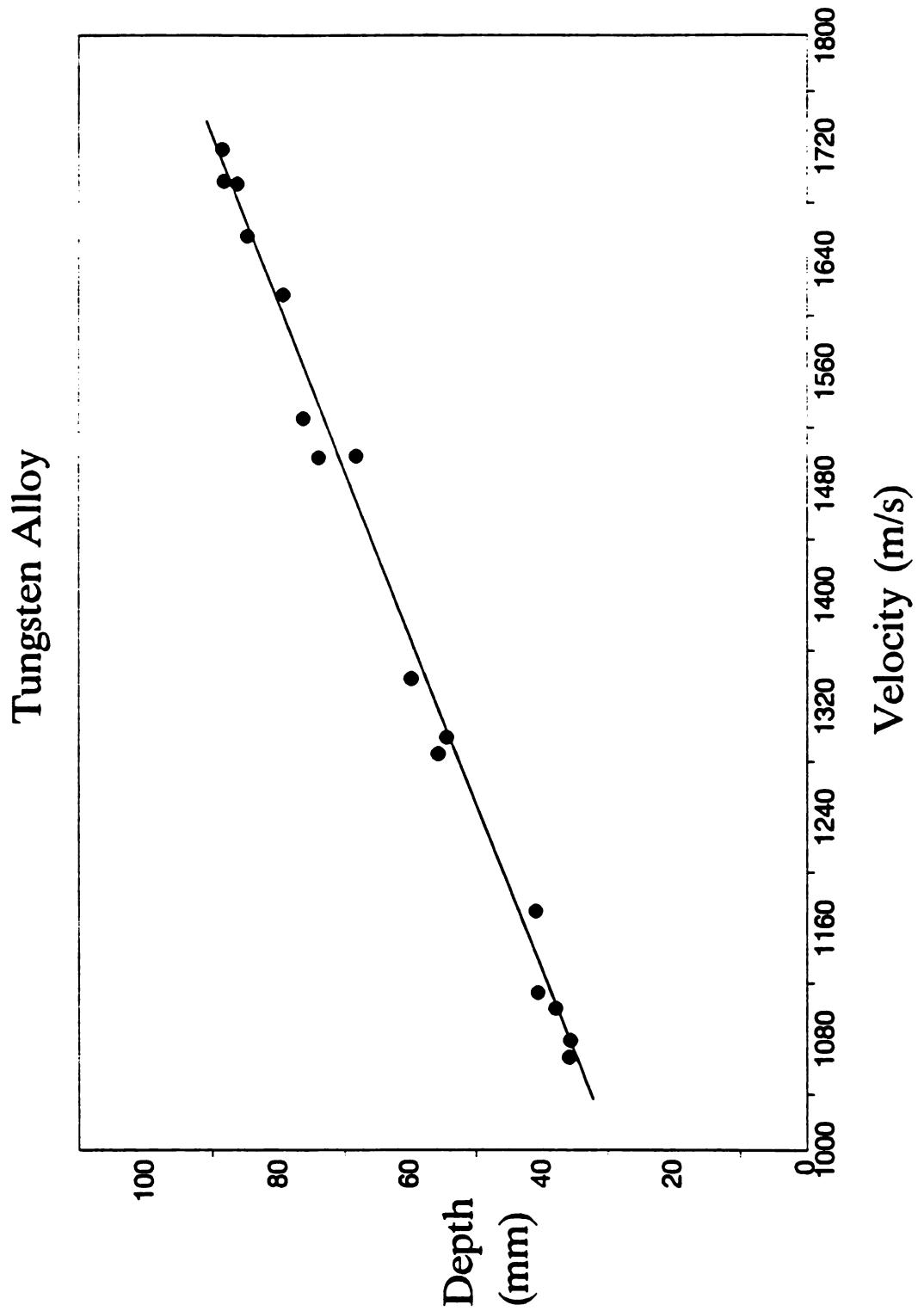
Figure 3. Tungsten Alloy Depth of Penetration Plotted Against Velocity.

the regression. Substituting this for $\mu_{d/v}$ in Equation 3 yields

$$d_{ij} = \mu + \gamma(v_{ij} - \overline{v}_{..}) + (\mu_j - \mu) + (d_{ij} - \gamma(v_{ij} - \overline{v}_{..}) - \mu_j), \tag{4}$$

which the reader recognizes as the common form of an analysis of covariance model.

Certainly, the analysis of covariance model in Equation 4 has appeal in that it can account for the contribution to penetration depth from individual velocities; whereas, in the factorial design the contribution of nominal velocities are counted as being the same regardless of noise. Further, even if the nominal velocities were exactly achieved, there is advantage to be gained in introducing the linearity information in the model. In that case, degrees of freedom for estimating error are saved. The factorial design allows (s-2) fewer degrees of freedom for error, where s represents the number of nominal velocities. This follows directly from the fact that the factorial design requires s-1 degrees of freedom be assigned to velocity; whereas, the simple linear regression needs only 1 degree of freedom assigned to the slope to account for the influence of velocity. If the regression is perfect (fits exactly to the mean response for each nominal velocity), the sum of squares associated with error for both models are identical, leaving analysis of covariance with a decided advantage. If the regression is not perfect, a tradeoff is made wherein degrees of freedom for the error term denominator are gained at the expense of the regression lack-of-fit being added in the numerator. In consideration of data with a strong linear relationship like those displayed in Figure 3, an analysis of covariance approach would be a more appropriate choice than the two-factor factorial.

Using the analysis of covariance model to describe the problem structure, questions regarding material comparisons can be answered through the study of contrasts. If the experimenter is interested in the difference in the effect of any two materials k and l, the contrast M_k-M_l would be estimated and then tested for significance.

3.2.2 Regression Formulation

It is convenient to formulate Equation 4 in terms of a regression model. From an applications perspective, the least-squares approach is more widely understood and accepted by practitioners.[2] Moreover, the parameters have greater intuitive appeal, and their meaning conforms to how experimenters at the Ballistic Research Laboratory currently think of the problem.

The change is accomplished easily. Replace the t-level treatment factor with indicator variables m_k, $k = 1, 2, \cdots$ t-1, defined such that

$$m_k = 1 \text{ if the observation is of material k;}$$

$$= 0 \text{ otherwise.}$$

[2] Regression is also of use, computationally, when the design matrix is unbalanced.

The columns in the regression design matrix corresponding to the indicator variables will be mutually orthogonal. Thus, Equation 4 may be expressed in terms of a regression model as

$$d_{ij} = \beta_0 + \beta_1 m_{ij1} + \beta_2 m_{ij2} + \cdots + \beta_{t-1} m_{ij,t-1} + \beta_t (v_{ij} - \overline{v}_{..}) + e_{ij}, \text{ where} \qquad (5)$$

$$\beta_0 = \mu + M_t,$$
$$\beta_k = M_k - M_t,$$
$$\beta_t = \gamma.$$

The coefficients β_k, $k = 1, 2, \cdots t-1$ represent the difference between the effect of the kth and t^{th} material (i.e., the vertical difference between the regression lines $d_k[v] - d_t[v]$). The designation of the tth material is arbitrary, determined by how the indicator variables are defined. In the design matrix for the regression model, the tth material would have zeros in the sub-rows corresponding to the $t-1$ indicator variables. The interpretation of the β_k's would be most natural if a reference group or an historical control was denoted the tth material. Other comparisons may also be of interest. The general contrast $M_k - M_l$, $k, l \neq t$ is obtained through the difference $\beta_k - \beta_l$.

In this section the treatment effects were expressed in the context of a regression formulation of the analysis of covariance model. Estimation of these effects can be accomplished after first determining the least squares estimate of the coefficient vector. The next step – and the main focus of this effort – is to determine the significance of these effects. To begin, a careful consideration of the assumptions is made.

3.2.3 Assumptions

Several assumptions are required to support the usual analysis of covariance for this problem. They appear as follows: 1) the regression slopes are nonzero and homogeneous among materials, 2) velocity is unaffected by material, 3) velocity is precisely measured, 4) model errors are distributed with zero mean and common variance, and 5) the responses are considered jointly independent normal random variables. The practical implication of 4) and 5) together is that penetration depths to be allowed by the semi-infinite blocks constitute a random sample from some conceptual normal population.

The first four assumptions are accepted; the last is not. Velocity obviously affects penetration depth, and data support the similar-slopes claim. All test penetrators are identical in dimension; there is no reason to expect that velocity will be influenced by which material composition is being tested. Velocity, though not completely controlled, is precisely measured using an x-ray multiflash system. As for the last assumption, there is no reason to expect that penetration depths are normal, and because of the individual-study nature of the experiments, they do not constitute a random sample.

In Section 4 we relax this last assumption to require only that the penetration depths be pairwise uncorrelated. With that change, the least-squares estimation of the parameters in Equation 5 will retain the usual properties of uniform minimum variance among linear unbiased estimators but without any known distribution on which to base tests of significance. Under these revised model assumptions, an alternative test for significance is given.

4. A Randomization Test

In this section a randomization test is proposed as a means to discern among statistically different materials. Its principal advantages are freedom from the assumption that data under consideration constitute a random sample from a normal population and the ability to provide exact significance levels. Some basic foundation is followed by a description of the test.

4.1 Foundation

A randomization test is a method through which significance testing is accomplished, with the sampling distribution of the test statistic derived from permutations (combinations) of the data. A test of significance measures the numerical evidence against a conjecture. Data, conveyed through a suitable test statistic, are examined as to their consistency with the conjecture by comparing the observed value of the test statistic to its sampling distribution — formed assuming the conjecture is true. Degrees of inconsistency are reflected in how unusual the observed test statistic appears. This appearance is measured in terms of the p-value, the probability that a value of the test statistic is at least as unusual (large or small) as the one observed.

A classical analysis in this 1/4-scale penetrator environment, based on the model of Section 3.2.2, suggests that a conjecture (null hypothesis) of either $H_0:\beta_k = 0$ or $H_0:\beta_k - \beta_l = 0$ might be tested to compare two materials. Consider the latter hypothesis, a claimed equivalence between materials k and l. Letting b denote a least-squares estimate for β, $b_k - b_l$ is the estimated difference between materials k and l (i.e., the estimated vertical distance between their parallel regression lines). To determine whether the distance is statistically significant, one need only compare $b_k - b_l$ to its sampling distribution. This distribution is readily attainable, but only if one is willing to assume a normal random sample — not satisfied here.

Useful significance tests are possible without benefit of assumption 5). In what follows, this assumption is replaced with the less restrictive condition that penetration depths be pairwise uncorrelated, thus guaranteeing nice properties for the least-squares estimators. Before proceeding we should note that others have circumvented the normality requirement. Nonparametric approaches to this problem include papers by Quade [1967], Puri and Sen [1969], Shirley [1981], Conover and Iman [1982], and Stephenson and Jacobson [1988]. All focus on the rank transforms of either the response variable, the concomitant variables, or both. For example, Conover and Iman [1982] transform both sets of variables to ranks and then conduct a parametric analysis of covariance, eventually relying on the F-distribution to determine significance. An exception to complete reliance on ranks is found in Puri and Sen [1969]. In that paper general scores, including ranks, are adjusted for regression on the concomitant variables, and the asymptotic distribution of the test statistic based on those scores is developed using permutation. The hypothesis tested is that no difference exists overall among the treatments (materials) studied. A related approach is now described, focusing on the pairwise comparison of materials.

4.2 Description

Consider first $H_0: \beta_k = 0$.[3] The geometrical interpretation of β_k is that it is the vertical distance between the parallel regression lines $d_k(v)$ and $d_t(v)$. This fact is evident from Equation 5. The linear effect of velocity can be removed by adjusting the penetration depth values for the velocities used to achieve them – the remaining difference among the adjusted values, excluding random variation, is attributable to material and is expressed β_k. This difference is estimated as b_k by subtracting the average of the residuals resulting from material t from those of material k, the residuals being computed relative to $d_t(v)$ in each case. Thus, once the two groups of residuals are formed, we are interested in the difference in location between them.

To determine if this difference is significant, we need only establish a reference distribution and compare the observed difference to it. Under the null hypothesis, $d_k(v)$ and $d_t(v)$ are coincident. Thus, the residuals computed after adjusting for the linear effect of velocity should be homogeneous. Therefore, in computing b_k, the distinction of which residuals resulted from assignment (association) with material k or material t should make little difference. The reference distribution is constructed by computing b_k under all possible assignments of residuals (effectively ignoring material distinction) to the two materials, the cardinality of each material set being preserved.[4] For example, if material k had five data values and material t had four, there would be $_{(5+4)}C_5$ values computed for b_k. The p-value for the two-sided alternative hypothesis is simply the ratio of the number of values in the reference distribution which equal or exceed in absolute value the observed $|b_k|$ to the total number of combinations, $_{(5+4)}C_5$.

Significance testing for the hypothesis $H_0: \beta_k - \beta_l = 0$ is achieved similarly. Adjust penetration values for the linear effect of velocity and compute residuals in the same manner, still computing the residuals relative to $d_t(v)$. The difference between materials is estimated by $b_k - b_l$ and computed by subtracting the average of the residuals resulting from material l from those of material k. The reference distribution arises from computing $b_k - b_l$ under all possible assignments of residuals between materials k and l.

Before turning to examples, some more detail is required as to how these residuals, relative to $d_t(v)$ are computed. From Equation 5, the model $d_t(v)$ can be expressed

$$d_t(v) = \beta_0 + \beta_t(v - \overline{v}_{..}). \tag{6}$$

(The indices has been suppressed to emphasize that this is a model for penetration depth.) Both β_0 and β_t must be estimated. Begin with slope. Assuming parallel

---

[3] Specifically, the null hypothesis for the randomization test is that penetration depth measurements are stochastically independent of the penetrator having been formed from material k or material t [Edgington 1987].

[4] This rationale presupposes random allocation of subjects to treatments. However, as pointed out by Edgington [1987], random allocation principally guards against undue influence resulting from between or within subject variability. Such variability in the context of semi-infinite steel blocks is considered negligible relative to the material differences under study.

penetration-against-velocity models, d(v), the common slope is taken as the average within-materials regression slope, b_t, which can be delivered by any regression subroutine fitting the regression expressed as Equation 5 in its complete form. The estimate is computed as

$$b_t = \frac{\sum_i \sum_j (v_{ij} - \overline{v}_{.j})(d_{ij} - \overline{d}_{.j})}{\sum_i \sum_j (v_{ij} - \overline{v}_{.j})^2}.$$ (7)

The ordinate at $\hat{v} = \overline{v}_{..}$, β_0, is taken as an adjusted mean penetration depth for the tth material, appearing as Equation 8.

$$\overline{d}_{.t(adj)} = \overline{d}_{.t} - b_t(\overline{v}_{.t} - \overline{v}_{..}).$$ (8)

This too will be delivered by a regression of Equation 5 when zeros are used as the values for the t-1 indicator variables in the data rows corresponding to the tth material. Using estimates for β_0 and β_t, the estimated model for the tth material takes the form

$$\hat{d}_{ij(t)} = \overline{d}_{.t(adj)} + b_t(v_{ij} - \overline{v}_{..}).$$ (9)

Equation 9 is merely the least-squares fit for the tth material, taking into consideration the common slope. The residuals for the jth material relative to the tth material appear as

$$r_{ij(t)} = d_{ij} - \hat{d}_{ij(t)}.$$ (10)

The residuals $r_{ij(t)}$ are then manipulated in the manner described above.

5. Examples

In this section two examples are discussed. The purpose of the first is to provide a detailed synopsis of how the randomization test is performed. In that example, data are characteristically sparse. The purpose of the second is to illustrate performance when data are slightly more abundant and when the data collection does not exactly follow the template discussed earlier. Data for both examples were extracted from an unpublished manuscript provided by Mr. Timothy Farrand of the Ballistic Research Laboratory.

5.1 Example 1

Figure 4 displays data arising from the firing of four penetrator (material) types against semi-infinite steel blocks. All penetrators were manufactured with a common mass of 65 g and with the length-over-diameter ratio (L/D) equal to 15. The depleted uranium (DU) penetrators are separated according to Rockwell hardness (Rc). It is apparent that the template for data collection given in Figure 2 was approximately followed, save duplicate 97%-tungsten results at 1500 m/s and no result for Du Rc=45 at 1100 m/s. Four data points are the most recorded for any material. Data are listed in Table 1.

Two tasks must be accomplished on the way to significance testing. The first step is the estimation of $d_t(v)$. In this example, material t is 93% tungsten. Estimates for the parameters β_0 and β_t will result from regressing penetration depth on velocity and the three indicator variables found in Equation 5. The values for the indicator variables m_{ij1}, m_{ij2}, and m_{ij3} are
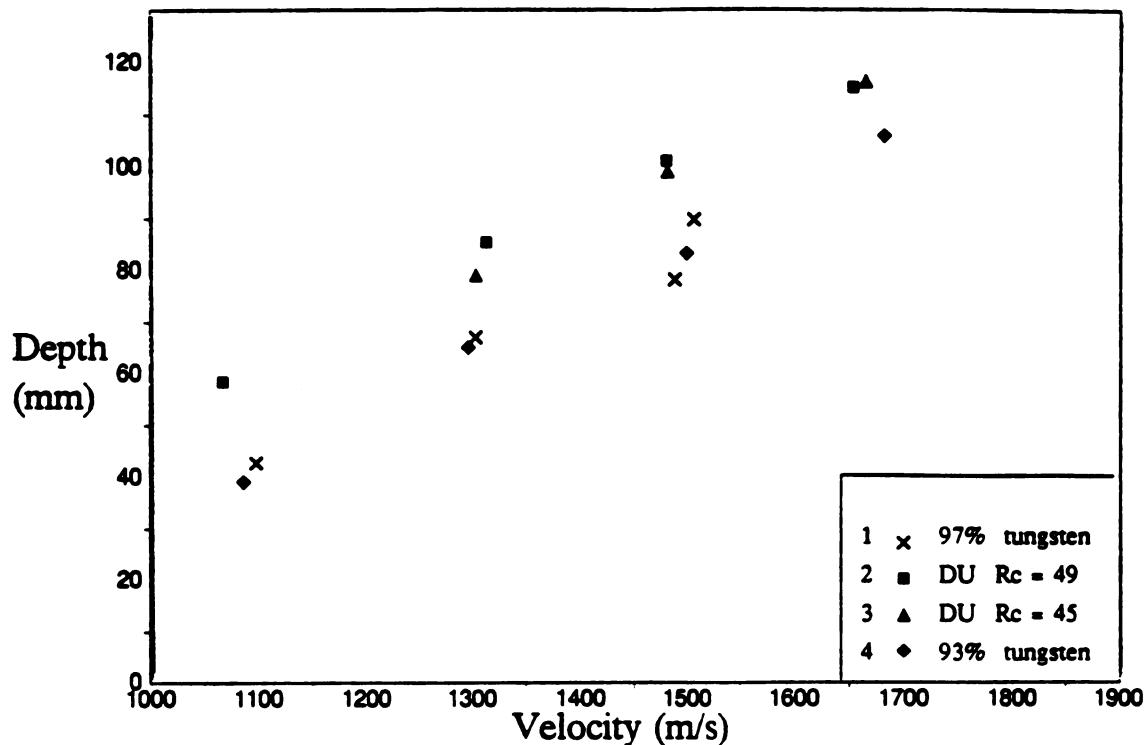
Figure 4. Depth of Penetration for Four Materials, L/D = 15.
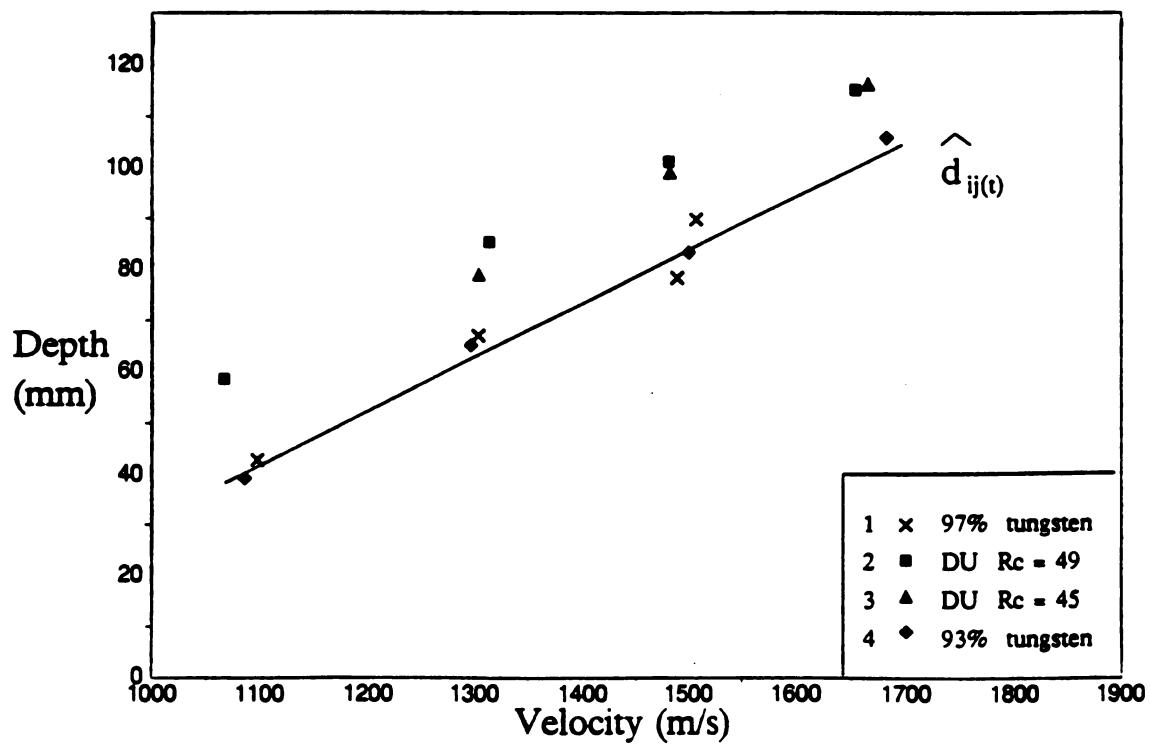
Figure 5. The Relationship of Penetration Depth to the Fit $\hat{d}_{ij(t)}$.

63

Table 1. Data Matrix for L/D = 15

	d_{ij} (mm)	v_{ij} (m/s)	m_{ij1}	m_{ij2}	m_{ij3}
97% tungsten	42.70	1098	1	0	0
	66.80	1304	1	0	0
	78.20	1489	1	0	0
	89.70	1507	1	0	0
DU Rc=49	58.42	1067	0	1	0
	85.34	1314	0	1	0
	101.09	1481	0	1	0
	115.06	1654	0	1	0
DU Rc=45	78.99	1304	0	0	1
	99.06	1482	0	0	1
	116.33	1660	0	0	1
93% tungsten	39.12	1086	0	0	0
	65.02	1297	0	0	0
	83.31	1500	0	0	0
	105.92	1682	0	0	0

Table 2. Residuals Relative to the tth Material for L/D = 15

	97% tungsten	DU Rc=49	DU Rc=45	93% tungsten
$r_{ij(t)}$	-0.29	18.65	14.68	-2.63
	2.49	20.00	16.32	1.43
	-5.26	18.46	15.16	-1.29
	4.38	14.52		2.49

shown in Table 1. It follows that β_1, β_2, and β_3 represent differences from 93% tungsten (our control) and 97% tungsten, DU Rc=49, and DU Rc=45, respectively. The estimated penetration depths for material t are given by

$$\hat{d}_{ij(t)} = 73.7310 + 0.1035(v_{ij} - 1395)$$

and are plotted as $\hat{d}_{ij(t)}$ in Figure 5. Next, the residuals relative to the tth group, $r_{ij(t)}$, are computed as $d_{ij} - \hat{d}_{ij(t)}$. In Figure 6 these residuals are plotted about the horizontal line, $r_{ij(t)} = 0$. Table 2 lists the residual values for each material.

To determine significance, the $r_{ij(t)}$ are permuted between the materials being compared. Consider, for example, the two DU materials. Their difference is estimated by $b_2 - b_3$ and takes on the value 2.514, the average of the residuals of DU Rc=49 less the average of the residuals of DU Rc=45. The reference distribution for determining significance is constructed by computing $b_2 - b_3$ for each possible combination of the residuals. Figure 7 depicts one such combination where four residuals were reassigned. In that instance $b_2 - b_3 = 1.535$. Figure 8 displays the reference distribution in the form of a stem-plot. The observed value for $b_2 - b_3$ is circled. There are six distribution values which equal or exceed in absolute value $| b_2 - b_3 |$ (denoted by bold type in Figure 8), hence a p-value of 6/35 or 0.171. Table 3 includes the results of each pairwise material comparison. In consideration of the data, all p-values appear reasonable and act to quantify the differences observed.

5.2 Example 2

A second data set is displayed in Figure 9. Three 65-g penetrators were tested, each with L/D=10. Unlike in the previous example, data were not collected strictly according to the template in Figure 2. They need not be for the randomization test to be valid. Also, the distinction between groups do not appear as great as in Example 5.1. It is in this situation that an explicit quantification of any differences is most needed because it becomes even less clear how much observed difference is real and how much is attributable to chance variations.

Table 4 lists the results for all pairwise comparisons between materials. The increased sample sizes over the previous example allows for a finer resolution in the number of reference distribution values. There are 12,870 values comprising the reference distribution for b_1, the estimated difference between 97% tungsten and 93% tungsten. The p-value for the randomization test is 0.192, meaning that the probability is 0.192 of observing a value for b_1 at least as unusual as 1.4050. Generally, such a p-value would not be considered significant, suggesting that 97% tungsten and 93% tungsten are performing similarly for L/D=10 penetrators.

A second contrast $\beta_1 - \beta_2$, signifying the difference between 97% tungsten and DU, is estimated to be -4.5113. [5] It is not clear from the examination of Figure 9 that this constitutes a real difference in performance. The randomization test, however, yields a p-value of 0.0040 and provides solid justification for the metallurgist's claim that 97% tungsten and DU materials are performing differently. Such a difference was observed by Magness [1990].

---

[5] No discussion in this report is devoted to controlling the error rate for multiple contrasts. For more explanation, see Kirk [1982].
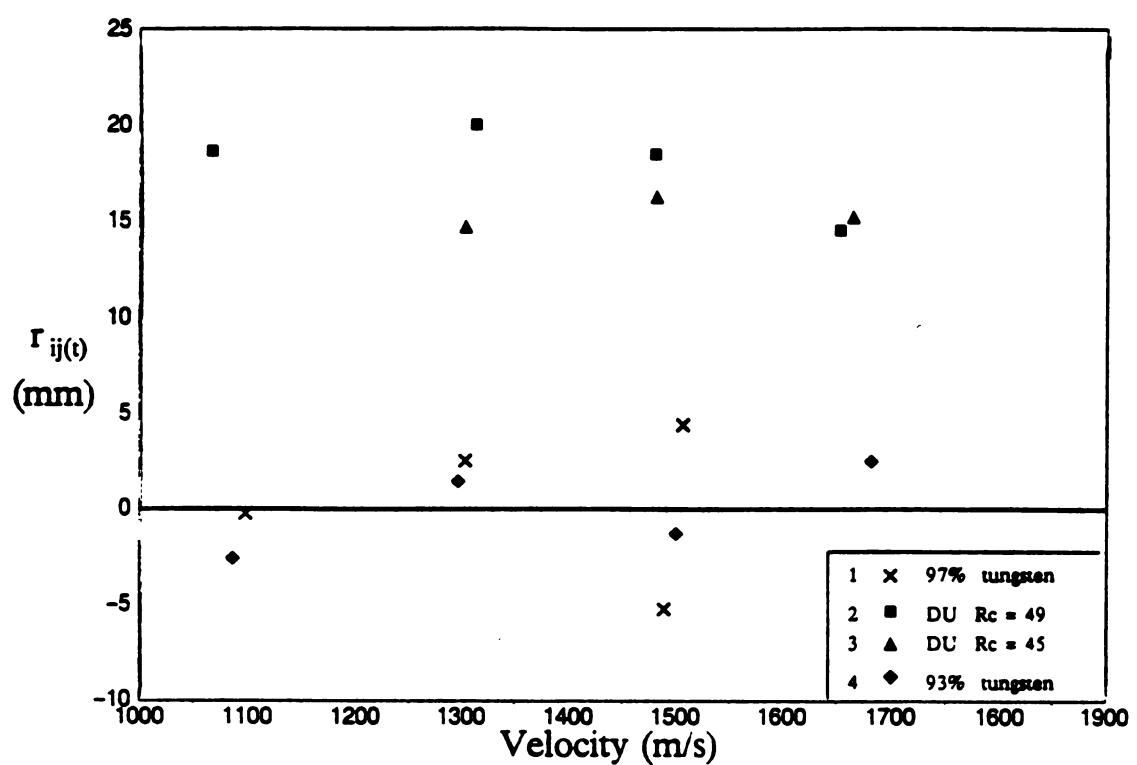
Figure 6. Residuals Relative to the tth Material.

$$r_{i2(t)} \qquad r_{i3(t)}$$

18.65   14.68

20.00   16.32

18.46   15.16

14.52
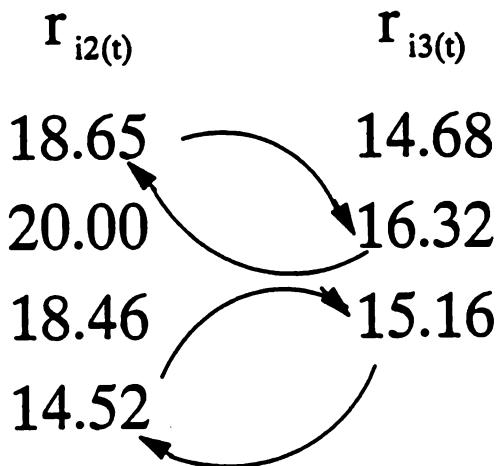
Figure 7. One Possible Reallocation of Residuals
Between Materials 2 and 3.

```
-3 | .9
-2 | .6 .5
-1 | .9 .8 .7 .7 .6 .6 .5 .0
-0 | .8 .7 .6 .3 .2
 0 | .2 .3 .4 .5 .5 .6 .6 .7 .7
 1 | .2 .3 .3 .4 .5 .6
 2 | (.5) .6 .9
 3 | .6
```
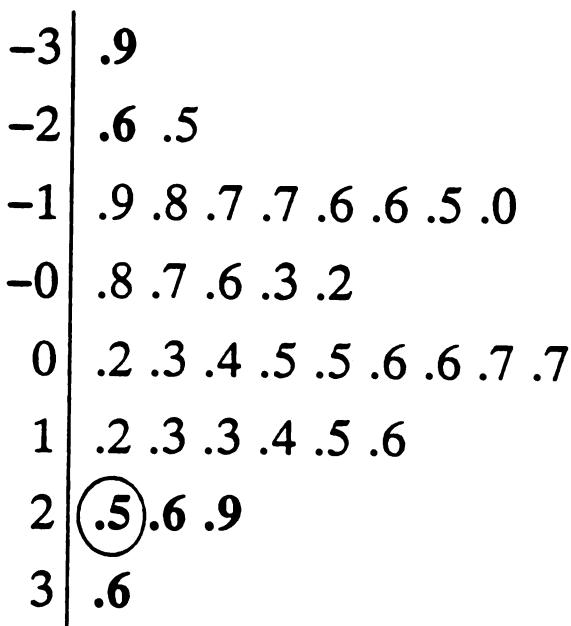
Figure 8. Stem-plot Representation of the Distribution
of the Test Statistic, b_2-b_3.

Table 3. Significance of the Differences Observed in Example 5.1

		L/D = 15		
			Randomization	
Contrast	Estimate	# unusual	# permutations	p-value
β_1	0.3298	60	70	0.857
β_2	17.9032	2	70	0.029
β_3	15.3890	1	35	0.029
β_1-β_2	-17.5734	2	70	0.029
β_2-β_3	2.5142	6	35	0.171
β_1-β_3	-15.0592	1	35	0.029
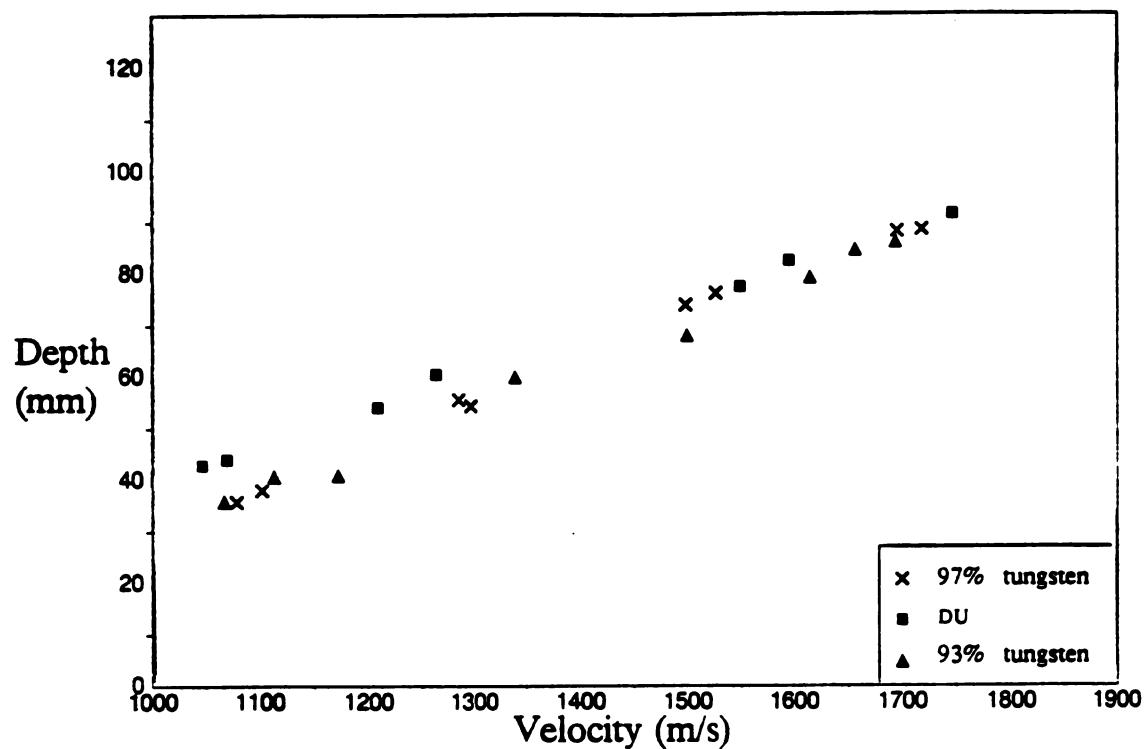
Figure 9. Depth of Penetration for Three Materials, L/D = 10.

Table 4. Significance of the Differences Observed in Example 5.2

		L/D = 10		
			Randomization	
Contrast	Estimate	# unusual	# permutations	p-value
β_1	1.4050	2472	12870	0.1921
β_2	5.91630	3	6435	0.0005
$\beta_1 - \beta_2$	-4.5113	26	6435	0.0040

6. Conclusion

For the testing of 1/4-scale kinetic energy penetrators against semi-infinite steel blocks, the technical considerations and the procedures addressing them are long established. It is the intent of this effort to enhance the inferential process within the presiding experimental structure. Presently, once data are collected, inferences principally consist of an engineering judgment as to the meaning of an observed vertical gap between linear functions representing the penetration performance of two materials. The initial motivation for pursuing this problem was the engineer's lament that, occasionally, when his judgment was questioned, he had little recourse but to stand firm on his opinion forged from years of experience. The linear functions themselves are usually established subjectively and are considered parallel over the range of 1100 m/s to 1700 m/s. Such subjectivity does bring into question the consistency of the assessment process. An objective method for fit, such as least squares, is seldom used, and then not in such a way as to incorporate the common slopes assumption. Nor need it be in all instances. Often, the differences are so great as to allow for the approximate fitting of the linear functions with no loss to the outcome, but perhaps equally as often they are not great, occurring when only marginal improvements are made over an historical (control) material.

In summary, the report identifies the experimental situation as being similar to that in which an analysis of covariance model is usually employed and then expresses the linear model in a manner conforming to how practitioners currently view the problem, even to the extent of automatically incorporating the parallel lines assumption. The report then explores some important problems, such as data arising from independent studies, in implementation of the classical method for significance testing and recommends an alternative to surmount these problems in the form of a randomization test. This test is implemented on two sets of real data, and its application in the context of those data is demonstrated.

The approach presented is an attempt at a unifying structure within which inferences in this environment can be made both quantifiable and consistent. The recommended procedure combines existing techniques such as least squares with a new application of a randomization test in determining the significance of observed material differences. With this test supporting, practitioners can make definitive statements as to the statistical significance of material differences observed.

7. References

Conover, W., and Iman, R. "Analysis of Covariance Using the Rank Transformation." *Biometrics*. 38 (1982) 715-724.

Edgington, E. Randomization Tests (2nd ed.). New York: Marcel Dekker, Inc., 1987.

Farrand, T. Unpublished manuscript, Ballistic Research Laboratory.

Fisher, R. Statistical Methods for Research Workers (4th ed.). London: Oliver and Boyd, 1932.

Hedges, L., and Olkin, I. Statistical Methods for Meta-Analysis. San Diego: Academic Press, Inc., 1985.

Kirk, R. Experimental Design (2nd ed.). Belmont, California: Brooks/Cole, 1982.

Magness, L. Proceedings of the 1990 Army Science Conference, 2 (1990), 465-479.

Neter, J., and Wasserman, W. Applied Linear Statistical Models. Richard D. Irwin, Inc., 1974.

Puri, M., and Sen, P. "Analysis of Covariance Based on General Rank Scores." *The Annals of Mathematical Statistics*, 40 (1969) 610-618.

Quade, D. "Rank Analysis of Covariance." *Journal of the American Statistical Association*, 62 (1967) 1187-1200.

Shirley, E. "A Distribution-free Method for Analysis of Covariance Based on Ranked Data." *Applied Statistics*. 30 (1981) 158-162.

Stephenson, W., and Jacobson, D. "A Comparison of Nonparametric Analysis of Covariance Techniques." *Communications in Statistics - Simulation*. 17(2) (1988) 451-461.

Taylor, M., and Bodt, B. "Randomization Procedures Applied to Analysis of Ballistic Data." BRL-TR-3245, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1990.

Tippet, L. The Method of Statistics. London: Williams & Norgate, 1931.

One-Sided Nonparametric Tolerance Limits Based on Two Order Statistics

Mark G. Vangel
U.S. Army Materials Technology Laboratory, SLCMT-MRS,
Watertown, MA 02172-0001

February 14, 1992

Abstract

A method was proposed by Hanson and Koopmans (Annals of
Math. Stat., 35, 1964) for obtaining conservative one-sided tolerance
limits for large classes of distribution functions. Let the cdf of the
underlying population be denoted F. The Hanson-Koopmans result
provides upper tolerance limits if $-\log(1-F)$ is convex (IFR distribu-
tions) and lower tolerance limits if $-\log(F)$ is convex. Any two order
statistics can be used to obtain these limits. The method is particu-
larly useful for small samples for which the nonparametric tolerance
limits do not exist.

The Hanson-Koopmans procedure was originally implemented for
consecutive order statistics. This was apparently done for computa-
tional simplicity in determining the weights. Unfortunately, this choice
results in *extremely* conservative limits. In this paper, we evaluate the
performance of the Hanson-Koopmans method for various pairs of or-
der statistics and suggest combinations for which the conservatism is
greatly reduced. In addition, we suggest a substantial further reduction
in conservatism for distributions with positive support.

An important application of this method in the aircraft industry
is to determining 95% lower confidence limits on the first and tenth
population percentiles of material strength, and this application has
been the motivation for the present study.

1   Introduction

In structural design, an *allowable stress*, *working stress*, or *design al-
lowable* for a material is the maximum stress at which one can be rea-

sonably certain that failure will not occur. For the design of structures for which weight is not a primary consideration, allowables are typically calculated by dividing a stress level at which failure is known to often occur by a sufficiently large constant called a *safety factor* (e.g., Gere and Timoshenko, 1984, p.29). The structure is then designed so as to ensure that the sresses do not exceed the allowables for the materials.

This approach is too conservative for many aircraft applications, however. Since weight is an important consideration in aircraft design, this industry long ago established two one-sided tolerance limits to supplement the use of safety factors in determining allowables. These tolerance limits are a 95% lower confidence limit on the tenth percentile and a 95% lower confidence limit on the first percentile of the strength distribution of a material. These are referred to as 'B-basis' and 'A-basis' values, respectively (Mil Handbook 5E, 1987; Mil Handbook 17C, 1992).

In this article, we discuss methods for determining one-sided tolerance limits nonparametrically. The motivation for this study has been Mil Handbook 17, and therefore the lower tolerance limits corresponding to 'A-basis' and 'B-basis' values will be used when it is desirable to fix ideas. However, the methods to be discussed are applicable to *any* one-sided tolerance limits, and to all sample sizes greater than one.

2 Preliminaries

Let F be the absolutely continuous distribution function of a continuous random variable X, and let x_β be the βth quantile of F; that is, $F(x_\beta) = \beta$. Assume that we have an *iid* random sample $\{X_i\}_{i=1}^n$ from F, and let the order statistics of this sample be

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}. \tag{1}$$

We will also need to make use of the order statistics of standard uniform and standard exponential random samples of size n, and we define these random variables as

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)} \tag{2}$$

and

$$E_{(1)} \leq E_{(2)} \leq \cdots \leq E_{(n)}, \tag{3}$$

72

respectively. We will adopt the usual convention of denoting a realization of a random variable by the corresponding lower case letter.

A (β, γ) *lower tolerance limit* is a random variable $T_l = T_l(X_1, \ldots, X_n)$ such that

$$P(T_l \leq x_{1-\beta}) \geq \gamma. \tag{4}$$

A (β, γ) *upper tolerance limit* is a random variable $T_u = T_u(X_1, \ldots, X_n)$ such that

$$P(T_u \geq x_\beta) \geq \gamma. \tag{5}$$

It is easy to see that upper (lower) tolerance limits are precisely upper (lower) confidence limits on x_β or $x_{1-\beta}$, respectively.

3 Nonparametric Limits Based on One Order Statistic

Let $\{X_i\}_{i=1}^n$ be independent, identically distributed random varibles with continuous cdf $F(x)$. We will demonstrate how a sample order statistic can sometimes be used as a lower tolerance limit. A similar argument can be made for upper tolerance limits. The material in this section is well known; a recent reference is Conover (1980, 118-121).

The probability that the ith order statistic, $X_{(i)}$, is less than the $(1-\beta)$th quantile of F, $x_{1-\beta}$, is easily seen to be

$$P(X_{(i)} \leq x_{1-\beta}) = P(F(X_{(i)}) \leq F(x_{1-\beta})) = P(U_{(i)} \leq 1 - \beta), \tag{6}$$

where $U_{(i)}$ is the corresponding order statistic from a uniform sample on $[0, 1]$ of size n. Since

$$U_{(i)} \sim \text{Beta}(i, n - i + 1) \tag{7}$$

(e.g., Hogg and Craig, 1978, p. 159), we have that

$$P(X_{(i)} \leq x_{1-\beta}) = \text{Beta}(1 - \beta; i, n - i + 1), \tag{8}$$

where Beta $(t; \nu_1, \nu_2)$ is the Beta cumulative with parameters ν_1 and ν_2. Let i_0 denote the largest rank for which

$$P(U_{(i_0)} \leq 1 - \beta) \geq \gamma. \tag{9}$$

73

Then $X_{(i_0)}$ is a lower (β, γ) tolerance limit. Note that, for any β and n, if

$$P(U_{(1)} \leq 1 - \beta) \leq \gamma_*, \tag{10}$$

then a (β, γ) lower tolerance limit based on a single order statistic *does not exist* for $\gamma = \gamma_*$.

3.1 Limitations of Single Order Statistic Tolerance Limits

This highlights the first, and most serious, limitation of nonparametric single order statistic limits. For a given tolerance limit (β, γ) and sample size n, these limits need not exist.

A second limitation is a consequence of the fact that, for given β and n, there are only n values of γ for which tolerance limits with *exact* confidence are available, namely $\{\gamma_i\}_{i=1}^{n}$, where, for lower limits,

$$\gamma_i = \gamma_i(\beta, n) = \text{Beta}(1 - \beta, i, n - i + 1). \tag{11}$$

Usually, the desired γ is not one of these n values, and one uses instead the largest i for which $\gamma_i \geq \gamma$ (or, for upper limits, the smallest i for which $\gamma_i \geq \gamma$). This can result in tolerance limits that are exremely conservative, as the following example illustrates.

Consider the problem of determining a lower tolerance limit when $n = 44$, $\beta = .9$ and $\gamma = .95$. Since

$$P(U_{(2)} \leq .1) = \text{Beta}(.1; 2, 43) \doteq .943 < .95 \tag{12}$$

and

$$P(U_{(1)} \leq .1) = \text{Beta}(.1; 1, 44) \doteq .990 > .95, \tag{13}$$

$X_{(1)}$ provides the desired tolerance limit. However, the actual confidence of this tolerance limit is .99 – substantially greater than $\gamma = .95$.

A natural way to circumvent both of these limitations is to consider tolerance limits which *interpolate* between two order statistics, or, if necessary, *extrapolate* beyond $X_{(1)}$ or $X_{(n)}$. It is necessary then to calculate probabilities such as $F(cX_{(i)} + (1 - c)X_{(j)})$, where $c > 0$. We need to make additional assumptions on F in order to relate $F(cX_i + (1 - c)X_j)$ to $F(X_{(i)}) = U_{(i)}$ and $F(X_{(j)}) = U_{(j)}$. In particular, if we assume that $-\log(F)$ is convex, then we can determine lower limits based on two order statistics, and if we assume that $-\log(1 - F)$ is convex, then we can find upper limits. This is the approach taken by Hanson and Koopmans (1964).

4 The Hanson-Koopmans Theorem

Hanson and Koopmans (1964) showed that, given certain pairs of order statistics from an *iid* sample of any size, it is possible to find a linear combination of these two random variables which is (either an upper or a lower) tolerance limit for any quantile β and any confidence γ.

The lower limits require the assumption that $-\log(F)$ be a convex function, and the upper limits require the convexity of $-\log(1-F)$. Both of these classes are large enough for one to legitimately regard these tolerance limits as nonparametric. The condition that $-\log(1-F)$ is convex is equivalent to an increasing hazard rate assumption. Hanson and Koopmans show that the *intersection* of the class of distributions for which $-\log(F)$ is convex with the class of distributions for which $-\log(1-F)$ is convex includes the *Polya Type II* distributions, which includes many of the distributions which are often used in practice.

We will derive the Hanson-Koopmans lower limits first, and then we will use a very similar argument to obtain upper limits.

4.1 Lower Tolerance Limits

Assume $n \geq 2$ and let $1 \leq i < j \leq n$. We will consider lower tolerance limits of the form

$$T_l = X_{(j)} + k_l(X_{(i)} - X_{(j)}), \qquad (14)$$

where $k_l \geq 1$. We make the assumption that $-\log(F)$ is convex. Note that, if $k_l > 1$, then Jensen's inequality does *not* hold, and we have, for any quantile $x_{1-\beta}$,

$$P(T_l \leq x_{1-\beta}) \qquad (15)$$
$$= P[F(T_l) \leq F(x_{1-\beta})]$$
$$= P[F(T_l) \leq 1 - \beta]$$
$$= P\{-\log[F(T_l)] \geq -\log(1-\beta)\}$$
$$\geq P\{-\log[F(X_{(j)})] + k_l \log[F(X_{(j)})] - k_l \log[F(X_{(i)})] \geq -\log(1-\beta)]$$
$$= P[\log(U_{(j)}) - k_l \log(U_{(j)}) + k_l \log(U_{(i)}) \leq \log(1-\beta)]$$
$$= P\left[U_{(j)} \left(\frac{U_{(i)}}{U_{(j)}} \right)^{k_l} \leq 1 - \beta \right].$$

75

The joint density of $U_{(i)}$ and $U_{(j)}$ is (Hogg and Craig, 1978, p. 160)

$$F_{U_{(i)},U_{(j)}}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} x^{i-1}(y-x)^{j-i-1}(1-y)^{n-j}.$$

$$(16)$$

For fixed $k_l > 1$, we can integrate this density over the region

$$\mathcal{D}(k_l) \equiv \{(x,y)|x \le y \text{ and } y(x/y)^{k_l} \le 1-\beta\}, \qquad (17)$$

and thereby express (15) as the following monotone function of k_l:

$$H(k_l) \equiv \int_{\mathcal{D}(k_l)} F_{U_{(i)},U_{(j)}}(x,y)dxdy. \qquad (18)$$

As $k_l \downarrow 1$, $T_l \to X_{(i)}$ and $H(k_l) \to \text{Beta}(1-\beta; i, n-i+1) = \gamma_i(\beta,n)$. As $k_l \to \infty$, $y(x/y)^{k_l} \to 0$ for all $x < y \in (0,1]$, so $H(k_l) \to 1$. So, for any β, i, j, and n, and for any $\gamma \ge \gamma_i(\beta,n)$, there exists a k_l^* such that $H(k_l^*) = \gamma$, and hence

$$T_l = X_{(j)} + k_l^*(X_{(i)} - X_{(j)}) \qquad (19)$$

is a lower (β,γ) tolerance limit.

4.2 Upper Tolerance Limits

We now sketch the corresponding argument for upper tolerance limits. The upper limit, based on $X_{(i)}$ and $X_{(j)}$ for $i < j$, is of the form

$$T_u = X_{(i)} + k_u(X_{(j)} - X_{(i)}). \qquad (20)$$

We assume that $-\log(1-F)$ is a convex function, and we would like to derermine k_u so that, for given β and γ, T_u provides a (β,γ) upper tolerance limit. We see that

$$P(T_u \ge x_\beta) \qquad (21)$$

$$= P[F(T_u) \ge F(x_\beta)]$$

$$= P[1 - F(T_u) \le 1-\beta]$$

$$= P\{-\log[1-F(T_u)] \ge -\log(1-\beta)]\}$$

$$\ge P\{-\log[1-F(X_{(i)})] + k_u\log[1-F(X_{(i)})] - k_u\log[1-F(X_{(j)})] \ge -\log(1-\beta)]\}$$

$$= P[-\log(U_{(n-i+1)}) - k_u\log(U_{(n-j+1)}) + k_u\log(U_{(n-i+1)}) \ge -\log(1-\beta)]$$

$$= P\left[U_{(n-i+1)}\left(\frac{U_{(n-j+1)}}{U_{(n-i+1)}}\right)^{k_u} \le 1-\beta\right].$$

The probability (21) is the following monotone function of k_u :

$$V(k_u) \equiv \int_{\mathcal{D}(k_u)} F_{U_{(n-i+1)}, U_{(n-j+1)}}(x, y) dx dy, \qquad (22)$$

where $F_{U_{(i)}, U_{(j)}}(x, y)$, the joint density of two uniform order statistics, is given by (16), and $\mathcal{D}(k)$ is defined in (17).

5   Which Order Statistics are Best?

The discussion of the previous section follows from results in Hanson and Koopmans (1964). However, the Hanson-Koopmans results are not widely known, and so it has been necessary to derive them again in this article.

It seems likely that the reason why the very useful work of Hanson and Koopmans has been virtually ignored by applied statisticians follows from their unfortunate choice of the ranks i and j. Probably because the numerical double integrals required in order to evaluate (18) represented substantial computation for the time, Hanson and Koopmans decided to only consider consecutive order statistics, that is they let $j = i + 1$. We then have that

$$F_{U_{(i)}, U_{(i+1)}}(x, y) = \frac{n!}{(i-1)!(n-i-1)!} x^{i-1}(1-y)^{n-i-1}, \qquad (23)$$

and the necessary calculations can be done with the aid of tables of the gamma and incomplete beta functions, and without the need for numerical integration. The tolerance limits which result from the use of consecutive order statistics, as we shall see, are usually conservative to the point of being useless for most applications.

Woodward and Frawley (1980) is apparently the only article in the applied literature which builds on the work of Hanson and Koopmans. These authors wanted to use the Hanson-Koopmans limits for data which had ties, so that the required consecutive order statistics were not always available. They noted that considering the *range*, that is, letting $i = 1$ and $j = n$, also leads to a single integral in (18). They computed tolerance limit factors for the range, and provided tables. It turns out that the Wordward and Frawley limits provide a *substantial*

77

improvement over the use of consecutive order statistics, although these authors did not comment on this point.

In fact, there is reason to believe that, if one uses closeness of the expected value of the tolerance limit to the relevant quantile as a criterion, then consecutive order statistics constitute the *worst possible* choice of ranks, at least for the important cases of (.90,.95) and (.99,.95) lower tolerance limits. A more careful selection of which i and j to use for given β, γ, and n leads to a dramatic improvement over consecutive order statistics, as we will show in the next section.

6   Exponential Spacings and the Calculation of $H(k_l)$ and $V(k_u)$

The computational burden due to the integral (18) is completely unnecessary, since the functions H and V can be obtained in closed form. The reason for this is that, if $X_i \sim F$ for any continuous F, then it is well known, and easy to show, that $-\log(X_i)$ has the standard exponential distribution.. Consequently, $-\log(U_{(l)})$ has the same distribution as $E_{(n-l+1)}$ and $-\log(1 - U_{(l)})$ has the same distribution as $E_{(l)}$, for any l, where $1 \le l \le n$, and the $E_{(l)}$ are order statistics from a standard exponential sample of size n.

Define the *spacings* in a standard exponential sample as follows:

$$D_s \equiv E_{(s)} - E_{(s-1)}, \tag{24}$$

where $E_{(0)} \equiv 0$. Then, by Theorem 2.5 of Barlow and Proschan (1981, p. 59), we have that:

1. The $\{D_s\}_{s=1}^n$ are mutually independent, and

2. $P(D_s \le t) = 1 - e^{-(n-s+1)t}$.

Let $\{E_s\}_{s=1}^n$ be an *iid* sample from a standard exponential distribution. For $1 \le l \le n$, we can write $E_{(l)}$ as a sum of the spacings, and therefore as a linear combination of $\{E_s\}_{s=1}^l$:

$$E_{(l)} = \sum_{s=1}^l D_s = \sum_{s=1}^l \frac{E_s}{n - s + 1}. \tag{25}$$

78

Digitized by Google

Let

$$
\begin{aligned}
\tilde{T}_l &\equiv (1 - k_l)[-\log(F(X_j)] + k_l[-\log(F(X_i)] \qquad (26)\\
&= (1 - k_l)E_{(n-j+1)} + k_l E_{(n-i+1)}\\
&= \sum_{s=1}^{n-j+1} \frac{E_s}{n-s+1} - k_l \sum_{s=1}^{n-j+1} \frac{E_s}{n-s+1} + k_l \sum_{s=1}^{n-i+1} \frac{E_s}{n-s+1}\\
&= \sum_{s=1}^{n-j+1} \frac{E_s}{n-s+1} + k_l \sum_{s=n-j+2}^{n-i+1} \frac{E_s}{n-s+1}
\end{aligned}
$$

so that

$$
H(k_l) = P[\tilde{T}_l \geq -\log(1 - \beta)]. \qquad (27)
$$

Similarly define

$$
\begin{aligned}
\tilde{T}_u &\equiv (1 - k_u)[-\log(1 - F(X_{(i)})] + k_u[-\log(1 - F(X_{(j)}))] \quad (28)\\
&= (1 - k_u)E_{(i)} + k_u E_{(j)}\\
&= \sum_{s=1}^{i} \frac{E_s}{n-s+1} - k_u \sum_{s=1}^{i} \frac{E_s}{n-s+1} + k_u \sum_{s=1}^{j} \frac{E_s}{n-s+1}\\
&= \sum_{s=1}^{i} \frac{E_s}{n-s+1} + k_u \sum_{s=i+1}^{j} \frac{E_s}{n-s+1} \qquad (29)
\end{aligned}
$$

so that

$$
V(k_u) = P[\tilde{T}_u \geq -\log(1 - \beta)]. \qquad (30)
$$

If

$$
Y \equiv \sum_{i=1}^{l} \lambda_i E_i \qquad (31)
$$

where the $\{E_i\}_{i=1}^{l}$ are *iid* standard exponential random variables, if the λ_i are distinct, and if $\lambda_i > 0$ for all i, then the cdf of Y is (Johnson and Kotz, 1970, p. 222)

$$
G_Y(y) = \sum_{i=1}^{l} \left[\prod_{\substack{j=1\\j \neq i}}^{l} (\lambda_i - \lambda_j)^{-1} \right] \lambda_i^{l-1}(1 - e^{-y/\lambda_i}). \qquad (32)
$$

By substituting the coefficients in the linear combinations of the $\{E_s\}_{s=1}^{n}$ given by (26) and (28) into (32), we obtain closed form expressions for the probabilities $H(k_l)$ and $V(k_u)$, respectively.

79

7 An Example: Determining a Good k_l for a (.99,.95) Lower Tolerance Limit With $n = 30$.

The first order statistic, $X_{(1)}$, provides a nonparametric (.99,.95) lower tolerance limit for $n \geq 299$. We will investigate the tolerance limit T_l given by (14) when $n = 30$. We introduce the more elaborate notation $k_l^{i,j}$ and $T_l^{i,j}$ in order to emphasize the dependence on the ranks of the data values used in the tolerance limits. Corresponding to each $i < j \leq 30$ there is a $k_l^{i,j}$ which provides the desired lower tolerance limit. Since $V(k_l)$ is available in closed form, it is straightforward to find these $k_l^{i,j}$ values numerically. We will use as a criterion for choosing i and j the expected value of $T_l^{i,j}$ under a normal model, the objective being to choose the tolerance limit for which $x_{1-\beta} - E(T_l)$ is as small as possible.

In Figure 1, $-\log |E(T_l)|$ is plotted for each $i < j \leq 30$. Since the tolerance limit problem for the normal distribution is invariant to a change in location and/or scale, we can, without loss of generality, consider a N(0,1) population. For this population we have $x_{1-\beta} = x_{.01} = -2.327$. We can see from this plot that, according to our criterion, the best tolerance limits have $i = 1$. In Figure 2, $-E(T_l^{1,j})$ is displayed for $1 < j \leq 30$. The Hanson-Koopman's limit, with $j = i + 1 = 2$, has expected value -15.11. The optimal limit $T_l^{1,12}$ has expected value -4.743 – a gain of *over ten standard deviations* in the expectation. The expected value of the parametric normal tolerance limit (e.g., Owen, 1968) is equal to -3.038. The Woodward-Frawley limit $T^{1,30}$ is somewhat more conservative than the optimal limit, but it also is much better than $T^{1,2}$.

We now make a case for using nonparametric over parametric tolerance limits for extreme quantiles when the sample size is small. We consider a Weibull population with a shape parameter of 10 and a scale parameter of 1, that is

$$F_X(x) = 1 - e^{-x^{10}}. \tag{33}$$

From Figure 3, we see that this Weibull population looks 'roughly normal', but, of course, any such resemblance breaks down in the tails. Let $n = 30$, $\beta = .99$, and $\gamma = .95$. We will compare the nonparametric lower tolerance limit $T_l^{1,12}$ with the parametric normal limit $\bar{x} - 3.063s$,

where $\bar{x}$ and s are the sample mean and standard deviation. The consequences of using a normal tolerance limit procedure when the population is (33) is examined by taking 500 random samples of size 30 from (33) and calculating $\bar{x} - 3.063s$ and $T^{1,12}$ for each. In Figure 4, a histogram of the normal tolerance limit values is displayed. Note the exreme anticonservatism – the actual confidence of about 68% is much less than the nominal confidence of 95%. However, in Figure 5, the tolerance limit $T^{i,j}$ has actual confidence of 98%, which is quite respectable, considering that we have only *one tenth* the data that would be required for the usual single order statistic nonparametric limit.

8   Log Transformation for Data With Positive Support

If the random variables $\{X_i\}_{i=1}^n$ have positive support, then a positive tolerance limit is desirable. By taking logarithms of the data, calculating T_l or T_u, and exponentiating the result, we obtain the lower and upper limits

$$L_l = x_j(x_i/x_j)^{k_l} \tag{34}$$

and

$$L_u = x_i(x_j/x_i)^{k_u}, \tag{35}$$

respectively. Although these tolerance limits are valid for classes of distributions more narrow than the classes corresponding to T_l and T_u, this is not likely to be a problem in practice, and the limits (34) and (35) can be substantially closer to the quantile in expectation than (14) and (20). Still, the transformation remains ad-hoc, pending further investigation. In Figure 6, (34) was used instead of (14) for the 500 simulated datasets, with some reduction in conservatism of the tolerance limit.

9   Two Order Statistic Tolerance Limits for All Sample Sizes

On the basis of extensive computation for the cases of (.90, .95) and (.99, .95) lower tolerance limits, we conjecture that the 'best' two order

statistic lower tolerance limit for any sample size $n > 1$, any β, and any γ is $T_l^{i_0,j}$, where $i_0 - 1$ is the rank corresponding to the single order statistic limit, where we adopt the convention of associating the rank 0 to situations where a single order statistic limit does not exist. Different criteria for what is best will result in different choices of j, but most reasonable criteria will probably result in a j for which $j - i_0$ is fairly large, and the precise j selected is of secondary importance.

To fix ideas, we adopt here the expected value under a normal model as a criterion, as discussed in previous sections. For this criterion, if we fix β and γ, choose $i = i_0(n, \beta, \gamma)$, and find the optimal j, then we can regard k_l to be a function of n alone. If $\tilde{n}$ is the smallest value of n for which a certain order statistic provides a single order statistic nonparametric limit, then $k_l(\tilde{n})$ will equal 1.

The values $k_l(n)$ of the previous paragraph provide two order statistic tolerance limits which reduce to single order statistic limits at certain sample sizes, and which are less conservative then the single order statistic limits in between these sample sizes. We illustrate this idea in Figure 7, for the case $\beta = .90$ and $\gamma = .95$. For this case, the first order statistic provides a nonparametric lower limit for $n = 29$ and the second order statistic provides a limit when $n = 46$. In between, the first order statistic provides a tolerance limit which increases in conservatism as n increases from 30 to 45. The 'optimal' two order statistic limit, however, equals the first order statistic when $n = 29$, equals the second order statistic when $n = 46$, and steadily increases in expected value at intermediate sample sizes. The expectation of a lower tolerance limit should increase monotonically with n. This is an appealing characteristic of the two order statistic limit, of this and the previous section, which is *not* shared by the usual nonparametric limit.

A similar argument can probably be made for upper limits and for other percentiles and confidences than (.90, .95) and (.99, 95), and this will be a subject of future work

10   Tables

A short table for lower (.90, .95) tolerance limits is provided in Appendix A, and a FORTRAN subroutine for determining $H(k_l)$ is given in Appendix B. Additional tables can easily be created. However, if

one wishes to use two order statistic tolerance limit factors beyond the cases covered by this table, and if one does not want to compute the $k_l^{i,j}$, the tables in Woodward and Frawley (1980) can be used.

11 Conclusion

The main point of this article is that the commonly held notion that one-sided nonparametric tolerance limits do not exist for certain β, γ, and n is misleading. We have shown that *useful* one-sided tolerance limits involving two order statistics can be obtained for any situation, and that these limits are valid over large classes of distributions.

In this article, we also discuss the usefulness of a log transformation, the choice of which order statistics to use, and the idea of using the proposed limits for *all* sample sizes – not merely for those cases for which the usual nonparametric method is not available. These discussions are somewhat ad-hoc, and considerable work remains to be done. But the potential usefulness of the proposed limits is clear, and one should not hesitate to make use of them even before all of the theoretical details are worked out.

Figure 1

Hanson-Koopmans Beta = .99, Gamma = .95, n = 30

Expected Values for N(0,1) Population
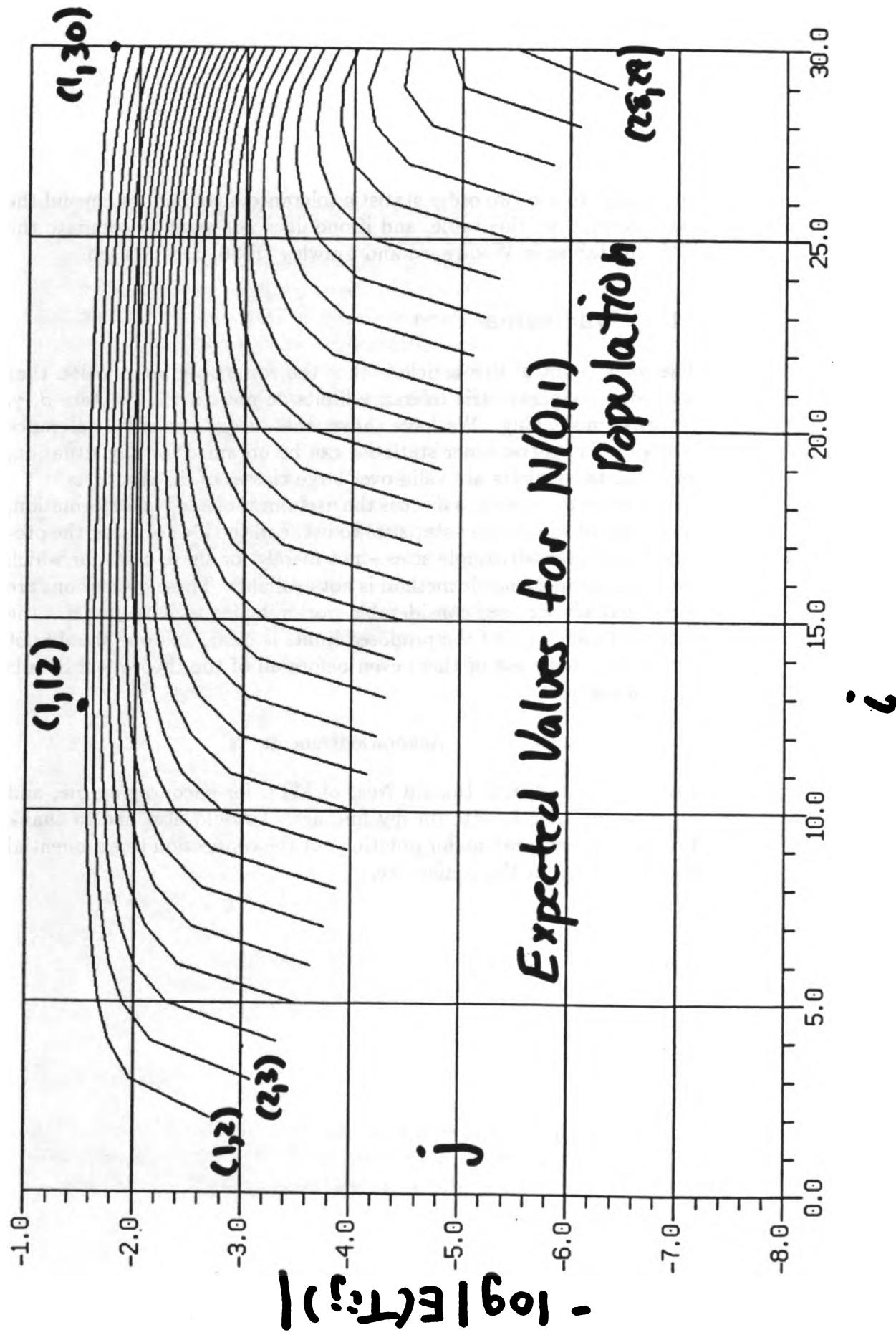
Figure 2

Optimizing the method with respcet to normal(0,1)
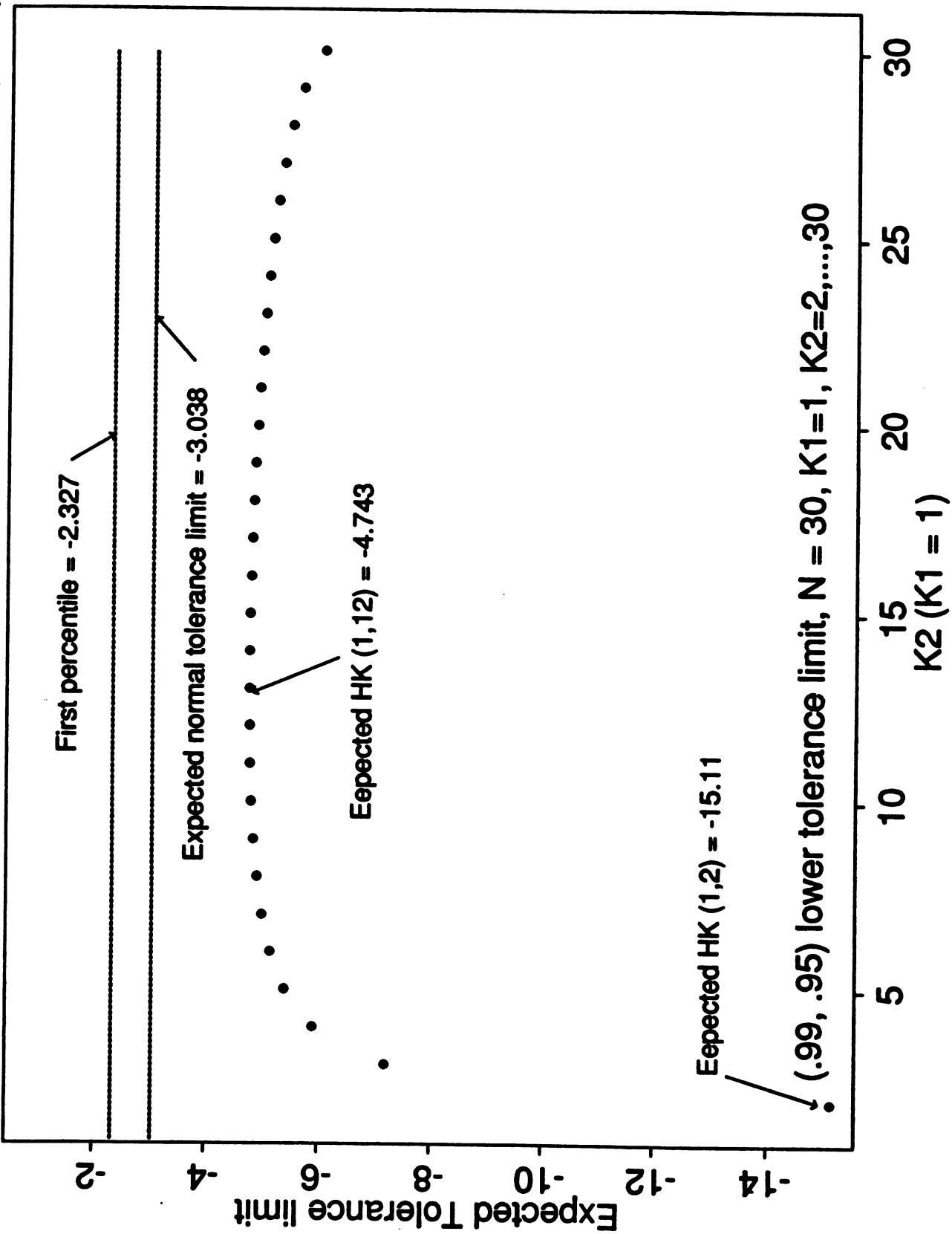
Figure 3

Comparison of normal and Weibull densities

Weibull (shape=10, scale=1)

Normal (mean =.951, SD =.115)

1st %ile (normal)

x

Probability

Figure 4

Normal (.99, .95) tolerance limits

Population 1st %ile

Weibull population, shape = 10

n = 30; number replicates = 500

p = .99, g = .95

Nominal confidence = .950

Actual confidence = .676

Tolerance limit using normal method

87
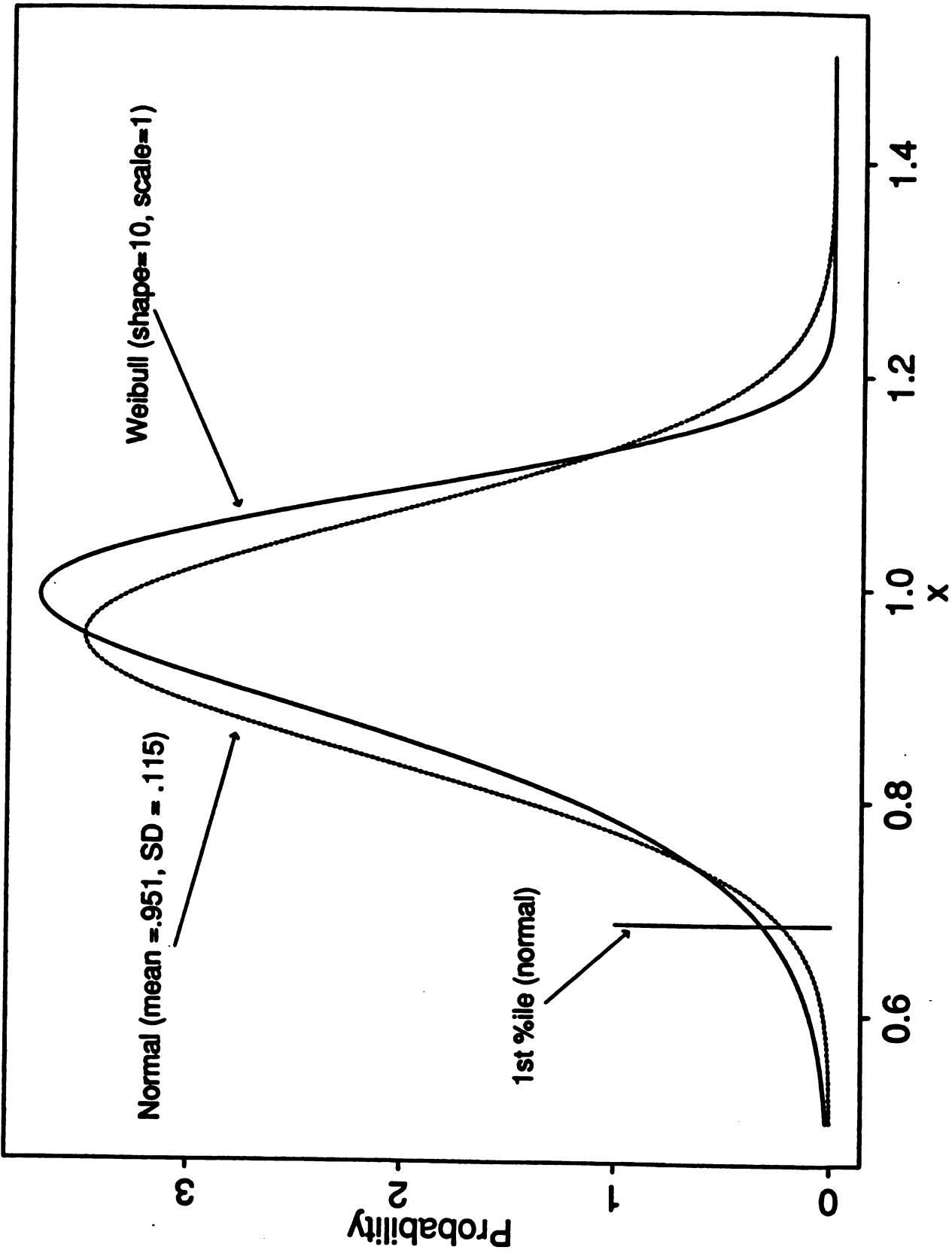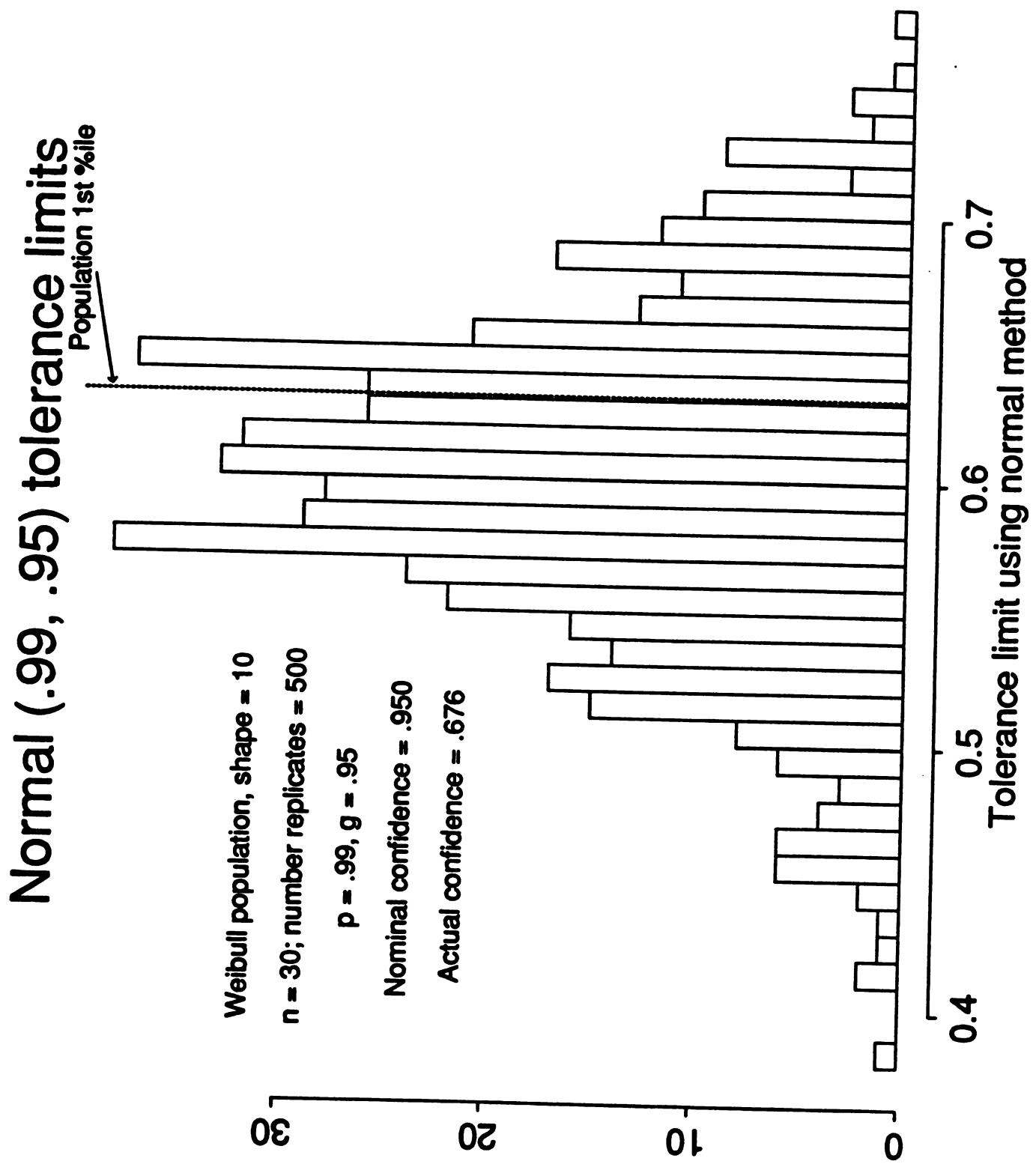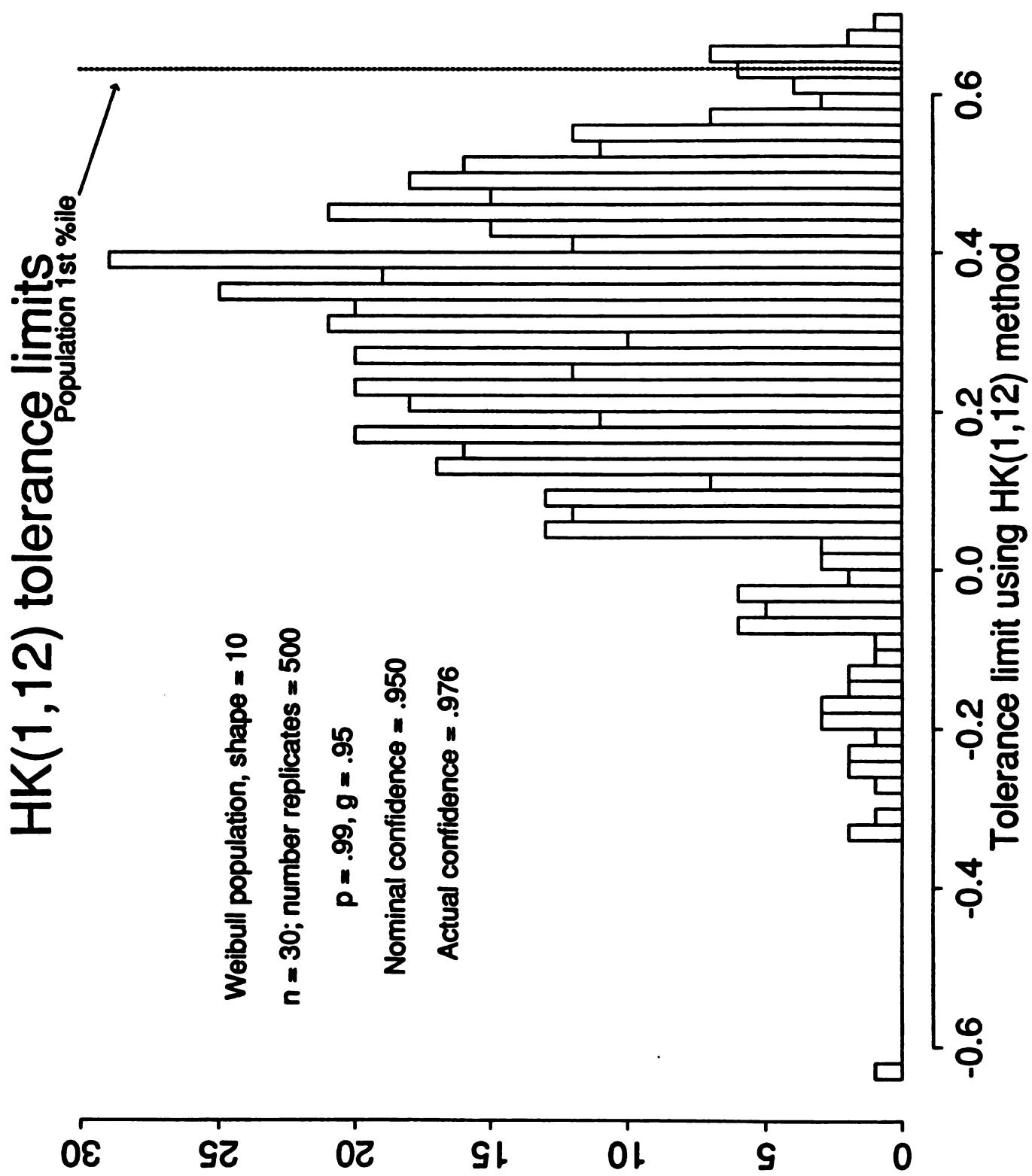
Figure 5

HK(1,12) tolerance limits
Population 1st %ile

Weibull population, shape = 10

n = 30; number replicates = 500

p = .99, g = .95

Nominal confidence = .950

Actual confidence = .976

Tolerance limit using HK(1,12) method

Figure 6

HK(1,12) tolerance limits (log trans.)

Weibull population, shape = 10

n = 30; number replicates = 500

p = .99, g = .95

p = .99, g = .95

Nominal confidence = .950

Actual confidence = .958

Population 1st %ile

Tolerance limit using HK(1,12) method (log trans.)
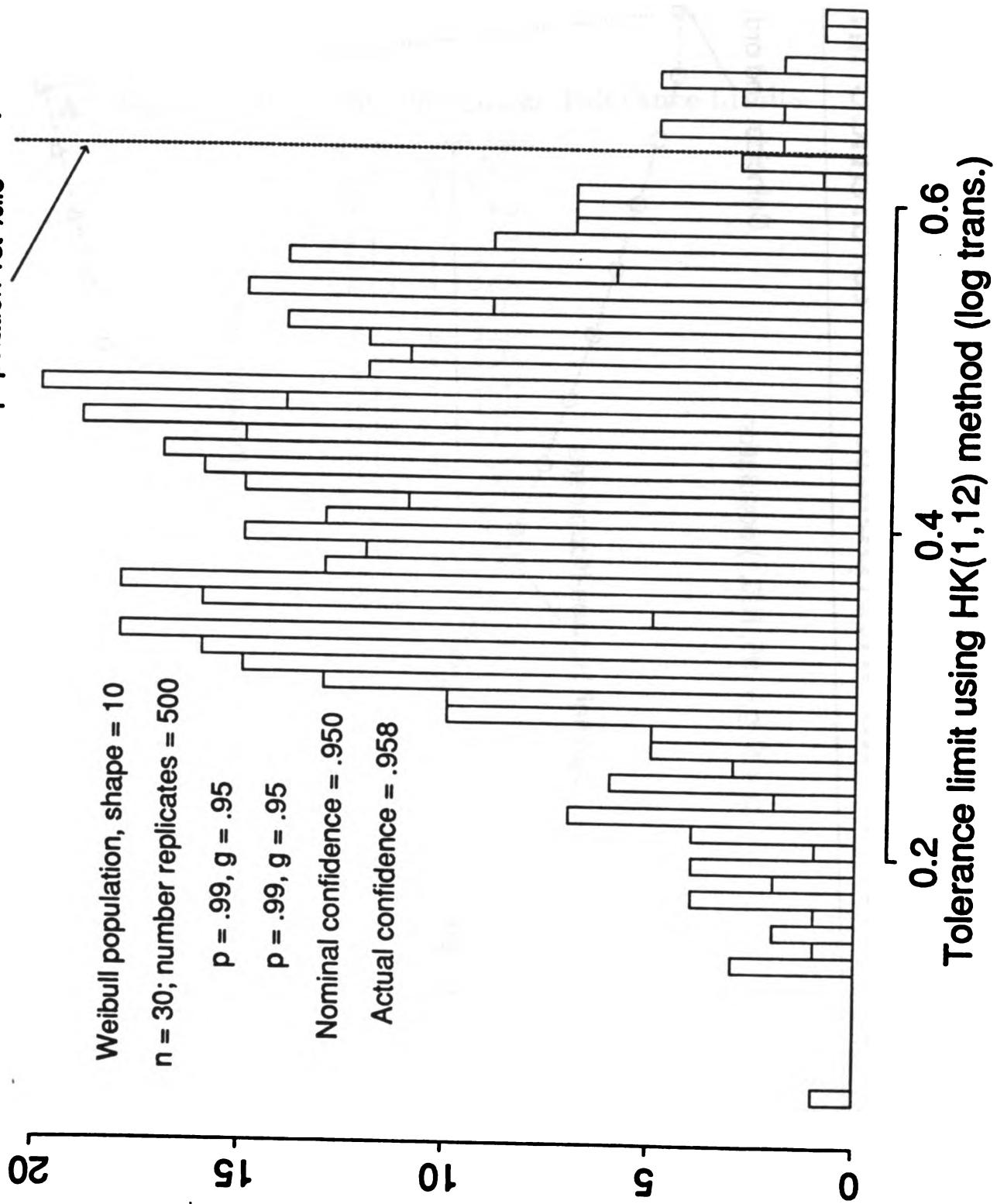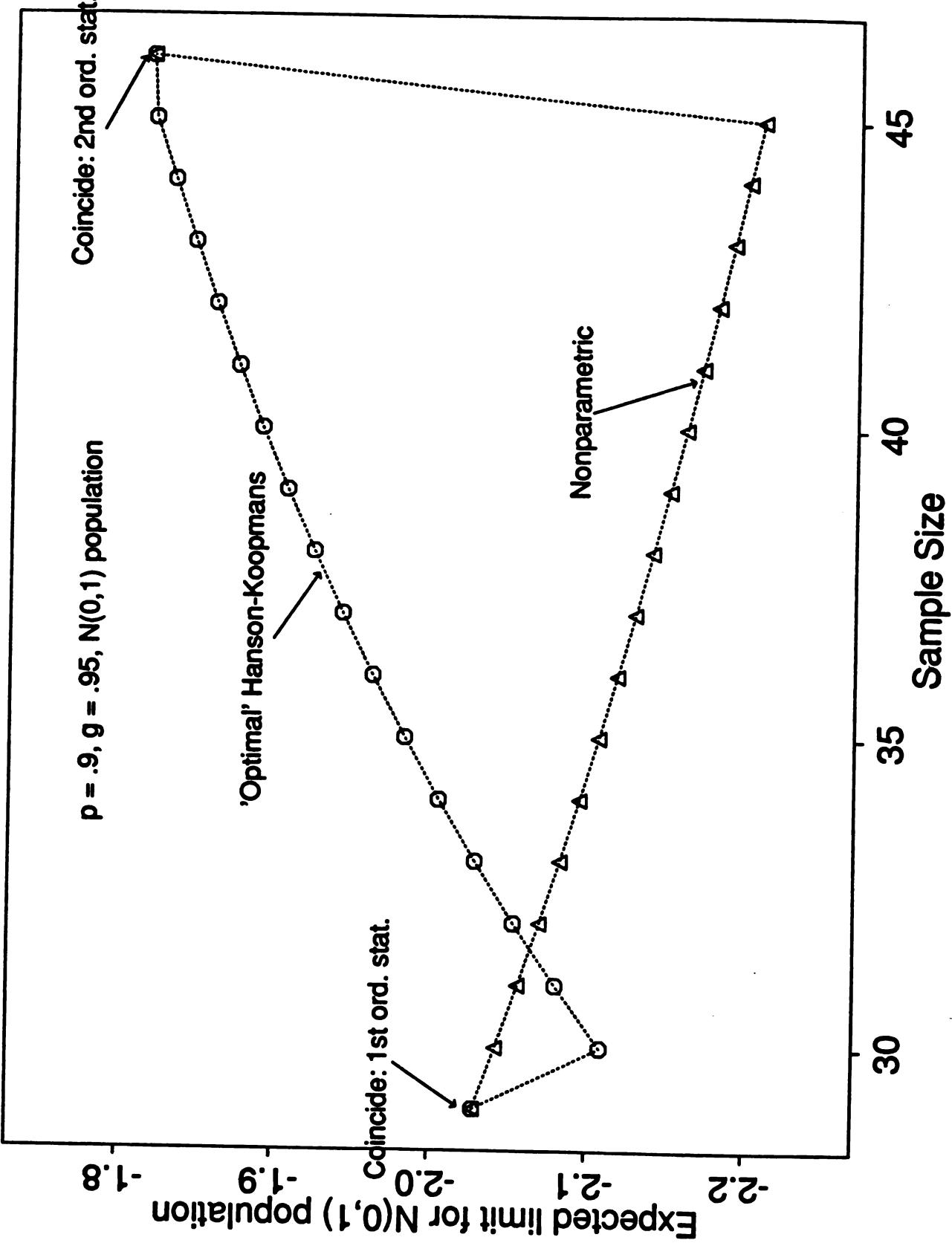
HK tolerance limits 'interpolate' NP tolerance limits

Figure 7

90

A   Factors k_l for (.90, .95) Lower Tolerance Limits

n	i	j	$k_l^{i,j}$
2	1	2	35.177
3	1	3	7.859
4	1	4	4.505
5	1	4	4.101
6	1	5	3.064
7	1	5	2.858
8	1	6	2.382
9	1	6	2.253
10	1	6	2.137
11	1	7	1.897
12	1	7	1.814
13	1	7	1.738
14	1	8	1.599
15	1	8	1.540
16	1	8	1.485
17	1	8	1.434
18	1	9	1.354
19	1	9	1.311
20	1	10	1.253
21	1	10	1.218
22	1	10	1.184
23	1	11	1.143
24	1	11	1.114
25	1	11	1.087
26	1	11	1.060
27	1	11	1.035
28	1	12	1.010
29	1	–	1
30	2	12	1.373
31	2	12	1.344
32	2	12	1.315
33	2	13	1.270
34	2	13	1.245
35	2	13	1.221
36	2	13	1.197
37	2	13	1.174
38	2	13	1.151
39	2	13	1.129
40	2	13	1.108
41	2	14	1.083
42	2	14	1.064
43	2	14	1.045
44	2	14	1.027
45	2	14	1.009
46	2	–	1

B A FORTRAN Subroutine for Determining $H(k_l)$

The following subroutine determines $H(k_l)$ using the closed form expression. With trivial modifications, this subroutine can be adapted for calculating $V(k_u)$. By using this subroutine along with a standard algorithm for finding a zero of a nonlinear function, a program can be written for determining the constants k_l and k_u.

```
         double precision function hk(p, n, i, j, xk)
         parameter (nmax = 1000)
         implicit double precision (a-h, o-z)
         dimension a(0:nmax), b(0:nmax), s(0:nmax)
         a(0) = 0
         do 10 m=1, n-j+1
            a(m) = 1.d0/(n-m+1)
10       continue
         do 20 m=n-j+2, n-i+1
            a(m) = xk/(n-m+1)
20       continue
         do 30 m=0, n-i+1
            b(m) = 0
            s(m) = 1
            do 40 mm=0, n-i+1
              if (mm .eq. m) go to 40
              r    = abs (a(m)-a(mm))
              s(m) = s(m)*(a(m)-a(mm))/r
              b(m) = b(m) +log(r)
40          continue
30       continue
c
         q    = -log(p)
         hk = 1
         do 60 m=1, n-i+1
            r    = abs (a(m))
            s1   = a(m)/r
            con  = s(m) *s1 *exp((n-i+1)*log(r) -b(m)-q/a(m))
            hk = hk -con
60       continue
         return
         end
```

References

[1] Barlow, R. E. and Proschan, F. (1981), *Statistical Theory of Reliability and Life Testing*, To Begin With, Sliver Spring, Maryland.

[2] Conover, J. (1980), *Practical Nonparametric Statistics*, 2nd ed., John Wiley and Sons, New York.

[3] Gere, J.M. and Timoshenko, S.P. (1984), *Mechanics of Materials*, Prindle, Weber & Schmidt, Boston.

[4] Hanson, D. L. and L. H. Koopmans (1964), "Tolerance Limits for the Class of Distributions With Increasing Failure Rates",*Annals of Mathematical Statistics*, 35, p. 1561-1570.

[5] Hogg, R. V. and A. T. Craig (1978), *Introduction to Mathematical Statistics*, 4th ed., Macmillan, New York.

[6] Johnson, N. L. and S. Kotz (1970), *Continuous Univariate Distributions – 1*, Houghton Mifflin, Boston.

[7] Mil Handbook 5E (1987), *Metallic Components for Aircraft Structures*, Naval Publications and Forms Center, Philadelphia.

[8] Mil Handbook 17C (1992), *Polymer Matrix Composites, Volume I: Guidelines*, Naval Publications and Forms Center, Philadelphia.

[9] Owen, D. B. (1968), "A Survey of Properties and Applications of the Noncentral t-Distribution", *Technometrics*, 10, 445-478.

[10] Woodward, W. A., and W. H. Frawley (1980),"One-Sided Tolerance Limits for a Broad Class of Distributions",*Journal of Quality Technology*, 12, p. 130-137.

ROBUST STATISTICAL DECISIONS
(AN EMPIRICAL INVESTIGATION)

EUGENE DUTOIT
U.S. ARMY INFANTRY SCHOOL
FORT BENNING, GEORGIA

ABSTRACT: THE PURPOSE OF THIS STUDY WAS TO DETERMINE, IN AN EMPIRICAL WAY, THE ROBUSTNESS OF DATA BASED STATISTICAL DECISIONS WHEN SEVERAL REASONABLE, BUT NOT EXACT, ANALYSES METHODS WERE USED ON THESE DATA. THE AUTHOR PRESENTED THE EMERGING RESULTS OF THIS EMPIRICAL STUDY THROUGH THE ANALYIS OF SEVERAL DATA SETS. AT LEAST TWO METHODS OF HYPOTHESIS TESTING WERE APPLIED TO EACH DATA SET AND THE PROBABILITIES OF REJECTING THE NULL HYPOTHESIS WERE COMPUTED AND COMPARED. THESE DATA SETS WERE OBTAINED FROM ACTUAL FIELD/SIMULATION EXPERIMENTS AND RANDOMLY SELECTED FROM STATISTICAL TEXTS. THE CONFERENCE ATTENDEES WERE ASKED TO SHARE THEIR OWN EXPERIENCES USING REASONABLE APPROACHES OF ANALYSES ON THE SAME DATA SETS. AN ANALYSIS FORM WAS HANDED OUT TO ANY ATTENDEE WHO MIGHT LIKE TO COLLABORATE.

THE FOCUS OF THIS STUDY IS ON HYPOTHESIS TESTING, NOT ESTIMATION. IT IS IMPORTANT TO EMPHASIZE THAT THE SIGNIFICANCE LEVELS THAT ARE OFTEN USED IN THE WEAPONS ACQUISITION PROCESS APPROACH 10%. RELATIVELY SMALL SAMPLE SIZES ARE USED BECAUSE THE ITEMS ARE EXPENSIVE. THE TOPIC OF THE PAPER MAY SOUND LIKE HEARISEY. THE STUDY IS NOT ADVOCATING WRONG ANALYSIS. THE GOAL IS SIMPLY TO GET A "FEELING" FOR THE SENSITIVITY OF THE DECISION MAKING PROCESS WHEN ALTERNATIVE METHODS OF ANALYSES ARE USED TO EXAMINE THE DATA. THIS QUEST HAS COME ABOUT BECAUSE SOMETIMES STUDY REVIEWERS AND DECISION MAKERS WILL ASK IF THE "SO AND SO" METHOD WAS USED IN THE STUDY BECAUSE THEY HEARD OR READ IN ANOTHER STUDY THAT THIS ALTERNATIVE METHOD WAS AN APPROPRIATE WAY TO GO OR WAS THE MOST CONSERVATIVE OR LIBERAL FOR THE STATED CONDITIONS. THESE COMMENTS MAY, AND OFTEN DO, HAVE SOME MERIT. IT WOULD BE NICE TO HAVE SOME ANSWER AND BE ABLE TO SAY SOMETHING LIKE THIS..."ALTHOUGH I CANNOT SAY THAT WE USED YOUR SPECIFIC METHOD TO ANALYZE THESE DATA WE DID USE AT LEAST TWO APPROACHES TO THE DATA ANALYSIS. WE HAVE SOME HISTORICAL BASIS TO SAY THAT THE DECISION THAT WOULD HAVE RESULTED IF WE HAD USED AN ALTERNATIVE METHOD WOULD HAVE NOT (OR HAVE) BEEN DIFFERENT." IT IS IMPORTANT TO POINT OUT TO THE READER THAT THIS PAPER CONTAINS NO CLASSIFIED WEAPONS DATA. THE EXAMPLES THAT ARE GIVEN IN THESE PARAGRAPHS USE DATA OBTAINED FROM TEXTBOOKS OR GRADUATE STUDENT RESEARCH.

FIGURE 1 IS THE BASIC "DATA COLLECTOR" FOR THIS EMPIRICAL STUDY. THIS IS THE FORM THAT WAS DISTRIBUTED AT THE ARMY DESIGN OF EXPERIMENTS CONFERENCE IN ORDER TO GET ADDITIONAL INPUT FROM SOME OF THE ATTENDEES. THE FORM SHOULD BE SELF-EXPLANITORY. THERE IS SPACE FOR A BRIEF DESCRIPTION OF THE PROBLEM AND THE VARIABLES THAT ARE BEING COMPARED IN THE STUDY. THESE ARE THE VARIABLES THAT ARE IMPORTANT IN THE STATISTICAL DECISION. THERE IS A SPACE TO ENTER THE VALUE OF "P" FOR EACH STATISTICAL METHOD USED FOR HYPOTHESIS

TESTING.THE DECISION INDICATES EITHER A STATISTICALLY SIGNIFICANT
DIFFERENCE (SD) OR NO SIGNIFICANT DIFFERENCE (NSD). THE AUTHORS'
COMPLETE MAILING ADDRESS, TELEPHONE NUMBERS AND FAX NUMBER APPEAR ON
THE BOTTOM OF THE FORM.

      FIGURE 2 GIVES A SUMMARY OF SOME OF THE STATISTICAL DECISIONS
THAT CAN BE PRESENTED IN THESE PROCEEDINGS. METHOD 1 IS THE PREFERRED
TECHNIQUE BASED ON THE UNDERLYING DISTRIBUTIONS OF THE DATA. METHOD 2
IS A REASONABLE ALTERNATIVE METHOD (USUALLY NONPARAMETRIC) THAT IS
NORMALLY DONE AS A MATTER OF COURSE. THIS CONVENTION HOLDS FOR THE
FIRST THREE STUDIES LISTED IN FIGURE 2. THE FLEITAS STUDY HAD A MIX
OF PARAMETRIC AND NONPARAMETRIC TECHNIQUES AS THE PREFERRED METHOD 1.
THIS FIGURE SHOWS THAT THERE IS A HIGH DEGREE OF AGREEMENT BETWEEN
THE "P" VALUES FOR THE ALTERNATIVE STATISTICAL METHODS. THEREFORE
THERE IS A CORRESPONDING DEGREE OF AGREEMENT IN THE RESULTANT
STATISTICAL DECISIONS. THIS SMALL SAMPLE OF RESULTS WOULD INDICATE
THAT THESE STATISTICAL DECISIONS ARE FAIRLY ROBUST WITH RESPECT TO
THESE DATA SETS (AND TESTING TO THE 10% LEVEL OF SIGNIFICANCE).IN A
SENSE I WILL "FLING DOWN THE GAUNTLET" AND CAUTIOUSLY AND TENTATIVELY
SAY THAT THE REASONABLE ALTERNATIVE METHOD WILL PROVIDE THE DECISION
MAKER A BASIS FOR ARRIVING AT A CONSISTENT CONCLUSION.

      I WOULD APPRECIATE YOUR THOUGHTS AND COMMENTS ABOUT THIS
PROBLEM. IF YOU COULD OR WOULD LIKE TO PROVIDE INSIGHTS, THOUGHTS,
COMMENTS AND EXAMPLES CONCERNING ALTERNATIVE METHODS OF HYPOTHESIS
TESTING I WOULD BE DELIGHTED TO HEAR FROM YOU. YOU CAN REACH ME USING
THE INFORMATION PROVIDED AT THE BOTTOM OF FIGURE 1.

FIGURE 1

PROBLEM DESCRIPTION

_____
_____
_____
_____
_____

VARIABLES COMPARED

_____
_____
_____
_____
_____

	"P VALUE"	DECISION	COMMENTS

PREFERRED METHOD(ALT # 1)

ALTERNATIVE METHOD(ALT 2)

ALTERNATIVE METHOD(ALT 3)

PLEASE SEND THIS COMPLETED FORM TO
COMMANDANT
U S ARMY INFANTRY SCHOOL
ATTN: ATSH - CDC - O (DUTOIT)
FORT BENNING GA 31905-5400

(404) 545-3165/3166
   DSN 835-3165/3166
   FAX: (404) 545-2517

THANK YOU FOR YOUR HELP,

GENE DUTOIT

FIGURE 2
SUMMARY OF STATISTICAL DECISIONS

PROBLEM & VARIABLES	METHOD #1	"P"	METHOD #2	"P"
TATSUOKA(MULTIVARIATE ANALYSIS) PG 200. TWO INDEPENDENT VARIABLES & TWO DEPENDENT VARS	NORMAL SCORES ANOVA	.76 .82 NSD	KRUSKAL-WALLIS	.86 .86 NSD
ELLIOT & TABOR(MULTI-VARIATE ANAL) PG.278. THREE INDEPENDENT VAR & TWO DEPENDENT VARS	NORMAL SCORES ANOVA	.07 .08 SD	KRUSKAL-WALLIS	.08 .09 SD
VALENTINE P.; 1991 STUDY; 2 INDEPENDENT GROUPS & 1 DEPENDENT VARIABLE	NORMAL SCORES VARIANCES EQUAL T-TEST	.075 SD	MANN-WHITNEY U	.080 SD
FLEITAS D.;1991 STUDY ON EATING DISORDERS 16 DIMENSIONS OF BEHAVIOR & PERSONALITY COMPARISONS OF TWO GROUPS	MIXTURE OF PARAMETRIC AND NONPARAMETRIC (T-TEST & MANN-WHITNEY)		THE SAME AS METHIOD # 1 THE "P" VALUES WERE CLOSE & THE DECISIONS MATCHED 15/16 TIMES	

:

REFERENCES

ELLIOTT,M.; TABOR,C.;MULTIVARIATE ANALYSIS,GEORGIA STATE
     UNIVERSITY,1978
FLEITAS,D.; "AN INDEPENDENT STUDY OF EATING DISORDERS",1991
TATSUOKA,M.;MULTIVARIATE ANALYSIS, JOHN WILEY,1971
VALENTINE,P.; MODERATION OF TEST ANXIETY THROUGH PHYSIOLOGICAL
     SELF AWARENESS AND CONTROL, A DISSERTATION PROPOSAL, GEORGIA
     STATE UNIVERSITY,1991

THE EFFECT OF SAMPLE SIZE ON THE VARIABLITY OF SAMPLE MATERIAL PROPETY VALUES

Bernard Harris

1. Introdutction and Summary. In certifying materials for aircraft construction, a commonly used certification criterion is called a statistically based material property value. For the specific application being considered, the tests are both destructive and expensive to carry out. Consequently, very small samples are usually employed. Therefore, it is desirable to study the small sample behavior of these statistically based material property values. In the usual statisical terminology (rather than the terminology of materials testing used above), a statistically based material property value is a lower tolerance limit. Specifically, a B-basis value is a 95% lower confidence limit on the tenth percentile of the probability distribution which has been assumed, and an A-basis value is a 95% lower confidence limit on the first percentile.

In this report, we assume that the data is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. It is well-known that these statistically based material property values (tolerance limits) can be written as

$$T_{n,\alpha,\gamma}(\overline{X}, s) = \overline{X} - k(n, \alpha, \gamma)s,$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \quad s = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}},$$

and $k(n, \alpha, \gamma)$ is $1/\sqrt{n}$ times the $100(1 - \alpha)$th percentile of the noncentral t-distribution with $n - 1$ degrees of freedom and non-centrality parameter $\sqrt{n}\Phi^{-1}(1 - \alpha)$, where Φ is the standard normal cumulative distribution function. For an A-basis value, $\alpha = .01$; for a B-basis value, $\alpha = .05$. For notational simplicity, we will denote (1) by T, omitting the subscripts. In order to provide a simple picture of the behavior of T as n changes, the moments of T are derived in the Appendix to this report.

In Section 2, some numerical tabulations are presented which provide concrete illustrations of the material given in the Appendix. These illustrate the variability of T for sample sizes from 2 to

500 for both A-basis values and B-basis values and the changes in such variability with increasing sample sizes.

The preparation of this report was motivated by the following considerations. The author has been a member of a working group concerned with the certification of advanced composites for aircraft construction. In the course of this activity, the author has come to suspect that some of the materials engineers conclude that the B-basis (A-basis) values are intrinsic material properties and not random variables. It is the purpose of this report to demonstrate that these are random variables and possess statistical variability.

It is not difficult to extend the calculations in Section 2 to obtain the probability density function of T. However, the simple description of the variability given by the mean, variance, and two-sigma limits for T provides the user with a suitable description of the variability that will be encountered.

2. Numerical Illustrations of the Variability of Statistically Based Material Property Values. In this section, we provide numerical illustrations of the mean and standard deviation of A-basis and B-basis values for samples of size $2, 3, \ldots, 20, 30, 35, \ldots, 50, 60, \ldots, 100, 200, 500, \infty$ for normally distributed data from a population with $\mu = 200, \sigma = 10$ (values which are reasonable for some composite materials). Also, two-sigma limits for such A-basis and B-basis values are given. The k-values employed have been taken from MIL-HDBK-17-1C, draft dated 7 September 1990, Materials Technology Laboratory, U.S. Army.

Table 1

The Expected Value, Standard Deviation, and Two-Sigma Limits of B-basis Values in Samples of Size n from a Normally Distributed Population with Mean $\mu = 200$ and Standard Deviation $\sigma = 10$.

Sample Size n	Expected Value	Standard Deviation	Two-Sigma Limits
2	035.787	124.266	(0.000, 284.316)
3	145.435	29.101	(87.233, 203.637)
4	161.676	16.941	(127.794, 195.558)
5	167.965	12.459	(142.867, 193.063)
6	171.387	10.109	(151.169, 191.605)
7	173.560	8.646	(156.268, 190.852)
8	175.073	7.639	(159.795, 190.351)
9	176.213	6.893	(162.427, 189.999)
10	177.094	6.318	(164.458, 189.730)
11	177.801	5.858	(166.085, 189,517)
12	178.386	5.480	(167.426, 189.346)
13	178.884	5.162	(168.560, 189.208)
14	179.311	4.890	(169.531, 189.091)
15	179.676	4.655	(170.366, 188.986)
16	179.996	4.450	(171.096, 188.896)
17	180.290	4.267	(171.756, 188.824)
18	180.548	4.104	(172.340, 188.756)
19	180.779	3.958	(172.863, 188.695)
20	180.982	3.826	(173.330, 188.634)
21	181.177	3.705	(173.767, 188.587)
22	181.353	3.594	(174.173, 188.533)
23	181.511	3.493	(174.525, 188.492)
24	181.660	3.399	(174.862, 188.458)
25	181.801	3.312	(175.177, 188.425)
30	182.373	2.956	(175.461, 188.285)
35	182.797	2.691	(177.415, 188.179)
40	183.128	2.484	(178.160, 188.096)
45	183.405	2.317	(178.771, 188.039)
50	183.624	2.180	(179.264, 187.984)
60	183.978	1.962	(180.054, 187.902)
70	184.237	1.799	(180.639, 187.835)
80	184.449	1.669	(181.111, 187.787)
90	184.623	1.563	(181.497, 187.749)
100	184.769	1.475	(181.819, 187.719)
200	185.518	1.014	(183.498, 187.538)
500	186.147	.630	(184.887, 187.407)
∞	187.180	0.000	(187.180, 187.180)

Table 2

The Expected Value, Standard Deviation, and Two-sigma Limits of A-basis Values in Sample of Size n from a Normally Distributed Population with Mean $\mu = 200$ and Standard Deviation $\sigma = 10$.

Sample Size n	Expected Value	Standard Deviation	Two-Sigma Limits
2	-95.967	223.718	(0.000, 351.469)
3	106.476	50.913	(4.650, 208.302)
4	137.592	27.833	(81.926, 193.258)
5	146.035	20.093	(105.849, 186.221)
6	151.833	16.094	(119.645, 184.021)
7	155.466	13.632	(128.202, 182.730)
8	157.983	11.949	(134.085, 181.881)
9	160.019	10.716	(138.587, 181.451)
10	161.278	9.771	(141.736, 180.820)
11	162.430	9.019	(144.392, 180.468)
12	163.371	8.405	(146.561, 180.181)
13	164.164	7.891	(148.382, 179.946)
14	164.832	7.456	(149.920, 179.744)
15	165.422	7.078	(151.266, 179.578)
16	165.932	6.750	(152.432, 179.432)
17	166.389	6.459	(153.471. 179.307)
18	166.792	6.201	(154.390, 179.194)
19	167.149	5.970	(155.209, 179.089)
20	167.481	5.761	(155.959, 179.003)
21	167.775	5.571	(156.633, 178.917)
22	168.052	5.397	(157.258, 178.846)
23	168.302	5.238	(157.826, 178.778)
24	168.534	5.091	(158.352, 178.716)
25	168.747	4.955	(158.837, 178.657)
30	169.623	4.402	(160.819, 178.427)
35	170.269	3.994	(162.281, 178.257)
40	170.779	3.677	(163.425, 178.133)
45	171.184	3.422	(164.340, 178.028)
50	171.526	3.212	(165.102, 177.950)
60	172.049	2.884	(166.281, 177.817)
70	172.450	2.636	(167.178, 177.722)
80	172.756	2.442	(167.872, 177.640)
90	173.016	2.283	(168.450, 177.582)
100	173.228	2.152	(168.924, 177.532)
200	174.332	1.469	(171.394, 177.270)
500	175.263	.902	(173.459, 177.067)
∞	176.740	0.000	(176.740, 176.740)

Appendix

Let $X_1, X_2, \ldots, X_n$ be independent normally distributed random variables with mean μ and variance σ^2. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2 = \sum \frac{(X_i - \overline{X})^2}{n-1}.$$

Then $\overline{X}$ and s^2 are independent, $\overline{X}$ is normally distributed with mean μ and variance σ^2/n, $(n-1)s^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom.

To simplify notation, we set $T_{n,\alpha,\gamma}(\overline{X}, s) = T$, and write

$$E(T^v) = \sum_{j=0}^{v} \binom{v}{j} E(\overline{X})^j (-1)^{v-j} k_{n,\alpha,\gamma}^{v-j} E(s)^{v-j}. \tag{1}$$

To evaluate $E(\overline{X})^j$, write

$$\overline{X} = \overline{X} - \mu + \mu.$$

Then,

$$E(\overline{X}^j) = E\left\{\sum_{i=0}^{j} \binom{j}{\tau}(\overline{X} - \mu)^\tau \mu^{j-\tau}\right\}$$

$$= \sum_{\tau=0}^{j} \binom{j}{\tau} \mu^{j-\tau} E(\overline{X} - \mu)^\tau$$

$$= \sum_{\tau=0}^{[j/2]} \binom{j}{2\tau} \frac{\mu^{j-2\tau} 2^\tau \sigma^{2\tau} \Gamma(\frac{2\tau+1}{2})}{n^\tau \sqrt{\pi}}. \tag{2}$$

Also, since

$$s^2 = \frac{u\sigma^2}{m}$$

where u has the chi-square distribution with $m = n - 1$ degrees of freedom,

$$Es^k = \int_0^\infty \frac{u^{k/2}\sigma^k}{m^{k/2}} \frac{e^{-u/2} u^{m/2-1}}{2^{m/2}\Gamma(\frac{m}{2})} du$$

$$= \frac{\sigma^k 2^{k/2}}{m^{k/2}} \frac{\Gamma(\frac{m+k}{2})}{\Gamma(\frac{m}{2})}. \tag{3}$$

Substituting (2) and (3) into (1), we obtain

$$E(T^v) = \sum_{j=0}^{v} \frac{\binom{v}{j}(-1)^{v-j}\sigma^{v-j}2^{(v-j)/2}\Gamma(\frac{n+v-j-1}{2})}{(n-1)^{(v-j)/2}\Gamma(\frac{n-1}{2})\sqrt{\pi}} \cdot$$

$$k_{n,\alpha,\gamma}^{v-j}\sum_{\tau=0}^{[j/2]}\frac{\binom{j}{2\tau}\mu^{j-2\tau}2^{\tau}\sigma^{2\tau}\Gamma\left(\frac{2\tau+1}{2}\right)}{n^\tau} \tag{4}$$

In particular, the variance of T is easily written as

$$\sigma_T^2 = \sigma_{\bar{X}}^2 + k^2\sigma_S^2$$

$$= \frac{\sigma^2}{n} + k^2(Es^2 - (Es)^2)$$

$$= \frac{\sigma^2}{n} + k^2\left(\sigma^2 - \frac{2\sigma^2}{(n-1)}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)^2\right)$$

$$= \sigma^2\left(\frac{1}{n} + k^2\left(1 - \frac{2}{n-1}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)^2\right)\right). \tag{5}$$

Small Sample Design Allowables
From Paired Data Sets

Donald M. Neal   Trevor D. Rudalevige   Mark G. Vangel

U.S. Army Materials Technology Laboratory
SLCMT-MRS-MM Arsenal Street
Watertown, Massachusetts 02172-0001

Abstract

This paper identifies an acceptable statistical procedure
for obtaining design allowable values from a small set of
material strength data. The allowable represents a ma-
terial design number defined as the 95% lower confidence
bound on the specified percentile of the population of ma-
terial strength data. The percentiles are the first and tenth
for the A and B allowables. The proposed method reduces
the penalties commonly associated with small sample allow-
able computation by accurately maintaining the definition
requirements and reducing variability in the estimate. Ap-
plication of very small samples will obviously reduce costs in
testing and manufacturing which is the primary motivation
for this study.

In the evaluation process five methods were considered for
computing the design allowable. Three of these methods in-
volved certain statistical distribution assumptions while the
other two were nonparametric procedures. The latter meth-
ods introduced a pooling process such that the small sample
was combined with a larger, previously obtained sample.

Monte Carlo studies showed that the nonparametric pro-
cedures are the most desirable for computing the design al-
lowable value.

1 Introduction

The A or B statistically based design allowable value is a statistic which is less than the first or tenth percentile of the population with probability .95. That is, the value is a 95% lower tolerance limit for the percentile. In Figures 1A and 1B, a graphical display is shown for the B allowable value probability density function for sample sizes of n equal to 10 and 50 from a standard normal population. The dotted vertical lines indicate the tenth percentile of the population and the probability that the allowable is less than or equal to the tenth percentile is .95 for the design allowable value probability density function. The graphical display of the allowable value density functions show much less dispersion for $n = 50$ than for $n = 10$. Therefore, small samples will usually result in lower allowable values. In,[1, 2, 3, 4, 5], various procedures are described for determining the statistical design allowable values.

The motivation for the work described in this paper resulted from a need by the aircraft industry to obtain a less conservative, statistically based material design value from a small sample of composite material strength data. Here, 'conservative' is to be interpreted to mean 'excessively low', which corresponds to a design engineer's use of the word. Statistical conservatism, that is a confidence exceeding the nominal level of .95, need not be present for 'engineering conservatism' to be a problem.

The use of small samples reduces the amount of testing and consequently the manufacturing cost of composite aircraft structures. For example, in order to qualify a composite material to be used in the manufacture of a commercial aircraft, the FAA,[6] requires property values for tension, compression, and shear tests subjected to the enviromental conditions: hot-wet, cold-dry, and room temperature for three separate batches of material. In the development of a composite tail section by one of the major aircraft companies the cost of testing was more than 20 million dollars. In addition to the cost, excessively conservative allowable values can also result in an over-design situation, since the value often provides information in determining a structural design.

In order to avoid the penalty associated with using small samples in the tolerance limit computation, a procedure is introduced in this paper involving pooling a large

[1]Military Handbook 17B, Army Materials Technology Laboratory, Polymer Matrix Composites, Volume 1, Guidelines, 1988.

[2]Neal, D. M., Vangel, M. G., and Todt, F., "Determination of Statistical Based Composite Material Properties" in Engineered Materials Handbook, Composites, C.A. Dostal, ed., American Society of Metals Press, Metals Park, Ohio, Vol. 1, 1987.

[3]Neal, D. M., Vangel, M. G., "Statistical Based Material Properties - A Military Handbook-17 Perspective", MTL TR 90-5, U.S. Army Material Technology Laboratory, Watertown, Massachusetts 02172-0001, 1990.

[4]Neal, D. M. and Spiridigliozzi, L., "An Efficient Method for Determining the 'A' and 'B' Design Allowables", ARO Report 83-2, U.S. Army Laboratory Command, Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709-2211, 1983.

[5]Shyprykevich, P.,"The Role of Statistical Reduction in the Development of Design Allowables for Composites", Test Methods for Design Allowables for Fibrous Composites: 2nd Vol., ASTM STP 1003, pp. 111-135, 1989.

[6]Soderquist, Joseph, National Resource Specialist for Composites (FAA), Private Conversation

sample with a smaller one in order to obtain the allowable value. This is done in order to reduce the inherent variability that occurs from applying the smaller data set alone.

In the pooling process the larger data set should be obtained from prior available test results or from less expensive tests. Ideally, both samples should be from the same material, test, and enviromental conditioning process. In the pooling process it is assumed that for a given material (eg., graphite-epoxy) there are similar classes of failure modes.

In order to avoid the uncertainties involved in identifying a statistical model from a small sample when computing the allowable value, this paper introduces two nonparametric methods (Ferguson,[7] and the Modified Hanson-Koopmans,[8]) In applying the Bayesian nonparametric method, the larger set represents the prior and the smaller one the empirical data. In the Modified Hanson-Koopmans method an ordered array of strength measurements is obtained from the pooled data sets. The tolerance limit is determined from a specific ratio of ordered values multiplied by a factor determined from the sample size of the pooled data.

The Reduced Ratio Method,[9] another procedure for computing small sample design allowables, was also evaluated. This method is commonly used by the aircraft industry. For example, a U.S. helicopter company routinely uses this method for obtaining allowables from six specimens tested in tension at 180°F. In the analysis an additional, previously obtained sample of at least thirty room temperature tension test results are included in order to reduce variability in the allowable estimate.

2 Determination of Allowable Values Nonparametric Bayesian Method

The nonparametric Bayesian,[7] allowable value is obtained from the following. Let $\{x_i\}_1^n$ represent the current empirical data which the allowable value is to represent and $\{t_j\}_1^m$ the larger prior data set obtained from previous test results.

In the analysis the cummulative density function (CDF) of the prior (larger data set) is written as

$$F_0(t) = \alpha((-\infty, t])/\alpha(R) \tag{1}$$

where $\alpha(R)$ is the sample size and $\alpha((-\infty, t])$ represents the number of values less than t from $\{t_j\}_1^m$. The CDF of the smaller sample $\{x_i\}_1^n$ is

$$F_n(t \mid x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \delta_{x_i}((-\infty, t])/n \tag{2}$$

---

[7]Ferguson, T. S., "A Bayesian Analysis of Some Nonparametric Problems", Annals of Statistics, Vol. 1, No. 2, 209-230, 1973.

[8]Vangel, M. G., "Lower Tolerance Limits for Log-Convex Distributions", to be published.

[9]Metallic Materials and Elements for Aerospace Vehicle Structures, MIL-HDBK-5C, 15 September 1976, pp. 9-14.

where n is the sample size and the sum over i of $\delta_{x_i}(t)$ is equal to the number of x_i values less than or equal to t. For example,

$$\text{if } t = 1, 2, 3, 4, 5$$
$$\text{and } x = 6, 7, 8, 9, 10$$
$$\text{then } F_n(5 \mid 6, 7, 8, 9, 10) = 0.$$
$$\text{If } t = 11, 12, 13, 14, 15$$
$$\text{then } F_n(11 \mid 6, 7, 8, 9, 10) = 1.$$

The posterior distribution for $\{x_i\}_1^\infty$ is then written as

$$\hat{F}_n(t \mid x_1, x_2, ..., x_n) = P_n F_0(t) + (1 - P_n) F_n(t \mid x_1, x_2, ..., x_n), \tag{3}$$

where

$$P_n = \frac{\alpha(R)}{\alpha(R) + n}. \tag{4}$$

An example of a Bayes estimate for $x = 1$ when

$$t = 1, 2, 3, 4, 5$$
$$\text{and } x = 1, 2, 3, 4, 5 \text{ is}$$
$$\hat{F}_n = P_n F_0 + (1 - P_n) F_n = (.5)(.2) + (.5)(.2) = .2.$$

3 Nonparametric Tolerance Limit on the Bayesian Quantile Estimate

The allowable value as described previously is a tolerance limit on the quantile estimates. The process for obtaining that limit is shown in this section. Initailly, a random sample $F(Y)$ of size $M = \alpha(R) + n$ is assumed independent of the mixture of the prior and empirical data sets shown in Equation 3. By ordering a sample of $Y_1, Y_2, ..., Y_M$ values, the probability density function for Y_i, $1 \leq i \leq M$ can be written as a Beta distribution,

$$f_{x_i}(z) = \frac{\Gamma(M) z^{uM-1} (1 - z)^{(1-u)M-1}}{\Gamma(uM) \Gamma((1 - u)M)}, \tag{5}$$

where $z_{(i)} = F(Y_{(i)})$ and $i = uM$ with u representing the CDF value corresponding to the ith ordered number. The tolerance limit Y^* for Y_q is

$$P(Y_q \geq Y^*) = 1 - \alpha = P[F(Y_q) \geq F(Y^*)] \tag{6}$$

where Y_q is the $100q^{th}$ percentile of Y. Since

$$P(Y \geq Y_q) = \int_0^q \frac{\Gamma(M) z^{uM-1} (1 - z)^{(1-u)M-1}}{\Gamma(uM) \Gamma((1 - u)M)} dz \tag{7}$$

from Equation 5, a $1 - \alpha$ tolerance limit on Y_q can be obtained by solving for u from the following. In the case of the B allowable computation, $\alpha = .05$ and $q = .10$, Equation 7 can be written as

$$\int_0^{.10} \frac{\Gamma(M) z^{uM-1} (1 - z)^{(1-u)M-1}}{\Gamma(uM) \Gamma((1 - u)M)} dz = .95. \tag{8}$$

See Table I for tabulation of u and M values that satisfy Equation 8.

Solving for u in Equation 8 determines the lower tolerance limit of the CDF of sample size M where the ith ordered value is equal to uM. Obtaining a lower ordered CDF value from Equation 3 that is approximately equal to a u determines the $1 - \alpha$ tolerance limit of the qth quantile of the posterior CDF for a sample size M.

An example of this would be if there were only prior data $\{t_j\}_1^{30}$ and a B allowable value is required where

$$t = 5, 6, 7, 8, 12, 16, 20, 25,...,40 \text{ and}$$
$$F_0(t) = .033, .066, .099,...,1.0,$$

then $M = 30$ and $u = .034$ from Table I. The allowable value t_j is determined from the approximate solution of $u \approx F(t)$ resulting in $t_1 = 5$; therefore, the first ordered value of the prior represents the B allowable value, which is the same as the nonparametric quantile sign test,[10] result, when the sample size is 30.

4 The Nonparametric Modified Hanson-Koopmans (MHK) Procedure

A nonparametric procedure (MHK),[8] for estimating the allowable value is introduced for any sample size greater than or equal to 2. The method is a modification of Hanson-Koopmans,[11] process. The modification has reduced the conservatism in computing property values when compared with the original method.

The method involves the following. Let $x_1,...,x_n$ be the order statistics of an independent and identically distributed sample from a continuous distribution F. Assume that F is log-convex, that is -log $F(x)$ is a convex function. The class of log-convex functions includes a large enough group of distributions so that the following procedure involving log-convex functions can be considered nonparametric for most purposes.

The Hanson-Koopmans lower tolerance limits are of the form

$$T_{rs} = kx_r + (1 - k)x_s, \tag{9}$$

where $r < s$ and $k \geq 1$. The tolerance limit T_{rs} can be negative, even if the distribution F is zero for any negative values. A practical solution to this problem is to apply the Hanson-Koopmans approach to the log of the data x, that is,

$$T_{rs}^* = k \log x_r + (1 - k) \log x_s, \tag{10}$$

and then obtain by exponentiation the following

$$T_{rs}' = e^{k \log x_r} + e^{(1-k) \log x_s} = x_s \left(\frac{x_r}{x_s}\right)^k. \tag{11}$$

---

[10]Conover, W. J., "Practical Nonparametric Statistics", John Wiley and Sons, 1980, p. 111.

[11]Hanson, D. L. and Koopmans, L. H., "Tolerance Limits for the Class of Distributions with Increasing Hazard Rates", Annals of Mathematical Statistics, Vol. 35, 1964.

For most distributions of interest, $\tilde{T}_{rs}$ still provides conservative tolerance limits, although technically $\tilde{T}_{rs}$ is valid for a class of distributions smaller than the log-convex class corresponding to T_{rs}.

In order to determine the B allowable value, the r, s, and k values are obtained for a given n in Table II. Tables are also available for the A allowable in Reference 8.

5 Allowable Computation for Normal and Weibull Models

The following small, single sample, data set allowable computation procedures were included for comparison purposes. This comparison is made with respect to the results obtained from the other methods described in this paper.

The normal PDF is

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, \tag{12}$$

where μ and σ are the mean and standard deviation. The normal allowable is

$$A_N = \bar{X} - K_A s, \tag{13}$$

where K_A is a factor obtained from Reference 1 and $\bar{X}$ and s are the sample mean and standard deviation.

The Weibull allowable computation is as follows. The Weibull PDF is

$$f_W(x) = \frac{\beta x^{\beta-1}}{\alpha^\beta}e^{-(x/\alpha)^\beta}, \tag{14}$$

where β and α are the shape and scale parameters and the Weibull allowables can be written as

$$A_W = \hat{\alpha}[-\log(P_A)]^{1/\beta}, \tag{15}$$

where the P_A's are tabulated in Reference 3 with $\hat{\alpha}$ and $\hat{\beta}$ being the maximum likelihood estimates for α and β obtained from an algorithem also shown in Reference 3.

6 The Reduced Ratio Method (RRM)

The Reduced Ratio Method,[9] determines an allowable value for a smaller data set $\{S_i\}_1^n$ by introducing an indirect computation procedure involving a larger, previously obtained set of data, $\{L_j\}_1^m$.

The first step is to determine the mean of L, that is $L = \frac{1}{m}\sum_{i=1}^{m} L_j$. The second step requires obtaining the ratios $R_1 = S_1/L$, $R_2 = S_2/L$, ..., $R_n = S_n/L$ and the mean ($\bar{R}$) of the R_i's. The reduced mean, $\bar{R}^*$ is then obtained from

$$\bar{R}^* = \bar{R} - t_{(.95)}V_R/\sqrt{n}, \tag{16}$$

112

where $t_{(.95)}$ is the .95 quantile of the t distribution for $n - 1$ degrees of freedom and V_R is the standard deviation of the R_i's. The next step is to compute an allowable (L_B) from the L sample using some single sample procedure such as described in the previous section. After obtaining L_B the allowable S_B for S is determined as follows

$$S_B = L_B \bar{R}^*. \tag{17}$$

7 The Pooling Process

The pooling process, as previously mentioned, requires combining a smaller data set S (the one represented by the allowable) with a larger set L obtained from prior test results. In the MHK process the objective is to represent S with a combined data set of S and L with sample size $m = n_S + n_L$. In the Bayes method the prior is represented by L and the empirical data by S.

If both the means and variances of S and L are known to be equal, then the pooling process can be easily justified. Unfortunately, this is seldom the case. Therefore, the following transformation is suggested. Let L_i and S_i be the data from sets L and S respectively and define the new data sets S* and L* by

$$S_i^* = \frac{S_i - \bar{S}}{\bar{S}} \tag{18}$$

and

$$L_i^* = \frac{L_i - \bar{L}}{\bar{L}}, \tag{19}$$

where $\bar{L}$ and $\bar{S}$ are the data set means. This procedure involves reducing the mean of S and L to a common mean of zero for S* and L*. In addition, the transformed data sets, S* and L*, have standard deviations equal to the CV's of S and L. Schematics of this transformation are shown in Appendices A and B.

It is suggested that an equality of variance test between S* and L* be made in order to determine if an excessively large difference in variance exists. The Siegel-Tukey nonparametric rank sum method,[12] proved effective in testing for equality of variance although for small samples (less than ten), the test on equality of variance will result in a certain amount of uncertainty.

8 Allowable Values for S* from Pooled Data

8.1 Bayes Solution

In the Bayes application let the smaller sample x (newly obtained data) of size n_S be represented by the S* values and the larger sample t (the prior) with n_L values by L*. Initially, u in Equation 8 is obtained from Table I for M equal to the combined sample sizes of S* and L* in order to determine the allowable for S*. CDF values are

---

[12]Siegel, S. and Tukey, J. W., "A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples", Journal of American Statistical Association, September, 1960.

determined from Equation 3 where $t = L^*$ and $x_i = S_i^*$ $i = 1,2,...,n$. Equating the CDF value that corresponds to u determines the ordered (uM) value of $\hat{F}_n$. Inverting $\hat{F}_n$ so that the corresponding ordered test result is obtained then determines the allowable value S_B^*.

8.2   The Modified Hanson-Koopmans Method

The nonparametric,[8] solution for obtaining allowable values involves pooling the values from S^* and L^* and letting the combined ordered array of values be x in Equation 11 with sample size $n = n_S + n_L$. Let this value be denoted S_B^*(in place of $\tilde{T}_{r_s}$). This method is very simple to apply yet provides results for any sample size greater than 2.

9   Transformation Procedure in Determining Allowable

The allowable value for S^* is not sufficient since S and L were the original data sets involved in the analysis and their magnitudes differ from S^* and L^*. Therefore, the following transformation is required:

$$S_B = S_B^* \bar{S}_{.95} + \bar{S}_{.95} \tag{20}$$

where S_B is the required allowable value for the small sample S. The $\bar{S}_{.95}$ values represent the lower 95% confidence value for the mean of the S values. The purpose in using $\bar{S}_{.95}$ instead of $\bar{S}$ is to adjust for the variability in estimating the mean $\bar{S}$ of the small sample S. This variability in $\bar{S}$ directly effects the computation of the allowable S_B. This often results in S_B values being greater than the pth percentile of the population of S more than 5% of the time. This is counter to the requirement for an allowable value as described in the introduction.

10   Results and Discussions

10.1   Coverage Rates from MHK, Bayes, and RRM

In Tables III, IV, and V the coverage rate results are tabulated from the application of the MHK, Bayes, and RRM procedures, as functions of the coefficient of variations $(CV(i))$ for both the small sample S and the large sample L. The coverage rate represents the percent of values less than the 10% pt. (B allowable) or the 1% pt. (A allowable) of a population of values representing the data set. The data was obtained by randomly selecting values from either a normal or Weibull distribution with the specified CV's.

The mean and standard deviation are identified as: m(1) and s(1) for the larger sample L and m(2) and s(2) for the smaller sample S. The sample sizes are usually $n(1) = 30$ and $n(2) = 6$ for the large and small data sets respectively. CV(1) and CV(2) have similar representation for the two samples.

In Table III results from randomly selected values obtained from normal distributions with sample sizes of 30 and 6 show that for differences in the CV's less than 20% an acceptable coverage rate can be obtained from all methods since the rates are greater than 95%. The MHK and Bayes methods provide acceptable results even for a 60% difference in the CV values although they fail to obtain the desired 95% minimum. The RRM coverage results with 40% differences in CV's fail to provide acceptable coverage as shown in both the A and B allowable computation. The A allowables could not be computed using the Bayes method since an amount of data much greater than 36 would be required. The A allowable tables for u and M have not been computed because of the excessively large data set requirements. When $CV(1) = .12$ and $CV(2) = .10$, greater variability in L than S, the coverage is much greater than required, therefore, resulting in potentially over conservative estimates for the S allowable. This will usually be the case when $CV(1) > CV(2)$.

The MHK and Bayes methods' ability to provide acceptable coverage when the CV's are .16 for the small sample and .10 for the large sample shows that the methods are quite robust with respect to differences in the spread of the data sets. In actual engineering application it is unlikely that the material being considered in the design (small sample) would have a variability 60% greater than that of the previously tested, similar material (large sample).

In Table IV, the small sample data set was randomly selected from a Weibull distribution where the shape and scale values were computed so that they were equivalent to the tabulated mean and CV's. The larger data set was obtained from a normal distribution. The results are similar to those in Table III for the MHK and Bayes methods. The Table IV RRM results show a reduction in the coverage when compared with those in Table III, an example is the 78.8% coverage for the A allowable in Table III compared to 48% in Table IV for differences in the CV's of only 40%.

These results indicate that the RRM is sensitive to the statistical model assumption in representing the test data while the MHK and Bayes methods are much less sensitive. Since MHK and Bayes are nonparametric methods, this robustness to the model assumption could be expected.

In Table V data was obtained from normal distributions with CV's of .10 and .16 for L and S respectively. The coverage percent and range of allowable values are tabulated with respect to increasing sample sizes of both L and S for the RRM and MHK procedures. Results show that increasing sample size for L with constant small sample size for S of 6 causes the RRM process to perform poorly since the coverage is reduced from 86.6% to 73%. The only advantage is the reduction in the range of the allowable from 17 to 14 which is not very significant. Increasing the sample size of S from 6 to 15 also shows a somewhat unsatisfactory result since a 81% to 72.8% reduction in the coverage occurs. These coverage reductions are the inherent weakness in the method which is vulnerable to situations where L has a much smaller CV than S. The range reduction from 15 to 10 could be considered an improvement since there is less spread in the allowable estimate. Unfortunately, this advantage is

115

removed because of the coverage loss. This implies that many more (much greater then 5%) allowable values will be greater than the 10% pt. of the population of material strength measurements. *This situation could result in an overly optimistic allowable value and therefore a potential under-design situation.*

MHK results provide reasonably acceptable coverage for all the combinations of sample size for both L and S. That is, results show, at least for the cases considered, that the method is robust to a variety of sample sizes for both L and S. The range of the allowables is affected by the sample sizes particularly for the case MHK(15,6) vs. MHK(30,15). The MHK method can provide a smaller range on the allowable but will not make significant improvements on the coverage capability when the sample sizes are increased. In the results for MHK(60,6) and MHK(15,6) the coverage is 88% and 92.8% showing that increasing the sample size of L can reduce the coverage. This is the result of sample L's increased influence in the allowable computation which the analyst should be aware of when applying the MHK method. It is suggested that the ratio of sample sizes $n(2)/n(1)$ should not be any smaller than .2.

10.2 A Comparison Study: Single Sample Vs. Two-sample Allowable Computation

In Figures 2 through 5 a comparison is made between the multi-sample methods (MHK, RRM, and Bayes) and the single sample Weibull and normal methods with respect to the coverage percentage and the spread in the allowable estimates. In Figures 2 and 3 results were obtained by using a random selection of data from normal distributions. The N(6) and W(6) designations represent results from applying 6 data values to the Normal and Weibull allowable computation procedures. MHK(36) results are for the Modified Hanson-Koopmans method using a single sample with 36 data values from the S population distribution. CV's of .10 and .14 are introduced for L and S in order to represent a possible difference in the spread of the two data sets. The ordinate values ($\triangle$) shown in the figure represent the 95th percentile value of the allowable simulation results. Ideally, the values should be located on the dotted line for optimum coverage. Values above the line indicate that coverage has not been achieved. Those below the line provide the coverage. This can also identify an excessively low allowable value. In the second part of the figure the vertical dotted lines represent the spread in the allowable estimates (1 to 99 percent of all the data from the simulation results).

The Figure 2 results show that the MHK and Bayes methods can provide an almost optimum computed B allowable. The RRM approach fails to provide the coverage since results show an 87% rate. Normal distribution for single sample (S) of 6 provided reasonably good coverage as expected since the data was originally obtained from a normal model. The Weibull results were overly conservative, possibly, because an incorrect model was assumed for the data (normal). MHK(36) results were excellent as expected since the 36 values applied to the model were all from the normal distribution representing the data sample S.

116

Evaluation of the models' capabilities with respect to spread in the allowables showed the two-sample methods' allowable values to have much less variability than those of the single data set methods.

The results in Figure 3 are similar to those in Figure 2 except that the A allowable was computed. The Bayes method was omitted since a very large data set would have been required. A spread in excess of 50 was determined from applying the single sample normal analysis with the 1% point showing an allowable of -12. This result can discourage the engineers from using statistical procedures for obtaining design allowables. *In this case, the single sample method, although statistically correct, provides a design number that is incorrect from an engineering perspective.* This result has been the primary motivating factor in the authors' examination of alternate small sample procedures. The results from MHK and RRM show a more reliable range of values for the allowable.

In Figures 4 and 5 random samples were obtained from a NASA contractor report,[13] on composite material strength measurements. The figures identify the names of the companies that manufactured the material and the number of specimens tested. In Figure 4, the CV's of .10 and .13 were obtained from unidirectional tension and crossply tension data. The results show that the MHK and Bayes methods are effective in obtaining a desirable allowable estimate. The RRM results are greater than the 10% point and therefore fail to provide an acceptable allowable estimate. The normal and Weibull perform well in obtaining the proper coverage but as shown previously the spread in allowables for N(6) and W(6) is much greater than that of the MHK, Bayes, and RRM results.

In Figure 5, the random samples for both S and L were obtained from 230 data values (composite short beam shear test). The results are similar to those in Figure 4 except that the normal analysis, N(6), fails to provide acceptable coverage and the MHK and Bayes allowables are more conservative (excessive coverage). A relatively good agreement between the coverages can be identifed by comparing MHK(36) and MHK results. A reasonable correlation also exists for the spread in the allowable estimates. This implies that MHK can perform almost as well as if 36 values from S were applied to the MHK analysis instead of only 6 from S and 30 from L.

11 Conclusions

Results from this comparison study show that the nonparametric MHK method is superior in determining small sample design allowables when compared to the the results from the other procedures evaluated in this paper. The allowable values obtained from the MHK method application consistently meet the coverage requirement (95% of values less than a specified percentile of the population of all test data) for a relatively wide spectrum of data sets. The variability of the MHK values is much

---

[13] Reese, C. and Sorem, J. Jr., "Statistical Distribution of Mechanical Properties for Three Graphite-Epoxy Material Systems", NASA Contract Report No. 165736, 1981.

lower than that of the values resulting from the small, single sample normal or Weibull analysis.

The nonparametric Bayesian method provides acceptable allowable values although this method is limited by the sample size requirements. This limitation prevents the method from being as desirable as the MHK process. Another undesirable feature is the complexity involved in applying the method.

The Reduced Ratio Method, which is currently used by the aircraft industry, is not effective due to its failure in providing the required coverage when there are relatively small differences between the CV's of the prior large data set and the smaller empirical set from which the allowable is obtained. Also, increasing the sample size of empirical data and incorrectly assuming statistical models for the data sets prevents proper coverage.

Application of the small, single sample analysis (Normal and Weibull) results in extremely large variability in the allowable estimate. In addition, the methods fail to provide acceptable coverage when incorrect models are assumed.

The proposed pooling process introduced in this paper provides a desirable method for combining the small and large data sets when there is a difference in their mean values. Application of this process in the MHK and Bayesian analysis results in an effective solution in obtaining economical allowable values.

Figure 1A

n = 10

Population

Design Allowable
probability density
function

95 %

Population 10% pt.

Figure 1B

n = 50

Population

Design Allowable
probability density
function

95 %
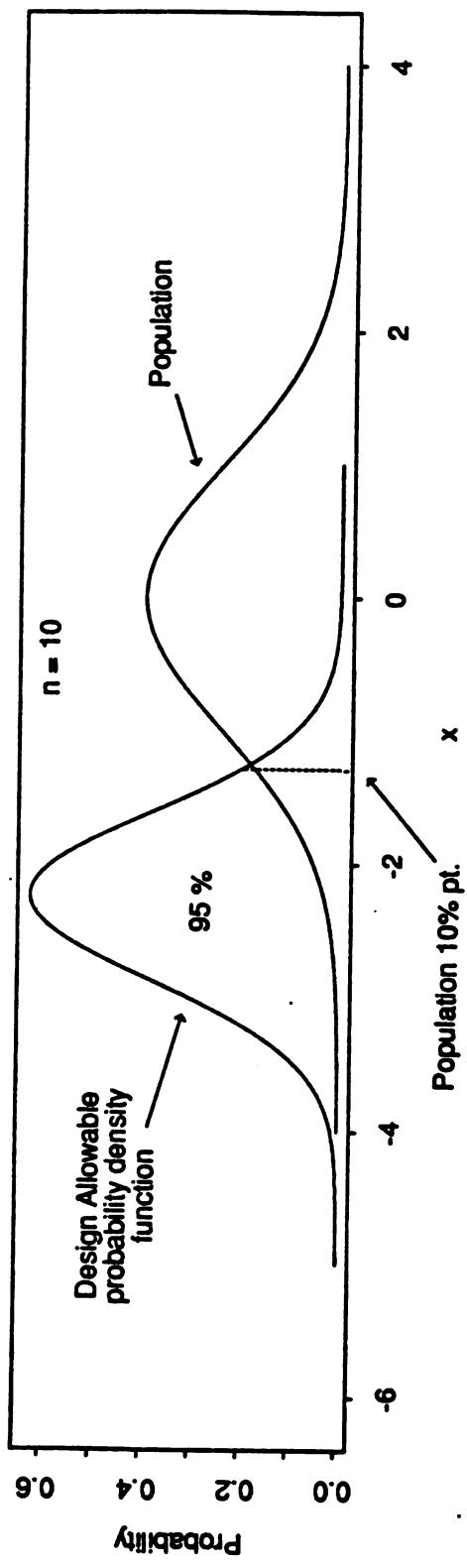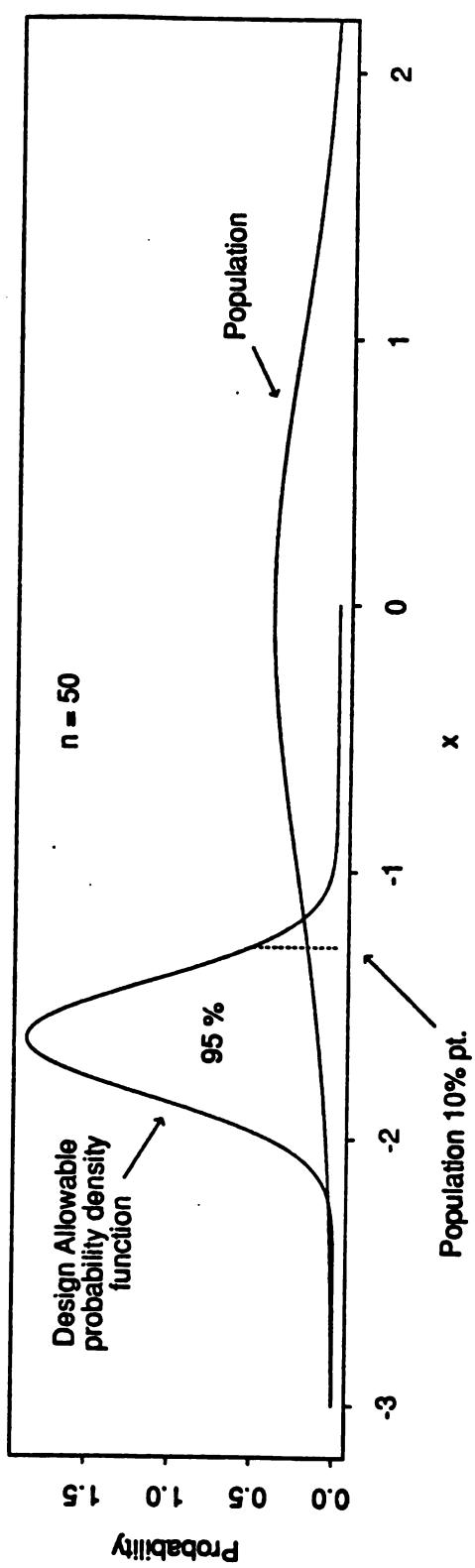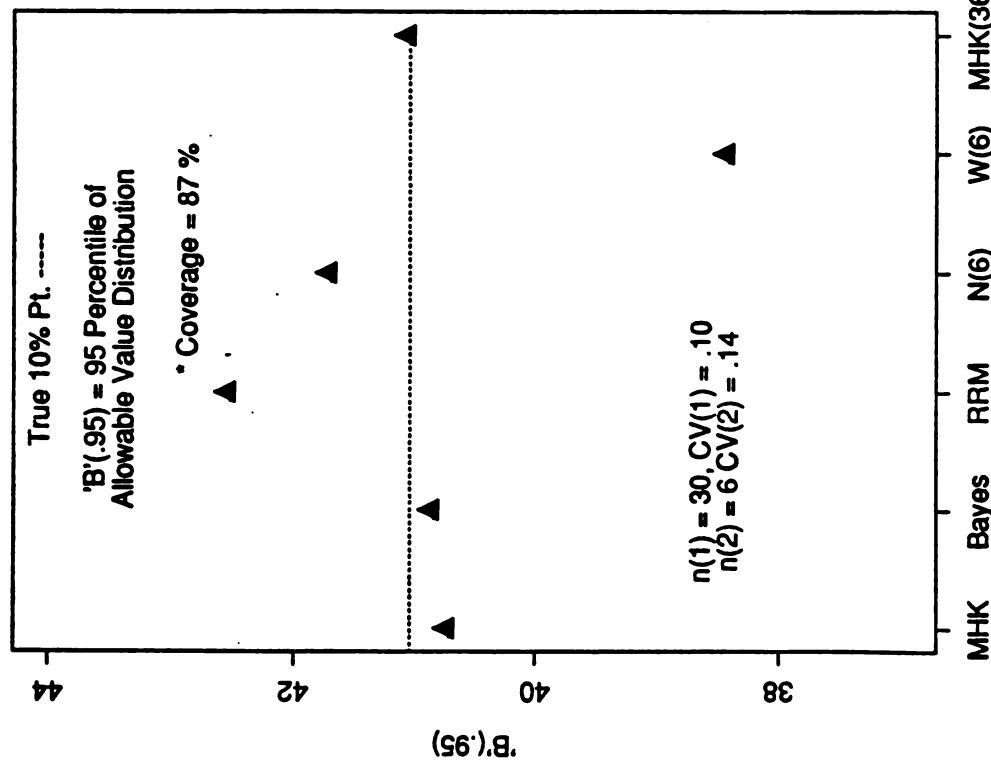
Population 10% pt.

Figure 1. Tolerance Limit for N(0,1) Population

Digitized by Google
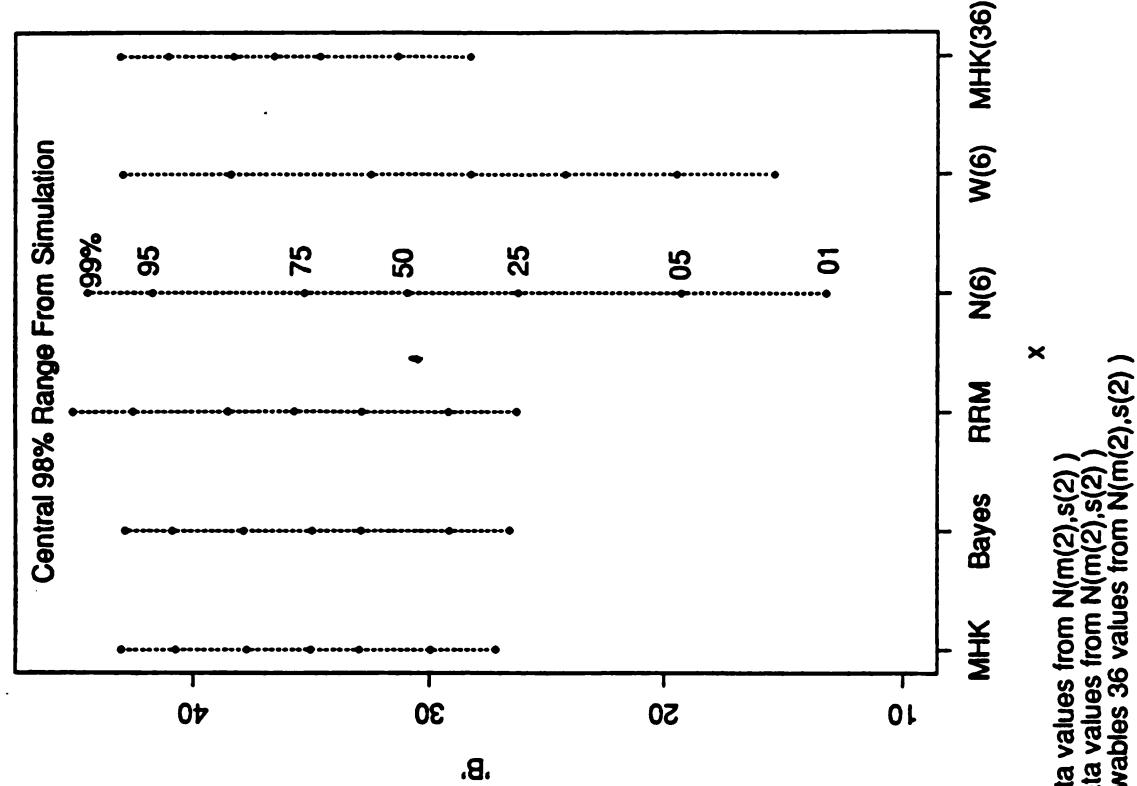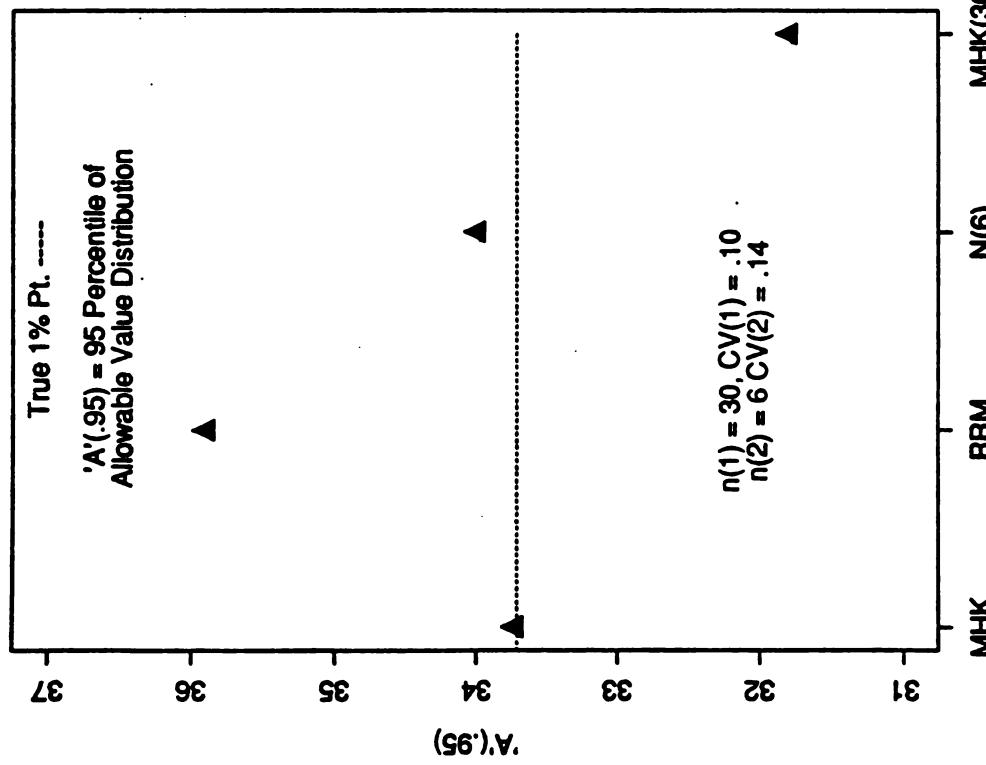
Figure 2. Monte Carlo Study - Allowables Versus Coverage/Range of Allowables
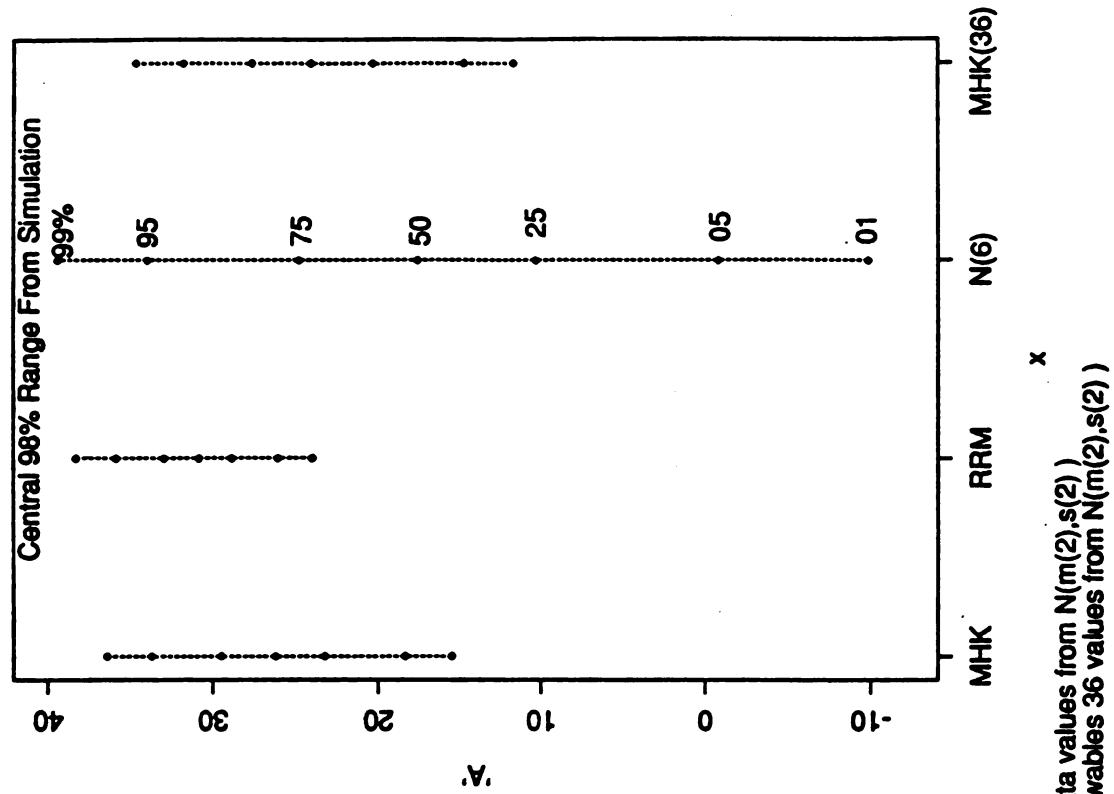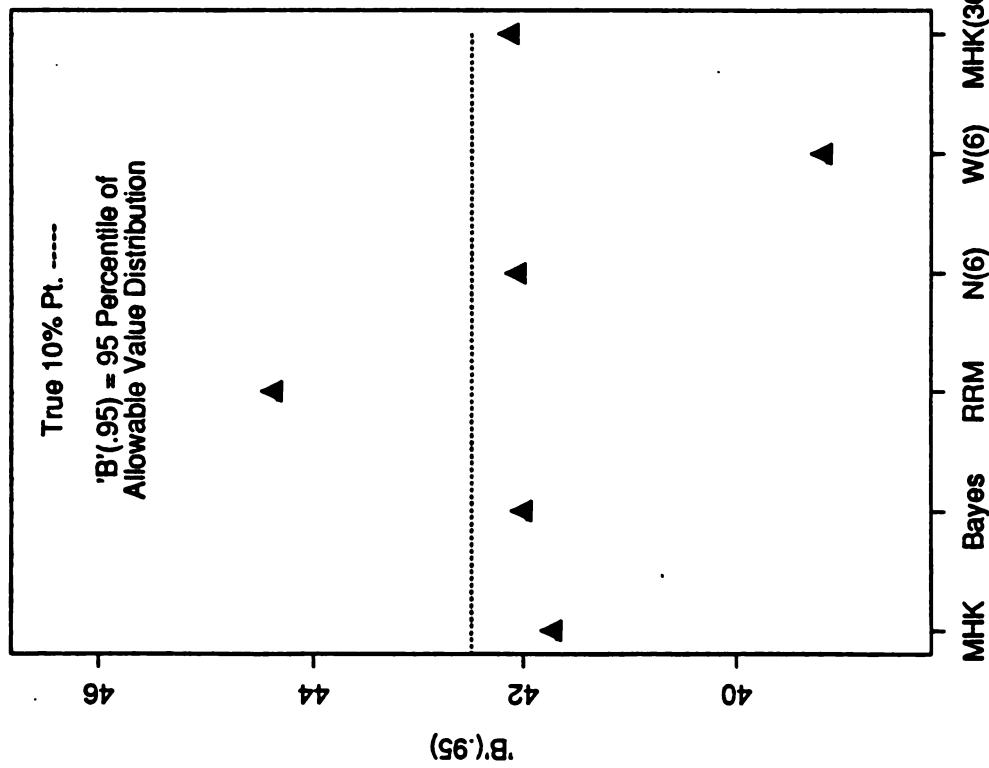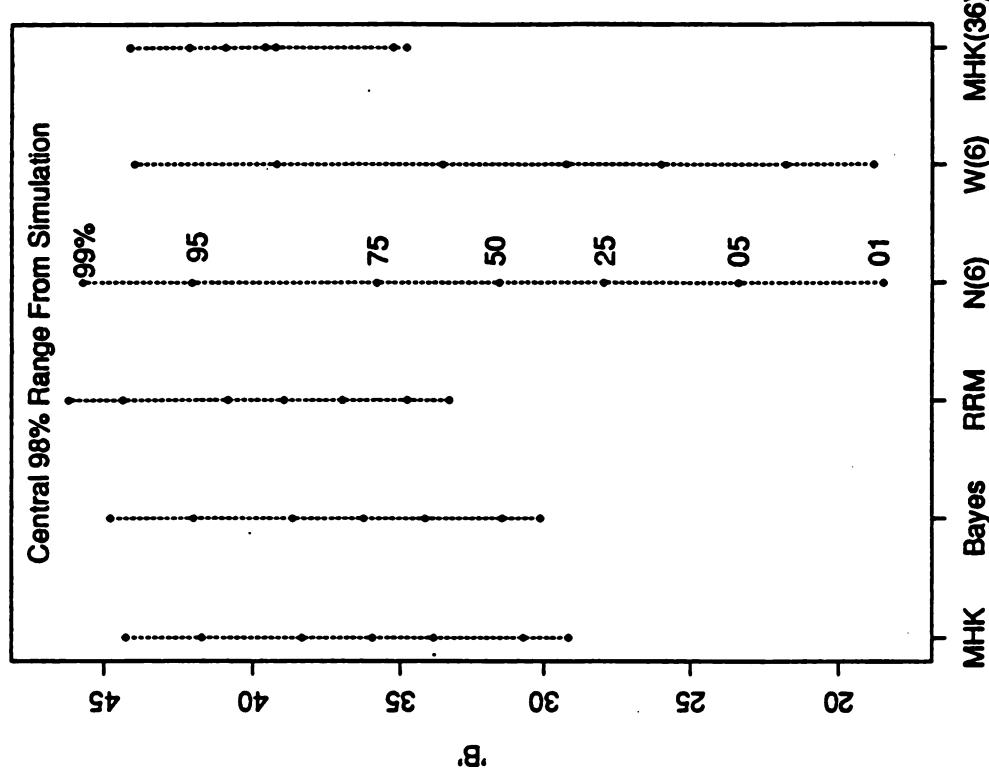
Figure 3. Allowables Versus Coverage/Range of Allowables
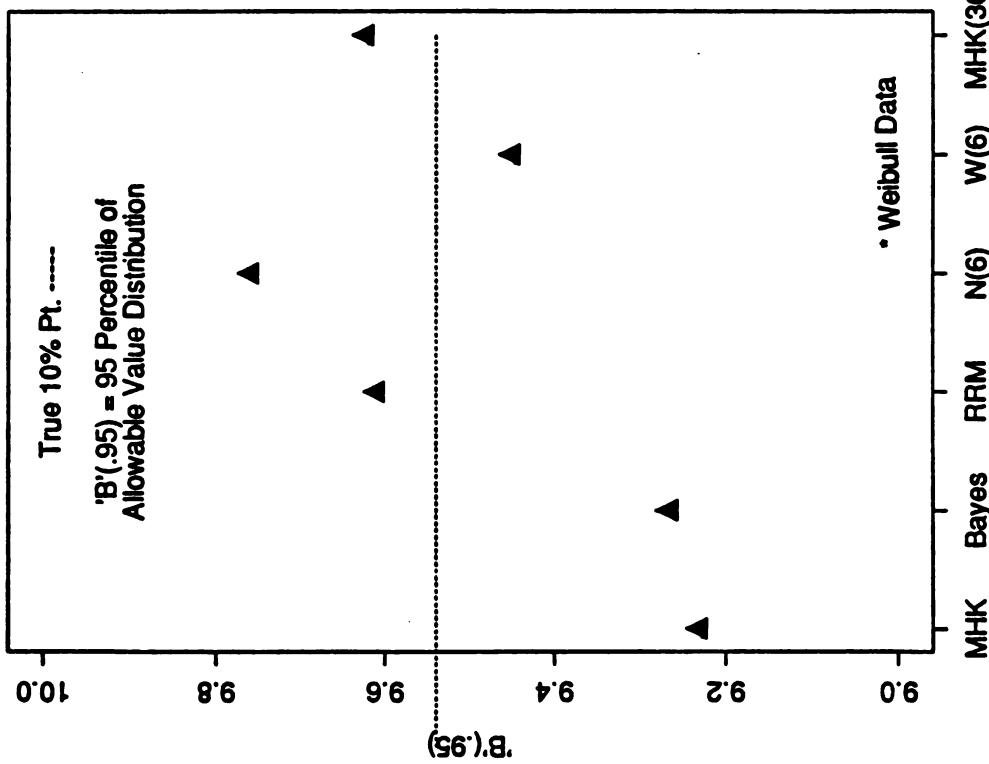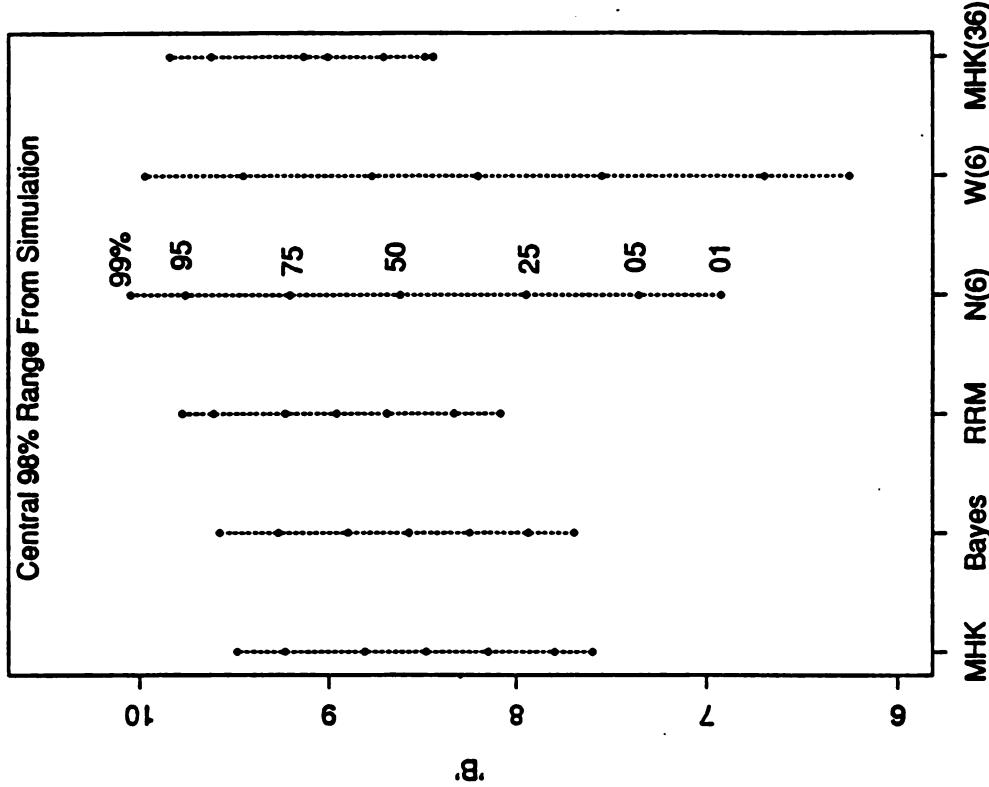
121

Figure 4. Allowables Versus Coverage/Range of Allowables

Figure 5. Allowables Versus Coverage/Range of Allowables

Table I. M and u Values for Bayesian Basis Value Computation

M	u	M	u	M	u	M	u
1	0.021953	51	0.044804	101	0.057686	151	0.064302
2	0.017855	52	0.045192	102	0.057856	152	0.064395
3	0.016529	53	0.045565	103	0.058023	153	0.064514
4	0.016140	54	0.045937	104	0.058188	154	0.064609
5	0.016199	55	0.046301	105	0.058352	155	0.064717
6	0.016516	56	0.046648	106	0.058517	156	0.064814
7	0.016997	57	0.046996	107	0.058670	157	0.064912
8	0.017590	58	0.047339	108	0.058837	158	0.065010
9	0.018264	59	0.047673	109	0.059006	159	0.065099
10	0.018996	60	0.048011	110	0.059156	160	0.065193
11	0.019769	61	0.048318	111	0.059313	161	0.065273
12	0.020570	62	0.048642	112	0.059454	162	0.065382
13	0.021391	63	0.048945	113	0.059619	163	0.065462
14	0.022223	64	0.049255	114	0.059761	164	0.065555
15	0.023060	65	0.049563	115	0.059914	165	0.065658
16	0.023897	66	0.049848	116	0.060051	166	0.065734
17	0.024729	67	0.050144	117	0.060192	167	0.065822
18	0.025554	68	0.050421	118	0.060344	168	0.065910
19	0.026368	69	0.050695	119	0.060480	169	0.065996
20	0.027171	70	0.050968	120	0.060628	170	0.066108
21	0.027959	71	0.051238	121	0.060754	171	0.066192
22	0.028734	72	0.051506	122	0.060883	172	0.066277
23	0.029491	73	0.051771	123	0.061031	173	0.066384
24	0.030233	74	0.052034	124	0.061162	174	0.066449
25	0.030959	75	0.052284	125	0.061292	175	0.066530
26	0.031666	76	0.052530	126	0.061420	176	0.066613
27	0.032361	77	0.052773	127	0.061547	177	0.066705
28	0.033033	78	0.053017	128	0.061679	178	0.066789
29	0.033695	79	0.053244	129	0.061802	179	0.066872
30	0.034339	80	0.053479	130	0.061933	180	0.066934
31	0.034967	81	0.053702	131	0.062065	181	0.067007
32	0.035577	82	0.053932	132	0.062179	182	0.067098
33	0.036172	83	0.054160	133	0.062293	183	0.067176
34	0.036754	84	0.054375	134	0.062430	184	0.067258
35	0.037328	85	0.054600	135	0.062553	185	0.067333
36	0.037884	86	0.054808	136	0.062667	186	0.067418
37	0.038420	87	0.055017	137	0.062784	187	0.067486
38	0.038952	88	0.055221	138	0.062894	188	0.067569
39	0.039461	89	0.055435	139	0.063010	189	0.067628
40	0.039964	90	0.055634	140	0.063128	190	0.067720
41	0.040459	91	0.055831	141	0.063245	191	0.067794
42	0.040944	92	0.056024	142	0.063344	192	0.067871
43	0.041409	93	0.056215	143	0.063459	193	0.067952
44	0.041864	94	0.056417	144	0.063550	194	0.068022
45	0.042314	95	0.056599	145	0.063666	195	0.068103
46	0.042751	96	0.056781	146	0.063763	196	0.068178
47	0.043182	97	0.056960	147	0.063899	197	0.068237
48	0.043596	98	0.057153	148	0.063985	198	0.068315
49	0.044009	99	0.057332	149	0.064101	199	0.068388
50	0.044413	100	0.057502	150	0.064197	200	0.068459

Table II.   Modified Hanson-Knoopmans Constants for Basis Value

n	r	s	k
2	1	2	35.177
3	1	3	7.859
4	1	4	4.505
5	1	4	4.101
6	1	5	3.064
7	1	5	2.858
8	1	6	2.382
9	1	6	2.253
10	1	6	2.137
11	1	7	1.897
12	1	7	1.814
13	1	7	1.738
14	1	8	1.599
15	1	8	1.540
16	1	8	1.485
17	1	8	1.434
18	1	9	1.354
19	1	9	1.311
20	1	10	1.253
21	1	10	1.218
22	1	10	1.184
23	1	11	1.143
24	1	11	1.114
25	1	11	1.087
26	1	11	1.060
27	1	11	1.035
28	1	12	1.010
29	1	--	1
30	2	12	1.373
31	2	12	1.344
32	2	12	1.315
33	2	13	1.270
34	2	13	1.245
35	2	13	1.221
36	2	13	1.197
37	2	13	1.174
38	2	13	1.151
39	2	13	1.129
40	2	13	1.108
41	2	14	1.083
42	2	14	1.064
43	2	14	1.045
44	2	14	1.027
45	2	14	1.009
46	2	--	1

Table III. Simulation Results/Computing Allowable Value
Coverage Rate (%) Versus CV Differences Normal - Normal Distributions

CV		Coverage Rate (%)				
		'B' Allowables			'A' Allowables	
CV(1)	CV(2)	MHK	Bayes	RRM	MHK	RRM
.10	.10	99.0	99.2	98.6	99.4	99.0
.10	.12	97.0	98.0	94.8	98.4	94.4
.10	.14	95.4 (94.6)*	95.6 (93.8)*	86.6 (84.2)*	95.8 (92.8)*	78.8 (72.8)*
.10	.16	91.8	92.6	81.0	89.4	59.4
.10	.18	88.2	89.2	72.2	83.4	41.2
.10	.20	83.2	83.0	64.2	72.6	24.6
.12	.10	99.8	99.8	99.6	99.8	100

CV(i) = s(i)/m(i) , i = 1,2   m(1) = 200 , m(2) = 50

Assumed Distributions are N(m(1),s(1)), N(m(2),s(2)) = Normal distribution
for prior and current data sets respectively

MHK - Modified Hanson-Koopmans
Bayes - Nonparametric Bayes (Ferguson)
RRM - Reduced Ratio Method (Mil-5)

Sample size n(1) = 30(prior), n(2) = 6(data) for cases except ( )*
* sample size n(1) = 60, n(2) = 6

Table IV. Simulation Results/Computing Allowable Values
Coverage Rate (%) Versus CV Differences Normal - Weibull Distributions

CV		Coverage Rate (%)				
		'B' Allowables			'A' Allowables	
CV(1)	CV(2)	MHK	Bayes	RRM	MHK	RRM
.10	.10	98.6	99.2	97.8	99.6	88.6
.10	.12	98.0	98.4	90.8	98.4	68.6
.10	.14	94.0	94.2	82.2	94.4	48.0
		(94.2)*	(------)*	(90.6)*	(96.0)*	(63.6)*
.10	.16	89.0	89.6	73.4	89.2	29.0
.10	.18	84.8	86.8	65.0	82.0	19.6
.10	.20	76.0	76.6	57.4	69.0	14.0
.12	.10	99.8	99.8	99.4	99.6	98.2

CV(i) = s(i)/m(i) , i = 1,2   m(1) = 200 , m(2) = 50

Distributions N(m(1),s(1)), W(a(2),b(2))
where N and W are Normal and Weibull models
for prior and current data sets respectively
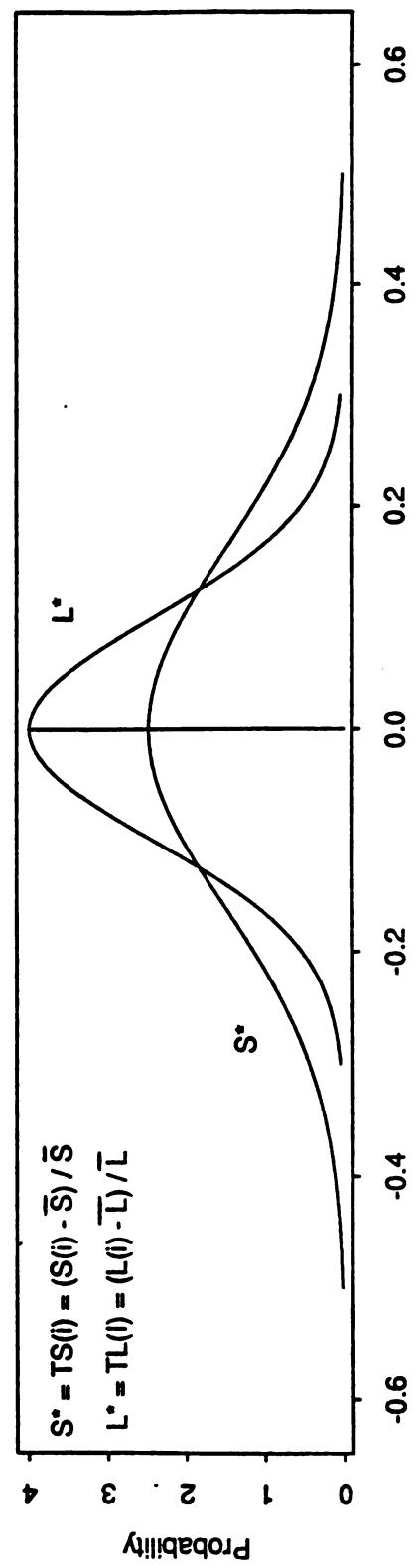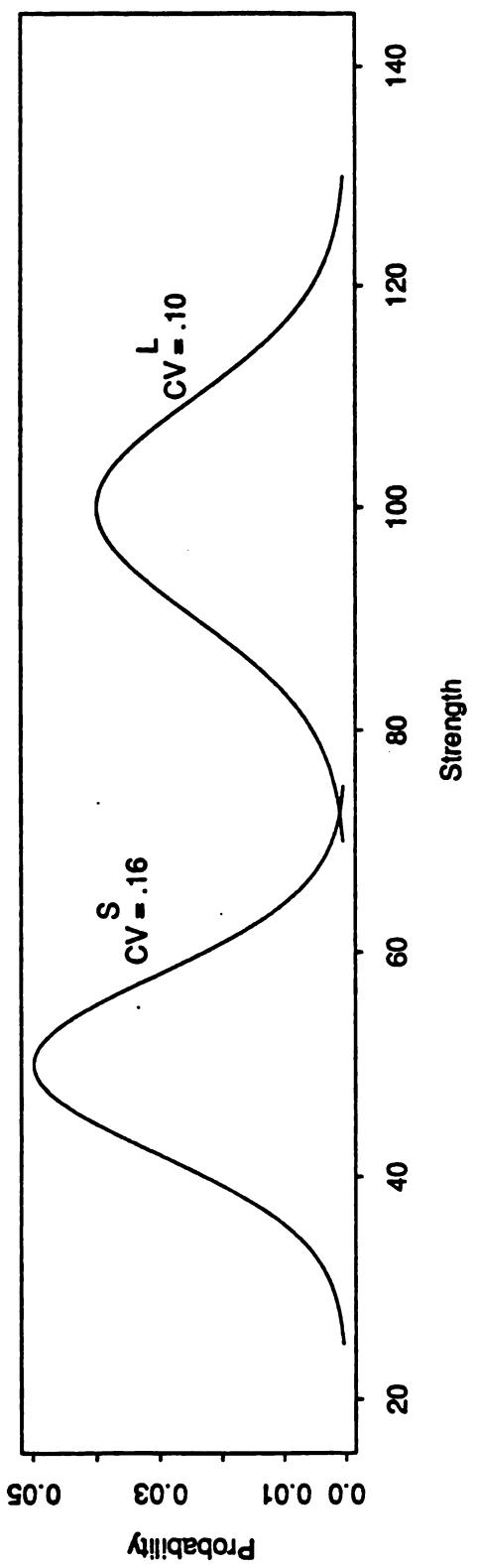
a(2) = shape parameter and b(2) = scale
determined for prescribed CV in columns 1 and 2

* sample size n(1) = 15, n(2) = 6
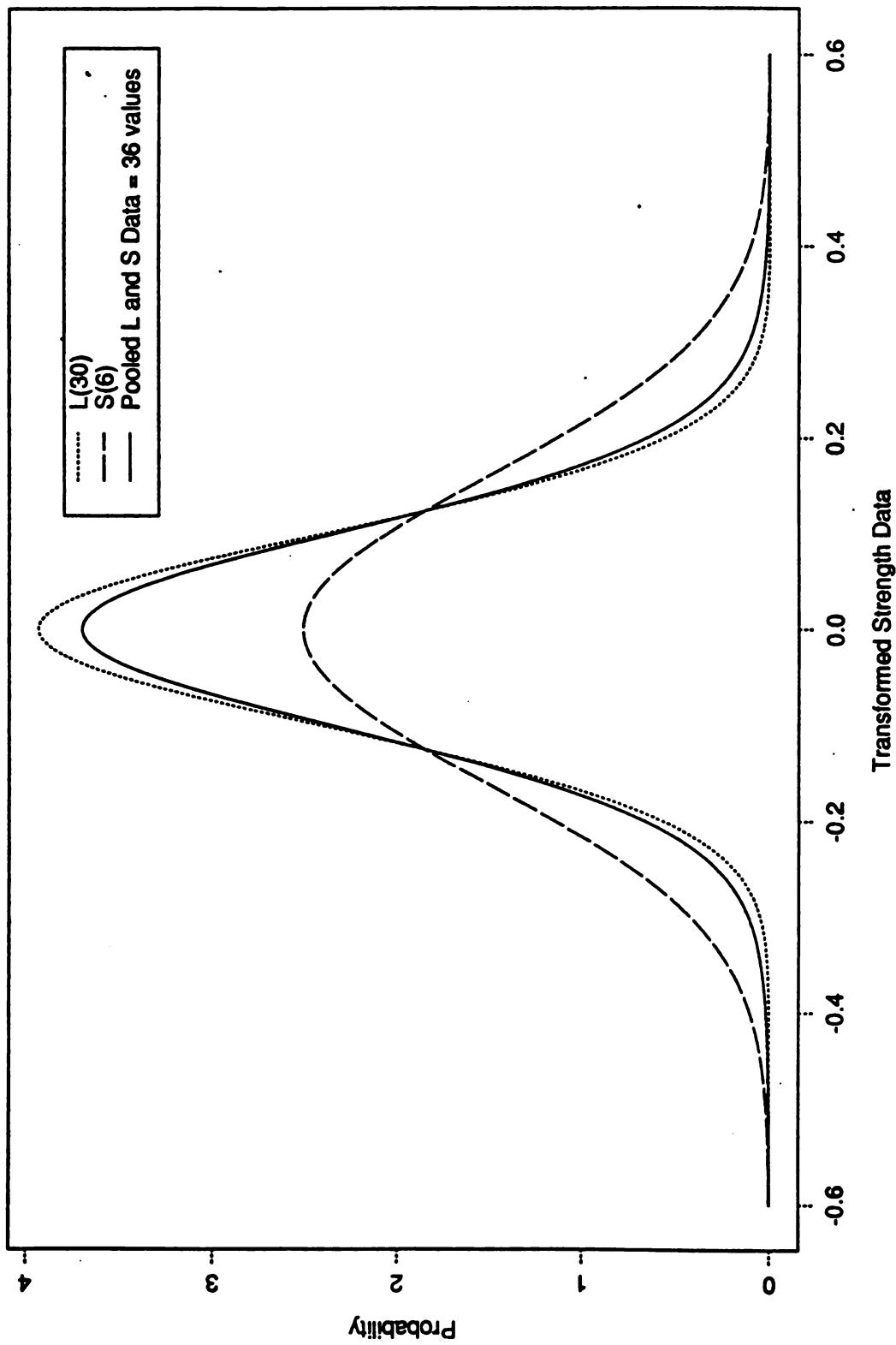
Table V. Range and Coverage (%) Versus Sample Size/Methods Normal Distributions

Method	Range (%) of 'B' Allowable			Coverage (%)
(n(1), n(2))	01	50	99	'B' Allowable
RRM (15,6)	27.19	35.55	44.28	86.6
RRM (30,6)	29.10	36.80	44.58	81.0
RRM (60,6)	30.32	37.52	44.70	73.0
RRM (30,15)	33.42	38.23	43.84	72.8
MHK (15,6)	20.49	32.95	42.87	92.8
MHK (30,6)	25.12	34.51	43.25	91.8
MHK (60,6)	27.40	35.19	42.98	88.4
MHK (30,15)	29.06	36.17	42.07	90.0
n(1) = L sample size CV(1) = .10				
n(2) = S sample size CV(2) = .16				

Appendix A. Transformation Process for Pooling PDF's of Small Data Set(S) and Large Data Set(L)

129

Appendix B. PDF's for Both Pooled and Single Data Sheets

Legend:
- L(30)
- S(6)
- Pooled L and S Data = 36 values

X-axis: Transformed Strength Data
Y-axis: Probability

THE PRINCIPLES OF SKIP-LOT SAMPLING AND A COMPARISON OF SAMPLING FREQUENCY OPTIONS

Charles J. Hunter
Defence Scientist, Operational Research and
Analysis Establishment, Ottawa, K1A 0K2

and

Jay M.H. Adamsson
Graduate Student, Department of Mathematics,
University of Ottawa

INTRODUCTION

1.   The American, British, Canadian and Australian (ABCA) Armies, plus New Zealand in an observatory role, jointly participate in a standardization programme on proofing, inspection, and quality assurance (PIQA). This quadripartite working group (QWG) is mandated to investigate areas where opportunities for standardization exist and develop appropriate standardized procedures for implementing specific PIQA techniques in the field.

2.   A large component of these areas of investigation is the application of statistical sampling methods for inspection purposes. Skip-lot sampling is one statistical technique currently being examined under project number QA/29 "Skip-Lot Sampling." A quadripartite advisory paper (QAP 28) has recently been published that explains the principles of this technique [1].

3.   The purpose of the present report is two-fold. The first section briefly introduces the principles of skip-lot sampling by means of summarizing the contents of QAP 28. This will include identifying the conditions under which implementing a skip-lot sampling plan may be warranted or beneficial.

4.   The second part focuses on the comparison of three different sampling frequency options that could be used during the skipping phase of any skip-lot sampling plan. The quality assurance properties of each sampling frequency option are derived by utilizing the simplified Markov Chain approach developed by Brugger [2 and 3]. Some advice is provided regarding which option is most appropriate under economical constraints or when one wishes to protect against quality deterioration.

131

OVERVIEW OF SKIP-LOT SAMPLING[1]

5.   Skip-lot sampling inspection can be defined as "an acceptance sampling procedure in which some lots in a series are accepted without inspection, when the sampling results for a stated number of immediately preceding lots meet stated criteria."[2] Interpreting this definition in a less formal sense, a skip-lot sampling plan involves sampling some lots according to established procedures at a specified sampling frequency level, while other lots are accepted without inspection.

6.   Herein lies both the purpose of skip-lot sampling and its disadvantages. There are various ways of reducing the inspection effort on products where a demonstrated high quality is being maintained by the supplier. At one extreme, sampling could be discontinued altogether. This is not recommended for the obvious reason that a deterioration in process quality could remain undetected for some time. Another option is to reduce the sample size of items inspected within a lot, which is the method recognized in ABCA Quadripartite Standardization Agreements (QSTAGs) 105 and 330. Skip-lot sampling offers yet another option which is applied to the lots as opposed to the individual items. This is not to say that a skip-lot sampling plan must be applied in place of reduced inspection when it is deemed more cost-effective.[3] Because one is applying these at different stages on the inspection process it is entirely possible to overlay a skip-lot sampling plan over a variety of inspection plans that are applied to the individual items. This is analogous to the commonly used statistical technique known as two-stage sampling since the sample drawn takes place in two steps.[4] Reference [6] indicates that for two-stage sampling, any type of sampling can be employed at each step and combined to form an overall sampling plan.

7.   Implementing a skip-lot plan also adds another degree of risk, since an occasional bad lot might be accepted without inspection. As will be illustrated later when the construction of skip-lot plans are discussed, this risk can be controlled through close monitoring of the production process,

---

[1]   This section draws heavily from the contents of reference 1.

[2]   Proposed International Standards Organization (ISO) definition, reference 4.

[3]   p. 252 reference 5.

[4]   p. 274 reference 6.

plus the inclusion of intermediate retrial stages.

8.   Skip-lot sampling is beneficial when the lot inspection or testing is destructive or costly. Also, if availability of inspection personnel or test equipment is limited, skip-lot sampling may be useful. There are three conditions which must be satisfied before skip-lot sampling is applied to a product:

     a.   the product must be grouped into identifiable lots as it is being produced or presented for inspection;

     b.   production must be stable, to assure a homogeneous series of lots; and

     c.   quality must be high, as demonstrated by previous history of the item.

9.   One of course should always investigate whether there are any incremental administration complications or costs associated with incorporating a skip-lot plan, and ensure that they are less than the benefits to be garnered by skip-lotting.
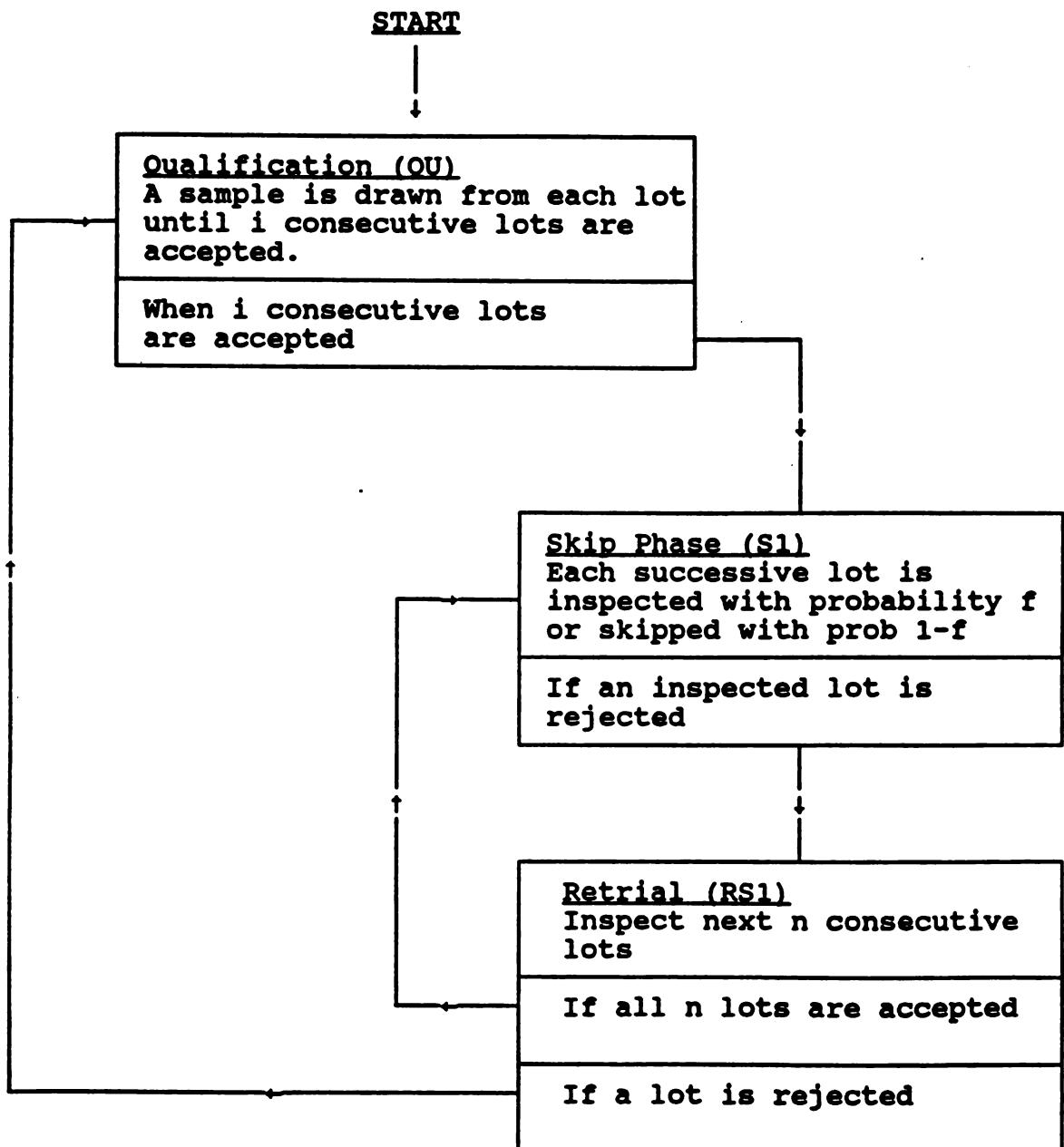
CONSTRUCTION OF SKIP-LOT PLANS

10.   The most convenient way to understand skip-lot sampling is by means of illustration. Figure 1 shows a skip-lot plan in its simplest form, a qualification stage and one level of sampling. The process begins at the qualification phase and one remains in this phase until i consecutive lots have been accepted. When this condition is satisfied, one switches to the skipping phase.

11.   For this example, there is only one level of sampling in the skipping phase, hence one either samples each successive lot with probability f, or the lot is skipped (accepted without inspection) with probability $1-f$.

12.   There are some key features in Figure 1 that should be part of any skip-lot plan. First, the process should always begin with 100% inspection of lots. This ensures that the product has a suitably high quality history for skip-lotting to be applied. Next, the selection of lots to be sampled during the skip phase should be done at random with probability f of being selected.

FIGURE 1

SIMPLE SKIP-LOT PLAN

START

```
┌─────────────────────────────────────┐
│ Qualification (QU)                   │
│ A sample is drawn from each lot      │
│ until i consecutive lots are         │
│ accepted.                            │
├─────────────────────────────────────┤
│ When i consecutive lots              │
│ are accepted                         │
└─────────────────────────────────────┘

            ┌─────────────────────────────────────┐
            │ Skip Phase (S1)                      │
            │ Each successive lot is               │
            │ inspected with probability f         │
            │ or skipped with prob 1-f             │
            ├─────────────────────────────────────┤
            │ If an inspected lot is               │
            │ rejected                             │
            └─────────────────────────────────────┘

            ┌─────────────────────────────────────┐
            │ Retrial (RS1)                        │
            │ Inspect next n consecutive           │
            │ lots                                 │
            ├─────────────────────────────────────┤
            │ If all n lots are accepted           │
            ├─────────────────────────────────────┤
            │ If a lot is rejected                 │
            └─────────────────────────────────────┘
```

134

13.   Perhaps most important, one should determine how responsive the process must be to an adverse shift in lot quality. A properly designed plan should respond to an adverse shift by resuming 100% lot inspection. The idea of delaying the decision of changing inspection phases is usually added to a skip-lot plan to protect against a premature action of reinstituting 100% inspection and requalification of the product when only a single rejection occurs, as shown in Figure 1.

14.   The intermediate step acts as a checkpoint to scrutinize the next few lots (based on experience, a minimum of four lots are recommended) to determine whether the lot quality has experienced an abrupt change or not. If during this step no additional lots are rejected the process returns to the skipping phase. If however, the process fails to meet the criterion for returning to the skip phase, 100% inspection resumes at the qualification phase and the process of requalifying by accepting i consecutive lots starts over.

15.   In practice one needs to consider several factors to determine how responsive the process should be. Some of the factors are:

        a.   The cost of nonconforming lots passing through the
             process without inspection and onto the consumer;

        b.   The cost of inspecting a lot; and

        c.   The likelihood of significant sudden change in
             quality occurring.

16.   Assessing the weight or importance of each of the above factors will assist the designer of a skip-lot plan in deciding whether they require responsive action in the form of immediate resumption of 100% inspection, an increase in sampling frequency, or postponing a decision by performing an intermediate inspection step.

17.   There are obviously many possible combinations of designing responsiveness into a skip-lot plan. If one is unsure about the consequences an abrupt shift in lot quality would have, they should consult a statistician or quality engineer that is familiar with acceptance sampling methods.

QUALITY ASSURANCE CHARACTERISTICS

18.   There are three quality assurance characteristics that are normally used to evaluate the effectiveness of lot sampling plans. Each of these statistics can be graphed

135

against the expected quality of the materiel entering the sampling procedure. The resulting graphs can be used as guidelines in designing an effective skip-lot plan, or in evaluating a plan already in operation.

19.   The first of these statistics is the average fraction inspected (AFI). This is the expected percentage of lots which would be chosen for inspection for a given lot quality. Figure 2 gives a graph of AFI vs lot quality for the sampling plan in Figure 1 with $i = 10$, $f = 0.5$ and $n = 4$. If there was no skip phase the AFI would equal one. As can be seen in the graph, the skip-lot sampling plan reduces the number of inspected lots for a high incoming lot quality, but it increases rapidly once the lot quality deteriorates. A procedure for calculating the AFI is covered later in paragraph 27.

20.   The other two statistics are dependent on the AFI and the incoming lot quality. The average output quality (AOQ) is calculated here under the assumption that a lot which is inspected and found to be non-conforming is replaced by a lot which conforms to the desired quality. As such, a number of non-conforming lots are inspected and removed, thereby improving the outgoing lot quality. The graph in Figure 3 shows the output lot quality. The AOQ is expected to remain high for the following reason. When incoming lot quality is low, the AFI is high. Thus, a large number of non-conforming lots are inspected and replaced, thereby raising the AOQ. This statistic is used mainly to show that outgoing lot quality remains high under skip-lot sampling even when incoming lot quality decreases. This would be important in processes where a high outgoing quality is essential. Given the incoming lot quality, p, and the AFI, the AOQ is calculated by:

$$AOQ = p + AFI(1-p) \tag{1}$$

21.   The third statistic is the operating characteristic (OC) curve. This curve gives the probability of a random lot being accepted under this procedure. A lot can be accepted in one of two ways. Either it can pass through the system without being inspected, in which case both conforming and non-conforming lots are accepted, or it can be chosen for inspection and passed. In the latter instance, any non-conforming lots are detected and removed. Figure 4 gives the OC curve for the system in Figure 1. This curve is useful in that it gives the expected percentage of lots that will be accepted for a given incoming lot quality. The separation between the diagonal

Figure 2

AVERAGE FRACTION INSPECTED

137
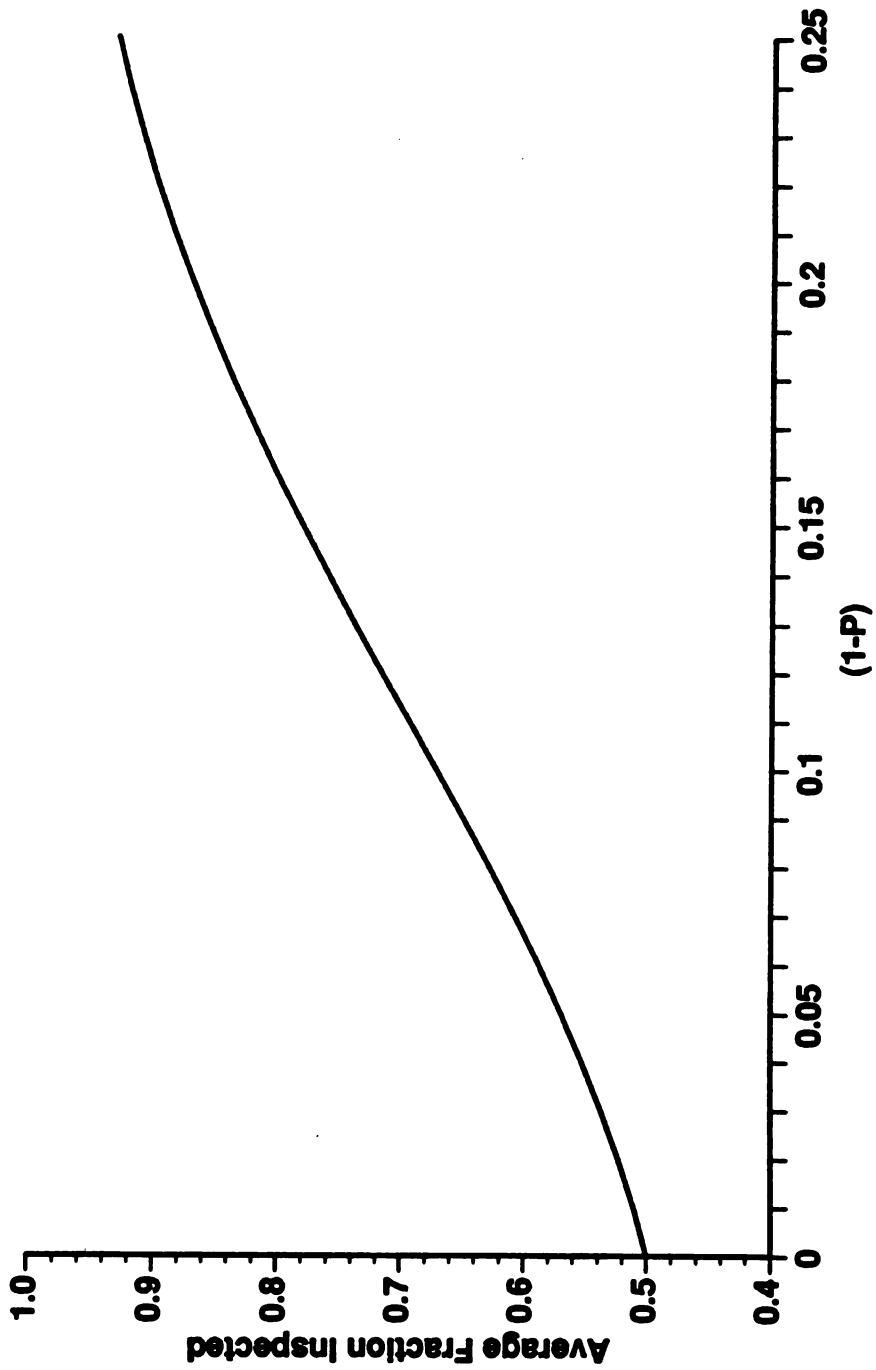
Figure 3

AVERAGE OUTPUT QUALITY

138
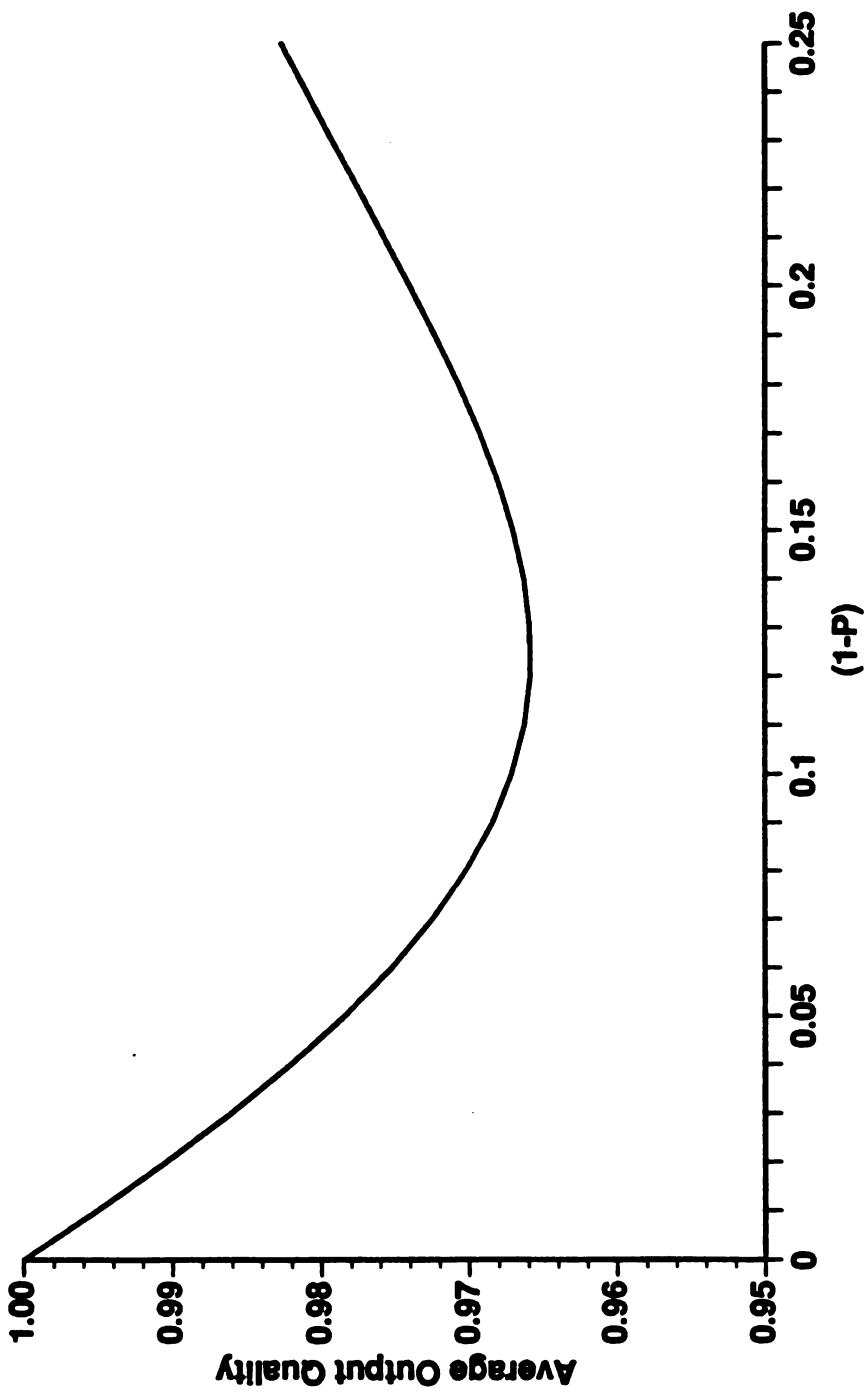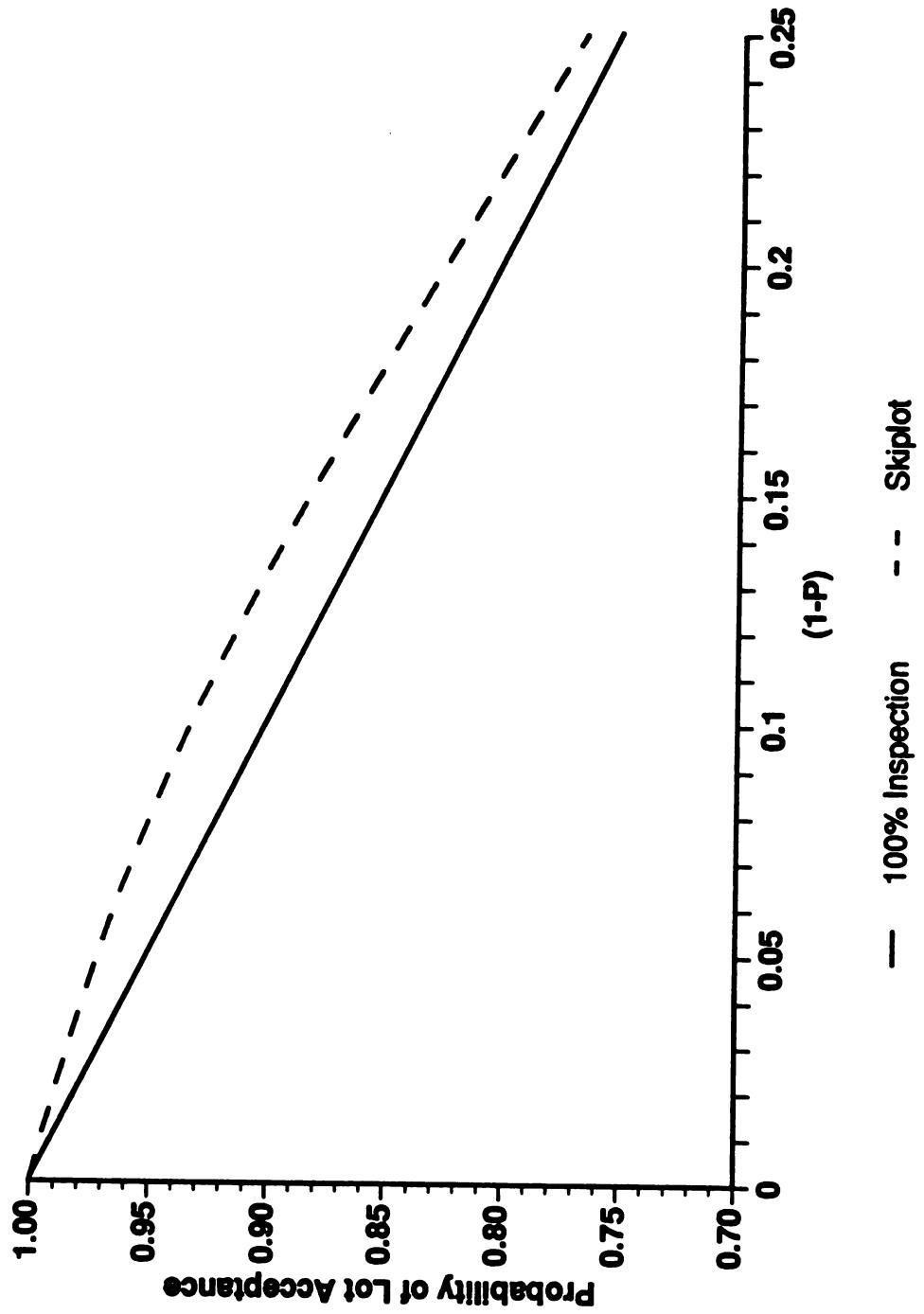
Figure 4

OPERATING CHARACTERISTIC CURVE

— 100% Inspection   - - Skiplot

line and the skip-lot curve gives an indication of the additional risk one takes of accepting lots as the lot quality decreases. The OC is calculated by:

$$OC = 1 - AFI(1-p)$$ (2)

22.   The AFI curve can be used to study the effect which each of the parameters has on the overall procedure. The AFI curve is the best choice of the three in most cases since the other two statistics are dependant on the AFI. By comparing the AFI curves for different choices of parameters, a number of different scenarios for skip-lot sampling can be analyzed and the best choice used in the design of a plan.

23.   In order to calculate the quality assurance curves, one must first create the transition probability matrix for a skip-lot sampling plan. Table 1 gives the transition matrix for the sample plan in Figure 1. The entries in the matrix represent the probability of traversing from one stage in the plan to the next. Note, from the qualification phase (QU) one can only move to the sampling step, hence there is only one entry in the QU row. Similarly, from S1 one can only transfer to the retrial step (RS1). There are two possibilities from the RS1 step. One could return to S1 with probability p^x, or transfer back to the QU phase with probability $1-p^x$.

TABLE 1

TRANSITION MATRIX

	QU	S1	RS1
QU	--	1	--
S1	--	--	1
RS1	$1-p^x$	p^x	--

24.   In general interpreting the transition matrix can be accomplished by reading across the rows for the step or phase which one is exiting, and down the column of the step which one is entering.

25.   From the transition matrix, a simplified Markov chain approach can be used to calculate the steady state probabilities. First, the number of times which the process

140

enters each stage, represented by T(S), can be calculated in terms of the qualification stage as follows:

$$T(QU) = T(QU) \tag{3}$$
$$T(S1) = T(QU) + p^x T(RS1)$$
$$T(RS1) = T(S1)$$

solving in terms of T(QU):

$$T(S1) = T(QU) + p^x T(S1)$$
$$T(S1) = T(QU) / (1-p^x) \tag{4}$$

and

$$T(RS1) = T(S1) = T(QU) / (1-p^x) \tag{5}$$

26.   A working table can be created once these calculations are completed, shown as Table 2. Column 1 of the table gives the relative number of times the process enters each stage. These terms are given as the coefficients of T(QU) in equations (3) - (5) above. Column 2 is a simplification of the terms in column 1.

27.   The terms in the first two columns only give the number of times entering each phase. Column 3 gives the expected number of lots in each phase, and column 4 is a simplification of column 3. Hence, column 4 is the relative number of lots expected each time the process enters the phase. By multiplying column 2 by column 4, we get the relative total number of lots in each phase. This is recorded in column 5. The steady state probabilities in terms of lots to be inspected during each phase is then obtained by taking each entry in column 5, and dividing it by the sum of the values from column 5 (denoted by D). The AFI is calculated from column 5 of the working table by multiplying each term in the column by the corresponding sampling frequency for that step, then summing these terms together and divide by D to obtain the AFI.

141

TABLE 2

WORKING TABLE

Phase	(1) Coefficient of T	(2) Simplification of (1)	(3)* Expected # of lots in phase	(4) Simplification of (3)	(5) column (2) x column (4)
QU	1	$1-p^x$	$(1-p^l)/p^l(1-p)$	$f(1-p^l)$	$f(1-p^x)(1-p^l)$
S1	$1/(1-p^x)$	1	$1/f(1-p)$	p^l	p^l
RS1	$1/(1-p^x)$	1	n^{**}	$nfp^l(1-p)$	$nfp^l(1-p)$

D   = Sum of terms in column (5)
        = $f(1-p^x)(1-p^l) + p^l + nfp^l(1-p)$

*   The derivation of the form of the expected values in this column are given in references [7] and [8]

**   If inspection in this step is done sequentially with the action of returning to the qualification step occurring immediately upon a lot being rejected the expected value would be $(1-p^n)/(1-P)$

142

28. If specific values are given for i, f, and x (ie: 10, .5, and 4 respectively), column(5) would look like:

TABLE 3

SAMPLE STEADY STATE PROBABILITIES

	Steady State Probabilities for	
	$p=0.90$	$p=0.95$
Phase		
Column (5) from Table 2		
QU	0.211	0.053
S1	0.657	0.860
RS1	0.131	0.086

Phase	Column (5) from Table 2
QU	$.5(1-p^4)(1-p^{10})$
S1	p^{10}
RS1	$2p^{10}(1-p)$

$$D = .5(1-p^4)(1-p^{10}) + p^{10} + 2p^{10}(1-p)$$

and the corresponding AFI for p = .90 and p = .95 would be,

$$AFI_{.90} = 0.672$$
$$AFI_{.95} = 0.568$$

143

SAMPLING FREQUENCY COMPARISON

29.   The examination of sampling frequency options for skip-lot sampling plans originated as a by-product of the first author's involvement as the Canadian project officer for QWG/PIQA project QA/29. During the process of producing an advisory paper [1] on the principles of skip-lot sampling it was noted that it was common to choose the sampling frequencies for plans in an ad-hoc fashion. This lack of formal thought in selecting sampling frequencies prompted the initiation of this study.

30.   In an effort to provide quality assurance personnel with some concrete guidance on selecting suitable sampling frequencies, three different options have been defined. They are based on simple generating functions for determining the sampling frequency at each successive stage in the skipping phase. Table 4 lists the generating functions for the three sampling options that are to be compared.

31.   The form of the generating functions were derived from those sampling frequencies most often applied for skip-lot plans, and are based on the premise of conservative sampling initially and more economical (less frequent) sampling at subsequent stages. Each option therefore starts the skip phase with sampling at the 50% level, or 1 in every 2 lots being inspected. Option A retains its

TABLE 4

SAMPLING FREQUENCY GENERATING RULES

Option	Generating Function	
A	$f_n = \dfrac{1}{n+1}$	for $n = 1,2,3,4$
B	$f_n = (\dfrac{1}{2})^n$	for $n = 1,2,3,4$
C	$f_n = \dfrac{1}{\dfrac{1}{f_{n-1}} + \dfrac{1}{f_{n-2}}}$	where $f_0 = f_{-1} = 1$ for $n = 1,2,3,4$

144

conservativeness throughout, reducing the sampling frequency slowly, whereas option B is the most liberal option in terms of the fraction of lots being inspected. Option C (also referred to as the Fibonacci option), meanwhile, duplicates the conservativeness of Option A for the early sampling stages, then begins to relax the inspection rate with sampling frequencies set between the two extremes.

32.   These features of Option C make it a good candidate as a compromise solution for someone who is concerned with both keeping inspection costs down and retaining good responsiveness for detecting quality changes.

33.   Figures 5, 6, and 7 illustrate skip-lot sampling plans with a qualification phase, followed by a skip phase that contains four sampling stages. Each sampling stage includes an intermediate retrial step to check for quality degradation. Figure 5 depicts the plan for sampling frequencies generated from option A. Figure 6 gives the equivalent plan for frequencies obtained from option B, and similarly, Figure 7 provides the plan derived from option C.

34.   The remaining portion of this report compares the quality assurance properties of these three sampling plans in terms of their respective operating characteristic (OC), average outgoing quality (AOQ), and average fraction of lots inspected (AFI) curves. Initial values are assigned to the clearance numbers (i, x_1, x_2, and x_3). Later, a sensitivity analysis is performed on the clearance numbers to measure their impact on the shape of the aforementioned curves.

35.   The transition matrix and working table for this four stage process is derived in Annex A. The graphs in Figures 8, 9, and 10 are the AFI, AOQ, and OC curves for the three sampling frequencies. In creating these graphs, the values for i, x_1, x_2, and x_3 are 10, 5, 10, and 15 respectively.

36.   The AFI curves shown in Figure 8 demonstrate the differences between the three options. When the lot quality is very high (above 95%), the average fraction inspected is very low. Also, the rate at which the AFI rises as lot quality decreases is very slow. However, when the lot quality drops lower (below 90%), the AFI increases rapidly. Therefore, a small change in lot quality in this range will result in a large change in the average fraction of lots inspected. This is characteristic of all three curves. At high quality levels, the AFI for sampling option A is greater than for option C, which in turn is larger than option B. The lower AFI is more economical, however, it

145

FIGURE 5

STEPPING RULE : $1/(n+1)$

SAMPLING LEVELS

QUALIFICATION STAGE

Inspect 100% of the lots. If 1 consecutive lots are approved

1

Inspect $f_1=1/2$ of the lots

x_1 consecutive lots approved

If a lot is rejected

Inspect each of the next 4 lots

If all 4 lots are approved

If a lot is rejected

2

Inspect $f_2=1/3$ of the lots

x_2 consecutive lots approved

If a lot is rejected

Inspect each of the next 4 lots

If all 4 lots are approved

If a lot is rejected

3

Inspect $f_3=1/4$ of the lots

x_3 consecutive lots approved

If a lot is rejected

Inspect each of the next 4 lots

If all 4 lots are approved

If a lot is rejected

4

Inspect $f_4=1/5$ of the lots until a lot is rejected

Inspect each of the next 4 lots

If all 4 lots are approved

If a lot is rejected

146

FIGURE 6

STEPPING RULE : $(\frac{1}{2})^n$

QUALIFICATION STAGE

SAMPLING LEVEL

QUALIFICATION STAGE	1	2	3	4
Inspect 100% of the lots. If 1 consecutive lots are approved	Inspect $f_1=1/2$ of the lots	Inspect $f_2=1/4$ of the lots	Inspect $f_3=1/8$ of the lots	Inspect $f_4=1/16$ of the lots until a lot is rejected
	X_1 consecutive lots approved	X_2 consecutive lots approved	X_3 consecutive lots approved	
	If a lot is rejected	If a lot is rejected	If a lot is rejected	
	Inspect each of the next 4 lots	Inspect each of the next 4 lots	Inspect each of the next 4 lots	Inspect each of the next 4 lots
	If all 4 lots are approved	If all 4 lots are approved	If all 4 lots are approved	If all 4 lots are approved
	If a lot is rejected	If a lot is rejected	If a lot is rejected	If a lot is rejected

147

STEPPING RULE : FIBONACCI NUMBERS

QUALIFICATION STAGE

Inspect 100% of the lots. If 1 consecutive lots are approved

SAMPLING LEVEL

	1	2	3	4
	Inspect $f_1=1/2$ of the lots	Inspect $f_2=1/3$ of the lots	Inspect $f_3=1/5$ of the lots	Inspect $f_4=1/8$ of the lots until a lot is rejected
	X_1 consecutive lots approved	X_2 consecutive lots approved	X_3 consecutive lots approved	
	If a lot is rejected	If a lot is rejected	If a lot is rejected	

Inspect each of the next 4 lots	Inspect each of the next 4 lots	Inspect each of the next 4 lots	Inspect each of the next 4 lots
If all 4 lots are approved	If all 4 lots are approved	If all 4 lots are approved	If all 4 lots are approved
If a lot is rejected	If a lot is rejected	If a lot is rejected	If a lot is rejected

148

Figure 8

AVERAGE FRACTION INSPECTED

also allows more non-conforming lots to pass without inspections. As can be seen in the graph, the AFI for option C increases more quickly as the lot quality decreases, eventually approaching the curve for option. This feature is desirable since it will permit an easier detection of a shift in quality by monitoring the AFI.

37.   The proportion of non-conforming lots passed is shown by the AOQ curve in Figure 9. This curve demonstrates that the average output quality is lowest for option B and largest for option A, with the AOQ for option C in between. Another characteristic shown by this graph is that the AOQ remains high, even when lot quality decreases. The reason for this was explained earlier in paragraph 20.

38.   The OC curve in Figure 10 shows that for superior lot quality, there is a high probability of a lot being accepted, however this decreases rapidly when lot quality decreases. Option A has the lowest probability of acceptance of the three choices, followed by option C, then option B. The options with the lower AFI would be expected to reject a smaller number of lots, thereby a larger number of lots would be accepted, giving a higher OC curve.

39.   This comparison procedure can also be used to study the effects of altering the clearance variables, denoted by x_1, in the skip-lot procedure. The graph given in Figure 11 shows the AFI curve for variations in x_1, when the sampling frequencies are generated from option C (Fibonacci). As can be seen in the graph, when the quality level is high, the AFI is changed very little for the different choices. However, when the quality of the lots decreases, the AFI shows a marked difference. When the value for x_1 increases, the AFI increases simultaneously. Thus, a larger value for x_1 will result in larger changes in the value of the statistics. This in turn increases the chance of detection when negative changes in quality occur.

40.   It must be remembered, however, that the values for AFI and the other statistics are given for a steady state process. If the situation is that of a limited number of lots being produced, then a large value for x_1 would have an adverse effect on the skip-lot procedure. If the value is too large, more time will be wasted in the first stage than is necessary. This is another factor which must be taken into account when a skip-lot plan is being designed.

41.   The effect of changing the other clearance variables is much the same as that of changing x_1, but to a smaller extent. By raising the value of x_2, for instance, the effect on the curves are very much similar, but the differences in the AFI are smaller than for a similar change in x_1.

Figure 9

AVERAGE OUTPUT QUALITY
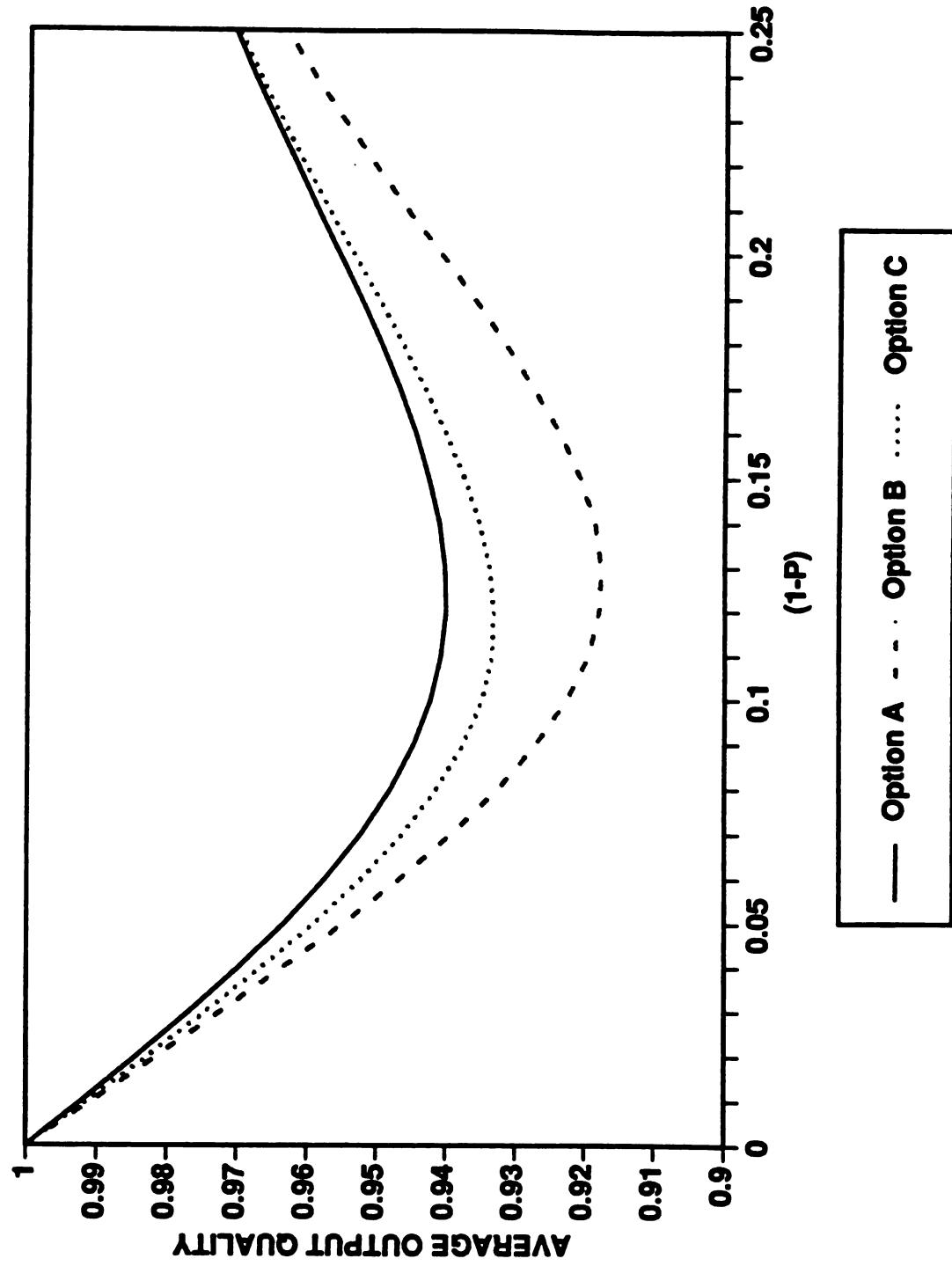
Figure 10

OPERATING CHARACTERISTIC
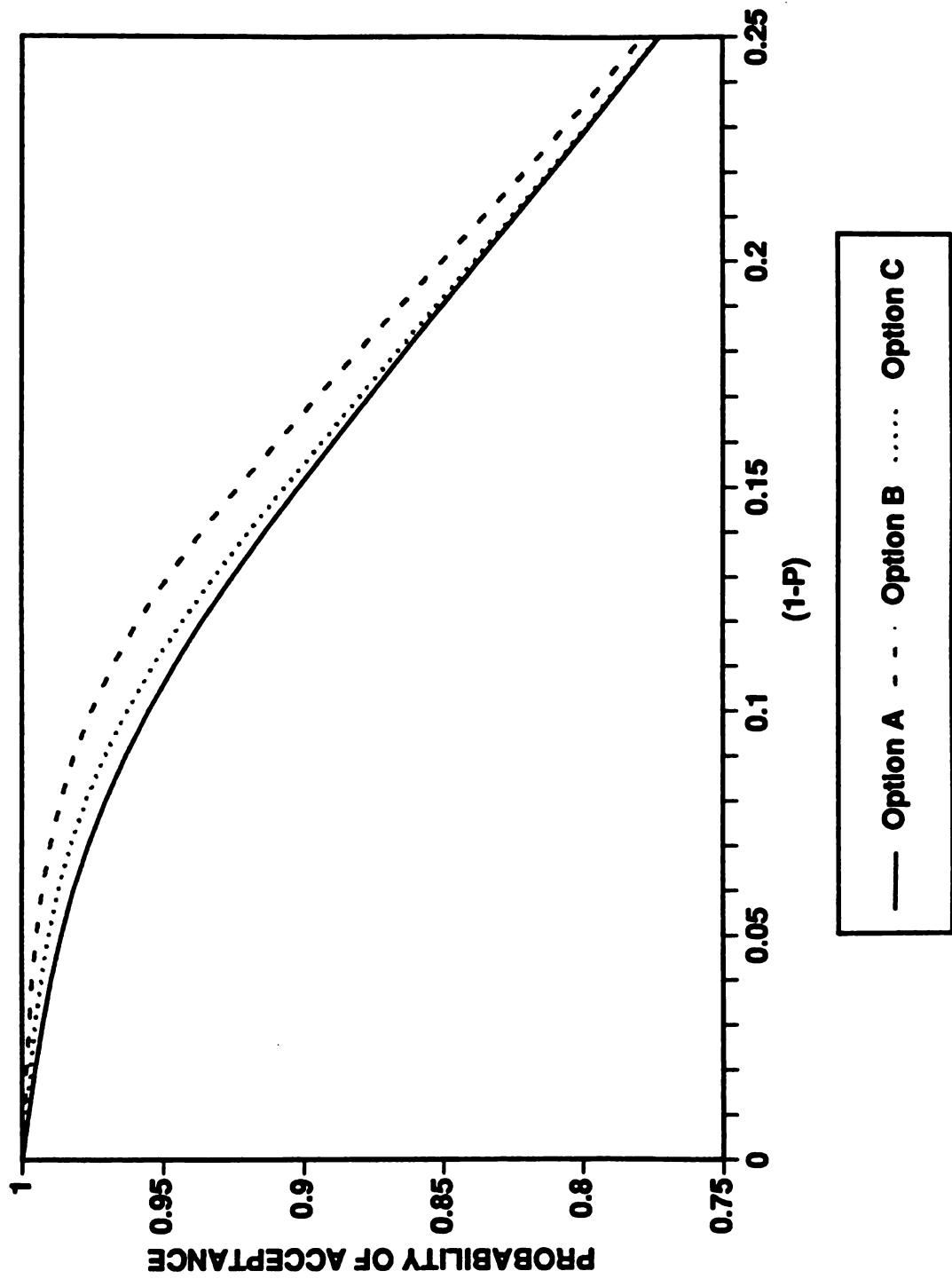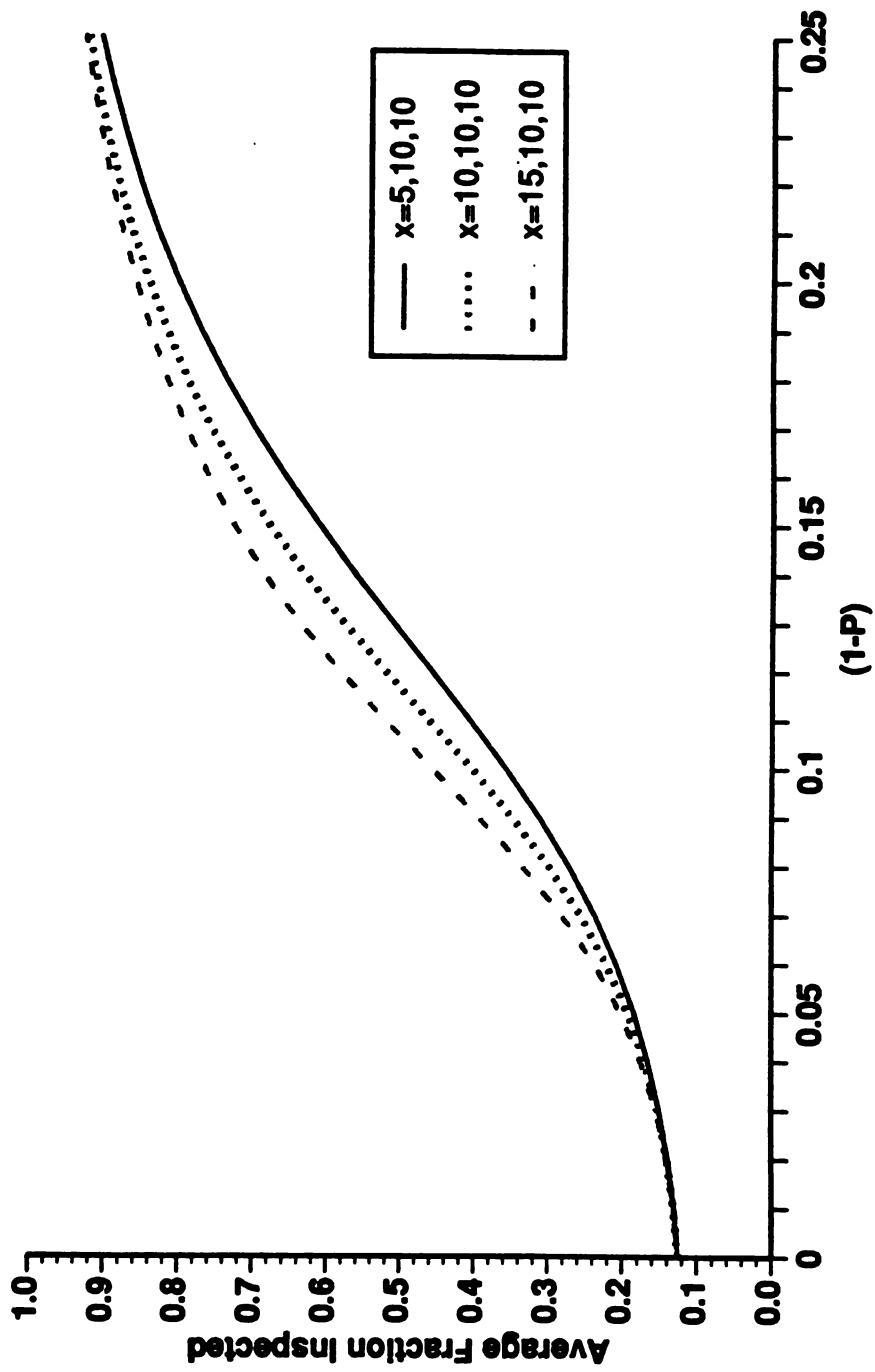
Figure 11

CLEARANCE NUMBER SENSITIVITY
(Varying X1)

153

SUMMARY

42.   This report has introduced the reader to the statistical lot inspection technique known as skip-lot sampling. The principles behind the method were explained by using a simple one-stage skip-lot plan as an example. The conditions under which implementing a skip-lot plan into a quality assurance process would be justified were also provided.

43.   The second part of the report dealt with three different sampling frequency options that could be used for multi-stage skip-lot sampling plans. These options were compared by deriving their respective quality assurance characteristics using the simplified Markov chain approach. It was noted that the sampling frequency option C, generated from the Fibonacci sequence of numbers offered a good compromise between minimizing the inspection effort and maintaining a high degree of responsiveness to lot quality deterioration.

44.   A sensitivity analyses on the clearance numbers showed that the clearance number x_1, associated with the first sample stage, had the largest impact on the shape of the AFI curve. It was also demonstrated that one can customize any skip-lot plan through manipulation of the clearance numbers.

REFERENCES

[1] ABCA Armies, QWG/PIQA Quadripartite Advisory Paper 28, "Principles of Skip-Lot Sampling", May 1991

[2] Brugger, R.M., "A Simplification of the Markov Chain Approach to Continuous Sampling Plan Formulation", QEM 21-230-12 U.S. Army Ammunition Procurement and Supply (1972)

[3] Brugger, R.M., "A Simplification of Skip-Lot Procedure Formulation", Journal of Quality Technology 7, pg 165-167 (1975)

[4] D/DQA/PSD/SR/5/6/10 U.K. Letter to Project QA/29 Project Officers, 21 Sept 1988

[5] Duncan, A.J. Quality Control and Industrial Statistics, Richard Irwin Inc. (1986) edition 2

[6] Cochran, W.G. Sampling Techniques, John Wiley & Sons (1977) edition 3

[7] ABCA Armies, QWG/PIQA Quadripartite Advisory Paper 16, "The Mathematical Background of QSTAG 340" December 1982

[8] Brugger, R.M., "A Simplified Markov Chain Analysis of ANSI/ASQC Z1.4 Used Without Limit Numbers", Journal of Quality Technology 21, pg 97-102, April 1989

MATHEMATICAL COMPUTATIONS

1.　　　This Annex provides the mathematical details for computing the steady state probabilities in terms of lots for the four-stage skip-lot plans given in Figures 5, 6 and 7.　Table A-1 shows the transition matrix with non-zero cell entries where it is possible to transit from the step given by the row to the corresponding column, representing the step being entered.

TABLE A-1

NEW VERSION

	QU	S1	RS1	S2	RS2	S3	RS3	S4	RS4
QU	-	1	-	-	-	-	-	-	-
S1	-	-	$1-P^{X1}$	P^{X1}	-	-	-	-	-
RS1	$1-P^{6}$	P^{6}	-	-	-	-	-	-	-
S2	-	-	-	-	$1-P^{X2}$	P^{X2}	-	-	-
RS2	$1-P^{6}$	-	-	P^{6}	-	-	-	-	-
S3	-	-	-	-	-	-	$1-P^{X3}$	P^{X3}	-
RS3	$1-P^{6}$	-	-	-	-	P^{6}	-	-	-
S4	-	-	-	-	-	-	-	-	1
RS4	$1-P^{6}$	-	-	-	-	-	-	P^{6}	-

156

2.   The stationary probabilities in terms of sequences or runs of each step can be determined by summing down the respective columns of this matrix. These equations can be restated in terms of T(Qu) which gives the expressions identified by the bracketed numbers for each step.

(1)

$$T(QU) = T(QU)$$

$$T(S1) = T(QU) + P^4 \cdot T(RS1)$$

$$T(RS1) = (1 - P^{X_1}) \, T(S1)$$

$$T(S1) = T(QU) + P^4 (1 - P^{X_1}) \, T(S1)$$

(2)

$$T(S1) = \frac{T(QU)}{1 - P^4 (1 - P^{X_1})}$$

(3)

$$T(RS1) = \frac{(1 - P^{X_1}) \cdot T(QU)}{1 - P^4 (1 - P^{X_1})}$$

$$T(S2) = P^{X_1} T(S1) + P^4 T(RS2)$$

$$T(RS2) = (1 - P^{X_2}) \, T(S2)$$

$$T(S2) = P^{X_1} T(S1) + P^4 (1 - P^{X_2}) \, T(S2)$$

$$T(S2) = \frac{P^{X_1} T(S1)}{1 - P^4 (1 - P^{X_2})}$$

(4)

$$T(S2) = \frac{P^{X_1}T(QU)}{[1-P^4(1-P^{X_1})] \cdot [1-P^4(1-P^{X_2})]}$$

(5)

$$T(RS2) = \frac{(1-P^{X_2})P^{X_1}T(QU)}{[1-P^4(1-P^{X_1})][1-P^4(1-P^{X_2})]}$$

$$T(S3) = P^{X_2}T(S2) + P^4 T(RS3)$$

$$T(RS3) = (1-P^{X_3})T(S3)$$

$$T(S3) = P^{X_2}T(S2) + P^4(1-P^{X_3})T(S3)$$

$$T(S3) = \frac{P^{X_2}T(S2)}{1-P^4(1-P^{X_3})}$$

(6)

$$T(S3) = \frac{P^{X_1}P^{X_2}T(QU)}{[1-P^4(1-P^{X_1})][1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})]}$$

(7)

$$T(RS3) = \frac{P^{X_1}P^{X_2}(1-P^{X_3})T(QU)}{[1-P^4(1-P^{X_1})][1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})]}$$

$$T(S4) = P^{X_3}T(S3) + P^4 T(RS4)$$

$$T(RS4) = T(S4)$$

$$T(S4) = \frac{P^{X_3}T(S3)}{1-P^4}$$

(8) (9)

$$T(RS4) = T(S4) = \frac{P^{X_1}P^{X_2}P^{X_3}T(QU)}{[1-P^4(1-P^{X_1})][1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})](1-P^4)}$$

3.   In order to obtain the steady state probabilities in terms of lots one must extract the coefficients from equations 1-9 and multiply them by the expected length of each step in terms of lots. A working table can be constructed (Table A-2) to assist in keeping track of the algebraic operations that need to be carried out. Reference [2] provides the justification for using this working table method.

	COLUMN 1
STEP	COEFFICIENT OF T(QU)
QU	1
S1	$\dfrac{1}{1-P^4(1-P^{X_1})}$
RS1	$\dfrac{(1-P^{X_1})}{1-P^4(1-P^{X_1})}$
S2	$\dfrac{P^{X_1}}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]}$
RS2	$\dfrac{P^{X_1}(1-P^{X_2})}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]}$
S3	$\dfrac{P^{X_1}P^{X_2}}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]\,[1-P^4(1-P^{X_3})]}$
RS3	$\dfrac{P^{X_1}P^{X_2}(1-P^{X_3})}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]\,[1-P^4(1-P^{X_3})]}$
S4	$\dfrac{P^{X_1}P^{X_2}P^{X_3}}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]\,[1-P^4(1-P^{X_3})]\,(1-P^4)}$
RS4	$\dfrac{P^{X_1}P^{X_2}P^{X_3}}{[1-P^4(1-P^{X_1})]\,[1-P^4(1-P^{X_2})]\,[1-P^4(1-P^{X_3})]\,(1-P^4)}$

COLUMN 2	COLUMN 3
SIMPLIFICATION OF COLUMN 1	**EXPECTED NO. OF LOTS**
$[1-P^4(1-P^{X_1})][1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})](1-P^4)$	$\dfrac{(1-P^4)}{P^4(1-P)}$
$[1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})](1-P^4)$	$\dfrac{(1-P^{X_1})}{f_1(1-P)}$
$(1-P^{X_1})[1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})](1-P^4)$	4
$P^{X_1}[1-P^4(1-P^{X_3})](1-P^4)$	$\dfrac{(1-P^{X_2})}{f_2(1-P)}$
$P^{X_1}(1-P^{X_2})[1-P^4(1-P^{X_3})](1-P^4)$	4
$P^{X_1}P^{X_2}(1-P^4)$	$\dfrac{(1-P^{X_3})}{f_3(1-P)}$
$P^{X_1}P^{X_2}(1-P^{X_3})(1-P^4)$	4
$P^{X_1}P^{X_2}P^{X_3}$	$\dfrac{1}{f_4(1-P)}$
$P^{X_1}P^{X_2}P^{X_3}$	4

COLUMN 4	COLUMN 5
SIMPLIFICATION OF COLUMN 3	(2) X (4)

$f_1 f_2 f_3 f_4 (1-P^i)$ $\qquad$ $f_1 f_2 f_3 f_4 (1-P^i)[1-P^4(1-P^{X_1}][1-P^4(1-P^{X_2})](1-P^4(1-P^{X_3})](1-P^4)$

$P^i f_2 f_3 f_4 (1-P^{X_1})$ $\qquad$ $f_2 f_3 f_4 P^i (1-P^{X_1})[1-P^4(1-P^{X_2})](1-P^4(1-P^{X_3})](1-P^4)$

$4P^i(1-P)f_1 f_2 f_3 f_4$ $\qquad$ $f_1 f_2 f_3 f_4 4P^i(1-P)(1-P^{X_1})[1-P^4(1-P^{X_2})][1-P^4(1-P^{X_3})](1-P^4)$

$P^i f_1 f_3 f_4 (1-P^{X_2})$ $\qquad$ $f_1 f_3 f_4 P^i (1-P^{X_2})P^{X_1}[1-P^4(1-P^{X_3})](1-P^4)$

$4P^i(1-P)f_1 f_2 f_3 f_4$ $\qquad$ $f_1 f_2 f_3 f_4 4P^i(1-P)P^{X_1}(1-P^{X_2})[1-P^4(1-P^{X_3})](1-P^4)$

$P^i f_1 f_2 f_4 (1-P^{X_3})$ $\qquad$ $f_1 f_2 f_4 P^i (1-P^{X_3})P^{X_1}P^{X_2}(1-P^4)$

$4P^i(1-P)f_1 f_2 f_3 f_4$ $\qquad$ $f_1 f_2 f_3 f_4 4P^i(1-P)P^{X_1}P^{X_2}(1-P^{X_3})(1-P^4)$

$P^i f_1 f_2 f_3$ $\qquad$ $f_1 f_2 f_3 P^i P^{X_1}P^{X_2}P^{X_3}$

$4P^i(1-P)f_1 f_2 f_3 f_4$ $\qquad$ $f_1 f_2 f_3 f_4 4P^i(1-P)P^{X_1}P^{X_2}P^{X_3}$

162

D = SUM OF ENTRIES IN COLUMN 5

$$D_1 = f_2 f_3 f_4 (1-P^4) [1-P^4(1-P^{x_2})] [1-P^4(1-P^{x_3})] [f_1(1-P^1) [1-P^4(1-P^{x_1})] + P^1(1-P^{x_1}) + f_1 4 P^1 (1-P)(1-P$$

$$D_2 = f_1 f_3 f_4 P^1 P^{x_1} (1-P^{x_2}) [1-P^4(1-P^{x_3})] (1-P^4) [1+4(1-P) f_2]$$

$$D_3 = f_1 f_2 f_4 P^1 P^{x_1} P^{x_2} (1-P^{x_3}) (1-P^4) [1+4(1-P) f_3]$$

$$D_4 = f_1 f_2 f_3 P^1 P^{x_1} P^{x_2} P^{x_3} [1+4(1-P) f_4]$$

$$D = D_1 + D_2 + D_3 + D_4$$

163

4.   Column 1 of Table A-2 lists the coefficients of T(QU) from equations 1-9.   Column 2 is a simplification of column 1 by getting rid of the denominators.   This multiplication or division performed on all the values in the column does not effect the end result.   Column 3 gives the expected length for each step in terms of lots.   Column 4 is a simplification of column 3.   Column 5 is the product of the values from columns 2 and 4.

5.   The steady state probabilities for lots are obtained by taking each of the terms in column 5 and dividing them by D, the sum of all the terms in column 5.

6.   To compute the AFI curve for each of the sampling frequency options from Table A-2 one proceeds in the same manner as was explained in paragraph 27 of the main report for the simple skip-lot case.   Each of the terms in column 5 are first multiplied by the sampling frequency corresponding to the respective skip-lot plan.   This equates to multiplying the column 5 entries for the qualification stage and the retrial stages by one (100% inspection) and the sampling stages by their respective sampling frequencies (f_i, i = 1,2,3,4).   These values are then summed together and divided by D to obtain the AFI.

164

Toxic Fumes Reduction Modifications
to the M2A2 Bradley Fighting Vehicle

Linda L. C. Moss and William P. Johnson
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

ABSTRACT

Carbon monoxide concentrations measured in the M2A2 Bradley Fighting Vehicle were high enough to cause the crew's predicted carboxyhemoglobin level to exceed the Military Standard during firings of the the acceptance test scenario. Five vehicle modifications were proposed to correct this deficiency, and experiments were conducted to identify carboxyhemoglobin levels of the modified vehicles relative to the baseline vehicle. The sample data was generalized to determine the probability that each configuration would pass the toxic fumes acceptance test. Recommendations for permanent vehicle modifications were made to the Program Manager-Bradley Fighting Vehicles based on the study results.

1. INTRODUCTION

The last Quality Performance Test for new or modified Bradley Fighting Vehicles is the Toxic Fumes Acceptance Test. This test assures that the crew's exposure to toxic fumes from the propellant gases of the weapon system is within safety limits. The safety limits for carbon monoxide stated in the Military Standard 1472C, paragraph 5.13.7.4.2 are:

> *Carbon monoxide in personnel areas shall be reduced to the lowest levels*
> *feasible. Personnel shall not be exposed to concentrations of carbon monoxide*
> *(CO) in excess of values which result in carboxyhemoglobin levels in their blood*
> *greater than the following percentages: 5% COHb, all system design objectives*
> *and aviation systems performance limits; 10% COHb, all other system*
> *performance limits.*

Prior to March 1989, five of the ten vehicles tested with 500-horsepower engines failed the toxic fumes acceptance test. That is, the predicted carboxyhemoglobin level for one or more crew members exceeded the 10% limit. Consequently, the Program Manager-Bradley Fighting Vehicles formed a working group to develop and test modifications for reducing the toxic fumes concentrations within the Bradley crew compartment during firings of the Bradley weapon system. The working group members included representatives from the Office of Program Manager-Bradley Fighting Vehicles (PMO), the Ballistic Research Laboratory (BRL), the Combat Systems Test Activity (CSTA), the Test and Evaluation Command (TECOM), the Army Materiel Systems Analysis Activity (AMSAA), the Training and Doctrine Command (TRADOC) and the FMC Corporation.

The working group concluded that several proposed vehicle modifications offered sufficient potential for reducing the toxic fumes in the Bradley to warrant testing. Descriptions of the recommended modifications are presented in subsequent sections of this report.

The test program objectives were:

(1) To quantify the level of toxic fumes produced in the unmodified, 600-horsepower M2A2 Bradley Fighting Vehicle during firings of the acceptance test.

(2) To quantify the relative effectiveness of proposed Bradley Fighting Vehicle modifications for reducing toxic fumes concentrations within the Bradley Fighting Vehicle during firings of the acceptance test.

(3) To provide a recommendation to the Program Manager-Bradley Fighting Vehicles for a solution to the Bradley toxic fumes problem based on test results and data analysis.

The acceptance test scenario is a combination of TRADOC events 12 and 15. The scenario, listed in Table 1, shows the sequence for firing the M242 25-millimeter (mm) gun and the M240C 7.62-mm coaxial machine gun. The method of firing is either a single-shot mode or a burst mode. At the conclusion of the 46-minute scenario, 320 25-mm rounds and 300 7.62-mm rounds are expended.

Table 1. Bradley Fighting Vehicle Acceptance Test (TRADOC Events 12 & 15)

TIME (min)	NUMBER OF ROUNDS 25-mm	7.62-mm	MODE	TIME (min)	NUMBER OF ROUNDS 25-mm	7.62-mm	MODE
0	10	–	SS	29	10	–	SS
2	10	–	SS	30	10	–	SS
4	10	–	SS	31	10	–	SS
6	10	–	SS	32	10	–	SS
8	10	–	SS	33	10	–	SS
10	10	–	SS	34	10	–	SS
11	10	–	SS	35	10	–	SS
12	10	–	SS	36	10	–	SS
13	–	21	B	36.5	10	–	SS
14	–	21	B	37	10	–	SS
15	–	21	B	38	10	–	SS
17	–	21	B	39	10	–	SS
19	–	42	B	40	5	–	SS
20	–	21	B	40.5	5	–	B
21	–	43	B	41	5	–	SS
21.5	5	–	SS	41.5	5	–	B
22	5	–	SS	42	5	–	SS
22.5	5	–	SS	42.5	5	–	B
23	5	–	B	43	5	–	SS
23.5	5	–	SS	43.5	5	–	B
24	5	–	B	44	5	–	SS
25	5	–	B	44.5	10	–	B
26	10	–	SS	45	–	55	B
27	10	–	SS	46	–	55	B
28	10	–	SS				

B = Burst / SS = Single Shot

2. PRETEST DATA ANALYSIS

In preparation for the test program, the toxic fumes data collected during the period July 1983 to March 1989 were examined to determine the variability of the carboxyhemoglobin (COHb) levels computed for the Bradley Fighting Vehicle (BFV) from the earlier tests. This information was sought to guide the test design and sample size recommendations.

MIL-STD-1472C dictates that passage of the BFV system is contingent on "all" vehicle positions having COHb values less than 10%. Consequently, our analysis focused on the highest COHb value calculated within the test scenario, regardless of crew position. This value will be subsequently referred to as the vehicle COHb level. We reasoned that if the highest COHb level satisfied the military standard, all other COHb levels within the vehicle satisfied the standard. Our approach was conservative, taken to

166

guarantee that the baseline and/or modified vehicles provided protection to every crew position. Therefore, all analyses to determine the best vehicle configuration used the vehicle COHb level.

Percentage data are frequently transformed using the arc sine transformation to stabilize the variances. However, the arc sine transformation neither increased stability of the variances nor changed the results in terms of significant differences. Consequently, all calculations within this report were performed with the vehicle COHb values in percentage units.

The mean vehicle COHb value calculated from the historic data is 8.9% with a standard deviation of 2.37. This information was used to determine the maximum mean COHb level that could be tolerated with a given sample size while retaining a predetermined level of confidence that the population mean COHb level would be less than 10%.

If we assume the sample standard deviation in our tests is no more than 2.37, then to verify the hypothesis that the true mean of the BFV COHb level was less than 10% with the desired 99.5% level of confidence, the sample mean calculated from a minimum of six experimental replicates can not exceed 6.1%. This was a reasonable goal for at least one of the proposed modifications, the BIGGRS modification (described in section 5), since its projected performance in the A0 vehicle was 4.0%[1].

The discussion to this point has addressed the criterion established to provide a 99.5% level of confidence that the "mean" vehicle COHb values would not exceed 10%. To establish tolerance limits with this high level of confidence that a large percentage of the population of COHb values would not exceed 10%, a much larger sample size would have been required. For example, if we had wished to state with 99.5% confidence that at least 75% of the population of vehicle COHb values would be less than 10%, a minimum of 30 experimental replications of each vehicle configuration would have been necessary when assuming a normal distribution. Resource constraints prohibited the testing of larger sample sizes within this test program.

3. DATA COLLECTION

Although the primary objective of this test program was the documentation of carbon monoxide at the crew positions, other toxic fumes (such as carbon dioxide, ammonia and oxides of nitrogen) were monitored to insure that their levels were below the applicable standards. These additional data will be stored in a CSTA database for future reference and analysis.

Toxic fumes data were collected by the Chemistry Branch of CSTA by placing sampling tubes in the breathing zones of the driver, gunner, commander, and in the center of the crew compartment. The air was continuously analyzed for ammonia and carbon monoxide at all four locations and for carbon dioxide at the commander and crew positions. All measurements were made with rapid response, non-dispersive, infrared gas analyzers. Oxides of nitrogen (nitric oxide and nitrogen dioxide) were continuously analyzed, at the commander and crew positions, by chemiluminescent analyzers. Concentration data were recorded at a minimum of four times per second.

Differential pressures (inside - outside) were measured in the turret by a capacitance type differential pressure sensor (-0.1 to +0.1 psid) and recorded at a minimum of four times per second. The interior temperature was measured in the turret by T-type thermocouples.

All testing was completed with 25-mm M793 TP-T and 7.62-mm Ball/Tracer ammunition. Firings were not conducted if the wind speed exceeded 10 mph or if the relative humidity exceeded 90%. All firings were conducted under the same general meteorological conditions and as close together as possible.

4. ANALYTICAL PROCEDURES

Carboxyhemoglobin levels, for each crew position, were calculated from the measured carbon monoxide by the modified Coburn-Forster- Kane equation[2,3]:

$$\%COHb_f = \%COHb_0 \, e^{-t/A} + 218 \, (1 - e^{-t/A}) \left[\frac{1}{B} + \frac{CO}{1403} \right]$$

where:

$\%COHb_f$ = COHb concentration at the end of the computation interval (percent).
$\%COHb_0$ = COHb concentration at the beginning of the computation interval (percent).
t = computation interval (15 seconds).
CO = average concentration of carbon monoxide detected during the computation interval (in parts per million).

A & B are regression constants determined by the work level.

The carbon monoxide data were integrated at 15 second intervals throughout the test scenario. Calculations began at the initial detection of carbon monoxide and continued until the concentration returned to pre-fire levels. The COHb increment was calculated using the average carbon monoxide concentration, a computation interval of 15 seconds, and a work level characteristic of the physical exertion required in weapon firing. An initial COHb concentration of 1.0% was assumed per MIL-STD-1472c. The computed COHb concentration from one iteration became the beginning COHb concentration for the successive iteration. Final COHb concentrations were computed after the test scenario for each vehicle configuration. Statistical comparisons of the mean COHb levels were made both parametrically, under the assumption that the COHb levels were normally distributed, and nonparametrically. The results of these comparisons, presented in subsequent sections of this report, served as the basis for the recommendations included herein.

5. PHASE I TESTING
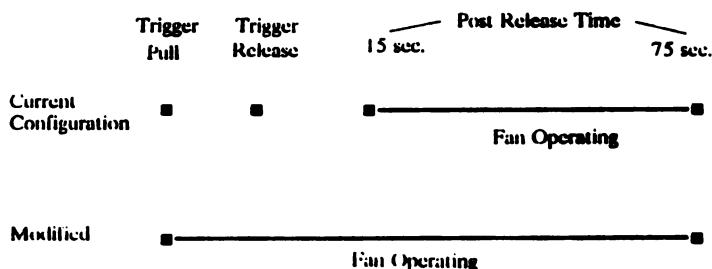
5.a. VEHICLE CONFIGURATIONS

Phase I of the test program included testing of three vehicle configurations within TRADOC Events 12 & 15. These included the baseline vehicle, a modification of the rear hull fan, and the vehicle with the Bradley Improved Gun Gas Removal System (BIGGRS or BG) kit attached.

The baseline test vehicle recommended for this study was the 600-horsepower M2A2 Bradley with a constantly operating Noah-Howden engine cooling fan. The test vehicle was considered a "worst" case for toxic fumes because of the internal under-pressure condition expected during concurrent operation of the engine, the Noah-Howden cooling fan and the vanaxial rotor fan.

The modification of the rear hull fan, referred to as the Fan Mod, modified the operational sequence of the rear hull fan. The activation time of the rear hull fan was changed from 15 seconds after gun trigger release to activation at trigger pull. The rear hull fan remained operational throughout trigger depression and continued to operate for a minimum of one minute fifteen seconds after trigger release (Figure 1). This modification increased the rear hull fan operational time from 37% of the scenario time to 90%. The operational sequence of the driver's hull fan was not changed; it activated 15 seconds after trigger release, remained on high speed for a maximum of one minute or until the trigger was pulled for the next scenario event. Both fans directed the air flow from the exterior of the vehicle to the interior.

168

Figure 1. Fan Control Box Operational Sequence

```
                        Trigger   Trigger      ──── Post Release Time ────
                        Pull      Release    15 sec.                    75 sec.

          Current
          Configuration   ■         ■           ■─────────────────────────■
                                                          Fan Operating


          Modified            ■─────────────────────────────────────────■
                                             Fan Operating
```

The third vehicle configuration tested in Phase I was the BIGGRS kit, which consisted of three parts: (a) a removable cover (or sock) for the coaxial machine gun feed chute, (b) improved seals for the coaxial machine gun access doors and (c) a deflector to direct the gases from the 25-mm chain gun breech to the vanaxial rotor fan located between the main gun and the coaxial machine gun. However, an independent decision by the PM-BFV to include improved door seals on all M2A2 future production vehicles and on the test vehicle eliminated the need for the seals as part of the BIGGRS kit. The improved seals for the coaxial machine gun access doors used with the baseline and all vehicle modifications were developed and installed by the FMC Corporation.

An experimental test design was constructed to test the hypothesis that the mean COHb values of the baseline vehicle (BL), the fan modification (FM), and the Bradley Improved Gun Gas Removal System (BG) were equal. The alternative hypothesis was that at least one mean value differed from the others. Six replicates of each configuration were planned for a total of 18 tests. A "pseudorandom" design was planned in two blocks. Three replicates from each configuration made up a block. The design is considered pseudorandom because the first block would be completed before the start of the second block. This approach provided a checkpoint on the efficiency of the potential vehicle modifications. At the halfway point, a cursory analysis was performed to determine if the tested modifications reduced the vehicle COHb levels as desired.

5.b. PARAMETRIC ANALYSIS

Analysis of variance is robust to the assumption of normality; however, it is not robust to heterogeneity of variance. Therefore, Cochran's test was implemented on the variances of the three configurations. The results indicated no significant difference among the variances; hence, the assumption of homogeneity of variance appeared justified.

At the $\alpha = 0.10$ level of significance chosen by the working group, the critical F-value, $F_{0.10}(2,6)$, equals 3.46. As a result of the ANOVA, the F-statistic equaled 1.00. Since the F-statistic was less than the critical value, we failed to reject the null hypothesis of equal means. This implied that within the constraints of the statistical test, neither the Fan Mod nor BIGGRS provided significantly lowered COHb levels than the baseline configuration.

A plot of the percent COHb obtained from each crew position in each vehicle configuration tested is shown in Figure 2. Here we observed that each configuration had at least one sample point close to or greater than the limit specification of 10%. Figure 3 shows the mean vehicle COHb value and confidence interval for each configuration. The overlapping intervals indicate that the mean COHb levels are not significantly different. The estimate of the standard deviation, Sp, used to construct the confidence intervals is the square root of the mean squared error from the ANOVA, which is 1.40.

Figure 2. Percent COHb by Position
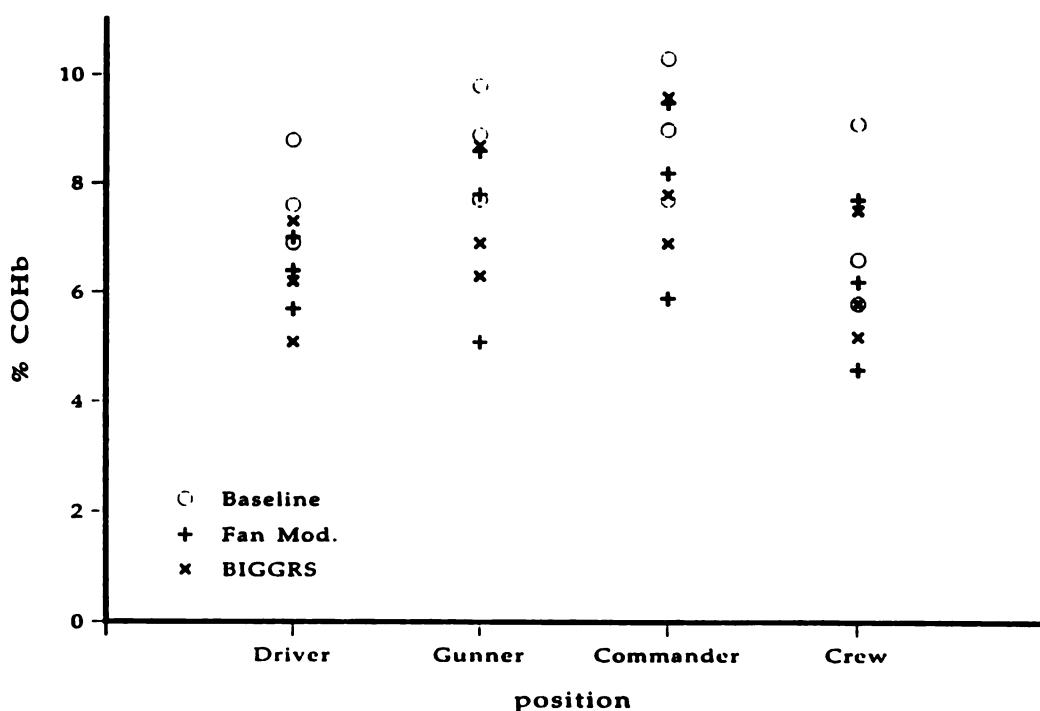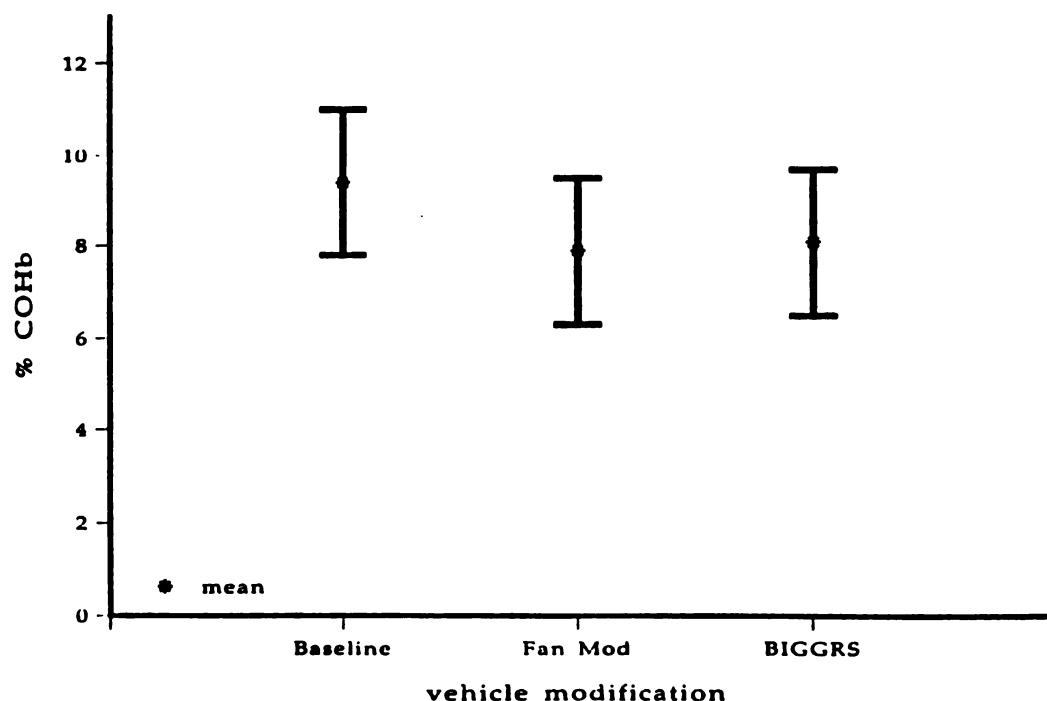
Figure 3. 90% Confidence Intervals on Vehicle Means, Phase I

$Sp = 1.40$

5.c. NONPARAMETRIC ANALYSIS

In addition to the parametric analysis, the Kruskal-Wallis Nonparametric Test was performed on the ranks of the data. Since three tests were performed on each configuration, a total of nine data points (nine COHb values) were ranked, with the lowest receiving a rank of one. The test statistic, T, computed as described by Conover[4] is 3.29. This value is less than the exact $1 - \alpha$ (0.90) quantile which equals 4.6. Therefore, as with the parametric analysis, we would fail to reject the hypothesis that the configurations have equal mean COHb values.

6. ADDITIONAL MODIFICATIONS CONSIDERED

The conclusions in Phase I forced consideration of alternative vehicle modifications. No growth potential appeared to exist for the BIGGRS modification; therefore, it was eliminated from further consideration. A comparison of the baseline and Fan Mod data collected during Phase I indicated that the mean COHb level, at each position and for the vehicle, was reduced from the baseline levels during the Fan Mod configuration tests. The baseline test conditions required normal operation of both the driver's and the rear hull fans; the Fan Mod configuration required normal operation of the driver's fan and increased operating time of the rear hull fan. The increased operating time of the rear hull fan increased the airflow within the test vehicle from that of the baseline configuration. These observations led to the hypothesis that the vehicle COHb level could be reduced further by increasing the airflow within the test vehicle.

Two options were recognized for increasing the vehicle internal airflow. The first, hereafter referred to as the Dual Fan (DF) modification, increased the airflow by changing the operating pattern of both the rear and the driver's hull fans as previously illustrated in Figure 1. Both fans were activated simultaneously at trigger pull and remained operating for a minimum of one minute fifteen seconds after trigger release.

The second option, hereafter referred to as the Reversed Fan (RF), activated both the rear and driver's hull fans at trigger pull and physically reversed the driver's hull fan. The reversal of the fan directed the airflow from the interior of the vehicle to the exterior.

Three replications each of the Reversed Fan and the Dual Fan modifications, within TRADOC Events 12 & 15, were recommended for Phase II testing. Subsequent to the completion of Phase II, it was proposed that the vehicle modification that appeared to offer the greatest potential for reducing the vehicle COHb level below the desired 6.1% limit be tested at least three additional times within Phase III.

7. PHASE II TESTING AND ANALYSIS

7.a. PARAMETRIC ANALYSIS
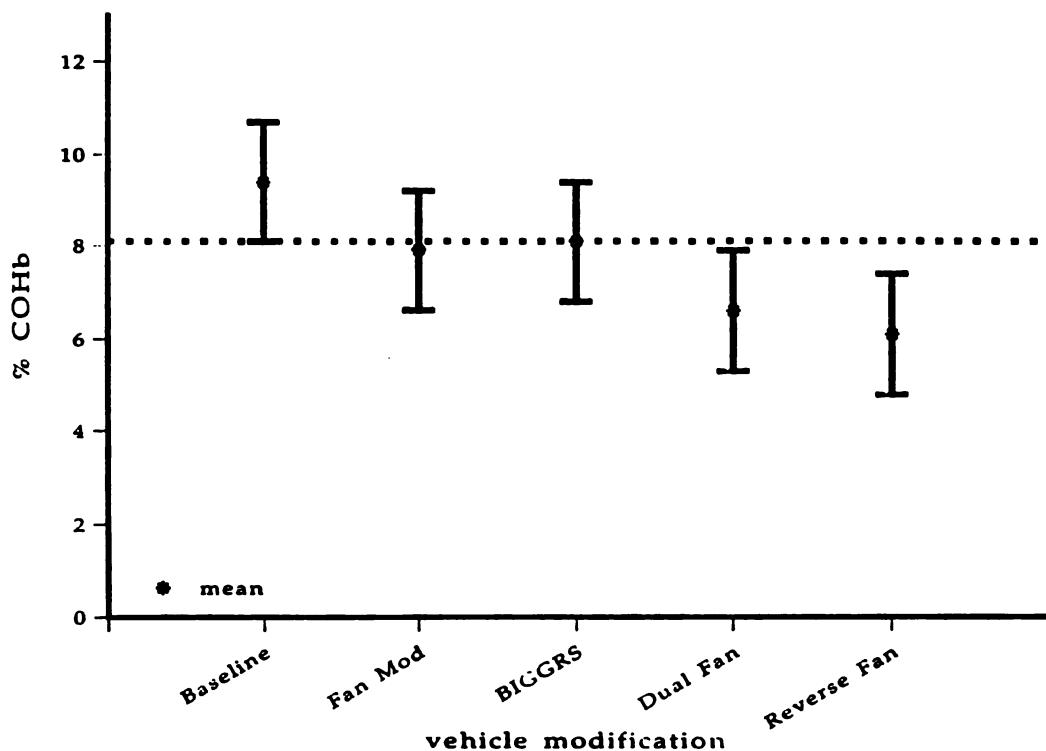
Three replicates each of the Dual Fan and Reversed Fan modifications were tested during Phase II within TRADOC Events 12 & 15. An ANOVA was performed on both Phase I and Phase II data. Since the F-statistic, 3.35, was greater than the critical F-value, $F_{0.10}$ (4,10) = 2.61, the hypothesis of equal means was rejected. To determine which means were different, the 90% confidence interval for each mean was plotted with Sp, the pooled estimate of the variance, equal to 1.23. Failure of the confidence intervals to overlap indicates that the means of the associated modifications are significantly different from each other. The horizontal dotted line in the figure provides a visual reference for the lower confidence bound for the mean of the baseline configuration. If a confidence interval falls below the dotted line, the mean of the associated modification is significantly different from the baseline configuration.

The confidence intervals in Figure 4 show no significant difference among the tested "modified" vehicles; however, the confidence intervals for both the Dual Fan and the Reversed Fan mods fall below the lower confidence bound of the baseline. This observation led to the conclusion that both the Dual Fan and the Reversed Fan modifications were significantly better than the baseline in reducing toxic fumes within the test vehicle. However, because the observed vehicle COHb mean of the Reversed Fan data was (1) less than the observed vehicle COHb mean of the Dual Fan data and (2) less than the predetermined

171

vehicle COHb mean specified in the pretest analysis, the Reversed Fan modification was selected for supplemental testing in Phase III.

Figure 4. 90% Confidence Intervals on Vehicle Means, Phases I & II

Sp = 1.23

7.b. NONPARAMETRIC ANALYSIS

As in Phase I, the nonparametric Kruskal-Wallis Test was performed on the ranks of the data. The T-value for the test was 7.63. The critical value, estimated by the chi-square distribution with four degrees of freedom at the $1 - \alpha$ (0.90) quantile is 7.78. Therefore, by this test method, we failed to reject the hypothesis of equal means. However, since the T-value was very close to the critical value, we proceeded to Phase III with the Reverse Fan modification.

8. PHASE III

8.a. TESTING AND PARAMETRIC ANALYSIS

The primary objective of Phase III testing was to obtain a better estimate of the vehicle mean COHb level of the Reversed Fan modification and the baseline vehicle by increasing the sample size of each. Another configuration was included in Phase III testing at the request of the Bradley Program Manager's Office. This configuration required a complete shutdown of the driver's hull fan (Fan Off). During one of the earlier phases, one test run was considered bad because the driver's fan was not operating. The calculated COHb value was less than that of the baseline vehicle. Although it was not near the desired COHb value of 6.1%, the Program Manager's Office insisted that this modification be tested.

172

The mean COHb values and standard deviation for each configuration in the three phases are listed in Table 2. Since the test for homogeneity of variance could not be rejected, ANOVA was performed on all the data with the degrees of freedom adjusted to account for the second set of three baseline and three reverse fan tests conducted several days after the initial tests in Phase I.

Table 2. Phase I - III Statistics for Each Configuration

Configuration	n	Mean	Std. Dev.
Baseline	6	9.5	1.02
Rear Fan	3	7.9	1.82
BIGGRS	3	8.1	1.37
Dual Fan	3	6.6	1.27
Reverse Fan	6	5.5	0.97
Driver's Fan Off	4	7.6	0.51

The 90% confidence intervals about the means are presented in Figure 5 for each vehicle configuration tested within the program. The pooled estimate of the variance used to construct the confidence intervals was Sp = 1.13. Figure 5 reveals that at 90% Confidence Level the Dual Fan, Reverse Fan, and Fan Off modifications are significantly different than the baseline vehicle. A similar plot drawn for the 99% confidence intervals about the means is shown in Figure 6. Here only the Reverse Fan modification is significantly different from the baseline vehicle. However, sample size was a contributor to the relative lengths of the intervals and the conclusions drawn. To answer the question "What conclusions would have been drawn if the sample size of the Dual Fan and Fan Off modifications were equaled to that of the Reverse Fan (six)?" We conclude from the dotted lines about the means shown in Figure 6 that if the sample size were increased to six for both the Dual Fan and Fan Off modifications (assuming the mean were no greater than the mean calculated in the sample of size three and that the variance remained the same) the Dual Fan would have been determined to be significantly different than the baseline vehicle; however, the Fan Off modification would not have been significantly different from the baseline vehicle.

8.b. NONPARAMETRIC RESULTS

Examining the results of the Kruskal-Wallis test with all the data from each of the three phases, the T-statistic is significant at the α = 0.01 level. Multiple comparisons were then performed on the average rank of each vehicle configuration. Figure 7 shows the pairwise results for α = 0.10, 0.05 and 0.01 levels. For each level of significance, a line begins at each configuration until a significant difference occurs. For example, referring to the α = 0.10 level, the average rank for the baseline vehicle, BL, is significantly different from all the other configurations. This is indicated by the horizontal line breaking when it reaches the next lowest average rank, BG. Then the line continues at BG thru DF, meaning that BG is not significantly different from FM, FO, or DF. However, BG is significantly different from RF; therefore, the line does not continue. A line beginning at the Fan Mod is not shown, since its line is a subset of the previous line. That is, the average rank for FM is not significantly different from FO or DF, but it is significantly different from RF. Likewise, a line is not drawn for FO, because its line is also a subset of the line for BG. Indicating FO is not significantly different than DF, but is significantly different from RF. Lastly, a horizontal line is drawn under DF and RF to represent that there is no significantly difference between the average ranks of the two modification at the α = 0.10 level.

Similarly, the results are shown for α = 0.05 and 0.01. At α = 0.01, DF and RF are not significantly different from each other, but both are significantly different from the baseline vehicle.

173

Figure 5. 90% Confidence Intervals on Vehicle Means, Phases I - III
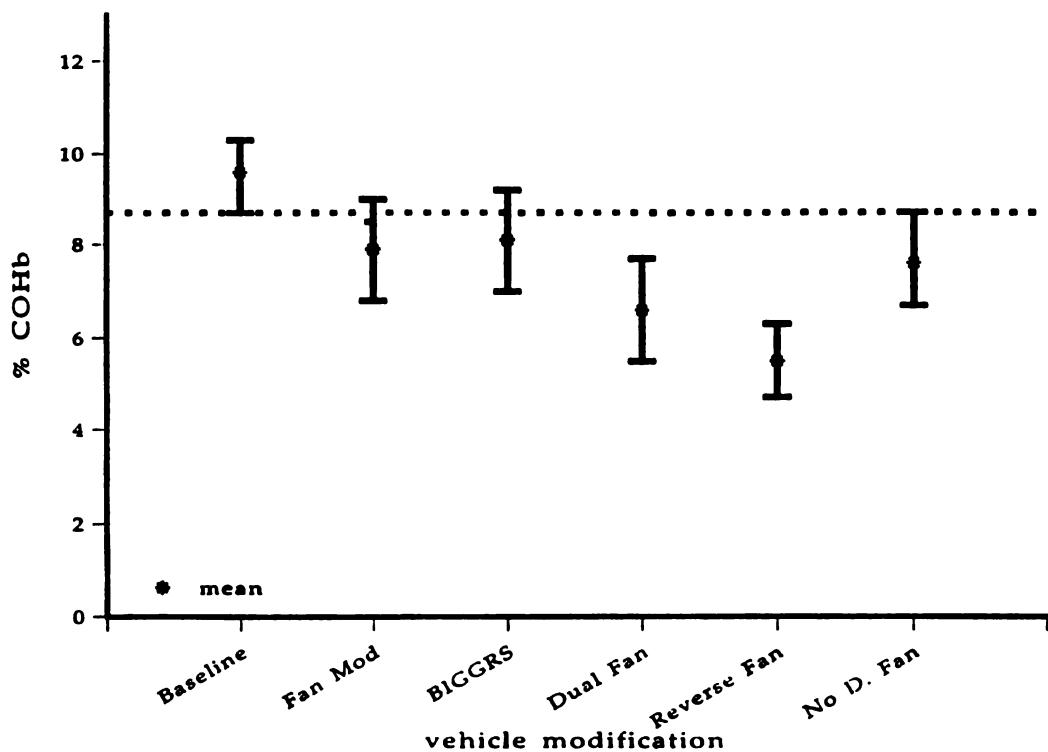
Sp = 1.13

Figure 6. 99% Confidence Intervals on Vehicle Means, Phases I - III

Sp = 1.13
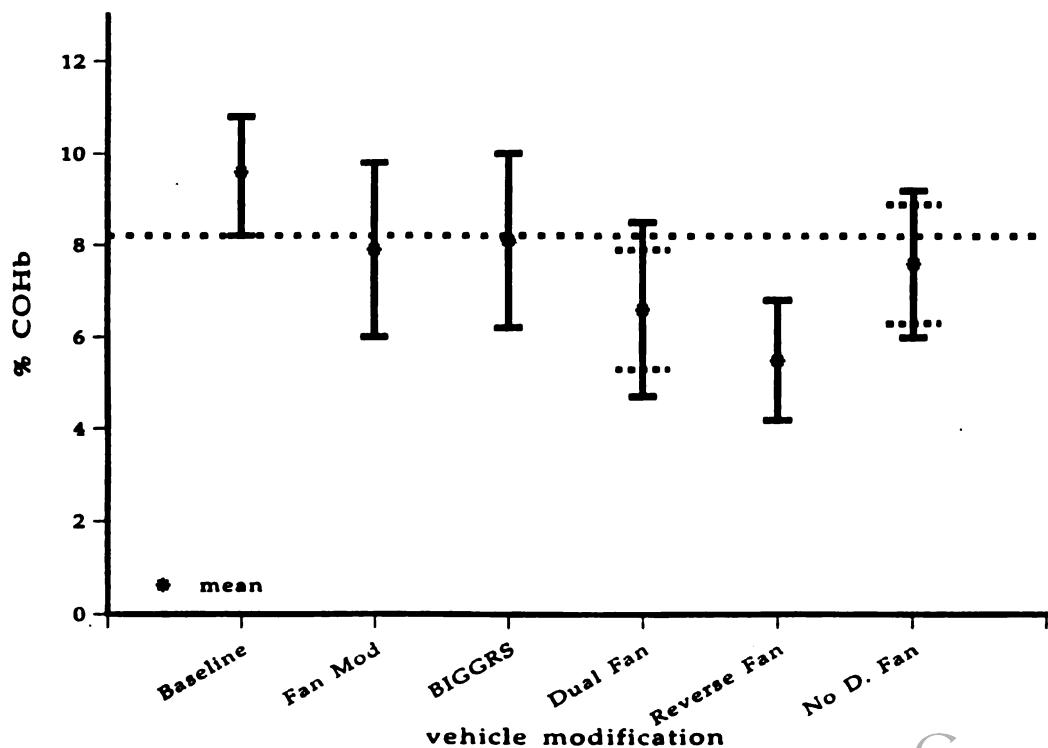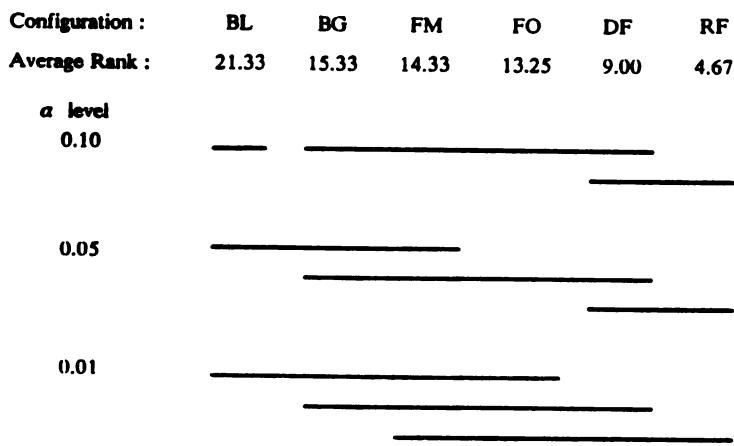
Figure 7. Nonparametric Pairwise Results

Configuration :	BL	BG	FM	FO	DF	RF
Average Rank :	21.33	15.33	14.33	13.25	9.00	4.67

α level

0.10

0.05

0.01

9. TOLERANCE LIMITS

The analysis up to this point was concerned with significant differences among the means. We discussed at the beginning of this paper the large sample size required if we were to guarantee that a large percent of the population were to remain under 10% with a very high level of confidence. However, we will report on the tolerance limits for the normal distribution expected from this small sample of six and three for the Reverse Fan and Dual Fan, respectively (see Table 3). The nonparametric tolerance limits for sample sizes this small will have much lower values either in confidence level and/or population proportion. The exact values are being investigated.

Table 3. One-Sided Tolerance Limits for
the Normal Distribution

Configuration	$\gamma = .90$	$\gamma = .95$	$\gamma = .99$
Reverse Fan	P = .98	P = .96	P = .86
Dual Fan	P = .82	P < .75	–

γ = confidence level that P proportion of
the population will have a COHb $\leq$ 10%

10. RECOMMENDATIONS

Testing of the modified and unmodified Bradley Fighting Vehicle within TRADOC Events 12 & 15 and subsequent analysis of the collected data leads to the following recommendations:

(1) that the Reversed Fan modification be considered the primary solution to the Bradley Fighting Vehicle toxic fumes acceptance test problem.

(2) that the Dual Fan modification be considered the secondary solution to the Bradley Fighting Vehicle toxic fumes acceptance test problem.

175

(3) that additional acceptance testing, with other M2A2 vehicles and with the M2A0 vehicle, be conducted to confirm the results obtained and the hypotheses formed within this test program.

(4) that a test program to investigate the effectiveness of the Reversed Fan modification for maintaining Bradley vehicle COHb levels below the limit specification, during the performance of the training and combat scenarios, be initiated.

(6) that vehicle-to-vehicle variability be investigated

11. POST-TEST PROGRAM ACTIONS

Subsequent to the completion of the test program, the conclusions and recommendations contained within this report were briefed to the Toxic Fumes Working Group and to the Program Manager-Bradley Fighting Vehicles.

The Program Manager-Bradley Fighting Vehicles concluded that the Dual Fan Mod was equally as effective as the Reversed Fan mod in reducing toxic fume concentrations within the M2A2 vehicle during the performance of the acceptance test scenario and that the Dual Fan mod was a more economical and logistically simplier modification than the Reversed Fan mod. For these reasons, the decision was made to include the Dual Fan modification on all future A2 Bradleys and to retrofit the modification to all existing A1 Bradley fighting Vehicles.

REFERENCES

1. Lyon, David H. and Dennis C. Kelham, *Bradley Improved Gun Gas Removal System*, Technical Report BRL-TR-3035, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, September 1989.

2. MIL-HDBK-759A, *Human Factors Engineering - Design for Army Materiel*, 30 June 1981.

3. DF, AMXHE-cc 26 February 1985, Subject: *Proposed revisions to MIL-HDBK-759A*.

4. Conover, W. J., Practical Nonparametric Statistics, 2nd edition, John Wiley & Sons, Inc., 1980.

5. Kirk, Roger E., Experimental Design: Procedures for the Behavioral Sciences, 2nd edition, Brooks/Cole Publishing Company, Monterey, California, 1982.

SOME LIMITATIONS OF THE RANK TRANSFORMATION TEST
FOR INTERACTION

W. J. CONOVER
College of Business Administration
Texas Tech University
Lubbock, Texas 79409

ABSTRACT. The rank transformation is used widely to convert parametric tests, such as the t-test and the F-test, to nonparametric tests such as the Wilcoxon test and the Kruskal-Wallis test. It is also widely used in experimental designs to convert analysis of variance procedures to robust procedures that have superior power in some cases. As a test for interaction, the rank transformation can lead to a test that is not valid. Some discussion of the limitations of this use of the rank transformation is given in this paper.

1. INTRODUCTION. The Rank Transformation Methods refer to standard classical statistical procedures that are applied to the ranks of the data rather than to the data themselves. The following two steps are involved.

Step 1: Replace the data by their ranks, from rank 1 for the smallest, to rank N for the largest observation.

Step 2: Use a standard statistical procedure, such as the t-test or analysis of variance F-test, on the ranks.

Example: Three steel mills are being monitored for the amount of smokestack contaminants to see if there is a difference in mean level of contamination. Five randomly selected times for observation lead to the following measurements.

Factory A	Factory B	Factory C
46.3 (6)	48.6 (8)	45.1 (5)
43.7 (4)	52.3 (13)	46.7 (7)
51.2 (12)	50.9 (11)	41.8 (2)
49.6 (10)	53.6 (14)	40.4 (1)
48.8 (9)	55.7 (15)	42.6 (3)

The classical F statistic computed on the data gives

$$F = \frac{SST/(k-1)}{SSE/(N-k)} = \frac{198.1/2}{89.6/12} = 13.27$$

which is compared with the F distribution with 2 and 12 degrees of freedom. Because the upper .05 quantile is 3.885, the observed value is significant.

The classical F statistic computed on the ranks, given in parentheses, gives

$$F = \frac{SST/(k-1)}{SSE/(N-k)} = \frac{185.2/2}{94.8/12} = 11.72$$

177

This is compared with the same quantile from the F distribution as before, $F_{.05,2,12} = 3.885$, to determine that the statistic is significant, as before.

In some cases the Rank Transform Procedure is a <u>nonparametric</u> test

Classical <u>Procedure</u>	Rank Transform <u>Procedure</u>
The t-Test on 2 Independent Samples	The Wilcoxon Rank Sum Test
One-way Analysis of Variance	The Kruskal-Wallis Test
Correlation Test for Bivariate Independence	Spearman's Rho Test for Independence

In other cases this results in a <u>robust</u> test.

The Randomized Complete Block Analysis of Variance

The Balanced Incomplete Blocks Analysis of Variance

The Two-way Analysis of Variance <u>Without</u> Interaction

In all of these cases the Rank Transformation Procedure provides an alternative to the classical procedure that has <u>almost</u> as much power as the classical procedure when the usual normality assumption is met, and <u>much greater</u> relative power when outliers are present, or the distributions are heavy-tailed and non-normal.

There is a temptation to use the Rank Transformation in other experimental designs, where the theory has not yet caught up with practice, such as in a two-way layout with interaction to test for interaction.

    2. EARLY RESULTS. Early simulation results (Iman, 1974) indicated that the rank transformation led to a valid test for interaction. Iman used the linear model

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

$i = 1, 2, 3, 4$ (4 *treatments*)

$j = 1, 2, 3$ (3 *blocks*)

$k = 1, 2, 3, 4, 5$ (5 *observations per cell*)

Under the null hypothesis of no interaction ($\gamma_{ij} = 0$) the F statistic for interaction, computed on ranks, followed the usual F distribution closely in the two cases studied:

178

Case 1: $\alpha_i = 0$, $\beta_j = 0$ (No treatment or block effects)

Case 2: $\alpha_i = 0$, $\beta_1 = \beta_2 = 0$, $\beta_3 = 1$ (No treatment effects)

This provided the basis for suggestions to use the rank transformation - suggestions in papers by Conover and Iman, and suggestions in the users manuals of SAS and IMSL.

3. RECENT RESULTS. Later simulation results (Blair, Sawilowsky and Higgins, 1987) showed that under some extreme conditions, in the presence of both block and treatment effects, the performance of the rank transform test for interaction could be both non-robust, and lacking in power. They examined the following case, and varied the constant c as described in the table below.

Case 1: $\alpha_1 = \alpha_4 = 0$, $\alpha_2 = c$, $\alpha_3 = -c$ (treatment effects)
$\quad\quad\quad \beta_1 = c$, $\beta_2 = -c$, $\beta_3 = 0$ (block effects)
$\quad\quad\quad \gamma_{ij} = 0$ (no interaction, the null case)

The linear model used is the same one used by Iman, and given above. The error terms were taken to be standard normal, and the number of observations per cell was n, which was varied along with c as described in the following table. The entry in this table is an estimate of the true level of significance, at a nominal alpha level of .05, obtained by simulation with 1000 runs.

	n = 2	n = 5	n = 10	n = 20	n = 50
c = 0.5	.056	.049	.053	.050	.054
1.0	.046	.053	.073	.101	.193
1.5	.053	.076	.132	.309	.848
2.0	.044	.105	.326	.803	1.000
2.5	.053	.186	.682	.997	1.000

The table shows that for reasonably large shifts in treatment effects, the rank transformation test for interaction is not robust, even for fairly small sample sizes.

These results inspired a study of the theory behind this test, by Thompson (1991). She assumed the model was linear, as above, and found that the rank transformation test for interaction was valid if and only if at least one main effect (blocks and/or treatments) was not present. Further, she found that if both effects were present, the mean of the F statistic on ranks has a term that increases without bound as n increases, thus forcing α to 1.0.

Choi (1991) found similar results for the general model

$$X_{ijk} \sim F(x - \alpha_i - \beta_j - \gamma_{ij})$$

This explains why Iman (1974) detected no problem with the rank transform test for interaction, because he looked only at cases where one main effect was zero, and why Blair et.al. (1987) found serious problems, because they reported only cases where both main effects were present.

4. DISCUSSION. The conclusion of this paper is that the rank transformation procedure is very useful in some designs, but that it needs to be used with care, and not applied thoughtlessly to every analysis for which a classical procedure exists.

REFERENCES

1.   Blair, R.C., Sawilowsky, S.S., and Higgins, J.J. (1987). Limitations of the Rank Transform Statistic in Tests for Interaction. *Commun. Statist.-Simula.*, 16(4), 1133-1144.

2.   Choi, Y.H. (1991). The Rank Transform Procedure in the Two-Way Layout with Interaction. Ph.D. Dissertation, Texas Tech University.

3.   Iman, R.L. (1974). A Power Study of a Rank Transform for the Two-Way Classification Model When Interactions May Be Present, *Canad. J. Stat.*, 2, 227-239.

4.   Thompson, G.L. (1991). A Note on the Rank Transformation for Interactions. *Biometrika*, 78 (3), 697-701.

ON A NEW SYSTEM OF
MULTIVARIATE DISTRIBUTIONS

Major Kevin M. Beam
U.S. Army TRADOC Research Associate
RAND
1700 Main Street, P.O. Box 2138
Santa Monica, CA 90407 - 2138

Albert S. Paulson
School of Management
Rensselaer Polytechnic Institute
Troy, NY 12180 - 3590

ABSTRACT. There has been substantial interest in multivariate probability distributions with given margins since Galton's (1885) investigations. For a multivariate system to be useful, we require it to be flexibly constructed and easily used in the modeling of multivariate data. Furthermore, it should possess computational ease and be intuitively appealing. Its parameters should represent important physical properties, e.g., measures of scale, location, shape and correlation. We propose such a multivariate system: the Diagonal Perturbation System.

We construct the system as a multivariate refinement of the intuitively appealing framework of the Neyman alternative. We demonstrate the system's flexibility by presenting several univariate and bivariate models, and bivariate constructions in which the margins are univariate versions of the system. Additionally, we find the system's parameters to represent physical properties. We show the system to be readily implemented numerically. We demonstrate the system's utility through the successful modeling of multivariate data that has eluded fitting for over a decade. We employ nonlinear minimization to produce the least squares parameter estimates while capturing not only the usual sum of squared errors but also the margin's first two moments and the first mixed moment.

1. INTRODUCTION. There has been an interest in multivariate distributions with given margins for several years. These include the systems of Morgernstern (1956), Gumbel (1960,1961), and Farlie (1960); Sibuya (1960); Plackett (1965); Ali, Mikhail, and Haq (1978); Frank (1979); Clayton (1978), Cook and Johnson (1981), Clayton and Cuzick (1985); and Marshall and Olkin (1988). For a general discussion see Mardia (1970a), Johnson and Kotz (1972), and Johnson (1987). We propose a system of distributions which is flexibly constructed, intuitively appealing and easily used in the fitting of multivariate data.

Neyman (1937) proposed for the density of any random variable, X, the alternative;

$$f(x) = c \, \exp[(\sum_{j=1}^{k} \theta_j \, l_j(x)]; \ 0 \le x \le 1 \text{ and } k=1,2,\dots .$$

The l_j are Legendre polynomials; θ_j are parameters; and c, a function of the θ_j, is a normalizing constant. We construct our system as a multivariate refinement of this intuitively appealing framework.

In this paper we present two univariate constructions and consider examples when constructed from the uniform and gaussian distributions. We then consider two analogous margin preserving bivariate constructions, again constructed from the uniform and gaussian distributions. The bivariate distributions with uniform margins are copulas. We consider a third bivariate construction which does not preserve the margins. As a final example of the system's flexibility, we consider a bivariate construction in which the margins are univariate versions of the system. We successfully model two bivariate data sets and give direction to future research.

2. THE DIAGONAL PERTURBATION SYSTEM. Let $X_1, \ldots, X_n$ be n random variables (r.v.'s) with marginal distribution functions (d.f.'s), $\Theta_1(X_1), \ldots, \Theta_n(X_n)$; survival functions (s.f.'s), $\bar{\Theta}_1(X_1), \ldots, \bar{\Theta}_n(X_n)$; where $\bar{\Theta}_i(X_1) = 1 - \Theta(X_n)$; probability density functions (p.d.f.'s), if continuous - or probability mass functions (p.m.f.'s), if discrete, $\theta_1(X_1), \ldots, \theta_n(X_n)$. The system is constructed as

$$\Delta_\pi(X_1, \ldots, X_n) = \left[\prod_{i=1}^{n} \Theta_i(X_i) \right] \exp\{A[\bar{\Theta}_i(X_i)]\}; \quad i=1, \ldots, n. \tag{1}$$

Here Π denotes product and $A[\bar{\Theta}_i(X_i)]$ is a function of the s.f.'s. We now suppress the arguments of the distribution functions. If $A(\bar{\Theta}_i) = 0$, we have the product of the distributions, the uncorrelated case. If $A(\bar{\Theta}_i) \neq 0$, we have correlation, that is, changes in the variance-covariance structure. The notation, Δ_π, is appropriate as we have, in a sense, a change in the product of the marginal distributions as $A(\bar{\Theta}_i)$ varies from zero.

3. CONSTRUCTIONS CONSIDERED. Although there are infinite forms of (1) we consider the following univariate and bivariate constructions where $A(\bar{\Theta}_i)$ is a polynomial function. F is the diagonal perturbation d.f. of X and G is the original d.f. of X in the univariate constructions. In the bivariate constructions, F is the joint d.f. of X and Y, G is the d.f. of X, and H is the d.f. of Y.

<u>Univariate Constructions</u>   <u>Bivariate Constructions</u>

$$F = G \exp(\alpha_p \bar{G}^p) \tag{2}$$
$$F = GH \exp[\alpha_p(\bar{G}\bar{H})^p] \tag{4}$$

$$F = G \exp(\alpha_1 \bar{G} + \alpha_2 \bar{G}^2) \tag{3}$$
$$F = GH \exp[\alpha_1 \bar{G}\bar{H} + \alpha_2(\bar{G}\bar{H})^2] \tag{5}$$

$$F = GH \exp(\alpha_{10}\bar{G} + \alpha_{01}\bar{H} + \alpha_{11}\bar{G}\bar{H}) \tag{6}$$

Bivariate constructions (4) and (5) preserve the margins as $\int f \, dy = g$, where $f = \frac{\delta^2}{\delta x \delta y} F$. Construction (6) does not preserve the margins as $\int F \, dx = \frac{d}{dx} G \exp(a_{10}\bar{G})$; which is the univariate version (2) if $\alpha_p = \alpha_{10}$ and p = 1, or version (3) if $\alpha_1 = \alpha_{10}$ and $\alpha_2 = 0$.

182

4. UNIVARIATE CONSTRUCTIONS.

4.1. Examples of the First Univariate Construction.
The density for (2) is

$$f = (1 - p\alpha_p \bar{G}^{P-1}G)g \exp(\alpha_p \bar{G}^P). \tag{7}$$

For (2) to be a d.f. we require:

a) $F(-\infty) = 0$,   b) $F(\infty) = 1$, and   c) $F(x+h) \geq F(x)$ for $h \geq 0$.

The first two requirements are immediate and the final requirement reduces to $(1 \leq \alpha_p < \infty)$; $(-\infty < \alpha_p \leq 1)$ for $p = 1$, or $(-\infty < \alpha_p \leq (\frac{p}{p-1})^{P-1})$ for $p > 1$. We see $\lim_{p \to \infty} \alpha_p = e$, the natural exponent.

If G is the uniform d.f. on $[0,1]$, $f = (1 - p\alpha_p(1-x)^{P-1} x)g \exp(\alpha_p(1-x)^P)$. If $p = 1$, we have

$$E(X^n) = 1 - \frac{n}{\alpha_1} \left[\frac{n!}{\alpha_1^n} e^{\alpha_1} - \sum_{i=0}^{n} \frac{n!}{(n-i)! \, \alpha_1^i} \right], \text{ where } \lim_{\alpha_1 \to 0} E(X^n) = \frac{1}{n+1}, \text{ the } n^{th} \text{ moment for } X \sim U(0,1);$$

and $\text{Var}(X) = (\alpha_1^2 + 2\alpha_1 - 1 + 2\alpha_1^2 e^{\alpha_1} - 2\alpha_1 e^{\alpha_1} + 2e^{\alpha_1} - e^{2\alpha_1}) / \alpha_1^4$, where $\lim_{\alpha_1 \to 0} \text{Var}(X) = \frac{1}{12}$.

Various plots of (7) constructed from $U(0,1)$, where $p = 1$, are given in Figure 1. Figure 2 shows the effect of α_1 on μ, σ^2, γ_1, and γ_2; where γ_1 is the coefficient of skewness, $\mu_3/(\mu_2)^{3/2}$, and γ_2 is the coefficient of kurtosis, $\mu_4/(\mu_2)^2$; where μ_i is the ith moment about μ. As α_1 increases we see μ decreasing, a small decrease in σ^2, a large increase in γ_1 and a decrease then increase in γ_2. Further analysis of the effect of the parameter α_1 on the moment ratios is presented in Figure 3 where we see (7) graphed in the $\beta_1 - \beta_2$ plane for varying values of α_1 where $\beta_1 = \gamma_1^2$ and $\beta_2 = \gamma_2$. The $\beta_1 - \beta_2$ plane is presented for reference in Figure 4 where the equations for the bounding curves and the location of the Pearson densities are taken from Pearson and Hartley (1970). As expected, $(\mu, \sigma^2, \gamma_1, \gamma_2) = (\frac{1}{2}, \frac{1}{12}, 0, 1.8)$ when $\alpha_1 = 0$. We see extreme values of α_1 effecting J-shaped beta type I densities.

As another example of (7) constructed from the uniform distribution, we consider $p = 2$. Graphs of (7) are presented in Figure 5 for various values of α_2. The effect of α_2 on the moments and moment ratios is presented in Figure 6 and Figure 7. We see more extreme distortion of the density. Of real interest is the U-shaped beta type I densities for $\alpha_2 > 0$ and beta type II densities for $\alpha_2 < 0$.

If (7) is constructed from $N(0,1)$, the expectations are not tractable in closed form and quadrature is required for all results. For $p = 1$ we again see positive skewness for positive values of α_1 and negative skewness for negative values of α_1 in Figure 8. The moments and moment ratios are presented in Figure 9. As expected, $(\mu, \sigma^2, \gamma_1, \gamma_2) = (0, 1, 0, 3)$ when $\alpha_1 = 0$. We see σ^2 decreasing as α_1 changes from 0. The effect of α_1 on the gaussian
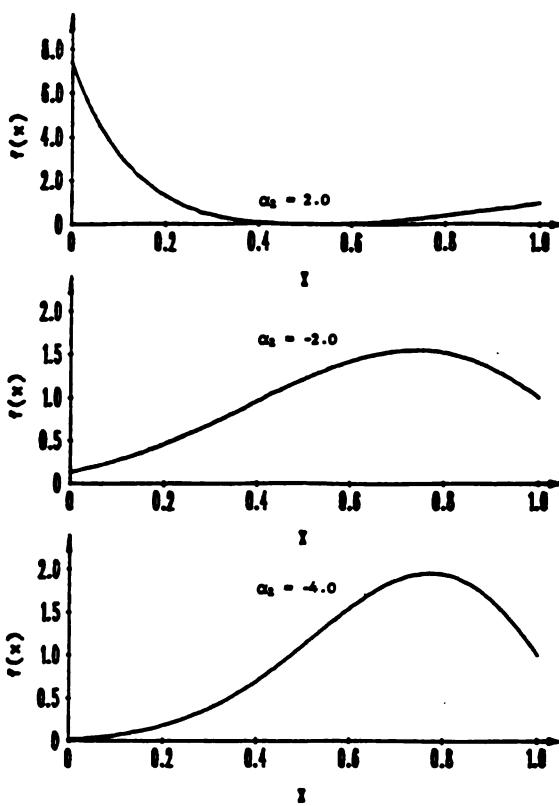
FIGURE 1   Density (7) Constructed from U(0,1);
          p = 1; α_1 = 1.0, 0.0, - 1.0
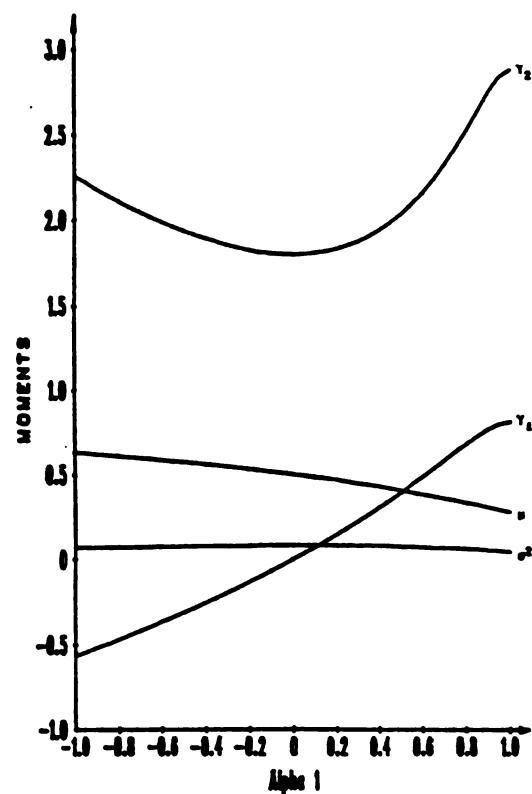
FIGURE 2   Moments and Moment Ratios of
          Density (7) Constructed from
          U(0,1) as a Function of α_1
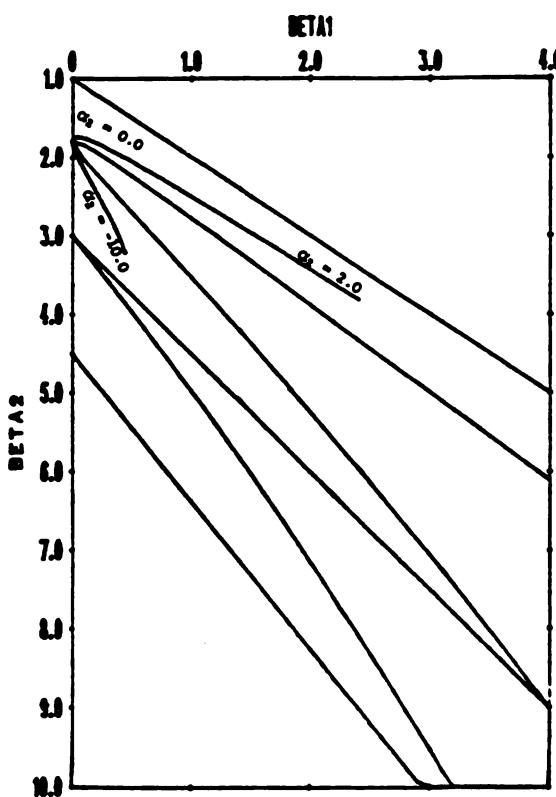
FIGURE 3   The Effect of α_1 on Density (7)
          Constructed from U(0,1) in the
          β_1 - β_2 Plane
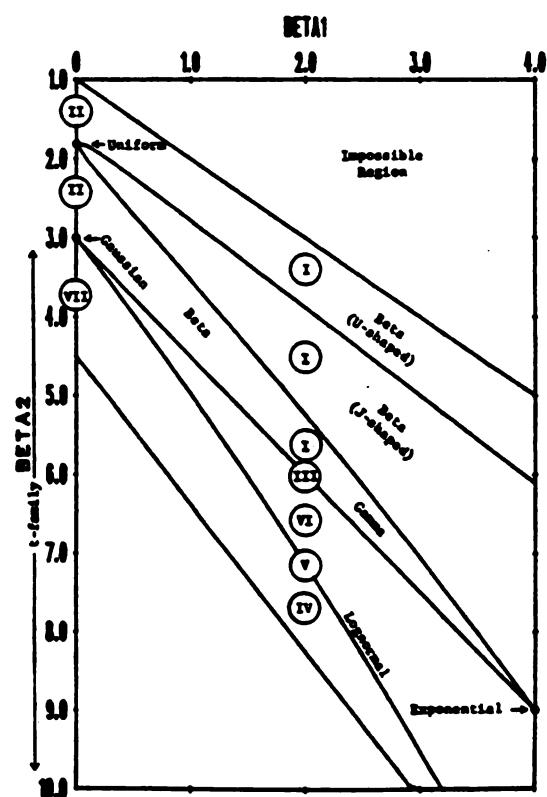
FIGURE 4   The β_1 - β_2 Plane with Locations of
          Some Common and Pearson-Type
          Distributions

184
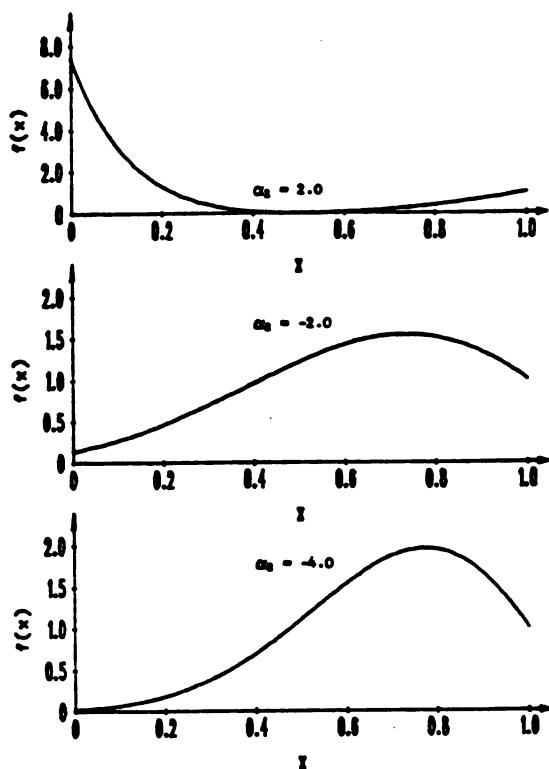
FIGURE 5 Density (7) Constructed from U(0,1);
p = 2; α_2 = 2.0, - 2.0, - 4.0
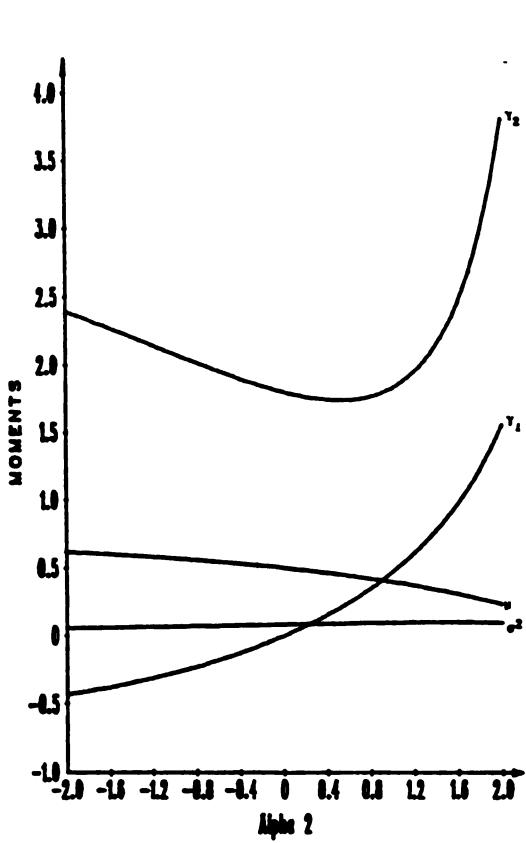
FIGURE 6 Moments and Moment Ratios of
Density (7) Constructed from
U(0,1) as a Function of α_2
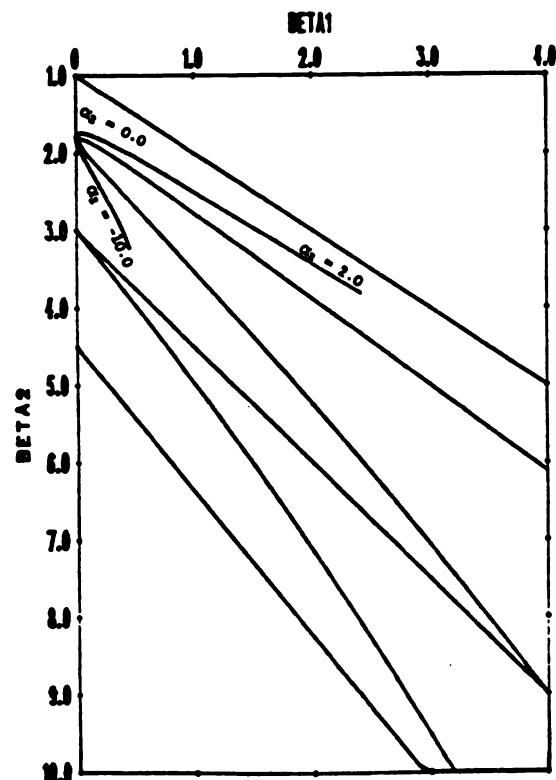
FIGURE 7 The Effect of α_2 on Density (7)
Constructed from U(0,1) in the
β_1 - β_2 Plane

185

distribution in the β_1 - β_2 plane is excluded as there is very little movement away from $(\beta_1, \beta_2) = (0,3)$.

We again consider (7) when constructed from the gaussian distribution when p = 2. Graphs for various positive values of α_2 are given in Figure 10. Graphs for negative values of α_2 are omitted as one can ascertain from the graph of the moments and moment ratios in Figure 11 that the density is perturbed as μ is increasing, σ^2 is decreasing, γ_1 is decreasing and γ_2 is increasing as α_2 decreases.

As a third example constructed from the gaussian distribution, we consider (7) when p = 3 and obtain results much like when p = 2. These are included in Figure 12 which shows the effect of varying values of α_2 and α_3 on the gaussian density in the β_1 - β_2 plane. We see densities appearing throughout the beta regions, to include the U-shaped type I and type II beta.

4.2. Examples of the Second Univariate Construction. The density for (3) is

$$f = (1 - \alpha_1 G - 2\alpha_2 G\bar{G})g \exp[\alpha_1 \bar{G} + \alpha_2 \bar{G}^2]. \tag{8}$$

For (3) to be a d.f., again, the first two requirements are immediate and the final requirement, from Section 4.1, reduces to $(-\infty \le \alpha_1 \le 1)$ and $(-\infty \le \alpha_2 \le 1 - \alpha_1/2 + \sqrt{1 - \alpha_1})$. The densities possible by construction of (8) with the gaussian density when only the maximum values of α_2 are considered are exhibited in Figure 13. This two version system encompasses nearly one-half of the admissible region of the β_1 - β_2 plane presented.

5. BIVARIATE CONSTRUCTIONS.

5.1. Examples of the First Bivariate Construction. The density for (4) is

$$f = \{1 + p\alpha_p(\bar{G}\bar{H})^{p-1}[pGH - G\bar{H} - \bar{G}H + p\alpha_p GH(\bar{G}\bar{H})^p]\} gh \exp[\alpha_p(\bar{G}\bar{H})^p]. \tag{9}$$

The sufficient conditions for (4) to be a d.f. (Mardia, 1970a) are

a) $F(\infty,\infty) = 1$, b) $F(x+0,y) = F(x,y+0) = F(x,y)$,
c) $F(-\infty,y) = F(x,-\infty) = 0$, and d) $F(x,y) + F(x+h,y+h) - F(x+h,y) - F(x,y+h) \ge 0$.

It is clear the d.f. satisfies the first three conditions. The final condition reduces to conditions on the parameters p and α_p. Without loss of generality, we consider (9) as a copula. As (9) is symmetric, the minimum value of f occurs on the axes, if $\alpha_p > 0$, or on y = x if $\alpha_p < 0$. If y = 0, we have (7) if G is the uniform d.f. on [0,1]. Thus, as in Section 4.1, we have $(1 \le \alpha_p < \infty)$; $(\alpha_p \le 1)$ for p = 1, or $(\alpha_p \le (\frac{p}{p-1})^{p-1})$ for p > 1. We see the upper limit of α_p as a function of p coincides with that for univariate construction (2). If y = x, the lower limit is

$$(\alpha_p \ge \frac{(p^2x^2 + 4px^2 - 4px)^{1/2} - px - 2x + 2}{2px(1-x)^{2p}}).$$
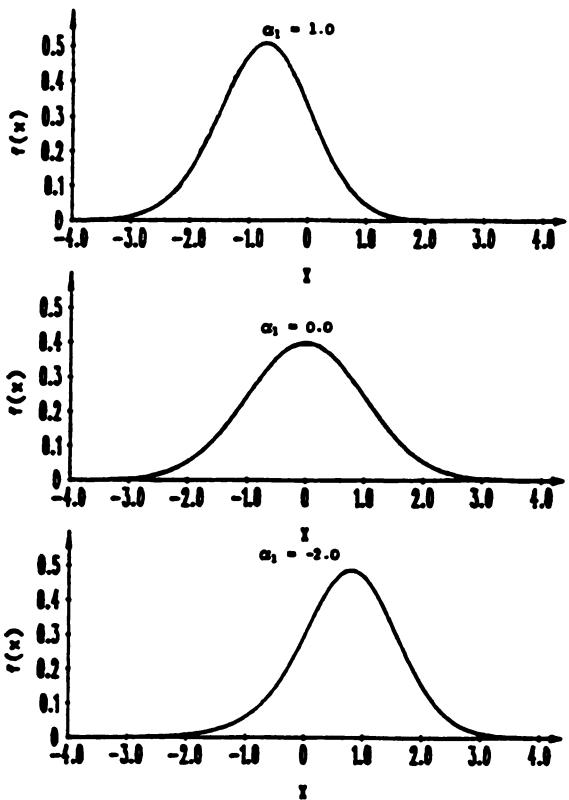
FIGURE 8 Density (7) Constructed from N(0,1); p = 1; α_1 = 1.0, 0.0, - 2.0
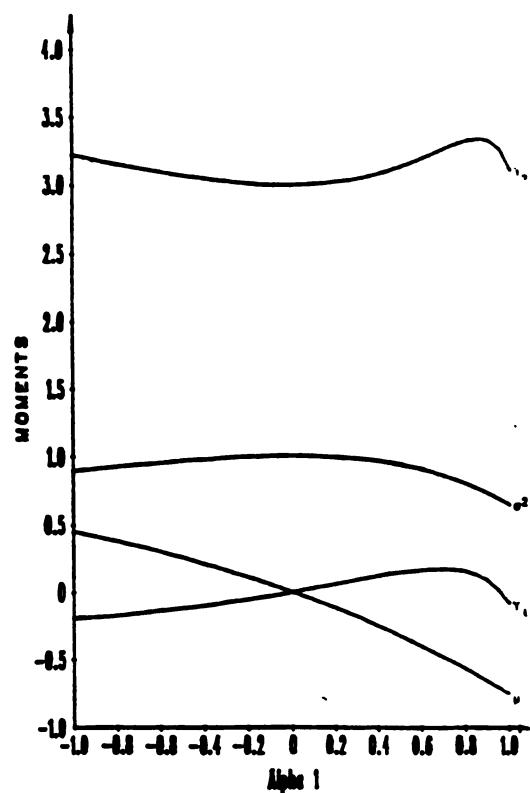

FIGURE 9 Moments and Moment Ratios of Density (7) Constructed from N(0,1) as a Function of α_1
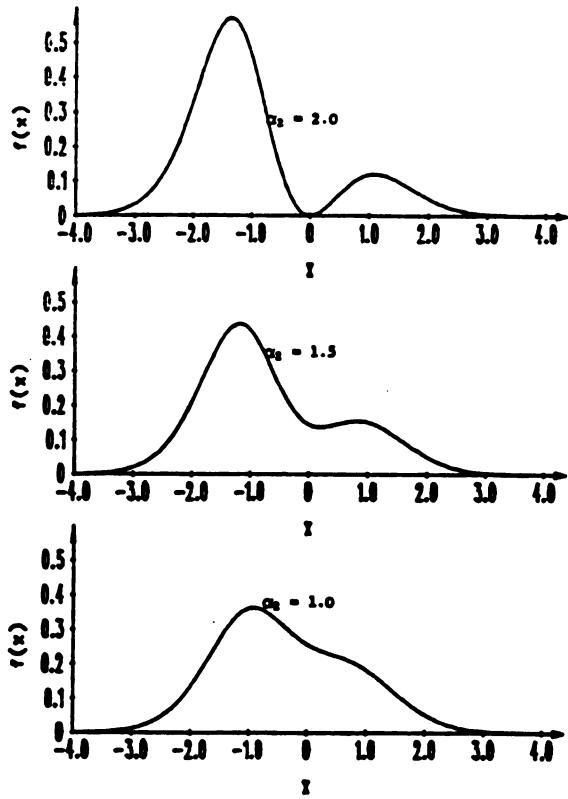
FIGURE 10 Density (7) Constructed from N(0,1); p = 2; α_2 = 2.0, 1.5, 1.0
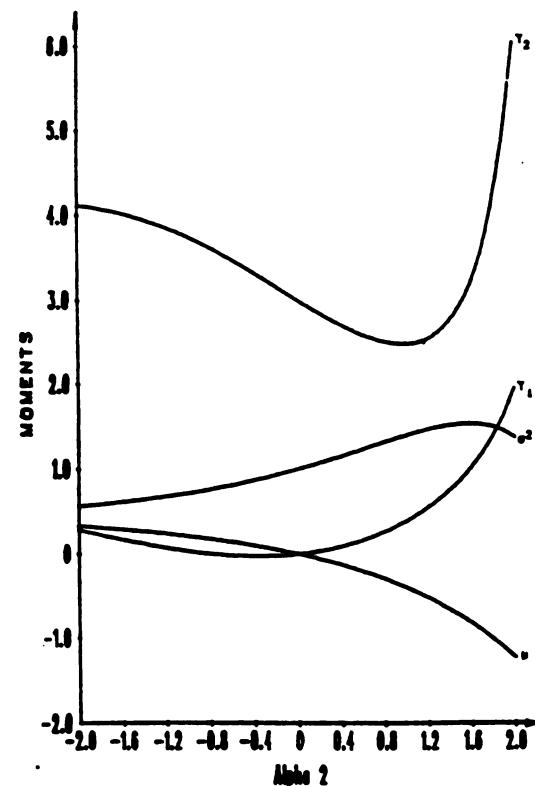
FIGURE 11 Moments and Moment Ratios of Density (7) Constructed from N(0,1) as a Function of α_2
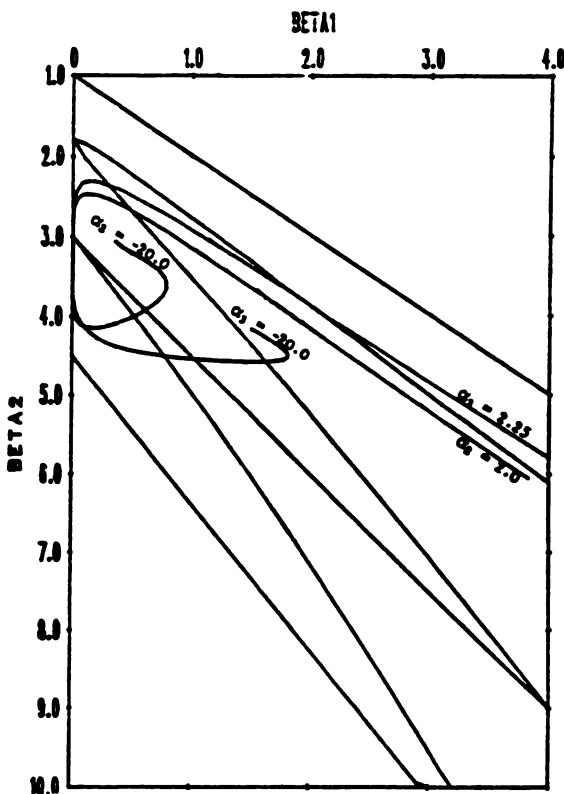
FIGURE 12 The Effect of α_2 and α_3 on Density (7)
Constructed from N(0,1) in the β_1 - β_2 Plane
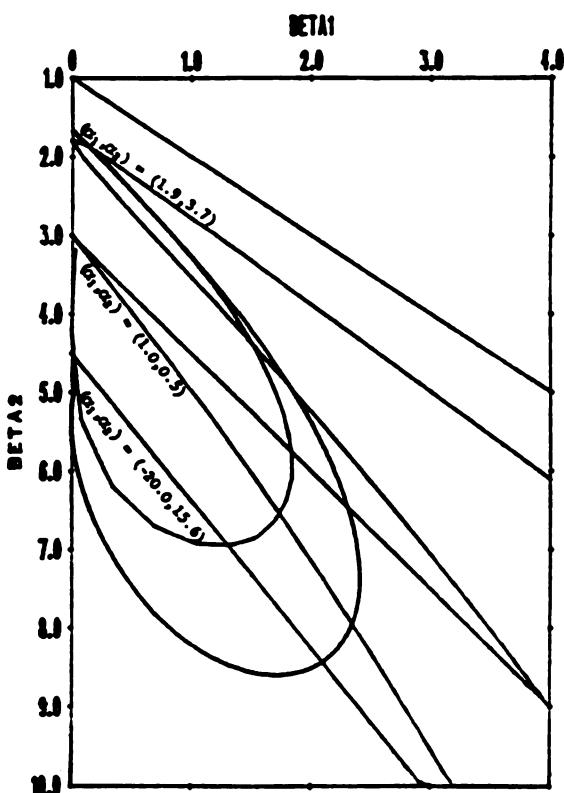
FIGURE 13 The Effect of α_1 and Maximumm of α_2 on Density (8)
Constructed from N(0,1) in the β_1 - β_2 Plane

188

Here the minimum value of f is at $x = |A| \dfrac{C + B/C}{A} - \dfrac{D_1}{3}$, where $A = \dfrac{2D_1^3 - 9D_1D_2 - 27/D_0}{54}$,

$B = \dfrac{D_1^2 - 3D_2}{9}$, and $C = [(A^2 - B^3)^{1/2} + |A|]^{1/3}$. Here, $D_0 = p^3 + 4p^2$, $D_1 = -\dfrac{5p^2 + 4p}{D_0}$, and

$D_2 = \dfrac{4p + 1}{D_0}$. We see $\lim\limits_{p \to 1} \alpha_p = -1$ and $\lim\limits_{p \to \infty} \alpha_p = 2 - \beta + \dfrac{(\beta^2 - 4\beta)^{1/2}}{2\beta} e^{\beta^2} = -646.7719769\ldots$,

where $\beta = [\dfrac{97}{54} + \dfrac{(23/3)^{1/2}}{6}]^{1/3} + [\dfrac{97}{54} - \dfrac{(23/3)^{1/2}}{6}]^{1/3} + \dfrac{5}{3}$ (Myers, 1990).

We give an example of (9) with gaussian margins in Figure 14. We see non-elliptic contours, unlike those of the standard correlated bivariate gaussian. In Figure 15, we present the effect of p and α_p on the Pearson product-moment correlation coefficient, ρ, when (10) is constructed from gaussian margins. We see negative/positive values of α_p producing negative/positive values of ρ. Figures 16 and 17 show the parameters effect on Mardia's (1970b, 1974) multivariate measures of skewness and kurtosis, $\beta_{1,2}$ and $\beta_{2,2}$. As expected, $(\rho, \beta_{1,2}, \beta_{2,2}) = (0,0,8)$ when $\alpha_p = 0$ for all values of p.

5.2. Examples of the Second Bivariate Construction. The density for (5) is

$$f = \{1 + \alpha_1(GH - G\bar{H} - \bar{G}H) - 2\alpha_2(G + H)\bar{G}\bar{H} + [\alpha_1^2 + 8\alpha_2 + 4\alpha_1\alpha_2\bar{G}\bar{H} + 4(\alpha_2\bar{G}\bar{H})^2]GH\bar{G}\bar{H}\}$$

$$gh \exp[\alpha_1\bar{G}\bar{H} + \alpha_2(\bar{G}\bar{H})^2]. \tag{10}$$

For (5) to be a d.f., again, the first three requirements are immediate and the final requirement, from Section 5.1, yield limits on the parameters α_1 and α_2. We find $(-1 \le \alpha_1 < 1)$ and $(\alpha_2 \le 1 - \alpha_1/2 + \sqrt{1 - \alpha_1})$. We see the upper limit of α_2 as a function of α_1 coincides with that of univariate construction (3). As a closed form for the lower limit of α_2 has eluded us, numerical computations have provided values for lower bounds. Table 1 gives the minimum α_2 as a function of representative values of α_1.

α_1	-1.0	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.0
Min α_2	-9.927	-10.223	-10.460	-10.655	-10.815	-10.947	-11.053	-11.138	-11.204	-11.252	-11.284

α_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Min α_2	-11.302	-11.306	-11.298	-11.278	-11.247	-11.206	-11.155	-11.095	-11.027	-10.951

TABLE 1 Minimum α_2 as a function of α_1

Graphs of (10) constructed from U(0,1) and N(0,1) margins are presented in Figures 18 through 21. Figures 22 and 23 are contour graphs of r as a function of α_1 and α_2 for the uniform and gaussian cases respectively. Figures 24 and 25 present contours of Mardia's multivariate moment ratios, $\beta_{1,2}$ and $\beta_{2,2}$, for (10) constructed from N(0,1). Again, $(\rho, \beta_{1,2}, \beta_{2,2}) = (0,0,8)$ when $(\alpha_1, \alpha_2) = (0,0)$. We see (10) to be very flexible, allowing for the construction of skewed and kurtic surface with negative and positive correlation.
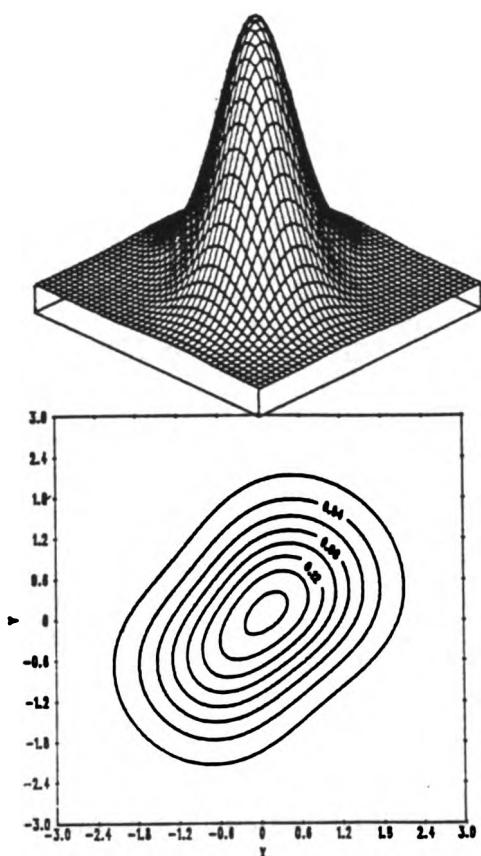
FIGURE 14 Density (9) Constructed from
N(0,1), Preserving the Margins,
$(p, \alpha_p) = (1.0, 1.0)$, $\rho = 0.36$
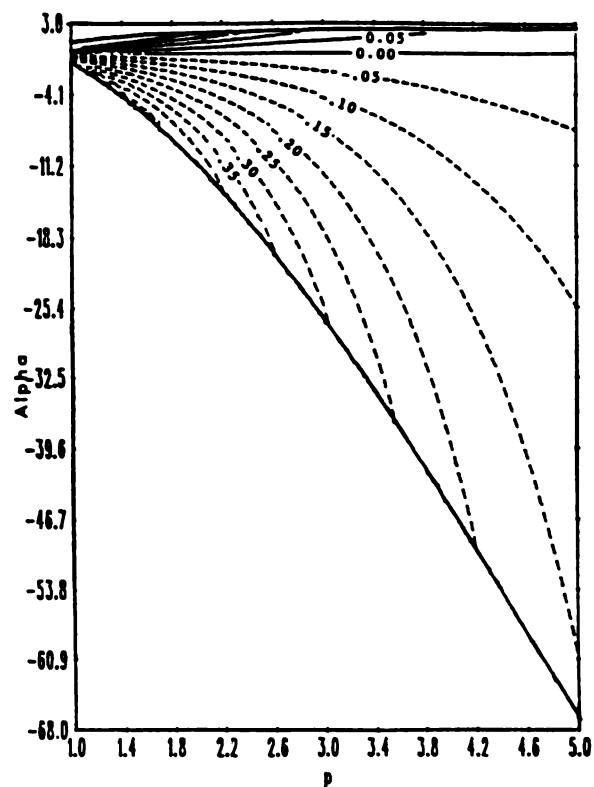
FIGURE 15 Contour Graph of ρ as a function of
p and α_p for Density (9)
Constructed from N(0,1)
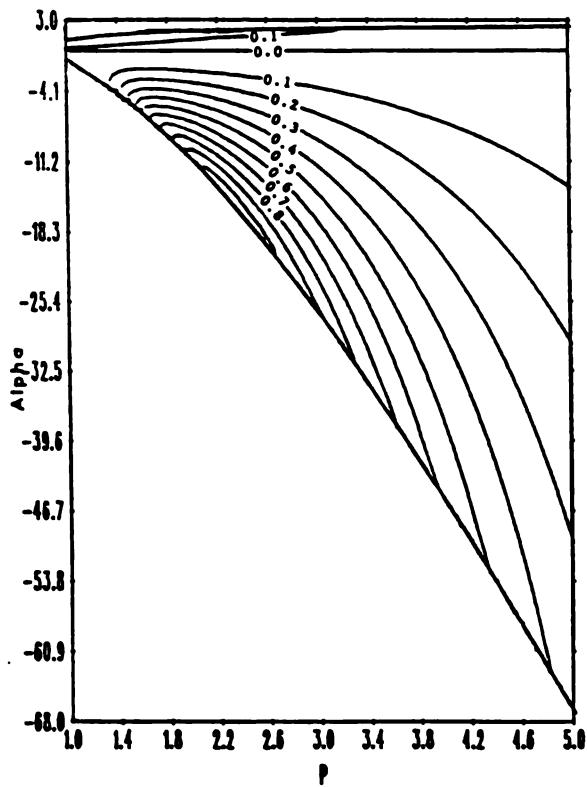
FIGURE 16 Contour Graph of $\beta_{1,2}$ as a function
of p and α_p for Density (9)
Constructed from N(0,1)
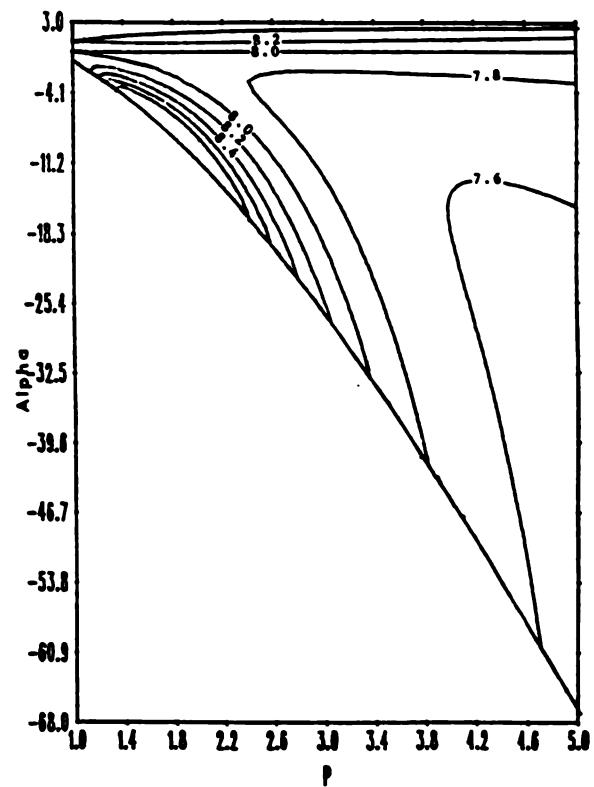
FIGURE 17 Contour Graph of $\beta_{2,2}$ as a function
of p and α_p for Density (9)
Constructed from N(0,1)
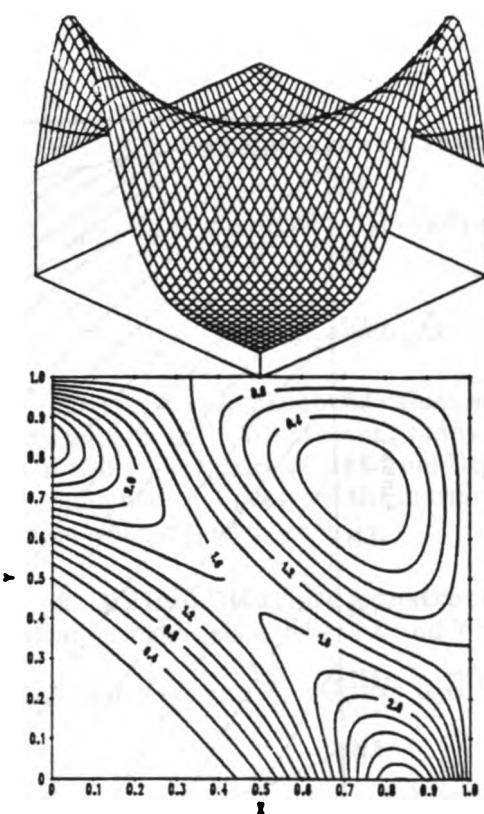
FIGURE 18 Density (10) Constructed from U(0,1), Preserving the Margins, $(\alpha_1, \alpha_2) = (0.0, -11.25)$, $\rho = -0.52$
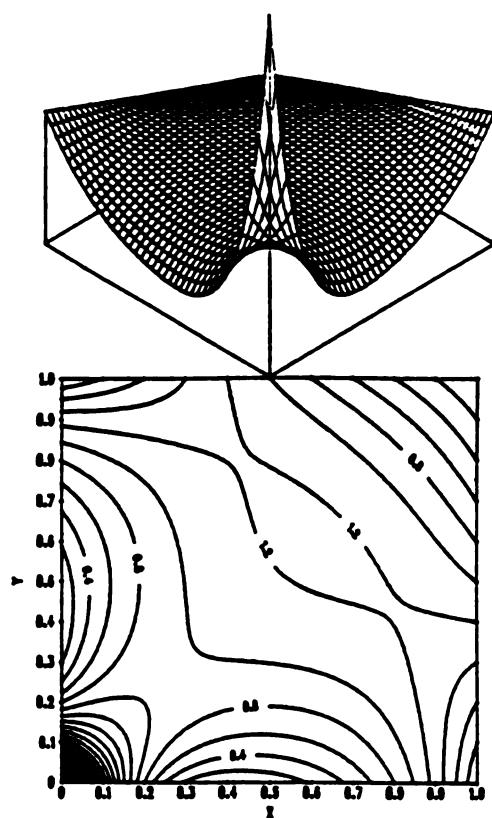
FIGURE 19 Density (10) Constructed from U(0,1), Preserving the Margins, $(\alpha_1, \alpha_2) = (-1.0, 2.90)$, $\rho = -0.07$
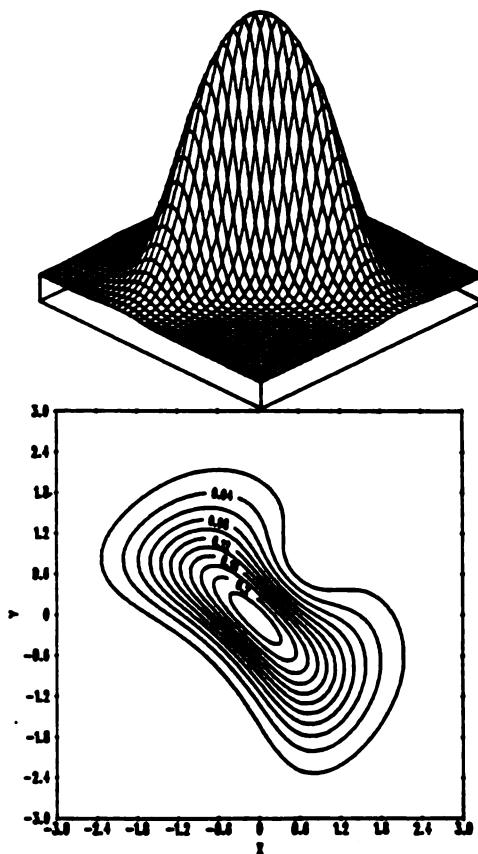
FIGURE 20 Density (10) Constructed from N(0,1), Preserving the Margins, $(\alpha_1, \alpha_2) = (-1.0, -9.90)$, $\rho = -0.59$
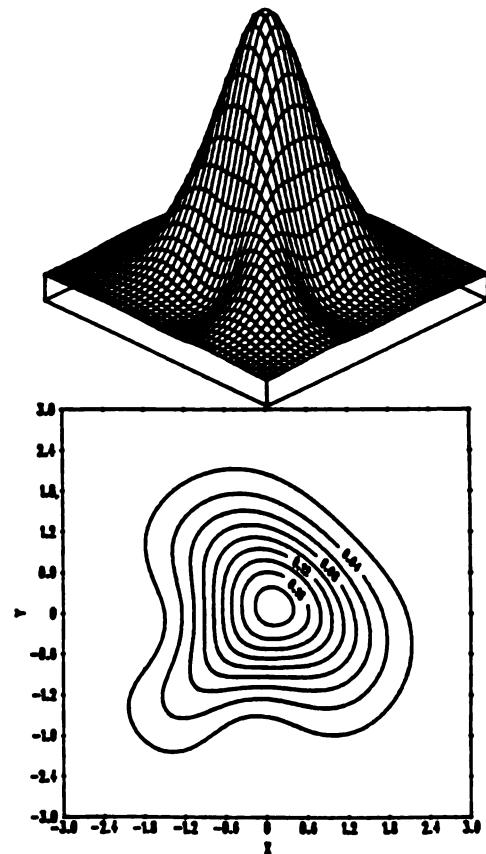
FIGURE 21 Density (10) Constructed from N(0,1), Preserving the Margins, $(\alpha_1, \alpha_2) = (-1.0, 2.90)$, $\rho = -0.06$

191
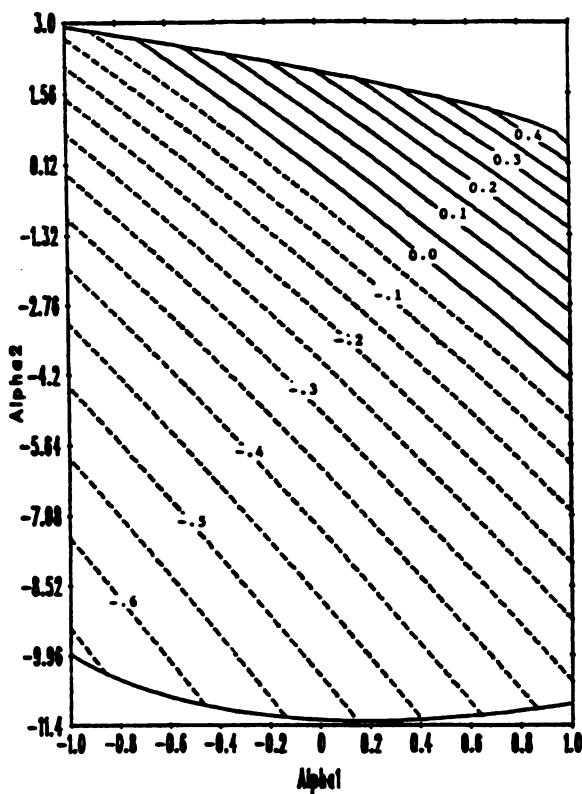
FIGURE 22 Contour Graph of ρ as a function of
α_1 and α_2 for Density (10)
Constructed from U(0,1)
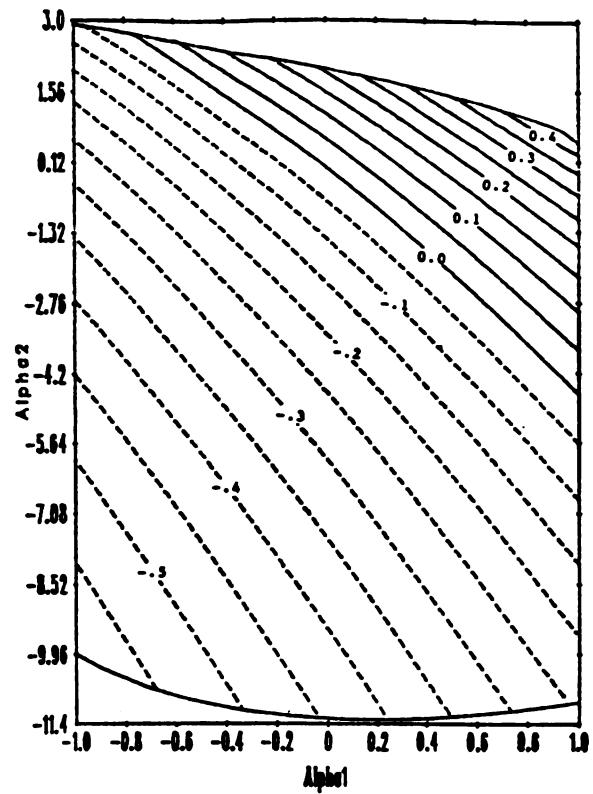
FIGURE 23 Contour Graph of ρ as a function of
α_1 and α_2 for Density (10)
Constructed from N(0,1)
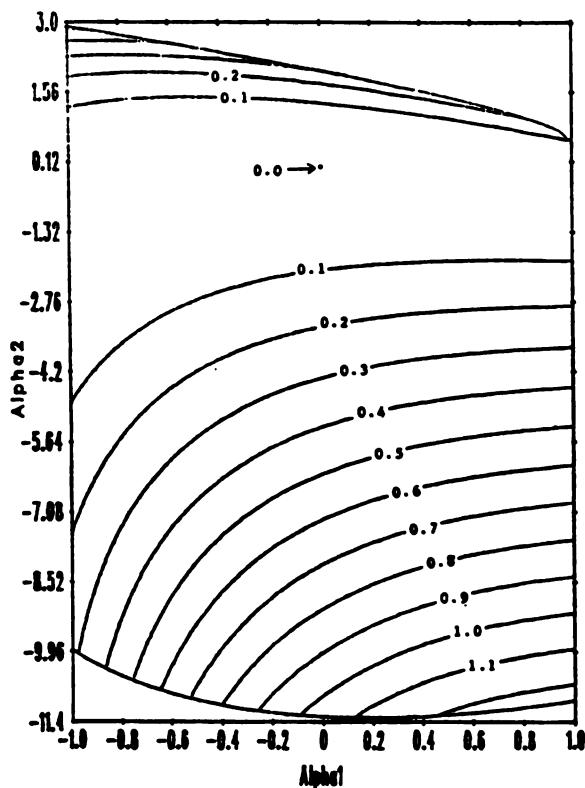
FIGURE 24 Contour Graph of $\beta_{1,2}$ as a function
of α_1 and α_2 for Density (10)
Constructed from N(0,1)
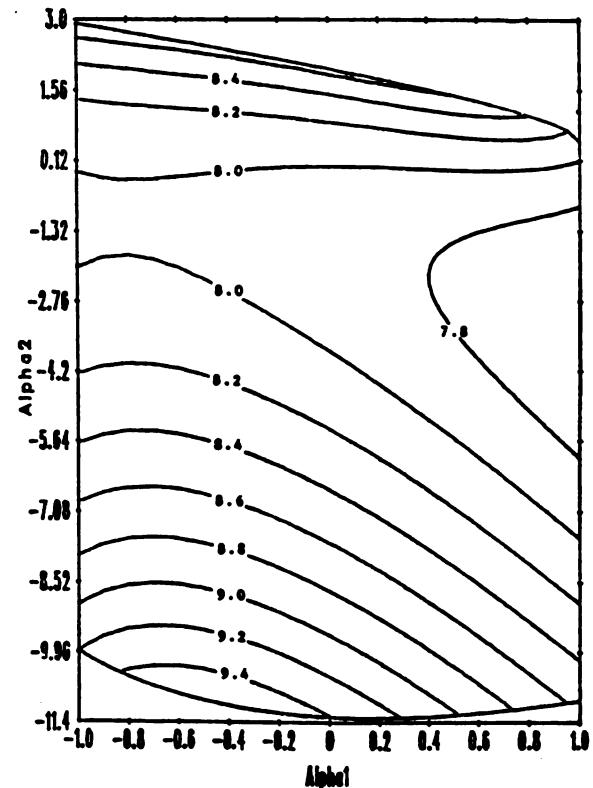
FIGURE 25 Contour Graph of $\beta_{2,2}$ as a function
of α_1 and α_2 for Density (10)
Constructed from N(0,1)

5.3. Examples of the Third Bivariate Construction. The density for (6) is

$$f = [1 - \alpha_{10}G - \alpha_{01}H + \alpha_{11}(GH - G\bar{H} - \bar{G}H) + (\alpha_{10}\alpha_{01} + \alpha_{10}\alpha_{11}\bar{G} + \alpha_{01}\alpha_{11}\bar{H} + \alpha_{11}^2\bar{G}\bar{H})GH]$$

$$gh \exp(\alpha_{10}\bar{G} + \alpha_{01}\bar{H} + \alpha_{11}\bar{G}\bar{H}). \tag{11}$$

Little is known about the values of the parameters required to preserve the nonnegativity of (11). We present graphs of (11) constructed from U(0,1) in Figures 26 and 27, from N(0,1), in Figure 28 and from Exp(1) in Figure 29. We recall from part 3 that this construction does not preserve the margins. We note changes in the marginal means and variances when α_{10} or $\alpha_{01} \neq 0$.

6. A Final Bivariate Construction. The final bivariate form of (1) we examine is a bivariate construction, where X and Y have marginal d.f.'s (2) and joint d.f. (4). We have
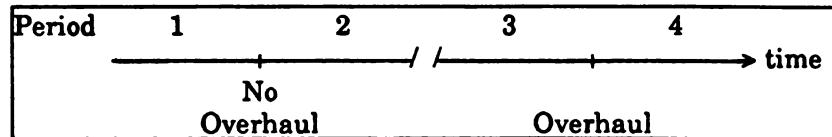
$$F_x = G_x \exp(\alpha_x \bar{G}_x^{P_x}) \text{ and } F_y = G_y \exp(\alpha_y \bar{G}_y^{P_y}),$$ resulting in the bivariate distribution,

$$F_{xy} = F_x F_y \exp[\alpha_{xy}(\bar{F}_x\bar{G}_y)^{P_{xy}}]. \tag{12}$$

Several graphs of (12)'s density constructed from the gaussian distribution are presented in Figures 30 through 33. The robustness of this construction is exhibited in these graphs. We are able to construct both noncorrelated and negative or positive correlated surfaces which are skewed or non-skewed and either uni-, bi-, tri-, or quadra-modal.

While not presented in this paper, we have investigated several forms of (9), (10), (11) and (12) with margins (7) and (8) constructed from Beta, Cauchy, Gamma, Gaussian, Laplace, Logistic, Uniform, and Weibull densities. It was recreative to observe the resulting surfaces.

7. An Innovative Approach to the Modeling of Bivariate Data. Here we exhibit the utility of the proposed system by fitting aircraft operations and maintenance data. Periodically aircraft undergo large scale overhaul programs. This presumes that the aircraft are restored to a better operating condition. The r.v.'s are defined to be the number of aircraft which suffer n aborts in a six month period. Aborts are mission interruptions occurring during pre-flight or in-flight operations. We consider the bivariate case of two consecutive six month periods. The following diagram indicates that between periods 1 and 2 there is no intervening overhaul and that an overhaul occurs between periods 3 and 4.

Period	1	2	3	4	
		No Overhaul		Overhaul	→ time

Thus the bivariate r.v.'s considered are A_{ij}, $(i,j) = (1,2)$ or $(3,4)$; the number of aircraft suffering x aborts in period i and y aborts in period j. Table 2 gives the data for 203 aircraft in periods (1,2). The sample is obtained by considering the entire inventory of a particular aircraft (about 500) and excluding those which are overhauled during periods 1 and 2, and those without 12 full months of data during this time - 203 result (Mitchell, 1976). As an illustration of the data, three aircraft have ten aborts in period 1 followed by six in period 2.

193
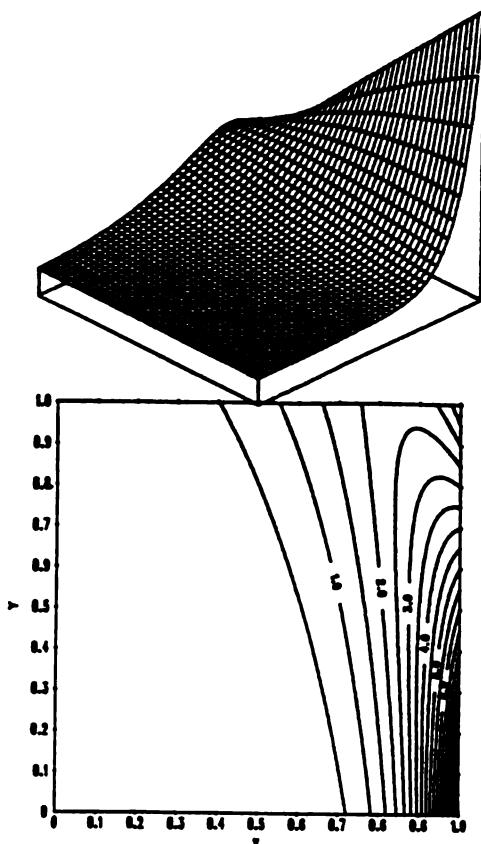
FIGURE 26 Density (11) Constructed from U(0,1),
$(\alpha_{10}, \alpha_{01}, \alpha_{11}) = (-5.0, 0.0, -5.0)$, $\rho = -0.34$,

$(\mu_x, \mu_y) = (0.84, \frac{1}{2})$, $(\sigma_x^2, \sigma_y^2) = (0.02, \frac{1}{12})$
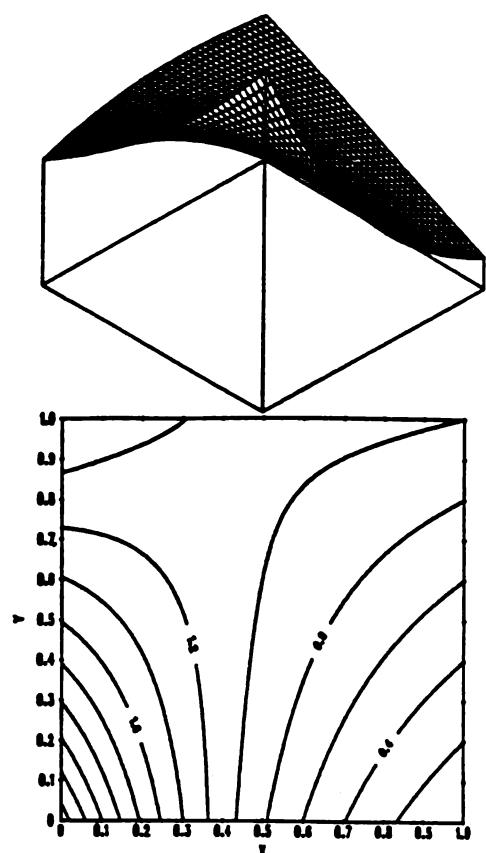
FIGURE 27 Density (11) Constructed from U(0,1),
$(\alpha_{10}, \alpha_{01}, \alpha_{11}) = (0.5, 0.0, 0.5)$, $\rho = 0.24$,

$(\mu_x, \mu_y) = (0.41, \frac{1}{2})$, $(\sigma_x^2, \sigma_y^2) = (0.08, \frac{1}{12})$
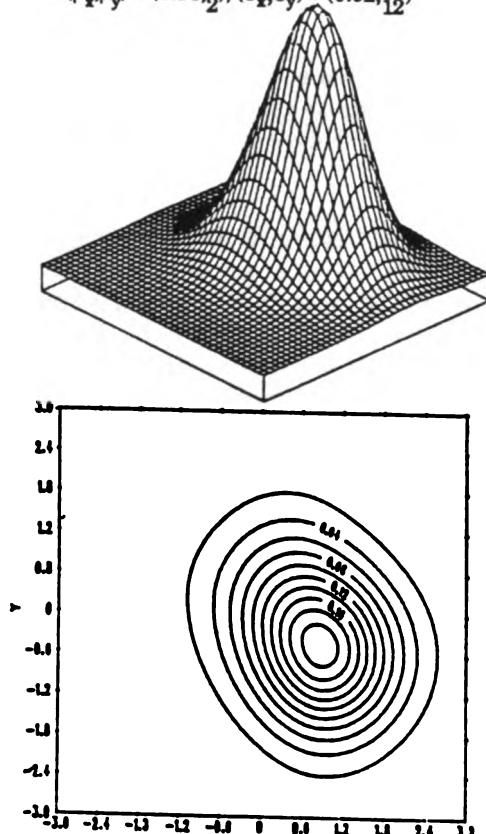
FIGURE 28 Density (11) Constructed from N(0,1),
$(\alpha_{10}, \alpha_{01}, \alpha_{11}) = (-2.0, 0.5, -1.0)$, $\rho = -0.47$,

$(\mu_x, \mu_y) = (0.74, -0.32)$, $(\sigma_x^2, \sigma_y^2) = (0.76, 0.94)$
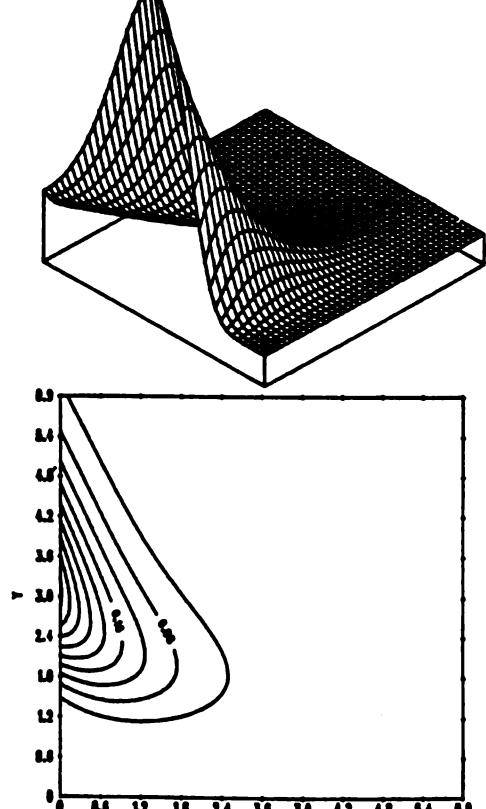
FIGURE 29 Density (11) Constructed from Exp(1),
$(\alpha_{10}, \alpha_{01}, \alpha_{11}) = (0.0, -10.0, -10.0)$, $\rho = -0.31$,

$(\mu_x, \mu_y) = (1.00, 2.98)$, $(\sigma_x^2, \sigma_y^2) = (1.00, 1.63)$

194
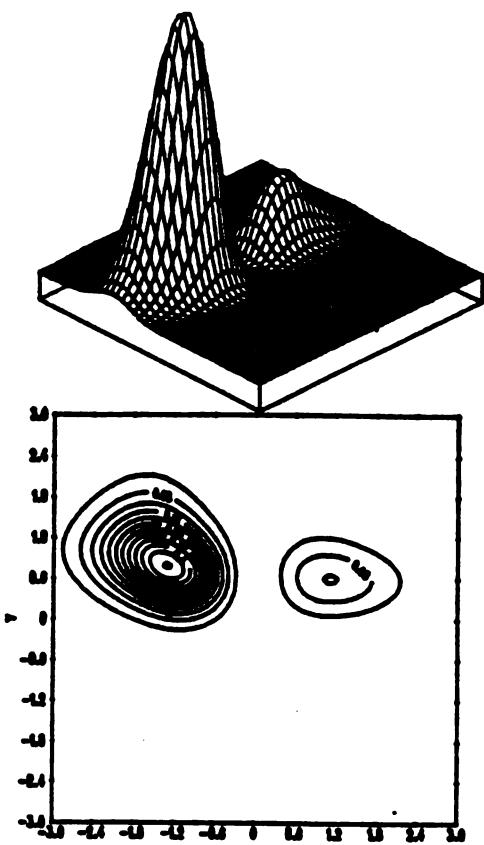
FIGURE 30 Density of (12) Constructed from N(0,1),
$(p_x, p_y, p_{xy}) = (2.0, 2.0, - 1.0)$,
$(\alpha_x, \alpha_y, \alpha_{xy}) = (2.00, - 11.28, 1.00)$, $\rho = - 0.27$,
$(\mu_x, \mu_y) = (- 1.00, 0.87)$, $(\sigma_x^2, \sigma_y^2) = (1.47, 0.27)$,

FIGURE 31 Density of (12) Constructed from N(0,1),
$(p_x, p_y, p_{xy}) = (4.0, 2.0, 3.0)$,
$(\alpha_x, \alpha_y, \alpha_{xy}) = (2.37, 2.00, - 26.95)$, $\rho = - 0.21$,
$(\mu_x, \mu_y) = (- 0.63, - 1.00)$, $(\sigma_x^2, \sigma_y^2) = (2.11, 1.47)$,

FIGURE 32 Density of (12) Constructed from N(0,1),
$(p_x, p_y, p_{xy}) = (2.0, 2.0, 0.0)$,
$(\alpha_x, \alpha_y, \alpha_{xy}) = (2.00, 2.00, 0.0)$, $\rho = 0.0$,
$(\mu_x, \mu_y) = (- 1.00, - 1.00)$, $(\sigma_x^2, \sigma_y^2) = (1.47, 1.47)$

FIGURE 33 Density of (12) Constructed from N(0,1),
$(p_x, p_y, p_{xy}) = (4.0, 4.0, 4.0)$,
$(\alpha_x, \alpha_y, \alpha_{xy}) = (2.00, 2.00, - 46.11)$, $\rho = - 0.23$,
$(\mu_x, \mu_y) = (- 0.47, - 0.47)$, $(\sigma_x^2, \sigma_y^2) = (1.85, 1.85)$

195

Digitized by Google

Figure — axis labels and data grid for Table 2:

Number of Aircraft with Y Aborts in Period 2: A_2 (vertical axis)
Number of Aircraft with X Aborts in Period 1: A_1 (horizontal axis, 0 to 30)

```
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
1 0 0 1 0 0 0 0 0 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 3 0 2 0 0 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 4 3 1 0 1 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 1 2 1 3 2 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
3 2 1 1 1 0 1 1 0 2 3 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
1 0 2 2 1 3 0 2 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 3 3 0 1 2 3 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0
3 4 4 1 3 3 1 0 4 2 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 3 4 3 0 0 2 0 4 2 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 2 2 2 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 3 0 1 1 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

TABLE 2 Observed Values of Abort Data for 203 Aircraft, No Intervening Overhaul

Figure — axis labels and data grid for Table 3:

Number of Aircraft with Y Aborts in Period 4: A_4 (vertical axis)
Number of Aircraft with X Aborts in Period 3: A_3 (horizontal axis, 0 to 35)

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 1 1 0 0 2 1 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 2 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 1 0 1 2 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 2 1 0 2 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 0 1 0 2 0 1 0 0 0 1 0 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 2 4 1 1 0 1 2 2 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 3 3 4 1 1 2 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 4 1 1 1 0 3 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 4 2 1 0 2 0 2 1 2 1 2 0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 1 2 1 2 0 0 2 0 2 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 0 2 3 0 1 0 2 1 0 0 1 0 0 1 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
2 5 4 3 3 0 1 0 0 0 1 3 2 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
5 1 2 4 3 6 1 3 1 4 1 0 0 1 0 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0
1 1 2 4 1 2 4 2 2 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 4 2 2 0 1 3 1 3 2 1 0 0 3 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 1
3 2 4 1 3 2 1 2 0 1 0 0 2 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 5 6 3 2 2 4 3 0 3 0 1 2 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 2 1 1 2 2 3 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 3 1 1 1 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

TABLE 3 Observed Values of Abort Data for 387 Aircraft, Intervening Overhaul.

Table 3 gives the data for 387 aircraft of the same type. Here an aircraft is included if it has six full months of reported data during adjacent time periods, (3,4), to a common overhaul.

The margins of each of the bivariate distributions are first assumed to be the negative binomial p.m.f. given by Johnson and Kotz (1969),

$$\Pr[X = x] = \frac{(x + a - 1)!}{x!(a - 1)!} \frac{1}{1 + b} \frac{b}{1 + b} x \; ; a > 0, b > 0, x = 0,1,2,\ldots . \tag{13}$$

Table 4 gives descriptive statistics for the data and it is apparent that the univariate negative binomial distribution is an appropriate choice to begin modeling the margins.

We select distribution (12) with margins (2) constructed from (13) to fit the bivariate data. We simultaneously estimated the ten parameters using least squares invoking a NAG (Numerical Algorithm Group 1983) nonlinear minimization routine. We originally estimated the parameters by considering only the sum of squared errors, (SSE), between the empirical and expected p.d.f.'s. While the SSE's were smaller than those presented in Table 4, the resulting bivariate models failed to approximate the marginal and bivariate moments. Thus we minimized an objective function composed of the usual SSE of the p.d.f.'s plus weighted SSE's of the margins' first two moments and the first bivariate mixed moment. The parameter estimates and results are given in Table 4. Thus we were able to consider not only the usual ordinary least squares but also the discrepancies between the margins' means and variances and the bivariate correlation. We attempted to include the higher univariate moments, β_1 and β_2, and Mardia's bivariate moments, $\beta_{1,2}$ and $\beta_{2,2}$, in the weighted SSE's , but these efforts were not fruitful due to the large variances of the higher order moments. Typically we observed agreement between the observed and expected moments and moment ratios but the expected observations were about ten percent less than the empirical.

We see in Table 4 a very nice fit, not only of total observations, but of the margins' means and variances and of the bivariate correlations. Of interest is the unexpected values of the estimated parameters for the 203 aircraft data set. We attempted to set α_x for margin $A_1(x)$ to zero but found the results to be extremely sensitive to this parameter. We can see the effects α_p and p, as indicated in part 3 above, by the estimates for margin $A_2(4)$. The original method of moment parameters are (a,b) = (2.01,2.98).

8. Summary. We have presented an intuitively appealing system of multivariate distributions which has the Neyman alternative as its genesis. We have considered several univariate and bivariate versions and a combination of versions to demonstrate the flexibility of construction. Finally, we used an innovative approach which captures the moments to model data which had eluded successful fitting for over two decades.

Further versions of the system need to be considered. We have investigated versions of (1) when $A(\bar{\Theta}_i)$ consisted of transcendental functions, but little progress has been made. Higher dimensional versions of the system must be explored. While we have obtained two trivariate densities through arduous differentiation, little is known about them. A clearer fitting technique is required, such as maximum likelihood. Finally a test of fit specific to the system is required. The distribution of the statistic is not known at this time. If random variates are required, we can simulate the system through the simulation of the margins used in its construction. With these advances, we will have a new and complete multivariate system for modeling and inference.

| Aircraft Total Aborts Without Intervening Overhaul (A_1, A_2) | | | | | | Aircraft Total Aborts With Intervening Overhaul (A_3, A_4) | | | | |

Observed Values

Margin	Total	μ	σ^2	ρ		Margin	Total	μ	σ^2	ρ
$A_1(x)$		5.61	25.77			$A_3(x)$		7.07	45.05	
	203			.29			387			.23
$A_2(y)$		6.00	23.58			$A_4(y)$		8.95	45.18	

Expected Values and Sum of Square Errors

Margin	Total	μ	σ^2	ρ	SSE	Margin	Total	μ	σ^2	ρ	SSE
$A_1(x)$		5.90	24.59			$A_3(x)$		7.26	40.58		
	201.5			.29	175.6		390.0			.23	387.6
$A_2(y)$		6.34	22.32			$A_4(y)$		9.34	44.65		

Estimated Parameters

Margin	Equation (13)		Equation (7)		Equation (9)	
	a	b	ρ	α	ρ	α
$A_1(x)$	2.38	2.34	103.71	33.39×10^{-5}		
					1.00	0.85
$A_2(y)$	5.49×10^{-2}	17.94	2.09	-384.33		
$A_3(x)$	1.81	4.49	1.70	0.58		
					1.00	0.86
$A_4(y)$	2.00	4.47	1.00	-0.15		

TABLE 4.   Descriptive Statistics and Estimates of Parameters for Bivariate Abort Data.

REFERENCES

Abramowitz, M. and Stegun, I. A., eds. (1973). Handbook of Mathematical Functions. Dover Publications, Inc., New York.

Ali, M. M., Mikhail, N. N. and Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. Journal of Multivariate Analysis, 8, 405-412.

Beam, K. M. (1989). Constructive multivariate methods for data analysis and modeling. Unpublished doctoral dissertation, Rensselaer Polytechnic Institute.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65, 141-151.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. Journal of the Royal Statistical Society, Ser. B, 47, 82-117.

Cook, R. D. and Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. Journal of the Royal Statistical Society, Ser. B, 43, 210-218.

Farlie, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution. Biometrika, 47 307-333.

Frank, M. J. (1979). On the simultaneous associativity of F(x,y) and x + y - F(x,y). Aequationes Mathematicae, 19, 194-226.

Galton, F. (1885). Family Likeness in Stature. Proceedings of the Royal Society, 40, 42 - 73.

Gumbel, E. J. (1960). Bivariate exponential distributions. Journal of the American Statistical Association, 55, 698-707.

Gumbel, E. J. (1961). Bivariate logistic distribution. Journal of the American Statistical Association, 56, 335-349.

Johnson, M. E. (1987). Multivariate Statistical Simulation. John Wiley and Sons, New York.

Johnson, N. L. and Kotz, S. (1969). Discrete Distributions. John Wiley and Sons, New York.

Johnson, N. L. and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions. John Wiley and Sons New York.

Mardia, K. V. (1970a). Families of Bivariate Distributions. Charles Griffin and Co., London.

Mardia, K. V. (1970b). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, 519-530.

Mardia, X. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya, Ser. B, 36, 115-128.

199

Marshall, A. W. , and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association.* 83, 834-841.

Mitchell, C. R. (1976). *On some new negative binomial and gamma distributions with applications to queues. inventories. and maintenance.* Unpublished doctoral dissertation, Rensselaer Polytechnic Institute.

Morgernstern, D. (1956). Simple examples of bivariate distributions. *Mitteilingsblatt fur Statistik,* 8, 234-235.

Myers, J. D. (1990). Derived the analytic lower limit of α_p as a function of p.

Neyman, J. (1937). 'Smooth' tests for goodness of fit. *Skandinavisk Aktuarietidskrift,* 20, 149-199.

Numerical Algorithms Group (1983). *NAG FORTRAN Library Manual: Mark 10.* Numerical Algorithms Group, Downer's Grove, IL.

Pearson, E. S. and Hartley, H. O., eds. (1970). *Biometrika Tables for Statisticians, Volume 1.* Cambridge University Press, London.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association,* 60, 516-522.

Sibuya, M. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics,* 19, 195-210.

EVALUATION OF COMMUNICATIONS THROUGHPUT

Virginia A.T. Kaste, Ann E.M.Brodeen, and Barbara D. Broome*
U S Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

Abstract

A controlled laboratory experiment was conducted during the summer of 1991 to evaluate the Combat Net Radio Network (CNRN) performance of several emulated Advanced Field Artillery Tactical Data System (AFATDS) nodes communicating via existing Tactical Fire Direction System (TACFIRE) protocol using Single Channel Ground and Airborne Radio System(SINCGARS). The purpose of this experiment was to examine the effects of four levels of message length and four levels of message transmission rate on network throughput and delay. In addition, radio transmission mode was considered by running the test once using single channel transmissions and once using frequency hopping. Three replications of a 4 x 4 full factorial design were made for each test. Analysis of variance techniques as well as other forms of network analysis were utilized to examine the significance and measure the effects of the three network parameters. These analyses provide information on communications thresholds for the TACFIRE protocol using CNRNs. These thresholds should be considered when designing AFATDS communication architectures and protocols. This paper examines the results with respect to throughput.

1. INTRODUCTION

The purpose of a network is to serve as a carrier of information from one point to another. To measure a network's effectiveness, one must determine whether the messages the network services arrive at their destination correctly and in time to be useful. We will refer to the amount of correctly passed information as "throughput" and the amount of time required to pass that information as "delay." There are a number of parameters that can impact throughput and delay, for example, the number of messages to transmit, the size of those messages, the number of nodes on the network, the communications protocol, and the communications hardware.

Simulation is a widely accepted means of examining the changes in network performance resulting from a change in hardware or communications protocol. Simulations, however, require input. They take information like the probability a message will collide, the expected delay in message transmission, or the arrival rate of messages at a given node,

and extrapolate those estimates, as appropriate, to a larger scenario. Only if the original assumptions and input are correct will simulation predictions have any validity.

On a limited bandwidth combat network, the number of nodes and the amount of information to pass can be large, especially during peak battle periods. Further, the impact of making decisions from old information can be catastrophic. If the interaction of the various network parameters is understood, the network's effectiveness can be optimized.

Protocols are often too complex to model precisely. Drastic assumptions are usually made to simplify the model resulting in an unrealistic representation of the protocol. Controlled experimentation with the actual protocol on the intended hardware offers much insight into the behavior of the protocol under various conditions and, therefore, facilitates modeling.

The purpose of this task was to quantify the effect of three parameters, message length, message arrival rate, and transmission mode, on throughput and delay on a small scale CNRN. An experiment was designed to allow an analysis of variance of these factors and to examine their significance. Further data summaries helped quantify the differences these factors made in network performance. The results of this experiment provide statistically sound baseline information which can be used as input for network simulations, as guidelines for designing communications architectures and protocols, and for future experiments on CNRNs.

The test was conducted on a CNRN, comprised of SINCGARS, using the TAC-FIRE protocol. Projecting these results to other hardware and other protocols is not necessarily valid. TACFIRE protocol was chosen for this experiment because the new AFATDS protocol is not yet available. This substitution provides an appropriate baseline, since future AFATDS will still be required to support the TACFIRE protocol.

2. TEST CONFIGURATION

The BRL's Firepower Control Facility was used to build, load, and monitor nodes emulating AFATDS communicating over Combat Net Radios.

There were four nodes, each of which was a SUN workstation. Each contained a message driver, providing a communications loading and a data collection program to log the sending and receipt of messages and acknowledgements as well as information on queues.

The nodes were connected to Magnavox Tactical Communications Modems (TCMs) to enable communications via SINCGARS. The TCMs allowed communication using the TACFIRE net sensing algorithm and communications protocol. Each TCM was then connected to a SINCGARS, which was pre–loaded with F200 hopsets for frequency hopping. Figure 1 illustrates this test configuration.

Potential message rates were reported in Magnavox's Level 1 AFATDS system performance study ("AFATDS Performance Modeling System Level for Multiservice Advanced Field Artillery Tactical Data System" 1991). BRL picked four values within the range of the reported potential message rates to emulate the rate of user generated messages as well as the user's system response to incoming messages. For this experiment, the arrival rate, λ, represented the number of messages generated during the hour and queued for transmission on the net, *not the number of messages actually transmitted during the hour*.

Digitized by Google
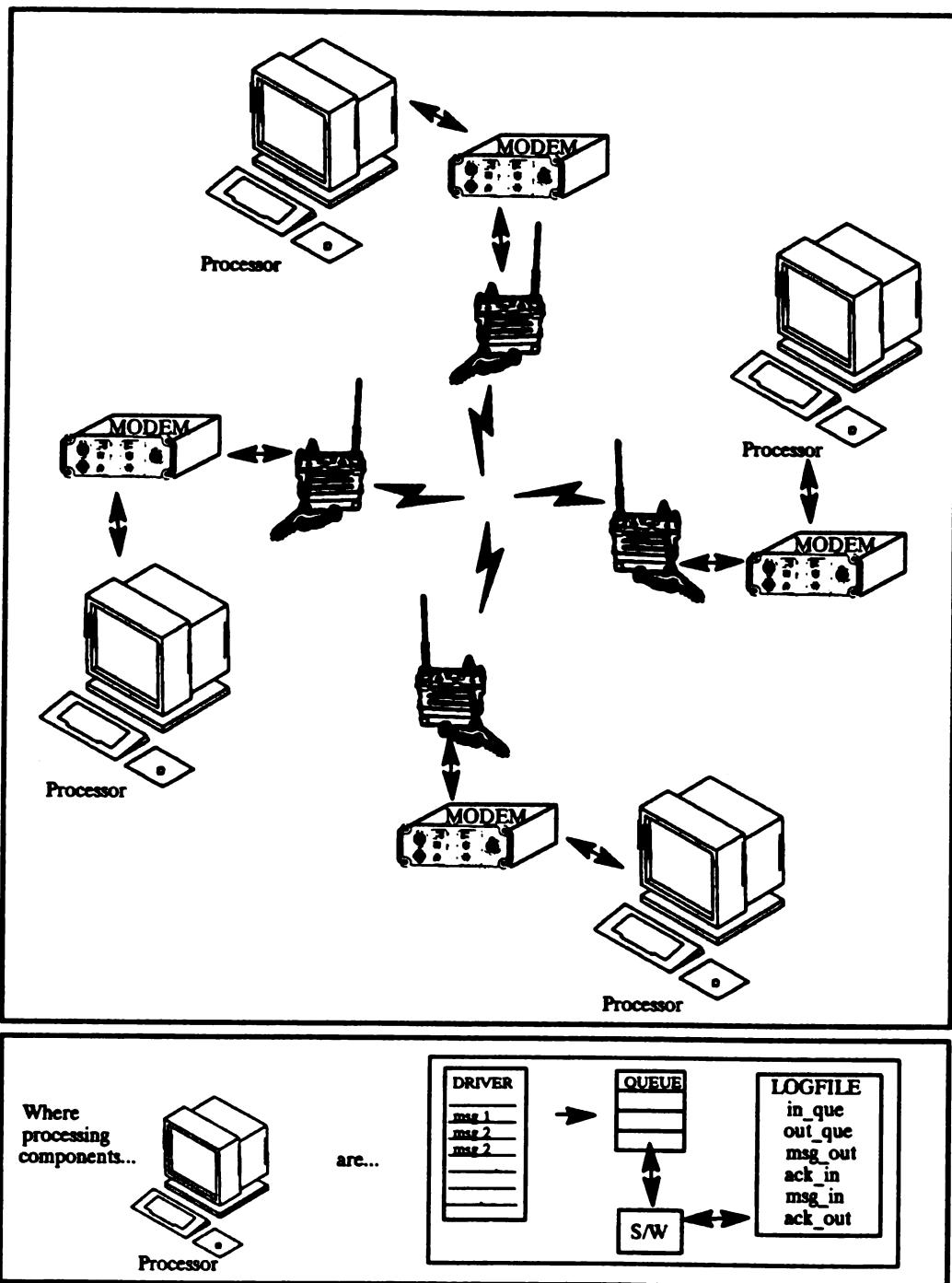
Figure 1. Experiment Configuration.

A scenario generator was written to create "messages" of character strings of a specified length and rate over a 1 hour period. In this baseline experiment message priorities were not considered.

Once the message is generated, the protocol adds several layers of information to insure the message arrives at its destination. This includes five error correction/detection bits for each seven–bit character, four synchronization characters, and a preamble (key-time) to bring the transmitter to full power before the message is sent. Acknowledgements, though shorter in length, are wrapped with similar overhead bits. Figure 2 illustrates the various portions of a transmitted message.
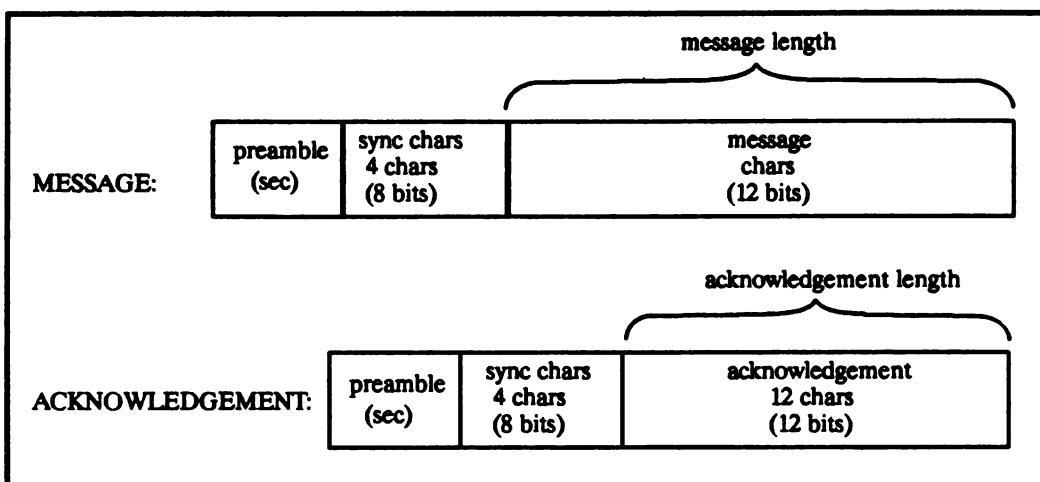
Figure 2. The components of messages and acknowledgements.

The numbers of messages generated for transmission each hour by each node, X_1, X_2, X_3, and X_4, were assumed to be mutually independent Poisson distributed random variables, X_i, with parameter λ_i. Using the property of convolutions and the uniqueness property of factorial moment generating functions, we can derive the sum of independent, nonnegative, integer valued random variables. The distribution of a sum of independent Poisson distributed random variables is a Poisson distributed random variable with a parameter equal to the sum of the individual parameters, $\Sigma\lambda_i$. From this, we can state that the arrival rate of messages for transmission to a network is the sum of the message arrival rates of each node on the network. Since our primary interest was to study the expected AFATDS network loading, this same theory allowed us to equally distribute the total expected loading among the four nodes during the experiment. For example, if the arrival rate was 2000 msgs/hr, then the scenario generator created four files of 500 msgs/hr for each node.

A network monitor with graphical displays provided testers a means of identifying operation anomalies, allowing early problem correction and effective test control. Figure 3 shows the net monitor display configuration. Messages and acknowledgements transmitted between nodes are dynamically displayed on the middle frame; the changing message queue
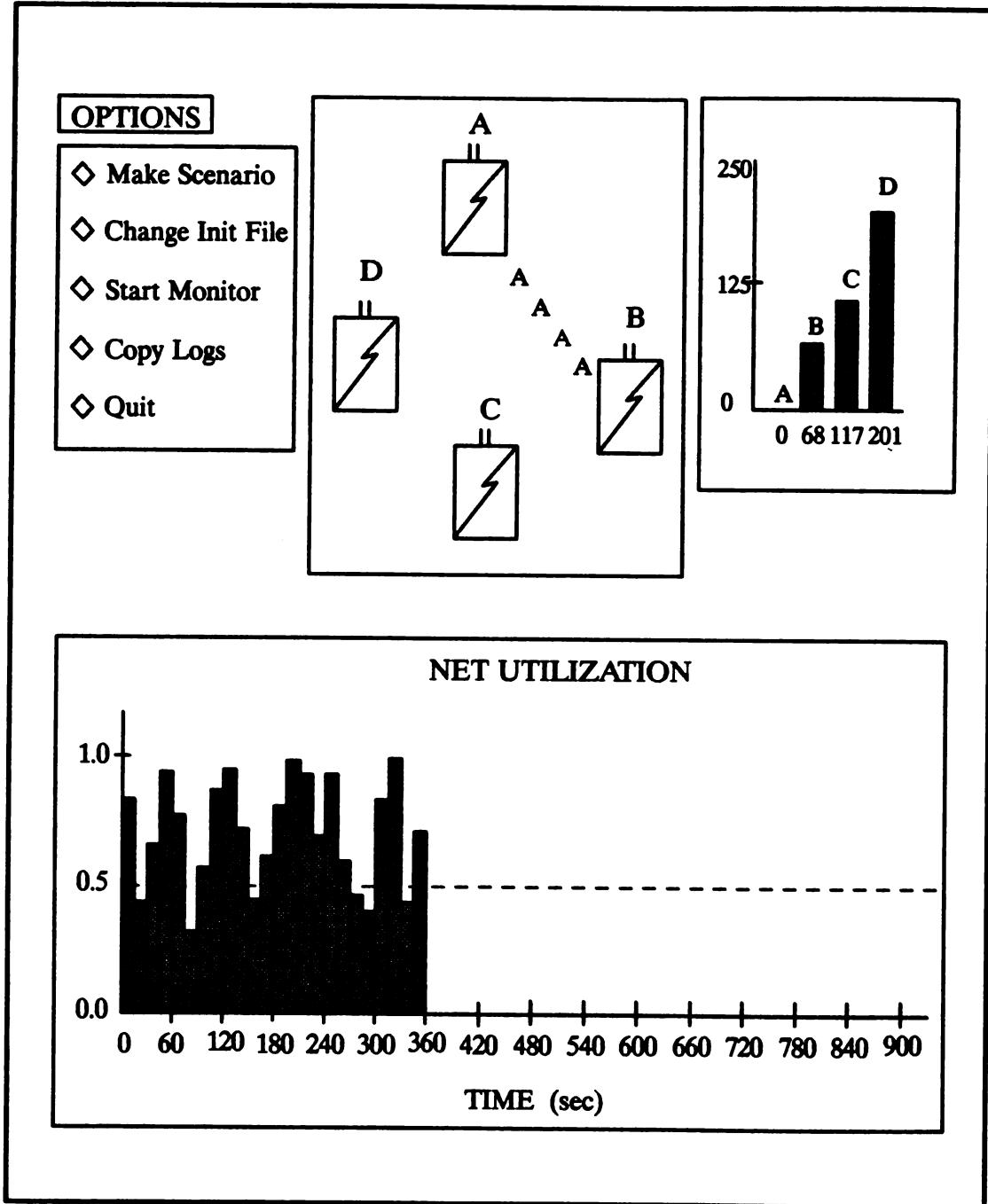
Figure 3. Net Monitor Displays

sizes are shown in the top right frame; and network utilization during the experiment is displayed in the bottom frame.

A message was assumed to enter network service when it reached the modem, as depicted by the area inside the dashed line in Figure 4. Thus, the server was considered a combination of modem and CNRN. The queue was the area outside the dashed line.

3. EXPERIMENTAL DESIGN

3.1 Factors

The two factors tested in each of the two separate experiments were message arrival rate and message length. Four levels of message arrival rate were tested with each of the four levels of message length, i.e. , full factorial design, yielding sixteen test combinations. Due to limitations in the TCMs, message lengths greater than 352 characters could not be considered, as the TCMs would lock up.

The levels of each factor are listed below:

1. Message Arrival Rate (for each node)
   100 per hour
   250 per hour
   350 per hour
   500 per hour
2. Message Length (in characters)
   48
   144
   256
   352

The maximum rate of 2000 messages per hour represented the heaviest traffic expected for AFATDS.

3.2 Design Matrix

It was decided that the shortest reasonable time to test any one of the 16 test combinations was one hour. Since the testing of all 16 test combinations would have required a minimum of 16 hours (for a single replication), which realistically could not be completed in one day, a randomized incomplete block design was constructed in order that day–to–day variability would not influence the results. The 16 test combinations were divided into blocks of size four, and the four blocks were run over a four day period. The assignment of the test combinations into blocks was based on a confounding scheme. This scheme, in which a different set of three of the nine degrees of freedom for the interaction term were completely confounded within each replication, assured that the effects of message arrival rate and message length and the interaction of these two factors on communication throughput could be measured. Three replications of the design matrix were made to insure the incomplete block design was balanced, thereby facilitating the analysis, although part of the precision of the estimate of the interaction effect was sacrificed (i.e., the relative information available with regard to the interaction term was 2/3).

206
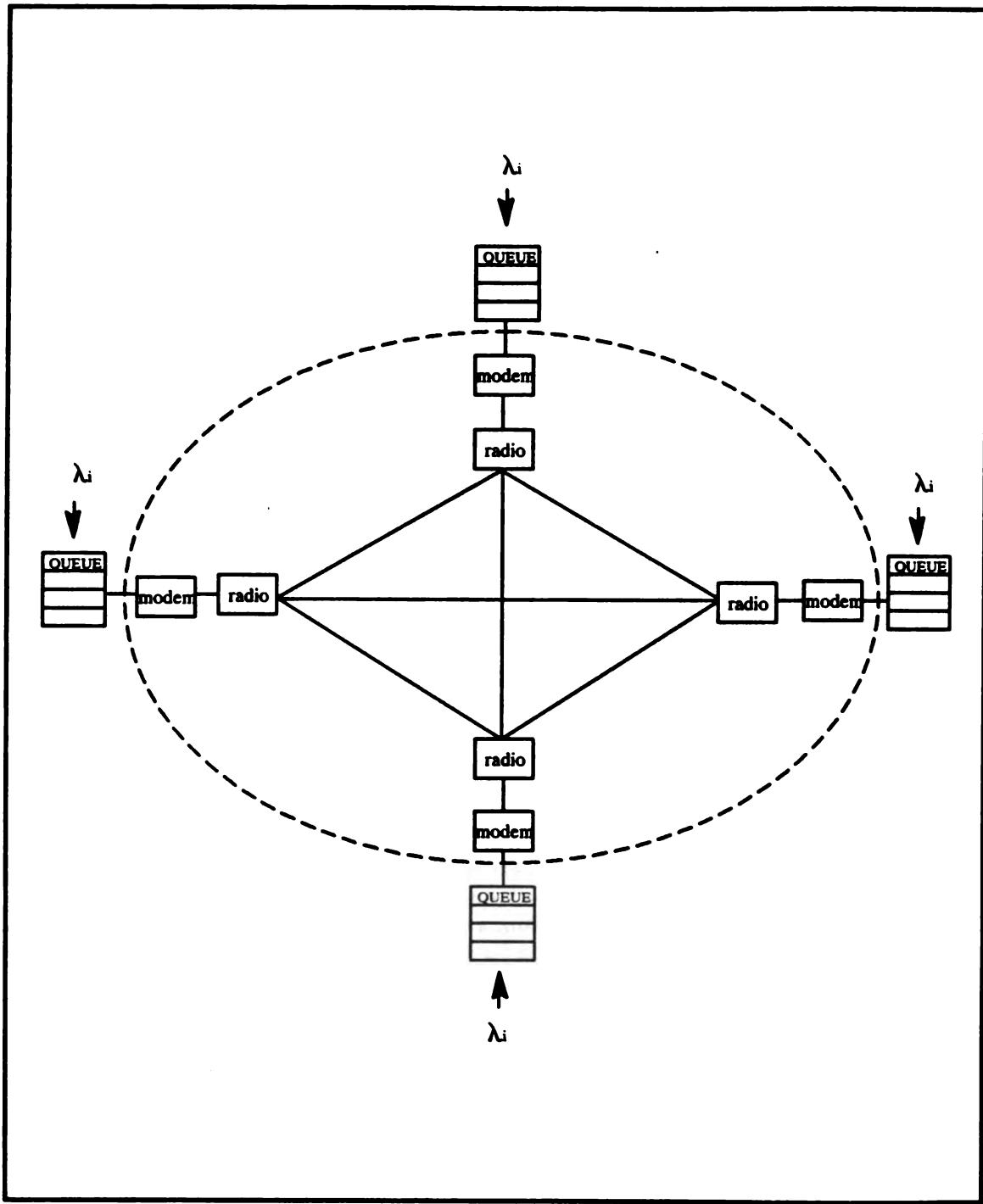
Figure 4. The Server and Its Queues.

3.3 Design Limitations

At the time the experiment was designed, no frequencies had been assigned to the hopset device; therefore, the frequency hopping portion of the experiment was in doubt. Frequencies were assigned just as the second replication under single channel mode was nearing completion. A second experiment was performed to evaluate communication throughput for the frequency hopping mode utilizing the same design as for the single channel experiment. Thus, single channel and frequency hopping were tested sequentially, one at a time, for twelve days each.

4. EFFECTS OF FACTORS ON THROUGHPUT

The effects of message length and message arrival rate on network throughput were evaluated for both single channel and frequency hopping modes using the following definition:

> Network throughput is the average number of bits per second that were successfully transmitted and acknowledged over a one hour test cell. This does not include such overhead as acks or, in the event of collisions, message retransmissions. It does, however, include error detection/correction bits and sync characters.

4.1 Analysis of Variance

As the throughput data was checked for completeness, it was noticed that some of the variances under various experimental conditions exhibited minor discrepancies. The assumption of homogeneity of variance was checked using the Burr–Foster Q–test and results confirmed this observation for both the single channel and frequency hopping data sets. (Note: The customary Bartlett test for homogeneity of variance could not be used, as a large portion of the experimental cells exhibited zero variance, a condition which negates the use of this particular test.) The mean throughput for each experimental condition is given in Table 1.

Unequal variances are often accompanied by skewness in the distribution within each experimental cell. When this occurs, the initial form of the observations is unsuitable for standard ANOVA procedures, and it is often desirable to transform the data. To assist in selecting an appropriate transformation, a plot of the cell standard deviations (y–axis) versus the cell means (x–axis) was made for each data set. The plots revealed a positive correlation between the cell standard deviations and the cell group means in each case. When a relationship such as this exists, a suggested transformation to obtain approximately equal variances can be obtained from the regression of the logarithm of the cell deviations on the logarithm of the cell means. Using this procedure, the suggested transformation was the logarithm; however, the natural logarithm transformation did a better job in reducing the correlation between the cell standard deviations and the cell means for both data sets. This was the transformation selected. The transformed data became more normal and the assumption of homogeneity of variance was verified.

An analysis of variance procedure was performed on each data set. The final ANOVA tables for the single channel and frequency hopping data are presented as Tables 2 and 3, respectively. The most significant term in the analysis of each data set was message length, although message arrival rate also influenced throughput. In each case the interaction of the main factors was negligible compared with these effects.

Rep	Mode	Message Length	Message Arrival Rate			
			400	1000	1400	2000
I	SC	48	64	126	137	148
		144	192	324	344	369
		256	341	492	495	541
		352	463	591	609	629
	FH	48	64	119	127	137
		144	192	298	314	325
		256	341	439	448	495
		352	460	507	542	566
II	SC	48	64	124	137	148
		144	192	321	350	364
		256	341	479	506	527
		352	461	587	603	647
	FH	48	64	118	127	134
		144	192	292	318	323
		256	279	450	442	479
		352	455	515	527	546
III	SC	48	64	125	136	148
		144	192	321	348	359
		256	341	492	511	513
		352	469	583	603	608
	FH	48	64	116	125	131
		144	192	292	297	310
		256	339	422	426	451
		352	463	481	508	563

Table 1. Mean Throughput (bits/sec) by Experimental Condition

Looking at the mean throughput for each level of message length in Table 1, one sees a significant increase in throughput as message length increased from 48 to 352 characters. Similarly, an increase in arrival rate increased the mean throughput. Comparing the change in mean throughput for the different levels of message arrival rate for each level of message length, the only meaningful change appears to occur between the 400 and 1000 message arrival rates. This is probably because with an arrival rate of 1000 and above, the queues were generally building, so network throughput reached its maximum. With an arrival rate of 400, network throughput did not reach a sustained maximum, as the queues were generally empty.

When a factor is significant, techniques are available by which to partition its overall sums of squares with $(p-1)$ degrees of freedom into $(p-1)$ separate sums of squares, each with one degree of freedom. The different components can then be interpreted as

sums of squares due to linear, quadratic, etc., components of regression of response on the factor of interest.

The individual degrees of freedom were separated as linear, quadratic, and cubic effects for both message length and message arrival rate ("cubic" included higher powers which could not be separately estimated from this data). This further ANOVA showed the effects of both message length and message arrival rate were linear, quadratic, and cubic, with the linear effect being strongest. This trend was evident for both single channel and frequency hopping.

There were no statistically significant differences among the three replications of either experiment. Typically the "blocks within replication" source of variation is not calculated, but it was evident from the ANOVA that it did not influence throughput.

4.2 The Smirnov Nonparametric Test

The Smirnov nonparametric test for two independent samples was utilized to test for differences between the single channel and the frequency hopping throughput data. The Smirnov test is useful in situations where two samples have been drawn, as in the AFATDS experiment, one from each of two possible different continuous distributions, and it is desired to determine whether the two empirical distribution functions associated with the two populations are identical.

The hypothesis for the comparison was based on a two–sided test. The null hypothesis was that the two populations, single channel and frequency hopping, have identical distribution functions. The respective distribution functions were denoted by $F(x)$ and $G(x)$, and the null hypothesis written as

$$H_0 : F(x) = G(x) \text{ for all x from } -\infty \text{ to } +\infty.$$

The alternative hypothesis was

$$H_1 : F(x) \neq G(x) \text{ for at least one value of x.}$$

For the two–sided test, the test statistic T_1 was defined as the greatest vertical distance between the two empirical distribution functions, and H_0 was to be rejected at some level of significance α if T_1 exceeded its $1 - \alpha$ quantile, as given in a table of quantiles of the Smirnov test statistic for two samples of equal size. The test statistic for the two–sided test was calculated as 4/16. From the appropriate table, the .95 quantile of T_1 was 7/16. Therefore, for the throughput data, H_0 was accepted at the 0.05 level of significance. From the table values, the critical level $\hat{\alpha}$ was estimated as larger than 0.20. *The effect of mode of radio operation on network throughput was not statistically significant.*

4.3 Data Summary

In this section the following definition for normalized throughput is used:

Normalized throughput is the number of bits transmitted and acknowledged divided by the number of bits that could be transmitted if there were no overhead. As with throughput, this does not include such overhead as acks or, in

210

SOURCE	DEGREES OF FREEDOM		SUM OF SQUARES	MEAN SQUARE	F RATIO
Replications	2		0.00	0.00	n.s.d.
Blocks within Replications	9		0.10	0.01	
Message Length (L)	3		18.45	6.15	s.s.d.
Linear		1	16.68	16.68	s.s.d.
Quadratic		1	1.71	1.71	s.s.d.
Cubic		1	0.06	0.06	s.s.d.
Arrival Rate (R)	3		2.32	0.77	s.s.d.
Linear		1	1.77	1.77	s.s.d.
Quadratic		1	0.46	0.46	s.s.d.
Cubic		1	0.09	0.09	s.s.d.
Message Length x Arrival Rate (LR)	9		0.22	0.02	n.s.d.
Error	21		0.19	0.01	
Total	47		21.28		

Table 2.   Analysis of Variance (Effect on Throughput – Single Channel)

n.s.d. – no significant difference
s.s.d. – statistically significant difference

SOURCE	DEGREES OF FREEDOM		SUM OF SQUARES	MEAN SQUARE	F RATIO
Replications	2		0.01	0.00	n.s.d.
Blocks within Replications	9		0.11	0.01	
Message Length (L)	3		17.41	5.80	s.s.d.
Linear		1	15.72	15.72	s.s.d.
Quadratic		1	1.62	1.62	s.s.d.
Cubic		1	0.07	0.07	s.s.d.
Arrival Rate (R)	3		1.54	0.51	s.s.d.
Linear		1	1.20	1.20	s.s.d.
Quadratic		1	0.27	0.27	s.s.d.
Cubic		1	0.07	0.07	s.s.d.
Message Length x Arrival Rate (LR)	9		0.21	0.02	n.s.d.
Error	21		0.18	0.01	
Total	47		19.46		

Table 3.   Analysis of Variance (Effect on Throughput – Frequency Hopping)

211

the event of collisions, message retransmissions. In this experiment the transmission rate was 1200 bps, so if throughput is provided in bits per second, the divisor is 1200 bps.

Figure 5 shows normalized throughput vs message length for SC transmissions. In the figure, normalized throughput has been multiplied by 100, so a value of 50 implies the throughput achieved is 50% of the network's throughput capacity, or about 600 bps. (To identify throughput from the normalized throughput graphs, simply multiply the normalized throughput value by 12.) *Normalizing throughput allows comparison of throughput results from experiments using different transmission rates.*

Normalized throughput increased as message length increased. With longer messages more time was spent on the actual transmission of data and less on the overhead associated with accessing the net. Further, as the arrival rate increased, throughput increased. At the 400 msg/hr rate the network was often idle, contributing to lower throughput. At higher arrival rates, there was an increase in throughput; however, the increase was small once the rates went beyond 1000 msg/hr. At these higher rates the queues were rarely empty. The maximum normalized throughput for this experiment was 52%, a throughput of about 625 bps. *Loading the system with more messages only increased the queue sizes, not the throughput.*

Figure 6 depicts normalized throughput vs message length for FH transmissions. The trends are similar to those for SC transmissions. The maximum normalized throughput was 47%, a throughput of about 560 bps. This was less than that of SC transmissions.

The lenient definition for throughput in this analysis does not consider the added overhead of the extra five Hamming bits (for error correction/detection) for each seven information bits. In reality the normalized throughput of *information* bits for SC is 52% * 7/12 = 30%, a throughput of about 360 bps. The normalized throughput for FH is 47% * 7/12 = 27%, a throughput of about 325 bps.

Figure 7 provides a slightly different perspective on the effects of message length and arrival rate on throughput for SC transmissions. Increased message length clearly improved throughput. The three message length increases (all around 100 characters) each improved throughput by a factor close to ten. Beyond the 1400 msg/hr arrival rate, there is very little increase in throughput. At this point the network appears saturated. Figure 8 depicts throughput vs arrival rate for FH.

5. CONCLUSIONS

For throughput, the ANOVA has shown that message length is a more significant factor than arrival rate; their interaction is negligible. Future experiments will consider longer message lengths. The mode of radio operation was not a statistically significant factor.

The advantages of experimentation are evident. The results provide valuable information on CNRN thresholds with the TACFIRE protocol, albeit in a "perfect" environment. Experimentation with actual hardware and protocols has shown that even under laboratory conditions, there is inherent communications degradation. The results have shown even best case conditions to be worse than assumed in previous modeling studies. SC throughput never exceeded 648 bps. In frequency hopping mode, it never exceeded 566
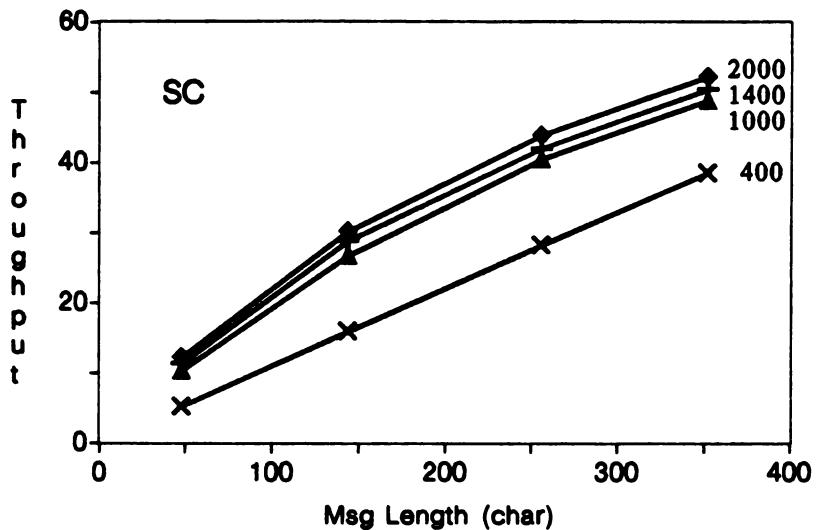
Figure 5.   Normalized Throughput vs Messge Length
Grouped by Arrival Rate (Single Channel)
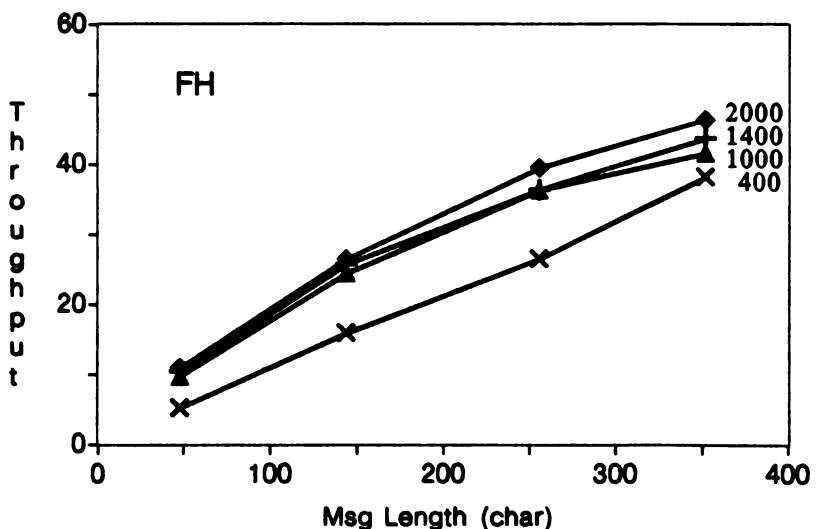
Figure 6.   Normalized Throughput vs Message Length
Grouped by Arrival Rate (Frequency Hopping)

213
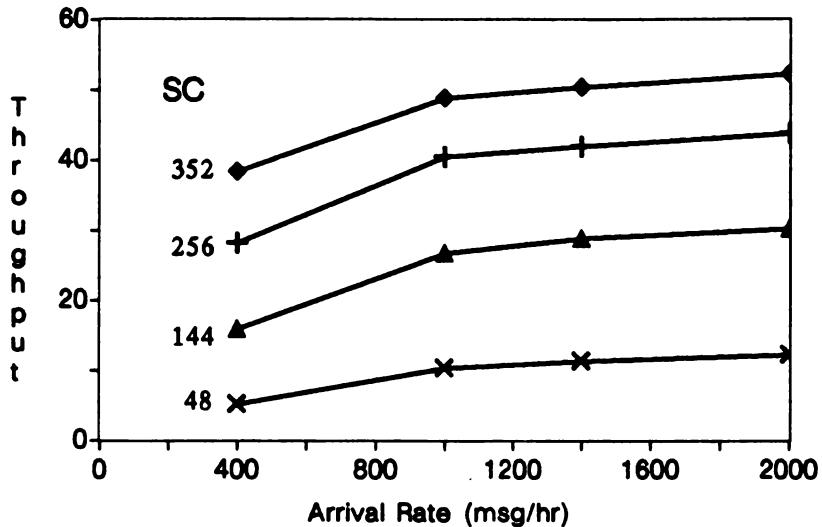
Figure 7. Normalized Throughput vs Arrival Rate
Grouped by Message Length (Single Channel)
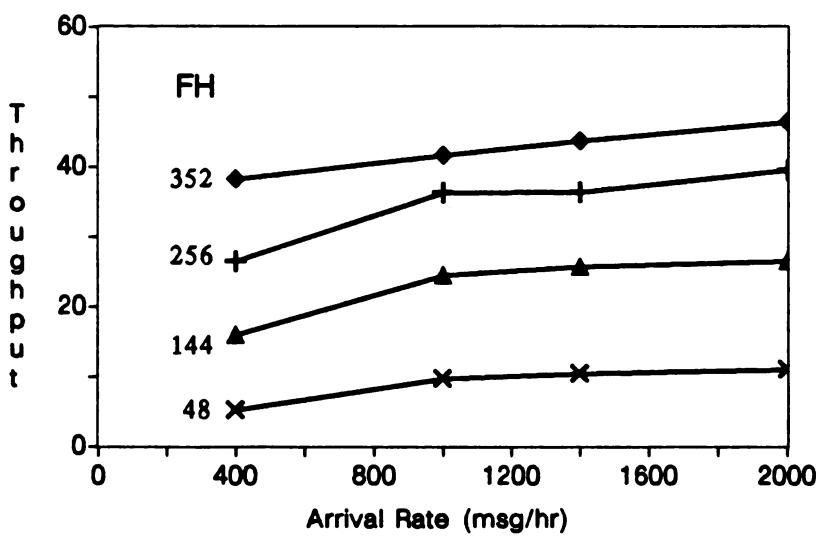
Figure 8. Normalized Throughput vs Arrival Rate
Grouped by Message Length (Frequency Hopping)

214

bps. If the Hamming code is considered overhead, the throughput dropped to 360 bps for SC and 325 bps for FH. The limits of the CNRN using TACFIRE protocol identified in this experiment should be considered when modeling or designing communications architectures and protocols.

6.   REFERENCES

Magnavox Government and Industrial Electronics Company, "AFATDS Performance Modeling System Level for Multiservice Advanced Field Artillery Tactical Data System," Contract No. DAAB07–89–G–E424–0001, CDRL Sequence No. NA, Fort Wayne, IN, 1991.

7.   BIBLIOGRAPHY

Samuel C. Chamberlain, "The Information Distribution System: An Overview", Ballistic Research Laboratory Technical Report No. 3114, July 1990.

W.G. Cochran and G. M. Cox, Experimental Designs, Wiley & Sons: New York, 1971.

W. J. Conover, Practical Nonparametric Statistics, Wiley & Sons: New York, 1971.

L. Cooper, U. N. Bhat, and L. J. LeBlanc, Introduction to Operations Research Models, W. B. Saunders Company: Philadelphia, 1977.

N. L. Johnson and F. C. Leone, Statistics and Experimental Design in Engineering and the Physical Sciences, Wiley & Sons: New York, 1977.

Virginia A.T. Kaste, " The Information Distribution System: The Fact Exchange Protocol, A Tactical Communications Protocol", Ballistic Research Laboratory Memorandum Report No. 3856, August 1990.

I. Olkin, L. J. Gleser, and C. Derman, Probability Models and Applications, Macmillan Publishing Co. Inc: New York, 1980.

M. Schwartz, Telecommunications Networks: Protocol, Modeling and Analysis, Addison–Wesley Publishing Co.: Massaachusetts, 1987.

W. Stallings, Data and Computer Communications, Macmillian Publishing Co., Inc: New York, 1985.

A. S. Tanenbaum,Computer Networks, Prentice–Hall Inc.: New Jersry, 1981.

Nonlinear time series analysis
via cumulative regressograms

Ian W. McKeague[1]
Department of Statistics
Florida State University
Tallahassee, Florida, 32306

Mei-Jie Zhang[2]
Department of Mathematics
New Mexico Institute of Technology
Socorro, New Mexico, 87801

Abstract: We consider the problem of identifying the class of time series model to which a series belongs based on observation of part of the series. Techniques of nonparametric estimation have been applied to this problem by various authors using kernel estimates of the one-step lagged conditional mean and variance functions. We study cumulative versions of Tukey regressogram estimators of such functions. These are more stable than estimates of the mean and variance functions themselves and can be used to construct confidence bands. Goodness-of-fit tests for specific parametric models are also briefly discussed.

1 Introduction

Currently one of the most challenging problems in nonlinear time series analysis is to identify the class of time series model to which a series $\{X_t\}$ belongs based on observation of part of the series, $\{X_t, \ t = 0, 1, \ldots, n\}$. Techniques of nonparametric estimation have been applied to this problem by Robinson (1983), who studied the large sample properties of kernel estimators of lagged conditional means $E(X_t|X_{t-j})$ and $E(X_t|X_{t-j}, X_{t-k})$ for various j and k values. Such estimators are useful for detecting nonlinearities graphically, see Tong (1990, p. 12). This approach has been further developed by Auestad and Tjøstheim (1990) who focused on kernel estimates of the one-step lagged conditional mean and variance functions $\lambda(x) = E(X_t|X_{t-1} = x)$ and $\gamma(x) = \text{var}(X_t|X_{t-1} = x)$ for the purpose of identifying common nonlinear models such as threshold (Tong, 1983) and exponential autoregressive (Ozaki, 1980).

In the present paper we discuss an approach to this problem based on estimation of *cumulative* versions of the conditional mean and variance functions, $\Lambda(\cdot) = \int_a^\cdot \lambda(x) \, dx$ and $\Gamma(\cdot) = \int_a^\cdot \gamma(x) \, dx$, where a is an appropriately chosen point in the state space. These estimators, denoted $\hat{\Lambda}$ and $\hat{\Gamma}$, are obtained by integrating Tukey regressograms for λ and γ. The reason for considering cumulative versions of the conditional mean and variance is that it is possible to derive functional limit theorems, whereas available asymptotic results for kernel or regressogram estimators of λ and γ are only useful pointwise. We advocate $\hat{\Lambda}$ and $\hat{\Gamma}$ as natural 'signatures' of a time-series in preference to estimates of λ and γ.

We present a functional limit theorem for $\hat{\Lambda}$ which holds under conditions that can be readily checked when $\{X_t\}$ is a Markov chain. This result can be used to

---

construct confidence bands, which are more helpful than confidence intervals in assessing plots. This is the chief benefit from estimating *cumulative* conditional means and variances rather than λ and γ themselves. Another benefit is that $\hat{\Lambda}$ and $\hat{\Gamma}$ are relatively insensitive to variations in bandwidth compared to the kernel or regressogram estimators.

We also briefly describe an application of our approach to omnibus goodness-of-fit testing for parametric models of the form $\lambda(x) = g(\theta, x)$, where g is a known function and θ is an unknown parameter, e.g. a linear model. Robinson (1983) has given a test for linearity, but his test has the diadvantage that it is only applicable at a small number of discrete locations. Other formal tests for linearity found in the literature are parametric—constructed by arranging the linear model to be nested within various larger parametric models, see Tong (1990, Section 5.2). We propose an omnibus test based on a comparison of $\hat{\Lambda}$ and a smoothed version of $\int_a^{\cdot} g(\tilde{\theta}, x)\, dx$, denoted $\tilde{\Lambda}$. Here $\tilde{\theta}$ is the conditional least square estimator of θ, see Klimko and Nelson (1978).

There are some connections between the present paper and cumulative hazard function estimation in survival analysis; see the survey articles of Andersen and Borgan (1985) and McKeague and Utikal (1990a). In fact $\hat{\Lambda}$ is closely related to an estimator introduced by McKeague and Utikal (1990b). Martingale techniques play an important role here, as they do survival analysis.

2 Estimation of Λ and Γ

Assume that the conditional mean and variance of X_t given $X_0, X_1, \ldots, X_{t-1}$ only depend on X_{t-1}. This property holds, for example, if $\{X_t\}$ is a Markov chain. In particular, an important example is the nonlinear autoregressive process

$$X_t = \lambda(X_{t-1}) + \sigma(X_{t-1})\epsilon_t, \tag{1.1}$$

where $\{\epsilon_t\}$ are iid with zero-mean and unit variance and $\gamma = \sigma^2$. In this case the time series is characterized by the triplet $(\lambda, \gamma, \text{distribution of } \epsilon_0)$. We are primarily interested in λ and γ. It is assumed throughout that $\{X_t\}$ is stationary with a marginal density denoted f.

We restrict attention to estimation of Λ and Γ on a fixed interval $[a, b]$. The regressogram estimators $\hat{\lambda}$ and $\hat{\gamma}$ are defined as follows. Let $\mathcal{I}_1, \ldots, \mathcal{I}_{d_n}$ be a partition of $[a, b]$ made up of intervals of equal length w_n, the bins of the regressogram, and denote $\mathcal{I}_x = \mathcal{I}_j$ for $x \in \mathcal{I}_j$. Set

$$\hat{\lambda}(x) = (nw_n \hat{f}(x))^{-1} \sum_{t=1}^{n} I\{X_{t-1} \in \mathcal{I}_x\} X_t,$$

$$\hat{\gamma}(x) = (nw_n \hat{f}(x))^{-1} \sum_{t=1}^{n} I\{X_{t-1} \in \mathcal{I}_x\} (X_t - \hat{\lambda}(x))^2,$$

where $\hat{f}$ is the histogram estimator of f given by

$$\hat{f}(x) = (nw_n)^{-1} \sum_{t=1}^{n} I\{X_{t-1} \in \mathcal{I}_x\},$$

and $I(\cdot)$ is the indicator function. Regressogram estimators were introduced by Tukey (1961) and have been studied recently by Diebolt (1990).

Introduce the estimators

$$\hat{\Lambda}(\cdot) = \int_a^{\cdot} \hat{\lambda}(x)\,dx \quad \text{and} \quad \hat{\Gamma}(\cdot) = \int_a^{\cdot} \hat{\gamma}(x)\,dx.$$

Although it is possible to use the more sophisticated kernel estimators to yield better estimates of λ and γ, there is little to be gained from using them in $\hat{\Lambda}$ and $\hat{\Gamma}$, which are less sensitive to variations in $\hat{\lambda}$ and $\hat{\gamma}$. We prefer the regressogram estimators due to their computational simplicity. In practice, care needs to be taken in choosing the interval $[a, b]$ and the bins to ensure that the regressogram estimates are not too unstable. For good results, the binwidths should be of comparable size (we have taken them to be of equal size merely to simplicity the notation), and there should be at least 5 observations per bin.

Ideally, in order to carry out inference on λ, using a confidence band for Λ say, we would like to find the limiting distribution of $\sqrt{n}\,(\hat{\Lambda} - \Lambda)$. However, for technical reasons we are only able to obtain a satisfactory weak convergence theory when Λ is replaced by the smoothed version of Λ given by $\Lambda^*(z) = \int_a^z \lambda^*(x)\,dx$, where

$$\lambda^*(x) = \int_{\mathcal{I}_x} f^*(u)\lambda(u)\,du \bigg/ \int_{\mathcal{I}_x} f^*(u)\,du$$

and f^* is the histogram estimator of f determined by a finer partition of $[a, b]$ consisting of intervals of equal length w_n^*. We regard Λ^* as a 'surrogate' for Λ, which is reasonable since Λ^* converges uniformly in probability to Λ. However $\sqrt{n}(\Lambda^* - \Lambda)$ may not be asymptotically negligible. If it is (for example if λ is piecewise constant over $\mathcal{I}_1, \ldots, \mathcal{I}_{d_n}$ for some n) then Λ^* is not needed and we can deal with Λ directly.

The asymptotic distribution of $\hat{\Lambda}$ is given by the following result, for which we assume that λ is Lipschitz, $EX_0^8 < \infty$, (X_0, X_t) has a bounded joint density for all $t \geq 1$, and the marginal density f is continuous and does not vanish on $[a, b]$.

THEOREM. *Suppose that* $\sup_{x \in [a,b]} \mathrm{var}[\hat{f}(x)] = o(w_n)$, $nw_n \to \infty$, $nw_n^4 \to 0$ *and* $w_n^* \sim w_n^2$ *as* $n \to \infty$. *Then* $\sqrt{n}\,(\hat{\Lambda} - \Lambda^*)$ *converges in distribution a continuous Gaussian martingale with mean zero and variance function* $H(z) = \int_a^z \gamma(x)/f(x)\,dx$.

A proof of this result can be found in McKeague and Zhang (1991). A large class of stationary Markov processes $\{X_t\}$ that satisfy the first condition of the

theorem is described in Auestad and Tjøstheim (1990), who show that strong mixing with a geometric mixing rate implies $\text{var}[\hat{f}(x)] \sim O((nw_n)^{-1})$ uniformly over $[a, b]$ provided that f is bounded there. Thus the condition of the theorem holds in this case if $nw_n^2 \to \infty$. In a particular example it will be easier to check geometric ergodicity (Nummelin, 1984), which implies strong mixing with a geometric mixing rate. Geometric ergodicity is in turn implied by a readily checkable condition of Tweedie (1983).

We now turn some possible applications of this result.

Confidence bands. It can be shown that $\hat{H}(\cdot) = \int_a^{\cdot} \hat{\gamma}/\hat{f} \, dx$ is a uniformly consistent estimator of H, so that an asymptotic $100(1 - \alpha)\%$ confidence band for Λ^* is given by

$$\hat{\Lambda}(x) \pm c_\alpha n^{-1/2} \hat{H}(b)^{1/2} \left(1 + \frac{\hat{H}(x)}{\hat{H}(b)}\right) \qquad x \in [a, b],$$

where c_α is the upper α quantile of the distribution of $\sup_{t \in [0, 1/2]} |B^0(t)|$ and B^0 is the Brownian bridge process, see Andersen and Borgen (1985, p. 114). Tables for c_α can be found in Hall and Wellner (1980).

Testing for a difference between two regression functions. Consider the two-sample problem of testing whether two independent time series have identical regression functions λ. Denote the various functions, sample sizes estimators etc. associated with the two series by using a subscript 1 or 2, as in λ_j, $j = 1, 2$. Let $n = n_1 + n_2$. Then, if $n_j/n \to p_j$ for $j = 1, 2$, and the conditions of the theorem are satisfied for the two series, $\sqrt{n}(\hat{\Lambda}_1 - \hat{\Lambda}_2)$ converges in distribution to a continuous Gaussian martingale with mean zero and variance function

$$p_1^{-1} \int_a^{\cdot} \frac{\gamma_1(x)}{f_1(x)} \, dx + p_2^{-1} \int_a^{\cdot} \frac{\gamma_2(x)}{f_2(x)} \, dx,$$

provided that $\lambda_1 = \lambda_2$ on $[a, b]$ and $\sqrt{n}(\Lambda_1^* - \Lambda_2^*)$ converges uniformly in probability to zero. The latter condition holds if the common λ is piecewise constant, as mentioned earlier. Confidence bands for $\Lambda_1^* - \Lambda_2^*$ are constructed as above. Some plots of such bands are given in Section 3.

Goodness-of-fit testing. Consider the problem of testing whether λ belongs to a parametric family $\{g(\theta, \cdot) : \theta \in \Theta\}$ of regression functions. Here g is a known deterministic function, and Θ is a closed, bounded subset of $\mathbf{R}^p$. Let $\tilde{\Lambda}(z) = \int_a^z \tilde{\lambda}(x) \, dx$, where

$$\tilde{\lambda}(x) = \int_{\mathcal{I}_x} f^*(u) g(\tilde{\theta}, u) \, du \Big/ \int_{\mathcal{I}_x} f^*(u) \, du$$

and $\tilde{\theta}$ is the conditional least squares estimator minimizing $\sum_{t=1}^{n}(X_t - g(\theta, X_{t-1}))^2$. McKeague and Zhang (1991) have shown that, under the parametric model, the process $\sqrt{n}(\hat{\Lambda} - \tilde{\Lambda})$ converges weakly to a Gaussian process having a covariance that can be estimated consistently. This result can be used to develop graphical methods for detecting departures from the parametric model based on plots of $\hat{\Lambda} - \tilde{\Lambda}$, or to give formal chi-squared goodness-of-fit tests.

3 Simulation study and example

We have carried out simulations using three model examples taken from Auestad and Tjøstheim (1990):

Model 1: linear autoregressive, $X_t = 0.8X_{t-1} + \epsilon_t$;

Model 2: threshold autoregressive,

$$X_t = \begin{cases} -0.3X_{t-1} + \epsilon_t, & \text{if } X_{t-1} \leq 0, \\ 0.8X_{t-1} + \epsilon_t, & \text{if } X_{t-1} > 0; \end{cases}$$

Model 3: exponential autoregressive, $X_t = \{0.8 - 1.1\exp(-50X_{t-1}^2)\}X_{t-1} + \epsilon_t$.

Here ϵ_t is Gaussian white noise with mean zero and standard deviation 0.1. Auestad and Tjøstheim (1990) checked geometric ergodicity and stationarity for these examples.

We restricted estimation of Λ to the interval $[-0.3, 0.3]$. The binwidth was taken as $w_n = 0.05$ (same as Auestad and Tjøstheim, who plotted point estimates of λ for these three models). Inspecting the plots of $\hat{\Lambda}$ in Figure 1, we find that the three models are easily distinguishable, even for sample size as low as 250. The parabolic shape of the linear autoregressive model, and the 'squashed' parabola of the exponential autoregressive are especially distinct.

Figure 2 shows plots of differences between the estimates of the cumulative regression functions in the two sample problem, for various pairs of the above models. In the first plot in each row, the two series are generated using the linear model and the zero function is contained within the band, so our test would correctly conclude that the regression functions are identical. In the other plots, the zero function is well outside the bands and the test correctly concludes that the regression functions are different.

We conclude with an example involving real data. Consider the set of IBM daily closing stock prices from late 1959 to mid 1960 (period I) and mid 1961 to early 1962 (period II) given in Tong (1990). The daily relative change in price appears to be stationary and is used in place of the raw data. Tong (1990) tested for linearity and decided that period I is linear and period II is nonlinear. Figure 3 gives a plot of the difference between the estimates of the cumulative regression functions in the two periods, along with the 95% confidence band, using $d_n = 10$.

221
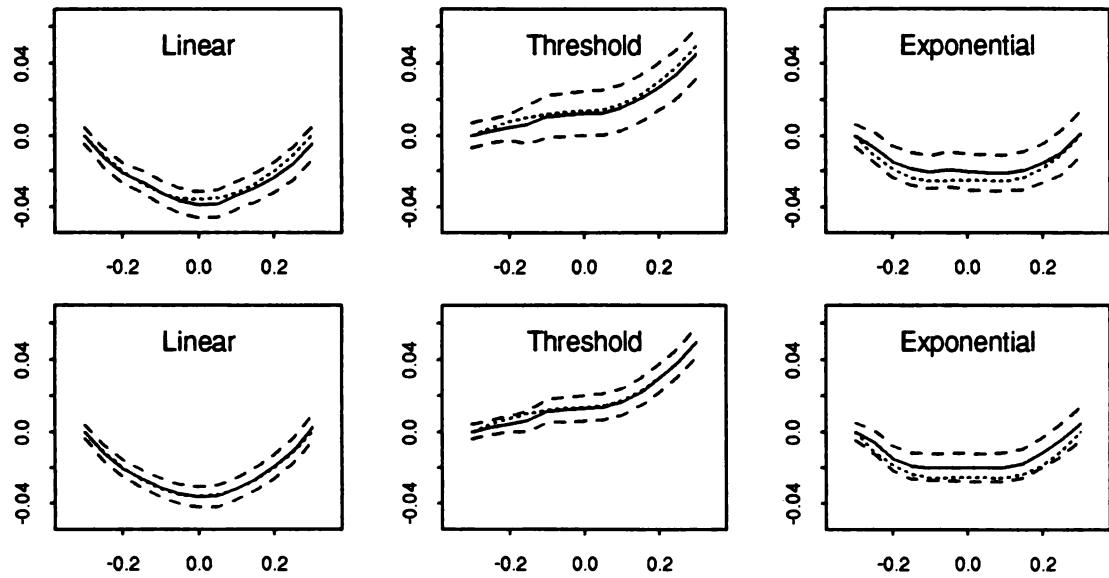
Figure 1. $\hat{\Lambda}$ with 95% confidence bands; solid lines, $\hat{\Lambda}$; dotted lines, Λ; dashed lines, confidence bands; first row, $n = 250$; second row, $n = 500$.
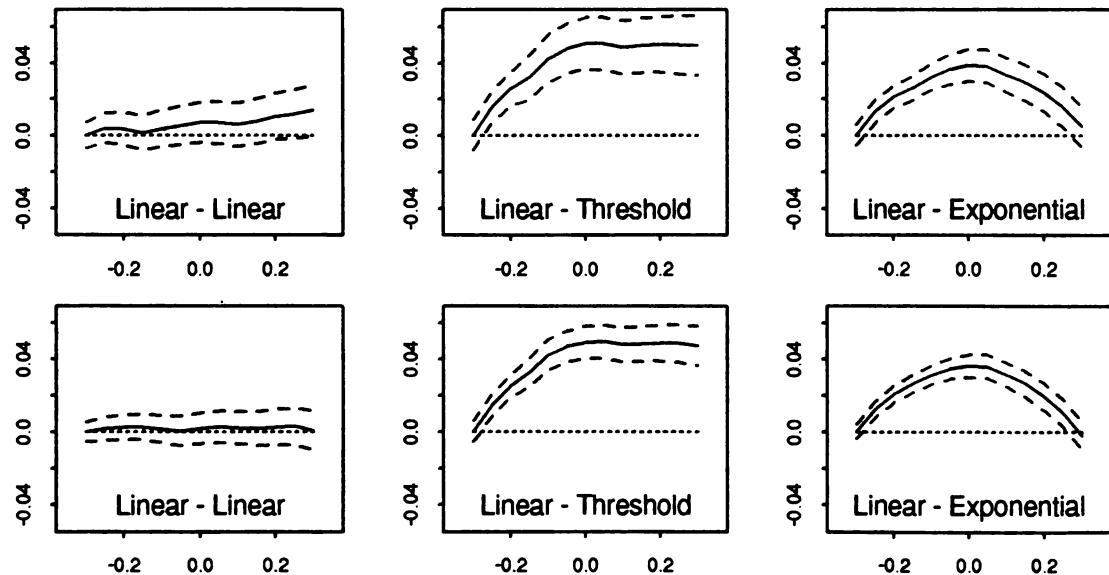
Figure 2. $\hat{\Lambda}_1 - \hat{\Lambda}_2$ with 95% confidence bands; first row, $n = 250$; second row, $n = 500$.

The confidence band does not contain the zero function, so we conclude that the regression functions for the two periods differ significantly from one another. Our chi-squared test with $d_n = 8, 10$ and 12, and degrees of freedom $L = 2$ and 4, gave the same result.
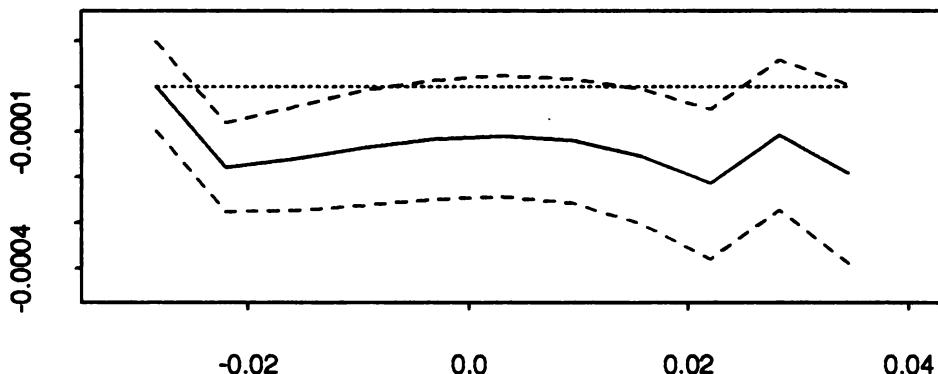
Figure 3. $\hat{\Lambda}_1 - \hat{\Lambda}_2$ with 95% confidence band for IBM stock price data; $d_n = 10$; $\hat{\Lambda}_1 = $ period I, $\hat{\Lambda}_2 = $ period II.

References

Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: a review. *Scand. J. Statist.* **12** 97–158.

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77** 669–687.

Diebolt, J. (1990). Testing the functions defining a nonlinear autoregressive time series. *Stoch. Process. Appl.* **36** 85–106.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67** 133–143.

Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6** 629–642.

McKeague, I. W. and Utikal, K. J. (1990a). Stochastic calculus as a tool in survival analysis: a review. *Appl. Math. Comput.* **38** 23–49.

McKeague, I. W. and Utikal, K. J. (1990b). Identifying nonlinear covariate effects in semimartingale regression models. *Probab. Theory Related Fields* **87** 1–25.

McKeague, I. W. and Zhang, M.- J. (1991). Identification of nonlinear time series from first order cumulative characteristics. Submitted to *Annals of Statistics*.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.

Ozaki, T. (1980). Non-linear time series model for non-linear random vibrations. *J. Appl. Prob.* **17** 84–93.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* **4** 185–207.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Lecture Notes in Statistics **21**. Springer, New York.

Tong, H. (1990). *Non-linear Time Series: a Dynamical System Approach.* Clarendon.

Tukey, J. (1961). Curves as parameters and touch estimation. *Proc. 4th Berkeley Symp. Math. Statist. Probab.* **1** 681–684.

Tweedie, R. L. (1983). Criteria for rates of convergence of Markov chains, with application to queueing theory. In *Papers in Probability, Statistics and Analysis,* Eds. J. F. C. Kingman and G. E. H. Reuter, pp. 260–276. Cambridge University Press.

CHANGE ANALYSIS

by Emanuel Parzen
Department of Statistics, Texas A&M University
College Station, Texas 77843-3143

Abstract

Change Analysis "in the strict sense" is concerned with the problem of detecting and estimating slow and abrupt changes in the probability distributions of successive observations $Y(t)$ of a variable or system. This paper has two goals (1) introduce an approach to Change problems by introducing analysis of Score Change Processes (whose idea is to study if a model to a whole data set fails to fit it by "random walking" the parameter estimating equations); (2) develop analogies between four basic statistics problems, corresponding to the standard assumptions made about a sequence of observations $Y(t)$, $t = 1, \ldots, n$; test the hypothesis: A: Distribution of specified parametric form, B: Independence, C: Identical distribution, For a sequence of bivariate observations $X((t), Y(t))$ one would like to test D: Independence of X and Y. Contents are: Introduction, Change analysis in the strict sense (test Assumption C), Goodness of fit (test Assumption A), Spectral Analysis (test Assumption B), Four phases of change analysis, Parametric scores change analysis, Nonparametric scores change analysis.

1. Introduction

Data $Y(1), \ldots, Y(n)$ which can be regarded as continuous random variables observed sequentially can be called indexed data or a time series. Classic statistical inference makes three basic assumptions:

Assumption A. Probability law of each Y has probability density belonging to a known parametric family of probability densities $f(y; \theta)$.

Assumption B. Random variables $Y(1), \ldots, Y(n)$ are independent.

Assumption C. Random variables $Y(1), \ldots, Y(n)$ are identically distributed.

Methods for detecting (and estimating) the fit (and the nature of violations) of these assumptions in our opinion can be respectively related to three parallel theories:

Theory C. Changepoint analysis or change analysis (in the strict sense).

Theory B. Spectral analysis (time series analysis in the frquency domain).

Theory A. Goodness of fit.

We believe that one can define a theory, called Comparison Change Analysis, which is intended to study analogies between theories A,B,C (and bring the insights of the theories that are more developed, such as spectral analysis, to less developed ones). General accounts of this theory are given in Parzen (1992), (1991).

The assumption that the data is observed sequentially, which may seem to limit the applicability of Change Analysis, is dropped when the analogies are extended to the bivariate data analysis problem which considers independent bivariate data $(X(t), Y(t))$, $t = 1, \ldots, n$, and desires to model the relation between X and Y and in particular to test

Assumption D. X and Y are independent random variables.

A general non-parametric theory of testing assumption D can be related to

---

Theory D. Change analysis (random effect).

Analogies between theories A to D are obtained from the facts that in each problem the first step in analysis is to define a dynamic statistic which is a function on the unit interval [0,1] whose asymptotic distribution (under the null hypothesis that the assumptions are true) is either a Brownian Bridge or a related process. The test statistics in each theory are analogous to the nonparametric test statistics that statisticians have developed to test goodness of fit for equality of two distributions.

Textbooks imply it is difficult to choose among the many test statistics for goodness of fit and analogous testing problems; we believe we should be optimistic about our ultimate ability to develop procedures for adaptively choosing appropriate test statistics which not only test the null hypothesis but also suggest likely models instead of only rejecting the null hypothesis.

2. Change analysis in the strict sense (test Assumption C)

The theory of change analysis in the strict sense considers data $Y(t), t = 1, \ldots, n$ which represents a transformation of observed data (the identity transformation leaves the data unchanged).

Let Y^- denote the sample mean (an estimator of the true mean μ if the data are identically distributed). Let $\sigma_Y\hat{}$ denote a suitable estimator (such as the sample standard deviation) of the true standard deviation σ_Y of the data under the assumption of identical distribution (which is assumed to be finite).

The data Y is transformed to normalized data

$$Y^\sim(t) = (Y(t) - Y^-)/\sigma_Y\hat{}.$$

We plot the normalized data as a sample *change density* $c^\sim(\tau)$, $0 < \tau < 1$, defined to be a piecewise constant function whose value is equal to $Y^\sim(j)$ on the interval $(j-1)/n < \tau < j/n$, for $j = 1, \ldots, n$. Note that $\int_0^1 c^\sim(\tau)d\tau = 0$, $\int_0^1 c^{\sim 2}(\tau)d\tau = 1$.

CUSUMS (cumulative sums) are becoming increasingly important diagnostic tools to look for patterns in indexed data. They are related to the sample *change process* on $0 < \tau < 1$

$$C^\sim(\tau) = \int_0^\tau c^\sim(t)dt.$$

The points $\tau = j/n$ for $j = 1, \ldots, n$ are called "exact" values of τ; at these points $C(\tau)$ equals a cumulative sum:

$$C^\sim(j/n) = (1/n)\sum_{k=1}^{j} Y^\sim(k) = (j/n)Y_j^-.$$

To understand why the change process is an effective means of detecting change in the data consider its behavior under two models for $Y(.)$.

If $Y(.)$ is deterministic and linear, say $Y(t) = t$, then at exact $\tau = j/n$ approximately

$$c^\sim(\tau) = Y^\sim(j) = 12^{.5}(\tau - .5),$$
$$C^\sim(\tau) = (-.5)12^{.5}\tau(1 - \tau).$$

The graph of $C^\sim(\tau)$ when $Y(.)$ is linear is a parabola that goes from 0 to 0 with mimimum value at $\tau = .5$.

226

If $Y(.)$ is random (independent identically distributed), the stochastic process $C^{\sim}(\tau)$, $0 < \tau < 1$, can be shown to be asymptotically distributed (as the sample size n tends to infinity) as a Brownian Bridge stochastic process $B(\tau)$, $0 < \tau < 1$, which is a zero mean Gaussian process with covariance kernel $E[B(s)B(t)] = min(s,t) - st$. Note that $B(0) = B(1) = 0$, and

$$\text{Variance}[B(\tau)] = \tau(1 - \tau).$$

To test for departures from Assumption C (identical distribution) one tests if the observed change process $C^{\sim}(\tau)$ is significantly different from a sample curve of a Brownian Bridge which can be expected to be a wiggly (non-smooth) curve oscillating about the horizontal axis.

A related process that plays a central role in change analysis is the Change Test Process

$$CT^{\sim}(\tau) = C^{\sim}(\tau)/(\tau(1-\tau))^{.5}.$$

The fundamental role of the change test process starts with the fact that for fixed $\tau = j/n$, $CT^{\sim}(\tau)$ can be shown to be a monotone transformation of the classic two-sample Student's t-test statistic of the null hypothesis $\mu_1 = \mu_2$ in the model $Y(1), \ldots, Y(j)$ is Normal(μ_1, σ^2) and $Y(j+1), \ldots, Y(n)$ is Normal(μ_2, σ^2). The sample means and variances of the two samples $Y(1), \ldots, Y(j)$ and $Y(j+1), \ldots, Y(n)$ are respectively denoted $\mu_1\hat{}, S_1\hat{}^2$ and $\mu_2\hat{}, S_2\hat{}^2$. The pooled sample variance is

$$S\hat{}^2 = \tau S_1\hat{}^2 + (1-\tau)S_2\hat{}^2.$$

One can verify that

$$\hat{\sigma}_Y^2 = S\hat{}^2 + (\tau(1-\tau))(\mu_1\hat{} - \mu_2\hat{})^2,$$
$$\mu_1\hat{} - Y^{\sim} = (1-\tau)(\mu_1\hat{} - \mu_2\hat{})$$

The classic two-sample Student's t-test statistic is $n^{.5} T$, defining

$$T = (\tau(1-\tau))^{.5}(\mu_1\hat{} - \mu_2\hat{})/S\hat{}.$$

Define R, a "correlation version" of T, by

$$R^2 = T^2/(1+T^2), T^2 = R^2/(1-R^2).$$

Then

$$R^2 = \tau(1-\tau)(\mu_1\hat{} - \mu_2\hat{})^2/\hat{\sigma}_Y^2$$

and one concludes that $CT^{\sim}(\tau)$ is, like R, a correlation type statistic since

$$R^2 = (\tau/(1-\tau))(\mu_1\hat{} - Y^{\sim})^2/\hat{\sigma}_Y^2$$
$$= |CT^{\sim}(\tau)|^2.$$

We can consequently express Student's t-test statistic T as a monotone function of $CT^{\sim}(\tau)$ since $T = R/(1-R^2)^{.5}$.

Let $\tau\hat{}$ denote the value among the exact values $\tau = j/n$ (for $j = 1, \ldots, n-1$) at which the absolute value of $CT^{\sim}(\tau)$ achieves its maximum. Under the assumption of at most one change in the distribution of $Y(.)$, $CT^{\sim}(\tau\hat{})$ is a test statistic for change

227

and its time of occurence is consistently estimated by $\tau\hat{}$ (a result established by Carlstein (1988)).

3. Goodness of fit (Test Assumption A)

One of the most extensive and least applied branches of statistical theory is the theory of goodness of fit of probability models to observed data. Despite its importance (both for theory and practice) it appears to be sparsely taught to graduate students in statistics. The chi-squared goodness of fit test introduced by Karl Pearson in 1900 is regarded as one of the top 20 achievements in modern science. How can one explain the neglect of instruction in its theory? One explanation may be that its theory is often taught rigorously as a study in pure probability theory rather than developed vigorously for its statistical interpretation.

Let $Y(t), t = 1, \ldots, n$, be a random sample of a continuous random variable with true distribution $F(y) = F(y; \theta_0)$ belonging to a finite parametric family $F(y; \theta)$. The true quantile function is $F^{-1}(u; \theta_0)$, $0 < u < 1$. The sample distribution function is denoted

$$F\hat{}(y) = \text{fraction of sample} \leq y.$$

Let $\theta\hat{}$ denote the maximum likelihood estimator of θ. Stochastic processes whose asymptotic properties are of interest (for both theory and practice) are

$$F\hat{}(y) - F(y; \theta_0),$$
$$F\hat{}(y) - F(y; \theta\hat{}),$$
$$F(y; \theta\hat{}) - F(y; \theta_0),$$

evaluated at $y = F^{-1}(u; \theta_0)$, $0 < u < 1$. We denote such a process $C\hat{}(u), 0 < u < 1$, to emphasize its analogy to a sample change process. We use functions of u to study changes of distribution, and functions of τ to study changes of models fitting data.

The testing and estimation procedures of goodness of fit theory can be organized into four phases summarized (in section 5) in our discussion of the four phases of change analysis.

4. Spectral Analysis (Test Assumption B)

One approach to testing the assumption of independence is to consider as an alternative hypothesis for the data $Y(t), t = 1, \ldots, n$, that it is a zero mean stationary time series with covariance function, defined for $v = 0, \pm 1, \pm 2, \ldots$,

$$R(v) = E[Y(t)Y(t - v)]$$

and spectral density function, defined for $0 < \omega < 1$,

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v)e^{-i2\pi\omega v}$$

The sample spectral density is defined

$$f\hat{}(\omega) = (2\pi n)^{-1}|\sum_{t=1}^{n} Y(t)e^{i2\pi\omega t}|^2$$

228

with sample distribution function (on $0 < \omega < 1$)

$$F^{\sim}(\omega) = \int_0^\omega f(\lambda)d\lambda$$

Normalized versions of these functions are

$$f^{*\sim}(\omega) = f^{\sim}(\omega)/F^{\sim}(1),$$
$$F^{*\sim}(\omega) = F^{\sim}(\omega)/F^{\sim}(1).$$

Analogues of the sample change density and sample change process are

$$c^{\sim}(\omega) = f^{*\sim}(w) - 1,$$
$$C^{\sim}(\omega) = F^{*\sim}(w) - w.$$

5. Four phases of change analysis

A sample change process $C^{\sim}(\tau), 0 < \tau < 1$, is a dynamic statistic (sample path of a stochastic process) which often can be shown to satisfy under the null hypothesis of "no change" the null hypothesis $H_0 : C^{\sim}(.)$ is a Brownian Bridge (or a related hypothesis). The statistical analysis of $C^{\sim}(.)$ has four phases:

Phase 1: *Graphical analysis*; is the plot of $C^{\sim}(\tau)$, $0 < \tau < 1$, oscillatory, a deterministic parabola, other patterns.

Phase 2: *Non-linear functionals.* One tests H_0 by computing the values of test statistics (whose asymptotic distributions under H_0 can be deduced from the theory of empirical processes)

$$\int_0^1 |C^{\sim}(\tau)|^2 d\tau,$$

$$\int_0^1 (|C^{\sim}(\tau)|^2/\tau(1-\tau))d\tau,$$

$$\max_{0<\tau<1} |C^{\sim}(\tau)|,$$

$$\max_{\tau=j/n} |C^{\sim}(\tau)|/\tau(1-\tau).$$

Phase 3: *Linear functionals.* For various score functions $K(\tau)$, called *change score* functions, one computes the linear functional (or component)

$$C^{\sim}(K) = \int_0^1 K(\tau)dC^{\sim}(\tau) = \int_0^1 K(\tau)c^{\sim}(\tau)d\tau$$

One can often write approximately

$$C^{\sim}(K) = (1/n)\sum_{j=1}^n K((j-.5)/n)c^{\sim}((j-.5)/n)$$

The score function is usually chosen as a sequence of Orthonormal functions $\psi_1(.), \psi_2(.), \ldots$, especially the Legendre polynomials, which test against patterns in the change density $c^{\sim}(\tau)$.

229

The key to change analysis is to choose transformations of data (score the data) which are most powerful for detecting change. From the sample change processes, suitable linear functionals (score the change) are formed. These linear functionals are called "double score components". One can define bivariate density functions $d(\tau, u)$, $0 < \tau < 1, 0 < u < 1$, of which double score functions are diagnostics. Choice of data score functions are motivated in sections 6 and 7 parametrically and non-parametrically, respectively.

Phase 4: *Density estimation.* By one of the many methods available in the theory of curve smoothing (kernel methods, splines, exponential methods, wavelets, etc.) form a smooth estimator $\hat{c}(\tau)$ of the change density.

An exposition of the theory of these phases would require a book and is beyond the scope of this paper. Our goal in this paper is to outline the phases and to explain how we choose transformations of the original data from which to form a change process.

6. Parametric scores change analysis

To detect change over time in a sequence one must have some prior opinion about the ways in which the probability distribution of the observations may be changing (such as in location, scale, skewness, etc). Sample change processes are formed for transformed data, where the transformation is called intuitively a *data score* function. The most powerful data transformations are essentially the sufficient statistics, or more precisely the Fisher score functions, when one has a parametric model $f(y; \theta)$ for a random sample $Y(t), t = 1, \ldots, n$, where $\theta = (\theta_1, \ldots, \theta_k)$.

The maximum likelihood estimator $\theta^{\wedge}$ is obtained by maximizing the average log-likelihood

$$L(\theta) = (1/n) \sum_{t=1}^{n} \log f(Y(t); \theta)$$

Define score functions

$$S_j(Y; \theta) = \partial/\partial\theta_j \log f(Y; \theta)$$

The maximum likelihood estimator is the solution of the *estimating equations* for $j = 1, \ldots, k$

$$(1/n) \sum_{t=1}^{n} S_j(Y(t); \theta^{\wedge}) = 0.$$

Our approach to change analysis asks if for every potential changepoint $\tau = m/n$ the parametric model with $\theta = \theta^{\wedge}$ fits the data $Y(t)$, $t = 1, \ldots, m$, up to the time m in the sense that approximately

$$(1/n) \sum_{t=1}^{m} S_j(Y(t); \theta^{\wedge}) = 0.$$

We define the *score change* process to linearly interpolate its values at $\tau = m/n$, for $m = 1, \ldots, n$

$$C^{\sim}(\tau; S_j) = (1/n) \sum_{t=1}^{m} S_j^*(Y(t); \theta^{\wedge})$$

where

$$S_j^*(Y; \theta^{\wedge}). = S_j(Y; \theta^{\wedge})/E_{\theta^{\wedge}}^{.5}[S_j(Y; \theta^{\wedge})].$$

230

We form k score change processes, for $j = 1, \ldots, k$.

We call this approach "random walk your normalized scores." We are developing the probability theory of the score change processes.

These theoretical concepts can best be understood through examples. Consider a gamma distribution model

$$f(y; \nu, \theta) = (\theta^\nu \Gamma(\nu))^{-1} x^{\nu-1} \exp(-y/\theta)$$

where θ is a positive scale parameter, assumed unknown, and ν is a positive shape parameter, assmed known. One can show that the score function of the parameter θ is

$$S(Y; \theta) = (1/\theta)((Y/\theta) - \nu);$$

the maximum likelihood estimator is

$$\theta^{\hat{}} = Y^-/\nu;$$

the normalized score function evaluated at the maximum likelihood estimator of the parameter may be shown to be

$$S^*(Y(t); \hat{\theta}) = \nu^{.5}((Y(t)/Y^-) - 1).$$

To test the observations $Y(.)$ for change, one forms the maximum likelihood score change process $C^{\sim}(\tau; S^*)$, $0 < \tau < 1$, and tests if this dynamic statistic is significantly different from a sample path of a Brownian Bridge stochastic process. A linear functional of the change process corresponding to the score function

$$K(\tau) = 12^{.5}(\tau - .5)$$

is

$$C^{\sim}(K, S^*) = (1/n) \sum_{t=1}^{n} (12\nu)^{.5}((Y(t)/Y^-) - 1)((t - .5)/n)$$

$$= (12\nu)^{.5}(1/n) \sum_{t=1}^{n} Y(t)((t - .5)/n)/Y^-$$

Under the null hypothesis of no change the asymptotic distribution of $n^{.5} C^{\sim}(K, S^*)$ is Normal(0,1).

An example of an application of this statistic is in Hsu (1979) where it is presented as a test designed for a small change in the scale parameter θ of an independent Gamma distributed sequence, derived by Kander and Zacks (1966) by a Bayesian analysis assuming the changepoint τ is uniformly distributed in time. This test statistic is derived in our approach as analogous to a component in standard goodness of fit analysis.

7. Nonparametric scores change analysis

Our approach to change analysis recommends that one compute and interpret several change processes formed from several transformations of the original data. In addition to (or instead of) various parametric score change processes, one can define various nonparametric score processes for a data sequence $Y(t)$, $t = 1, \ldots, n$. Define:

sample distribution function $F^\sim(y)$;

sample probability mass function $p^\sim(y)$ =fraction of sample equal to y;

mid-distribution function $P^\sim(y) = F^\sim(y) - .5p^\sim(y)$.

The mid-rank data transformation forms $P^\sim(Y(t))$, $t = 1,\ldots,n$. When all Y values are distinct, $P^\sim(Y(t)) = (\text{Rank}(Y(t)) - .5)/n$; we recommend this definition of mid-ranks over the most used definition Rank $(Y(t))/(n+1)$.

One chooses a data score function $J(u)$, $0 < u < 1$, suitable for testing non-parametrically various types of changes in the distribution of the data (especially changes in location or scale parameters). A typical choice for $J(u)$ is a Legendre polynomial normalized to satisfy

$$\int_0^1 J(u)du = 0, \int_0^1 J^2(u)du = 1.$$

Apply the four phases of change analysis to the transformed data sequence $J(P^\sim(Y(t)))$. In the third phase one examines and interprets linear functional tests for change of the form

$$C^\sim(K, J) = (1/n) \sum_{t=1}^{n} K((t - .5)/n)J(P^\sim(Y(t)))$$

for suitable change score functions $K(\tau)$. One can usually show that under the null hypothesis of no change the asymptotic distribution of $n^{.5}C^\sim(K, J)$ is Normal(0,1).

References

Carlstein, E. (1988). "Nonparametric change-point estimation, *The Annals of Statistics*, **16**, 188-197.

Hsu, D. A. (1979) "Detecting shifts of parameter in gamma sequences with applications to stock price and air traffice flow analysis," *Journal of the American Statistical Association*, **74**, 31-40.

Kander, Z., and Zacks, S. (1966). "Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points," *Annals of Mathematical Statistics*, **37**, 1196-1210.

Parzen, Emanuel. (1991). "Unification of statistical methods for continuous and discrete data," *Proceedings Computer Science–Statistics INTERFACE '90*, (eds. C. Page and R. LePage), Springer Verlag: New York, 235-242.

Parzen, Emanuel. (1992). "Comparison change analysis." *Nonparametric Statistics Internatinal Symposium*, (ed. A. K. Saleh), Elsevier: Amsterdam.

AN ANALYSIS OF THE EFFECT OF THE DEGRADED-STATES METHODOLOGY FOR VULNERABILITY ASSESSMENT

William E. Baker

U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD

ABSTRACT

Degraded-states methodology represents a fundamental change in the procedure for assessing the vulnerability of armored fighting vehicles. Results of such assessments serve as input to Army wargame models—large stochastic computer simulations which, for a given scenario, calculate the number of kills for both friendly and enemy forces. In an attempt to determine the ultimate effect of this new methodology, one such wargame was modified to accept the degraded-states input. Three hundred replications of the model were run using conventional input, degraded-states input, and some aggregate of the two. Outputs from the wargame, in the form of partial kills, complete kills, and ratio of kills, were compared using contingency tables to determine whether differences in the results were statistically significant.

I. INTRODUCTION

Degraded-states methodology represents a fundamental change in the procedure for assessing the vulnerability of armored fighting vehicles. The traditional metric for vulnerability analysis was derived from a mapping procedure applying the Standard Damage Assessment List (SDAL) to the calculated damage state of the vehicle. However, studies had shown that the theory behind this metric was problematical in a conceptual sense.[1,2] The degraded-states metric was proposed as an alternative procedure (more appealing because the mathematical foundation is more rigorous).[3]

---

[1] Rapp, J. R. "An Investigation of Alternative Methods for Estimating Armored Vehicle Vulnerability," BRL-MR-03290, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, July 1983.

[2] Starks, M. W. "New Foundations for Tank Vulnerability Analysis." The Proceedings of the Tenth Annual Symposium on Survivability and Vulnerability of the American Defense Preparedness Association (ADPA), Naval Ocean Systems Center, San Diego, CA, 10–12 May 1988.

[3] Abel, J., L. Roach, and M. Starks. "Degraded-States Vulnerability Analysis." BRL-TR-3010, U.S. Army Ballistic Research Laboratory Aberdeen Proving Ground, MD, June 1989.

An immediate concern was how the vulnerability results obtained using degraded-states methodology could be incorporated into the large wargames so prevalent in U.S. Army studies, since these results did not conform to those from traditional analyses. Accordingly, the U.S. Army Ballistic Research Laboratory (BRL) and the U.S. Army Material Systems Analysis Activity (AMSAA) commenced a joint program to further develop and implement the degraded-states methodology. Consequently, the Degraded-States Weapons Analysis Research Simulation (DSWARS) was written.[4] It is an adaptation of a previous stochastic ground-combat simulation and will accept as input the degraded-states metric as well as the traditional SDAL-based metric.

For a specific set of scenarios, DSWARS was run in three different modes—standard damage assessment list, degraded states, and some aggregate derived from the first two. AMSAA and BRL were interested in whether the differences in results were statistically significant, and that is the question which will be addressed in this paper.

II. WARGAME RESULTS

DSWARS was run at two visibility ranges using two types of attacker bullets and two types of defender bullets. Thus, eight (2^3) different scenarios were examined. The simulation was run 300 times for each scenario using each methodology—a total of 7,200 runs. Summary results are shown in Table 1. The first column indicates the scenario, where the first digit represents the attacker bullet (1 = penetrator, 2 = nonpenetrator), the second digit represents the defender bullet, and the third digit represents the range of visibility (3 km, 7 km). A penetrator bullet will penetrate frontal armor; a nonpenetrator bullet will not but may penetrate other armor. The second column indicates the methodology employed, while the third column lists the side. Blue is always defending with three armored vehicles in a hull-defilade mode, and red is always attacking with nine armored vehicles in a fully-exposed mode. The simulation commenced at a range of 4,000 m and continued until the nearest attacker closed to within 500 m of any defender.

The results presented are for a firepower kill on the armored vehicles. The table shows the number of complete kills for both red and blue. The degraded-states methodology calculates partial kills on vehicles, and so the column labeled "any" is the sum of the complete and partial kills. Notice that the

---

[4] Comstock, G. R. "The Degraded-States Weapons Analysis Research Simulation (DSWARS): An Investigation of the Degraded States Vulnerability Methodology in a Combat Simulation." TR-495, U.S. Army Materiel Systems Analysis Activity, Aberdeen Proving Ground, MD, February 1991.

"complete" column is identical to the "any" column for the remaining methodologies. The final two columns are the ratios of red kills to blue kills for both "complete" and "any."

I was asked to compare the three methodologies in a pairwise fashion for four different measures. 1) Attacker and defender complete-firepower kills, 2) attacker and defender any-firepower kills, 3) the ratio of attacker complete-firepower kill to defender complete-firepower kill, and 4) the ratio of attacker any-firepower kill to defender any-firepower kill. Merely comparing the summary results would not provide a clear indication of how DSWARS was reacting to the different methodologies; while the average number of kills might be close over the 300 runs, individual replications could be considerably different. There were also advantages to comparing each scenario individually. The results might be similar at the longer visibility range but disagree significantly at the shorter one. Therefore, I requested the intermediate results (i.e., the number of red kills and the number of blue kills in each replication for each scenario). Furthermore, since these numbers are not independent, the data were categorized into the number of "x" blue kills and "y" red kills, where x = 0, 1, ..., 3 and y = 0, 1, ..., 9.

Intermediate results for the ratio data were a bit more difficult to separate. In an individual replication, if the number of blue kills totaled zero, then the ratio of red kills to blue kills went to infinity. Furthermore, there were a large number of different ratios that were possible. Therefore, the intermediate results for ratios were grouped into ten specified intervals (0.0–0.9, 1.0–1.9, ..., 8.0–8.9, 9.0 and over), placing those ratios equal to infinity into the top interval.

III. STATISTICAL ANALYSIS

In attempting to evaluate the consistency of DSWARS results under the three different methodologies, I have used the chi-square test for differences in probabilities within contingency tables to determine whether or not the differences in output are statistically significant. This procedure can be found in most elementary statistical textbooks.[5] The chi-square test is used to test a hypothesis—in this case, that there is no difference between results obtained using the SDAL method, the degraded-states method, and the aggregate method. These three methods can be considered populations, and output from the 300 replications of DSWARS can be considered random samples from these populations.

---

[5] Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley & Sons, 1971.

Table 2 displays the contingency table from the BMDP statistical software package for one of the scenarios using any-firepower kill ratios, comparing the SDAL method and the degraded-states method. The rows represent the different methodologies; the columns are bins that represent intervals $(x, x + 0.9)$ for $x = 0, 1, ..., 9$, into which the ratio of red kills to blue kills can conceivably fall. The top bin includes ratios defined as infinity. When examining firepower kills rather than ratios, the columns of the contingency tables are bins that represent "x" blue kills and "y" red kills for $x = 0, 1, ..., 3$ and $y = 0, 1, ..., 9$. Recall that blue kills and red kills are combined in the tables since they are not independent (i.e., a large number of red kills would imply a tendency for a small number of blue kills). The intersection of rows and columns form cells that contain the number of occurrences out of 300 replications that such a "column" output results from such a "row" method.

Three assumptions should be satisfied when using the chi-square test for differences in probabilities:

    (1) each sample is a random sample,

    (2) the various samples are all mutually independent, and

    (3) each observation falls into exactly one cell.

Consider the probability that a randomly selected value from the i th population falls into the j th bin. The null hypothesis is that all such probabilities in the same column of the contingency table are equal (i.e., results obtained using the different methodologies are identical). The expected number of observations for each cell, assuming a true null hypothesis, is calculated. The test statistic is defined as

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

where

    r = number of rows

    c = number of columns

    O_{ij} = number of actual observations in cell (i,j)

    E_{ij} = number of expected observations in cell (i,j)

and

$$E_{ij} = \frac{n_i C_j}{N}, \tag{2}$$

where

$$n_i = \text{number of observations in population } i$$

$$C_j = \text{number of observations in bin } j$$

$$N = \text{total number of observations.}$$

This test statistic is computed and subsequently compared with the chi-square distribution.

The chi-square distribution is used as a large-sample approximation, since the exact distribution of the test statistic is difficult to determine. A widely held belief is that this approximation is good, provided that $< 20\%$ of the cells have an expected number of observations < 5. If this is not true, or if there are cells where the expected number of observations is zero, then several categories should be combined to overcome the problem.

The results are shown in Table 3. The left half of the table pertains to kills, both complete-firepower (CF) and any-firepower (AF). The right half of the table pertains to kill ratios, complete-firepower (CFX) and any-firepower (AFX). Each number in the table represents a p-value, which is the probability of being in error if the null hypothesis is rejected. In other words, if the methods are equivalent, the probability of seeing differences equal to or greater than those observed in these samples of 300 is equal to p. A p-value ≤ 0.05 would generally lead to rejection of the null hypothesis. This would lead to the conclusion that results obtained using the different methodologies are dissimilar. Therefore, the methodologies should not be mixed in DSWARS, since it would be unclear whether differences in any subsequent comparisons of war-game results were true differences or merely a manifestation of these dissimilar methodologies. The column labeled "overall" in the table refers to a single comparison of all three methods; columns to the right show the results of pairwise comparisons. The results for the kill ratios seem to mimic those for the kills themselves, with the notable exception of scenario 213 AF.

In summary, the p-values indicate that for complete-firepower kills, the three methods generally give different results. For any-firepower kills, the p-values are slightly larger, but the general conclusions remain unchanged. In each case, the SDAL method and the aggregate method seem to agree more often than the other two pairs—but still not consistently.

237

IV. PROBLEM

I had decided to compare the intermediate results produced by DSWARS because I was concerned that merely examining the final results might mask some subtle difference in methodologies. I also thought it reasonable to compare each scenario individually, hoping this might provide some additional insight. Final results sometimes agree, but for reasons other than similarity of the methodologies. Eventually, such a problem arose.

Comparison of the SDAL method with the degraded-states method provided, what at first appeared to be, inconsistencies. For scenario 123 CF, the complete-firepower kill ratios were close (i.e., 1.4 and 1.3 [Table 1]). For this case, the p-value was 0.013 (Table 3), indicating that the null hypothesis of no difference in results obtained from the two methodologies should be rejected. Contrast that to scenario 223 AF, where the any-firepower kill ratios were quite different (3.5 and 4.3); but the p-value was 0.313, implying that the null hypothesis should not be rejected.

After discussing these outcomes with the developer of DSWARS, I realized that the kill ratios were calculated differently from what I had expected. While the final results for kills are merely the intermediate results averaged over the number of replications, the final ratios are not such averages, but simply the ratio of the final kill results. In other words, I expected

$$Ratio = \left[R_1/B_1 + R_2/B_2 + ... + R_n/B_n \right]/n$$
$$= R_1/nB_1 + R_2/nB_2 + ... + R_n/nB_n \,. \tag{3}$$

Instead, I received

$$Ratio = \frac{[R_1 + R_2 + ... + R_n]/n}{[B_1 + B_2 + ... + B_n]/n}$$
$$= \frac{R_1 + R_2 + ... + R_n}{B_1 + B_2 + ... + B_n} \,. \tag{4}$$

Unless the number of blue kills is the same in each replication, Equation (3) and Equation (4) will give different results for the ratio. If the individual kill ratios (R_i/B_i) include high values, Equation (4) will be less than Equation (3). Of course, Equation (3) breaks down when an intermediate ratio becomes infinity. Apparently, the pros and cons of these two equations had been discussed years ago; Equation (4) remains the procedure of choice for this simulation.

Table 4 supports the p-value from Table 3. For the case of interest, it shows means and standard deviations of the number of kills along with 99% confidence intervals. The first column represents the scenario, while the second indicates the side (blue or red), and the third indicates the type of kill (CF or AF). The final two groups of columns represent the SDAL method and the degraded-states method. Notice that for scenario 123 CF, the 99% confidence intervals for mean blue kills do not overlap. For mean red kills, the confidence intervals overlap slightly. This is an indication that the two methodologies disagree in terms of the number of kills. In spite of this, the ratio of their means is close. The nonoverlapping confidence intervals lend credence to the low p-value of 0.013. Now look at scenario 223 AF. The 99% confidence intervals overlap in both cases, even though the ratio of their means is quite different. Recall that the p-value for this case was 0.313, again intuitive based on the overlapping confidence intervals.

Kill ratios appear to be the most important output of the ground-combat simulations. It is these exchange ratios that the decision makers want to see. Therefore, it was particularly desirable to evaluate how they are affected by the different methodologies. Since the kill ratios calculated in the individual replications are not used in the evaluation of the final ratios, the contingency table approach was inappropriate. Given the procedure for calculating such ratios, a pairwise comparison would seem reasonable, but then the methodologies must be compared over all scenarios. Also, in using these data, there are only eight comparisons, indicating that the statistical test may not have much power (i.e., may not have great ability to detect a false null hypothesis). I did use the Wilcoxon signed-ranks test to examine the null hypothesis between the SDAL method and the degraded-states method. The p-values were 0.484 for the CF ratio and 0.050 for the AF ratio.

V. CONCLUSIONS

DSWARS, a stochastic ground-combat simulation, was written to accept input from three different methods of vulnerability assessment—the standard-damage-assessment-list method, the degraded-states

method, and an aggregate method combining portions of the other two. I used statistical procedures to test the hypothesis that there is no difference in results from using the various methodologies. Three hundred replications were run using each methodology for each of eight different scenarios—a total of 7,200 runs.

The numbers of red kills and blue kills were compared for individual scenarios using contingency tables. The intermediate (individual replication) results were used to test the null hypothesis. Because kill ratios are merely a function of final kill results, no intermediate kill ratios could be used. For this case, I employed the Wilcoxon signed-ranks test to test the null hypothesis. However, it was necessary to make pairwise comparisons over all scenarios, and the paucity of the data (only eight different scenarios) lessens the confidence in the test results.

The primary conclusion is that the SDAL method, the degraded-states method, and the aggregate method produce different results in DSWARS, especially when examining the number of kills for both red and blue. (Exchange ratio results are inconclusive and would benefit from additional data.) Of course, this is not always the case; there are scenarios where they agree quite well. However, in general, we should guard against mixing results from these three methodologies.

Table 1. DSWARS Results - Firepower Kill

Scenario	Methodology	Side	Kills			Ratios	
			Complete	Partial	Any	Complete	Any
113	SDAL	R	7.59	0.00	7.59	5.7	5.7
		B	1.33	0.00	1.33		
113	DS	R	6.17	1.28	7.45	4.8	5.3
		B	1.27	0.12	1.39		
113	AGG	R	7.79	0.00	7.79	6.0	6.0
		B	1.29	0.00	1.29		
123	SDAL	R	3.52	0.00	3.52	1.4	1.4
		B	2.57	0.00	2.57		
123	DS	R	2.92	1.37	4.29	1.3	1.8
		B	2.25	0.18	2.43		
123	AGG	R	4.32	0.00	4.32	1.7	1.7
		B	2.53	0.00	2.53		
213	SDAL	R	8.39	0.00	8.39	13.0	13.0
		B	0.64	0.00	0.64		
213	DS	R	7.25	1.08	8.33	18.6	15.7
		B	0.39	0.14	0.53		
213	AGG	R	8.55	0.00	8.55	14.7	14.7
		B	0.58	0.00	0.58		
223	SDAL	R	5.76	0.00	5.76	3.5	3.5
		B	1.65	0.00	1.65		
223	DS	R	4.32	1.90	6.22	3.9	4.3
		B	1.10	0.34	1.44		
223	AGG	R	6.23	0.00	6.23	4.1	4.1
		B	1.52	0.00	1.52		

Table 1. DSWARS Results - Firepower Kill (Continued)

Scenario	Methodology	Side	Kills			Ratios	
			Complete	Partial	Any	Complete	Any
117	SDAL	R	6.75	0.00	6.75	3.5	3.5
		B	1.91	0.00	1.91		
117	DS	R	5.94	1.01	6.95	3.2	3.4
		B	1.85	0.18	2.03		
117	AGG	R	7.19	0.00	7.19	4.1	4.1
		B	1.76	0.00	1.76		
127	SDAL	R	2.68	0.00	2.68	0.9	0.9
		B	2.89	0.00	2.89		
127	DS	R	2.28	1.08	3.36	0.8	1.2
		B	2.84	0.06	2.90		
127	AGG	R	3.29	0.00	3.29	1.2	1.2
		B	2.80	0.00	2.80		
217	SDAL	R	8.25	0.00	8.25	9.3	9.3
		B	0.89	0.00	0.89		
217	DS	R	7.70	0.75	8.45	12.4	10.6
		B	0.62	0.18	0.80		
217	AGG	R	8.49	0.00	8.49	11.4	11.4
		B	0.74	0.00	0.74		
227	SDAL	R	5.36	0.00	5.36	2.7	2.7
		B	1.98	0.00	1.98		
227	DS	R	4.49	1.70	6.19	2.9	3.3
		B	1.56	0.32	1.88		
227	AGG	R	5.66	0.00	5.66	2.8	2.8
		B	2.05	0.00	2.05		

Table 2. BMDP Output for Contingency Table Example

OBSERVED FREQUENCY

MODEL	KILLS										
	0-0.9	1-1.9	2-2.9	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-	TOTAL
DS	10	30	50	48	9	9	18	35	35	56	300
SDAL	23	37	56	53	6	9	12	31	29	44	300
TOTAL	33	67	106	101	15	18	30	66	64	100	600

MINIMUM ESTIMATED EXPECTED VALUE IS 7.50

STATISTIC	VALUE	D.F.	PROB.
PEARSON CHI-SQUARE	10.485	9	0.3127

EXPECTED VALUES

MODEL	KILLS										
	0-0.9	1-1.9	2-2.9	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-	TOTAL
DS	16.5	33.5	53.0	50.5	7.5	9.0	15.0	33.0	32.0	50.0	300.0
SDAL	16.5	33.5	53.0	50.5	7.5	9.0	15.0	33.0	32.0	50.0	300.0
TOTAL	33.0	67.0	106.0	101.0	15.0	18.0	30.0	66.0	64.0	100.0	600.0

COMPONENTS OF CHI-SQUARE = (OBS - EXP)**2 /EXP

MODEL	KILLS										
	0-0.9	1-1.9	2-2.9	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-	TOTAL
DS	2.6	0.4	0.2	0.1	0.3	0.0	0.6	0.1	0.3	0.7	5.2
SDAL	2.6	0.4	0.2	0.1	0.3	0.0	0.6	0.1	0.3	0.7	5.2
TOTAL	5.1	0.7	0.3	0.2	0.6	0.0	1.2	0.2	0.6	1.4	10.5

Table 3. DSWARS Comparisons: P-Values

Scenario		Kills					Ratios				
	Type	Overall	DS SDAL	SDAL AGG	AGG DS	Type	Overall	DS SDAL	SDAL AGG	AGG DS
113	AF	.559	.283	.940	.352	AFX	.742	.599	.748	.575
	CF	.000	.000	.940	.000	CFX	.000	.000	.748	.000
117	AF	.000	.001	.018	.006	AFX	.001	.029	.044	.001
	CF	.000	.000	.018	.000	CFX	.000	.000	.044	.000
123	AF	.005	.017	.002	.434	AFX	.000	.000	.002	.558
	CF	.000	.000	.002	.000	CFX	.000	.013	.002	.000
127	AF	.001	.000	.004	.070	AFX	.000	.000	.090	.014
	CF	.000	.000	.004	.000	CFX	.000	.012	.090	.000
213	AF	.001	.035	.457	.001	AFX	.713	.790	.355	.732
	CF	.000	.000	.457	.000	CFX	.000	.000	.355	.000
217	AF	.079	.071	.250	.235	AFX	.062	.050	.111	.357
	CF	.000	.000	.250	.000	CFX	.000	.000	.111	.000
223	AF	.012	.342	.006	.014	AFX	.027	.313	.028	.020
	CF	.000	.000	.006	.000	CFX	.000	.000	.028	.000
227	AF	.001	.000	.413	.092	AFX	.002	.000	.881	.004
	CF	.000	.000	.413	.000	CFX	.000	.000	.881	.000

Table 4. Selected Confidence Intervals

Scenario	Side	Type	SDAL			DS		
			$\overline{X}$	S	99% CI	$\overline{X}$	S	99% CI
123	B	CF	2.57	0.72	(2.46, 2.68)	2.25	0.82	(2.13, 2.37)
123	R	CF	3.52	2.25	(3.19, 3.85)	2.92	1.99	(2.62, 3.22)
223	B	AF	1.65	0.93	(1.51, 1.79)	1.44	0.92	(1.30, 1.58)
223	R	AF	5.76	2.02	(5.46, 6.06)	6.21	1.80	(5.94, 6.48)

VARIANCE COMPONENTS IN TANK GUN ACCURACY RESEARCH

David W. Webb
Probability & Statistics Branch
Systems Engineering & Concepts Analysis Division
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
webb@brl.mil   (410) 278-6646

Abstract

Using a nested-factorial design, the Ballistic Research Laboratory conducted a 480-round test to determine if a proposed manufacturing process referred to as dynamic indexing would reduce tube-to-tube variability, σ^2_{Tube}, of the U.S. Army's M256 tank cannon. For each of four different 120-mm ammunition types and for both horizontal and vertical axes, variance component estimates of σ^2_{Tube} were calculated for dynamically indexed tubes and for standard tubes. Using an indirect hypothesis test, estimators for the two tube types were compared. The analysis showed that dynamic indexing does not reduce σ^2_{Tube} as hoped.

I. INTRODUCTION

Through the analysis of several statistically designed experiments, researchers at the U.S. Army Ballistic Research Laboratory (BRL) have learned that tube-to-tube variability* is one of the major contributors to the overall variation in jump for the M1A1 Tank System. Jump is defined as the difference between the aimpoint and impact location after corrections have been made for all known sources of error (e.g., muzzle velocity and wind conditions).

In an ongoing effort to reduce tube-to-tube variation in the M1A1, a concept refered to as dynamic indexing was developed for the M256 cannon. A brief description of this concept follows. Using data describing the gun dynamics during firing, a reference profile is defined that compensates for tube droop *and* minimizes the perturbations to sabot projectiles at ambient temperature. This differs from the standard indexing procedure in which the reference profile only compensates for tube droop. Under both indexing procedures, measurements of the tube centerline profile are taken and the tube mathematically rotated until the best match to the reference profile is obtained. Once this rotation is accomplished, the muzzle upstand, breech-interrupted threads, and breech locking locking plug are machined, thus irrevocably defining "up" for the gun tube.

As part of a prototype test, five gun tubes were randomly selected from the production line at Watervliet Arsenal and were dynamically indexed. These tubes were

---

* "Tube-to-tube variability" is a term used by researchers in the tank gun accuracy community. It is equivalent to what statisticians might refer to as "between-tube variability."

shipped to Aberdeen Proving Ground and fired from a single M1A1 Tank using 105 rounds of 120-mm ammunition. The results of this test showed "a tendency to reduce tube-to-tube variability, providing an improvement in fleet hit probability" (Schmidt, et. al., 1989). Based on these initial results, a more complete proof-of-principle test was recommended.

II. TEST PLAN

The primary objective of the test was to determine if the tube-to-tube variability of dynamically indexed tubes (DITs) is statistically smaller than the tube-to-tube variability of standard tubes (STs). In other words, "Does dynamic indexing of the M256 cannon reduce the dispersion of centers-of-impact?" (see Figure 1).

The Probability & Statistics Branch of the BRL devised several experimental design plans and presented them before the Project Manager for Tank Main Armament Systems. The relative merits of the design plans were discussed with concerns expressed for economy of available resources and robustness of test conditions.
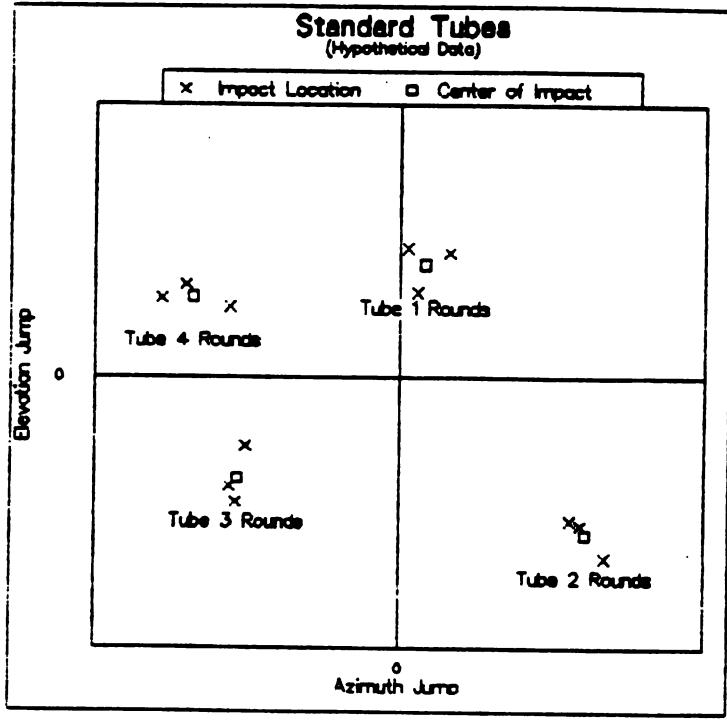
Ultimately, one of the plans was selected and slightly modified for a detailed comparison of DITs and STs. The design called for four types of 120-mm ammunition, twenty DITs, and twenty STs. To make the experiment and the results as robust as possible, it also called for four tanks and four ammunition temperatures. However, in order to reduce the amount of testing, the tank and ammunition temperature factors were confounded. Therefore, if this confounded factor was found to be statistically significant, one would be unable to distinguish between the Tank effect, the Ammunition Temperature effect, or their interaction. For each treatment combination, three rounds were fired. The complete test design matrix is shown in Figure 2.

III. DATA MATRIX

The trajectory of each round was monitored at four different ranges: 800 m, 1500 m, 2400 m, and 3000 m. From measurements taken at these ranges, azimuth and elevation jumps were computed and recorded. For various reasons, such as short-landing rounds or range equipment failures, the jumps were not always obtained at each of the four ranges. The percentage of missing data at the 1500-m and 3000-m targets precluded any formal analysis at these ranges. Exploratory analysis and the computation of simple descriptive measures of the 800-m and 2400-m data indicated that the 2400-m azimuth jumps were strongly affected by wind. Therefore, the analysis was concentrated on the data recorded at 800 m, since at this shorter range the data are believed to be less influenced by wind and other flight conditions.

Using a procedure that relies on the assumption that rounds fired on the same occasion follow a similar trajectory profile, jump estimates were made for those rounds in which the 800-m field data were missing.

Because of the substantial differences (e.g., aerodynamics and threat capabilities) between the ammunition types, the comprehensive data set was divided into four 120-round subsets to be analyzed separately. Furthermore, the azimuth and elevation jump values were assumed to be independent, so that separate analyses were performed on the azimuth jumps and the elevation jumps.

246

Standard Tubes
(Hypothetical Data)

Hypothetically, dynamic indexing of the M256 tank cannon lowers tube-to-tube variability, σ^2_{Tube}, thereby reducing the dispersion of the centers-of-impact.


Dynamically Indexed Tubes
(Hypothetical Data)
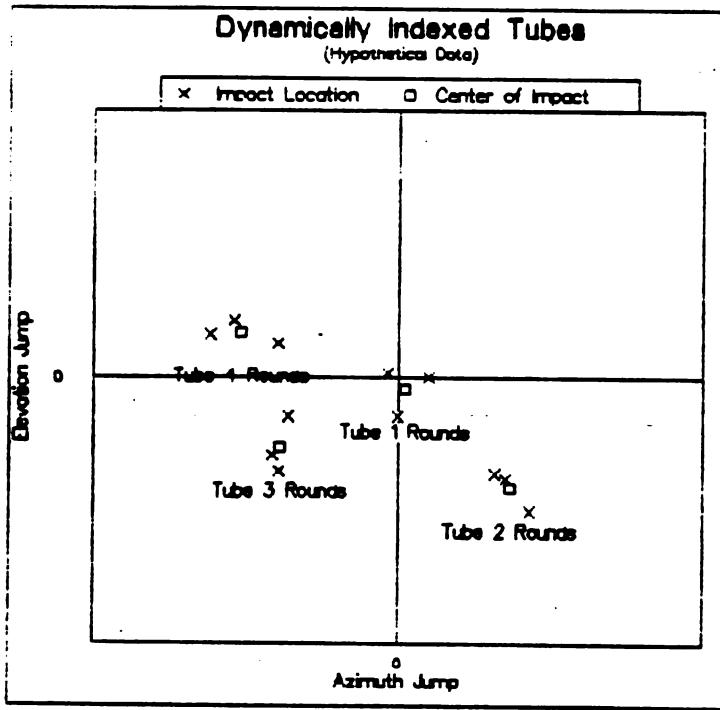
Figure 1. The Goal of Dynamic Indexing

Tank/Temp	Tube Type	Tube ID	Ammunition Type			
			1	2	3	4
1	ST	1	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		2	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		3	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		4	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		5	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
	DIT	1	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		2	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		3	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		4	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		5	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
2	ST	6	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		7	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		8	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		9	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		10	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
	DIT	6	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		7	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		8	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		9	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		10	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
3	ST	11	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		12	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		13	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		14	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		15	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
	DIT	11	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		12	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		13	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		14	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		15	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
4	ST	16	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		17	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		18	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		19	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		20	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
	DIT	16	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		17	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		18	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		19	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)
		20	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)	(x,y) (x,y) (x,y)

Figure 2. 480-Round Test Design

With simple exploratory procedures, some potential outliers in the data were flagged. After conferring with test directors and other researchers involved in the study, some of these jump values were corrected while others remained unchanged. No round was deleted from the analysis just because it seemed to be a "flier". All jump values were eventually accepted as accurate and included in the analysis.

IV. STATISTICAL ANALYSIS

The following discussion will examine how the comparison between DIT and ST tube-to-tube variability was made for any of the eight (four ammunitions for azimuth and elevation) subsets of the entire data set.

The analysis strategy was to obtain an estimate of the tube-to-tube variance for those rounds fired from DITs, and to compare it with the tube-to-tube variance estimate

248

for STs, (i.e., $\hat{\sigma}^2_{Tube\text{-}DIT}$ versus $\hat{\sigma}^2_{Tube\text{-}ST}$). This required dividing the data in half into 60 DIT rounds and 60 ST rounds. The test design for such a set of 60 jump values appears in Figure 3. The model for the standard tubes azimuth data, for example, is

$$x_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{k(ij)}$$

where,

x_{ijk} = jump of kth round from j^{th} standard tube on i^{th} tank/temp, measured in mils;

μ = overall mean;

α_i = effect of ith tank/temp, $i = 1, 2, 3, 4$;

$\beta_{j(i)}$ = effect of jth standard tube on i^{th} tank/temp, $j = 1, 2, 3, 4, 5$;

$\epsilon_{k(ij)}$ = error of kth round from j^{th} standard tube on i^{th} tank/temp, $k = 1, 2, 3$.

Tank/Temp	Tube ID	Jump		
1	1	x	x	x
	2	x	x	x
	3	x	x	x
	4	x	x	x
	5	x	x	x
2	6	x	x	x
	7	x	x	x
	8	x	x	x
	9	x	x	x
	10	x	x	x
3	11	x	x	x
	12	x	x	x
	13	x	x	x
	14	x	x	x
	15	x	x	x
4	16	x	x	x
	17	x	x	x
	18	x	x	x
	19	x	x	x
	20	x	x	x

Figure 3. 60-jump test design from which the variance component estimate $\hat{\sigma}^2_{Tube}$ is derived. For each combination of Direction (azimuth or elevation), Tube Type (DIT or ST), and Ammunition Type (1, 2, 3, or 4), the test design is as shown above where each "x" represents a jump value.

The confounded factor, Tank/Temp, and the nested factor, Tube, were each considered to be random. Furthermore, each α_i, $\beta_{j(i)}$, and $\epsilon_{k(ij)}$ was assumed to be normally distributed with mean zero and variance $\sigma^2_{Tank/Temp}$, $\sigma^2_{Tube\text{-}ST}$, and $\sigma^2_{Rnd\text{-}ST}$, respectively.

249

For an analysis-of-variance under these model assumptions, the degrees-of-freedom and expected mean-squares (EMS) associated with each factor are:

Source	df	EMS
Tank/Temp	3	$\sigma^2_{Rnd\text{-}ST} + 3\sigma^2_{Tube\text{-}ST} + 15\sigma^2_{Tank/Temp}$
Tube	16	$\sigma^2_{Rnd\text{-}ST} + 3\sigma^2_{Tube\text{-}ST}$
Round	40	$\sigma^2_{Rnd\text{-}ST}$

The method-of-moments estimator for the variance component $\sigma^2_{Tube\text{-}ST}$ is given by

$$\hat{\sigma}^2_{Tube\text{-}ST} = \frac{MS_{Tube\text{-}ST} - MS_{Rnd\text{-}ST}}{3}.$$

Likewise, for the DIT data, an estimator for $\sigma^2_{Tube\text{-}DIT}$ is

$$\hat{\sigma}^2_{Tube\text{-}DIT} = \frac{MS_{Tube\text{-}DIT} - MS_{Rnd\text{-}DIT}}{3}.$$

Once each estimate was computed, there was still the task of how to determine if there was a significant reduction in tube-to-tube variance with dynamic indexing. Unfortunately, this could not be done directly since the distributions of each $\hat{\sigma}^2_{Tube\text{-}DIT}$ and $\hat{\sigma}^2_{Tube\text{-}ST}$ are extremely complex and there was no known method for comparing two random variables coming from such distributions.

An indirect test for the equality variance components was needed, and after conferring with several academicians a novel approach was attempted (Mathew 1991). For each data set, the round-to-round variability is denoted by either $\sigma^2_{Rnd\text{-}DIT}$ or $\sigma^2_{Rnd\text{-}ST}$ depending upon the tube type. While there is no reason to believe that dynamic indexing has any effect on the round-to-round variance, this assumption can be tested. Estimators for $\sigma^2_{Rnd\text{-}DIT}$ and $\sigma^2_{Rnd\text{-}ST}$ are $MS_{Rnd\text{-}DIT}$ and $MS_{Rnd\text{-}ST}$, which are multiples of χ^2 random variables with 40 degrees of freedom. Therefore,

$$F = \frac{MS_{Rnd\text{-}DIT}/40}{MS_{Rnd\text{-}ST}/40} = \frac{MS_{Rnd\text{-}DIT}}{MS_{Rnd\text{-}ST}} \sim F_{40,40}$$

is a test statistic for the two-tailed hypothesis test

$$H_O: \ \sigma^2_{Rnd\text{-}DIT} = \sigma^2_{Rnd\text{-}ST}, \quad \text{against}$$

$$H_A: \ \sigma^2_{Rnd\text{-}DIT} \neq \sigma^2_{Rnd\text{-}ST}.$$

Next, note that

$$F' = \frac{MS_{Tube\text{-}DIT}/16}{MS_{Tube\text{-}ST}/16} = \frac{MS_{Tube\text{-}DIT}}{MS_{Tube\text{-}ST}} \sim F_{16,16}$$

250

is a test statistic for the hypothesis test

$$H_O': \quad \sigma^2_{Rnd\text{-}ST} + 3\sigma^2_{Tube\text{-}ST} = \sigma^2_{Rnd\text{-}DIT} + 3\sigma^2_{Tube\text{-}DIT}, \quad \text{against}$$

$$H_A': \quad \sigma^2_{Rnd\text{-}ST} + 3\sigma^2_{Tube\text{-}ST} > \sigma^2_{Rnd\text{-}DIT} + 3\sigma^2_{Tube\text{-}DIT}.$$

However, if the ratio $F = MS_{Rnd\text{-}DIT}/MS_{Rnd\text{-}ST}$ is insignificant, then F' becomes an indirect test statistic for the equality of tube-to-tube variability; that is, $F' = MS_{Tube\text{-}ST}/MS_{Tube\text{-}DIT}$ tests the hypothesis

$$H_O'': \quad \sigma^2_{Tube\text{-}ST} = \sigma^2_{Tube\text{-}DIT}, \quad \text{against}$$

$$H_A'': \quad \sigma^2_{Tube\text{-}ST} > \sigma^2_{Tube\text{-}DIT}.$$

On the other hand, if the test for equality of round-to-round variabilities is significant, then this would contaminate the indirect approach suggested above. In such case, graphical comparison of the variance components estimates, $\hat{\sigma}^2_{Tube\text{-}DIT}$ and $\hat{\sigma}^2_{Tube\text{-}ST}$ with other pairs of estimates that are known to differ could provide a subjective evaluation of $\sigma^2_{Tube\text{-}DIT}$ and $\sigma^2_{Tube\text{-}ST}$.

V. RESULTS

Due to the classification of this subject matter, numerical results cannot be divulged. Qualitatively, the overall conclusions were quite surprising and somewhat disappointing. For the 120-mm ammunition type that was used in the prototype test, which showed that dynamic indexing resulted in a lowering of σ^2_{Tube}, the azimuth *and* elevation tube-to-tube variance estimates were statistically greater with DITs than they were with STs. The reason for this is still unknown.

Among the other 120-mm ammunition types, only one showed a significant reduction in σ^2_{Tube} with dynamically indexed tubes and that occurred only in azimuth. All other cases were insignificant.

VI. REFERENCES

[1] Schmidt, E., J. Bornstein, J. Thomas, and T. McCloskey. "A Method for Indexing Tank Cannon." BRL-IMR-912, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, October 1988 (CONFIDENTIAL).

[2] Mathew, Thomas. Letter to the author and Jerry Thomas. February, 1992.

IDENTIFYING EXPERIMENTAL VARIABILITY AND ITS INFLUENCE ON THE RESEARCH, DEVELOPMENT, TESTING, AND EVALUATION PROCESS

Jock O. Grynovicki
U.S. Army Laboratory Command
Human Engineering Laboratory
Aberdeen Proving Ground, Maryland
21005-5001

ABSTRACT

Uncontrolled sources of variability during experimentation may mask significant results and may consequently hinder the ability of the Department of Defense (DoD) decision makers to draw conclusions concerning weapon systems or standing operating procedures (SOPs). To protect against drawing the wrong conclusions, one must use experimental methodology that reduces, identifies, or controls sources of variability. To illustrate this point, a field exercise, which is designed to determine how quickly and accurately soldiers can identify enemy ordnance using a prototype expert system compared to the standard 60 series technical manual, is presented, along with results and conclusions. A design strategy that allows experimenters to estimate and test these uncontrolled sources of variability is provided.

INTRODUCTION

Since decision making based on the use of statistical tools almost always involves collection of data, the way in which the data are collected becomes extremely important. The design of an experiment has been defined very simply as the order in which a combination of experimental variables is run or controlled.

Experimental designs are employed in the Army's research, developmental testing, and evaluation efforts to assure that unbiased and correct decisions are made regarding equipment, SOPs, and weapon systems.

Variation produced by disturbing factors, both known and unknown, is called experimental error. Important effects may be wholly or partially obscured by experimental error. Conversely, through experimental error, the experimenter may be misled into believing in effects that do not exist. Experimental designs are used to help reduce the experimental error in the data collected. Randomization or counterbalancing is employed to try to help average the effect of many extraneous factors which may be present in an experiment.

Unfortunately, uncontrolled sources of variability during experimentation may mask significant results and may consequently hinder the ability of the Department of Defense (DoD) decision makers to draw conclusions concerning weapon systems or SOPs. To protect against drawing the wrong conclusions, one must use experimental methodology that reduces, identifies, or controls sources of variability. One tool is to use "Model-Based Diagnostics" associated with variance-component estimations to assess the data and the experimental model as proposed by Hocking (1985), demonstrated by Grynovicki and Green (1988), and published by Hocking, Green, and Brener (1989), and Grynovicki (1990).

These new closed form expressions for the estimators of variance components have been developed, based on the equivalence shown in Hocking, Green, and Brener (1990) of the variance component estimation problem to the problem of estimating the covariance, θ_t, between appropriately related observations. In addition, these estimators have been shown to provide information that will be useful in diagnosing problems and to suggest simple graphical procedures for examining the influence of the treatment levels. (Grynovicki, and Green (1988)).

To illustrate this point, a field exercise designed to determine how quickly soldiers can identify enemy ordnance, using a prototype expert system compared to the standard 60 series technical manual, is presented, along with results, diagnostics, and conclusions. A design strategy that allows experimenters to estimate, test, and identify these uncontrolled sources of variability is provided.

GENERAL VARIANCE COMPONENT ESTIMATES
AND MODEL DIAGNOSTIC METHODOLOGY

Before introducing the field study, a general description of the model diagnostics is provided. For brevity, this discourse is limited to a hierarchical model with factor 2 random and nested in factor 1 and crossed with factor 3. All other factors will be considered fixed.

The number of levels of factor (i) is designated by a_i. The traditional univariate repeated measures mixed model is

$$Y_{ijkm} = \mu + A_i + AS_{j(i)} + C_k + AC_{ik} + CAS_{kj(i)} + E_{(ijk)m} \qquad 2.1$$

in which μ is the grand mean, A_i is the effect of level i of treatment or factor A, $AS_{j(i)}$ is the effect of factor 2 nested into factor 1, C_k is the effect of level k of the third treatment, AC_{ik} is the effect of the AC treatment combination of level ik, $CAS_{kj(i)}$ is the effect of levels $kj(i)$ of treatment 3 crossed with treatment 2 nested in 1, and $E_{(ijk)m}$ is the random experimental error. For the traditional univariate repeated measures approach, it is assumed that A_i, C_k, AC_{ik}, and μ are fixed treatment means, and $AS_{i(j)}$, $CAS_{kj(i)}$, and $E_{(ijk)m}$ are zero mean, independent normal random variables with variances ϕ_{12}, ϕ_{123}, and ϕ_0, respectively. While the variables are independent, the responses are correlated with the covariance structure.

$$\text{Cov} (\overline{Y}_{ijkm} \overline{Y}_{i*j*k*m*}) =$$

$$
\begin{array}{lll}
0 & \text{if} & i \neq i* j \neq j* \\
\theta_{12} = \phi_{12} & \text{if} & ij = i*j*, \ k \neq k* & 2.2 \\
\theta_{123} = \phi_{12} + \phi_{123} & \text{if} & ijk = i*j*k*, \ m \neq m* \\
\phi_0 + \theta_{123} & \text{if} & ijkm = i*j*k*m*
\end{array}
$$

The covariance θ_t is between observations at the same level of factors indexed by t and different levels of all other factors in the model. If we think of the data as arranged in a two-way table with a_1 times a_2 rows and a_3 columns, we see that ϕ_{12} is the covariance between observations in the same row but different columns. Under the assumption that there are no uncontrolled sources of variability and the model and its assumptions are

254

correct, these covariances are assumed to be the same for all rows and all pairs of columns. Theta $_{123}$ is the covariance between observations in the same cell and is assumed to be the same for all cells. If the experiment is not replicated, this estimate is not estimable and considered zero.

This suggests examining the corresponding sample covariance. These sample covariances, or averages thereof, yield the estimators of the θ_t. The sample covariance yielding the estimate of ϕ_{12} is

$$\hat{\theta}_{12} = (a_1\, a_3\, r_3)^{-1} \sum_{\substack{k \neq k^* \\ i}} (r_2)^{-1} \sum_j (\bar{Y}_{ijk\bullet} - \bar{Y}_{i\bullet k\bullet}) (\bar{Y}_{ijk^*\bullet} - \bar{Y}_{i\bullet k^*\bullet}) \qquad 2.3$$

in which a_i is the number of levels of factor i, $r_i = (a_i-1)$, and $\bar{Y}_{ijk\bullet}$ is the cell mean for cell (ijk).

From equation 2.3, one recognizes the θ_{12} estimator as an average of $a_1a_3r_3$ equal sample covariances corresponding to all combinations of $k \neq k^*$, i=1 to a_1. These covariances that are averaged contain the diagnostic power in determining if the data contains outliers, there is uncontrolled sources of variability, or if the data does deviate from the underlying assumptions.

The distribution theory for these diagnostics is developed in Grynovicki (1989). In this dissertation, the covariance is written in matrix notation which is referred to in the literature as a bilinear form. For the above model, θ_{12} can be written $Z_1'\, A\, Z_2$ in which

$$Z_1' = (\bar{Y}_{i1k\bullet},\, \bar{Y}_{i2k\bullet},\, \bullet \bullet \bullet,\, \bar{Y}_{ia_2k\bullet}),$$

$$2.4$$

$$Z_2 = (\bar{Y}_{i1k^*\bullet},\, \bar{Y}_{i2k^*\bullet},\, \bullet \bullet \bullet,\, \bar{Y}_{ia_2k^*\bullet})$$

and $A = (I_{a_2} - J_{a_2}\, J_{a_2}')$. I_{a_2} is the identity matrix of dimension a_2, and J_{a_2} is a row of ones of length a_2. The variance covariance structure of (Z_1, Z_2) determines the probability density function. If the variance covariance matrix has the form

$$\begin{pmatrix} a(I) & c(I) \\ c(I) & a(I) \end{pmatrix} \qquad 2.5$$

with a and c being linear combinations of the treatment variances in the model, one can consider Z_1 and Z_2 as independent, and the distribution of the diagnostics is

$$\frac{n}{(2\pi)^{1/2} a^{1/2} (1-\rho^2)^{1/2}} \int_0^\infty \frac{S^{(n-4)/2} \exp(-S/(2a))}{(2a)^{(n-3)/2} ((n-1)/2)!} \times \exp\left[-\frac{(nc-(\rho a^{1/2}S)/a)^2}{2a(1-\rho^2)S}\right] dS. \quad 2.6$$

If (Z_1, Z_2) have a variance-covariance structure

$$\begin{pmatrix} a(I)+b(JJ') & c(I)+d(JJ') \\ c(I)+d(JJ') & a(I)+d(JJ') \end{pmatrix} \quad 2.7$$

one can consider this case as dependent, and the distribution as derived in Grynovicki (1989) is

$$\sum_i \sum_j \frac{q_i q_j t^{(2M+2i+2j-2)/2} e^{(t/2a_1)}}{2^{(M+2i+M+2j)/2} a_1^{(M+2i)/2} b_1^{(M+2j)/2} \Gamma((M+2i)/2) \Gamma((M+2j)/2)} \times \quad 2.8$$

$$\left[\int_0^\infty e^{-((a_1+b_1)/(2a_1 b_1))w} w^{(M+2j-2)/2} (1+w)^{(M+2i-2)/2}\right] dw.$$

An interactive computer program that calculates the cumulative distribution of the variance estimates and diagnostics for either distribution has been written and is described in Grynovicki and Green (1990). With the use of this program, Army experiments can determine if the covariances are consistent or abnormal given the particular experiment. Thus, inconsistencies with the data or inadequacies with the model that may lead to erroneous conclusions can be identified and adjusted.

The program is written in Turbo-Pascal® and can be run on any IBM-compatible personal computer.

ARMY FIELD EXPERIMENT

To illustrate how uncontrolled experimental error can influence the decision process and how the model diagnostics can reduce, identify, or control sources of variability, a field exercise designed to determine how quickly and accurately soldiers can identify enemy ordnance, using a prototype expert system compared to the standard 60 series technical manual is presented, along with results and conclusions. For illustration, only the time data will be discussed. It is worth mentioning that both systems were fairly accurate with AIRES being slightly but not significantly better (Grynovicki, Miller, and Krass (1990)).

The U.S. Army Electronics Technology and Devices Laboratory (ETDL) developed a prototype explosive ordnance disposal (EOD) automated information

retrieval and expert system (AIRES) to assist an EOD team to disarm, detonate, or otherwise "render safe" ordnance. Traditionally, the EOD team's source of technical information is the technical manual (TM) 60 series manual which has 15,000 pages of paper and 1,500 sheets of microfiche. AIRES transforms a portion of the TM 60 series manual into a portable computer-based system capable of assisting the EOD technician in the identification and render safe procedure.

This proof-of-principle prototype system (see Figure 1), which acts as an intelligent interface between the user and a large complex data base, features (a) expert system software for information retrieval, (b) an optical disk for data storage, (c) a flat panel display for presenting graphics and text data, and (d) touch interaction for operator input.

The operation of AIRES requires the user to enter the known characteristics of the unidentified ordnance. The ordnance is identified when a match between the characteristics entered and the attributes of an item in the data base has been achieved. As new characteristics are entered, the system monitors the number of items in the data base that match the user input. In the prototype system, data are entered by touching menu choices and keypad entries presented on the display. When all known data have been entered or when a unique match has been obtained, the user is presented with an engineering drawing of the ordnance and is asked to confirm the identification. This pairing was because the test participants (TPs) could only identify the same ordnance once.

Text Matrix

A repeated measures design was used to expose each TP to the manual and automatic methods of identifying ordnance for disarmament. Groups were the between-subject factor, and methods were the within-subject factor. The experiment was counterbalanced for site, method, and ordnance (see Table 1).

Since the subject could not identify the same ordnance more than once, the six items of ordnance were paired and grouped into two groups of three by difficulty (i.e., a mortar round (a) in group A had the same general identification characteristics and difficulty as a mortar round (f) in group B). The pairing resulted in the following initial grouping:

Group A   a b c
Group B   d e f

The ordnance was then counterbalanced to assure that the same ordnance was tested an equal number of times for each method.

Subjective Evaluation

The independent variables were method of classification (standard versus AIRES) group (five groups of subjects evaluated at different times), and subjects. The dependent variables were the time to identify the ordnance and number of correctly identified trials.

257

Figure 1. Explosive ordnance disposal automated information retrieval and expert system (EOD AIRES)

Table 1

Randomization Scheme and Experimental
Design for AIRES Evaluation

Experiment I
August 1988

Group	Subject	STANDARD			AIRES		
					Method		
1	1	F	A	B	C	D	E
	2	E	F	A	B	C	D
	3	C	D	E	F	A	B
	4	A	B	C	D	E	F
	5	D	E	F	A	B	C
	6	B	C	D	E	F	A
2	1	C	D	E	F	A	B
	2	B	C	D	E	F	A
	3	A	B	C	D	E	F
	4	F	A	B	C	D	E
	5	E	F	A	B	C	D
	6	D	E	F	A	B	C
3	1	A	B	C	D	E	F
	2	B	C	D	E	F	A
	3	C	D	E	F	A	B
	4	D	E	F	A	B	C
	5	E	F	A	B	C	D
	6	F	A	B	C	D	E
4	1	F	A	B	C	D	E
	2	E	F	A	B	C	D
	3	D	E	F	A	B	C
	4	C	D	E	F	A	B
	5	B	C	D	E	F	A
	6	A	B	C	D	E	F
5	1	D	E	F	A	B	C
	2	C	D	E	F	A	B
	3	B	C	D	E	F	A
	4	A	B	C	D	E	F
	5	F	A	B	C	D	E
	6	E	F	A	B	D	D

ROUNDS:
A Model 500/1, 122mm, fuze model RGM-2
B Model D-832 Smoke Mortar, 82mm, fuze model M-6
C Model OF-A, HE Mortar, 120mm, fuze model MZ-31
D Model M59, 122mm, fuze model M51A5
E Model EBK-5M or ZBK-5K, Heat, USSR, 100mm, fuze model GPV-2
F Model 365-K, Fixed Frag., USSR, 85mm, fuze model KTM-1

To try to control for subject variability, all participants had to be graduates from the EOD Advanced Training Course at the U.S. Army Ordnance Missile and Munitions Center and School (USAOMMCS) and had to have experience using the TM 60 manual. In addition, all participants were given individual training about AIRES by the Armament Research, Development, and Engineering Center (ARDEC) using munitions that were different but equal in difficulty to the munitions used in the field exercise. The TPs were given detailed instruction about entering known characteristics of the unidentified on AIRES using the touching menu choices and keypad entries. They were administered a pretest trial and had to correctly identify fire ordnance before they were considered proficient.

Originally, 40 military personnel from USAOMMCS and the U.S. Army Forces Command (FORSCOM) were requested, but only 30 TPs were obtained. Based on the pool of subjects that were available, the participants in this experiment had a diversity of military occupational specialties (MOSs) from various armed forces as well as a diversity of EOD experience as shown in Table 2.

Table 2

Summary of EOD Experience by Armed Forces

Service	Percent of total	TPs	EOD average experience
Army	26.7	8	10.5
Navy	20.0	6	8.2
Marines	16.7	4	9.5
Air Force	26.6	8	7.6
Unknown	10.0	1[a]	12.0

[a]Three additional TPs with no report of years of experience.

RESULTS AND DISCUSSION

The results of the analysis for identification time are shown in Table 3.

The traditional analysis of variance (ANOVA) indicated no significant difference in time between the two methods. The average time to identify ordnance using the standard method was 8.41 minutes as compared to 7.46 minutes to identify the ordnance using AIRES. One reason for being unable to determine a significant difference between the classification methods may have been because of a significant subject-within-group-by-method effect which is used as an error term for method. Possible experimentation for this inflated error term may be attributable to uncontrolled sources of variability, and outliers.

260

Table 3

ANOVA
August 1988
(time in minutes)

Source	DF	SS	F	Significance (.05)
Groups	4	215.7	1.91	N.S.
Subject within group	21	593.7	0.94	N.S.
Method	1	35.1	0.69	N.S.
Method by group	4	97.4	0.48	N.S.
Method by subject within group	21	1067.8	1.69	S.
Error	104	3132.0		

Note. S. = significant; N.S. = nonsignificant

To obtain insight into the nature of this significant interaction and the underlying linear model, the variance components diagnostics for the random effect subject within group (ϕ_{12}) will be evaluated for consistency. Based on the underlying assumption of sphericity associated with the repeated measures model, the variance components should be statistically similar. These four components are equal to the sample covariance across subjects for a specific group between the two different methods and are presented graphically in Figure 2.
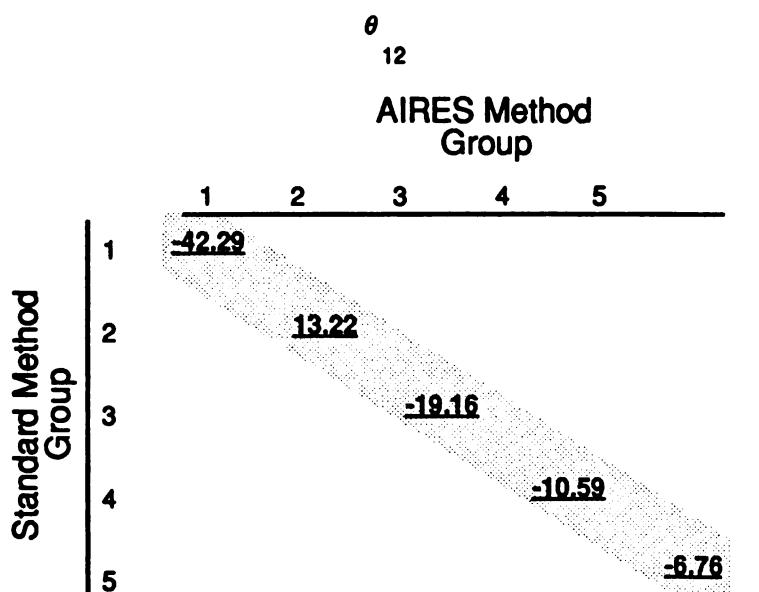

Figure 2. Variance component diagnostics for the random effect subject-within-group (Experiment I).

In rewriting these covariances as a bilinear form, it can be shown that the variance-covariance matrix of the two response vectors exhibits the independent form as outlined in equation 2.5 and therefore, the distribution of the diagnostics follows equation 2.6.

A computer program to calculate the cumulative distribution of the variance component diagnostics is presented in Grynovicki (1989). The program is written in Turbo-Pascal Version 5.0®, and can be compiled and run on any IBM personal computer, provided Turbo-Pascal 5.0 is available.

The program uses Simpson's integration method and requires calculation of the modified Bessel function for both integer and fractional order.

When the program is run, a menu appears that prompts the user about whether the diagnostic being investigated should be considered the dependent or independent case. Once the distribution is selected, the user is asked to enter the standard deviations of each response vector

$$\overline{Y}_{ijk\bullet} \quad \text{and} \quad \overline{Y}_{ijk\bullet}^{\star}$$

as well as the covariance between the selected pairs. The program also asked for the sample size and range of integration.

For the diagnostics for ϕ_{12}, the variance was estimated as 11.2 and covariance as −0.523 based on the method of variance component estimate found in Hocking, Green, and Brenner (1990). The sample size used was six which is conservative, and range was decided interactively until the cumulative distribution reached one.

The 5th percentile for the diagnostic distribution was estimated as −23.12, and the 95th percentile associated with a diagnostic had a value of 3.73. Thus, −42.29 and 13.22 fall outside this 95th percentile confidence interval and indicate a problem with the underlying linear model and its assumptions.

In investigating this unexplained square of variability, the average time to identify the three ordnances was examined by the demographics of the military personnel. As shown in Table 4, the Army military personnel identified the ordnance slightly quicker than military personnel from other armed forces. In addition, experience seemed to influence identification time. Military personnel with more than 5 years' experience identified ordnance more quickly with both methods than military with 5 years' experience or less (see Table 5). For the less experienced group, identification time was decreased an average of 3.3 minutes using AIRES versus the manual system, while the experienced group had a decrease of only 1.1 minutes. Thus, AIRES seems to benefit the less experienced soldier more. To investigate these trends, a second experiment was conducted using soldiers with less than 5 years' experience.

262

Table 4

Average Identification Time of Three Ordnance by Armed Forces

Armed Forces	Manual	AIRES
Army	23.5 minutes	22.0 minutes
Other[a]	24.7 minutes	23.0 minutes

[a]Others include Air Force, Navy, and Marine Corps personnel.

Table 5

Average Identification Time of Three Ordnance by EOD Experience

Experience	Number of Subjects	Manual	AIRES
1 to 5 years	6	26.8 minutes	23.5 minutes
More than 5 years	21	23.8 minutes	22.7 minutes

Experiment II

Speculating that the inflated error term for method may be because of the wide range of EOD experience and service background of the subjects, it was decided that a second field experiment with TPs having an equal amount of EOD experience and service background (all Army) be conducted. All participants had 2 years 6 months or less EOD experience as summarized in Table 6. A repeated measures design was also chosen for this experiment with two groups of six subjects each. Time was the dependent measure. Only 11 subjects successfully completed the field trial. The assumption of sphericity using Box's M could not be rejected. The diagnostics for ϕ_{12} for the second experiment is shown in Figure 3. The variance for the response vectors was estimated using the method previously discussed as 17.91 with a correlation of 0.323. The diagnostic was determined to fall under the independent distribution. The sample size of six was used. The 95th percentile confidence interval was calculated to be between and including -6.68 and 21.08. As seen from Figure 3, the data appear consistent, and the linear model and underlying assumption of sphericity appear to hold; the assumption of sphericity using Box's M statistic could not be rejected.

θ_{12}

Standard Method Group
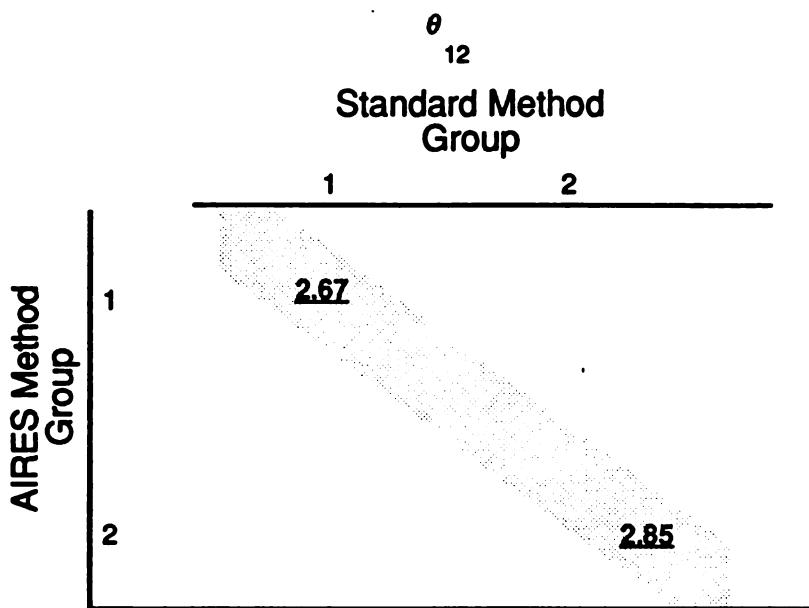
Figure 3. Variance component diagnostics for the random effect subject-within-group (Experiment II).

Table 6

Subjects' Training and Experience
Experiment II
November 1988

Group	Subject	Service and MOS	EOD experience (years and months)	
1	1	USA/55D10	1	11
	2	USA/55D10	2	2
	3	USA/55D10	2	5
	4	USA/55D10	1	0
	5	USA/55D10	1	11
2	1	USA/55D10	0	11
	2	USA/55D10	1	6
	3	USA/55D10	1	2
	4	USA/55D10	2	0
	5	USA/55D10	2	6

A univariate analysis for repeated measures was then performed. The results are shown in Table 7.

264

Table 7

ANOVA
Experiment II
November 1988
(time in minutes)

Source	DF	SS	F	Significance (.05)
Groups	1	29.99	0.42	N.S.
Subject within group	9	640.85	1.95	N.S.
Method	1	351.82	9.70	S.
Method by group	1	46.94	1.29	N.S.
Method by subject within group	9	326.13	1.00	N.S.
Error	43	1568.50		

Scheffé's Test

Scheffé Grouping Mean

A	10.85	STANDARD
B	6.19	AIRES

Note. S. = significant; N.S. = nonsignificant

The F variable for method (F [1,9] = 9.70) indicated that there were statistically significant differences between the automatic and manual methods of identification. A soldier using the standard method took almost 11 minutes to identify the ammunition, while using the AIRES method decreased the average time by 4.6 minutes to 6.18 minutes. There were no significant differences between soldiers or interactions regarding time of identification. All soldiers but one identified the ordnance faster using AIRES as shown in Table 8.

Thus, there was a significant identification time difference between the automatic (AIRES) and manual (standard) methods when soldiers with limited experience and the same MOS (Experiment II) were evaluated. Soldiers in this group identified ordnance significantly faster using AIRES.

Table 8

Experiment II
November 1988
Mean Completion Time by Group, Subject, and Method
(minutes)

Group	Subject	AIRES	STANDARD
1	1	5.67	13.67
	2	3.67	5.00
	3	9.50 ·	12.67
	4	4.67	7.33
	5	7.00	24.00
	6	---	---
2	1	7.67	11.67
	2	5.33	12.00
	3	6.67	13.33
	4	5.33	6.33
	5	4.67	13.67
	6	9.00	10.00

CONCLUSIONS

This paper has demonstrated that uncontrolled sources of variability during experimentation, attributable to difference in experience, age, MOS, and Armed Forces, mask significant results and hinder decision makers from drawing the correct conclusions concerning the explosive ordnance disposal automated information retrieval and expert system. Model-based diagnostic procedures have been demonstrated to be effective in assessing the data and experimental model, and indicating probable causes for the violation of the model assumptions. Through these diagnostic procedures, the researcher was able to control additional sources of variability so that the data conformed to the standard assumption of compound symmetry. Thus, the researcher was able to conclude that there was a significant identification time difference between the automatic (AIRES) and manual (standard) method when soldiers with limited and the same MOS were evaluated. Soldiers in this group identified ordnance significantly faster using AIRES.

REFERENCES

Grynovicki, J. O. (1989). Variance component estimation diagnostics and associated distribution theory for all random and mixed designs: Application to repeated measures. (Doctoral dissertation, University of Delaware, 1989) Dissertation Information Service

Grynovicki, J. O., & Green, J. W. (1988). Estimation of variance components and model-based diagnostics in a repeated measures design. Proceedings of the Thirty-Third Conference on the Design of Experiments in Army Research Development and Testing, 65-89.

Grynovicki, J. O., Miller, R. F., & Krass, A. E. (1990). An evaluation of the ability of soldiers to quickly and accurately identify ordnance using the prototype explosive ordnance disposal automated information retrieval and expert system (EOD AIRES) (Technical Note 7-90). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Hocking, R. R. (1985). The analysis of linear models, Monterey, CA: Brooks-Cole Publishing Company.

Hocking, R. R., Green, J. W., & Bremer, R. H. (1989). Estimation of fixed effects and variance components in mixed factorial models including model-based diagnostics. Technometrics, 31 (2), 227-239.

MODELING GUNFIRE:
A PROBLEM IN MAXIMUM LIKELIHOOD ESTIMATION

Henry B. Tingey

Statement of the Problem.

$$- \leftarrow x \rightarrow +$$
$$M$$

$$d$$
$$\phi$$

$$-\frac{\pi}{2} \qquad\qquad\qquad\qquad\qquad\qquad \frac{\pi}{2}$$
$$0$$
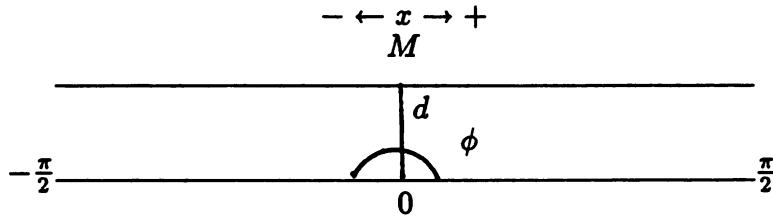
Figure 1. Schematic

A gun at 0 fires at a wall d units away. The angle of fire, ϕ, that the gun makes with OM is a random variable having the distribution function given by

$$F(\phi) = \begin{cases} 0 & \phi \leq -\pi/2 \\ \left(\frac{\phi+\pi/2}{\pi}\right)^{\alpha} & -\pi/2 < \phi < \pi/2 \\ 1 & \phi \geq \pi/2 \end{cases} \tag{1}$$

Suppose we can observe only the hit made on the wall. Let x be the distance measured from M to a hit. From a random sample $x_1, x_2, \ldots, x_n$ of such hits: i) Determine the maximum likelihood estimator of α, $\hat{\alpha}$, ii) Find the distribution of $\hat{\alpha}$, iii) When $\alpha = 1$ what is the distribution of x, the distance form M to a hit? iv) Set up a uniformly most powerful test for the hypothesis, $H : \alpha = \alpha_0$ with alternatives $\alpha < \alpha_0$ for an arbitrary significance level.

a) Given the sample of hits, $-1.306\,d$, $4.921\,d$, $2.865\,d$, $-6.512\,d$, $-3.984\,d$, -0.782 d, $3.542\,d$, $5.552\,d$, $-4.375\,d$, $-2.107\,d$. Test $H_0 : \alpha = 1$ versus $H_1 : \alpha < 1$ at the .05 level of significance.

b) Plot the power function of the test $H_0 : \alpha = 1$ against the alternative $\alpha < 1$ at the .05 significance level using the points $\alpha = 7/4$, $(1/4)$, $1/4$.

Determination of the Maximum Likelihood Estimator.

The principle of maximum likelihood estimation is: If $f_{X_1,\ldots,X_n}(X_1,\ldots,X_n;\theta)$ is the density for a random sample of size n drawn from a population with an unknown parameter θ, then the maximum likelihood of θ is the value, $\hat{\theta}$, if it exists, such that $f_{X_1,\ldots,X_n}(X_1,\ldots,X_n;\theta) > f_{X_1,\ldots,X_n}(X_1,\ldots,X_n;\theta')$ where θ' is any other value of θ, $\Pi f(x_i;\theta) = L(\theta;x_i,\ldots,x_n) = L(\theta)$.

269

The function, $\pi f(x_i, \theta)$, the likelihood function, is a function of θ when the x_i are known. When this is true the function is regarded as a likelihood function of θ and the maximum likelihood estimate of θ is therefore that value of θ for which maximizes $\pi f(x_i, \theta)$ as a function of θ.

The procedure is as follows: We are given the distribution function (cdf) of which the first derivative with respect to the random variable is the density.

Recall then, (1) and for simplicity let

$$z = \frac{\phi + \frac{\pi}{2}}{\pi}.$$

That is;

$$G(z) = \begin{cases} 0, & z \leq 0 \\ z^\alpha, & 0 < z < 1 \\ 1, & z \geq 1 \end{cases}. \tag{2}$$

Then

$$G'(z) = g(z, \alpha) = \alpha z^{\alpha - 1} \tag{3}$$

where $g(z, \alpha)$ is defined over $0 < z < 1$ and is zero otherwise.

Consider now the likelihood function,

$$L(\alpha) = \prod_{i=1}^{n} g(z_i, \alpha).$$

Following the customary procedure of finding the maximum of the logarithm of the likelihood function rather than of the likelihood function itself, maximize

$$L(\theta)^\star = \ln L(\theta) = \ln \prod_{i=1}^{n} g(z_i, \alpha) = n \ln \alpha + (\alpha - 1) \sum_{i=1}^{n} \ln z_i.$$

Then

$$L'(\theta)^\star = \frac{n}{\alpha} + \sum_{i=1}^{n} \ln z_i,$$

and setting $L'(\theta)^\star = 0$, we obtain,

$$\hat{\alpha} = -\frac{n}{\sum \log z_i}.$$

270

The Distribution of $\hat{\alpha}$.

The distribution of $\hat{\alpha}$ may be determined exactly by two different approaches and approximately in at least one way. The two methods of finding the exact distribution are, i) consider a change of variable in the density function and ii) by means of moment generating function. The approximate method consists of considering the normal approximation

i) Recall (3), let $y = -\log z$. Then

$$h(y)\,dy = \alpha e^{-\alpha y}\,dy \quad 0 < y < \infty. \tag{4}$$

It is easily verified that the integral of (4) is 1. Consider the transformation $\alpha y = -u/2$. Again by substitution into (4) yields,

$$q(u)\,du = \frac{1}{2}\,e^{-u/2}\,du$$

and

$$Q(V) = \int_0^V \frac{1}{2}\,e^{-u/2}\,du \tag{5}$$

describes the χ^2 distribution with two degrees of freedom.

Hence $\log z$ is distributed as χ^2 with two degrees of freedom. Therefore $1/\hat{\alpha}$ is distributed as χ^2 with $2n$ degrees of freedom since it is the sum of n independent χ_2^2 variables.[1]

ii) Recall:

$$\hat{\alpha} = -\frac{n}{\sum \ln z_i} \quad 0 < z < 1.$$

Let

$$-\frac{\sum \ln z_i}{n} = y \quad \hat{\alpha} = 1/y.$$

The moment generating function is defined by $m_Y(t) = E[e^{tY}]$ with t as a parameter, $t \neq \alpha n$. From the sample $z_1, \ldots, z_n$ of n independently and identically distributed variates we may make the following calculations, by substitution $E[e^{tY}]$

$$\{E\,[e^{-(t/n)\,\log z}]\}^n,$$

since the distribution of any one z describes the distribution of all the others. Hence the result here

$$= \{E[z^{-t/n}]\}^n.$$

---

[1] The exact distribution of $\hat{\alpha}$ is considered in Appendix I.

271

But

$$\{E[z^{-t/n}]\}^n = \left[\int_0^1 z^{-t/n} \alpha z^{\alpha-1} \, dz\right]^n = \left[\frac{\alpha z^{\alpha-(t/n)}}{\alpha - (t/n)} \Big|_0^1\right]^n .$$

Therefore

$$m_z(t) = \left[\frac{1}{1 - (t/\alpha n)}\right]^n , \qquad (6)$$

This is the moment generating function for a gamma distribution with parameters, $\beta = \frac{1}{\alpha n}$ $\alpha' = n - 1$.

By virtue of the uniqueness theorem: If two random variables have the same generating function, they have the same density function. thus, an attempt must be made to identify this density function with a known density. Otherwise, we go on to the next problem.

Consider $k(\omega) = \dfrac{1}{(n-1)! \left(\dfrac{1}{n\alpha}\right)^n} \omega^{n-1} e^{-\omega n \alpha}$ $\omega > 0$.

Let $\omega = \frac{1}{\hat{\alpha}}$ then $\ell\left(\frac{1}{\hat{\alpha}}\right) = \dfrac{n^n}{(n-1)!} \left(\dfrac{\alpha}{\hat{\alpha}}\right)^{n-1} e^{-n\alpha/\hat{\alpha}} \dfrac{\alpha}{\hat{\alpha}^2}$ $\hat{\alpha} > 0$.

Let $u = 2n\,\alpha/\hat{\alpha}$ $j(u) = \dfrac{1}{2^n(n-1)!} u^{n-1} e^{-u/2}$ $u > 0$. This is a χ_2^2 distribution, when $n = 1$.

iii) The approximate normal distribution of the estimator, $\hat{\alpha}$, for large samples is given by:

$$\Phi(\hat{\alpha}) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} e^{-\frac{(\hat{\alpha}-\alpha)^2}{2\tilde{\sigma}^2}}$$

where

$$\tilde{\sigma}^2 = \sigma_{\hat{\alpha}}^2 = \frac{1}{-nE\left[\frac{\partial^2}{\partial\alpha^2} \log g(z,\alpha)\right]} \cdot \log, g(z,\alpha) = \log\alpha + (\alpha - 1)\log z$$

$$\frac{d^2 \log g(z,\alpha)}{d\alpha^2} = -\frac{1}{\alpha^2}$$

Then

$$\tilde{\sigma}^2 = \frac{1}{-n\left(-\frac{1}{\alpha^2}\right)} = \frac{\alpha^2}{n}$$

$$\Phi(\hat{\alpha}) = \sqrt{\frac{n}{2\pi\alpha^2}} e^{-\frac{n}{2\alpha^2}(\hat{\alpha}-\alpha)^2} .$$

272

The Case When $\alpha = 1$.

Returning to the original density function,

$$f(\phi) = \left(\frac{\phi + \pi/2}{\pi}\right)^{\alpha-1} \quad -\pi/2 < \phi < \pi/2.$$

If $\alpha = 1$ it is then readily apparent that $f(\phi) = k$ and $k = \frac{1}{\pi}$ has the uniform density.

The distribution of x will be given by the expression

$$g(x) = h[f(x)]|h'[f(x)]|.$$

From the physical situation, let $x = d \tan \phi$.

$$h'(\phi) = \frac{1}{1 + x^2}.$$

Hence the distribution of x is given by

$$g(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} \qquad -\infty < x < \infty \tag{7}$$

or the familiar normalized Cauchy distribution.[2]

The Uniformly Most Powerful Test.

First the definition of a uniformly most powerful test. A uniformly most powerful test is defined as a test in which all alternatives give rise to the same critical region. That is:

$$f(x : \phi) > kf(x : \phi_0)$$

for any particular value of ϕ and all possible ϕ_0. This situation is, in general, not true. However, for the common distributions in statistics it is.[3]

A natural procedure, in the context of this paper is then to try the likelihood ratio test. For the situation here let $z_1, \ldots, z_n$ be a sample size of n from a population with density $g(z, \alpha)$. We wish to test the null hypothesis, $H_0 : g(z, \alpha)$ belongs to ω a subspace of Ω. The likelihood of the sample is:

$$L = \frac{1}{\pi} g(z_i, \alpha).$$

---

[2] Referred to in Box Hunter, Hunter Statistics for Researchers as a Mathematicians Toy, p. 69.

[3] Further discussion of the uniformly most powerful test is found in Appendix II.

Ordinarily the likelihood as a function of the parameters will have a maximum as the parameters are allowed to vary over the whole parameter space, Ω. The maximum in ω will occur in a like manner except here the parameters vary over a subspace of Ω. $L(\hat{\alpha})$, the maximum value, will be called $L(\hat{\Omega})$ in the whole space. $L(\alpha_0)$, the maximum value, will be called $L(\hat{\omega})$ in the subspace. The ratio of these two maxima thus form the likelihood ratio which is denoted by λ.

This fraction has the following obvious properties; it is positive since L is a product of positive density functions, $L(\hat{\omega})$ is always less than or equal to $L(\hat{\Omega})$ because of fewer degrees of freedom in ω. Also here it is a ratio of similar shaped monotonic functions that differ, at most, in location. Thus, the λ ratio is a monotonic function of sample observations only and has range zero to one.

Suppose for the distribution here we wish to test the null hypothesis

$$H_0 : \alpha = \alpha_0 \,.$$

This point is then ω while the whole α axis is Ω. The likelihood is

$$L = \prod_{i=1}^{n} \alpha z_i^{\alpha-1} \,.$$

The maximum value of the likelihood is Ω will be given by setting $\alpha = \hat{\alpha}$ to obtain

$$L(\hat{\Omega}) = \prod_{i=1}^{n} \hat{\alpha} z_i^{\hat{\alpha}-1}$$

and in ω

$$L(\hat{\omega}) = \prod_{i=1}^{n} \alpha_0 z_i^{\alpha_0-1} \,.$$

The likelihood ratio is

$$\lambda = \prod_{i=1}^{n} \alpha_0/\hat{\alpha} \; z_i^{\alpha_0-\hat{\alpha}} \,.$$

Taking the log of both sides we obtain

$$\frac{\log \lambda}{n} = \log \, \alpha_0/\hat{\alpha} - \alpha_0/\hat{\alpha} + 1 \qquad \text{where} \qquad \frac{\sum \log z_i}{n} = -\frac{1}{\hat{\alpha}} \,.$$

Consider the limiting cases for α_0 and $\hat{\alpha}$:[4]

---

[4] Here n will be fixed and have no real effect at the limit.

For $\alpha_0 \to 0 \qquad \lim_{\alpha_0 \to 0} \log \frac{\alpha_0}{\hat{\alpha}} - \frac{\alpha_0}{\hat{\alpha}} + 1 \to -\infty$.

For $\alpha_0 = \hat{\alpha} \qquad \log \frac{\alpha_0}{\hat{\alpha}} - \frac{\alpha_0}{\hat{\alpha}} + 1 = 0$.

For $\hat{\alpha} \to 0 \qquad \lim_{\hat{\alpha} \to 0} \log \frac{\alpha_0}{\hat{\alpha}} - \frac{\alpha_0}{\hat{\alpha}} + 1 \to -\infty$.[5]

Also it is readily apparent that $\log \lambda$ is monotone since all the values involved in the function are always positive or zero and monotone, as the *ratio of two* similar monotonic functions.

Using λ criterion as a test would yield a very clumsy procedure. An actual sampling distribution for λ would indeed be difficult to find. We then have two resources,

i) The large sample distribution of λ which states that under certain conditions, $-2 \log \lambda$ is distributed approximately as χ^2 with degrees of freedom equal to the dimensionality of Ω, say k, less the dimensionality of w, say r, if $r < k$, and H_0 is true.

ii) We may recall the result of the distribution of the reciprocal of the estimator, $\hat{\alpha}$, as determined via the moment generating function method.

Here we now see the quantity $2n \frac{\alpha_0}{\hat{\alpha}}$ which is distributed as χ^2 with $2n$ degrees of freedom is everywhere greater than the λ criterion and has the desirable properties of the χ^2 distribution. We need only specify an 'n' and a uniformly most powerful test will be determined. Further, $2n \frac{\alpha_0}{\hat{\alpha}}$ cannot have fewer than 2 degrees of freedom while the large sample λ has only 1, no matter what the sample size. The nature of the chi square distribution has been shown to give uniformly more power as the degrees of freedom increase.

Because the large sample approximation still has the inherent awkward calculation involved, we use the test criterion,

$$T = \frac{2n\,\alpha_0}{\hat{\alpha}} \overset{d}{=} \chi^2_{2n} \quad \text{under} \quad H_0 \tag{8}$$

a) For the sample given earlier the hypothesis is tested by

Test procedure:
  i) $H_0 : \alpha = 1$
  ii) $H_A : \alpha < 1$
  iii)   Significance level $= .05$
  iv) $T = 2n \frac{\alpha_0}{\hat{\alpha}} \chi^2$ with $2n$ degrees of freedom under H_0
  v)   Here $n = 10$ so we have 20 degrees of freedom
  vi)   Test Rule: Reject H_0 if $T > \chi^2_{20}\,.05 = 31.4$
  vii)   Calculation of T

---

[5] This result can be easily verified from elementary calculus.

275

Recall

$$\hat{\alpha} = -n / \sum \log \left(\frac{\phi + \pi/2}{\pi} \right)$$

and $x/d = \tan \phi$ $\phi = \arctan x/d$

For our sample

x	ϕ (nearest minute)[6]	Radians	$\left(\frac{\phi}{\pi} + \frac{1}{2} \right)$	$\log \left(\frac{\phi}{\pi} + \frac{1}{2} \right)$
$-1.306d$	$-52°34'$	-0.91746	.208	-1.57022
$4.921d$	$78°31'$	1.37037	.936	-0.06614
$2.865d$	$70°46'$	1.23512	.893	-0.11317
$-6.512d$	$-81°16'$	-1.41838	.049	-3.01593
$-3.984d$	$-75°55'$	-1.32499	.078	-2.55105
$-0.782d$	$-38°02'$	-0.66381	.289	-1.24133
$3.542d$	$74°14'$	1.29561	.412	-0.09212
$5.552d$	$79°48'$	1.39218	.943	-0.05869
$-4.375d$	$-77°08'$	-1.34673	.071	-2.64508
$-2.071d$	$-64°13'$	-1.12079	.143	-1.94491

Then $\sum \log \left(\frac{\phi + \pi/2}{\pi} \right) = -13.29864$ and

$$\hat{\alpha} = \frac{10}{13.29684} = .75196 .$$

So for T we obtain

$$T = \frac{2(10)(1)}{.75196} = 26.60 .$$

Since $T = 26.6 < \chi^2_{20}.95 = 31.4$ we do not reject H_0.

The Power Function.

For the case here, when we are testing the hypothesis, $H_0 : \alpha = \alpha_0$ versus all alternatives $H_A : \alpha < \alpha_0$, the power function will be given by

$$P(\alpha) \int_A^\infty f(x) \, dx ,$$

where $f(x)$ is the appropriate χ^2 density. This will be a function of α. Then the true α is far to the left of α_0 the power will be near 1. As α approaches α_0 the power approaches the significance level. As α goes to the right the power decreases to zero for this alternative.

276

For purposes of graphing the power function, consider the following argument.

$$\text{Power} = Pr\left\{T > \chi^2_{20}(.95)\right\}$$

where T is our test statistic and $\chi^2_{20}(.95)$ is the
critical value of the χ^2 distribution or

$$Pr\left\{\frac{2n\,\alpha_0}{\hat{\alpha}} > \chi^2_{20}(.95)\right\} = 1 - \beta.$$

Call the true value of α, α_T and multiply both sides of the inequality by $\frac{\alpha_T}{\alpha_0}$ obtaining:

$$Pr\left\{\frac{2n\,\alpha_T}{\hat{\alpha}} > \frac{\alpha_T}{\alpha_0}\chi^2_{20}(.95)\right\} = 1 - \beta.$$
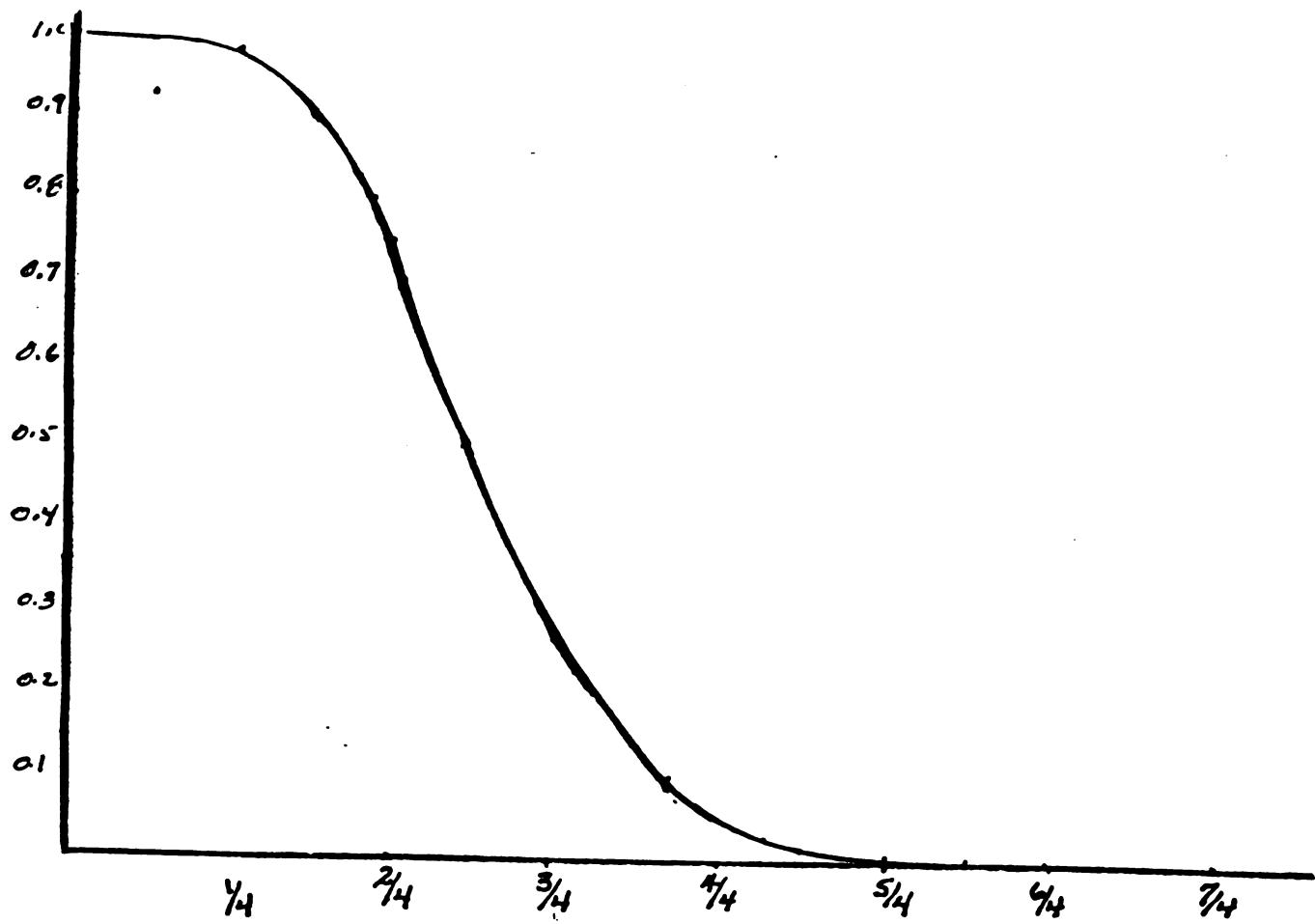
Then, since $\frac{2n\,\alpha_T}{\hat{\alpha}}$ is an exact χ^2 variate we have

$$Pr\left\{\chi^2_{2n} > \frac{\alpha_T}{\alpha_0}\chi^2_{20}(.95)\right\} = \text{Power}$$

α_T	$\frac{\alpha_T}{\alpha_0}\chi^2_{20}(.95)$	$Pr\left\{\chi^2_{2n} > \frac{\alpha_T}{\alpha_0}\chi^2_{20}(.95)\right\}$
7/4	54.95	0.0
6/4	47.10	.0005
5/4	39.25	.0033
4/4	31.40	.05
3/4	23.55	.27
2/4	15.70	.74
1/4	7.85	.9925

277

Power Function for

$$T = \frac{2n\, a_0}{\hat{a}}$$

$$\frac{a\, T}{a_2}$$

This is the uniformly most powerful test for a given alternative. Also it is more powerful that the λ criterion as will be shown later. As n increases the test becomes more sensitive. As n decreases so does the power of the test.

As a final note to this paper it may be of interest to repeat the above test procedure for the large sample distribution of the λ criterion.

Under this procedure $-2 \log \lambda \; (\overset{d}{=}) \; \chi_1^2 \quad$ under $\quad H_0$.

Reject here if $-2 \log \lambda > \chi_1^2(.95) = 3.84$.

Calculation:

$$-2 \log = -40 \left(\log \frac{1}{.75196} - \frac{1}{.76196} + 1 \right)$$
$$= -40(.28518 - 1.33 + 1)$$
$$= 1.7928$$

Therefore do not reject H_0 since $1.7928 < 3.84$.

A look at the power of the test in comparison to the above power function may take the following argument.

In order to satisfy the condition of a uniformly most powerful test for the test criterion $2n \frac{\alpha_0}{\hat{\alpha}}$ consider the following:

We need to show for

$$\frac{\alpha_0}{\hat{\alpha}} > A$$

that $T\chi_{2n}^2$ as the test statistic for the exact sampling distribution of $\frac{2n \, \alpha_0}{\hat{\alpha}}$ and T_λ as the statistic for the appropriate distribution of $-2 \log \lambda$ have the relationship $T_\lambda > T_{\chi_{2n}^2}$ or where

$$\text{for} \quad T_{\chi_n^2} : \; \frac{\alpha_0}{\hat{\alpha}} > A$$
$$\text{and} \quad T_\lambda : \; \frac{\alpha_0}{\hat{\alpha}} > A'$$

It is sufficient to show that $A' > A$ to conclude that $T_{\chi_{2n}^2}$ is a uniformly more powerful criterion than T_λ. It will be necessary to establish the inequality $A' > A$ for every A and A'.

Consider first the critical values and equate the respective test statistics to them

$$T_{\chi_{2n}^2} = \frac{2n \, \alpha_0}{\hat{\alpha}} = 31.4$$

here $\hat{\alpha} \leq .6394$ is sufficient for rejection and $\frac{\alpha_0}{\hat{\alpha}} = 1.5700$

$$T_\lambda = -2n \left(\log \frac{\alpha_0}{\hat{\alpha}} - \frac{\alpha_0}{\hat{\alpha}} + 1 \right) = 3.84$$

here $\hat{\alpha} \leq .5700$ for rejection and $\frac{\alpha_0}{\hat{\alpha}} = 1.754$[7]

Then $A' > A$ and clearly since the chi square distribution is monotonic and also has a greater rate of change as the degrees of freedom are increased the inequality will therefore hold for every A' and A it follows that $T_{\chi^2_{2n}}$ is the uniformly most powerful test criterion.

---

[7] The solution here is iterative. Usually 4 or 5 tries to give accuracy to the 3rd decimal place.

Appendix I

The Exact Distribution of $\hat{\alpha}$.

After performing the substitution in Part II, it was found that:

$$\phi(y)\, dy = \alpha e^{\alpha y}\, dy$$

which defines the distribution of the reciprocal of the estimator, $\hat{\alpha}$.

The distribution of $\hat{\alpha}$ then is actually the inverse of a chi square distribution with two degrees of freedom.

It is possible to find the form of the distribution of $\hat{\alpha}$ then by substitution into the above formula.

Let $x = 1/y$

then $dx = -\frac{1}{y^2}\, dy$

thus $y = \frac{1}{x}$ is the inverse of the above transformation and it follows that $h(x)\, dx = x^{-2}\, dx$. From the relationship

$$g(x) = f[h(x)]\, |h'(x)|$$

we obtain

$$\alpha x^{-2}\, e^{-\alpha/x}\, dx\,.$$

If we let

$$x = \hat{\alpha}$$
$$dx = d\hat{\alpha}$$

then

$$\frac{\alpha}{\hat{\alpha}^2}\, e^{\alpha/\hat{\alpha}}\, d\hat{\alpha}$$

describes the distribution of the maximum likelihood estimator, $\hat{\alpha}$.

Here the distribution of $\hat{\alpha}$ is not readily recognizable. It may be possible to identify this distribution via Pearson's Principle of Moments. However, if this were tried, one is still faced with getting to a convenient form of the distribution. Since it has already been shown that the distribution of $1/\hat{\alpha}$ is readily recognizable and has the desirable properties of a "good" test, it seems futile to pursue this approach any further.

Appendix II

Mathematical Proof of a Uniformly Most Powerful Test.

For a particular value of α' of α the best test of α_0 against α' is given by choosing as a critical region the set of points for which

$$f(z; \alpha') > kf(z; \alpha_0)$$

or for this particular distribution

$$\alpha' z^{\alpha'-1} > k\alpha_0 z^{\alpha_0-1}$$

$$z^{\alpha'-\alpha_0} > k \frac{\alpha_0}{\alpha'}$$

$$\log z > \frac{\log k + \log \frac{\alpha_0}{\alpha'}}{\alpha' - \alpha_0}$$

Hence we have as the best critical region an interval $\log z > A$. Since it is known that $\log z$ is distributed as chi square with two degrees of freedom, A is to be chosen so that

$$\int_A^\infty f(\chi^2_{df}\ d\chi^2 = P\ \{\text{Type I error}\}.$$

The value A is the abscissa for χ^2_2 distribution such that if $f(x)$ is the χ^2_2 distribution then

$$\int_A^\infty f(x)\ dx = P\ \{\text{Type I error}\}.$$

For example: For

$$P\ \{\text{type I error}\} = .05$$

$$A = 5.99$$

This derivation is on the basis of a single observation. The generalization to samples of size n is immediate. The observation $(z_i, \ldots, z_n)$ may be plotted in n dimensional space which may be divided into two regions, rejection and acceptance. The ratio of the products of the two density functions will yield a critical region which will be defined by a chi square value with $2n$ degrees of freedom at the level of the Type I error.

Bibliography

Box, George, G. P., Hunter, W. G., Hunter, B. S., *Statistics for Experimenters*, (New York: John Wiley & Sons, 1978).

Dixon, Wilfred J. and Massey, Frank, Jr., *Introduction to Statistical Analysis*, 4th Ed. (New York: McGraw–Hill Book Company, 1983).

Granville, William Anthony, Smith, Percy F. and Langly, William Raymond, *Elements of Calculus*, (Gern and Company, 1946).

Hoel, Paul G., *Introduction to Mathematical Statistics*, 5th Ed., (New York: John Wiley and Sons, Inc., 1984).

Mood, Alexander McFarlane, Graybill, F. A., and Boes, D. C., *Introduction to the Theory of Statistics*, 3rd Ed., (New York: McGraw–Hill Book Company, 1974).

Walker, Helen M. and Lev, Joseph, *Statistical Inference*, Dewey Holt and Co., 1953.

RESPONSE SURFACE METHODOLOGY FOR VALUE ADDED ANALYSIS

Captain William F. Mann III and LTC Andrew G. Loerch
Force Systems Directorate - Resources Analysis
U.S. Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, Maryland 201814-2797

ABSTRACT. The Value Added Analysis (VAA) study at the U.S.Army Concepts Analysis Agency will develop the capability to perform cost-benefit analysis of major item systems in support of the Army Planning, Programming, Budgeting, and Execution System (PPBES). A sub-problem of the study is the analysis of the contribution of the systems under consideration to different measures of effectiveness in a combat model. Modified Plackett-Burman designs are used to create general linear formulas of the form:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_rx_r .$$

While linear models are usually used for fitting a regression line along a range of variables, they can also be used with qualitative variables. The estimator b_0 is a collective measure of the worth of the base case weapons. The estimators b_1 through b_r represent the contribution of individual systems to the improvement of the Corps-level measures of effectiveness (MOEs); e.g., Loss Exchange Ratio, Fractional Exchange Ratio, Effective Battalions Remaining, etc.

There are several advantages of this method over previous practices. First a weapon's "additive" effect can be estimated. This "additive" effect can be used to determine the "benefit" of buying this system. Secondly, the general linear model can be used to evaluate different force packages or mixes without rerunning the combat model.

1. INTRODUCTION. The Value Added Analysis (VAA) methodology was developed by the U.S. Army Concepts Analysis Agency to provide the Army Staff with a rapid response analysis framework for performing cost-benefit analysis to compare competing investment alternatives during development of the Army Program. An important aspect of this methodology is the evaluation of major item systems performance using a combat simulation.

In the present environment of shrinking budgets and no dominant theater, program development has become increasingly difficult. Many competing systems must be evaluated in several theaters across several years. The number of simulation runs is potentially huge. The task, then,

is to keep the number of runs at a minimum while extracting as much information as possible from each one.

The purpose of this paper is to describe how the technique of response surface methodology (RSM) and the general linear model (GLM) can be used to accomplish this task. The use of these two statistical techniques allows the systematic development of the set of required simulations and its inputs, and facilitates the analysis of the output as well.

2. COMBAT MODELING. The Corps Battle Analyzer (CORBAN) is a complex combat simulation in which combat functions are represented mathematically. VAA requires the evaluation and quantification of certain combat and non-combat systems. It is also necessary to represent the systems in several different theaters and across a wide range of years. To get a broad spectrum of possible situations, the systems are evaluated in three different and plausible scenarios for three different years (1996, 2001, and 2008). For each of these nine scenario/year combinations, VAA requires approximately 50 CORBAN runs. We are interested in several measures of effectiveness (MOEs) derived from the model. Some examples of these MOEs are:

a. Loss Exchange Ratio (LER) - This MOE is a ratio of the number of major Red systems lost divided by the number of Blue systems lost. LER is computed as follows:

$$LER = \frac{Red\ Systems\ Lost}{Blue\ Systems\ Lost}$$

b. Fractional Exchange Ratio (FER) - This MOE measures the fractional red losses compared to fractional blue forces. FER is computed as follows:

$$FER = \frac{Red\ Systems\ Losses / Red\ Systems\ Started}{Blue\ Sytems\ Losses / Blue\ Systems\ Started}$$

c. Red and Blue Effective Battalions Remaining (EBR) - This MOE is a measure of the number of battalions (generally maneuver, artillery, rocket and helicopter) remaining on each side that are still combat effective at the conclusion of the simulated conflict.

d. Red and Blue Movement of Force Center of Mass (MFCM) - This MOE measures of the performance of an attacker by examining the distance the center of mass of his forces has travelled.

The inputs to CORBAN include scenario, terrain, representative Red forces and Blue forces, missions, and orders. The Base Case asset list

contains current systems that will be available during a particular year of interest. Excursions are developed by substituting base case systems with new weapon systems, or adding new systems, to produce new results that can be compared with the base case results.

3. THE STATISTICAL MODEL. If one imagines that the CORBAN model is a black box (see Figure 1), the use of response surface methodology becomes easier to understand. "Response surface methodology comprises a group of statistical techniques for empirical model building and model exploitation. By careful design and analysis of experiments, it seeks to relate a 'response', or output variable to the levels of a number of 'predictors', or input variables, that affect it" (Box and Draper, 1987; 1).
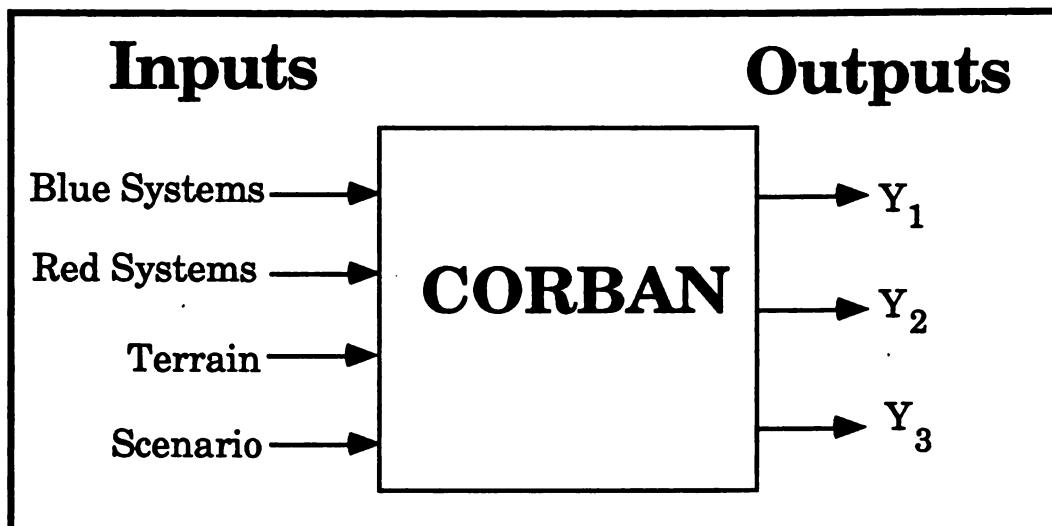

Figure 1. CORBAN with All the Inputs

The most common and simplest method of using a combat model is to establish a base line case, and then to add each new weapon system one at a time, measuring the changes in combat effectiveness. These changes from the base line case measure the amount a weapon system contributes to the outcome of the battle. While this method measures the contribution of each individual weapon, it **does not** allow the determination of the additive effect of weapon systems, i.e., if an attack helicopter raises the value of an MOE by "x" and a tank raises the value by "y", then it is **not** true that if both systems are present the resulting improvement would be "x+y".

The ideal solution would be to explore all possible combinations and find the combination of systems that yield the greatest increase in the MOEs values. While this method is practical with small situations, the number of combinations grows quickly. If one had to explore every combination of 40 different systems, the number of potential runs would be 2^{40}, or 109.9 billion runs.

287

RSM represents a compromise between the process of replacing weapons one at a time and the ideal solution of doing every combination. This compromise is a fractional design. This technique is based on taking specific combinations of the total combinations possible. The combinations' results are then averaged to find an estimate for a system's contribution to the MOE.

RSM **does** allow the measurement of the additive effects of weapons. This methodology fosters the construction of a "design matrix" that varies the inputs in an efficient manner so that a linear model can be built to forecast the effects of the systems with respect to the outputs. A set of coefficients is computed which are the mean or average improvement given the new weapon system. These coefficients can be used in an additive estimate.

In the following example, we hold the Red weapon systems, the terrain, and the scenario input variables constant and vary the Blue systems that we want to study. This technique reduces the inputs to the black box to a more manageable size as illustrated in Figure 2.
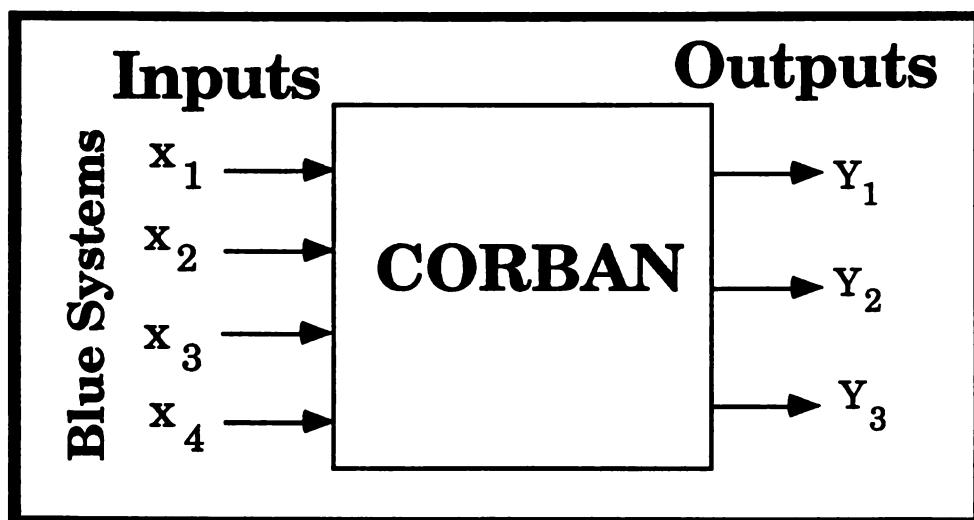

Figure 2. Controlling only the Desired Input

A general linear model is a method of estimating an output or dependent variable, Y, whose mean is a function of one or more independent variables (x_1, x_2, etc.). The general linear model is of the form:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_r x_r .$$

(1)

or in matrix form,

$$Y = XB,$$

288

where Y is an output MOE vector of the model run results, b_0 is the effect of the base case weapons, b_1 through b_r are the effects of systems 1 through r in the excursions, and X is the design matrix of binary independent variables whose construction is described below. This model isolates only the presence of main effects without considering any interaction effects.

Plackett-Burman designs are useful when the problem of determining the main effects with maximum precision is reduced to a combinatorial problem. They are useful when the problem has only two-level factors, i.e., when there are low and high variable setting or binary (0, 1) variables. If there is a need to consider a mixture of two and three factor treatments, orthogonal array designs may be employed.

Plackett-Burman's method specifies the construction of the design matrix, X. This matrix represents a map of all the independent variables' values for each computer run. Each row corresponds to a specific computer run and each column corresponds to a different factor. In the case of VAA, the systems being considered for procurement are the factors. The values of the matrix elements are either 1 or 0 which represents the presence or absence of the system, respectively.

To illustrate this coding scheme, we consider the following two cases. The first case involves a new system replacing an existing system. An example of this case would be the Squad Automatic Weapon (SAW) replacing the M60 machine gun. In excursions where soldiers are equipped with a SAW, a 1 would be entered in the design matrix. In excursions where the M60 is used, a 0 would appear.

The second case involves a new system that does not replace an existing system. JSTARS would be such a system. In this case, a 1 would indicate the presence of the new system, while a 0 would denote its absence.

Once the design matrix is formed, each excursion is performed as specified using the combat model, and the outputs are obtained, forming the Y vector. The coefficients (elements of B) for the linear model are obtained by matrix algebra. Then we have:

$$XB = Y. \tag{2}$$

To solve for B, we have two options. If a full Plackett-Burman matrix is used without deleting any columns such that X is of full rank, then:

$$B = X^{-1}Y. \tag{3}$$

If there is a need or desire to use a reduced matrix (modified Plackett-Burman design), then we use the standard formula for GLM:

$$B = (X^TX)^{-1}X^TY \tag{4}$$

289

(Neter, Wasserman, and Kutner, 1985; 239).

The above matrix manipulation can be done on spreadsheets, IMSL routines, FORTRAN programs, SPSS, BMDP, or other statistic programs.

While the GLM is normally used for fitting a regression line along a range of variables, it can be used with qualitative variables as well (Neter, Wasserman, and Kutner, 1985; 342). The estimator b_0 is the measure of worth of the base case weapons. The estimators b_1 through b_r represent the contribution of individual systems to the improvement of the Corps-level MOEs; e.g., LER, FER, EBR, etc. "These coefficients are often called **differential intercept coefficients** because they reveal by how much the value of the intercept term of the category that receives the value 1 differs from the intercept coefficient of the category that receives the value zero. The category that receives the value zero is often referred to as the **base** or **comparison** category" (Dillon and Goldstein, 1984; 245).

The advantage of the above formulation is that the general linear model can be used to evaluate different force packages or mixes of systems without rerunning the CORBAN model. The linear model can then be used to give an estimate of the values of the MOEs Y_1, Y_2, and Y_3 for the new combinations of systems without conducting additional simulation runs.

The number of computer runs needed to determine the linear model depends on the type of combat model that is being used. A stochastic combat model will give different answers for each random number seed used, all other inputs being held constant; therefore, multiple runs must be made for each combination. Deterministic (expected value) models are not random. Therefore, only one run of this model is needed to measure the output for each combination. VAA uses the deterministic version of CORBAN, so one more run than the total number of system contributions, r, to be estimated, or at least r+1 runs are needed using the Plackett-Burman design or a subset of Plackett-Burman (Plackett and Burman, 1946; 319). The additional run is required to determine the intercept, b_0 for the base case value. So with r+1 runs, coefficients can be computed for equation (1) above. It should be noted that each Plackett-Burman design is unique and based on multiples of four runs. However, one can use subsets of larger designs to tailor the design matrix. In these cases, one must use equation (4) to find the coefficients.

This general linear model will result in a perfect fit (no residuals) of the output data when we have exactly r+1 model runs for r weapon systems coefficients and one coefficient for the base case or intercept term. Therefore, we will have r+1 data points for r+1 estimators (b_0 to b_r). Statistically, we can not estimate the lack of fit due to the lack of residual terms. If desired, additional runs could be made on non-explored points or a subset of a larger Plackett-Burman matrix can be used and the general

linear model's aptness can be determined. This is necessary if statistical tests are needed to determine the significance of the coefficients.

An additional advantage of the use of this method, is the ability to add a small number of additional weapon systems later. The past work is not for naught, additional computer runs can be added, and the problem can grow from the initial set of runs (see example 2), into a more traditional multiple regression problem.

4. ILLUSTRATIVE EXAMPLES. Example 1 (See the appendix for additional computational details) demonstrates the proposed methodology. In this example there are 7 systems (r=7) that we want to compare to the base case of which four are new direct fire weapons (weapons are 1-4), two field artillery weapons (FA weapons 5 and 6), and one new attack helicopter (helicopter 7). Using the Plackett-Burman design, we set up the design matrix, X. The 8 (r+1) computer runs are performed and, in this case, the output of Blue Effective Battalions Remaining (EBRs) MOE is used, forming our output vector, Y. Using equation (3) our linear model becomes:

$$Y_2 = 3 + 18.5x_1 + 18.5x_2 - 3x_3 + 8x_4 - 14x_5 + 18.5x_6 + 30x_7 \qquad (5)$$

where $b_0 = 3$, the value of the base case systems, and

$$x_i = \begin{cases} 1, & \text{if weapon i is present,} \\ 0, & \text{otherwise.} \end{cases}$$

With the above general linear model, one can see that the largest improvement is incurred when the Attack Helicopter is introduced, followed by a three way tie for the next largest contribution among systems 1, 2, and 6. The next important weapon is 4. Finally note that systems 3 and 2 actually decrease the ability to obtain a higher EBR score.

Once the study is done, the sponsors wish to add another system. As shown by Example 2, this can be done. The disadvantage of adding a new system is that there is a loss of the structure of the original orthogonal design. This orthogonal design gave us advantages of balance and coverage of the sample space. The same basic methodology can still be used since the problem is still a multiple regression problem.

5. PROBLEMS USING THIS TECHNIQUE. This procedure is not a cure for input errors and intense weapon system interactions, nor does it immediately help to explain counter-intuitive results. Thorough analysis is still required to check the answers, and then serious thought is needed to interpret the results.

When this methodology was first used in the VAA study, the discovery of input errors was more difficult. When the traditional method of combat

291

modeling is used, only one system is adjusted at a time from the base case. The modelers can then easily compare the results from their excursions to the base case and determine the differences and find any errors. It is much harder to find modeling errors in an experimental design when there are many different new systems in each run. Once major errors are discovered, it may then be required to rerun all the runs to correct the error.

A major concern arose when some system coefficients were negative. Intuition tells the analyst that the replacement of an old system with a vastly superior system should not result in a decrease in capability. An examination of the systems then showed unnoticed modeling data problems with all the negative systems. Examples are search patterns larger than the arrays set up to store them, firing rates set to zero, or an incorrect portrayal of the weapon. The methodology has not yet, in its use in VAA, produced a result that was worse for a clearly superior system when the input data for the respective new and old systems were correct.

To assist in error checking, other summary or intermediate statistics were used to find problems. An example is the number of losses for each side. For instance, if the Fractional Exchange Ratio for a system is negative, the major components of FER, the number of Red losses and the number of Blue losses are examined. The regression was applied to these statistics and interesting discoveries could be made. The presence of some weapons was linked to Blue experiencing more losses during these runs. Calculations showed that it was the increase of Blue losses that made the system's FER turn negative. This technique helped lead to discovering the reason why there were more Blue losses.

In some cases, an interaction between two systems occurs. This interaction can be a synergism between two systems that greatly increases their ability. A pairing of a long range sensor and a long range missile is an example of a possible synergistic interaction. The interaction could also decrease the systems' MOE contribution, such as two long range weapons that when together compete for the same targets. A solution to this problem is called a "foldover" design that isolates the main effects of the interested systems (Box and Draper, 1987; 158 and 162). The cost of this foldover design is to double the number of computer runs. Other designs to specifically measure interactions are available, but their cost in the number of computer runs are very prohibitive except in problems with only a limited number of independent variables.

Interpretation of combat results has never been easy. Time and thought still need to be invested to determine why system "A" is doing so poorly, or doing so well, and then to present the results to the decision makers. Decision makers need to understand the reasons for an answer, and RSM may make it harder for the modeler to explain the results. In this particular situation, a database system is invaluable to manipulate the

large amount of data needed to perform these analyses. Explaining and verifying counter-intuitive results is a difficult and tedious process.

6. SUMMARY. The Plackett-Burman designs are ideally suited for use in the VAA methodology. The requirement to evaluate multiple systems in various combinations would be impossible in a timely manner if combat simulations would be needed for every conceivable grouping of weapon systems. Thus, the use of RSM to form general linear models will allow VAA to be a responsive tool needed in the PPBES process. While there are many benefits to this statistical approach, this method is not a panacea for combat modeling. There is still a great requirement for thorough and detailed analysis to be done.

BIBLIOGRAPHY

Box, G. and Draper, N. *Empirical Model-Building and Response Surfaces.* John Wiley, New York, 1987.

Dillon, William R. and Goldstein, Matthew. *Multivariate Analysis.* John Wiley, New York, 1984.

Neter, J., Wasserman W., and Kutner, M. *Applied Linear Statistical Models.* Irwin, Homewood, IL. 1985.

Plackett, R. L. and Burman, J. P. "The Design of Optimal Multifactorial Experiments." *Biometrika*, 33, 305-325 and 328-332.

APPENDIX

Example 1

This example is done using a software package called MATHCAD, which allows one to use matrix algebra. This example is for demonstration purposes only and does not represent actual data.

ORIGIN ≡ 1   PROPOSAL FOR DETERMINING COEFFICIENTS

$$
DESIGN\_MATRIX := \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
$$

BASED ON THE
PLACKETT-BURMAN
DESIGN

X := DESIGN_MATRIX

	Run #
$$Y := \begin{bmatrix} 23 \\ 45 \\ 24 \\ 34 \\ 56 \\ 67 \\ 78 \\ 3 \end{bmatrix}$$	1 2 3 4 5 6 7 8

Representative output from the CORBAN model
i.e. Effective Battalions Remaining

$$B := (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

$$
B = \begin{bmatrix} 3 \\ 18.5 \\ 18.5 \\ -3 \\ 8 \\ -14 \\ 18.5 \\ 30 \end{bmatrix}
$$

MEAN Value of the Base Case
WPN 1
WPN 2
WPN 3
WPN 4
FA 1
FA 2
ATK HELO

Check of the fit of the model

Y1 := X·B

$\Sigma(Y - Y1) = 0$

$$
Y - Y1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

294

Example 2

In this example, we have completed our computer runs, and
now have a new weapon system to add in. While we will lose
some of the benefit of the original orthogonal design,
we can still treat this as a multiple regression problem.

In this example, we add an additional column, and repeated
the first three rows of the design and with the presence of
the new weapon system.

ORIGIN ≡ 1

<!-- # of Runs -->

$$
\text{DESIGN\_MATRIX} := \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}
$$

# of Runs	
1	
2	BASED ON THE
3	ORIGINAL
4	PLACKETT-BURMAN
5	DESIGN WITH AN
6	ADDITIONAL COLUMN
7	AND THREE MORE ROWS
8	
9	
10	
11	

X := DESIGN_MATRIX

$$
Y := \begin{bmatrix} 23 \\ 45 \\ 24 \\ 34 \\ 56 \\ 67 \\ 78 \\ 3 \\ 33 \\ 55 \\ 33 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix}
$$

The output matrix with the three additional
results from the three additional computer runs.

$$
B := (X^T \cdot X)^{-1} \cdot X^T \cdot Y
$$

$$
B = \begin{bmatrix} 3 \\ 18.583 \\ 18.667 \\ -3 \\ 7.917 \\ -14.083 \\ 18.583 \\ 29.833 \\ 9.667 \end{bmatrix}
$$

MEAN of the Base Case Weapons	
WPN 1	
WPN 2	
WPN 3	This is our new coefficient matrix.
WPN 4	Notice the small changes in the
FA 1	coefficients when compared to
FA 2	Example 1.
ATK HELO	
NEW WPN	

295

We will now check the results of the model. Realize now that we have more points than coefficients, so there may be some variance in the model.

$$X \cdot B = \begin{bmatrix} 23.167 \\ 45.167 \\ 23.667 \\ 34 \\ 56 \\ 67 \\ 78 \\ 3 \\ 32.833 \\ 54.833 \\ 33.333 \end{bmatrix}$$

$E := Y - (X \cdot B)$

The variance is represented by the error of the predicted results, Y1, to the actual results, Y.
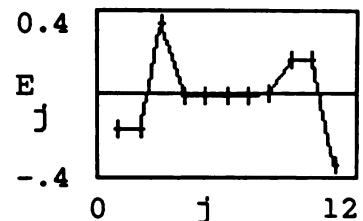
Check of the fit of the model

Y1 := X · B

$$Y - Y1 = \begin{bmatrix} -0.167 \\ -0.167 \\ 0.333 \\ -14 \\ -3.553 \cdot 10 \\ -14 \\ -8.527 \cdot 10 \\ -14 \\ -5.684 \cdot 10 \\ -14 \\ -5.684 \cdot 10 \\ -14 \\ 6.306 \cdot 10 \\ 0.167 \\ 0.167 \\ -0.333 \end{bmatrix}$$

j := 1 ..rows(X)

This graph shows the distribution of the error, Y1 - Y.

$\Sigma(Y - Y1) = -5.871 \cdot 10^{-13}$

In this example the error term is not exactly zero due to round off error.

APPLIED TIME SERIES ANALYSIS

H. Joseph Newton

Department of Statistics
Texas A&M University

37TH CONFERENCE ON DESIGN OF
EXPERIMENTS IN ARMY RESEARCH

Vicksburg, Mississippi, October 21–22, 1991

LECTURE 1

Descriptive Statistics and Tests For White Noise

Almost every area of scientific inquiry is concerned with data collected over time, that is, with time series. Figure 1.1 contains the graph of ten such time series. The aim of this lecture is to introduce some of the basic descriptive techniques used in time series analysis.

1.1. Time Series Data Types

Time series can be classified in many ways, including by the following four characteristics:

1. The dimension of the index set T over which observations are made. The set T can be one-dimensional or multidimensional. Daily births of females in California (Series I) have a one-dimensional index set, while wheat yields recorded over a regular grid of positions in a large field have a two-dimensional index set. This second example also illustrates that the index set need not literally be "time" but can also be "position."

2. Whether the index set is continuous or discrete. Thus an analog EEG record from one probe is a continuous one-dimensional series, while daily birth data are a discrete one-dimensional series.

3. How many variables are recorded at each element of the index set. Univariate (also called scalar) time series have one variable at each time, while multivariate time series have a vector of variables measured at each time. For example, a series consisting of monthly interest rates and gross national product is a one-dimensional bivariate time series.

4. Whether the variable or variables at each time are themselves discrete or continuous. Most series are continuous; that is, each variable can take on a continuum of values. An example of a discrete valued series is a binary time series, one in which an observation can take on only one of two values.

Note that discrete-time time series can also be either equally spaced or irregularly spaced in time. We will be almost exclusively concerned with time series that are one dimensional, equally spaced, discrete in time but continuous in space, and univariate.

1.2. Time Series Memory Types

The basic property of time series analysis is that it is concerned with repeated measurements on the same phenomenon at different times or places. Because of this, the analyst must take into account the correlation between successive observations. This is in marked contrast with the data analyzed in many areas of statistics where one assumes that the data are made up of independent and identically distributed observations obtained by randomly sampling some population or populations. The presence of correlation makes the analysis of time series data and the interpretation of the results much more difficult than in the independent case.
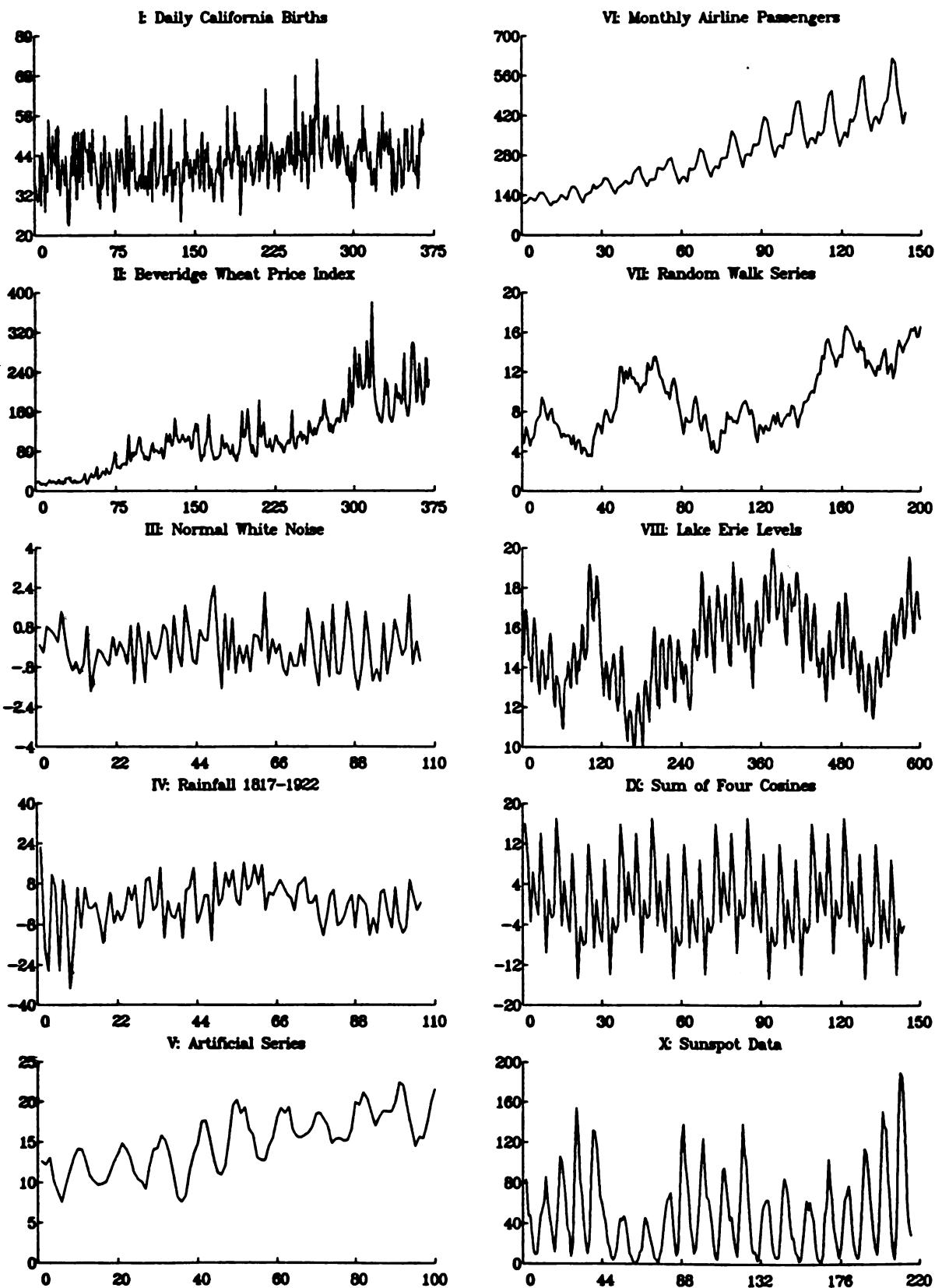
Course Materials

This handout is based on the book *TIMESLAB: A Time Series Analysis Laboratory* written by H. Joseph Newton and published by Wadsworth and Brooks/Cole of Pacific Grove, California 93950; (800) 354-9706. For more information, contact the publisher or H. Joseph Newton; Department of Statistics; Texas A&M University; College Station, Texas 77843; (409) 845-3141; email: jnewton@stat.tamu.edu.

Outline of Lectures

This tutorial is divided into six lectures. The aim of the lectures is to introduce the basic ideas and methods for analyzing time series from a statistical point of view. They will make extensive use of the computer program—also called TIMESLAB—that accompanies the speaker's book. The lectures are:

1. *Descriptive statistics and tests for white noise.* Numerical and graphical summaries of time series are discussed and illustrated on ten typical time series. Two tests are described for determining if a time series can be regarded as having no patterns (white noise).

2. *Transforming and forecasting time series.* As in regression analysis, it is sometimes necessary to transform a time series before proceeding with the analysis. Several transformations are described. Also, several simple forecasting methods are described. These methods do not use time series models.

3. *Time series models.* The basic properties of standard time series models (ARMA, ARIMA, etc.) are described. The basic theory of covariance stationary time series and their prediction are discussed.

4. *Estimation and model identification.* Estimation procedures and their properties are described for the mean and autocorrelations of a time series as well as for the parameters of the models introduced in Lecture 3.

5. *Model-based forecasting methods.* The models and estimates discussed in Lecture 4 are used to find forecasts and forecast intervals. Forecasting using regression models is also discussed.

6. *Searching for periodicities.* The problem of determining if cycles exist in time series is considered and illustrated on some famous time series, including sunspot data. The general problem of spectral density estimation is also briefly discussed.

Figure 1.1. Some Examples of Time Series.

In this section we classify time series into three broad classes according to what we will call their memory type:

1. **Purely Random Series.** This type of series shows no patterns over time. Series III is an example of such a series. It was generated by TIMESLAB using a random number generator so that it would be indistinguishable from a random sample from a standard normal distribution. Series IV, on the other hand, is a real data set (monthly total rainfall), but appears quite similar to Series III. We will see later that such data are aptly named white noise. Purely random series are also called no-memory series, as one characterization of statistical independence is that an observation at one time has no memory of the observations at any other time.

2. **Long-Memory Series.** This type of series is the opposite extreme of white noise; that is, a plot of the data looks to be almost that of a deterministic function of time. Series V and VI illustrate this type. The first was artificially generated as values lying on a cosine curve that goes through ten cycles with small random numbers added to each point. Series VI is a real economic time series (monthly total international airline passengers for 12 years). These two series have in common that both could be almost perfectly extrapolated far into the future unless something were to happen to the mechanism generating the data. This is the origin of the term "long memory." The dependence on the past does not die away quickly. Note that many of the time series in business and economics are long memory.

3. **Short-Memory Series.** This type lies between white noise and long memory, occurs often in the physical and engineering sciences, and comprises the bulk of the time series that can be most usefully analyzed by the sophisticated methods of time series analysis that we will study. Series VIII and IX appear to be short-memory series; clearly observations close together in time are more similar than those far apart in time, but there are no apparent deterministic patterns in the data (although upon closer inspection you might be able to tell that Series IX is actually the sum of four cosine curves). In a short-memory series the predictability of the observations at one place in time from past observations appears to die out quickly as time goes on.

1.3. Basic Descriptive Statistics

The first aim of any statistical procedure is to give a succinct description of the data being analyzed, both graphically and numerically. In time series analysis there are three basic graphical techniques for describing data: the correlogram, the partial correlogram, and the periodogram. In this section we introduce each of these quantities in turn and illustrate how they are used to describe data $z(1), \ldots, z(n)$.

1.3.1. The Sample Correlogram

The distinguishing characteristic of a time series is that it can exhibit serial correlation, that is, correlation over time. For example, Figure 1.2 contains scatterplots of $z(t)$ versus $z(t-1)$ for $t = 2, \ldots, n$ for Series I and II. Note that for Series I there appears to be little correlation in the plot, while in Series II there appears to be high positive correlation.

In a time series $z(1), \ldots, z(n)$, we usually want to measure the correlation of the data with themselves except "lagged" a certain number of time units. Thus for a lag v, we have $n - v$ pairs of z's that are separated by v time units, namely the pairs

$$(z(1), z(1+v)), (z(2), z(2+v)), \ldots, (z(n-v), z(n)) .$$

The traditional way to measure serial correlation in time series is by the sample autocorrelation coefficient as given in the following definition.
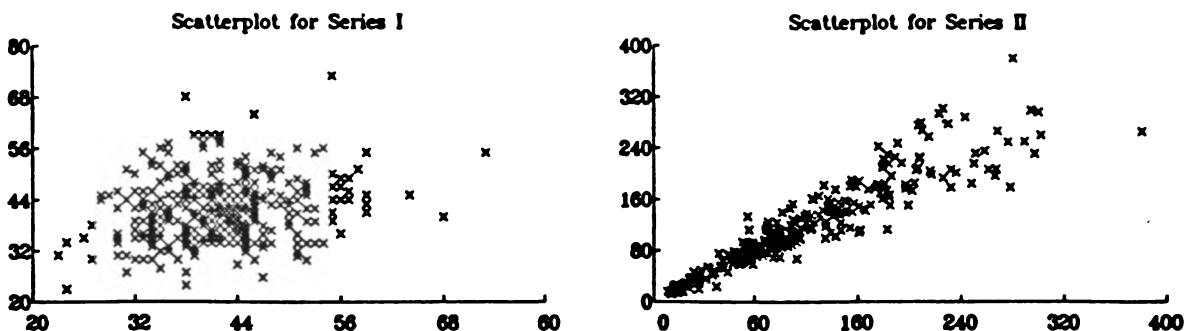
Figure 1.2. Scatterplot of $x(t)$ versus $x(t-1)$ for Series I and II.

Definition. *Let $x(1), \ldots, x(n)$ be a univariate time series data set. The sample autocorrelation coefficient of lag v is given by*

$$\hat{\rho}(v) = \frac{\sum_{t=1}^{n-v} (x(t) - \bar{x})(x(t+v) - \bar{x})}{\sum_{t=1}^{n} (x(t) - \bar{x})^2}, \qquad v < n,$$

where $\bar{x}$ is the sample mean of $x(1), \ldots, x(n)$. A plot of $\hat{\rho}(v)$ versus v for $v = 0, 1, \ldots, M$ for some maximum lag M is called the correlogram of the data.

Note that $\hat{\rho}(0) = 1$ and that the largest value of v that can be considered is $n - 1$, since if $v \geq n$, there are no pairs of x's separated by v time units. Note further that we will define $\hat{\rho}(-v) = \hat{\rho}(v)$.

To illustrate the use of the correlogram, consider Figure 1.3 where we have given the correlograms for the ten series introduced earlier. A large value of $\hat{\rho}(v)$ is indicative of a possible periodicity in the data of v time units. For example, the correlogram of a pure cosine curve is also sinusoidal. Notice that Series V and VI exhibit this behavior.

A correlogram that does (does not) decay rapidly to zero and then stay there as v increases means that the series is short (long) memory. Note that the wheat price index, the artificial series, the airline data, and the random walk series appear to be long memory by inspection of the correlogram.

1.3.2. The Sample Partial Correlogram

As in elementary statistics, high correlation does not imply causality. We can introduce partial and multiple correlations in an attempt to describe further the correlation in certain sets of variables. In elementary statistics, if we have random samples from populations Y, Z, and $X_1, \ldots, X_p$, we can find the residuals e_Y and e_Z of regressing Y on the X's and Z on the X's and then find the usual sample correlation coefficient between $e_Y(1), \ldots, e_Y(n)$ and $e_Z(1), \ldots, e_Z(n)$. The result is what is called the sample partial correlation coefficient between Y and Z given $X_1, \ldots, X_p$, and is often denoted $\hat{\rho}_{YZ|X_1 \cdots X_p}$. The sample multiple correlation coefficient $R^2_{Y|X_1 \cdots X_p}$ is the proportion of variability in Y that is 'explained' by its linear relationship with the X's. The analogous quantities in univariate time series analysis are the sample partial autocorrelation coefficients and residual variances. The sample partial autocorrelation coefficient of lag v is the correlation between $x(t)$ and $x(t+v)$ after having removed the common linear effect of the data in between (the lag one partial is just the usual lag one sample autocorrelation). We denote the lag v partial autocorrelation by $\hat{\theta}_v$. We will usually display what we call the standardized residual variances, that is, the residual variances divided by the sample variance. These values will always be between zero and one, and a small (large) value indicates that $x(t)$ is (is not) very predictable from its past, thus indicating that the data are long (short) memory. A useful rule of thumb is that data are long memory if their standardized residual variance sequence becomes less than $8/n$ for some lags.

In Figure 1.4 we display the partial correlogram and standardized residual variances for each of the univariate series that we have been considering. We can use the partial autocorrelations in the same way that usual partial correlations are used. In Series VII, for example, the correlogram takes some time to decay
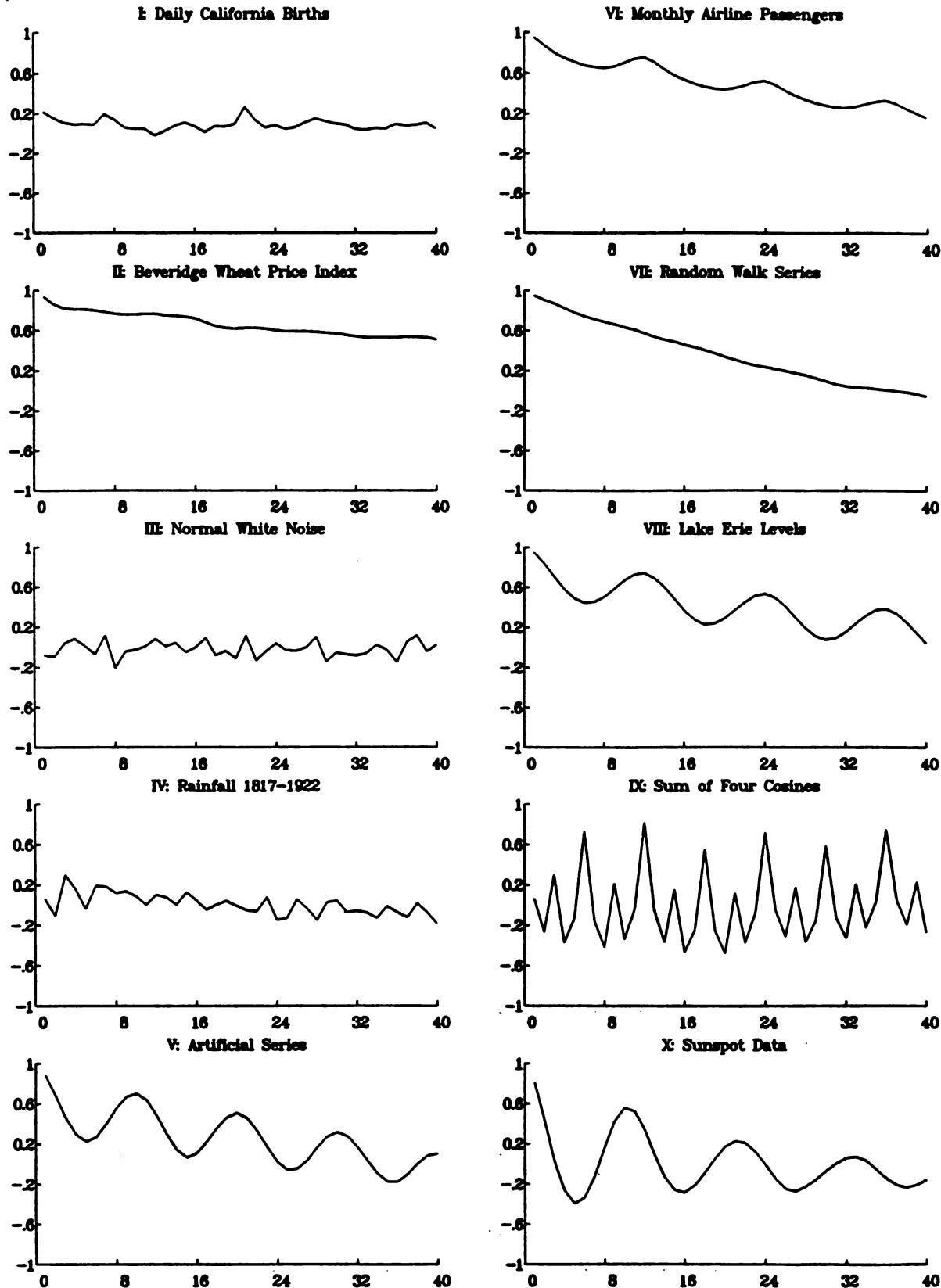
Figure 1.3. The Correlograms for Series I–X.

to zero, while the partial correlogram is large for lag one and then small henceforth. This indicates that the correlation between $x(t)$ and $x(t+2)$ is due only to the common relationship of $x(t)$ and $x(t+2)$ to $x(t+1)$.

1.3.3. The Periodogram and Sample Spectral Density

A recurring theme in science is the study of deterministic phenomena that have been observed over time (music, speech, radio signals, etc.) via harmonic analysis, that is, by decomposing a time record into the sum of sinusoids of various frequencies and amplitudes. Harmonic analysis is also important in the study of random phenomena observed over time. In this section we will look at what is called the periodogram of a data set $x(1),\dots,x(n)$.

The basic idea of harmonic analysis is that one can find a unique set of sinusoids (a cosine plus sine of the same frequency) that when added together reconstruct the mathematical object being studied. Thus one can study a complicated object by breaking it up into these simple "frequency components" and studying them separately. Further, in a certain sense, conclusions made about one sinusoid are independent of conclusions made about others.

A similar idea in statistics is that of orthogonal contrasts in the analysis of variance, wherein the results of applying different treatments to experimental units can be decomposed into the sum of linear, quadratic, and higher-order effects that are orthogonal, or statistically independent.

The mathematical objects that can be studied by harmonic analysis vary widely. Examples include sequences of finitely many numbers, infinitely long sequences, functions that are periodic, and general continuous or even discontinuous functions. No matter what object is being studied, the results are always of the form of a decomposition into independent sinusoids that can be studied separately.

Definition. *A sinusoid $g_k(x)$ of period k (or frequency $1/k$) is the function*

$$g_k(x) = a\cos\frac{2\pi x}{k} + b\sin\frac{2\pi x}{k}, \qquad x \in (-\infty,\infty).$$

Since cos and sin have period 2π, we have that $g_k(x + kl) = g_k(x)$, for any integer l. We can also write

$$g_k(x) = C\cos(\frac{2\pi x}{k} - \phi),$$

where $C = \sqrt{a^2 + b^2}$ and $\phi = \arctan(b/a)$ are called the amplitude and phase of g_k.

A data set $x(1),\dots,x(n)$ is an example of a finite sequence of numbers. The basic fact of the harmonic analysis of such a sequence is that we can calculate from $x(1),\dots,x(n)$ a set of orthogonal sinusoids of frequencies $1/n$, $2/n$, ..., $[n/2]/n$, where $[c]$ denotes the greatest integer less than or equal to c, together with a constant, so that when the constant and the sinusoids are added together, the resulting function coincides with the data at the indices $1,2,\dots,n$. Further, the sample variance of the x's is proportional to the sum of the squared amplitudes of the sinusoids.

Before presenting the sinusoidal decomposition of a time series data set, we define the discrete Fourier transform of a set of numbers.

Definition. *The discrete Fourier transform (DFT) of the (possibly complex) numbers $x(1),\dots,x(n)$ is the set of complex numbers $z(1),\dots,z(n)$ given by*

$$z(k) = \sum_{t=1}^{n} x(t)e^{2\pi i(t-1)\omega_k} = \sum_{t=1}^{n} x(t)\cos 2\pi(t-1)\omega_k + i\sum_{t=1}^{n} x(t)\sin 2\pi(t-1)\omega_k,$$
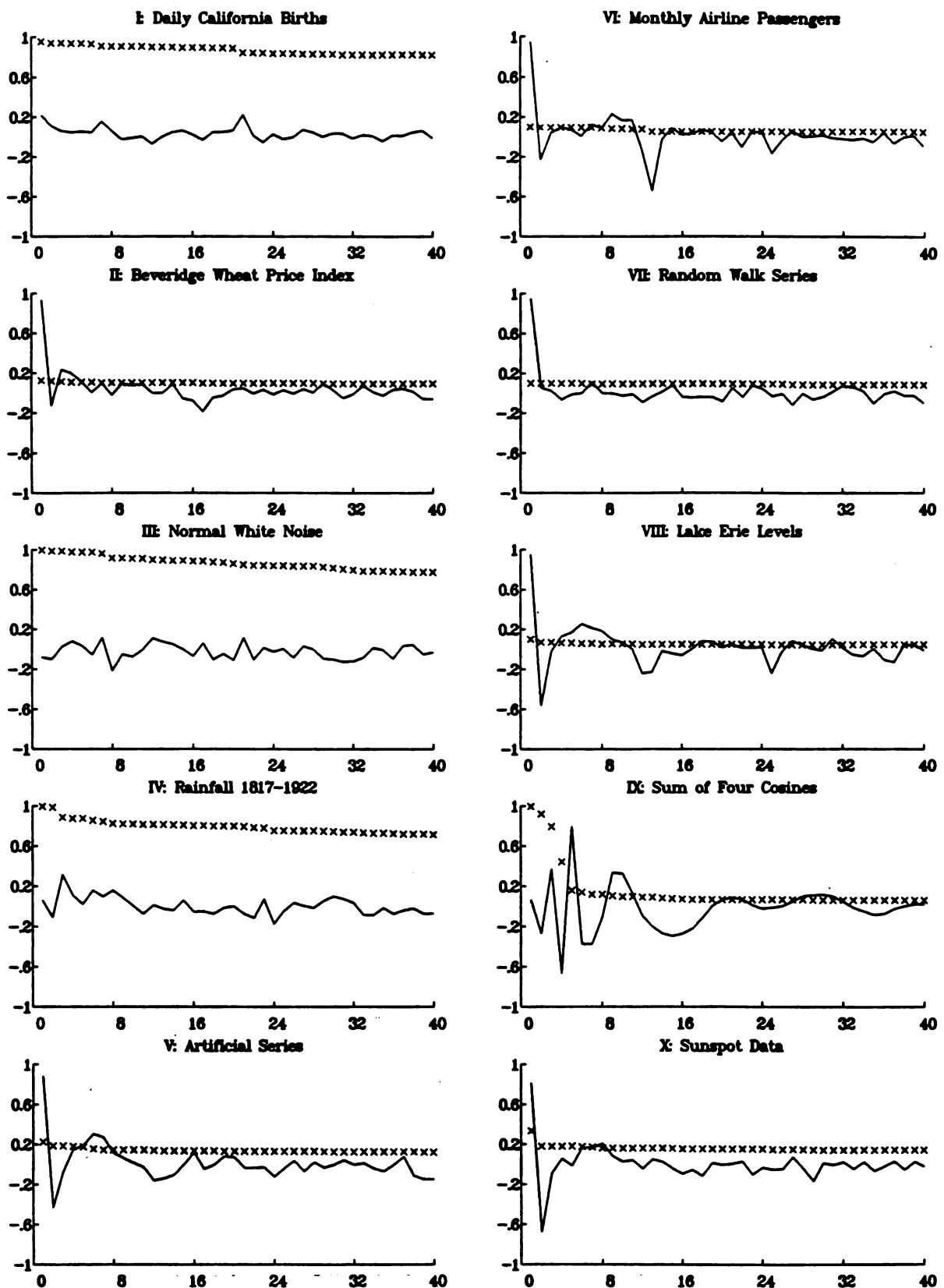
Figure 1.4. The Partial Autocorrelations (Solid Curve) and Standardized Residual Variances (Dotted Curve) for Series I–X.

Digitized by Google

Table 1.1. Interpreting the Periodogram

Appearance of Data	Nature of Periodogram
Smooth	Excess of low frequency; that is, amplitudes of sinusoids of low frequency (long period) are large relative to other frequencies
Wiggly	Excess of high frequency
Random (no pattern)	No frequencies dominate
Basically sinusoidal of period p time units	A peak at frequency $1/p$
Periodic of period p but not sinusoidal	A peak at fundamental frequency $1/p$ and peaks at some multiples of $1/p$ (harmonics)

where $\omega_k = (k-1)/n$, $k = 1, \ldots, n$.

If we insert a minus sign into the exponent in the first expression, we have what is called the inverse discrete Fourier transform (IDFT), since if one were to find the IDFT of the DFT of x, the result would be nx.

Theorem 1.3.1   SINUSOIDAL DECOMPOSITION OF DATA

Let $x(1), \ldots, x(n)$ be a finite sequence of numbers and let $z(1), \ldots, z(n)$ be the DFT of x. For $k = 1, \ldots, n$, let $\omega_k = (k-1)/n$ and define

$$a_k = \frac{1}{n} Re(z(k)), \qquad b_k = \frac{1}{n} Im(z(k)).$$

Let $\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (x(t) - \bar{x})^2$, and for $t = 1, \ldots, n$ define

$$g_k(t) = a_k \cos 2\pi(t-1)\omega_k + b_k \sin 2\pi(t-1)\omega_k$$
$$= C_k \cos[2\pi(t-1)\omega_k - \arctan(b_k/a_k)].$$

Then

a) $a_1 = \bar{x}$, $\quad b_1 = 0$, $\quad C_1^2 = \bar{x}^2$, $\quad g_1(t) = \bar{x}$

b) $x(t) - \bar{x} = \sum_{k=2}^{n} g_k(t)$

c) $\hat{\sigma}^2 = \sum_{k=2}^{n} C_k^2$

d) For $k = 2, \ldots, [n/2] + 1$, we have

$$a_k = a_{n-k+2}, \qquad b_k = -b_{n-k+2}$$

$$\cos 2\pi(t-1)\omega_k = \cos 2\pi\omega_{n-k+2}$$

$$\sin 2\pi(t-1)\omega_k = -\sin 2\pi\omega_{n-k+2}$$

and thus

$$g_k(t) = g_{n-k+2}(t), \qquad C_k^2 = C_{n-k+2}^2,$$

which gives

$$x(t) - \bar{x} = \begin{cases} 2\sum_{k=2}^{[n/2]+1} g_k(t), & n \text{ odd} \\[2ex] 2\sum_{k=2}^{n/2} g_k(t) + g_{(n/2)+1}(t), & n \text{ even} \end{cases}$$

$$\hat{\sigma}^2 = \begin{cases} 2\sum_{k=2}^{[n/2]+1} C_k^2, & n \text{ odd} \\[2ex] 2\sum_{k=2}^{n/2} C_k^2 + C_{(n/2)+1}^2, & n \text{ even.} \end{cases}$$

e) For $k = 1, \ldots, [n/2]+1$, the vectors $\mathbf{g}_k = (g_k(1), \ldots, g_k(n))^T$ are orthogonal, that is, $\mathbf{g}_j^T \mathbf{g}_k = 0$, $j \neq k$.

Implications: This theorem gives a decomposition of a time series data set into a set of orthogonal (part (e)) sinusoids of frequencies $1/n$, $2/n$, ..., $[n/2]/n$ (part (d)), and gives a decomposition of the variability in the x's (as measured by $\hat{\sigma}^2$) in terms of the sum of squared amplitudes of the sinusoids (part (c)). In terms of understanding why the x's vary then, these squared amplitudes play an important role. Note that $C_k^2 = |x(k)|^2/n^2$.

Definition. *For a time series data set $x(1), \ldots, x(n)$, let $\omega_k = (k-1)/n$, $k = 1, \ldots, [n/2]+1$, and define*

$$C_k^2 = \frac{1}{n^2} \left| \sum_{t=1}^{n} x(t) e^{2\pi i (t-1)\omega_k} \right|^2, \qquad k = 1, \ldots, [n/2]+1.$$

A plot of nC_k^2 versus ω_k is called the periodogram of x. The function

$$\hat{f}(\omega) = \begin{cases} \frac{1}{n} \left| \sum_{t=1}^{n} x(t) e^{2\pi i (t-1)\omega} \right|^2, & \omega \in [0, .5] \\[2ex] \hat{f}(1-\omega), & \omega \in [.5, 1] \end{cases}$$

is called the sample spectral density function of x.

Note that the periodogram is the sample spectral density evaluated at the so-called natural frequencies $\omega_1, \omega_2, \ldots, \omega_{[n/2]+1}$.

Interpreting the Periodogram

In Figure 1.5 we give a plot of three data sets of length 200 (and the log of their periodograms) that were constructed by

$$x(t) = \alpha \cos \frac{2\pi(t-1)}{100} + \beta \cos \frac{2\pi(t-1)}{10} + \delta \cos \frac{2\pi(t-1)}{4},$$

that is, as the sum of pure cosines of frequencies 1/100, 1/10, and 1/4. The three series were obtained by varying α, β, and δ (Series 1 has (10,3,1), Series 2 has (3,3,3), and Series 3 has (1,3,10)).

Recall that a sinusoid of frequency ω is the sum of a sine and a cosine of that frequency, but for simplicity we are considering sinusoids that have no sine part. A sinusoid of long period (low frequency) is very smooth in appearance relative to one of short period (high frequency). Thus when α is large relative to β and γ (such as Series 1), we would expect x to be relatively smooth in appearance; that is, the long-term rise and fall of the data should be large relative to short term oscillations. On the other hand, Series 3 appears quite wiggly since γ is large relative to α and β. In Series 2 we have used equal values of the three coefficients and the resulting data are between Series 1 and 3 in terms of wiggliness.
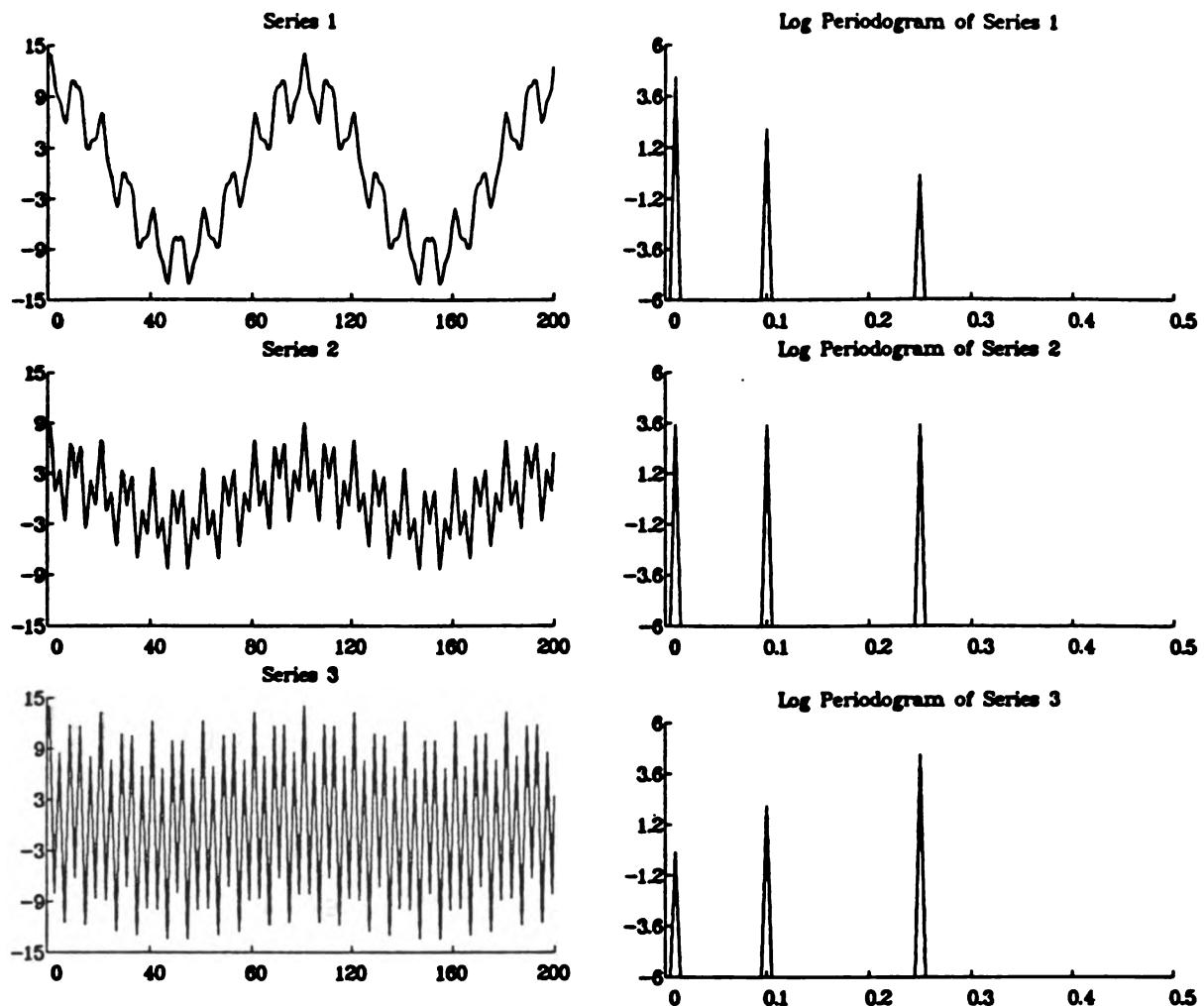
Figure 1.5. Sums of Pure Cosines and Their log Periodograms.

Note how the periodograms pick out the amplitudes of the component sinusoids. In general, of course, time series data sets are not made up of sums of only a few pure sinusoids. However, with the above example as motivation, we can make the qualitative statements given in Table 1.1.

In Figure 1.6 we give plots of the logarithm of a standardized form of the periodograms of the ten univariate series that we have been discussing throughout this lecture. Note how these graphs confirm the discussion above. In particular, note the harmonics in the periodogram of the airline data, the random pattern in those for Series III and IV, the excess of low frequency in Series VII, the periodicity in the birth data, and how the periodogram is able to tell that Series IX is in fact the sum of four pure cosines.

Displaying the Periodogram

It often happens that a few values of a periodogram are very large relative to the rest and thus dwarf them in the plot. Plotting the (natural) log of the values rather than the values themselves allows us to see other possible values of interest. We would also like to plot the log periodogram on some standard scale so that several such plots can be compared in a meaningful way. A useful standardization of the periodogram is suggested by the following theorem, which follows immediately from part (c) of Theorem 1.3.1.
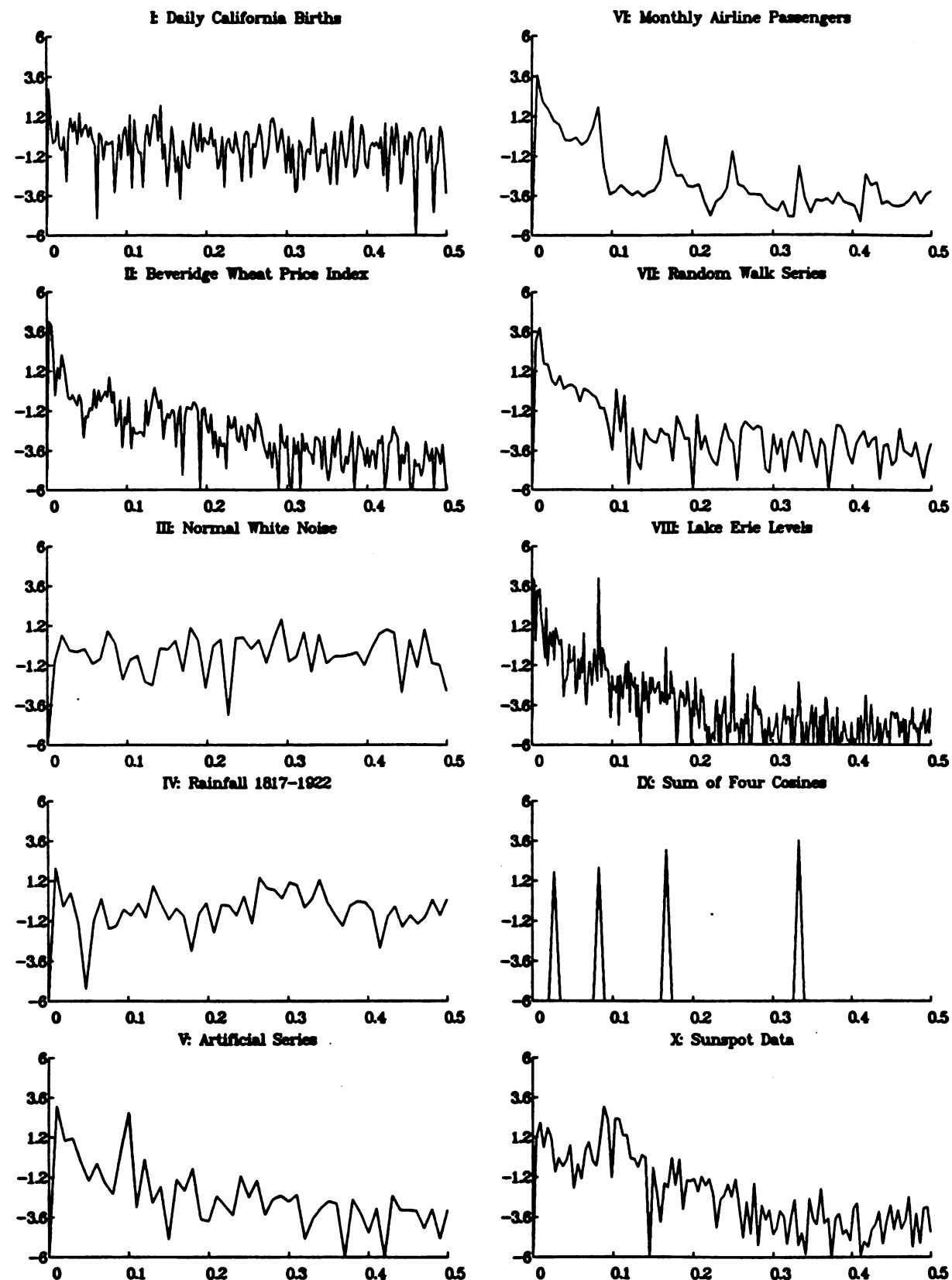
Figure 1.6. The Log of the Standardized Periodograms of Series I–X.

| Theorem 1.3.2 | STANDARDIZING THE PERIODOGRAM |

Let nC_k^2 for $k = 1, \ldots, n$ be the periodogram ordinates of a time series $x(1), \ldots, x(n)$, and let $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x(t) - \bar{x})^2$. Then

$$\frac{1}{n}\sum_{k=2}^{n}\frac{nC_k^2}{\hat{\sigma}^2} = 1;$$

that is, the average value of $nC_k^2/\hat{\sigma}^2$ is one.

Because of this theorem we will routinely display

$$\log\left(\frac{nC_k^2}{\hat{\sigma}^2}\right) \text{ versus } \frac{k-1}{n}, \qquad k = 1, \ldots, [n/2] + 1,$$

with the graph truncated at -6 and 6 on the vertical axis. We will rarely encounter a value of $nC_k^2/\hat{\sigma}^2$ that is greater than e^6 or less than e^{-6}.

The Sample Spectral Distribution Function

We have seen that data that have no obvious trends, cycles, or serial correlation should have a periodogram that oscillates roughly about a constant. On the other hand, smooth (wiggly) data have an excess of low (high) frequency, while cyclic data will have peaks in their periodogram. One useful way to display these ideas graphically is via the sample spectral distribution function.

Definition. Let $\hat{f}(\omega_1), \ldots, \hat{f}(\omega_Q)$ be the sample spectral density function of data $x(1), \ldots, x(n)$ at the frequencies $\omega_j = (j - 1)/Q$, $j = 1, \ldots, Q$. Let $q = [Q/2] + 1$ and

$$\hat{F}(\omega_k) = \frac{\sum_{j=1}^{k}\hat{f}(\omega_j)}{\sum_{j=1}^{q}\hat{f}(\omega_j)}, \qquad k = 1, \ldots, q.$$

Then $\hat{F}(\omega_1), \ldots, \hat{F}(\omega_q)$ is called the sample spectral distribution function of x.

If $Q = n$, then the $\hat{f}$'s are in fact the periodogram ordinates, and thus the $\hat{F}$'s could be called the cumulative periodogram of x. Note that $\hat{F}(\omega_q) = 1$, while if $\bar{x} = 0$, we have $\hat{f}(\omega_1) = 0$ and so $\hat{F}(\omega_1) = 0$. Thus

$$0 \le \hat{F}(w_k) \le 1, \qquad k = 1, \ldots, q,$$

and the spectral distribution function of random data will fluctuate about the straight line that connects the points $(0,0)$ and $(.5,1)$, that is, the line $y = 2x$.

In Figure 1.7 we display the cumulative spectral density of our example data sets. Note that these graphs are much less variable than their corresponding log periodograms. Also, an excess of low (high) frequency means that the cumulative periodogram starts out above (below) the line $y = 2x$ before catching up to it before $\omega = .5$, while peaks in the periodogram translate into jumps in the cumulative periodogram (such as at frequency $1/12$ in the airline data), and the cumulative periodograms of the white noise series (Series III and IV) vary little from the line $y = 2x$.
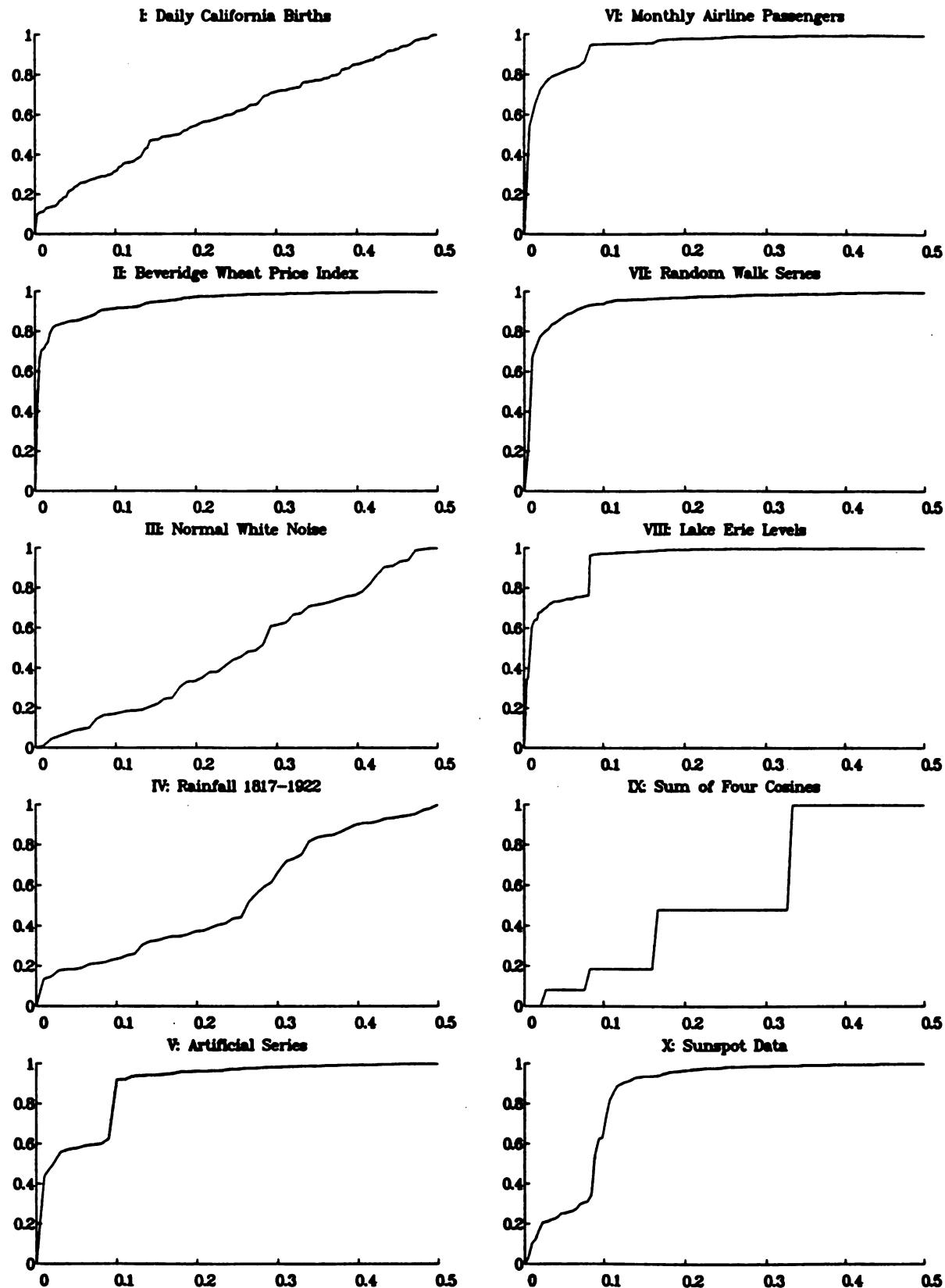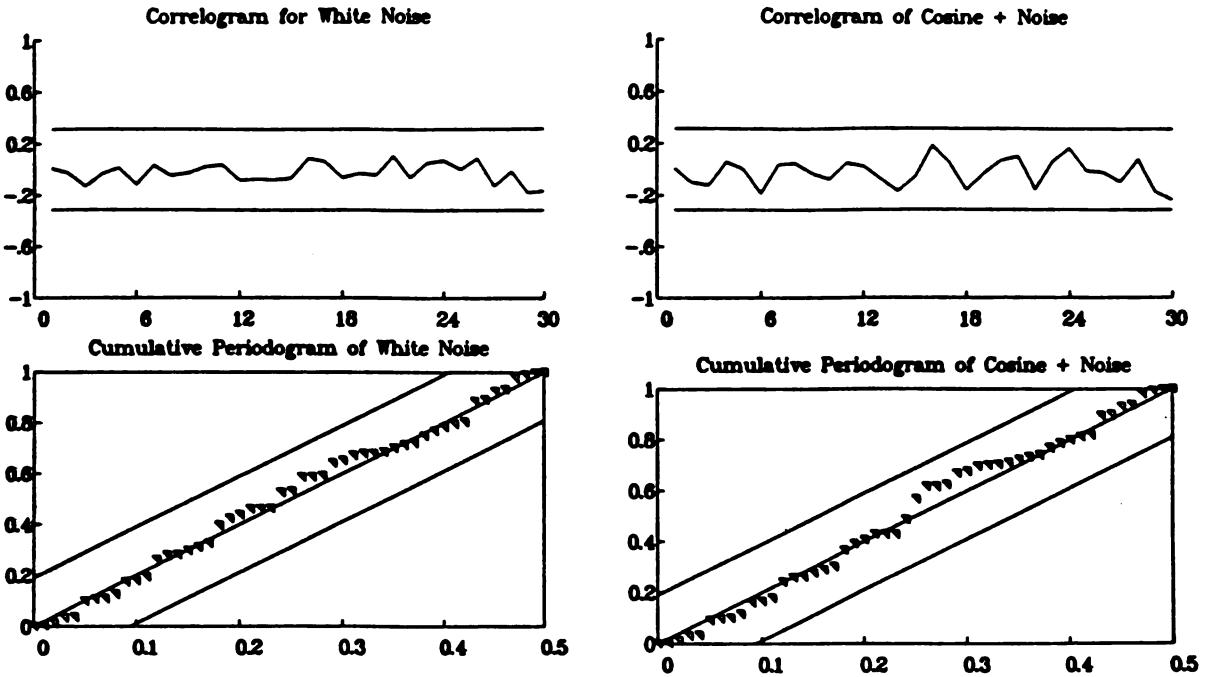
Figure 1.7. Cumulative Periodograms for Series I–X.

Figure 1.8. Example of White Noise Test.

1.3.4. The Relationship of the Correlogram and Periodogram

Recall that the sample autocorrelation coefficient of lag v for data z is given by

$$\hat{\rho}(v) = \frac{\sum_{t=1}^{n-|v|} (z(t) - \bar{z})(z(t+v) - \bar{z})}{\sum_{t=1}^{n} (z(t) - \bar{z})^2}, \qquad |v| < n.$$

Then $\hat{\rho}(v)$ can be written as

$$\hat{\rho}(v) = \frac{\hat{R}(v)}{\hat{R}(0)}, \qquad |v| < n,$$

if we define the sample autocovariance function $\hat{R}$ by

$$\hat{R}(v) = \frac{1}{n} \sum_{t=1}^{n-|v|} (z(t) - \bar{z})(z(t+v) - \bar{z}), \qquad |v| < n.$$

We note that it can be shown that

$$\hat{f}(\omega) = \sum_{v=-(n-1)}^{n-1} \hat{R}(v) e^{-2\pi i v \omega}, \qquad \hat{R}(v) = \int_0^1 \hat{f}(\omega) e^{2\pi i v \omega}.$$

The Fast Fourier Transform (FFT)

To calculate the discrete Fourier transform

$$w(k) = \sum_{t=1}^{n} z(t) e^{2\pi i (t-1)(k-1)/n}, \qquad k = 1, \ldots, n,$$

312

of data $x(1), \ldots, x(n)$ appears to require n^2 multiplications and additions (n for each of the n values of k) and many evaluations of complex exponentials. In the mid-1960s various researchers made use of a variety of trigonometric identities to obtain algorithms called fast Fourier transform (FFT) algorithms that essentially require only $n(p_1 + \cdots + p_k)$ multiplications and additions, where $p_1, \ldots, p_k$ are the prime factors of n, and a greatly reduced number of evaluations of complex exponentials. For example, if $n = 1024 = 2^{10}$, then the number of operations is $1024(2 + \cdots + 2) = 10 \cdot 2 \cdot 1024 = 2\log_2 1024$, which is approximately 50,000 as opposed to $1024^2 \sim 1,000,000$, a savings of a factor of about 50.

Note that if n is not very composite, that is, it has some large prime factors, the FFT saves very little over a straightforward DFT.

1.4. Testing for White Noise

The first inference that one should make about an observed time series is to whether or not it could be considered to be a realization from a white noise process. In this section we consider two graphs that are useful for making such an inference. If $x(1), \ldots, x(n)$ is a random sample from a population, then for large n:

a) The correlations $\hat{\rho}(1), \ldots, \hat{\rho}(m)$ are independent and identically distributed as $N(0, 1/n)$ variables. Thus there is approximately a 95 percent chance that an individual $\hat{\rho}(v)$ will be outside of $\pm 1.96/\sqrt{n}$. To produce simultaneous confidence bands having 95 percent confidence level, we must construct individual intervals having level $.95^{1/m}$. This is what is done in the first graph.

b) The cumulative periodogram has a 95 percent chance of falling entirely within the lines $y = 2x \pm 1.36/\sqrt{q}$, where $q = [n/2] + 1$. The second graph is the cumulative periodogram with these two lines and the line $y = 2x$.

In Figure 1.8 we give the result for two data sets; the first being a normal white noise series of length 100, and the second being the same series with a cosine of length 100, amplitude .5, and period 4 added to it. Notice that none of the boundary lines are crossed for either series, but that the cumulative periodogram seems to increase rapidly at $\omega = .25$, something that is unusual and should be a hint that this data set is in fact not white noise.

LECTURE 2

Transforming and Forecasting Time Series

2.1. Transformations

Some of the time series analysis techniques that we will introduce in later lectures assume (1) that the data being analyzed have no deterministic trends or cycles, and (2) that the variability in the data is constant over time. The traditional method of analyzing data that fail to meet these requirements is to do the analysis in three steps: (1) try various transformations until the result appears to meet the requirements, (2) analyze the result of step 1, and (3) do the inverse operation of what was done in step 1. This strategy is very similar to that used in regression analysis. In this section we will describe some of the transformations that are often used in time series analysis.

2.1.1. Stabilizing Variance

Suppose that the variability in a data set x appears to be increasing as time increases (see the airline data for example). If the mean level of the data is also increasing with time, then the variability in $y(t) = \log x(t)$ should appear fairly constant over time. This is a fairly common occurrence in real data. In general, one might try various power transformations to obtain a series having constant variability; that is,

$$y(t) = \big(x(t)\big)^r.$$

2.1.2. Removing Trends

Another common phenomenon in time series data is that the values appear to be growing in some polynomial fashion with time, particularly linearly; that is, the data appear to follow

$$x(t) = a + bt + \epsilon(t),$$

where $\epsilon(t)$ is white noise. Such polynomial trends are often removed by using differencing.

Definition. *The dth difference of* a *time series data set* x *having n elements is* a *data set* y *having* $n - d$ *elements obtained by*

$$y(t) = x(t + d) - x(t), \qquad t = 1, \ldots, n - d.$$

It is easy to see that if x contains a general dth-degree polynomial trend, applying first differencing d times will remove it, while if x has a cycle of length s time units, then taking sth differences will remove the cycle.
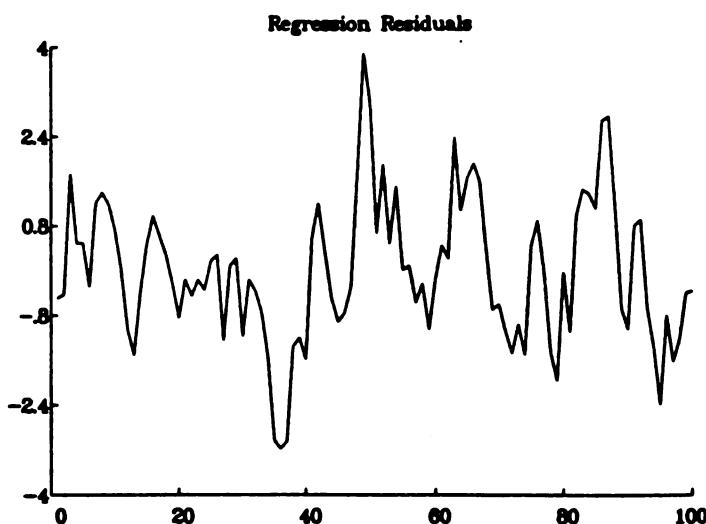
Figure 2.1. Residuals from Regression for Series V.

Perhaps the most well known example of differencing is the airline data that we have been analyzing. The variability in the data is increasing with time and thus a log transform is usually applied. The result of this transform has an obvious 12-month cycle and also a linear trend. Thus the traditional advice on a series such as this is to take the first difference of the 12th difference of the log of the original data.

Regression Analysis

A traditional method of removing trends and cycles from time series data is to do ordinary regression of the data on the deterministic functions of time. For example, Series V of the univariate series that we have been analyzing was formed by

$$x(t) = 10 + .1t + 3 \cos \frac{2\pi(t-1)}{10} + \epsilon(t), \qquad t = 1, \ldots, 100,$$

where $\epsilon(t)$ is a series having serial correlation If we did ordinary least squares regression of the form

$$y(t) = \beta_0 + \beta_1 t + \beta_2 \cos \frac{2\pi(t-1)}{10} + \epsilon(t),$$

we would obtain the residuals given in Figure 2.1, which can then be further analyzed by time series analysis.

2.1.3. Accounting for Seasonal Variability

Often it is of interest in seasonal time series to analyze how a data set differs from regular seasonal variation. For example, suppose that x consists of m years of monthly data, that is, $n = 12m$. We can define the monthly means and variances to be the means and variances of each of the 12 data sets consisting of like months; the Januaries, Februaries, and so on. Thus,

$$\bar{x}_k = \frac{1}{m} \sum_{t=1}^{m} x(k + 12(t-1))$$

$$s_k^2 = \frac{1}{m} \sum_{t=1}^{m} \left[x(k + 12(t-1)) - \bar{x}_k \right]^2,$$

315
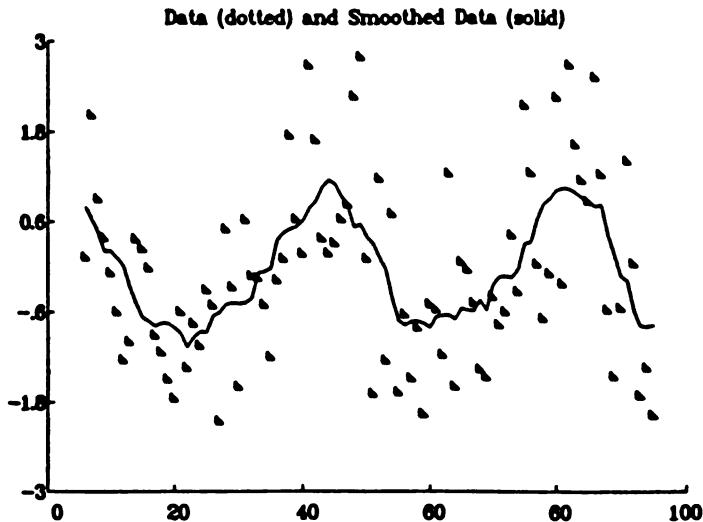
Figure 2.2. An Example of the Moving Average Smoother.

$k = 1, \ldots, 12$. Note that other seasonal (quarterly or hourly, for example) means and variances can be defined similarly.

2.1.4. General Smoothing Operations

Another approach to removing deterministic-looking parts of a time series is to use general methods of smoothing data and then analyze the deviations from the smooth version. For example, one way to smooth wiggly data is to use a moving average smoother, for example, the one of length 3:

$$y(1) = \frac{x(1) + x(2) + x(3)}{3}$$

$$y(2) = \frac{x(2) + x(3) + x(4)}{3}$$

$$\vdots$$

$$y(n-2) = \frac{x(n-2) + x(n-1) + x(n)}{3}.$$

Since consecutive y's have two of the x's in common in their average, we would expect that the y's won't vary as much as the original x's; that is, y will be much smoother in appearance. In Figure 2.2 we have generated 100 points of a cosine of amplitude one and period 40, added $N(0,1)$ white noise to it, and then used a moving average smoother of length 11. Note that the last smoothed value corresponds to the 95th data point, and so we have superimposed the 6th through 95th data points and the smoothed data. Note that the smoothed data clearly exhibit the cosine curve, which is not obvious in the original data.
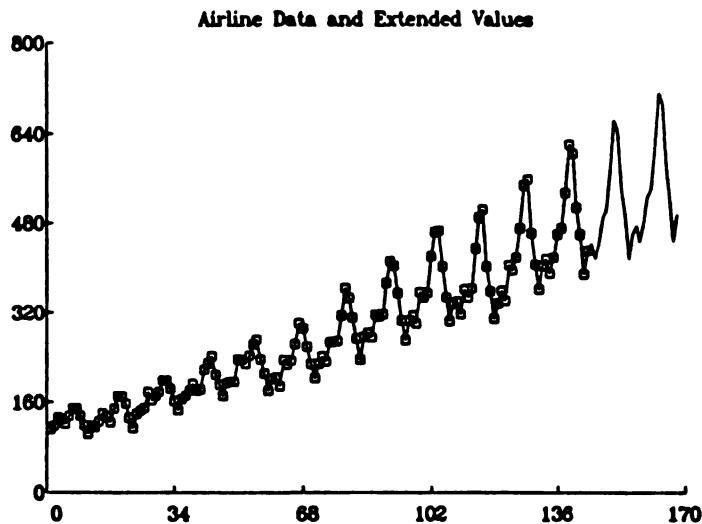
Figure 2.3. Airline Data and Extended Values.

2.2.3. Simple Moving Average

This method models an observation as a simple average of the previous m observations where m is to be chosen. For a given moving average length k we can calculate

$$S(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} \left(z(t) - \frac{1}{k} \sum_{j=1}^{k} z(t-j) \right)^2$$

as a measure of how well the simple moving average model of length k fits the observed data. We chose m as the value of k minimizing $S(k)$. Then we forecast future values recursively. For example, if $m = 3$ we calculate

$$\hat{z}(n+1) = \frac{1}{3} \left[z(n) + z(n-1) + z(n-2) \right]$$

$$\hat{z}(n+2) = \frac{1}{3} \left[\hat{z}(n+1) + z(n) + z(n-1) \right]$$

$$\hat{z}(n+3) = \frac{1}{3} \left[\hat{z}(n+2) + \hat{z}(n+1) + z(n) \right],$$

and so on.

2.2.4. Simple Exponential Smoothing

Instead of modeling an observation as a simple average of the previous m observations, exponential smoothing methods model $z(t)$ as a weighted average of all of the previous values. The type of weights that are used depends on the appearance of the data and leads to methods having a variety of names. We will discuss only the simple exponential smoothing technique which is most suitable for data that appear to have no linear or seasonal trends.

2.2. Some Simple Forecasting Methods

An important objective of time series analysis in many scientific areas is to forecast (predict) future values of a time series $x(1), \ldots, x(n)$. In this section we begin our discussion of forecasting methods by looking at some simple methods. In later lectures we will consider forecasting using time series models.

Each of the methods in this section can be viewed as having two parts: (1) a model-fitting part, and (2) a forecasting part.

2.2.1. The Inverse Differencing Method

A simple but effective method for extrapolating trends and cycles in data that can be handled by differencing is to extend the data set with M new values in such a way that if the extended series of length $n + M$ were then differenced, the last M values (the ones corresponding to the new values) would all equal the mean of the differences of the original data. For example, if a data set appears to have a linear trend, we might detrend it by applying first differences. We could then find a new data point $x(n + 1)$ so that $x(n+1) - x(n) = \bar{z}$, where $\bar{z}$ is the mean of the differences of $x(1), \ldots, x(n)$. Thus $x(n+1) = x(n) + \bar{z}$, and if we continue this process, we obtain

$$x(n + h) = x(n + h - 1) + \bar{z}, \qquad h \geq 1.$$

If we have data that appear to contain both a linear trend and a seasonal cycle, we might apply both first and dth differences, where d is the length of the seasonal cycle. For monthly data having an annual cycle, we would solve

$$x(n + 1) - x(n) - x(n - 11) + x(n - 12) = \bar{z},$$

where $\bar{z}$ is the mean of the first and 12th differences of $x(1), \ldots, x(n)$. This results in the predictor

$$x(n + h) = x(n + h - 1) + x(n + h - 12) - x(n + h - 13) + \bar{z}, \qquad h \geq 1.$$

In Figure 2.3 we give the result of inverse differencing applied to the airline data. The observed data values are represented on the plot by both a solid curve and an × at each point, while the extended values are represented only by the solid curve. The extended values are the last two cycles on the graph and are consistent with what we would expect from a predictor. It is not surprising that we could use this "naive" method on the airline data as it is a long-memory series and we would expect that almost any method would work well, including just drawing in the next two cycles by hand. For series that do not follow such a deterministic pattern we will have to use more sophisticated methods.

2.2.2. The Regression Method

If the data appear to follow a deterministic function of time that can be expressed as a simple or multiple linear regression model, we can predict future values of the series by substituting future values of t into the least squares regression function. For example, if

$$x(t) = a + bt + c \cos \frac{2\pi(t - 1)}{p} + \epsilon(t),$$

for $t = 1, \ldots, n$, where the period p is known, we can estimate a, b, and c by their least squares estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$, and forecast a value h steps into the future by

$$\hat{x}(n + h) = \hat{a} + \hat{b}(n + h) + \hat{c} \cos \frac{2\pi(n + h - 1)}{p}.$$

The simple exponential smoothing method consists of modeling $z(t)$ as a weighted average:

$$\hat{z}(t+1) = \sum_{j=1}^{t} \beta_j z(t+1-j), \qquad t \geq 1.$$

If we let $\beta_j = \alpha(1-\alpha)^{j-1}$, where $0 \leq \alpha \leq 1$; that is, we let the weights decay exponentially to zero, then $\sum_{j=1}^{\infty} \beta_j = 1$, and thus for large t, the weights will approximately sum to one. We can choose α as the value of a minimizing

$$S(a) = \sum_{t=2}^{n} \left(z(t) - \sum_{j=1}^{t-1} a(1-a)^{j-1} z(t-j) \right)^2,$$

and then forecast the value at time $n+1$ by

$$\hat{z}(n+1) = \sum_{j=1}^{n} \alpha(1-\alpha)^{j-1} z(n+1-j).$$

The calculations involved in this process can be greatly reduced by noting that

$$\hat{z}(t+1) = \alpha z(t) + (1-\alpha)\hat{z}(t),$$

where $\hat{z}(1)$ is defined to be $z(1)$. We can also write this as $\hat{z}(t+1) - (1-\alpha)\hat{z}(t) = \alpha z(t)$, which is called a difference equation of order one.

2.3. Difference Equations

Difference equations are very important in the study of time series analysis. In particular, many forecasting methods can be thought of as future values of difference equations. Further, many probabilistic models for time series are written as difference equations.

Definition. *Let $z(\cdot)$ and $w(\cdot)$ be sequences of real numbers. Then*

$$z(t) + \alpha_1 z(t-1) + \cdots + \alpha_p z(t-p) = w(t)$$

is called a difference equation of order p, coefficients $\alpha_1, \ldots, \alpha_p$, and forcing term $w(\cdot)$.

To calculate the values of z for all values of t it is sufficient to know all of the values of w and any p consecutive values of z. These p values are called starting values or initial conditions. If we know the starting values $z(1), \ldots, z(p)$ and the values $w(p+1), \ldots, w(n)$ for a difference equation, then we can find $z(p+1), \ldots, z(n)$ recursively by

$$z(p+j) = w(p+j) - \sum_{k=1}^{p} \alpha_k z(p+j-k), \qquad j = 1, \ldots, n-p.$$

Note that $w(p+1)$ and $z(1), \ldots, z(p)$ are used to find $z(p+1)$, which is in turn used in finding $z(p+2)$, and so on.

319

Definition. *Let*

$$\sum_{j=0}^{p} \alpha_j z(t-j) = w(t)$$

be a pth-order difference equation. If $w(t) = 0$ for all t, then the difference equation is said to be homogeneous. The polynomials

$$g(z) = \sum_{j=0}^{p} \alpha_j z^j \quad \text{and} \quad h(z) = \sum_{j=0}^{p} \alpha_j z^{p-j}$$

are called the characteristic and indicial polynomials, respectively, of the difference equation.

Let $z_1, \ldots, z_p$ be the zeros of h; that is, $h(z_i) = 0$. Note that the zeros of g are the reciprocals of the zeros of h. These zeros play an important role in solving difference equations.

If we were to plot the real and imaginary parts of a complex number $z = a + bi$ on an X-Y plane, then $|z| = \sqrt{a^2 + b^2}$ is greater than one if and only if the plotted point falls outside of a circle of radius one that is centered at the origin. This circle is called the unit circle, and we will henceforth refer to zeros being inside, outside, or on the unit circle.

LECTURE 3

Time Series Models

In this lecture we consider how the basic ideas of probability theory apply to the analysis of time series and we discuss some of the traditional time series models.

3.1. Introduction

As in many areas of statistics, our basic aim in time series analysis is twofold: descriptive and inferential. In Lecture 1 we considered some basic ways to describe time series data. To make inferences from data we will use the following strategy:

1. Assume that some member of a family of models will adequately represent the observed behavior of a time series data set.

2. Identify which member of the family best represents the data (model identification).

3. Estimate the parameters of the chosen model.

4. Check the adequacy of the fit of the estimated model.

5. Make statistical and scientific inferences based on the characteristics of the chosen model.

Note that we will also make inferences occasionally without assuming a particular model for the data. We will call these nonparametric in analogy with the usual nonparametric analysis in the random sampling case.

The aim of this lecture is to introduce both the quantities that we want to make inferences about (the correlogram, spectral density function, and predictors) and the models that are usually used to allow us to make meaningful inferences about these quantities.

We visualize that a data set $x(1), \ldots, x(n)$ is just one possible set out of many that could have been generated by some random mechanism that is producing data. The set of all possible realizations that could be observed is called the ensemble of realizations.

Definition. *A time series is an indexed collection $\{X(t), t \in T\}$ of random variables having finite second moments; that is, $E(X^2(t)) < \infty$ for each element of the index set T.*

Usually we will consider T to be the set Z of all integers; that is, we will assume that the phenomenon being observed has been going on for a long time and will continue indefinitely. We will often refer to a time series as $X(t)$ or just X if there is no possibility of confusion. We use capital letters to refer to random variables and lowercase letters for particular values for the random variables.

3.2. Covariance Stationary Time Series

Traditionally, two assumptions are made about the joint behavior of the random variables making up a time series. The first we will label as an assumption and the second is the definition of covariance stationarity.

Assumption. *The behavior of a time series X can be adequately described by a knowledge of its mean value function m and covariance kernel K, that is,*

$$m(t) = E(X(t)), \qquad t \in Z$$
$$K(s,t) = \text{Cov}(X(s), X(t)), \qquad s,t \in Z.$$

If X is a Gaussian time series then this assumption is valid. We will see later that a wide variety of non-Gaussian time series can also be effectively analyzed. There is currently a great deal of effort being made in the time series community toward developing methods for analyzing data from processes which do not satisfy this assumption.

Definition. *A time series X is said to be a Gaussian time series if the joint distribution of any finite number of $X(t)$'s is multivariate normal; that is, for any positive integer n and any n integers $t_1, \ldots, t_n$, we have that the joint distribution of $X(t_1), \ldots, X(t_n)$ is multivariate normal.*

The multivariate normal distribution plays a crucial role in the theory and analysis of time series.

Definition. *A time series X is said to be covariance (or weakly) stationary if its mean value function is a constant, that is, $E(X(t)) = \mu$, and if there exists a function $\{R(v), v \in Z\}$ such that*

$$K(s,t) = \text{Cov}(X(s), X(t)) = R(t-s);$$

that is, the covariance of any pair of X's that are the same distance apart is the same. The function R is called the autocovariance function of X.

We will often drop the prefix *auto* in the word *autocovariance*. Note that there is another (stronger) type of stationarity, namely strict stationarity.

The Autocorrelation Function

We usually concentrate on the correlation between two random variables rather than on their covariance.

Definition. *Let X be a covariance stationary time series having autocovariance function R. The autocorrelation function of X is given by*

$$\rho(v) = \text{Corr}(X(t), X(t+v)), \qquad v \in Z.$$

A plot of $\rho(v)$ versus v for $v = 0, \ldots, M$ is called the correlogram of X.

Note that by the definition of the correlation of two random variables, we have

$$\rho(v) = \frac{\text{Cov}(X(t), X(t+v))}{\sqrt{\text{Var}(X(t))\text{Var}(X(t+v))}} = \frac{R(v)}{\sqrt{R(0)R(0)}} = \frac{R(v)}{R(0)},$$

and thus $\rho(0) = 1$ and $\rho(-v) = \rho(v)$. Thus the autocorrelation function of X is just the autocovariance function divided by the variance $R(0)$ of the series.

We will be primarily concerned with covariance stationary time series and thus will be trying to make inferences about μ and the autocovariance function R. In Theorem 3.2.1 we summarize some of the basic facts about R, and introduce a functions that is mathematically equivalent to R and are of importance in their own right, both theoretically and in practice.

Theorem 3.2.1 | THE SPECTRAL DENSITY FUNCTION

Let $\{R(v), v \in \mathcal{Z}\}$ be the autocovariance function of the covariance stationary time series X. Then

a) $R(v) = R(-v), \qquad v \in \mathcal{Z}$.

b) If R is absolutely summable, that is, $\sum_{v=-\infty}^{\infty} |R(v)| < \infty$, then there exists a function $f(\omega), \omega \in [0, 1]$, symmetric about $\omega = .5$, called the spectral density function of X, such that

$$R(v) = \int_0^1 f(\omega) \cos 2\pi v \omega \, d\omega, \qquad v \in \mathcal{Z}$$

and

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v) \cos 2\pi v \omega, \qquad \omega \in [0, 1].$$

c) The equations relating R and f can also be written as

$$R(v) = \int_0^1 f(\omega) e^{2\pi i v \omega} d\omega$$

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v) e^{-2\pi i v \omega} = R(0) + 2 \sum_{v=1}^{\infty} R(v) \cos 2\pi v \omega.$$

Implications: We are assuming that all inferences about a covariance stationary time series X can be based on making inferences about μ and R. The theorem provides us with a function f that is mathematically equivalent to R. This function is important in its own right in many circumstances. We can also say that one possible realization $\{x(t), t \in \mathcal{Z}\}$ can be thought of as a sum of infinitely many sinusoids where within a particular realization the amplitudes of these sinusoids are fixed, but between possible realizations they vary according to a probability law. These sinusoids are called frequency components and we can think of $f(\omega)$ as being proportional to the average value (over many realizations) of the squared amplitudes of the frequency component of frequency ω.

Ensemble Mean Interpretation of ρ and f

Perhaps the most useful interpretation of ρ and f is as ensemble averages of the sample autocorrelation and sample spectral density functions. It can be shown that under very general conditions,

$$E(\hat{f}(\omega)) \to f(\omega) \qquad \text{and} \qquad E(\hat{\rho}(v)) \to \rho(v)$$

as the length of the sample realization goes to ∞. For example, if we are interested in studying the EEG of a patient, we would recognize that a sample time series consisting of an EEG record would vary from one

time to another in that patient's history and we would be most interested in making inferences about some average behavior of the observed time series. In a situation like this, viewing the spectral density function as the average of the sample spectral density over many long realizations of EEG records is a natural thing to do.

The White Noise Process

The simplest type of time series model is when X consists of uncorrelated random variables having a mean of 0 and a constant variance σ^2.

Definition. *A time series X is said to be a white noise process with variance σ^2 if*

$$E(X(t)) = 0, \qquad t \in Z$$

$$R(v) = \mathrm{Cov}(X(t), X(t+v)) = \begin{cases} \sigma^2, & v = 0 \\ 0, & v \neq 0. \end{cases}$$

Such a process is denoted by $X \sim \mathrm{WN}(\sigma^2)$.

If we define the Kronecker delta function

$$\delta_v = \begin{cases} 1, & v = 0 \\ 0, & v \neq 0, \end{cases}$$

we can write $R(v) = \delta_v \sigma^2$ when $X \sim \mathrm{WN}(\sigma^2)$. Note that

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v) \cos 2\pi v \omega = \sigma^2 \sum_{v=-\infty}^{\infty} \delta_v \cos 2\pi v \omega = \sigma^2;$$

that is, the spectral density of a white noise process is flat (a constant). Thus in analogy with the physical spectrum of white light, a sequence of uncorrelated random variables (which is often used to model physical noise) is referred to as white noise. Processes not having constant spectral densities are often called colored noise.

3.3. The Theory of Linear Filters

We will often express one time series Y as a linear function of the values of another series X. We would like to be able to derive the probabilistic properties of Y from known properties of X.

Definition. *The time series Y is a filtered version of the time series X with filter coefficients $\{\beta_j, j \in Z\}$ if we can write (as a limit in mean square)*

$$Y(t) = \sum_{j=-\infty}^{\infty} \beta_j X(t-j), \qquad t \in Z.$$

In Lecture 2 we defined the moving average smoother

$$Y(t) = \sum_{j=-M}^{M} \frac{1}{2M+1} X(t-j),$$

which is a filtering operation with coefficients

$$\beta_j = \begin{cases} \dfrac{1}{2M+1}, & |j| \le M \\ 0, & |j| > M. \end{cases}$$

Another simple example of filtering a time series is a moving average process (as opposed to the moving average smoother).

Definition. *The time series Y is called a moving average process of order q if*

$$Y(t) = \sum_{k=0}^{q} \beta_k \epsilon(t-k), \qquad t \in \mathcal{Z},$$

where $\beta_0 = 1$ and $\epsilon \sim \mathrm{WN}(\sigma^2)$. We will write $Y \sim \mathrm{MA}(q, \beta, \sigma^2)$ to denote such a series. We will also often write $Y \sim \mathrm{MA}(q)$ to mean that Y is a moving average process of order q without concern for its coefficients β or noise variance σ^2.

We have seen that if $\epsilon \sim \mathrm{WN}(\sigma^2)$, then $R_\epsilon(v) = \delta_v \sigma^2$ and $f_\epsilon(\omega) = \sigma^2$ where we have now put the subscript ϵ on R and f to indicate to which time series they correspond. The following theorem will allow us to easily find R_Y and f_Y for an MA process. In fact, the theorem provides expressions for R_Y and f_Y (as long as they exist) in terms of R_X and f_X whenever Y is a filtered version of X.

Theorem 3.3.1 UNIVARIATE FILTER THEOREM

Suppose that X is a covariance stationary time series with autocovariance function R_X and spectral density function f_X. Suppose that Y is a filtered version of X with filter coefficients β. Then assuming that the quantities involved exist, we have

a) Y is also covariance stationary.

b) The autocovariance function of Y is given by

$$R_Y(v) = \sum_{k=-\infty}^{\infty} R_\beta(k) R_X(v-k), \qquad v \in \mathcal{Z},$$

where

$$R_\beta(k) = \sum_{j=-\infty}^{\infty} \beta_j \beta_{j+|k|}, \qquad k \in \mathcal{Z}.$$

c) The spectral density function of Y is given by

$$f_Y(\omega) = |h(e^{2\pi i \omega})|^2 f_X(\omega),$$

where the function

$$h(z) = \sum_{k=-\infty}^{\infty} \beta_k z^k, \qquad z \in \mathcal{C},$$

325

is called the impulse response function of the filter.

We note that $h(e^{2\pi i \omega})$ is called the frequency transfer function of the filter. We will use the Filter Theorem extensively in the sequel, particularly in the next section where we introduce time series models having a finite number of parameters.

R and f for an MA(q)

Let $Y \sim \text{MA}(q, \beta, \sigma^2)$. Then part (a) of the Filter Theorem allows us to immediately conclude that Y is covariance stationary since it is a filtered version of white noise, which is certainly itself covariance stationary. Further we have

$$f_Y(\omega) = \sigma^2 |h(e^{2\pi i \omega})|^2, \qquad \omega \in [0, 1],$$

where $h(z) = \sum_{k=0}^{q} \beta_k z^k$ is a qth-degree complex valued polynomial. We also have

$$f_Y(\omega) = R_Y(0) + 2 \sum_{v=1}^{q} R_Y(v) \cos 2\pi v \omega.$$

The fact that $R_Y(v) = 0$ for $|v| > q$ is an important characterization of an MA(q) process.

The Effect of Differencing

The Filter Theorem is also important for studying the effects of some transformations. For example, we can see the effect of differencing very clearly. Thus suppose X is a covariance stationary time series with spectral density function f_X. Suppose Y is obtained as the dth difference of X; that is, $Y(t) = X(t) - X(t-d)$. Then Y is covariance stationary and

$$f_Y(\omega) = |1 - e^{2\pi i d \omega}|^2 f_X(\omega),$$

since $h(z) = 1 - z^d$. Thus $f_Y(\omega)$ will be zero anywhere that $e^{2\pi i d \omega}$ is one, namely at any ω such that $d\omega$ is an integer. Thus

$$f_Y(\frac{j}{d}) = 0, \qquad j = 0, 1, \ldots, d.$$

In particular, first differencing makes $f_Y(0) = f_Y(1) = 0$, while 12th differencing makes $f_Y(0) = f_Y(\frac{1}{12}) = \cdots = f_Y(1) = 0$. Thus differencing totally removes frequency components of these frequencies from a time series. In fact, any differencing makes $f_Y(0) = 0$.

What Does the MA Smoother Do?

Another example of using the Filter Theorem to study the effect of transformations is to consider the moving average smoother. Let X be a covariance stationary time series with spectral density f_X and let

$$Y(t) = \sum_{j=-M}^{M} \frac{1}{2M+1} X(t-j).$$

The frequency response function of this filter is

$$h(e^{2\pi i \omega}) = \frac{1}{2M+1} \sum_{j=-M}^{M} e^{2\pi i j \omega} = \frac{1}{2M+1} D_M(\omega),$$

where the function

$$D_M(\omega) = \sum_{j=-M}^{M} e^{2\pi i j \omega}$$
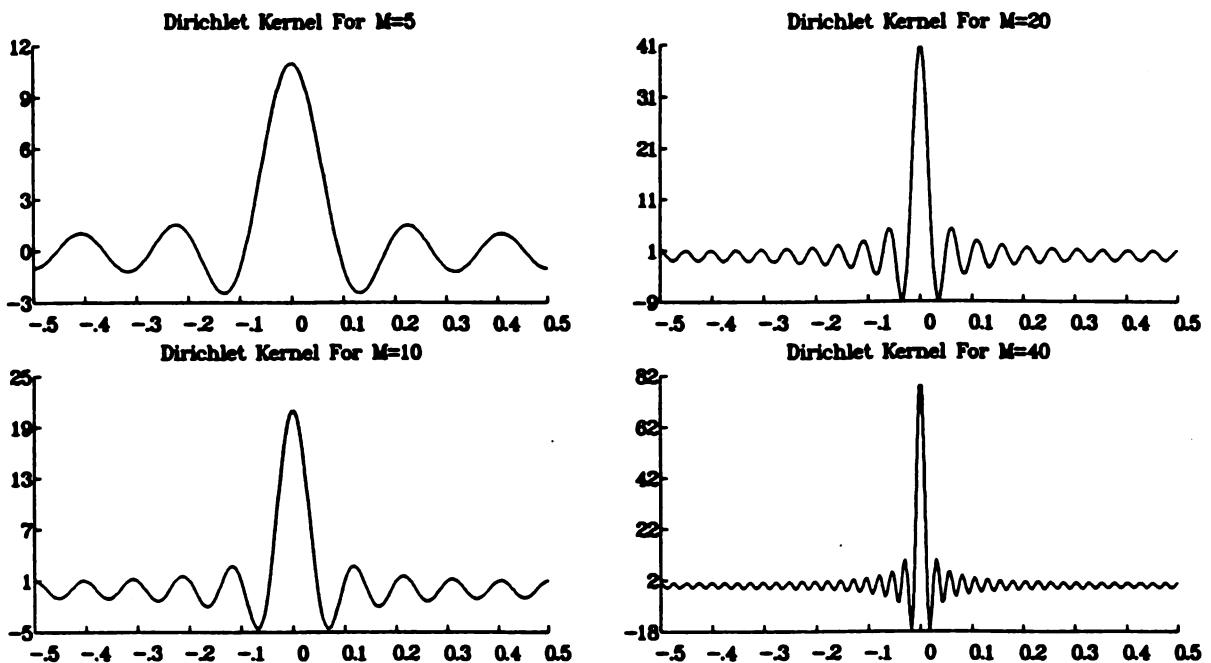
Figure 3.1. The Dirichlet Kernel.

is called the Dirichlet kernel and is well known in many scientific areas. We have

$$D_M(\omega) = \frac{\sin[(M + \frac{1}{2})2\pi\omega]}{\sin \pi\omega}.$$

Figure 3.1 gives graphs of D_M for $M = 5, 10, 20,$ and 40. Note that the kernel becomes more concentrated about zero as M increases. Note also that the kernel is negative for certain frequency ranges and has large "sidelobes," that is, secondary peaks. Thus we have

$$f_Y(\omega) = \left(\frac{1}{2M+1}\right)^2 |D_M(\omega)|^2 f_X(\omega),$$

and D_M becomes more and more concentrated around frequency zero as M (the number of terms on each side of $X(t)$ used in the average) gets large. Thus the moving average smoother is essentially allowing only frequency components of low frequency to be "passed" to Y from X.

Definition. *If Y is a filtered version of X with frequency transfer function $h(e^{2\pi i\omega})$, then the filter is called:*

a) a low (high) pass filter if only low (high) frequency components are passed through the filter, that is, if $f_Y(\omega) = 0$ for $\omega \geq \omega_1$ ($\omega \leq \omega_1$) for some frequency ω_1.

b) a bandpass filter if only frequency components in a certain interval (band) of frequencies are passed through the filter.

Thus the moving average smoother is an example of a low pass filter except that its frequency transfer function never becomes exactly zero for high frequencies. We can now see why $h(e^{2\pi i\omega})$ is called the frequency

transfer function of the filter; it determines what happens to the various frequency components in X as a result of the filter.

The Lag Operator

Linear filters can be succinctly represented if we introduce what is called the lag (or backshift) operator.

Definition. *The lag operator L operating on a time series X is defined by*

$$L^k X(t) = X(t-k), \qquad k \in \mathcal{Z}.$$

If Y is a filtered version of X we define the filter polynomial operator to be

$$h(L) = \sum_{k=-\infty}^{\infty} \beta_k L^k.$$

Thus we can write formally

$$Y(t) = \sum_{k=-\infty}^{\infty} \beta_k X(t-k) = \sum_{k=-\infty}^{\infty} \beta_k L^k X(t) = h(L) X(t).$$

3.4. Time Series Prediction

In this section we describe some of the basic theory for prediction. Several time series models are presented in Section 3.5.

If we have a realization $X(1), \ldots, X(n)$ from a time series X, we often wish to find a function of the data that is close to some future variable $X(n+h)$. We call n the memory or origin of the predictor and h the horizon or number of steps ahead being predicted.

Definition. *Let $\mathbf{X}_n = (X(1), \ldots, X(n))^T$ be a realization of length n from a time series X.*

a) *The best unbiased predictor of $X(n+h)$ given $\mathbf{X}_n$ is that function $\bar{X}_{nh}$ of $\mathbf{X}_n$ that has the same mean as $X(n+h)$ and has smaller prediction error variance than any other unbiased function of $\mathbf{X}_n$.*

b) *The best unbiased linear predictor of $X(n+h)$ given $\mathbf{X}_n$ is that linear function $\hat{X}_{nh}$ of $\mathbf{X}_n$ that has the same mean as $X(n+h)$ and has smaller prediction error variance than any other unbiased linear function of $\mathbf{X}_n$.*

c) *If $\hat{X}_{nh}$ converges in mean square as $n \to \infty$ to a random variable X_{nh}, then X_{nh} is called the infinite memory, h step ahead predictor of $X(n+h)$.*

d) *The error variances of $\bar{X}_{nh}$, $\hat{X}_{nh}$, and X_{nh} are denoted $\bar{\sigma}_{nh}^2$, $\hat{\sigma}_{nh}^2$, and σ_{nh}^2, respectively.*

The following theorem is a straightforward application of material on prediction of random vectors. Again, it is somewhat technical so we have included a discussion after the theorem for those not interested in the details.

Theorem 3.4.1 | UNIVARIATE PREDICTION

Let X be a zero mean time series. Then

a) The best unbiased predictor and its error variance are given by

$$\tilde{X}_{nh} = E\left(X(n+h)|X(1),\ldots,X(n)\right)$$

$$\tilde{\sigma}_{nh}^2 = \mathrm{Var}\left(X(n+h)|X(1),\ldots,X(n)\right).$$

b) If X is covariance stationary with autocovariance function R, then the best unbiased linear predictor and its error variance are given by

$$\hat{X}_{nh} = \lambda_{nh}^T P_n X_n$$

$$\hat{\sigma}_{nh}^2 = R(0) - r_{nh}^T \Gamma_n^{-1} r_{nh},$$

where P_n is the $n \times n$ matrix having zeros except on its main reverse diagonal which is made up of ones, $X_n = (X(1),\ldots,X(n))^T$, and the prediction coefficients

$$\lambda_{nh} = (\lambda_{nh}(1),\ldots,\lambda_{nh}(n))^T$$

satisfy the prediction normal equations

$$\Gamma_n \lambda_{nh} = r_{nh},$$

where $\Gamma_n = \mathrm{Toepl}(R(0),\ldots,R(n-1))$ and $r_{nh} = (R(h),\ldots,R(h+n-1))^T$. Further, these predictors and prediction error variances can be found using conditional means and variances as in part (a) but for a Gaussian time series having the same autocovariance function as X.

c) Let $\lambda_n(j)$ and $\hat{\sigma}_n^2$ denote the coefficients and prediction error variances for the best unbiased linear one step ahead predictor. Then the $\lambda_n(j)$ and $\hat{\sigma}_n^2$ satisfy Levinson's recursion:

$$\lambda_{j+1}(j+1) = \frac{R(j+1) - \sum_{k=1}^{j} \lambda_j(k)R(j+1-k)}{\hat{\sigma}_j^2}$$

$$\lambda_{j+1}(k) = \lambda_j(k) - \lambda_{j+1}(j+1)\lambda_j(j+1-k), \qquad k = 1,\ldots,j$$

$$\hat{\sigma}_{j+1}^2 = \hat{\sigma}_j^2(1 - \lambda_{j+1}^2(j+1)),$$

with $\lambda_1(1) = \rho(1)$ and $\hat{\sigma}_0^2 = R(0)$. Further, for $k > 1$, $\lambda_k(k)$ is equal to the correlation between the errors in predicting $X(t)$ from $X(t+1),\ldots,X(t+k-1)$ and predicting $X(t+k)$ from $X(t+1),\ldots,X(t+k-1)$. Thus $\lambda_k(k)$ is called the partial autocorrelation coefficient of lag k.

d) If X is covariance stationary and

$$\lim_{n\to\infty} \hat{\sigma}_{n1}^2 = \sigma_\infty^2 > 0,$$

then X is said to be purely nondeterministic and

i) There exists a white noise time series ϵ having variance σ_∞^2 and an infinite sequence of constants $\gamma_0 = 1, \gamma_1, \gamma_2, \ldots$ such that, as a limit in mean square,

$$X(t) = \sum_{k=0}^{\infty} \gamma_k \epsilon(t-k).$$

This is called the infinite order moving average representation of X, and the ϵ's are called the innovations of the process.

    ii) The infinite memory predictor X_{nh} exists and is given by, as a limit in mean square,

$$X_{nh} = \sum_{k=h}^{\infty} \gamma_k \epsilon(n+h-k),$$

while

$$\sigma_{nh}^2 = \sigma_\infty^2 \sum_{k=0}^{h-1} \gamma_k^2.$$

Further, for $v \geq 0$,

$$\sigma_{hv} = \mathrm{Cov}(X_{nh}, X_{n,h+v}) = \sigma_\infty^2 \sum_{j=h}^{\infty} \gamma_j \gamma_{j+v} = R(v) - \sigma_\infty^2 \sum_{j=0}^{h-1} \gamma_j \gamma_{j+v}.$$

    iii) A sufficient condition for X to be purely nondeterministic is that X have a spectral density f satisfying

$$S = \int_0^1 \log f(\omega) d\omega > -\infty,$$

in which case

$$\sigma_\infty^2 = e^S.$$

    e) WOLD DECOMPOSITION. Any covariance stationary time series X can be written as

$$X(t) = U(t) + V(t),$$

where U and V are uncorrelated with each other, U is purely nondeterministic, and V is purely deterministic.

Implications: Part (a) says that best predictors and their error variances are just conditional means and variances, while part (b) gives formulas for finding best linear predictors and their error variances. Note the similarity of the prediction normal equations to the normal equations in regression analysis, with $R(0)$ playing the role of $y^T y$, Γ_n playing the role of $X^T X$, and r_{nh} playing the role of $X^T y$. The last part of part (b) provides an often used method for calculating linear predictors, namely the device of pretending that the process is Gaussian and then using conditional means and variances.

    Part (c) of the theorem contains the often used Levinson algorithm. This algorithm allows us to find recursively the prediction coefficients. We will refer to this algorithm frequently. The last part of part (c) shows that for each memory n, the last prediction coefficient is indeed a partial correlation. These partials are used extensively in identifying time series models.

    Part (d) of the theorem provides algorithms for finding infinite memory linear predictors. These are often very easy to calculate. Further, the infinite order MA representation of a process in terms of its innovations is often used in other contexts. Note from part (ii) of (d) that the infinite memory one step ahead prediction error $X(n+1) - X_{n1} = \epsilon(n+1)$, which is the origin of the term *innovation*; that is, $\epsilon(n+1)$ is what is left over after having used the infinite past of X to predict $X(n+1)$. Finally, part (e) gives a way of decomposing any covariance stationary time series into two parts, one that can be perfectly predicted if we know enough of its past, and one part that cannot be perfectly predicted no matter how much of its past we know. Most of the time series models that we consider have only this unpredictable or nondeterministic part.

3.5. Time Series Models

In this section we present some of the models that have been used to represent data.

3.5.1. Random Walk Processes

The first process we discuss in this section is in fact not even covariance stationary, but occurs very frequently in the physical and economic sciences.

Definition. *Suppose*

$$X(t) = X(t-1) + \epsilon(t), \qquad t \geq 1,$$

where $\epsilon \sim \text{WN}(\sigma^2)$. Then X is called a random walk process and we write $X \sim \text{RW}(\sigma^2)$.

Actually a random walk process is not fully specified until the characteristics of the starting value $X(0)$ are given. We usually assume that $X(0)$ is a random variable that is uncorrelated with any of the ϵ's.

Theorem 3.5.1 | PROPERTIES OF RANDOM WALKS

Suppose $X \sim \text{RW}(\sigma^2)$ with

$$\text{E}(X(0)) = \mu_X, \quad \text{Var}(X(0)) = \sigma_X^2, \quad \text{Var}(\epsilon(t)) = \sigma_\epsilon^2, \quad \text{Cov}(X(0), \epsilon(t)) = 0.$$

Then

$$\text{E}(X(t)) = \mu_X \qquad and \qquad \text{Var}(X(t)) = \sigma_X^2 + t\sigma_\epsilon^2, \qquad t \geq 1.$$

Note that X is not covariance stationary since $\text{Var}(X(t))$ is not independent of t. Further, $\text{Var}(X(t)) \to \infty$ as $t \to \infty$. Figure 3.2 gives five realizations of length 200 from a Gaussian random walk process; that is, $X(0)$ and $\epsilon(1), \ldots, \epsilon(200)$ are iid $N(0,1)$ variables. As time progresses, the realizations get increasingly far apart. This is expected since $\text{Var}(X(t))$ is increasing linearly without bound as t increases. Note also that these realizations are similar in appearance to many price time series in economics such as stock market data. Finally, note that the first difference of a random walk process is a white noise process.

Prediction of Random Walks

If $X \sim \text{RW}(\sigma^2)$, then $\bar{X}_{n1} = X(n)$.

3.5.2. Moving Average Processes

In our discussion of linear filters above, we introduced the moving average process:

$$X(t) = \sum_{k=0}^{q} \beta_k \epsilon(t-k), \qquad \beta_0 = 1, \qquad t \in Z,$$

where $\epsilon \sim \text{WN}(\sigma^2)$, and showed that its autocovariance function R and spectral density function f are given by

$$R(v) = \begin{cases} \sigma^2 \sum_{k=0}^{q-|v|} \beta_k \beta_{k+|v|}, & |v| \leq q \\ 0, & |v| > q \end{cases}, \qquad f(\omega) = \sigma^2 |h(e^{2\pi i \omega})|^2, \qquad \omega \in [0,1],$$
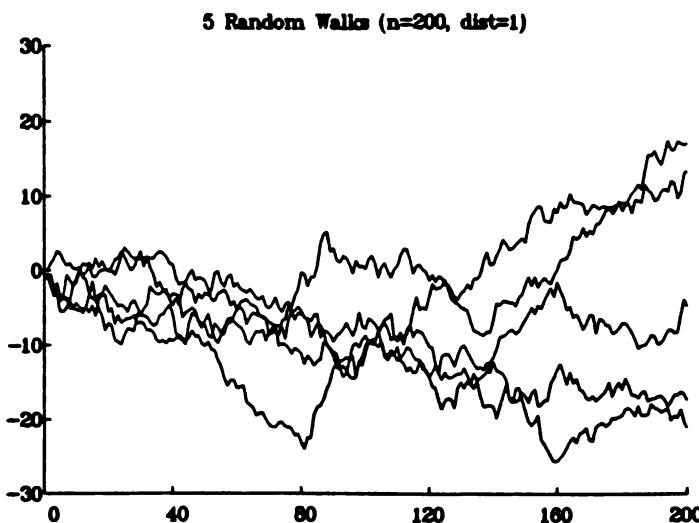
Figure 3.2. Five Realizations from a Gaussian Random Walk Process.

where the complex valued polynomial h is given by

$$h(z) = \sum_{k=0}^{q} \beta_k z^k.$$

If we write the process in operator notation as

$$X(t) = h(L)\epsilon(t),$$

it is of interest to determine whether we can invert the operator; that is, can we write in any sense $\epsilon(t) = (1/h(L))X(t)$? The answer is given by the following theorem.

| Theorem 3.5.2 | INVERTIBILITY OF MA PROCESSES |

Suppose $X \sim$ MA(q, β, σ^2). If the zeros of h are all greater than one in modulus, then

a) $f(\omega) > 0, \quad \omega \in [0, 1]$.

b) We have $\epsilon(t) = \sum_{j=0}^{\infty} \alpha_j X(t-j), \quad \omega \in [0, 1]$, where the α's are the coefficients of the polynomial $g(z) = 1/h(z) = \sum_{j=0}^{\infty} \alpha_j z^j$ and are given by $\alpha_0 = 1$ and $\alpha_j = -\sum_{i=1}^{\min(q,i)} \beta_i \alpha_{j-i}, \quad j \geq 1$.

We can now formalize our definition of an invertible MA process.

332

Definition. *An $MA(q, \beta, \sigma^2)$ process is said to be invertible if all of the zeros of its characteristic polynomial h are greater than one in modulus.*

Prediction for MA Processes

For a Gaussian MA process,

$$\tilde{X}_{nh} = \mathrm{E}\,(X(n+h)|\mathbf{X}_n) = \sum_{k=0}^{q} \beta_k \mathrm{E}\,(\epsilon(n+h-k)|\mathbf{X}_n) = 0, \qquad h > q,$$

since $\epsilon(t)$ is independent of $X(s)$ for $t > s$. Thus by the device described in part (b) of Theorem 3.4.1, we can write the following recursion for the best linear predictors:

$$\hat{X}_{nh} = \begin{cases} \sum_{k=0}^{q} \beta_k \hat{X}_{n,h-k}, & h = 1, \ldots, q \\ 0, & h > q. \end{cases}$$

Partial Autocorrelations for MA Processes

In Theorem 3.4.1 we saw that the partial autocorrelation of lag v of a time series X is the last coefficient $\lambda_v(v)$ in the best v step ahead linear predictor. Note that unlike the ordinary autocorrelations which become identically zero for lags greater than the order q, the partial autocorrelations of an MA process decay to zero exponentially.

MA Spectra and Trigonometric Polynomials

Before leaving MA processes, we note that since $R(v) = 0$ for $|v| > q$, we can write from Theorem 3.2.1,

$$f(\omega) = R(0) + 2\sum_{v=1}^{q} R(v) \cos 2\pi v\omega.$$

Now we can also write $\cos 2\pi v\omega$ as a polynomial of degree v in $\cos 2\pi\omega$ since for $v \geq 2$, we have the important trigonometric identity

$$\cos v\theta = 2\cos\theta \cos(v-1)\theta - \cos(v-2)\theta,$$

which if used recursively, ultimately expresses $\cos v\theta$ as a polynomial in $\cos\theta$. Thus the spectral density of an MA(q) can be written as a qth-degree trigonometric polynomial. The above identity is also important in other contexts in time series analysis. If we write $z(v) = \cos v\theta$, we have

$$z(v) - 2\cos\theta\, z(v-1) + z(v-2) = 0, \qquad v \geq 2,$$

with $z(0) = 1$ and $z(1) = \cos\theta$, which is a second-order difference equation with initial conditions $z(0)$ and $z(1)$. This equation has solution $z(v) = \cos v\theta$.

Examples of MA Processes

In this section we have seen:

1. The autocorrelation function for an MA(q) process is identically zero for lags greater than the order q.

2. The partial autocorrelation function decays to zero as v increases.

3. The spectral density function of an MA(q) process is a qth-degree trigonometric polynomial. Thus for small values of q, it is difficult for the spectral density to have sharp peaks.

In Figure 3.3 we display the correlogram, partial correlogram, spectral density, and one realization of length 200 from each of two MA processes. The two processes are:

$$\text{Model 1:} \qquad X(t) = \epsilon(t) - 0.70\epsilon(t-) - 0.10\epsilon(t-2) + 0.60\epsilon(t-3)$$

$$\text{Model 2:} \qquad X(t) = \epsilon(t) + 0.80\epsilon(t-4);$$

that is, the first process is an MA(3) process while the second is a (subset) MA(4). In both cases, the correlogram is zero for lags greater than the order of the process, while the partial correlogram is not. The partials for the subset model are zero except for lags that are multiples of four.

3.5.3. Autoregressive Processes

Perhaps the most often used time series model in practice is the autoregressive process. Intuitively, X is an autoregressive process of order p with coefficients $\alpha_1, \ldots, \alpha_p$ and noise variance σ^2 if it can be transformed to white noise by a filter of length p; that is,

$$X(t) + \alpha_1 X(t-1) + \cdots + \alpha_p X(t-p) = \epsilon(t), \qquad t \in Z,$$

where $\epsilon \sim WN(\sigma^2)$. One appeal of this process is that it is in the form of a regression model

$$X(t) = -\alpha_1 X(t-1) - \cdots - \alpha_p X(t-p) + \epsilon(t),$$

with past (or "lagged") values of X as the independent variables in the regression.

It is important to note that the autoregressive process is not defined explicitly, but rather as the solution to a stochastic difference equation, that is, a difference equation whose terms are random variables. The difficulty with this is in determining whether there is in fact a covariance stationary time series that satisfies such an equation. To illustrate this, suppose that X satisfies the difference equation with $p = 1$:

$$X(t) + \alpha X(t-1) = \epsilon(t).$$

Successively substituting for $X(t-1), X(t-2), \ldots$ gives

$$X(t) - \sum_{j=0}^{K-1} (-\alpha)^j \epsilon(t-j) = -(\alpha)^K X(t-k).$$

If $|\alpha| < 1$, then taking the limit of the expected value of the square of both sides of this results in

$$X(t) = \sum_{j=0}^{\infty} \beta_j \epsilon(t-j),$$

where $\beta_j = (-\alpha)^j$.

Now we could also write $\frac{1}{\alpha} X(t+1) + X(t) = \frac{1}{\alpha} \epsilon(t+1)$, which gives upon successive substitution

$$X(t) = -\frac{1}{\alpha}[X(t+1) + \epsilon(t+1)]$$

$$= -\frac{1}{\alpha}[-\frac{1}{\alpha}(X(t+2) + \epsilon(t+2)) + \epsilon(t+1)]$$

$$= \cdots$$

$$= (-\frac{1}{\alpha})^K X(t+K) + \sum_{j=1}^{K} (-\frac{1}{\alpha})^j \epsilon(t+j),$$
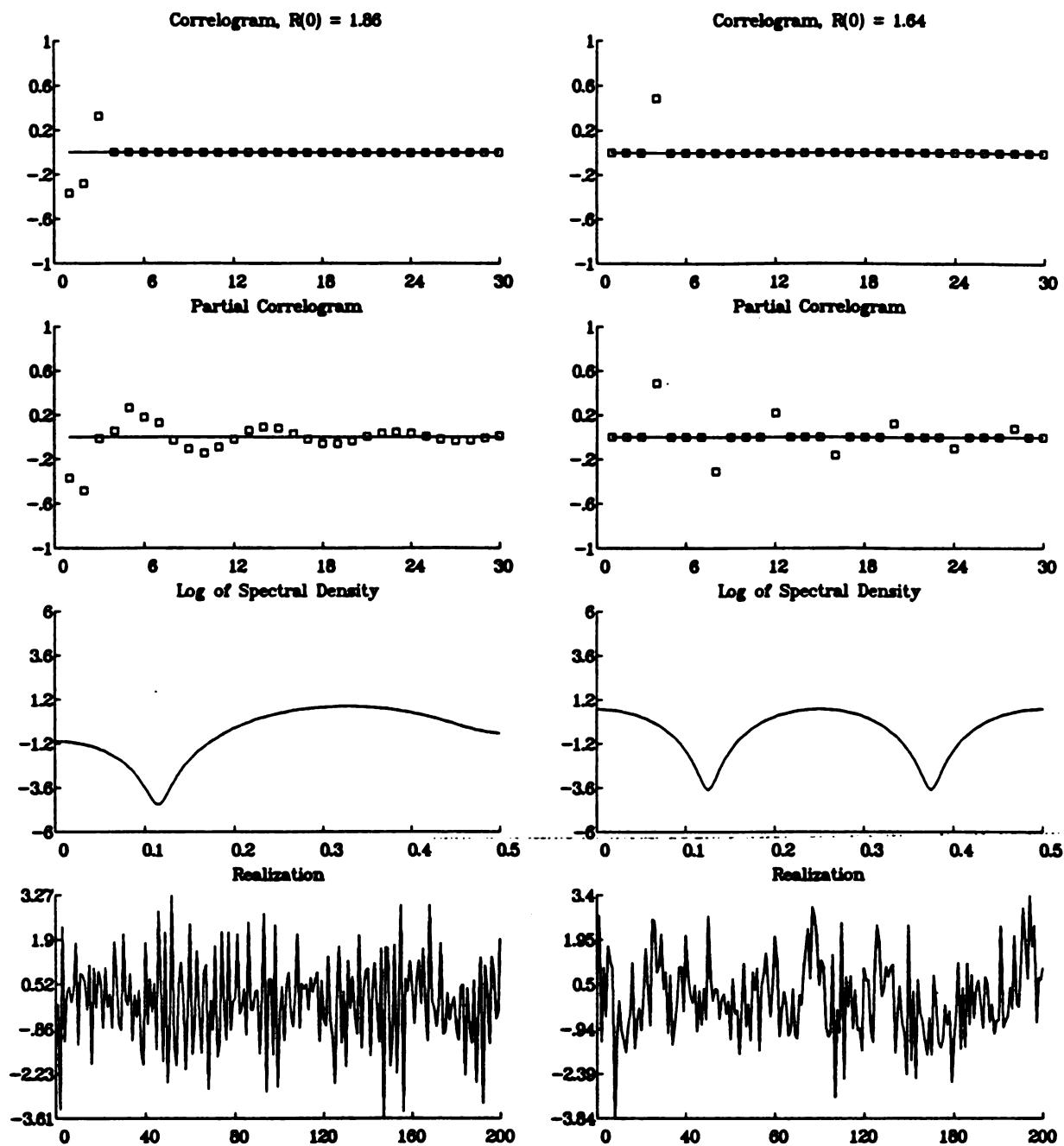
Figure 3.3. Correlogram, Partial Correlogram, Spectral Density, and a Realization of Length 200 for Each of Two MA Processes.

which, if $|\alpha| > 1$, gives as a limit in mean square

$$X(t) = \sum_{j=1}^{\infty} \beta_j \epsilon(t + j),$$

where $\beta_j = (-1/\alpha)^j$. If $|\alpha| = 1$ we are not able to do this process as neither $(-\alpha)^K$ nor $(-1/\alpha)^K$ goes to zero.

These arguments can be extended to $p > 1$ where instead of the location of $|\alpha|$ determining whether to write $X(t)$ as a function of past and present or future ϵ's, the location of the zeros of the complex valued polynomial

$$g(z) = \sum_{j=0}^{p} \alpha_j z^j, \qquad \alpha_0 = 1,$$

determines the representation. The following theorem summarizes the basic results.

Theorem 3.5.3 PROPERTIES OF AUTOREGRESSIVE PROCESSES

Let $\alpha_0 = 1$ and $\alpha_1, \ldots, \alpha_p$ be a set of real constants, and define the complex valued polynomial (called the characteristic polynomial of the stochastic difference equation) $g(z) = \sum_{j=0}^{p} \alpha_j z^j$. Let $\epsilon \sim \mathrm{WN}(\sigma^2)$. Then

a) If none of the zeros of g are equal to one in modulus, then

$$X(t) = \sum_{j=-\infty}^{\infty} \beta_j \epsilon(t - j), \qquad t \in Z,$$

exists as a limit in mean square where the β's are the coefficients of the polynomial

$$h(z) = \frac{1}{g(z)}.$$

This representation is called the (doubly) infinite order moving average representation of X. Further, X satisfies the autoregressive difference equation, is covariance stationary, has spectral density

$$f(\omega) = \sigma^2 \frac{1}{|g(e^{2\pi i \omega})|^2} = \sigma^2 |h(e^{2\pi i \omega})|^2,$$

and has autocovariance function R satisfying

$$\sum_{j=0}^{p} \alpha_j R(j - v) = \sigma^2 \beta_{-v}, \qquad v \in Z.$$

b) If the zeros of g are all less than one in modulus, then $\beta_j = 0$ for $j \geq 0$; that is, $X(t)$ can be expressed as a function of ϵ's at only future times.

c) If the zeros of g are all greater than one in modulus, then $\beta_j = 0$ for $j < 0$; that is, $X(t)$ can be expressed as a function of ϵ's at only the present and past times. Thus the equations relating the R and the α's can be written as

$$\sum_{j=0}^{p} \alpha_j R(j - v) = \delta_v \sigma^2, \qquad v \geq 0,$$

which are called the Yule-Walker equations.

From this theorem we can see that as long as none of the zeros of g are on the unit circle (that is, equal to one in modulus), the infinite order moving average process provides a covariance stationary solution to the stochastic difference equation and has the properties given by the theorem. If all of the zeros of g are outside the unit circle, then a stationary solution can be explicitly defined.

Since in practice we do not observe the future, we would like to define an autoregressive process so that $X(t)$ is only a function of $\{\epsilon(s), s \leq t\}$. We formalize this in the following definition.

Definition. *A time series X is said to be an autoregressive process of order p with coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ and noise variance σ^2 if X satisfies the stochastic difference equation*

$$X(t) + \alpha_1 X(t-1) + \cdots + \alpha_p X(t-p) = \epsilon(t)$$

where the zeros of $g(z) = \sum_{j=0}^{p} \alpha_j z^j$ are all greater than one in modulus. We denote such a time series by $X \sim \mathrm{AR}(p, \alpha, \sigma^2)$.

Recall for an MA(q) process that $R(v) = 0$ for $|v| > q$; that is, as soon as the lag v gets larger than the order of the process, $X(t)$ and $X(t+v)$ are no longer correlated. The autoregressive process has a correlation function that often gives a more realistic description of how $X(t)$ and $X(t+v)$ become uncorrelated as v increases. We will illustrate this by considering the AR(1) process. For an AR(1) process having coefficient α and error variance σ^2, we can write the Yule-Walker equations for $v = 0$ and $v = 1$ as

$$R(0) + \alpha R(1) = \sigma^2$$

$$R(1) + \alpha R(0) = 0,$$

since $R(1) = R(-1)$. From the second equation we have $R(1) = -\alpha R(0)$, which when substituted into the first equation gives

$$R(0) = \frac{\sigma^2}{1 - \alpha^2}.$$

From the Yule-Walker equations for $v > 0$, we have

$$R(v) = -\alpha R(v-1) = (-\alpha)^2 R(v-2) = \cdots = (-\alpha)^v R(0),$$

which gives

$$R(v) = \frac{(-\alpha)^v \sigma^2}{1 - \alpha^2}, \qquad \rho(v) = \frac{R(v)}{R(0)} = (-\alpha)^v, \qquad v \geq 0.$$

Finally, since $R(v) = R(-v)$, we have

$$R(v) = \frac{(-\alpha)^{|v|} \sigma^2}{1 - \alpha^2}, \qquad \rho(v) = (-\alpha)^{|v|}, \qquad v \in \mathcal{Z}.$$

Note that a sequence $\{a(v), v \geq 0\}$ is said to decay exponentially to zero if we can write

$$|a(v)| < c\delta^v,$$

where $|\delta| < 1$. Thus the autocovariance function (and hence the autocorrelation function) of an AR(1) process decays exponentially to zero. The rate of decay depends on how close α is to one.

337

Theorem 3.5.4 | EXPONENTIAL DECAY OF AR CORRELATIONS

Suppose $X \sim \text{AR}(p, \alpha, \sigma^2)$ and let R be the autocovariance function of X. Let $z_1, \ldots, z_p$ be the zeros of $g(z)$. Then

$$|R(v)| < \kappa \left(\max_i \frac{1}{|z_i|} \right)^v$$

for some constant $\kappa > 0$.

Note that the theorem says that $R(v)$ is an exponentially decaying function of the reciprocal of the modulus of the zero of g that is closest to the unit circle. If a zero is close to the unit circle, $R(v)$ may decay very slowly.

Autoregressive processes are very popular also because the estimation of their parameters and the forecasting of future values are rather easy from a computational point of view.

Prediction for AR Processes

Prediction is very simple for AR processes as we can write for a Gaussian AR(p) process

$$\tilde{X}_{nh} = \text{E}(X(n+h)|\mathbf{X}_n) = -\sum_{j=1}^{p} \alpha_j \tilde{X}_{n,h-j}.$$

Thus the AR predictors satisfy the homogeneous difference equation

$$\sum_{j=0}^{p} \alpha_j \tilde{X}_{n,h-j} = 0.$$

Further

$$\text{Var}\left(\tilde{X}_{n,h} - X(n+h) \right) = \sigma^2 \sum_{k=0}^{h-1} \beta_k^2,$$

where the β's are the coefficients of the MA(∞) representation of X.

From these facts, we also obtain an important fact about the partial autocorrelations θ of an AR(p) process, namely that for any lag v greater than the order p, we have that $\theta_v = 0$.

Examples of AR Processes

We have now seen that for an AR(p) process:

1. The autocorrelation function satisfies a difference equation of order p (namely the Yule-Walker Equations) and thus, since the zeros of the characteristic polynomial are all outside of the unit circle, the correlogram decays exponentially to zero.

2. The partial autocorrelation function is identically zero for lags greater than p.

3. The spectral density is the reciprocal of a pth degree trigonometric polynomial. This means that it is capable of having very sharp peaks.

In Figure 3.4 we give the correlogram, partial correlogram, spectral density, and a realization of length 200 for each of the two AR models:

$$\text{Model 1:} \quad X(t) - 0.80X(t-1) + 0.40X(t-2) = \epsilon(t)$$

$$\text{Model 2:} \quad X(t) - 0.90X(t-1) + 0.70X(t-2) = \epsilon(t).$$

Note how the qualitative features of the data correspond to the spectra and correlations. In particular, notice that the correlograms are of two types. In Series 1, the correlogram decays relatively rapidly to 0, while in Series 2, the correlogram follows a sinusoidal decay.

3.5.4. Autoregressive-Moving Average Processes

In the previous two sections we studied MA and AR processes. In this section we consider the autoregressive-moving average (ARMA) process which in a variety of ways is a combination of the AR and MA processes.

Definition. *A time series X is said to be an autoregressive-moving average process of orders p and q, coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ and $\beta = (\beta_1, \ldots, \beta_q)^T$, and noise variance σ^2 if X satisfies*

$$\sum_{j=0}^{p} \alpha_j X(t-j) = \sum_{k=0}^{q} \beta_k \epsilon(t-k), \qquad t \in \mathbb{Z},$$

where $\epsilon \sim \mathrm{WN}(\sigma^2)$ and the zeros of the complex polynomial $g(z) = \sum_{j=0}^{p} \alpha_j z^j$ are all outside the unit circle. We denote such a time series by

$$X \sim \mathrm{ARMA}(p, q, \alpha, \beta, \sigma^2).$$

If the zeros of the complex polynomial $h(z) = \sum_{k=0}^{q} \beta_k z^k$ are all outside the unit circle, then we say that X is invertible.

The operators $g(L)$ and $h(L)$ are called the AR and MA operators, respectively. Using arguments similar to those given in the previous two sections, we can show that an ARMA process has the MA(∞) representation

$$X(t) = \sum_{k=0}^{\infty} \gamma_k \epsilon(t-k), \qquad t \in \mathbb{Z},$$

where the γ_k's are the coefficients of the polynomial $h(z)/g(z)$, while if X is invertible, it has the AR(∞) representation

$$\sum_{j=0}^{\infty} a_j X(t-j) = \epsilon(t), \qquad t \in \mathbb{Z},$$

where the a_j's are the coefficients of the polynomial $g(z)/h(z)$. Thus we are able to justify the idea of writing $g(L)X(t) = h(L)\epsilon(t)$ as

$$X(t) = \frac{h(L)}{g(L)} \epsilon(t)$$

or

$$\frac{g(L)}{h(L)} X(t) = \epsilon(t).$$

In the next theorem, we express the autocovariance and spectral density functions of an ARMA process in terms of its parameters.

| Theorem 3.5.6 | R AND f FOR AN ARMA PROCESS |

If

$$X \sim \mathrm{ARMA}(p, q, \alpha, \beta, \sigma^2),$$

then
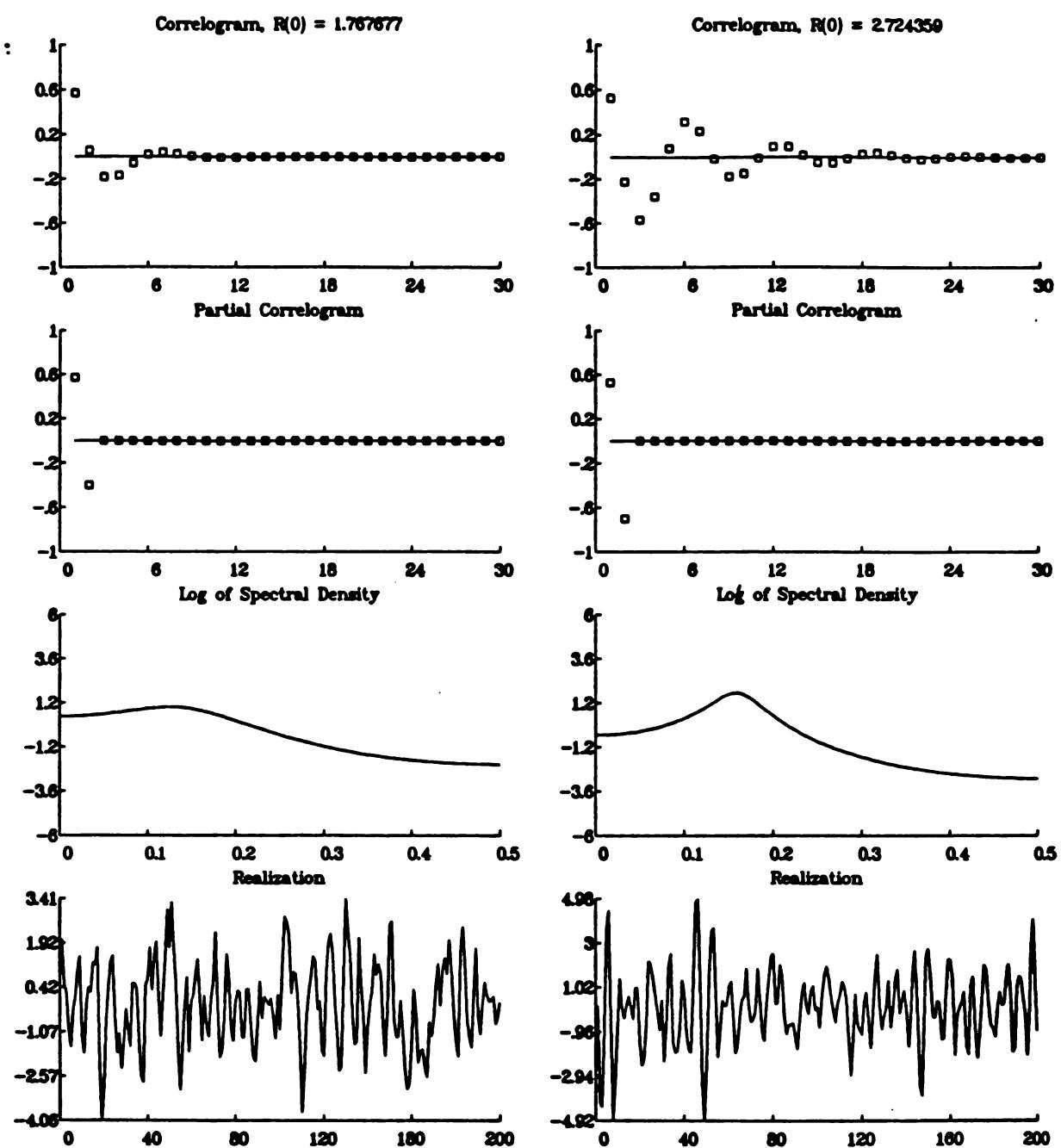
Figure 3.4. Correlogram, Partial Correlogram, Spectral Density, and a Realization of Length 200 for Each of Two AR Processes.

Table 3.1. Nature of R and f for ARMA Processes

Process	$R(v)$ or $\rho(v)$	$f(\omega)$
MA(q)	0 for $|v| > q$	qth-degree trigonometric polynomial
AR(p)	exponential decay for $v > 0$	reciprocal of pth-degree trigonometric polynomial
ARMA(p, q)	exponential decay for $v > q$	(q, p)th-degree rational trigonometric polynomial

a) $f(\omega) = \sigma^2 \dfrac{|h(e^{2\pi i \omega})|^2}{|g(e^{2\pi i \omega})|^2}$.

b) The autocovariances R satisfy

$$\sum_{j=0}^{p} \alpha_j R(j - v) = \sum_{k=0}^{q} \beta_k R_{X\epsilon}(v - k), \qquad v \geq 0$$

$$= 0, \qquad v > q,$$

where the cross-covariances $R_{X\epsilon}$ are given by

$$R_{X\epsilon}(v) = \operatorname{Cov}(X(t), \epsilon(t + v)) = \begin{cases} 0, & v > 0 \\ \sigma^2 \gamma_{-v}, & v \leq 0 \end{cases}$$

and the γ_v's are the coefficients of the MA(∞) representation of X.

Thus R satisfies a homogeneous difference equation of order p for $v > q$. The equations for $v = q + 1, \ldots, q + p$ are often called the high order Yule-Walker equations. We can now summarize the nature of R (or ρ) and f for AR, MA, and ARMA processes as in Table 3.1.

Examples of ARMA Processes

As we have seen, the ARMA process is a combination of the AR and MA processes. The basic characteristics of an ARMA model are:

1. Neither the correlogram nor the partial correlogram is identically zero past some lag; rather they each decay exponentially past some lag.

2. The spectral density function is the ratio of a qth-degree trigonometric polynomial to a pth-degree trigonometric polynomial. Thus it can have sharp peaks and/or sharp troughs.

In Figure 3.5 we give the correlogram, partial correlogram, spectral density, and a realization of length 200 from each of the two ARMA models:

1 : $X(t) - 1.20X(t - 1) + 0.78X(t - 2) = \epsilon(t) + 0.45\epsilon(t - 1) + 0.39\epsilon(t - 2)$

2 : $X(t) - 0.90X(t - 1) = \epsilon(t) + 0.80\epsilon(t - 4).$

That is, Model 1 is an ARMA(2,2), while Model 2 is an ARMA(1,4) where the first three MA coefficients are zero. In each of these models, both the correlogram and the partial correlogram exhibit an exponential decay, thus ruling out a pure AR or pure MA model.

Correlogram, R(0) = 10.39489

Correlogram, R(0) = 14.15663

Partial Correlogram

Partial Correlogram

Log of Spectral Density

Log of Spectral Density

Realization
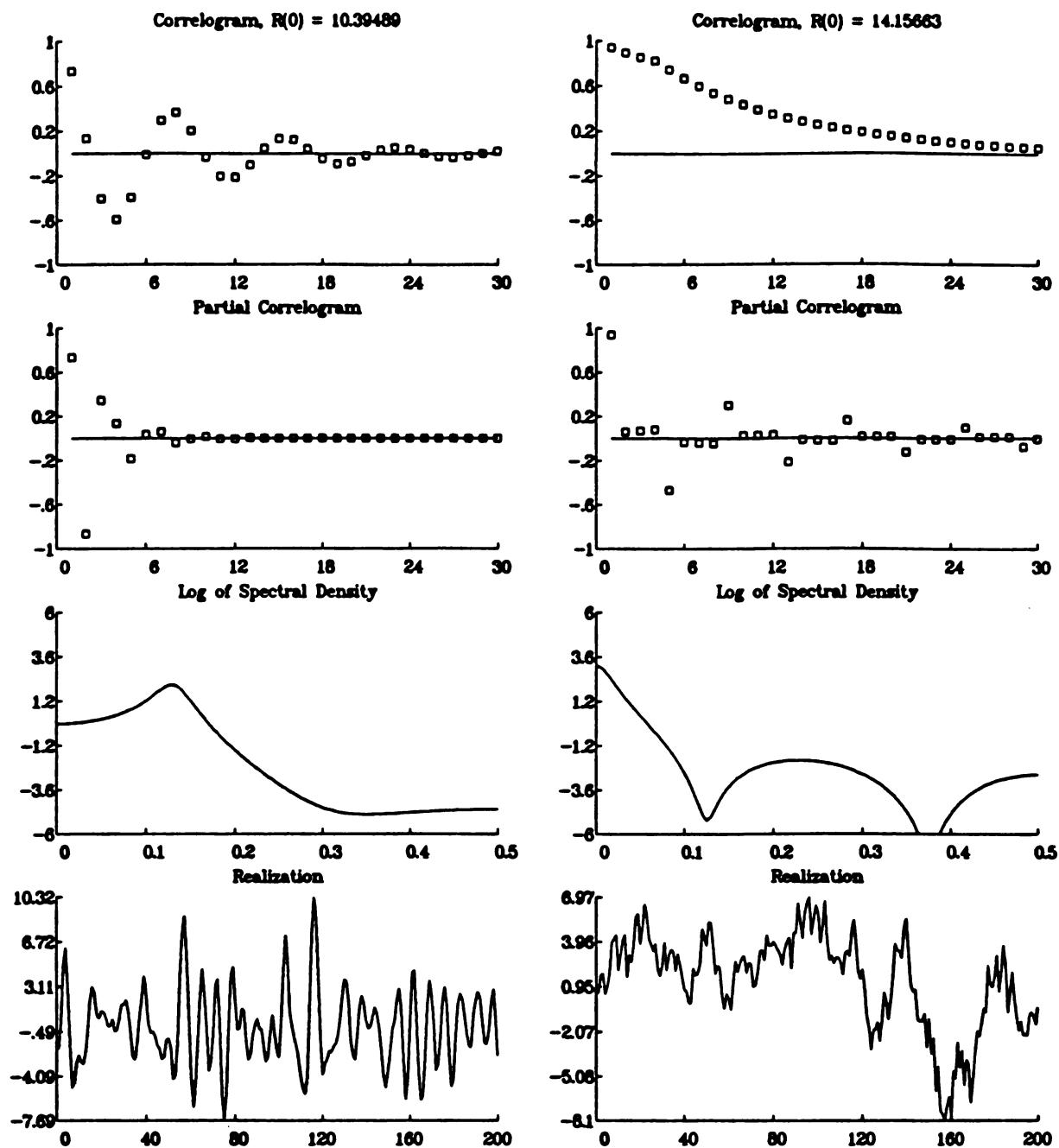
Realization

Figure 3.5. Correlogram, Partial Correlogram, Spectral Density, and a Realization of Length 200 from Two ARMA Processes.

3.5.5. Subset and Multiplicative Subset ARMA Processes

A natural extension of the ARMA model is the case where only a few of its coefficients are nonzero; that is, the subset ARMA models which we write as

$$\left(\sum_{k=0}^{P} \theta_k L^{u_k}\right) X(t) = \left(\sum_{m=0}^{Q} \gamma_m L^{v_m}\right) \epsilon(t),$$

where $u_1 \leq \cdots \leq u_P = p$ and $v_1 \leq \cdots \leq v_Q = q$ are called the AR and MA lags, respectively. We can go a step further and define a multiplicative subset ARMA process as a combination of a "full" ARMA and subset ARMA by

$$\left(\sum_{j=0}^{p} \alpha_j L^j\right)\left(\sum_{k=0}^{P} \theta_k L^{u_k}\right) X(t) = \left(\sum_{l=0}^{q} \beta_l L^l\right)\left(\sum_{m=0}^{Q} \gamma_m L^{v_m}\right) \epsilon(t).$$

Such models will be important later. Note that as long as the zeros of each of the four polynomials involved in the model are all outside the unit circle, then we can express the multiplicative subset ARMA model as an ordinary ARMA model of orders $p + u_P$ and $q + v_Q$, and use all of the results that we have obtained for ARMA processes.

3.5.6. ARIMA Processes

Many of the processes in economics and business are nonstationary but can be transformed to stationarity by differencing. This fact has led to the popularity of the so-called Box-Jenkins method of time series analysis. The method employs extensively what is called an ARIMA model.

Definition. *A time series X is called an autoregressive integrated moving average process of orders (p, d, q) if the series $Z(t) = (1 - L)^d X(t)$ is an ARMA(p,q) process. We denote such a process by*

$$X \sim ARIMA(p, d, q, \alpha, \beta, \sigma^2).$$

The simplest example of such a model is the random walk process where $d = 1$ and $p = q = 0$. Note that if we let g and h denote the AR and MA operators for the ARMA part of the model, then we have that X satisfies

$$g(L)(1 - L)^d X(t) = h(L)\epsilon(t).$$

Multiplicative Subset Seasonal ARIMA Models

A general model of the form

$$g(L)G(L)(1 - L)^d (1 - L^S)^D X(t) = h(L)H(L)\epsilon(t),$$

where g, G, h, and H are the four operators in the multiplicative subset ARMA model, provides a general framework within which many seasonal time series can be analyzed. If the AR and MA lags are multiples of the seasonality factor S, then this model is the one which is used in the Box-Jenkins forecasting method.

LECTURE 4

Estimation and Model Identification

In this lecture we describe some of the basic statistical inferences that can be made about univariate time series. In Section 4.1 we give the sampling properties of the descriptive statistics that were introduced in Lecture 1. In Section 4.2 we describe the basis of the two tests for white noise we used in Lecture 1. In Sections 4.3 and 4.4 we consider respectively the questions of estimating the parameters of an ARMA model and identifying the orders of the model.

4.1. Sampling Properties of Descriptive Statistics

In Lecture 1 we introduced the statistics

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X(t)$$

$$\hat{R}(v) = \frac{1}{n} \sum_{t=1}^{n-|v|} (X(t) - \bar{X})(X(t + |v|) - \bar{X}), \qquad |v| < n$$

$$\hat{\rho}(v) = \frac{\hat{R}(v)}{\hat{R}(0)}, \qquad |v| < n$$

$$\hat{f}(\omega) = \frac{1}{n} \Big| \sum_{t=1}^{n} (X(t) - \bar{X}) e^{2\pi i (t-1)\omega} \Big|^2, \qquad \omega \in [0, 1],$$

as well as the sample partial autocorrelation function. We will now use upper case letters to represent time series observations, as they are considered to be random variables, whereas in Lecture 1 we were treating them primarily as sets of numbers.

In this section we state a series of theorems summarizing the sampling properties (mean, variance, and asymptotic distribution) of $\bar{X}$, $\hat{R}$, $\hat{\rho}$, and $\hat{f}$. We will investigate the sampling properties of the sample partial autocorrelation function later. At the end of each theorem, we give summaries of the their implications.

4.1.1. The Sample Mean

| Theorem 4.1.1 | SAMPLING PROPERTIES OF $\bar{X}$ |

Let X be a covariance stationary time series with mean μ and autocovariance function R. Let $\mathbf{X}_n = (X(1), \ldots, X(n))^T$ be a realization of length n from X and let $\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X(t)$. Then

a) $E(\bar{X}_n) = \mu$.

b) $\text{Var}(\bar{X}_n) \to 0$ as $n \to \infty$ if and only if $\text{Cov}(X(n), \bar{X}_n) \to 0$ as $n \to \infty$, a sufficient condition for which is that $R(n) \to 0$ as $n \to \infty$.

c) If $\sum_{v=-\infty}^{\infty} R(v) < \infty$, then $\lim_{n \to \infty} n\text{Var}(\bar{X}_n) = \sum_{v=-\infty}^{\infty} R(v)$,

d) Under mild assumptions, $\sqrt{n}(\bar{X}_n - \mu) \overset{\mathcal{L}}{\longrightarrow} N(0, f(0))$.

Implications: Part (a) says that $\bar{X}_n$ is an unbiased estimator of μ. A sufficient condition for an estimator to be consistent is that it is asymptotically unbiased (that is, has expectation converging to the parameter being estimated) and has variance converging to zero. Thus part (b) gives a condition for $\bar{X}_n$ to be consistent. A time series is also said to be ergodic in the mean if $\text{Var}(\bar{X}_n) \to 0$ as $n \to \infty$, that is, if the average $\bar{X}_n$ over time of a single realization approaches the average $\mu = E(X(t))$ at a single time t of the ensemble of all possible realizations. Thus X is ergodic in the mean if and only if as n gets large, adding new observations in the calculation of $\bar{X}_n$ has no effect on it in the sense that $\text{Cov}(X(n), \bar{X}_n) \to 0$.

Part (c) gives two expressions that can be used to approximate $\text{Var}(\bar{X}_n)$ and also shows that $\text{Var}(\bar{X}_n)$ goes to zero at the rate of $1/n$. Finally, part (d) provides a general Central Limit Theorem for $\bar{X}_n$. This gives

$$\bar{X}_n \pm Z_{\alpha/2}\sqrt{\frac{f(0)}{n}}$$

as a $100(1-\alpha)\%$ confidence interval for μ. We will see in Lecture 6 that under the conditions of part (d), we can find an estimator $\tilde{f}(0)$ such that

$$\bar{X}_n \pm Z_{\alpha/2}\sqrt{\frac{\tilde{f}(0)}{n}}$$

is also a $100(1-\alpha)\%$ large sample confidence interval for μ.

Equivalent Number of Uncorrelated Observations

To illustrate the results of Theorem 4.1.1, suppose that X is an AR(1) process with coefficient α, noise variance σ^2, and mean μ. Then

$$f(0) = \frac{R(0)(1-\rho^2)}{(1-\rho)^2} = R(0)\frac{1+\rho}{1-\rho}.$$

Thus a 95% confidence interval for μ is given by

$$\bar{X}_n \pm 1.96\sqrt{\frac{R(0)}{n}\frac{1+\rho}{1-\rho}}.$$

This expression allows us to introduce the idea of an equivalent number of uncorrelated observations. Recall that a 95% confidence interval for the mean of a population having variance $R(0)$ based on a random sample of size N is given by

$$\bar{X} \pm 1.96\sqrt{\frac{R(0)}{N}}.$$

345

Thus to make sure that the confidence interval from the AR(1) process and the random sample have the same width, we would need

$$N = n\frac{1-\rho}{1+\rho}.$$

For example, a sample of size $n = 100$ from an AR(1) process having $\rho = .8$ is equivalent in this sense to a random sample of size 11 from a population having the same variance as that of the AR(1) process. On the other hand, if ρ is negative, there is actually more information about μ in an AR(1) realization than in a random sample of the same size. In general, a realization of length n from a time series having spectral density f is equivalent for estimating μ to a random sample of size N from a population having the same variance $R(0)$ as the time series, where

$$N = \frac{nR(0)}{f(0)} = \frac{n}{\displaystyle\sum_{v=-\infty}^{\infty} \rho(v)}.$$

Thus if $f(0)$ is large (small), estimation of μ for the time series case is less (more) accurate than for the corresponding random sample case.

4.1.2. The Sample Autocovariances and Autocorrelations

In Lecture 1 we stated that the sample autocovariance function

$$\hat{R}(v) = \frac{1}{n} \sum_{t=1}^{n-|v|} (X(t) - \bar{X})(X(t+|v|) - \bar{X}), \qquad |v| < n,$$

is generally accepted as the most satisfactory estimator of the autocovariance function R. The most popular alternative to $\hat{R}$ is the "unbiased" estimator

$$\tilde{R}(v) = \frac{1}{n-|v|} \sum_{t=1}^{n-|v|} (X(t) - \bar{X})(X(t+v) - \bar{X})$$

$$= \frac{n}{n-|v|}\hat{R}(v),$$

so called because if we knew μ and used it instead of $\bar{X}$ in the definition of $\tilde{R}$, the result would be unbiased, whereas $\hat{R}(v)$ with μ replacing $\bar{X}$ would have bias $-|v|/n$. However, we rarely know μ so neither $\hat{R}$ nor $\tilde{R}$ is unbiased, and in fact their bias is small relative to their variability. Further, it is generally believed that the mean square error of $\hat{R}(v)$ is smaller than that of $\tilde{R}(v)$.

$\boxed{\text{Theorem 4.1.2}}$ PROPERTIES OF $\hat{R}$ AND $\hat{\rho}$

Under mild assumptions,

a) $E(\hat{R}(v) - R(v)) = -\dfrac{|v|}{n}R(v) \to 0$ as $n \to \infty$.

b) We have

$$\lim_{n \to \infty} n\mathrm{Cov}(\hat{\rho}(h), \hat{\rho}(g))$$

$$= \sum_{r=-\infty}^{\infty} \Big[\rho(r+g)\rho(r+h) + \rho(r-g)\rho(r+h) - 2\rho(h)\rho(r)\rho(r+g)$$

$$- 2\rho(g)\rho(r)\rho(r+h) + 2\rho(g)\rho(h)\rho^2(r) \Big]$$

$$= \frac{2}{R^2(0)} \int_0^1 \big[\cos 2\pi h\omega - \rho(h)\big]\big[\cos 2\pi g\omega - \rho(g)\big] f^2(\omega)d\omega.$$

c) Any finite collection of $\hat{R}$'s or $\hat{\rho}$'s is asymptotically multivariate normal.

d) We have

$$\lim_{n \to \infty} n\mathrm{Var}(\hat{R}(v)) = \sum_{r=-\infty}^{\infty} [R^2(r) + R(r-v)R(r+v)] = 2\int_0^1 \cos^2 2\pi v\omega\, f^2(\omega)d\omega,$$

while

$$\lim_{n \to \infty} n\mathrm{Var}(\hat{\rho}(v)) = \sum_{r=-\infty}^{\infty} \Big[\rho^2(r) + \rho(r-v)\rho(r+v) - 4\rho(v)\rho(r)\rho(r+v) + 2\rho^2(v)\rho^2(r) \Big]$$

$$= \frac{2}{R^2(0)} \int_0^1 \big[\cos 2\pi v\omega - \rho(v)\big]^2 f^2(\omega)d\omega.$$

e) If X is a white noise series with variance $R(0) = \sigma^2$, we have, approximately for large samples, that the elements of the covariance and correlation sequences are uncorrelated, and for $v > 0$,

$$\hat{R}(v) \sim N(0, \frac{\sigma^4}{n}), \qquad \hat{\rho}(v) \sim N(0, \frac{1}{n}),$$

while $\hat{R}(0) \sim N(\sigma^2, 2\sigma^4/n)$.

4.1.3. The Sample Spectral Density Function

The next theorem verifies our observation in Lecture 1 that the sample spectral density function is too oscillatory to be useful for making statistical inferences.

| **Theorem 4.1.3** | PROPERTIES OF THE PERIODOGRAM |

Under mild assumptions,

a) The periodogram is unbiased.

b) We have

$$\lim_{n \to \infty} \mathrm{Var}(\hat{f}(\omega)) = \begin{cases} f^2(\omega), & \omega \neq 0, .5 \\ 2f^2(\omega), & \omega = 0, .5, \end{cases}$$

while

$$\lim_{n \to \infty} \mathrm{Cov}(\hat{f}(\omega_1), \hat{f}(\omega_2)) = 0, \qquad \omega_1 \neq \omega_2.$$

c) *For any integer $M \geq 1$ and fixed frequencies $\omega_1, \ldots, \omega_M$ in $(0, .5)$, we have that the random variables*

$$\frac{2\hat{f}(\omega_1)}{f(\omega_1)}, \ldots, \frac{2\hat{f}(\omega_M)}{f(\omega_M)}$$

converge in distribution to the M independent random variables $Q_1, \ldots, Q_M$, each having the chi-square distribution with two degrees of freedom.

d) *If X is a normal white noise process with variance σ^2, then the periodogram ordinates, that is, the values of the sample spectral density at the natural frequencies, are independently distributed and*

$$\frac{2\hat{f}(\omega_j)}{\sigma^2} \sim \chi_2^2, \qquad \omega_j \neq 0, .5 \qquad \frac{\hat{f}(\omega_j)}{\sigma^2} \sim \chi_1^2, \qquad \omega_j = 0, .5.$$

Implications: Part (b) shows that the sample spectral density function is not consistent since its variance does not go to zero. Further, the variance at frequency ω is a constant independent of sample length. The fact that the variance of $\hat{f}(\omega)$ is a function of $f^2(\omega)$, that is, the square of the mean of $\hat{f}(\omega)$, is another reason for plotting the log of the periodogram instead of the periodogram itself, as it is easy to show that the limiting variance of the log of $\hat{f}(\omega)$ is a constant independent of ω.

Part (b) also shows that the values of the sample spectral density are asymptotically uncorrelated at different frequencies. Thus the log periodogram appears to be a white noise series itself except that its level will change with the level of the true spectral density. Parts (c) and (d) provide information on the asymptotic distribution of the sample spectral density. In essence, the sample spectral density has the chi-square distribution for large n in general, while if X is a Gaussian white noise series, then the periodogram is exactly chi-square for all n. We will use this fact in the next section when we consider tests for white noise.

4.2. Tests for White Noise

In this section we consider two widely used methods for testing whether it is reasonable to conclude that a data set is a realization from a white noise time series.

4.2.1. Bartlett's Test

Suppose that $X(1), \ldots, X(n)$ is a realization from a WN(σ^2) process. Thus the spectral distribution function F of X is given by $F(\omega) = \sigma^2 \omega$ for $\omega \in [0,1]$, and we would expect that the cumulative periodogram

$$\hat{F}(\omega_k) = \frac{\sum_{j=1}^{k} \hat{f}(\omega_j)}{\sum_{j=1}^{q} \hat{f}(\omega_j)}, \qquad k = 1, \ldots, q = [n/2] + 1,$$

of the data should be close to the points

$$S_k = \frac{k}{q}, \qquad k = 1, \ldots, q.$$

In fact it is easy to show that

$$\lim_{q \to \infty} \Pr\left(\max_{1 \leq k \leq q} \sqrt{q}\left|S_k - \frac{k}{q}\right| \leq a\right) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2a^2 j^2} \equiv G(a)$$

to judge whether a cumulative periodogram has a maximum deviation from the expected straight line too extreme to be reasonable under the hypothesis of white noise.

4.2.2. The Portmanteau Test

Bartlett's test is done in the frequency domain and uses the fact that the spectral density of white noise is constant. Another approach to the problem of testing for white noise is to use the fact that its autocorrelation coefficients for nonzero lag are all zero. The most popular method of this type is often called the portmanteau or Q test. If $X(1), \ldots, X(n)$ is a sample realization from a white noise process, then the statistic

$$Q = n(n+2) \sum_{j=1}^{m} \frac{1}{n-j} \hat{\rho}^2(j) \xrightarrow{\mathcal{L}} \chi_m^2.$$

Thus the hypothesis of white noise is rejected if $Q > \chi_{\alpha,m}^2$. Note that this test is often applied to the residuals (one step ahead forecast errors) from a model that has been fit to a data set. In this case, one degree of freedom is subtracted from m for each parameter that has been estimated.

The Choice of m in the Q Test

One difficulty in using the Q test is in deciding what value of m to use. Asymptotically, this choice shouldn't matter, but to investigate its effect in small and moderate samples, we give in Table 4.1 the results of a simple simulation study. For samples of size 50, 100, and 200, we generated 500 Gaussian white noise series and counted how many times the null hypothesis of white noise was rejected for $\alpha = .01$, .05, and .10 for $m=5$, 10, and 20. For each sample size, these should be 5, 25, and 50 for the respective values of α. For each α, we also counted the number of times out of 500 that the conclusion was not the same for the three values of m. From these results, it appears that (1) the true Type I error probability of the Q test tends to increase with m, particularly for smaller sample sizes, and (2) the effect of choosing m is more important as α increases.

Table 4.1. Results of a Simulation Study of the Effect of the Choice of m in the Q Test

	alpha=.01			alpha=.05			alpha=.10		
	n=50	n=100	n=200	n=50	n=100	n=200	n=50	n=100	n=200
m=5	5	8	6	31	27	27	46	52	35
m=10	14	6	4	31	26	25	55	56	54
m=20	18	14	6	46	43	23	72	70	53
# diff.	21	20	11	57	48	45	81	82	74

4.3. Estimating the Parameters of ARMA Models

Given data $X(1), \ldots, X(n)$ from an ARMA process of orders (p, q), we would like to find estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$ of the parameters of the model as this would allow us to estimate f by

$$\hat{f}(\omega) = \hat{\sigma}^2 \frac{\left| \sum_{k=0}^{q} \hat{\beta}_k e^{2\pi i k \omega} \right|^2}{\left| \sum_{j=0}^{p} \hat{\alpha}_j e^{2\pi i j \omega} \right|^2},$$

and we could substitute the parameter estimates into the formulas for forecasting future values of X. In this section we assume that the orders p and q are known and discuss the estimation of the parameters. In the next section we consider the problem of determining p and q.

4.3.1. Maximum Likelihood Estimation

The first method that we consider is the application of the general principle of maximum likelihood. If we assume that the data are from a Gaussian ARMA process, the ARMA likelihood can be evaluated for specified values of its parameters via the Kalman Filter Algorithm. Much of the recent emphasis in calculating the maximum likelihood estimates for ARMA processes has been on combining a likelihood "evaluator" with a nonlinear "optimizer" that requires no derivatives.

The basic theoretical results for maximum likelihood estimation for time series are contained in the following theorem.

Theorem 4.3.1   PROPERTIES OF MLE's FOR TIME SERIES

Let $\mathbf{X}_n = (X(1), \ldots, X(n))^T$ be a finite realization from a Gaussian time series that has spectral density f which depends on the parameters $\theta = (\theta_1, \ldots, \theta_r)^T$. Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ based on $\mathbf{X}_n$. Then under certain regularity conditions

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N_r(0_r, \mathbf{I}^{-1}(\theta)),$$

where $\mathbf{I}(\theta)$ is called the information matrix of θ and has (j, k)th element

$$I_{jk}(\theta) = \frac{1}{2} \int_0^1 \frac{\partial \log f(\omega)}{\partial \theta_j} \frac{\partial \log f(\omega)}{\partial \theta_k} d\omega, \qquad j, k = 1, \ldots, r.$$

Asymptotic Distribution of ARMA MLE's

The next theorem applies Theorem 4.3.1 to ARMA processes.

Theorem 4.3.2   ASYMPTOTIC DISTRIBUTION OF ARMA MLE's

Suppose X is an invertible Gaussian $ARMA(p, q, \alpha, \beta, \sigma^2)$ process. Let R_Y and R_Z be the autocovariance functions of an $AR(p, \alpha, 1)$ and $AR(q, \beta, 1)$ process, respectively. Let $\theta = (\alpha^T, \beta^T, \sigma^2)^T$. Then

a) The information matrix of θ is given by

$$\mathbf{I}(\theta) = \begin{bmatrix} \mathbf{I}^{\alpha\alpha} & \mathbf{I}^{\alpha\beta} & \mathbf{0}_p \\ & \mathbf{I}^{\beta\beta} & \mathbf{0}_q \\ \text{Symmetric} & & \dfrac{1}{2\sigma^4} \end{bmatrix}$$

where for $j, k = 1, \ldots, p$ and $l, m = 1, \ldots, q$, we have

$$I_{jk}^{\alpha\alpha} = R_Y(|j - k|)$$

$$I_{lm}^{\beta\beta} = R_Z(|l - m|)$$

$$I_{jl}^{\alpha\beta} = -\int_0^1 \frac{e^{2\pi i(j-l)\omega}}{g(e^{2\pi i\omega})h(e^{-2\pi i\omega})} d\omega.$$

350

b) If $q = 0$, then $I^{-1}(\alpha) = S(\alpha)$, where $S(\alpha)$ is the Schur matrix corresponding to α, and thus

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{L}} N_{p+1} \left(0_{p+1}, \begin{bmatrix} S(\alpha) & 0_p \\ 0_p^T & 2\sigma^4 \end{bmatrix} \right).$$

c) If $p = 0$, then $I^{-1}(\beta) = S(\beta)$, where $S(\beta)$ is the Schur matrix corresponding to β, and thus

$$\sqrt{n} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{L}} N_{q+1} \left(0_{q+1}, \begin{bmatrix} S(\beta) & 0_q \\ 0_q^T & 2\sigma^4 \end{bmatrix} \right).$$

4.3.2. Approximate MLE's

Box and Jenkins show that if X is an invertible ARMA process with a Gaussian error term ϵ, then the likelihood function of α, β, and σ^2 for a realization X of length n can be expressed as

$$L(\alpha, \beta, \sigma^2 | X) = (2\pi\sigma^2)^{-n/2} |M_n(\alpha, \beta)|^{1/2} \exp\left(-\frac{S(\alpha, \beta)}{2\sigma^2} \right),$$

where M_n is a matrix that is a function of α and β, and

$$S(\alpha, \beta) = \sum_{t=-\infty}^{n} \hat{\epsilon}^2(t)$$

with $\hat{\epsilon}(t) = E\left(\epsilon(t) | X \right)$, and the conditional expectation is calculated assuming that α and β are the true values of the ARMA parameters. To find the α, β, and σ^2 that maximize the likelihood, Box and Jenkins state first that the term $|M_n(\alpha, \beta)|$ can be ignored, and second that the summation for $S(\alpha, \beta)$ can be truncated below at some finite limit, $-T$, say. The process they suggest for finding the $\hat{\epsilon}$'s is called back forecasting. Once $\hat{\alpha}$ and $\hat{\beta}$ are found to minimize $S(\alpha, \beta)$, the value of σ^2 that maximizes the likelihood (again ignoring the term $|M_n|^{1/2}$) is given by

$$\hat{\sigma}^2 = \frac{S(\hat{\alpha}, \hat{\beta})}{n}.$$

4.3.3. Method of Moments Estimators

Given the autocorrelations, the coefficients and error variance for AR, MA, and ARMA processes can be found. using the relationships between the true ρ's and true coefficients. A natural way to estimate the coefficients of these processes is to use estimates of the autocorrelations in these procedures. Such estimators are called method of moments estimators. For AR processes, the method of moments estimates use the Yule-Walker equations with sample autocorrelations substituted for true autocorrelations. The resulting estimates are thus called the Yule-Walker estimates. For an AR(p) process the Yule-Walker estimates have the same asymptotic properties as the maximum likelihood estimators;, that is, they are asymptotically efficient.

MA Processes

The method of moments for MA and ARMA processes are not guaranteed to exist. For an MA process, the factorization is feasible if the estimate

$$\hat{f}(\omega) = \sum_{v=-q}^{q} \hat{R}(v) \cos 2\pi v\omega$$

is positive for all $\omega \in [0, 1]$.

4.3.4. Estimation for AR Processes

Autoregressive processes are widely used in both spectral estimation In this section we consider some non-MLE estimation procedures for such processes. Throughout the section we assume that the sample mean has been removed from the data being analyzed.

Yule-Walker Estimators

Perhaps the most natural method of estimating the parameters α and σ^2 of an AR(p, α, σ^2) process is to substitute the sample autocovariances $\hat{R}$ for the true autocovariances R in the Yule-Walker equations and then solve the resulting sample Yule-Walker equations:

$$\hat{\Gamma}_p \hat{\alpha} = -\hat{r} \quad \text{and} \quad \hat{\sigma}^2 = \sum_{j=0}^{p} \hat{\alpha}_j \hat{R}(j),$$

to obtain what are called the Yule-Walker estimators (YWE) $\hat{\alpha}$ and $\hat{\sigma}^2$. The availability of Levinson's algorithm makes the YWE's attractive from a computational point of view. Unfortunately there is wide evidence that the YWE's perform poorly in some cases.

Theorem 4.3.3 | PROPERTIES OF YWE'S

Let $\hat{\alpha}$ and $\hat{\sigma}^2$ be the YWE's calculated from a sample $X(1), \ldots, X(n)$. Then

a) $\hat{\alpha}$ and $\hat{\sigma}^2$ are the ordinary least squares estimators for the regression problem $y = -X\alpha + \epsilon$ where $E(\epsilon) = 0_n$, $\text{Var}(\epsilon) = \sigma^2 I_n$, y is the $(n+p)$-dimensional vector $y = (X(1), \ldots, X(n), 0, \ldots, 0)^T$, and X is the $(n+p) \times p$ matrix $X = (x_1, \ldots, x_p)$, where $x_j = L^j y$.

b) The zeros of $\hat{g}(z) = \sum_{j=0}^{p} \hat{\alpha}_j z^j$ are guaranteed to be outside the unit circle.

c) If $X(1), \ldots, X(n)$ is a sample from a Gaussian AR(p, α, σ^2) process, then

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{L}} N_{p+1} \left(0_{p+1}, \begin{bmatrix} S(\alpha) & 0_p \\ 0_p^T & \frac{1}{2\sigma^4} \end{bmatrix} \right),$$

where $S(\alpha)$ is the Schur matrix corresponding to α.

Implications: This theorem gives two important results in addition to verifying our observation about the use of zeros as proxies for unobserved data. First, the fitted process is guaranteed to be stable. Thus the spectral density is sure to be positive. Further, if we use the $\hat{\alpha}$ in the AR prediction formula, we have

$$\hat{X}(n+j) = -\sum_{j=1}^{p} \hat{\alpha}_j \hat{X}(n+j-1), \qquad j = 1, 2, \ldots,$$

and the predictors will converge to zero. The second result is that the YWE's are asymptotically efficient; that is, they have the same large sample properties as the maximum likelihood estimators.

Properties of Sample Partial Autocorrelations

The sample partial autocorrelations play an important role in determining models for time series. In the next theorem we describe some basic sampling properties that they have.

Theorem 4.3.4 | PROPERTIES OF PARTIAL AUTOCORRELATIONS

If X is a Gaussian AR(p) process, then the sample partial autocorrelations for lags $p+1$ and higher are asymptotically independent and identically normally distributed, are asymptotically unbiased, and have asymptotic variance $1/n$.

4.4. Identifying ARMA Models

The problem of determining what type of ARMA model best fits a data set has become very important in time series analysis. In this section we describe a variety of diagnostics that have used to identify models.

4.4.1. Some Useful Diagnostics

Suppose that we have a realization $X(1), \ldots, X(n)$ from a covariance stationary time series X. We saw in Lecture 3 that the true autocorrelations, partial autocorrelations, and spectral density have certain characteristics for different model types. The statements below are of necessity rather general and somewhat vague. In later parts of this section we will describe some methods that have been developed to try to use these statements in a somewhat automatic way.

AR Processes

The autocorrelation function decays to zero and then oscillates about zero, while the partial autocorrelation becomes identically zero for lags greater than the true order p. The decay of ρ can follow a sinusoidal pattern if some of the zeros of its characteristic polynomial are complex. The spectral density of an AR process can contain very sharp peaks, while the troughs appear somewhat less sharp.

MA Processes

The autocorrelation function is identically zero for lags greater than the order of the process, while the partial autocorrelation function does not become identically zero. The peaks in the spectral density are smooth relative to what can be attained for AR models, while the troughs can be rather sharp.

ARMA Process

Here the autocorrelation function decays in the same way as that of an AR process but only after lag q. Again the partial autocorrelation does not become zero. The spectral density function can now have either sharp or rounded peaks and troughs.

The statements made above are for models that have nonzero coefficients. If only a few of the coefficients of a model are nonzero, then further statements can be made. We consider two examples of this, although there are a wide variety of possibilities. First, if we have an MA or ARMA model with only a few nonzero coefficients, then $\rho(v)$ can be small for some lags smaller than q. Second, if X is an AR(p) with only α_p

nonzero, then the spectral density of X will have a very sharp peak at frequency $1/p$. We will consider below automatic methods for determining if a subset model will adequately represent a time series.

4.4.2. The AIC and Related Criteria

The problem of order determination in time series analysis is similar to the general statistical problem of determining which of a set of possible models best fits a set of data. One very important method for making such a choice is due to Akaike and is thus called Akaike's Information Criterion (AIC). Although there are a variety of ways to describe AIC, one simple motivation is as follows. If we have two competing models for a data set and they have the same number of parameters, then we would probably choose the one that best "fits" the data in some sense. For example, in regression, if we have two possible models for the dependent variable and each is in terms of a subset of q independent variables, then we would probably choose that model that leads to the smaller sum of squares of residuals. If competing models have different numbers of parameters, then we would not want to choose one based merely on goodness of fit. Again, in regression, it is a well-known phenomenon that adding a variable to a model is guaranteed to decrease the sum of squares of residuals, but also tends to increase the variability in predictions resulting from the larger model. We have also seen this phenomenon in the problems of smoothing that we have considered. In such problems, increasing the complexity of a model may reduce variability, while at the same time increasing bias.

Using general principles of information theory, Akaike proposed measuring the "goodness" of a model that has r parameters $\theta = (\theta_1, \ldots, \theta_r)^T$ by

$$\text{AIC}(r) = -2 \log L(\hat{\theta}) + 2r,$$

where L is the likelihood function of θ, and $\hat{\theta}$ is the maximum likelihood estimate of θ. The first term is the measure of how well a model fits the data, while the second is a penalty for the number of parameters in the model. The AIC is then used to choose the model that has the smallest value of the criterion. Thus if a term in AIC is the same for all competing models, it can be deleted. For ARMA modeling, this leads to considering

$$\text{AIC}(r) = n \log \hat{\sigma}_r^2 + 2r,$$

where $\hat{\sigma}_r^2$ is the maximum likelihood estimate of the error variance for the model having r parameters.

The CAT Criterion and AR Models

Another criterion that is useful for determining the order of an autoregressive process is the CAT criterion due to Parzen. Instead of assuming that X is an AR(p) process, Parzen considered the problem of determining the order p of an autoregressive process that approximates in some optimal way the behavior of an arbitrary covariance stationary time series. We will discuss this further in Lecture 6.

Definition. Let $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2$ be the error variance estimates from the Yule-Walker equations based on a sample of size n, and let

$$\bar{\sigma}_j^2 = \frac{n}{n-j} \hat{\sigma}_j^2, \qquad j = 1, \ldots, M.$$

Then the criterion autoregressive transfer function (CAT) criterion for order k is defined to be

$$CAT(k) = \begin{cases} \left(\frac{1}{n} \sum_{j=1}^{k} \frac{1}{\bar{\sigma}_j^2} \right) - \frac{1}{\bar{\sigma}_k^2}, & k = 1, \ldots, M \\ \\ -\left(1 + \frac{1}{n} \right) \hat{R}(0), & k = 0. \end{cases}$$

Note that the natural way to define CAT(0) would be as just $-\hat{R}(0)$, in which case it has been shown that the AIC and the CAT criterion are asymptotically equivalent. Parzen has suggested the definition above so that CAT can be used as a test for white noise. As we pointed out above, if X is an AR(0) (that is, white noise) process, then AIC (and hence CAT with CAT(0)$= -\hat{R}(0)$) would asymptotically only choose order zero approximately 75% of the time.

AIC and ARMA Models

The AIC is also applicable for choosing the order of ARMA processes. Ideally, we should use the exact maximum likelihood method for each possible order and then calculate AIC based on the maximized likelihood. Unfortunately, this is very time consuming and in some cases (particularly when the process is close to nonstationary or noninvertible) it is difficult to get the maximization procedure to converge. An alternative is to use the estimate of the error variance obtained from the method of moments estimator for each order. This also can be difficult to implement since in some cases the factorization of the covariance generating function becomes infeasible. Thus in TIMESLAB, we suggest using the consistent but inefficient method described below to estimate the error variance for each order.

4.4.3. The Stepwise ARMA Method

Since an ARMA model looks very much like an ordinary regression model, it is natural to apply regression techniques in their analysis. In fact we saw above that applying ordinary least squares to autoregressive data leads to asymptotically efficient parameter estimates. In the MA case this was not true. In this section we extend the regression analogy to the ARMA case by describing the TIMESLAB command ARMASEL which performs a stepwise ARMA procedure.

Suppose we feel that an ARMA model of some order (p, q) should fit a data set $X(1), \ldots, X(n)$ and (P, Q) is an upper bound on the order. Then in analogy with stepwise regression analysis, we would write

$$X(t) = -\alpha_1 X(t-1) - \cdots - \alpha_P X(t-P) + \beta_1 \epsilon(t-1) + \cdots + \beta_Q \epsilon(t-Q) + \epsilon(t),$$

that is, we would regress $X(t)$ on previous X's and ϵ's. Since we cannot observe the ϵ's, we cannot do the regression directly. The command

        ARMASEL(x,n,m,s,k1,k2,kopt,pval,p,q,alpha,beta,rvar,ier)

carries out a stepwise version of the regression described above. The first eight arguments are input and the last six are output although alpha and beta can be used as both input and output if kopt is negative as described below. The final model chosen has orders p and q returned in the integers p and q, coefficients α and β returned in the arrays alpha and beta, and noise variance returned in the real scalar rvar. The output integer ier is 0 if no errors are encountered, while it is 1 if the matrix being swept is judged to be singular at any step.

The arguments x, n, m, and s contain the data, the sample size, and M and s as described above. The arguments k1, k2, kopt, and pval allow the user to use ARMASEL in a variety of ways. First, the real scalar pval is the "p-value to enter" at each step of the stepwise procedure. Thus using a smaller value for pval allows more variables to enter the model. Then m, s, k1, k2, and kopt (and possibly alpha and beta) together describe what lags are contenders for inclusion in the model. First, the largest AR or MA lag possible is m-s as described above. Then we have the following rules based on the value of kopt.

kopt=0   - The k1 AR lags (out of the possible m-s) having the largest $\hat{R}_{XX}(v)$ in absolute value and the k2 MA lags having the largest $\hat{R}_{X\epsilon}(-v)$ in absolute value are contenders for inclusion.

kopt=j>0   - This is the same as kopt=0 except that the stepwise procedure continues until j lags are included.

kopt=j<0   -   The k1 AR lags that are entered in alpha and the k2 MA lags that are entered in beta
             are forced into the model. Thus k1+k2= −j.

Subset Autoregression

Subset AR models have been found to be particularly useful in representing some time series. For example, for the log of the lynx data, an AR(11) model having only coefficients of lags 1, 2, 4, 10, and 11 nonzero has a value of AIC smaller than that of any full AR model. The ARMASEL command can be used to fit such models.

LECTURE 5

Forecasting Using Models

In this lecture we discuss forecasting based on both ARMA models and regression models.

5.1. Box-Jenkins Modeling and Forecasting

The most popular ARMA-model based forecasting method is one proposed by Box and Jenkins. The method uses the traditional statistical modeling strategy of (1) model identification, (2) parameter estimation, (3) diagnostic checking, and (4) consideration of alternative models, if necessary, within the class of multiplicative seasonal ARIMA models to obtain a model suitable for forecasting. In this section we describe the various steps in the method. The appeal of the method is its simplicity and the wide availability of computer software for using it.

The Models

The Box-Jenkins method for data $X(1), \ldots, X(n)$ assumes that if we transform X to a series W by

$$W(t) = (1 - L)^d (1 - L^S)^D Y(t),$$

where

$$Y(t) = \begin{cases} (X(t) + m)^\lambda, & \lambda \neq 0 \\ \log(X(t) + m), & \lambda = 0, \end{cases}$$

and m is chosen to make $X(t) + m$ positive for all t (unless $\lambda = 1$), then a model of the form

$$\left(\sum_{j=0}^{p} \alpha_j L^j \right) \left(\sum_{k=0}^{P} \theta_k L^{kS} \right) (W(t) - \mu) = \left(\sum_{l=0}^{q} \beta_l L^l \right) \left(\sum_{r=0}^{Q} \gamma_r L^{rS} \right) \epsilon(t),$$

with $\epsilon \sim WN(\sigma^2)$, will adequately represent W. Note that this model for W can be written as an ARMA($p + PS, q + QS$) model by multiplying the polynomials in the model. One of the appeals of the Box-Jenkins models is that many of the coefficients in this general ARMA model are allowed to be zero. Thus this class of models allow us to represent data parsimoniously, that is, with a small number of parameters. This is the same idea as in the stepwise ARMA modeling embodied in the ARMASEL command.

The model for W is called the multiplicative seasonal ARMA model with orders $\langle p, P, q, Q, d, D, S \rangle$ and coefficients $\langle \alpha, \theta, \beta, \gamma \rangle$. We will generalize this model slightly by removing the restriction that the powers of L in the second and fourth polynomials above are of the form kS and rS. This allows us to include the general subset ARMA model in the class of models that we can consider. Thus we have the model

$$g(L)G(L)[W(t) - \mu] = h(L)H(L)\epsilon(t),$$

357

where

$$g(L) = \sum_{j=0}^{p} \alpha_j L^j, \qquad G(L) = \sum_{k=0}^{P} \theta_k L^{u_k}, h(L) = \sum_{l=0}^{q} \beta_l L^l, \qquad H(L) = \sum_{m=0}^{Q} \gamma_m L^{v_m},$$

are called the full and subset AR and full and subset MA operators, respectively. Each of these is assumed to have all of their zeros outside the unit circle, so that the model is stationary and invertible. We can also write the model with a constant term τ_0, that is,

$$g(L)G(L)W(t) = \tau_0 + h(L)H(L)\epsilon(t),$$

where

$$\tau_0 = g(1)G(1)\mu = \left(\sum_{j=0}^{p} \alpha_j\right)\left(\sum_{k=0}^{P} \theta_k\right)\mu.$$

From our previous discussions, it should be clear that the Y transform is used to stabilize variance, the W transform removes trends and cycles, and the model for W represents the behavior in the data after this "preprocessing." In the remainder of this section we describe and illustrate the four steps in the Box-Jenkins procedure.

Model Identification

The model identification step itself consists of three parts:

1. Determine if the power transform from X to Y is required, and if so, what values of m and λ should be used. Note that m is only included to make $X(t) + m$ positive for all t.

2. Determine if the differencing transform from Y to W is necessary, and if so, what values of d, D, and S should be used.

3. Determine if an ARMA model is required to model W, and if so, what values of the orders and lags should be used.

Identifying λ

This step is used either (1) to stabilize the variance of X or (2) to produce a time series Y for which the differencing step can eliminate any trends and/or cycles. This step is not needed for any series whose variance appears constant and which has no trends or cycles. To illustrate a case where the power transform is needed, consider the series in Figure 5.1, which consists of the sales of an industrial company for 77 months. This data set consists mainly of a linear trend and an annual cycle whose amplitude is increasing with time at a faster than linear rate. If no power transformation is applied, no amount of differencing can remove the cycle. For example, we have included in Figure 5.1 a graph of the first and 12th difference of X. Since all of the data are positive, we let $m = 0$ and try to find a suitable value for λ.

There are a variety of possible methods for selecting λ. Unfortunately, which method to use is somewhat subjective and depends on the nature of the data. We will try a variety of λ's and chooses the one that has the highest value of R^2 for the regression model

$$X^\lambda(t) = a + bt + c\cos\frac{2\pi(t+1)}{12} + \epsilon(t),$$

that is, the value that leads to the series that is best fit by a linear trend plus a cosine curve. Note that the cosine is shifted two months because the data start in January and the maximum yearly value is in November.

358
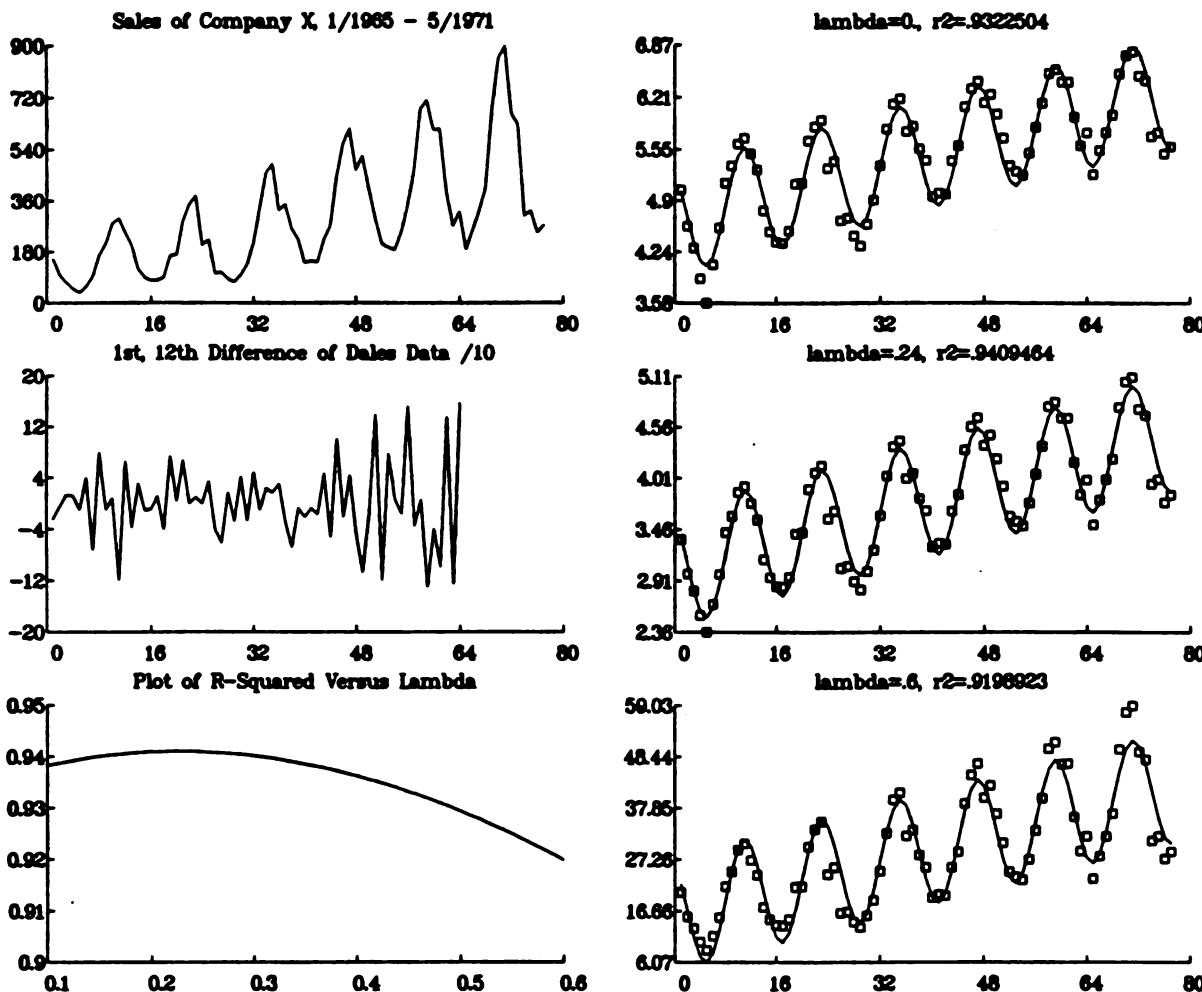
Figure 5.1. Example of Identifying λ.

In Figure 5.1 we have included a graph of R^2 for 51 values of λ between .1 and .6, and the graphs of X^λ for $\lambda = 0$, $\lambda = .24$ (the value giving the maximum R^2), and $\lambda = .6$. The log transform is ruled out because of how poorly the first cycle of the cosine fits the data.

The method of the previous paragraph is appropriate for these data because the transformed data do seem to be sinusoidal. In general, one chooses the transformation so that the resulting data most closely match some deterministic function that can be removed by differencing.

Identifying Differences

In this part of the process, one looks for the least amount of differencing required to produce a correlogram and partial correlogram that can be matched by that of a multiplicative subset ARMA process having only a few parameters. This is not always an easy process as it requires rather extensive experience. In Section 3.5 we described a variety of types of correlograms and partial correlograms and the corresponding models. Unfortunately no such description can be exhaustive. Note that doing more differencing leads to losing observations, while doing less may lead to having to use more parameters in the ARMA model. The probability limits for prediction that we will obtain below are a function of the number of observations minus the number of estimated parameters. Thus there is a tradeoff in degrees of freedom in these two parts of the

identification process.

To illustrate differencing, consider again the sales data. In Table 5.1, we give the values of the correlogram and partial correlogram for the power transformed data and for its first, 12th, and first and 12th differences.

Table 5.1. Table of Correlations and Partial Correlations for Various Differenced Versions of the Power Transformed Sales Data

```
Correlations of Y
   1 |  .89  .71  .45  .19 -.01 -.09 -.07  .06  .28  .48  .63  .69
  13 |  .61  .46  .25  .03 -.14 -.21 -.19 -.06  .07  .24  .37  .42
  25 |  .36  .26  .09 -.10 -.26 -.34 -.36 -.29 -.16 -.04  .07  .14
Partials of Y
   1 |  .89 -.42 -.50  .06  .29  .43  .12  .15  .13 -.06  .03 -.04
  13 | -.28  .04  .13 -.02 -.18 -.04  .11 -.06 -.06  .12 -.07  .06
  25 | -.06 -.06 -.10 -.11  .02  .02 -.11 -.06 -.06  .09 -.06 -.02

Correlations of 1st Difference
   1 |  .33  .40 -.07 -.29 -.57 -.49 -.54 -.25  .02  .24  .45  .61
  13 |  .36  .27 -.01 -.23 -.44 -.43 -.43 -.20 -.06  .24  .27  .50
  25 |  .29  .29  .04 -.13 -.35 -.34 -.35 -.20 -.10  .15  .18  .39
Partials of 1st Difference
   1 |  .33  .33 -.33 -.45 -.46 -.13 -.20 -.24 -.03 -.06 -.04  .24
  13 | -.06 -.16  .06  .19 -.03 -.14  .04  .13 -.16  .01 -.15  .06
  25 |  .06  .05  .00 -.09  .02  .11  .05  .04 -.14  .07  .01 -.01

Correlations of 12th Difference
   1 |  .40  .54  .23  .19  .07  .04 -.13 -.10 -.06 -.14  .06 -.22
  13 | -.10 -.22 -.14 -.12 -.06 -.10 -.04 -.03 -.02 -.06 -.14 -.20
  25 | -.17 -.06 -.06  .01  .00  .06  .10  .10  .00 -.01 -.13 -.14
Partials of 12th Difference
   1 |  .40  .45 -.10 -.12 -.02  .01 -.20 -.06  .22 -.11  .16 -.30
  13 | -.17  .03  .03  .10  .00 -.01 -.06 -.11  .03 -.23  .02 -.11
  25 |  .02  .22 -.17  .07  .00 -.06  .05 -.13  .01 -.15 -.03 -.16

Correlations of 1st,12th Difference
   1 | -.60  .36 -.22  .06 -.06  .11 -.17 -.02  .12 -.26  .44 -.35
  13 |  .18 -.14  .06 -.04  .09 -.09  .05  .00  .03  .04 -.02 -.06
  25 | -.06  .09 -.11  .10 -.06  .02  .04  .06 -.07  .06 -.06 -.06
Partials of 1st,12th Difference
   1 | -.60  .01  .00 -.10 -.13  .06 -.10 -.31  .04 -.22  .23  .06
  13 | -.16 -.12 -.15 -.06 -.06  .01  .05 -.11  .14 -.09  .02 -.11
  25 | -.29  .11 -.12 -.04  .01 -.10  .09 -.06  .06 -.04  .05 -.06
```

```
r0z=38863.330000
r0=.397435
r01=.083806
r012=.036166
r0112=.044060
```

Identifying ARMA Models

Once the data are differenced (if necessary), the next part of the method is to find an ARMA model that adequately represents the differenced data. Here only the sample autocorrelation and partial autocorrelation functions are used. To illustrate this process, we continue with a discussion of the sales data. Note that the correlogram and partial correlogram of the first and 12th difference series have an obvious pattern that the others don't. First, the lag one partial is large and is followed by several that are small, followed by an increase at lags 10 and 11. The first three correlations (−.6, .36, −.22) are decaying exponentially with alternating signs, indicating the presence of an AR(1) term in the model with coefficient approximately .6. The presence of large correlations at lags 10, 11, and 12 indicates the presence of a subset MA term in the

data. Since the maximum correlation is at lag 11, it is tempting to conclude that the MA term should be of lag 11. In Table 5.2 we give the true autocorrelation function for the best fitting models having an AR(1) term and then either an MA(11) term or an MA(12) term.

Table 5.2. True Correlations for Two Models for Sales Data

```
Coefficients From SEASEST
    1 |        .504522        .780130
rvar=.025497
Correlations for AR(1),MA(11)
    1 | -.50  .25 -.13  .06 -.02  .00  .02 -.06  .12 -.24  .46 -.24
   13 |  .12 -.06  .03 -.02  .01  .00  .00  .00  .00  .00  .00  .00
Coefficients From SEASEST
    1 |        .501941       -.801871
rvar=.024118
Correlations for AR(1),MA(12)
    1 | -.50  .25 -.13  .06 -.03  .01  .01 -.03  .06 -.12  .24 -.49
   13 |  .24 -.12  .06 -.03  .02 -.01  .00  .00  .00  .00  .00  .00
```

The residual variance for the model having the MA(12) term is smaller than that for the model having the MA(11) term but the correlations for the latter model appear to match those of the data better. There seems to be no clear optimal choice between the two models.

Estimation and Diagnostic Checking

Once a model has been determined, the approximate likelihood estimation procedure is used to estimate its parameters, while the portmanteau test is used to determine if the residuals of the fitted model are white noise.

Forecasting

Given estimates of the parameters of a model that has been judged to be adequate, the Box-Jenkins procedure then produces forecasts as follows. For simplicity we suppose that $\mu = 0$. Then we have

$$g(L)G(L)(1 - L)^d(1 - L^S)^D Y(t) = h(L)H(L)\epsilon(t),$$

where Y is the power transformed series. The result of using the SEASEST command is estimates of the coefficients of the four polynomials $g, G, h,$ and H, that is, the parameters of the model for

$$W(t) = (1 - L)^d(1 - L^S)^D Y(t).$$

This model for W can be translated into a model for Y, and thus we can write Y as an ARMA model of orders $p^* = p + u_P + d + SD$ and $q^* = q + v_Q$. Note that if differencing was done, then some of the zeros of the polynomial for the AR part of the model for Y will be on the unit circle.

Now given this ARMA(p^*, q^*) model for Y, a recursion for the forecasts of Y can be used. This recursion is similar to the ones used in the approximate MLE procedure. Once forecasts for Y are obtained they can be converted to forecasts for the original series X by doing the transformation that is the inverse of the one done in the power transform step.

Consider Figure 5.2 which contains the sales data with forecasts of the next 24 values of the series and 95% probability limits appended.

Sales Data and Forecasts
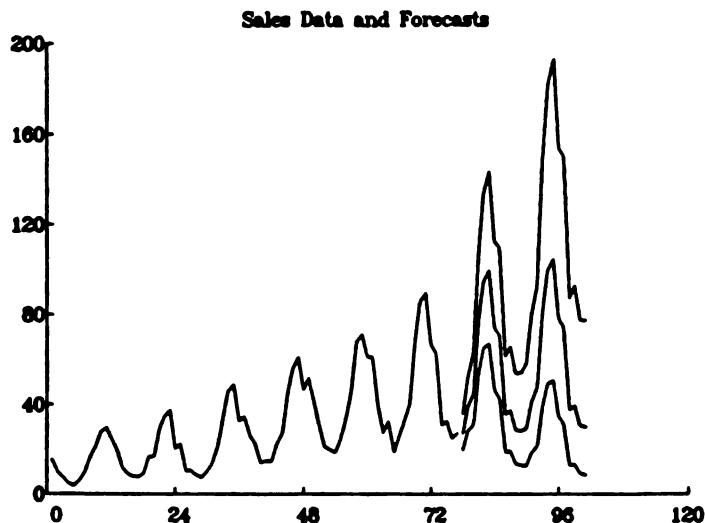
Figure 5.2. Sales Data and Forecasts with 95% Probability Limits.

5.2. Other Modeling Strategies

An important feature of the Box-Jenkins modeling and forecasting procedure is that the initial detrending and deseasonalizing transformations are done by differencing and not by regressing the data on deterministic functions of time.

5.2.1. Regression with Autoregressive Errors

Suppose we would like to model $X(t)$ as a deterministic function of time and the errors appear to be autocorrelated; that is, we have

$$y = X\beta + \epsilon,$$

where $y = (X(1),\ldots,X(n))^T$, X is an $(n \times m)$ matrix of fixed constants, and $\epsilon = (\epsilon(1),\ldots,\epsilon(n))^T$ has covariance matrix $\sigma^2 V$. In this section we assume that ϵ can be adequately modeled as an AR(p) process for some order p. We can estimate the parameters by the following iterative procedure.

1. Use ordinary least squares to find initial estimates $\hat{\beta}_0$ and residuals e_0.

2. Determine the order $\hat{p}_1$, coefficients $\hat{\alpha}_1$, and error variance $\hat{\gamma}_1^2$ of an autoregressive model for e_0.

3. Create a new observation vector s_1 and regression matrix W_1 by applying the AR filter found in step 2 to y and X.

4. Now apply ordinary least squares to s_1 and W_1 to obtain the coefficient estimate $\hat{\beta}_1$ and residuals $e_1 = y - X\hat{\beta}_1$.

5. Return to step 2 with e_1 replacing e_0.

This process continues until successive iterations result in the same value of the AR order. The successive values of the quantities involved will have subscripts 1,2,....

One possible model for the log (base 10) of the lynx data is

$$x(t) = \mu + a \cos \frac{2\pi t}{9.5} + b \sin \frac{2\pi t}{9.5} + \epsilon(t),$$

where ϵ is an AR(p). Using the procedure described above, we find $p = 2$ with coefficients -1.037 and 0.351, while the estimates of the regression coefficients are 2.901, 0.354, and -0.469.

5.2.2. ARARMA Modeling

One very simple method for automatically modeling and forecasting data has been suggested by Parzen (1982) and used successfully by Newton and Parzen in a forecasting competition organized by Makridakis (see Makridakis et al. (1984)). The method consists of two parts:

1. For some maximum lag M_1, calculate the regression coefficients

$$\hat{\beta}_k = -\frac{\sum_{t=1}^{n-k} X(t)X(t+k)}{\sum_{j=1}^{n} X^2(t)}, \qquad k = 1, \ldots, M_1.$$

Then let m be the value of k having the largest $|\hat{\beta}_k|$, and form

$$e(t) = X(t+m) + \hat{\beta}_m X(t), \qquad t = 1, \ldots, n - m.$$

2. Fit an autoregressive process to the e's, using maximum possible order M_2 and an order-determining criterion to determine the order p to use. In the general ARARMA procedure one would fit an ARMA model to the e's, but in most cases an AR process is adequate.

The first part is called the first AR since it is in essence fitting a one lag subset AR model to X. The second part is called the second AR. Note that the first AR may result in a coefficient that is greater than one, while in the second AR, it is recommended that the Yule-Walker or Burg estimators be used to guarantee that the process fit to the e's is stationary.

The result of this procedure is a model of the form

$$(1 + \hat{\beta}_m L^m)\hat{g}(L)X(t) = \epsilon(t),$$

where $\hat{g}$ is the AR operator determined in part 2 of the procedure. This model is similar in form to the Box-Jenkins model (with no MA terms) except that it is easily made automatic, and the data determine the nature of the first AR, which is analogous to differencing and is in fact sometimes referred to as "quasi-differencing". If the first AR turns out to be stable, that is, the coefficient has absolute value less than one, then the forecasts will eventually converge to the mean of the observed data. As we discussed in Lecture 1, forecasts that follow a difference equation of necessity either are explosive or must converge to some finite value. In the short run this may not be troublesome, but if the analyst has a feel for the long-run nature of forecasts, then this information should be incorporated into whatever model is used, either by modeling this behavior using a regression model, or by using differencing if polynomial growth is expected, or by insisting on a stable ARMA model if the series is expected to remain fairly constant.
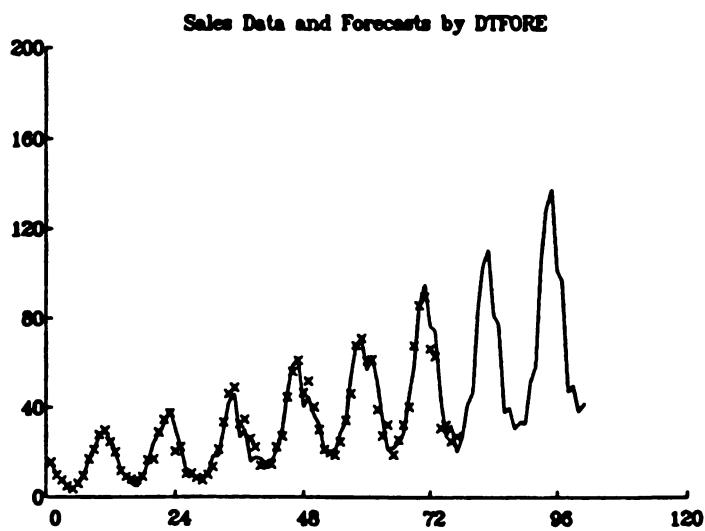
Figure 5.3. Sales Data and Forecasts As Determined by DTFORE.

The Sales Data Again

To illustrate the simplicity of the ARARMA procedure, we applied it to the fourth root of the sales data. It turned out that $a=12$, $p=2$, $beta=<-1.0566>$, and $alpha=<-.229,-.457>$. The first AR essentially chose 12th differences, except that the coefficient is actually more nonstationary than the -1 that would mean twelfth difference. In data such as this, this quasi-differencing will remove most of the variation in the data, and at least in the short term future, the model used in the second part of the model fitting has little impact on the nature of the forecast. Note further in this example the quasi-differencing will ultimately lead to more explosive growth than the actual 12th difference.

In Figure 5.3, we display the sales data, the values fitted by the model (the actual data are represented by the x's), and the forecasts of the next 24 values of the series. The values to be plotted were again divided by 10 prior to doing the plotting. Thus this figure is comparable to the one using the Box-Jenkins method.

If the first AR chosen by DTFORE is stable, then the SEASPRED command can be used to find probability limits for the forecasts. The intent of DTFORE however is to provide an easy-to-use command that gives forecasts that have been shown (in the forecasting competition) to compare very favorably with those given by more elaborate methods.

LECTURE 6

Searching for Periodicities

In this lecture we consider the classic problem of searching for periodicities in time series. The basic tools in this search are the spectral density function f and sample spectral density function $\hat{f}$. In Section 6.1 we consider nonparametric estimation of f, that is, estimation performed without making any assumptions about the form of the true f. Then in Section 6.2 we consider using ARMA models for estimating f. Finally, in Section 6.3 we use the results of the first two sections to actually search for periodicities in some famous series.

6.1. Nonparametric Spectral Density Estimation

We saw in Lecture 4 that the sample spectral density function $\hat{f}$ is an inconsistent estimator of the spectral density function f. In this section we describe how $\hat{f}$ can be modified to produce consistent estimators.

6.1.1. Smoothing the Sample Spectral Density

The basic problem with $\hat{f}$ is that it is too wiggly to be an adequate estimator of a function that is typically smooth over much of its domain. In this section we consider a simple averaging approach to smoothing the periodogram.

Recall from Theorem 4.1.3 that under general conditions

$$\hat{f}(0), \hat{f}(\tfrac{1}{n}), \ldots, \hat{f}(\tfrac{[n/2]}{n})$$

are asymptotically independent random variables with

$$\hat{f}(\omega_j) \sim \begin{cases} \dfrac{f(\omega_j)}{2}\chi_2^2, & \omega_j \neq 0, .5 \\[2mm] f(\omega_j)\chi_1^2, & \omega_j = 0, .5. \end{cases}$$

Consider estimating f at one of the natural frequencies $\omega_j = (j-1)/n$ by

$$\tilde{f}(\omega_j) = \frac{1}{2m+1} \sum_{k=-m}^{m} \hat{f}(\omega_{j+k}),$$

that is, by the average of $\hat{f}(\omega_j)$ and the m values of the periodogram on either side of it. If $\omega_{j-m} < 0$ or $\omega_{j+m} > .5$, we can use the fact that $\hat{f}$ is symmetric about 0 and .5 to obtain the elements in the sum.

Now suppose that $\omega_{j-m} > 0$ and $\omega_{j+m} < .5$. We can think of $\tilde{f}(\omega_j)$ as averaging all of the periodogram values in the frequency band $\omega_j \pm m/n$, that is, as a smoother having bandwidth m/n. We consider bandwidth to be half the width of the frequency band, that is, the width of the interval on each side of the

center frequency. Arguing heuristically, we assume that n is large and use the asymptotic properties of $\hat{f}$ to obtain

$$E(\bar{f}(\omega_j)) \doteq \frac{1}{2m+1} \sum_{k=-m}^{m} f(\omega_{j+k})$$

$$\text{Var}(\bar{f}(\omega_j)) \doteq \left(\frac{1}{2m+1}\right)^2 \sum_{k=-m}^{m} f^2(\omega_{j+k}) \leq \frac{1}{2m+1} \max_{-m \leq k \leq m} f^2(\omega_{j+k}).$$

These two expressions exhibit clearly the importance of the smoothness of f, as well as the tradeoff between variance and bias in $\bar{f}(\omega_j)$ in the choice of m. For a fixed value of n, increasing m will increase the bandwidth and lead to including values of f at frequencies farther from the target frequency ω_j in the expected value of $\bar{f}(\omega_j)$, thus increasing the possibility of a larger bias. If f is smooth in the neighborhood of ω_j, then this bias will not increase appreciably. On the other hand, increasing m has the effect of decreasing the variance of the estimator, again assuming that f does not significantly increase as frequencies get farther from ω_j. Note that the simplest case is when X is white noise, in which case f is constant over $[0,.5]$, and we can increase the bandwidth without increasing the bias while decreasing the variance at the rate of $1/(2m+1)$. At the other extreme, we might be estimating f at a frequency that is adjacent to a band where f rises sharply to a peak, in which case increasing m could increase both the bias and the variance. This phenomenon is known as leakage as high power in f in one frequency band "leaks" into estimates of f at other frequencies.

As long as f is bounded, we can let m go to ∞ as n goes to ∞ so that the variance of $\bar{f}(\omega_j)$ goes to zero. As m and n get large, we have

$$E(\bar{f}(\omega_j)) \doteq \frac{1}{2m+1} \sum_{k=-m}^{m} f(\omega_{j+k}) \doteq \frac{n}{2m} \int_{\omega_j-m/n}^{\omega_j+m/n} f(\omega)d\omega.$$

If we let m/n get small and assume that f is such that it is roughly constant over a small interval, then the above integral is approximately $\frac{2m}{n} f(\omega_j)$, and $\bar{f}(\omega_j)$ is asymptotically unbiased.

For two different fixed frequencies $\omega_j \neq \omega_k$, we have that $\bar{f}(\omega_j)$ and $\bar{f}(\omega_k)$ will become independent as m/n goes to zero since they will eventually be based upon averaging values of the periodogram for nonoverlapping frequency bands.

We next consider the asymptotic distribution of $\bar{f}(\omega_j)$. For $\omega_{j-m} > 0$ and $\omega_{j+m} < .5$, we have

$$\bar{f}(\omega_j) \approx \frac{1}{2m+1} \sum_{k=-m}^{m} \frac{f(\omega_{j+k})}{2} Q_{k+m+1},$$

where $\approx$ means has the same distribution as, and the Q's are independent χ_2^2 distributed random variables. We know that a pure sum of independent χ^2 random variables has itself a χ^2 distribution with degrees of freedom equal to the sum of the individual degrees of freedom. Thus if $X \sim WN(\sigma^2)$, we would have that $\bar{f}(\omega_j) \approx \{\sigma^2/[2(2m+1)]\} \chi^2_{2(2m+1)}$. The distribution of a general linear combination of χ^2 random variables is often approximated by that of $Q = c\chi_\nu^2$, where c and ν are chosen so that the mean and variance of Q are the same as those of the linear combination. We will use this device to approximate the distribution of $\bar{f}(\omega_j)/f(\omega_j)$. Thus we have

$$E(Q) = c\nu = E\left(\frac{\bar{f}(\omega_j)}{f(\omega_j)}\right) \doteq 1$$

$$\text{Var}(Q) = 2c^2\nu = \text{Var}\left(\frac{\bar{f}(\omega_j)}{f(\omega_j)}\right) \doteq \frac{1}{2(2m+1)}.$$

We then set $c\nu = 1$ and $2c^2\nu = 1/2(2m+1)$ and obtain $c = 1/\nu$ and $\nu = 2(2m+1)$, which gives

$$\frac{\nu\bar{f}(\omega_j)}{f(\omega_j)} \sim \chi_\nu^2, \qquad \omega_j \neq 0,.5.$$

Note that in this simple averaging case we could have obtained this immediately from our assumption that f is constant around ω_j.

Estimating $f(0)$ and $f(.5)$

The arguments given above must be modified when we consider estimating f at frequency 0 or .5, as then the values of $\hat{f}$ on each side of the center frequency are the same. If we write, for example,

$$\bar{f}(0) = \frac{1}{2m+1}\left[\hat{f}(0) + 2\sum_{k=1}^{m}\hat{f}(\omega_{j+k})\right],$$

we find

$$\frac{\nu}{2}\frac{\bar{f}(\omega_j)}{f(\omega_j)} \sim \chi^2_{\nu/2}, \qquad \omega_j = 0, .5.$$

6.1.2. Kernels and Fourier Series Approximations

The next natural step in estimating f is to insert a weight function into the averaging of the periodogram so that values of $\hat{f}$ at frequencies far removed from ω_j get less weight than values at frequencies close to ω_j. We consider integrated weighted averages of $\hat{f}$ of the form

$$\tilde{f}(\omega) = \int_0^1 K(\omega - \tau)\hat{f}(\tau)\,d\tau,$$

where K is a weight function called a (spectral) window. For a fixed frequency ω, this expression essentially says to superimpose the weight function onto the graph of the periodogram (with its value at zero centered at frequency ω), then multiply the two functions together (which down-weights values of $\hat{f}$ far removed from the "center" frequency ω), and finally do the integral. To find the estimate at a different frequency, the same process is carried out with the weight function moved over to be centered at the new frequency. Thus the function K essentially dictates what part of $\hat{f}$ can be "seen" when finding the estimate at a certain frequency. This is the origin of the term *window*. We will study windows of various shapes and discuss the effect of their width.

Using spectral windows then is a natural extension of the idea of averaging the sample spectral density. We could use weighted averages that are sums instead of integrals, but the integrated averages arise naturally from another point of view, which we now describe. Such averages arise naturally from considerations of the general theory of the Fourier series representation of a function on a finite interval. An expression of the form

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v)e^{-2\pi i v\omega}$$

where

$$R(v) = \int_0^1 f(\omega)e^{2\pi i v\omega}\,d\omega$$

is called the Fourier series representation of f, and the coefficients $R(v)$ are called its Fourier coefficients. The periodogram

$$\hat{f}(\omega) = \sum_{v=-(n-1)}^{n-1} \hat{R}(v)e^{-2\pi i v\omega}$$

is thus actually an estimator of the nth partial sum

$$f_n(\omega) = \sum_{v=-(n-1)}^{n-1} R(v)e^{-2\pi i v\omega}.$$

From this it is clear that there are two sources of error in $\hat{f}$ as an estimator of f: a truncation error due to using f_n to approximate f, and an estimation error due to having to estimate R. In the general theory of Fourier series it is well known that the approximation of some functions by their Fourier series can be improved by applying a sequence of weights $k_n(v)$ to R in f_n, giving

$$f_{n,k}(\omega) = \sum_{v=-(n-1)}^{n-1} k_n(v)R(v)e^{-2\pi i v\omega}$$

$$= \int_0^1 f(\tau)K_n(\omega - \tau)d\tau.$$

The sequence k_n is called a lag window, and the function

$$K_n(\omega) = \sum_{v=-(n-1)}^{n-1} k_n(v)e^{-2\pi i v\omega}, \qquad \omega \in (-\infty,\infty),$$

is called a spectral window. Thus the integrated averages that we have been discussing arise naturally from the idea of applying weights to the sample autocovariances. The weights $k_n(v)$ are usually an even function of v and thus K_n is real and symmetric about .5. Further, $K_n(\omega)$ is defined for all $\omega \in (-\infty,\infty)$ and is periodic of period 1. This also means that K_n is symmetric about zero.

Definition. *A spectral estimator of the form*

$$\hat{f}_{n,k}(\omega) = \sum_{v=-(n-1)}^{n-1} k_n(v)\hat{R}(v)e^{-2\pi i v\omega}$$

$$= \int_0^1 \hat{f}(\tau)K_n(\omega - \tau)d\tau$$

is called a window estimator with lag window k_n and spectral window K_n. If $k_n(v) = 0$ for $|v| > M$ for some integer M (called a truncation point), we say that the estimator is of truncated form.

Note that the sample spectral density function itself is of the above form with $k_n(v)$ identically one, which gives that K_n is the Dirichlet kernel.

Windows of Scale Parameter Form

A variety of weight functions have been suggested, most of the scale parameter form, that is,

$$k_n(v) = \lambda(\frac{v}{M}),$$

for some integer $M < n$ called the scale parameter, with λ being a function satisfying

1. $\lambda(0) = 1$.

2. $\lambda(-u) = \lambda(u)$.

3. $\int_{-\infty}^{\infty} \lambda^2(u)du < \infty$.

368

Definition. *The function* λ *defined above is called a lag window generator, while the function*

$$\Lambda(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \lambda(u) e^{iu\omega} du, \qquad \omega \in (-\infty, \infty),$$

is called the spectral window generator corresponding to λ.

For lag windows of scale parameter form, it can be shown for large M that

$$K_n(\omega) \doteq 2\pi M \Lambda(2\pi M \omega).$$

In Table 6.1 we give information about eight of the lag and spectral window generators that are commonly used. The last five columns in the table contain numerical quantities that are important for judging the adequacy of the window. We describe these quantities below.

The first five lag window generators are zero outside of $|u| \leq 1$ and thus $k_n(v) = 0$ for $|v| > M$ and $\hat{f}_{n,k}$ only involves values of $\hat{R}(v)$ for $|v| \leq M$, and M is then called a truncation point. This terminology is often used even when λ is not of truncated form.

The basic feature of the windows is that they decay rapidly from frequency zero and then rise back up again in what are called sidelobes. All of the windows converge to a delta function as M increases. This is analogous to letting the bandwidth of the simple averaging smoother of the previous section go to zero. The truncated periodogram, Tukey, and Parzen-Cogburn-Davis windows are negative in certain frequency bands, while the others are nonnegative. This is important because using a negative spectral window can result in a spectral estimator that is negative.

Since the main lobe of a window is not rectangular, it is difficult to measure how wide it is. In the last column of Table 6.1 we give what is called Parzen's measure of the bandwidth of a window, denoted B_P and defined to be the width of a rectangular window having the same value at $\omega = 0$ as K_n and the same area as K_n. Since the areas of all of the K_n are one, we have

$$B_P = \frac{1}{2\pi M \Lambda(0)} \doteq \frac{1}{K_n(0)}.$$

In Figure 6.1, we have superimposed the Parzen and Tukey windows for truncation points 24 and 18, respectively. Note how the main lobes seem to line up, which could be predicted based on the bandwidths of the two windows.

6.1.3. Sampling Properties of Estimators

As we have pointed out, the properties of a spectral estimator based on smoothing $\hat{f}$ will depend on the smoothness of f and the properties of the weighting function being used. In the case of a lag window of scale parameter form, the properties of the weighting function will depend on the scale parameter M and the lag window generator λ.

| **Theorem 6.1.1** | PROPERTIES OF WINDOW ESTIMATORS |

Let $X(1), \ldots, X(n)$ *be a sample realization from a covariance stationary time series* X *which has spectral density function* f *and let* λ *be a lag window generator. Let*

$$\hat{f}_{\lambda, M}(\omega) = \sum_{v=-\infty}^{\infty} \lambda\left(\frac{v}{M}\right) \hat{R}(v) e^{-2\pi i v \omega}$$

where M *is chosen as a function of* n *so that* $M \to \infty$ *as* $n \to \infty$. *Note that the limits on the sum are actually* $-M$ *to* M *if* λ *is of truncated form and* $-(n-1)$ *to* $n-1$ *otherwise. Then*

Table 6.1. Some Window Generators and Their Characteristics

| $\Lambda(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\lambda(u)e^{iu\omega}du$	q	$\lambda^{(q)}$	$\int_{-\infty}^{\infty}\lambda^2(u)du$	E_{min}	B_p
Generators of Truncated Form, i.e., $\lambda(u)=0$, $\|u\|>1$					
Truncated Periodogram Window: $\lambda(u)=1$					
$\frac{1}{\pi}\frac{\sin\omega}{\omega}$	∞	–	2	–	$\frac{1}{2}\frac{1}{T}$
Bartlett Window: $\lambda(u)=1-\|u\|$					
$\frac{1}{2\pi}\left(\frac{\sin\omega/2}{\omega/2}\right)^2$	1	1	$2/3$	$2/3$	$\frac{1}{T}$
Tukey Window: $\lambda(u)=0.54+0.46\cos\pi u$					
$\frac{1}{2\pi}\frac{\sin\omega}{\omega}\frac{\pi^2}{\pi^2-\omega^2}$	2	$0.23\pi^2$	0.795	1.20	$\frac{1}{T}$
Parzen Window: $\lambda(u)=\begin{cases}1-6u^2+6\|u\|^3, & \|u\|\le .5\\ 2(1-\|u\|)^3, & .5\le u\le 1\end{cases}$					
$\frac{3}{8\pi}\left(\frac{\sin\omega/4}{\omega/4}\right)^4$	2	6	0.539	1.32	$\frac{4}{3}\frac{1}{T}$
Bohman Window: $\lambda(u)=(1-\|u\|)\cos\pi u+\frac{1}{\pi}\sin\pi\|u\|$					
$\frac{2\pi(1+\cos\omega)}{(\pi^2-\omega^2)^2}$	2	$\pi/2$	0.586	0.734	$1.23\frac{1}{T}$
Generators Not of Truncated Form					
Daniell Window: $\lambda(u)=\frac{\sin\pi u}{\pi u}$					
$\frac{1}{2\pi}$, $\|\omega\|\le\pi$	2	$\pi^2/6$	1	1.28	$\frac{1}{T}$
Bartlett-Priestley Window: $\lambda(u)=\frac{3}{(\pi u)^2}\left(\frac{\sin\pi u}{\pi u}-\cos\pi u\right)$					
$\frac{3}{4\pi}\left(1-\left(\frac{\omega}{\pi}\right)^2\right)$, $\|\omega\|\le\pi$	2	$\pi^2/10$	1.2	1.19	$\frac{4}{3}\frac{1}{T}$
Parzen-Cogburn-Davis Window: $\lambda(u)=\frac{1}{1+u^{2r}}$, $s=\pi/2r$					
	$2r$	1	$\frac{s}{\sin s}\frac{2r-1}{r}$		
For $r=2$:					
$\frac{1}{2}e^{-\|\omega\|/\sqrt{2}}\sin\left(\frac{\|\omega\|}{\sqrt{2}}+\frac{\pi}{4}\right)$	4	1	1.66	1.66	$.45\frac{1}{T}$
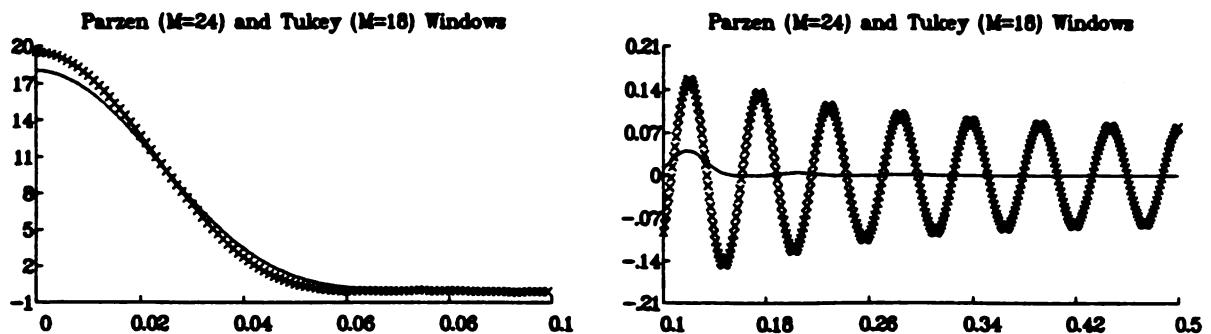
Figure 6.1. The Parzen (solid curve) and Tukey (x's) Spectral Windows for Comparable Bandwidths.

a) *The estimator $\hat{f}_{\lambda,M}(\omega)$ is asymptotically unbiased.*

b) ASYMPTOTIC VARIANCE. *If M is chosen so that $M \to \infty$ and $M/n \to 0$ as $n \to \infty$, then*

$$\lim_{n \to \infty} \frac{n}{M} \mathrm{Var}(\hat{f}_{\lambda,M}(\omega)) = \begin{cases} 2f^2(\omega) \int_{-\infty}^{\infty} \lambda^2(u)du, & \omega = 0, .5 \\[2mm] f^2(\omega) \int_{-\infty}^{\infty} \lambda^2(u)du, & \omega \neq 0, .5 \end{cases}$$

$$\lim_{n \to \infty} \frac{n}{M} \mathrm{Cov}(\hat{f}_{\lambda,M}(\omega_1), \hat{f}_{\lambda,M}(\omega_2)) = 0, \qquad \omega_1 \neq \omega_2.$$

c) *Confidence intervals for $f(\omega)$ are given by*

$$\hat{f}_{\lambda,M}(\omega) \pm Z_{\alpha/2} \sqrt{\frac{M}{N} \hat{f}_{\lambda,M}^2(\omega) \int_{-\infty}^{\infty} \lambda^2(u)du},$$

$$(\frac{1}{c}\hat{f}_{\lambda,M}(\omega), c\hat{f}_{\lambda,M}(\omega)),$$

where $c = \exp(Z_{\alpha/2}\sqrt{\frac{M}{n} \int_{-\infty}^{\infty} \lambda^2(u)du})$, *and*

$$\left(\frac{\nu \hat{f}_{n,k}(\omega)}{\chi^2_{\alpha/2}}, \frac{\nu \hat{f}_{n,k}(\omega)}{\chi^2_{1-\alpha/2}} \right).$$

These intervals are for f at a single frequency. A large sample simultaneous confidence band on f for all frequencies can be obtained by multiplying both terms in the second confidence interval by $\exp((2\log M)^{1/2})$.

The Choice of M

The choice of M determines the amount of smoothing done, with too large (small) a value resulting in undersmoothing (oversmoothing). In general the choice of M is very difficult unless some information about the function being estimated is known. Basically, if f has a narrow peak, we would like the bandwidth of the spectral window to be narrow so that leakage doesn't occur. This however leads to undersmoothing in other frequencies. Thus nonparametric spectral estimators have trouble "resolving" peaks without introducing spurious peaks in other ranges. We will see later that parametric spectral estimators can solve this problem in many cases. In any event, the prevailing view on the choice of M is to try more than one value and use

the resulting plots to make general statements about f. For example, one could use three values, M_1, M_2, and M_3, where

$$.05 \leq \frac{M_1}{n} \leq .10, \qquad .10 \leq \frac{M_2}{n} \leq .25, \qquad .25 \leq \frac{M_3}{n} \leq .75.$$

6.2. Parametric Spectral Density Estimation

In Section 6.1 we considered nonparametric methods for estimating the spectral density of a covariance stationary time series. We saw that (1) the choice of the scale parameter M is the key to finding a good estimator, and (2) it is difficult to produce an estimator having sharp peaks without introducing possibly spurious peaks as well. In this section, we consider using the spectral density of an ARMA model as an approximation to that of the process being analyzed. We will see that this parametric spectral density estimation procedure is easy to implement and under a wide variety of conditions leads to estimators superior to window estimators.

6.2.1. Autoregressive Spectral Estimation

In this section we assume that the process being observed can be written as an infinite order autoregressive process. This includes the cases of a finite order autoregressive process and an invertible MA or ARMA process.

Theorem 6.2.1 CONDITIONS FOR AR(∞) REPRESENTATION

If X is a covariance stationary time series having a spectral density satisfying

$$0 < \lambda_1 \leq f(\omega) \leq \lambda_2 < \infty, \qquad \omega \in [0,1],$$

for some constants λ_1 and λ_2, then X can be written as an infinite order AR process.

All ARMA processes except those that have a zero of the moving average polynomial on the unit circle have spectral densities that satisfy the conditions of this theorem. We consider approximating the true spectral density f of a process that can be written as an infinite order autoregression by that of a pth order AR process. Given data $X(1),\ldots,X(n)$, autoregressive spectral estimation consists of three steps: (1) determining the order k of the best approximating AR spectral density f_k, (2) finding estimates $\hat{\alpha}_k(1),\ldots,\hat{\alpha}_k(k)$ and $\hat{\sigma}_k^2$ of the parameters of f_k, and (3) estimating f by

$$\hat{f}_k(\omega) = \frac{\hat{\sigma}_k^2}{\left| \sum_{j=0}^{k} \hat{\alpha}_k(j) e^{2\pi i j \omega} \right|^2}.$$

To determine the order, we can use any of the methods described in Lecture 4. However, Parzen's CAT criterion was specifically proposed in the context of autoregressive spectral estimation, so we describe it here. The basic motivation for the criterion is contained in the next theorem.

Theorem 6.2.2 PROPERTIES OF CAT

Let $\hat{\sigma}_p^2$ be the error variance and

$$\hat{a}_p(z) = \sum_{j=0}^{p} \hat{\alpha}_p(j) z^j$$

372

be the transfer function of the estimated AR process of order p based on a realization of length n from a process which can be written as an infinite order autoregressive process having coefficients $\alpha_1, \alpha_2, \ldots$, transfer function

$$g_\infty(z) = \sum_{j=0}^{\infty} \alpha_j z^j,$$

and error variance σ_∞^2. Let $CAT(p)$ be the CAT criterion for order p. Then

$$\lim_{n \to \infty} E(CAT(p)) = \lim_{n \to \infty} J_p,$$

where

$$J_p = \int_0^1 E\left(\left|\frac{1}{\hat{\sigma}_p^2}\hat{g}_p - \frac{1}{\sigma_\infty^2}g_\infty\right|^2\right) f(\omega)\, d\omega.$$

This result shows that the CAT criterion chooses the order that results asymptotically in the spectral estimator that is closest (in the sense of the integrated relative mean square error measured by J_p) to the true $AR(\infty)$ transfer function. This is the origin of the name "criterion autoregressive transfer function."

In the next theorem we present the basic sampling properties of the autoregressive spectral estimator.

Theorem 6.2.3 | PROPERTIES OF AR SPECTRAL ESTIMATION

Let X be a covariance stationary time series that can be expressed as an infinite order autoregressive process

$$\sum_{j=0}^{\infty} \alpha_j X(t - j) = \epsilon(t),$$

where the errors ϵ are independent and satisfy $E(X^4(t)) < \infty$. Let $\hat{f}_p$ be the autoregressive spectral estimator of the spectral density f based on a realization of length n. Then

a) If p is chosen so that

$$i)\ p \to \infty, \quad ii)\ \frac{p^3}{n} \to 0, \quad \text{and} \quad iii)\ \sqrt{n} \sum_{j=p+1}^{\infty} |\alpha_j| \to 0,$$

as $n \to \infty$, then for any k fixed frequencies $0 < \omega_1 < \cdots < \omega_k < 0.5$, the joint asymptotic distribution of

$$\sqrt{n/p}(\hat{f}_p(0) - f(0)),\ \sqrt{n/p}(\hat{f}_p(\omega_1) - f(\omega_1)), \ldots,$$

$$\sqrt{n/p}(\hat{f}_p(\omega_k) - f(\omega_k)),\ \sqrt{n/p}(\hat{f}_p(0.5) - f(0.5)),$$

is that of independent, zero mean, normal random variables having variances

$$4f^2(0), 2f^2(\omega_1), \ldots, 2f^2(\omega_k), 4f^2(.5).$$

b) If X is a finite order autoregressive process, then letting p be any function g of n that satisfies $p \to \infty$ and $p^3/n \to 0$ as $n \to \infty$ satisfies the requirements of part (a), and the variance of $\hat{f}_p(\omega)$ goes to zero at the rate $g(n)/n$ as $n \to \infty$.

c) If X is an invertible ARMA process, then $p = \log n$ satisfies the requirements of part (a), and the variance of $\hat{f}_p(\omega)$ goes to zero at the rate of $\log n/n$ as $n \to \infty$.

Implications: This theorem shows that if we are estimating the spectral density of a process that can be written as an infinite order AR, then the autoregressive spectral estimator is consistent and asymptotically normal as long as the order that we use is chosen in the correct way. Requirement (iii) for the choice of p shows that the choice depends on how rapidly the coefficients of the AR(∞) go to zero. For a finite order AR process, these coefficients become zero past the order of the process, and we can choose p to go to infinity as slowly as we want. For an invertible MA or ARMA process, we could show that the coefficients of the AR(∞) representation go to zero at a rate bounded by ρ^k for some constant ρ and thus choosing $p = \log n$ is sufficient.

The next theorem (due to Newton and Pagano (1984)) gives confidence bands for AR(p) spectra.

Theorem 6.2.4 | CONFIDENCE BANDS FOR AR SPECTRA

If X is a Gaussian AR(p) process, then asymptotically the probability is $1 - \alpha$ that the true spectral density f lies entirely within the Scheffe-type bands

$$\frac{1}{\hat{h}_p(\omega) + s(\omega)} \leq f(\omega) \leq \frac{1}{\hat{h}_p(\omega) - s(\omega)}, \qquad \omega \in [0, 1],$$

where

$$\hat{h}_p(\omega) = \frac{1}{\hat{f}_p(\omega)} = \hat{\gamma}(v) + 2\sum_{v=1}^{p} \hat{\gamma}(v)\cos 2\pi v \omega$$

$$\hat{\gamma}(v) = \frac{1}{\hat{\sigma}_p^2}\sum_{j=0}^{p-v} \hat{\alpha}_p(j)\hat{\alpha}_p(j+v), \qquad v = 0,\ldots,p$$

$$s^2(\omega) = \frac{\chi_{\alpha,p+1}^2}{n}\mathbf{x}^T(\omega)\hat{\mathbf{D}}\mathbf{x}(\omega)$$

and

$$\mathbf{x}^T(\omega) = (1, 2\cos 2\pi\omega, \ldots, 2\cos 2\pi p\omega)$$

$$\mathbf{D} = \mathbf{BCB}^T$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{S}(\alpha) & \mathbf{0}_p \\ \mathbf{0}_p^T & 2/\sigma^4 \end{bmatrix}$$

$$B_{j,k} = \begin{cases} \sigma^2\gamma(j-1), & k = p+1, \quad j = 1,\ldots,p+1 \\ \sigma^{-2}(\alpha_{k+j-1} + \alpha_{k-j+1}), & k = 1,\ldots,p, \quad j = 1,\ldots,p+1, \end{cases}$$

$S(\alpha)$ *is the Schur matrix corresponding to α, and in forming $\hat{\mathbf{D}}$ estimators of α and σ^2 are substituted for the true values. For any frequency ω where $\hat{h}(\omega) - s(\omega)$ is negative, the upper limit is taken to be ∞.*

Lynx Data Divided by 1000


Wolfer Sunspot Numbers, 1749-1963
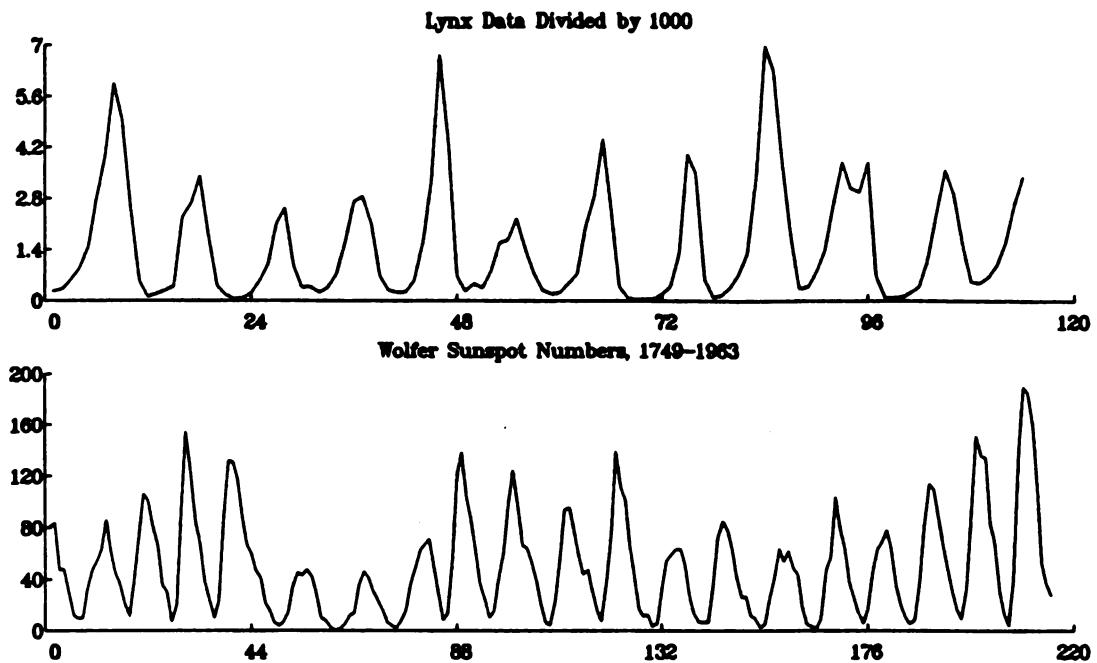
Figure 6.2. The Lynx and Sunspot Data.

6.2.2. MA and ARMA Spectral Estimation

The next natural step in spectral estimation is to use MA and ARMA models to obtain estimators. We note that MA spectral estimation is very similar to nonparametric spectral estimation in that it results in an estimator that is a finite-degree trigonometric polynomial. Thus it suffers from the same difficulty in estimating spectra that have sharp peaks. It does have the advantage of having the AIC available to aid in choosing the degree of the polynomial.

The use of ARMA models to estimate spectra does not have the weakness that nonparametric and moving average methods have and much attention has been paid recently to ARMA spectral estimation However, unless the true spectral density has both sharp peaks and troughs, autoregressive spectral estimation, with its computational simplicity, should be adequate.

6.3. Methods for Determining Periodicities

Several of the most famous time series that have been studied in the past have appeared to contain cyclical components. For example, in Figure 6.2, we display two data sets: (1) the annual number of Canadian lynx trapped on the Mackensie River for the years 1821-1934, and (2) the annual average value of the daily index of the number of sunspots (using the scale devised by Professor Rudolf Wolf in 1849) for the years 1755-1964. These data sets have been extensively studied over the past several years (see for example Part 4 of the 1977 volume of the *Journal of the Royal Statistical Society, Series A*). The basic property of these data is that there appear to be cyclic patterns but that these patterns are not perfectly cyclic. The "sunspot cycle" is a well known phenomenon and has an important effect on radio communications. The cycle in the lynx data is usually explained as being due to a predator-prey relationship between the Canadian lynx and the snowshoe hare, its most important source of food.

Time series such as the lynx and sunspot data are traditionally analyzed according to one of three models:

1. As the sum of a deterministic sinusoid plus an error that is a covariance stationary time series, that is,

$$X(t) = \mu + a\cos(\frac{2\pi t}{p}) + b\sin(\frac{2\pi t}{p}) + \epsilon(t),$$

where the error series ϵ is covariance stationary, and p is the period of the sinusoid.

2. As the values of a stochastic process referred to by names such as an 'outburst' model or a filtered Poisson process. The basic idea of this type of model is that a physical process builds up and then has a large surge or outburst, followed by a decay back to some baseline value. We will not consider this model further.

3. As the values of a covariance stationary time series of near cyclic type. For example, the AR(2) model

$$X(t) + \alpha_1 X(t-1) + \alpha_2 X(t-2) = \epsilon(t)$$

will appear roughly cyclic with period p if $\alpha_1 = -2\cos(2\pi/p)$ and α_2 is close to 1.

6.3.1. Deterministic Sinusoid Plus Error

If we believe that a time series is actually a deterministic sinusoid observed with additive error and we know the period of the sinusoid, then we can use regression analysis to estimate the coefficients of the sinusoid and test whether the amplitude is in fact zero. If the errors are uncorrelated, then ordinary least squares can be used, while if the errors are correlated, we can use the procedure described in Section 5.2 for doing regression with autoregressive errors. In fact, in Section 5.2 we used the lynx data to illustrate the procedure.

Estimating a Period

If one has data $X(1), \ldots, X(n)$ and suspects that they contain a deterministic sinusoid of some unknown frequency plus noise, then this problem is often referred to as a search for hidden periodicities. It seems natural to inspect the periodogram of the data and test whether the largest value of the periodogram is significantly different from zero. If not, then we would conclude that there are no deterministic sinusoids of any of the natural frequencies in the data. If the largest value is significantly different from zero, then we must ask whether it could happen that there is a sinusoid but it is actually at a different frequency. To do this test, we can use the results of the next theorem. In this theorem we assume that n is odd. The results are approximately correct if n is even.

| Theorem 6.3.1 | FISHER'S EXACT TEST |

Let $\hat{f}(\omega_j)$ be the periodogram of a realization of length n (n odd) from the process

$$X(t) = a\cos\frac{2\pi(t-1)}{p} + b\sin\frac{2\pi(t-1)}{p} + \epsilon(t),$$

where the period p is a factor of n, and ϵ is a Gaussian white noise series. Let $m = [n/2]$. Then

a) If $a = b = 0$, the exact distribution of

$$g = \frac{\max\limits_{j=2,[n/2]}(\hat{f}(\omega_j))}{\sum\limits_{j=2}^{[n/2]} \hat{f}(\omega_j)}$$

Log10 of Lynx Data   Lynx data, Period CI:(9.323975,10.05591)
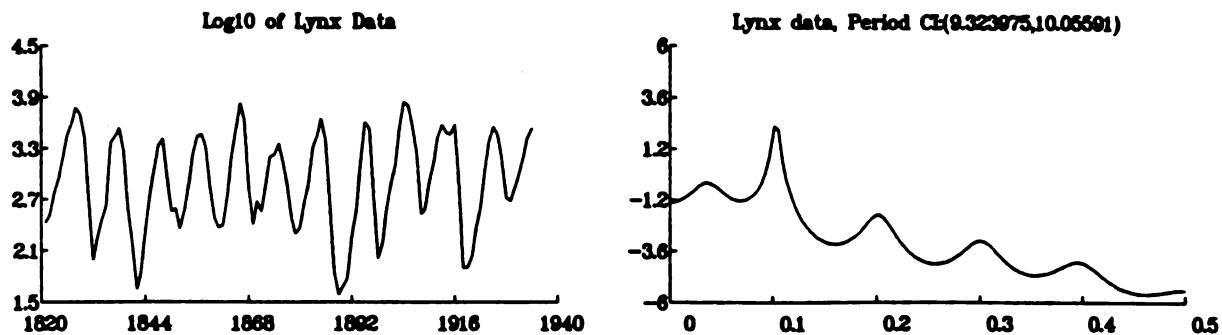
Figure 6.3. Log (base 10) of the Lynx Data and the Estimated AR(11) Spectral Density.

is given by

$$\Pr(g > z) = \sum_{j=1}^{K} (-1)^{j-1} \binom{m}{j} (1 - jz)^{m-1}, \qquad z > 0,$$

where $K = [1/z]$.

b) The chance that the test given in part (a) will choose a frequency other than $1/p$ is less than the level of significance α that is used in the test.

The test in part (a) is called Fisher's exact test. For the log of the lynx data, we find that $g = 0.5967$ which has a p-value less than 10^{-4}, and the maximum is at period $114/12 = 9.5$. If one suspects the existence of r sinusoids, then the above test can be modified.

A crucial assumption in the Fisher test is that the errors are white noise. If they are not, that is, the spectral density of the error time series is not a constant, then a large value of the periodogram at a particular frequency could be due to either (1) the existence of a sinusoid having that frequency, (2) the spectral density being large at that frequency, or (3) a combination of the two. This is referred to as the mixed spectrum case (see Section 6.3 of Priestley (1981)), and unless one has some prior knowledge of the nature of f or the sinusoids involved, separating the two parts of the model is very difficult.

6.3.2. Estimating Peak Frequencies in AR Spectra

For a process of near cyclic type, the data appear to be cyclic except that the lengths of cycles vary from one cycle to the next. Thus such a model is often referred to as a disturbed periodicity model, and the analog of determining the period of a deterministic sinusoid is to determine the frequency ω^* where the spectral density of the process has a peak. Searching for peak frequencies is not difficult if the process is a finite order MA or AR process as then f (or its reciprocal in the AR case) is a finite-degree trigonometric polynomial and finding the critical values (maxima and minima) of such a polynomial is not difficult. Given a realization of length n from an AR(p) process, we can estimate the peak frequency ω^* by finding the frequency where the reciprocal of $\hat{f}$ has a relative minimum. We will denote this estimator by $\hat{\omega}$. If the order p is unknown, then we can estimate it and the coefficients of the estimated order process and again use the process described above to estimate the peak frequency. This estimator is denoted by $\hat{\omega}_{\hat{p}}$. The properties of these estimators are given in the next theorem.

| Theorem 6.3.2 | ESTIMATING A PEAK FREQUENCY |

Let $\hat{\omega}$ and $\hat{\omega}_f$ be the autoregressive spectral estimates described above of a peak frequency ω^* based on a realization of size n from a Gaussian $AR(p, \alpha, \sigma^2)$ process. Then

a) $\sqrt{n}(\hat{\omega} - \omega^*) \xrightarrow{\mathcal{L}} N(0, \sigma^2(\omega^*))$, where

$$\sigma^2(\omega^*) = \frac{b^T(\omega^*)C(\alpha)S(\alpha)C^T(\alpha)b(\omega^*)}{\left(h''(\omega^*)/2\pi\right)^2},$$

where $S(\alpha)$ is the Schur matrix corresponding to α,

$$b^T(\omega) = (\sin 2\pi\omega, 2\sin 4\pi\omega, \ldots, p\sin 2\pi p\omega),$$

$h(\omega) = \sigma^2/f(\omega)$, and $C(\alpha)$ is the $p \times p$ matrix having (j, k)th element

$$C_{j,k}(\alpha) = \alpha_{j+k} + \alpha_{k-j},$$

with $\alpha_v = 0$ if $v > p$ or $v < 0$.

b) *If a consistent order-determining criterion is used for finding $\hat{p}$, then the results of part (a) continue to hold for $\hat{\omega}_f$.*

c) *If an order-determining criterion that is guaranteed asymptotically to not underestimate the order is used to find $\hat{p}$, then $\hat{\omega}_f$ is a consistent estimator of ω^*.*

From these results, we can find a large sample confidence interval for ω^* as

$$\hat{\omega} \pm Z_{\alpha/2}\sigma(\hat{\omega}),$$

while the lower and upper limits in a confidence interval for the reciprocal of ω^*, that is, for the period of the peak, are given by the reciprocals of the upper and lower limits for the frequency. Notice the important role played by the second derivative of h in the asymptotic variance of $\hat{\omega}$. For a sharp peak, h'' will be large and thus the confidence interval will be narrow. For a broad peak, the interval will be wide.

A crucial part of the method of this section is the fact that we have been finding a zero of a finite-degree polynomial. Thus if f is of the form of an ARMA spectral density, that is, as the ratio of two finite-degree polynomials, we cannot apply the above procedure. However, as long as f can be expressed as the spectral density of either an $AR(\infty)$ or $MA(\infty)$ process, we should be able to apply the above procedure and obtain asymptotically good properties.

To illustrate the use of this procedure, we consider the lynx data again. While there is no clear agreement among analysts what AR process best models the log of the lynx data, we consider the result of using the order 11 process determined in Section 5.2. In Figure 6.3 we give plots of the log (base 10) of the data and the estimated AR(11) spectral density. The confidence interval for the period corresponding to the largest peak in the spectral density is (9.32, 10.05) years.

378

REFERENCES

Anderson, T.W. (1971). "The Statistical Analysis of Time Series." Wiley, New York.

Ansley, C.F. (1979). An algorithm for the exact likelihood of a mixed autoregressive moving average process. *Biometrika*, 66, 59–65.

Bartlett, M.S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37, 1–16.

Bloomfield, P. (1976). "Fourier Analysis of Time Series: An Introduction." Wiley, New York.

Box, G.E.P., and Jenkins, G.M. (1970). "Time Series Analysis, Forecasting, and Control." Holden-Day, San Francisco.

Brillinger, D.R. (1975). "Time Series Analysis: Data Analysis and Theory." Holt, Rinehart & Winston, New York.

Brockwell, P.J., and Davis, R.A. (1987). "Time Series: Theory and Methods." Springer-Verlag, New York.

Chatfield, C., and Prothero, D.L. (1973). Box-Jenkins seasonal forecasting: problems in a case study. *J. Roy. Statist. Soc. Ser. A*, 136, 295–336.

Cooley, J.W., and Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19, 297–301.

Fuller, W.A. (1976). "Introduction to Statistical Time Series." Wiley, New York.

Gentleman, W.M., and Sande, G. (1966). Fast Fourier transforms–for fun and profit. AFIPS, 1966 Fall Joint Computer Conference, *Conf. Proc.*, 28, 563–578.

Grenander, U., and Rosenblatt, M. (1957). "Statistical Analysis of Stationary Time Series." Wiley, New York.

Jenkins, G.M., and Watts, D.G. (1968). "Spectral Analysis and Its Applications." Holden-Day, San Francisco.

Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389–395.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1984). "The Forecasting Accuracy of Major Time Series Methods." Wiley, New York.

Meinhold, R.J., and Singpurwalla, N.D. (1983). Understanding the Kalman filter. *Amer. Statist.*, 37, 123–127.

REFERENCES

Melard, G. (1984). A fast algorithm for the exact likelihood of autoregressive-moving average models. *Appl. Statist.*, 33, 104–114.

Newton, H.J., and Pagano, M. (1983a). A method for determining periods in time series. *J. Amer. Statist. Assoc.*, 78, 152–157.

Newton, H.J., and Pagano, M. (1984). Simultaneous confidence bands for autoregressive spectra. *Biometrika*, 71, 197-202.

Parzen, E. (1957a). On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Statist.*, 28, 329-348.

Priestley, M.B. (1981). "Spectral Analysis and Time Series," Vols. 1 and 2. Academic Press, New York.

Shibata, R. (1981). An optimal autoregressive spectral estimate. *Ann. Statist.*, 9, 300-306.

Tong, H. (1977). Some comments on the Canadian lynx data. *J. Roy. Statist. Soc. Ser. A*, 140, 432–436.

Ulrych, T.L., and Bishop, T.N. (1975). Maximum entropy spectral analysis and autoregressive decomposition. *Rev. Geo. Space Phys.*, 14, 183–200.

Whittle, P. (1963a). "Prediction and Regulation." English Universities Press, London.

Whittle, P. (1963b). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50, 129–134.

9th Annual Army Conference on Applied Mathematics Attendees

Name	Address	Phone	Email
George F. Adams	Director U.S. Army Ballistic Research Laboratory ATTN: George F. Adams Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6197	qta@brl.mil
Adair R. Aguiar		623-9096	.
Gerald R. Andersen	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4253	jerry@brl.mil
Bruce Anderson	Cornell University Mathematical Sciences Institute Department of Mathematics, White Hall Ithaca, NY 14853	607-272-6132	anderson@mssun7.msi.cornell.edu
A. Arvind	Massachusetts Institute of Technology Laboratory for Computer Science 545 Technology Square Cambridge, MA 02139		
Steven F. Ashby	Lawrence Livermore National Laboratory Computing & Mathematics Research Group Mail Station L-316 P.O. Box 808 Livermore, CA 94551	415-423-2462	ashby@lll-crg.llnl.gov
Donald Austin	Executive Director Army High Performance Computing Research Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	austin@ahpcrc.umn.edu
Harry Auvermann	Commander U.S. Army Atmospheric Sciences Lab. ATTN: SLCAS-AR-l (Dr. Harry Auvermann) White Sands Missile Range, NM 88002-5501	505-678-4224	
Marco Avellaneda	Courant Institute 251 Mercer Street New York, NY 10012	212-998-3141	avellane@math1.nyu.edu
Kenneth A. Bannister	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. Kenneth A. Bannister Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6121	kab@brl.mil
Romesh C. Batra	Department of Mechanical and Aerospace Engineering and Engineering Mechanics University of Missouri-Rolla Rolla, MO 65401-0249	314-341-4589	c2980@umrvmb.bitnet
Michael J. Belczynski	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (Michael J. Belczynski) Warren, MI 48397-5000	313-574-7816	
Simeon M. Berman	New York University Warren Seaver Hall 251 Mercer Street New York, NY 10012	212-998-3001	
Dimitris Bertsimas	Alfred P. Sloan School of Management Massachusetts Institute of Technology 50 Memorial Drive, E53-359 Cambridge, MA 02139		
Adam W. Bojanczyk	School of Electrical Engineering Cornell University Phillips Hall Ithaca, NY 14853	607-255-4296	adamb@ee.cornell.edu

Name	Address	Phone	Email
Daniel Boley	Department of Computer Science University of Minnesota Minneapolis, MN 55455	612-625-3887	boley@cs.umn.edu
Deborah Brandon	Carnegie Mellon University Mathematics Department Pittsburgh, PA 15213	412-268-2545	
Michael Brewer	Department of Mechanical Engineering Colorado State University Fort Collins, CO 80523	303-491-7479	
Roger Brockett	Division of Applied Sciences Harvard University Cambridge, MA 02138	617-495-3922	brockett@gramian.harvard.edu
Paul Broome	Director U.S. Army Ballistic Research Lab ATTN: SLCBR-SECAD (Dr. Paul Broome) Aberdeen Proving Ground, MD 21005-5066	301-278-6884	broome@brl.mil
Mel Brown	U. S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4336	brown@aro-emh1.army.mil
Jagdish Chandra	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4254	sso@aro-emh1.army.mil
Lang-Mann Chang	Director U.S. Army Ballistic Research Laboratory ATTN: SLCBR-IB-P (Dr. Lang-Mann Chang) Aberdeen Proving Ground, MD 21005-5066		
Peter C.T. Chen	Research Mechanical Engineer Benet Laboratories ATTN: SMCAR-CCB-RA Watervliet, NY 12189-4050	518-266-5907	
Shih C. Chu	Commander U.S. Army Armament R&D Center ATTN: SMCAR-CCL-EM (Dr. Shih C. Chu) Light Armament Division, CCAC Picatinny Arsenal, NJ 07806-5000	201-724-7316	pica shihchu
Kenneth D. Clark	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4256	clark@adm.csc.ncsu.edu
Susan Coates	Director U.S. Army Ballistic Research Laboratory ATTN: Susan Coates Vulnerability/Lethality Division Aberdeen Proving Ground, MD 21005-5066	301-278-6710	scoates@brl.mil
Norman Coleman, Jr.	ARDEC ATTN: SMCAR-FSF-RC Picatinny Arsenal, NJ 07806-5000	201-724-6275	ncoleman@pica.army.mil
Terry Cronin	Director U.S. Army CECOM Center of Signals Warfare ATTN: AMSEL-RD-SW-TRI (Mr. Terry Cronin) Vint Hill Farms Station Warrenton, VA 22186-5100	703-349-6939	
Benjamin E. Cummings	Director U.S. Army Human Engineering Lab ATTN: Dr. Benjamin E. Cummings Aberdeen Proving Ground, MD 21005-5066	301-278-5811	bigben@brl.mil

382

Yuefan Deng
State University of New York at Stony Brook
Applied Mathematics Department
StonyBrook, NY 11794-3600 .   516-632-8614

Keith Dennis
Cornell University
Mathematical Sciences Institute
Department of Mathematics, White Hall
Ithaca, NY 14853   607-255-4027   dennis@mssun7.msi.cornell.edu

Max Donath
University of Minnesota
Department of Mechanical Engineering
111 Church St. SE
Minneapolis, MN 55455   612-625-2304   donath@vx.acs.umn.edu

Devdatt P. Dubhashi
Cornell University
Mathematical Sciences Institute
Department of Computer Science, White Hall
Ithaca, NY 14853   607-255-9206   dubhashi@cs-cornell.edu

Godi Fischer
University of Rhode Island
Kelly Hall Annex
Room A-109
Kingston, RI 02881   401-792-5879   fischer@quahog.uri.edu

Donald French
Department of Mathematical Sciences
University of Cincinnati
Old Chemistry Building (ML 25)
Cincinnati, OH 45221-0025   513-556-4039   french@ucunix.san.uc.edu

James Glimm
State University of New York at Stony Brook
Department of Applied Mathematics and Statistics
Stony Brook, NY 11794-3600   516-632-8355   glimm@ams.sunysb.edu

Aaron Das Gupta
Director
U.S. Army Ballistic Research Laboratory
ATTN: SLCBR-TB-B (Dr. Aaron Das Gupta)
Aberdeen Proving Ground, MD 21005-5066 301-278-6026   dasgupta

Morton E. Gurtin
Department of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213-3890   412-268-2545

Harumi Hattori
West Virginia University
Department of Mathematics
Morgantown, WV 26506   304-293-2014   u1493@wvnvm.bitnet

Robert Heyman
U.S. Army Materials Technology Laboratory
ATTN: SLCMT-MRD
Arsenal Street
Watertown, MA 02172-0001   617-923-5274   rheyman@watertown-emh1.army.mil

Jieh Hsiang
Department of Computer Science
State University of New York
Stony Brook, NY 11794-4400   516-632-8449   hsiang@sbcs.sunysb.edu

Bill Hrusa
Carnegiie Mellon University
Mathematics Department
Pittsburgh, PA 15213   412-268-8487

John S. Hurley
Hampton University
Department of Mathematics
Hampton, VA 23668-00101   804-727-5352   hurley@hurley.math.hamptonu.edu

E. F. Infante
Dean, Institute of Technology
University of Minnesota
107 Walter Library
Minneapolis, MN 55455   612-624-2006   infante@mailbox.mail.umn.edu

Richard James
University of Minnesota
Department of Aerospace Engineering
Minneapolis, MN 55455   612-625-0706   rdjames@umnacvx

Arthur Johnson
Commander
U.S. Army Materials & Mechanics
  Research Center
ATTN: AMXMR-SMM (Dr. Arthur Johnson)
Watertown, MA 02172   617-923-5272

Name	Address	Phone	Email
Moon S. Jun	Physical Science Laboratory New Mexico State University Box 30002 Las Cruces, NM 88003-0002	505-522-9137	mjun@dante.nmsu.edu
Professor G. Kallianpur	Department of Statistics University of North Carolina-Chapel Hill Chapel Hill, NC 27514	919-962-2187	
R. L. Kashyap	Purdue University Department of Electrical Engineering W. Lafayette, IN 47907	317-494-3437	kashyap@ecn.purdue.edu
Kent D. Kimsey	Director U.S. Army Ballistic Research Laboratory ATTN: Kent D. Kimsey Terminal Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6083	kimsey@brl.mil
Robert V. Kohn	Courant Institute 251 Mercer Street New York, NY 10012	212-998-3217	kohn@math1.nyu.edu
Professor G. S. Ladde	Department of Mathematics University of Texas at Arlington Box 19408 Arlington, TX 76019	817-273-3261	
Steven P. Lalley	Department of Statistics Purdue University Mathematical Sciences Building West Lafayette, IN 47906	317-494-6036	
Professor E. B. Lee	University of Minnesota EE/CS Building, 200 Union Street Minneapolis, MN 55455	612-625-0125	eblee@ee.umn.edu
M. Howard Lee	Department of Physics and Astronomy University of Georgia Athens, GA 30602	404-542-3539	
Franklin T. Luk	School of Electrical Engineering Cornell University Phillips Hall Ithaca, NY 14853	607-255-5075	luk@ee.cornell.edu
Mitchell Luskin	School of Mathematics University of Minnesota 127 Vincent Hall Minneapolis, MN 55455		
Paul Muzio	Support Director Army High Performance Computing Research Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	muzio@ahpcrc.umn.edu
Madhura Nirkhe	University of Maryland Department of Computer Science A.V. Williams Building College Park, MD 20742	301-405-2716	madhura@cs.umd.edu
Jorge Nocedal	Department of Electrical Engineering and Computer Science Northwestern University Evanston, IL 60208	601-634-3987	
Michael J. Nusca	U.S. Army Ballistic Research Laboratory ATTN: SLCBR-LF-A Aberdeen Proving Ground, MD 21005 3909 Halls Ferry Road Vicksburg, MS 39180-6199	301-278-2057	nusca@brl.mil

Name	Address	Phone	Email
Robert E. Olson	Commander U.S. Army Waterways Engineer Waterways Experiment Station ATTN: CEWES-EN (Robert E. Olson) Environmental Laboratory 3909 Halls Ferry Road Vicksburg, MS 39180-6199		
Dave Olwell	Department of Mathematics United Stated Military Academy West Point, NY 10096-1786		
James L. Overholt	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (James L. Overholt) Analytical and Physical Simulation Br. Warren, MI 48397-5000	313-574-8633	
Thomas J. Pence	Department of Metallurgy, Mechanics, and Materials Science Michigan State University East Lansing, MI 48824-1226	517-353-3889	pence@frith.eng.msu.edu
Shietung Peng	Department of Computer Science University of Maryland, Baltimore County Catonsville, MD 21228	301-455-3540	
Linda R. Petzold	Lawrence Livermore National Laboratory L-316 P.O. Box 808 Livermore, CA 94550	415-423-6571	petzold@lll.crg.llnl.gov
Olivier Pironneau	Institut National De Recherche En Informatique Et En Automatique Domaine de Voluceau - Rocquencourt B.P. 105 - 78153 Le Chesnay CEDEX France	011-331-39635483	pironneau@menusin.mralfr
Alyson Reeves	Cornell University Mathematical Sciences Institute Department of Mathematics, White Hall Ithaca, NY 14853	607-564-9041	reeves@mssun7.cornell.msi.edu
J. R. Rice	Department of Computer Science Purdue University Computer Science Building West Lafayette, IN 47907	317-494-6003	rice@cs.purdue.edu
Jubaraj Sahu	Director U.S. Army Ballistic Research Laboratory ATTN: Jubaraj Sahu Launch and Flight Division Aberdeen Proving Ground, MD 21005-5066	301-278-3707	sahu@br-mil
S. Sathananthan	Division of Science and Mathematics Jarvis Christian College Hawkins, TX 75765	903-769-2174 ext. 304	
Robert B. Schnabel	Department of Computer Science University of Colorado at Boulder Boulder, CO 80309	303-492-7554	bobby@cs.colorado.edu
George R. Sell	Army High Performance Computing Research Center University of Minnesota Minnesota Tech Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	sell@ahpcrc.umn.edu
Leang S. Shieh	Cullen College of Engineering Department of Electrical Engineering University of Houston University Park Houston, TX 77204-4793	713-749-4418	

Name	Address	Phone	Email
Royce Soanes	Chief, Benet Laboratories ATTN: Mr. Royce Soanes Watervliet Arsenal Watervliet, NY 12189-4050	518-383-8203	
Janet Spoonamore	Construction Engineering Research Lab. Corps of Engineers Facility Systems Division P.O. Box 4005 Champaign, IL 61824-4005	217-373-7268	slpmma.@csrd@io.edu
Ram P. Srivastav	Department of Applied Mathematics and Statistics State University of New York Stony Brook, NY 11794-3600	516-632-8364	rsrivastav@ccmail.sunysb.edu
J. Michael Steele	Department of Statistics The Wharton School University of Pennsylvania 3010 Steinberg Hall-Dietrich Hall Philadelphia, PA 19104-6302		
Kim Stelson	University of Minnesota Mechanical Engineering 125 Mech. Eng. 111 Church Street SE Minneapolis, MN 55455	612-625-6528	
John C. Strikwerda	Department of Computer Science University of Wisconsin-Madison 1210 West Dayton Street Madison, WI 53706	608-262-0822	strik@cs.wisc.edu
Moss Sweedler	Mathematical Sciences Institute Cornell University White Hall Ithaca, NY 14853	607-255-4373	jcsj@cornellc.cit.cornell.edu
James C. Turner, Jr.	Hampton University Department of Mathematics Hampton, VA 23668	804-727-5352	
Donald Truhlar	University of Minnesota Department of Chemistry 207 Pleasant Street Minneapolis, MN 55455	612-624-7575	mf12101@sc.msc.edu
Gregory Tawa	University of Minnesota Department of Chemistry 139 Smith Hall Minneapolis, MN 55455	612-825-8656	tawa@sc.msc.umn.edu
Thomas Ting	University of Illinois at Chicago CEMM Department (M/C246) Box 4348 Chicago, IL 60680		
Philip Van Valkenberg	MSC 1200 Washington Ave. South Minneapolis, MN 55415		
Mark Vangel	U.S. Army Materials Technology Laboratory ATTN: SLCMT-MRS-MM Arsenal Street Watertown, MA 02172		
John D. Vasilakis	Commander US Army Armament R&D Center ATTN: SMCAR-CCB-RA AMCCOM Benet Weapons Laboratory Watervliet, NY 12189-4050	518-266-5019	vasilaki@pica.army.mil

Name	Address	Phone	Email
John W. Walter, Jr	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. John W. Walter, Jr. Aberdeen Proving Ground, MD 21005-5066	301-278-6051	jwalter@brl.mil
J.R. Walton	Department of Mathematics Texas A&M University College Station, TX 77840		
Yeih J. Wang	Cullen College of Engineering Department of Electrical Engineering University of Houston University Park Houston, TX 77204-4793	713-749-4418	
Roger Wehage	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (Dr. Roger Wehage) Warren, MI 48397-5000	313-574-5378	
Richard Weiss	USACE-WES 3909 Halls Ferry Road Vicksburg, MS 39180	601-634-2194	
J. R. Whiteman	BICOM, Institute of Computational Mathematics Brunel University Uxbridge, Middlesex, UB8 3 PH United Kingdom	44-895-274000	john.whiteman@brunel.ac.uk
Stephen A. Wilkerson	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. Stephen A. Wilkerson Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6131	swilker@ibd.brl.mil
Wendy A. Winner	Director U.S. Army Ballistic Research Laboratory ATTN: Wendy A. Winner Vulnerability/Lethality Division Aberdeen Proving Ground, MD 21005-5066	301-278-6655	wendy@brl.mil
Julian J. Wu	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4332	jjwu@brl. mil
Stephen S. T. Yau	University of Illinois at Chicago Department of Mathematics P.O. Box 4348, M/C 249 Chicago, IL 60680		
Ashraf Zeid	Computer Science Corporation P.O. Box 5156 Warren, MI 48090-5156	313-574-7816	azeid@tacom.emh2.army.mil

<table>
<tr><td colspan="3" align="center">REPORT DOCUMENTATION PAGE</td><td>Form Approved
OMB No. 0704-0188</td></tr>
</table>

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release: distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	

4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARO REPORT 92-2	5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION Army Research Office	6b. OFFICE SYMBOL (If applicable) AMXRO-MA	7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS (City, State, and ZIP Code) P.O. Box 12211 Research Triangle Park, NC 27709	7b. ADDRESS (City, State, and ZIP Code)

8a. NAME OF FUNDING/SPONSORING ORGANIZATION AMSC	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Proceedings of the Thirty-Seventh Conference of the Design of Experiments in Army Research, Development and Testing

12. PERSONAL AUTHOR(S)

13a. TYPE OF REPORT Technical	13b. TIME COVERED FROM Jan 92 TO Feb 93	14. DATE OF REPORT (Year, Month, Day) 1992 August	15. PAGE COUNT 387

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This is a technical report of the Thirty-Seventh Conference on the Design of Experiments in Army Research, Development and Testing. It contains most of the papers presented at this meeting. These articles treat various Army statistical and design problems.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Francis G. Dressel	22b. TELEPHONE (Include Area Code) 919-549-4319	22c. OFFICE SYMBOL AMXRO-MA

DD Form 1473, JUN 86   *Previous editions are obsolete.*

3558-16-4
22-02