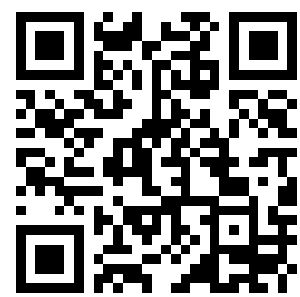


ARO Report 87-2

PROCEEDINGS OF THE THIRTY-SECOND 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. 87-2

June 1987

PROCEEDINGS OF THE THIRTY-SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS

Sponsored by the Army Mathematics Steering Committee

HOST
U. S. Army Combat Developments Experimentation Center
Fort Ord, California

29-31 October 1986

HELD AT
Hilton Inn Resort
Monterey, California

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

FOREWORD

The Army Mathematics Steering Committee (AMSC) sponsors annually the conferences entitled the Design of Experiments in Army Research, Development and Testing. The thirty-second one in this series had as its host the US Army Combat Development Experiment Center (USACDEC) and was held 29-30 October 1986 at the Hilton Inn Resort, Monterey, California. Dr. Marion R. Bryson, Director of USACDEC, served as the local host and conference coordinator not only for this conference but also for the twenty-third and twenty-eighth Design of Experiments meetings. The members of the AMSC appreciates his efforts and the efforts of his staff in coordinating the many details needed to conduct all three of these symposia.

The Special Session this year was entitled "Field Experimentation: The Analysis of Messy Data." There were three invited papers presented. The papers of Professors Dallas E. Johnson and John Tukey discussed the analysis of messy data, while the joint authored paper by Drs. Marion R. Bryson and Carl T. Russell presented some of the problems of scoring casualties in field trials. The titles of the technical and clinical sessions give some idea of the many statistical areas treated in the contributed papers: (1) Parametric Statistics,(2) Statistical Theory, (3) Design of Experiments, (4) Data Analysis and Modeling, (5) Theory and Probablistic Inference, (6) Fuzzy Statistics, (7) Forecasting and Prediction, (8) Small Sample Analysis, and (9) Regression and Smoothing. The program Committee, for the invited speaker phase of the conference, obtained the following nationally known scientists to talk on topics of current interest to Army personnel as well as other attendees.

Speaker and Affiliation	Titles of Address
Professor George E.P. Box University of Wisconsin	Statistical Design, Analysis for Quality Improvement
Professor Walter T. Federer Cornell University	Statistical Analysis for Intercropping Experiments
Professor Persi Diaconis Stanford University	The Search for Randomness
Professor Emanuel Parzen Texas A&M University	Quantile Statistical Data Analysis
Professor Stuart Geman Brown University	Some Applications of Bayesian Image Analysis

The conference was preceded by a two-day tutorial on "Density Estimation, Modeling and Simulation" by Professor James Thompson of Rice University. The dates for the tutorial were Monday and Tuesday 27-28 October 1986. Professor Thompson has conducted extensive research in the areas covered in his lectures. His approach to the presented material was excellent and he generated many interesting discussions.

Dr. Francis G. Dressel was the recipient of the Sixth Wilks Award for contributions to Statistical Methodologies in Army Research, Development and Testing. Dr. Dressel was uniquely qualified by virtue of his service in the Mathematical Sciences Division of the U.S. Army Research Office over three decades. He was one of the principals at the inception of the Army Design of Experiments Conference and along with Sam Wilks, planned and implemented the then-fledgling conference. Dr. Dressel currently serves as editor of the Conference Proceedings and continues to contribute to the advancement of statistics in the U.S. Army.

The AMSC would like to thank the members of the conference committee for guiding this excellent scientific conference, and to also thank the Mathematical Sciences Division of the Army Research Office, for preparing the proceedings of these meetings.

CONFERENCE COMMITTEE

Carl Bates
Robert Burge
David Cruess
Bernard Harris
Robert Launer
J. Richard Moore
Carl Russell
Douglas Tang
Malcolm Taylor
Jerry Thomas

TABLE OF CONTENTS*

*This Table of Contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Thirty-Second Conference on the Design of Experiments, see the Program of this meeting.

AGENDA

THIRTY-SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS

IN ARMY RESEARCH, DEVELOPMENT AND TESTING

29-31 October 1986

Host: US Army Combat Developments Experimentation Center
Fort Ord, California 93941
Dr. Marion R. Bryson, Director

Location: Hilton Inn Resort
1000 Aguajito Road
Monterey, California 93940

*** * * * * Wednesday, 29 October * * * * ***

0815-0915	REGISTRATION	**Vista Del Mar Room**
0915-0930	**CALLING OF THE CONFERENCE TO ORDER**	**Presidio Room**
	Dr. Marion R. Bryson, Director US Army Combat Developments Experimentation Center	
	WELCOMING REMARKS	
0930-1200	**GENERAL SESSION I**	**Presidio Room**
	Chairman: Dr. Marion R. Bryson	
0930-1030	**KEYNOTE ADDRESS:** STATISTICAL DESIGN, ANALYSIS FOR QUALITY IMPROVEMENT	**Presidio Room**
	Professor George E. P. Box, University of Wisconsin	
1030-1100	BREAK	**Vista Del Mar Room**
1100-1200	STATISTICAL ANALYSIS FOR INTERCROPPING EXPERIMENTS	**Presidio Room**
	Professor Walter T. Federer, Cornell University	
1200-1330	LUNCH	
1330-1530	**SPECIAL SESSION** - FIELD EXPERIMENTATION: THE ANALYSIS OF MESSY DATA	**Presidio Room**
	Chairman: Mr. William D. West, Director, Science and Technology, US Army Combat Developments Experimentation Center	

SPECIAL SESSION (cont'd)

TITLE: SCORING CASAULTIES IN FIELD TRIALS
Dr. Marion R. Bryson, Director, USACDEC and
Dr. Carl T. Russell, US Army Operational Test and Evaluation Agency

SOME TOPICS IN MESSY DATA ANALYSIS
Professor Dallas E. Johnson, Kansas State University

TITLE: To be announced
Professor John Tukey, Princeton University

1530-1545 BREAK — **Vista Del Mar Room**

1545-1705 **TECHNICAL SESSION ON PARAMETRIC STATISTICS** — **Presidio Room**

Chairman: Dr. Oskar Essenwanger, US Army Missile Command

MAXIMUM LIKELIHOOD ESTIMATION OF THE PARAMETERS OF A FOUR-PARAMETER CLASS OF PROBABILITY DISTRIBUTIONS
Dr. S. H. Lehnigk, US Army Missile Command

ON THE FITTING OF CLIMATOLOGICAL DATA SAMPLES BY A THREE PARAMETER DISTRIBUTION FUNCTION
Mr. Helmet P. Dudel, US Army Missile Command

1830-1930 CASH BAR — **Big Sur Room**

1930-2130 **BANQUET AND PRESENTATION OF WILKS AWARD** — **Big Sur Room**

*** * * * * Thursday, 30 October * * * * ***

0815-1000 **TECHNICAL SESSION 1 - STATISTICAL THEORY** — **Presidio Room**

Chairman: Dr. Francis Dressel, US Army Research Office

QUICK APPROXIMATIONS TO SYNTHESIS IN PATTERN THEORY
Professor Jayaram Sethuraman, Florida State University

ON ROTATION IN FACTOR ANALYSIS OF ATMOSPHERIC PARAMETERS
Dr. Oskar Essenwanger, Redstone Arsenal

AN EXACT METHOD FOR ONE-SIDED TOLERANCE LIMITS BASED ON A BALANCED ONE-WAY ANOVA RANDOM EFFECTS MODEL
Mr. Mark Vangel, US Army Materials Technology Laboratory

LIMIT THEOREMS FOR GENERALIZED RANDOM ALLOCATION PROBLEMS
Dr. Bernard Harris, MRC, University of Wisconsin

0815-1000 **CLINICAL SESSION I** - DESIGN OF EXPERIMENTS **Vista Del Mar Room**

Chairman: Dr. Malcolm Taylor, US Army Ballistic Research Laboratory

Panelists: Dr. Kaye Basford, Mathematical Sciences Institute Cornell University
Professor George E.P. Box, University of Wisconsin
Professor Walter Federer, Cornell University

APPLICATION OF EXPERIMENTAL DESIGN TO THE EVALUATION OF EXPERT OPINION
Mr. Franklin E. Womack and Mr. Carl B. Bates
US Army Concepts Analysis Agency

ANALYSIS OF AN INCOMPLETE BLOCK DESIGN OF EXPERIMENTS
Ms. Wendy A. Winner and Ms. Jill H. Smith, US Army Ballistic Research Laboratory

1000-1030 BREAK

1030-1200 **TECHNICAL SESSION 2** - DATA ANALYSIS AND MODELING **Presidio Room**

Chairman: Dr. William Baker, US Army Ballistic Research Laboratory

A HEURISTIC APPROACH TO POST-HOC COMPARISONS FOR SIGNIFICANT INTERACTIONS - A SIMPLIFIED NOTATION
Dr. Eugene Dutoit, US Army Infantry School

THE DESIGN OF EXPERIMENTS TO DETERMINE THE INCIDENCE OF SKIN BURNS UNDER CONTEMPORARY ARMY UNIFORMS EXPOSED TO THERMAL RADIATION FROM SIMULATED NUCLEAR FIREBALLS
Mr. Brian R. Shallhorn, Mr. Anthony J. Baba and
Mr. Stewart Share, Harry Diamond Laboratories

STATISTICAL EVALUATION OF DESERT INDIVIDUAL CAMOUFLAGE COVERS (ICC) BY GROUND OBSERVERS
Mr. George Anitole and Mr. Ronald L. Johnson, US Army Belvoir Research Development and Engineering Center
Mr. Christofer J. Neubert, US Army Engineer School

1030-1200 **CLINICAL SESSION II - THEORY AND PROBABILISTIC INFERENCE** **Vista Del Mar Roo**

Chairman: Mr. Carl Bates, US Army Concepts Analysis Agency

Panelists: Dr. Kaye Basford, Mathematical Sciences Institute
Cornell University
Professor Jayaram Sethuraman, Florida State University

COMBINATORIAL ISSUES ASSOCIATED WITH MACHINE FILTERING OF TACTICAL MESSAGES
Mr. Terry Cronin, US Army Signal Warfare Center

NONPARAMETRIC SMALL SAMPLE TOLERANCE LIMITS
Mr. Donald Neal and Mr. John Reardon, US Army Materials Technology Laboratory (Presented by Dr. Bernard Harris, MRC, University of Wisconsin)

1200-1330 LUNCH

1330-1515 **TECHNICAL SESSION 3 - FUZZY STATISTICS** **Presidio Room**

Chairman: Dr. Carl Russell, US Army Operational Test and Evaluation Agency

INCORPORATING FUZZY SET THEORY INTO STATISTICAL HYPOTHESIS TESTING
Mr. William E. Baker, US Army Ballistic Research Laboratory

PRACTICAL MODELING WITH FUZZY FUNCTIONS AND FUZZY DATA
Dr. Aivars Celmins, US Army Ballistic Research Laboratory

A CENTRAL LIMIT THEOREM FOR FUZZY RANDOM VARIABLES
Dr. Steven B. Boswell, Harvard University School of Public Health
Dr. Malcolm Taylor, US Army Ballistic Research Laboratory

AN APPLICATION OF A FUZZY RANDOM VARIABLE TO VULNERABILITY MODELING
Dr. Steven B. Boswell, Harvard University School of Public Health
Dr. Malcolm Taylor, US Army Ballistic Research Laboratory

1330-1515 **CLINICAL SESSION III - FORECASTING AND PREDICTION** **Vista Del Mar Room**

Chairman: Dr. Charles A. Correia, US Army Logistics Center

Panelists: Professor Persi Diaconis, Stanford University
Professor Emanuel Parzen, Texas A&M University

CLINICAL SESSION III (cont'd)

AN EVALUATION OF AUTOMATIC FORECASTING TECHNIQUES APPLIED TO ENLISTED PERSONNEL SEPARATIONS
Dr. Betsy Abbe and Mr. Frank Womack, US Army Concepts Analysis Agency

STUDY ON THE FEASIBILITY OF GENERATING "PREDICTIVE ANALYSIS FLAGGING SYSTEM" (PAFS) BY UTILIZING THE ARMY'S EXISTING DATA SOURCE
Ms. Li Pi Su, US Army Materiel Readiness Support Activity

1515-1530 BREAK

1530-1730 **GENERAL SESSION II** **Presidio Room**

Chairman: Professor Henry B. Tingey, University of Delaware

THE SEARCH FOR RANDOMNESS
Professor Persi Diaconis, Stanford University

QUANTILE STATISTICAL DATA ANALYSIS
Dr. Emanuel Parzen, Texas A&M Univeristy

*** * * * * Friday, 31 October * * * * ***

0830-1015 **TECHNICAL SESSION 4 - SMALL SAMPLE SURVIVAL ANALYSIS** **Presidio Room**

Chairman: Mr. John Robert Burge, Walter Reed Army Institute of Research

SAMPLE SIZE REQUIREMENTS IN QUANTAL RESPONSE TESTING
Mr. Berry A. Bodt, US Army Ballistic Research Laboratory
Professor Henry B. Tingey, Univeristy of Delaware

A TEST FOR CONSISTENCY OF A CLASS OF VULNERABILITY MODELS
Mr. David W. Webb and Mr. J. Richard Moore, US Army Ballistics Research Laboratory

MORE ON THE PERVERSITY OF MISSING POINTS IN 16-POINT DESIGNS
Dr. Carl T. Russell, US Army Operational Test and Evaluation Agency

0830-1015 **CLINICAL SESSION IV - REGRESSION AND SMOOTHING** **Vista Del Mar Room**

Chairman: Mr. Franklin E. Womack, US Army Concepts Analysis Agency

Panelists: Professor Stuart Geman, Brown University
Professor Walter Federer, Cornell University

A METHOD FOR THE STATISTICAL ANALYSIS OF THE STRESS-STRAIN PROPERTIES OF EARTH MATERIALS
Mr. G. Y. Baladi and Mr. Behzad Rohani, US Army Engineer Waterways Experiment Station

1015-1045 BREAK

1045-1200 **GENERAL SESSION III** **Presidio Room**

Chairman: Dr. Douglas B. Tang, Walter Reed Army Institute of Research, Chairman of the AMSC Subcommittee on Probability and Statistics

1045-1100 OPEN MEETING OF THE STATISTICS AND PROBABILITY SUBCOMMITTEE OF THE ARMY MATHEMATICS STEERING COMMITTEE

1100-1200 SOME APPLICATIONS OF BAYESIAN IMAGE ANALYSIS
Professor Stuart Geman, Brown University

A D J O U R N

Statistical Analyses for Intercropping Experiments

Walter T. Federer
Cornell University
Ithaca, N.Y. 14853

Abstract

Statistical methodology for analyzing intercropping experiments was developed over the last 20 years and is being developed at present. Considerably more research is required for the many and diverse types of experiments involving sole crops (crops grown alone) and mixtures of crops (intercrops) grown together or in sequence. The growing of two or more crops together or in sequence is known as intercropping. An outline of twenty chapters of a book on the statistical design and analysis of intercropping experiments is presented. A number of the statistical analyses in the book are briefly described. Sections 2 to 8 relate to analyses for two crops in a mixture along with sole crops. Sections 9 to 15 discuss analyses for three or more crops in a mixture in addition to sole crops and mixtures of two crops. It is stressed that it is dangerous to extrapolate from sole crop responses to mixtures of two crops and from mixtures of k crops to mixtures of k + 1 crops. Many of the data sets examined produced unexpected and sometimes surprising results. The last section discusses other areas of application, e.g., survey sampling, nutrition, education, medicine, and recreation, where these results can be utilized.

Statistical Analyses for Intercropping Experiments

Walter T. Federer
Cornell University
Ithaca, N.Y. 14853

BU-880-M*

1. Introduction

Intercropping investigations involves the growing of two or more crops on the same area of land either simultaneously, partially at the same time, or sequentially. It is a centuries old practice in tropical agriculture, and to some extent in temperate zone agriculture. Agricultural, biological, and statistical investigations have tended to ignore the problems of research in this area. Statistical analysis; of intercropping investigations is considered to be the most important unsolved statistical question related to research in tropical agriculture. It is an area neglected by all except a handful of statisticians. A computer search of statistical literature resulted in the single paper citation for Mead and Riley (1981). This is an excellent paper, though limited in outlook for the broad range of statistical analyses useful in intercropping research.

---

* In the Technical Report Series of the Biometrics Unit.

To acquaint the statistical profession with relevant procedures and to fill a need by intercropping researchers, a book is being published by this author on the topic. The table of contents is:

Part I – Two Crops

Chapter 1. Introduction

Chapter 2. One main crop grown with a supplementary crop

Chapter 3. Both crops main crops – density constant – analyses for each crop separately

Chapter 4. Both crops main crops – density constant – combined crop responses

Chapter 5. Both crops of major interest with varying densities

Chapter 6. Monocultures and their pairwise combinations when responses are available for each member of the combination

Chapter 7. Monocultures and their pairwise combinations when separate crop responses are not available

Chapter 8. Spatial and density arrangements

Chapter 9. Some variations for intercropping

Part II – Three or More Crops

Chapter 10. Introduction

Chapter 11. One main crop with more than one supplementary crop

Chapter 12. Three or more main crops – density constant

Chapter 13. Three or more main crops – density variable

Chapter 14. Monocultures and their combinations when responses are available for each crop

Chapter 15. Monocultures and their combinations when separate crop responses are not available

Chapter 16 Spatial and density arrangements for three or more crops

Chapter 17 Variations for intercropping of three or more crops

Part III — Additional Topics

Chapter 18 Experiment design for intercropping experiments

Chapter 19 Other areas of application

Chapter 20 Bibliography on intercropping investigations

It is necessary to fully comprehend the nature of two crop mixtures before proceeding to anything more difficult. The interpretational difficulty increases by an order in magnitude when going from sole crop (crops grown alone) experiments to experiments with sole crops and biblends (mixture of two crops.) It goes up another order in magnitude in going from intercropping experiments with two crops to experiments involving mixtures of three or more crops. In addition to the interpretational difficulty, it is dangerous to extrapolate from sole crops to biblends and from biblends to mixtures involving three or more crops. It is dangerous to extrapolate from lower densities to higher ones. Many, if not most, experiments contain an unexpected result.

A number of statistical analyses found useful for intercropping investigations are discussed below. The topics follow the table of contents of a forthcoming book that is outlined above.

2. One Main Crop Plus one Supplementary Crop

The experiment designs found useful for sole crops will be the same ones found useful for one main crop grown with a supplementary crop. The treatment design consists of the varieties of a main crop grown as sole crops and in combination with varieties of the supplementary crop. To

illustrate, suppose that five = c_m varieties of maize are to be grown alone and in combination with six = c_b varieties of beans. A single density for maize and for beans is selected, i.e. plant population per hectare is not a variable. The treatment design would be:

	Cropping System						
Maize Variety	Sole	Bean Variety 1	2	3	4	5	6 = c_b
1							
2							
3							
4							
5 = c_m							

There would be $v = c_m + c_b c_m = 36$ treatments composed of five sole crops and 30 biblends. Experiment designs appropriate for 36 treatments would be used (see e.g., Federer and Kirton, 1984.)

Statistical analyses for experiments in a given experiment design and for the above treatment design would involve the same types of statistical analyses as used for sole crop experiments (see e.g., Snedecor and Cochran, 1967.) Some common statistical procedures used would be

(i) single (or subsets of) degree(s) of freedom contrasts,

(ii) multiple comparisons procedures,

(iii) subset selection procedures,

(iv) covariance analyses, and

(v) multivariate analyses.

Some additional statistical analyses found useful for yields are:

(vi) Tukey's one-degree-of-freedom analysis for the crop one by crop two interaction,

(vii) Finlay-Wilkinson (1963) analysis for mixtures,

(viii) tests for interaction given that one or more of the c_m maize varieties are standards for comparison, and

(ix) yields of main crop are not to be reduced by more than a fixed percentage.

3. Two Main Crops – Density Constant

Experiment design considerations for biblends when both crops are main crops, are the same as discussed in Section 2. The treatment design would have sole crops of both crops included; otherwise, it is the same as discussed in Section 2. Statistical analyses on the yields of each crop separately would follow that outlined in the previous section.

In order to evaluate cropping systems and to compare biblend production with sole crop production, it is necessary to combine the yields of both crops in some meaningful manner. An economic point of view would place a value, v_i, on the produce from crop i, say Y_i and use $V = v_1Y_1 + v_2 Y_2$. If v_i are prices, it might be more realistic to use ratios of prices, which are more stable, and use relative values $V^* = Y_1 + Y_2(v_2/v_1)$. For sole crops, V(or V^* could be obtained by putting $Y_2 = 0$ for crop one and $Y_1 = 0$ for crop two. A nutritional point of view would convert the yield to calories and/or protein and use a measure of the form: $C = c_1Y_1 + c_2Y_2$, where c_i is a calorie (or protein) conversion factor. An agronomic or land use point of view would consider a linear combination of yields of the form:

$$L = \frac{Y_{b1}}{Y_{s1}} + \frac{Y_{b2}}{Y_{s2}} ,$$

where Y_{bi} is the yield of crop i in a biblend mixture and Y_{si} is the yield of crop i grown as a sole crop. There are many forms of L_i, which is called relative yield or land equivalent ratio. The component yields of the mixture are put into proportions of yields obtained from sole crop yields. Since yields may vary considerably, a ratio of sole crop yields might be more stable. In this case a "relative land equivalent" ratio would be computed as

$$L^* = Y_{b1} + Y_{b2}\left(Y_{s1}/Y_{s2}\right) \quad .$$

A statistical point of view would use a discriminant function analysis and construct a canonical variable of the form:

$$D = Y_{b1} + RY_{b2} \ ,$$

where R is chosen to maximize the ratio, treatment sum of squares divided treatment plus error sums of squares.

The first three linear combinations given above, i.e., V, C, and L are readily interpretable quantities by a researcher or a farmer. The last one D is not and sole crop yields cannot be compared with D, but can be with V, C, and L. Although a statistician's first thoughts in combining yields most likely would be to use multivariate analyses, this would not be the correct thing to do as comparisons of sole crop yields and farming system yields cannot be made and the canonical variable has no practical meaning in the sense that C, V, and L do. Some aspects of multivariate analyses have been found useful by Pearce and Gilliver (1978, 1979) in studying the nature of response from mixtures.

Statistical analyses for linear combinations C, V, and L, are straightforward. Those outlined in the previous section may be utilized. These created functions of yield may be used in the same manner as canonical variables from a discriminant function analysis, i.e., univariate analyses are performed on the canonical variables. It is possible to combine value and land use by taking the ratio $Y_{s1}v_1/Y_{s2}v_2 = R$ and using the created function of yields $Y_1 + RY_2$. It does not appear realistic to combine variables other than yield variables as described above.

4. Two Main Crops – Density Variable

Plant populations per hectare in sole crops and in biblends need to be considered seriously in conducting intercropping investigations. Crop densities maximizing yields Y_i, or linear combinations of yield V, C, and D, are desired. Using univariate analyses, a multiple comparisons or subset selection procedure may be used to pick the "optimal" densities for the crops. A useful procedure would be to model yield as a function of plant density. Within narrow ranges of densities, a linear approximation of the form has been found to be useful:

$$Y_{ii\ell k} = \beta_{0i} + \rho_k + \beta_{1i} d_{i\ell_i} + \varepsilon_{ii\ell k} \quad ,$$

where $Y_{ii\ell k}$ is the yield of the ith crop as a sole crop, β_{0i} is an intercept, β_{1i} is a linear regression coefficient, $d_{i\ell_i}$ is the density ℓ_i for crop i, ρ_k is the effect of block k, and $\varepsilon_{ii\ell k}$ is a random error term with mean zero and variance σ^2_ε. Note that a variety of other functional relations could be used to model yield as a function of density. Using the above form, the yields of crop i in the mixture ij of two crops may be

expressed as

$$Y_{i(j)\ell_1\ell_2 k} = \beta_{0i} + \rho_k + \beta_{1i} d_{i\ell_i} + \gamma_{i(j)}\left(d_{i\ell_i}, d_{j\ell_j}\right) + \varepsilon_{i(j)\ell_1\ell_2 k}$$

where $\gamma_{i(j)}(d_{i\ell_i}, d_{j\ell_j})$ is an additive effect on the yield of crop i due to its being intercropped with crop j at the corresponding densities $d_{i\ell_i}$ and $d_{j\ell_j}$. A large positive value of $\gamma_{i(j)}(d_{i\ell_i}, d_{j\ell_j})$ is desired. When there are many lines of a cultivar in an investigation, the above analysis may be conducted for each line. Then, analyses over all lines can be obtained.

5. Modeling Responses for Sole Crops and Biblends – Two Responses

In many situations, responses for both components of a mixture are available. The crops may be intermingled but distinct in type so that responses for each crop are obtained, or the crops may be spatially separated and again responses for each crop are available. For treatment designs containing all sole crops and all possible combinations of lines of crops in mixtures of two, response model equations can be constructed which have measures of a general mixing ability (gma) effect and of a specific mixing ability (sma) effect of a line or crop. To illustrate, suppose that it was desired to compare yields of v = five bean cultivars as sole crops and in mixtures of two. The v(v + 1)/2 = 15 combinations would be:

Cultivar	1	2	3	4	5
1	S	B	B	B	B
2		S	B	B	B
3			S	B	B
4				S	B
5					S

where S stands for sole crop and B denotes a biblend. With such a treatment design in a randomized complete block design, one possible linear model is:

Sole crop i:

$$Y_{hiis} = \mu + \rho_h + \tau_i + \varepsilon_{hiis},$$

where $\mu + \rho_k$ is a block mean effect, τ_i is cultivar effect, and ε_{hiis} have zero mean and common variance σ^2_ε.

Biblend ij:

$$Y_{hi(j)b} = \frac{1}{2}(\mu + \rho_h + \tau_i + \delta_i) + \gamma_{i(j)} + \varepsilon_{hi(j)b},$$

$$Y_{h(i)jb} = \frac{1}{2}(\mu + \rho_h + \tau_j + \delta_j) + \gamma_{(i)j} + \varepsilon_{h(i)jb}$$

where $Y_{hi(j)b}$ is the yield of cultivar i from the mixture ij, μ, ρ_h, and τ_i are as defined for sole crop, δ_i is a general combining ability effect for cultivar i when grown in biblends, γ_{ij} is an interaction effect for crop i in the presence of crop j, and the $\varepsilon_{hi(j)b}$ are error components for cultivar i responses which have zero mean and common variance $\sigma^2_\varepsilon/2$. The coefficient 1/2 is included in order to have the μ, ρ_h, τ_i, and δ_i from the biblends on the same basis as the corresponding parameters for sole crops. With two cultivars on the same area of land as the sole crops, each crop response can only contribute 1/2 to μ, ρ_h, and τ_i. Response model equations can easily be constructed for the case where one crop occupies a proportion p of the area and the second crop occupies 1 - p of the area. In this case, care must be taken in defining an interaction effect. An interaction is defined to relate to two items in equal proportions. To interact, both must be present. When $p < 1/2$, only 2p of the total material in an experimental unit is available to interact on a 1:1 basis;

1 - 2p of the material is not available. If some such definition as the above is taken, interaction effects will be invariant with respect to changing proportions p.

Note that when other treatment designs are used, other models can be constructed. For example, suppose that only a subset of the v(v - 1)/2 biblends were included in a experiment along with sole crops. The parameters μ, ρ_h, τ_i, and $\delta_i/2 + \gamma_{i(j)} = \gamma^*_{i(j)}$ can be estimated. It is not possible to obtain solutions for $\delta_i/2$ and $\gamma_{i(j)}$ but only their sum. If the experimenter were willing to assume that the $\gamma_{i(j)}$ not present were all zero, then solutions are possible. This is considered to be an unrealistic assumption.

6. Modeling Responses for Sole Crops and Biblends — One Response

For certain types of mixtures, such as, e.g., a diallel crossing experiment, it is impossible or difficult to obtain responses for both components of a biblend. Experiments involving sole crops and mixtures of two lines of a cultivar where the lines are not phenotypically distinct or are not spatially separated would be found for wheat, beans, and many other crops. In mixtures of grasses and legumes in hay it is difficult to obtain the separate responses for each member in the mixture. Several response models are available. For a randomized complete block design and the treatment design involving sole crops and all possible biblends, the following pair of equations for sole crop and biblend yields has been proposed (Federer *et al.*, 1982):

$$Y_{hiis} = \mu + \rho_h + \tau_i + \varepsilon_{hiis}$$

$$Y_{hijb} = \mu + \rho_h + (\tau_i + \delta_i + \tau_j + \delta_j)/2 + \gamma_{ij} + \varepsilon_{hijb}$$

where the effects are as defined in the previous section except for γ_{ij} which is an interaction component for specific mixing ability. Note that γ_{ij} is equal to the sum $\gamma_{i(j)} + \gamma_{(i)j}$. These last two components cannot be estimated unless individual responses are available whereas γ_{ij} can be estimated when only the combined response is available.

Another treatment design would be sole crops, all combinations, and all reciprocals. To illustrate, suppose that v = 5 wheat varieties are available, and the experimenter wishes to have all varieties bordered by every other variety and itself. Responses from border rows are not obtained. The $v^2 = 25$ treatments would be:

	Wheat Variety				
Border	1	2	3	4	5
1	S	B	B	B	B
2	B	S	B	B	B
3	B	B	S	B	B
4	B	B	B	S	B
5	B	B	B	B	S

where S denotes sole crop and B denotes the mixture. Note that variety 1 bordered by variety 2 is <u>not</u> the same as variety 2 bordered by variety 1. One set of response models for sole crop and biblends respectively is:

$$Y_{hii} = \mu + \rho_h + \tau_i + \epsilon_{hiis}$$

and

$$Y_{hijb} = \mu + \rho_h + \tau_i + \delta_i + \gamma_{ij} + \epsilon_{hijb, i \neq j}$$

where μ, ρ_h, τ_i, δ_i, ϵ_{hiis}, and ϵ_{hijb} are as defined as above and γ_{ij} is a within variety interaction term with $\gamma_{ii} = 0$; γ_{ij} is an interaction term for crop i when bordered by crop j.

A second response model equation for the above treatment design would be the one for a two-factor (crops and borders) factorial:

$$Y_{hij} = \mu + \rho_h + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{hij} \quad ,$$

where α_i is the effect of crop i, β_j is the effect of border j, and $\alpha\beta_{ij}$ is an interaction term. Such a model would not be too realistic in a variety of situations since sole crop responses may be quite different from biblend responses.

A third model is adapted from Martin (1980) and is the previous model with the following change:

$$\alpha\beta_{ij} = \eta_{ij} + \omega_{ij} + \kappa_{ij} \quad ,$$

where $\eta_{ij} = \eta$ for $i = j$ and $\eta_{ij} = -\eta/(v-1)$ for $i \neq j$, $\omega_{ij} = (\alpha\beta_{ij} + \alpha\beta_{ji})/2 + \eta/(v-1)$ for $i \neq j$, and $\kappa_{ij} = (\alpha\beta_{ij} - \alpha\beta_{ji})/2$.

A fourth model is a mixture of the previous ones and is

$$Y_{hijb} = \mu + \rho_h + \tau_i + \delta_i + \beta'_j + \omega'_{ij} + \kappa_{ij} + \varepsilon_{hijb} \quad ,$$

where β'_j and ω'_{ij} are similar to the above β_j and ω_{ij} but are conditional on the fact that $\alpha\beta_{ii} = 0$; the remaining parameters are as defined above.

Other situations will lead to the construction of other response model equations. Appropriate models will need to be constructed for the particular conditions encountered in an investigation.

7. Spatial and Density Arrangements

Spatial arrangements and density levels are very important items to consider in intercropping investigations. By spatial arrangement, we mean the pattern used for plants in a given area of land. The plants could be in rows, in hills, or drilled. The number of plants per hectare could be varied over a wide range. The following five items need to be studied for any intercropping investigation:

(i) spatial arrangement of crop one,

(ii) spatial arrangement of crop two,

(iii) density of crop one,

(iv) density of crop two, and

(v) intimacy of the two crops.

By intimacy we mean the closeness of plants of the two crops. If plants of the two crops are randomly mingled in the same row, we say that they are 100% intimate. Plants of the two crops in separate rows would be less intimate. If the two crops were isolated far enough to eliminate any interaction, they have zero intimacy. To illustrate, suppose that density is not a variable but intimacy and spatial arrangement are. One plan could be to have two crops, say maize and beans, in the same row with rows one meter apart. A second plan could be to double the density within rows and double the distance between rows. The density per hectare and intimacy would be the same but spatial arrangement would be different. A third plan would be to alternate rows of the two crops. The intimacy would be less than in the first two plans. Another plan commonly used for maize and

beans in Brazil is one row of maize and two rows of beans alternating as below (M = maize and B = beans):

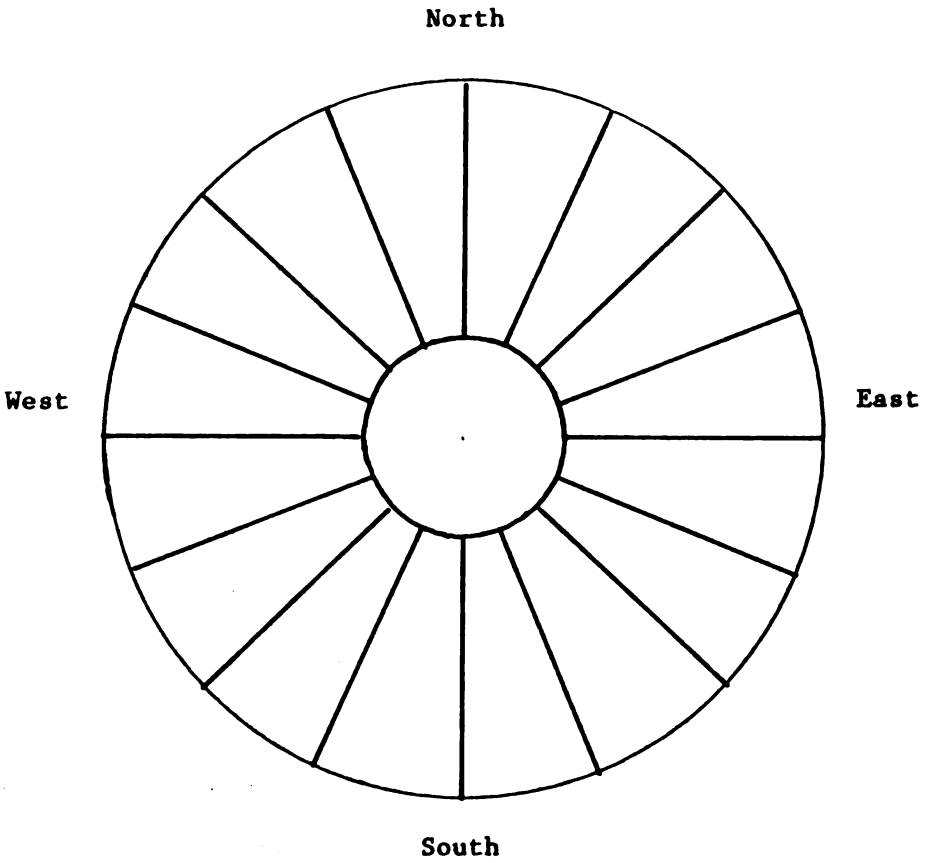
MBBMBBMBBMBBM··· .

The maize rows are one meter apart and the bean rows are one-half meter apart. A fifth plan would be:

MMBBBBMMBBBBMMBBBBMM··· .

The pairs of maize rows are 1.75 meters apart and the rows of a pair are 0.25 meter apart. The bean rows are one-half meter apart. The last plan could be the best as more light would be available for maize and for bean plants than in the previous plans. The rows should be oriented in a north-south direction in order to benefit from the additional light.

Several plans are available to study wide variations in density with a relatively small amount of material. They should be used to obtain information on ranges of density for future study. The best known of these is the fan design of Nelder (1962). There are several versions of this design. Another useful design has been suggested by B. N. Okigbo (1978). The design is a circle with orientation noted (see below). A small circle in the center is not used as some space is needed to start the rows. The row spacing becomes increasingly distant as one moves away from the center of the circle. The density within a row could be kept constant or the density per hectare could be kept constant by increasing the density within a row as one moves away from the center of a circle of the following nature:



The lines above could indicate the rows of plants. The above design could be for a single crop or for mixtures of two crops using the previously described plans for spatial arrangements and intimacy. A Nelder fan design would be one-quarter of the above and would be used if directional orientation were unimportant. Both the Okigbo-circle and the Nelder-fan designs are very parsimonious of space. One statistical analysis would be to divide the circle into concentric circles of equal areas. Yields would then be obtained for the areas of individual rows. The results could be plotted graphically to determine optimal yields or some regression function could be fitted to the yields. Optimal row distances and optimal densities for yield could then be obtained. These circles or fans could be constructed for various cropping systems and replicated over a range of

conditions to be encountered in practice. It may be possible to determine optimal density, spatial arrangement, and intimacy well enough so that future experimentation is not necessary. However, it is likely that future experimentation will be needed to more precisely determine optimal values.

8. Variations and Additional Analyses

Many and diverse situations exist in intercropping research. One such area is to study the effect of replacing one crop in a mixture with a second crop with proportions ranging from zero to one. Given that p_a is the proportion for crop a and $1 - p_a = p_b$ is the proportion of crop b in the mixture and Y_{si} = yield of sole crop i, the computed value for a strictly replacement series would be $p_a Y_{sa} + p_b Y_{sb}$. If the yield of the mixture at proportion (p_a, p_b) was greater than this value, this would be termed cooperation. If less, then denote this as inhibition. If one crop is inhibited and the other exhibits cooperation, this would be denoted as compensation since the yield of one crop is increased and the other is decreased. For intercropping, proportions and crops showing a large amount of cooperation are desired.

Several other statistics have been developed for competition studies. A number of them are related to a land equivalent ratio.

Let $Y_{bi}/Y_{si} = L_i$ which is the proportional yield of the crop in a mixture relative to the crop grown alone. A land equivalent ratio is $L = L_1 + L_2$. A statistic was developed to compute *total effective area* for the case where A_i = area devoted to sole crop i and A_m = area devoted to the mixture of the two crops. Then, total effective area is computed as $A_1 + A_2 + LA_m$. A *relative crowding coefficient* is computed as $L_1 L_2/(1 - L_1)(1 - L_2)$. A *coefficient of aggressivity* to measure the

dominance of one crop over another is computed as $L_1 - L_2$. A *competitive ratio index* is given by L_1/L_2. Each of these can be adjusted to the relative proportions $p_a:p_b$ of crop a and crop b in the mixture. Other coefficients have been suggested. A number relate to crop stability (an ill-defined term) and to "risk to farmers". Survival farming must take some form of these measures into account as a farmer needs to produce food every year in order to survive.

Another type of analysis suggested by B. R. Trenbath in his discussion of the Mead and Riley (1981) paper is linear programming. Here yields of the crops as sole crops and in mixtures is required. Then for a goal, say S units of starch and P units of protein, an optimal allocation of area to sole crops and to mixtures can be computed. A farmer can minimize land area needed to reach his primary goal (food production) and can use the remaining area of his farm for crops to achieve a secondary goal (say produce for sale). Economic studies make use of linear programming for some of their investigations.

9. One Main Crop with Two or More Supplementary Crops

Consideration of mixtures for more than two crops in the mixture would at first sight appear to be a straightforward extension of the procedures for two crops. This is not the case. To illustrate this for one main crop with supplementary crops, it would appear that one could simply follow the procedures described in Section 2, but consider the following treatment design and example. Barley was the main crop and only one barley variety was included in the experimental units along with barley in combinations of one cultivar plus barley, all possible combinations of three of the six cultivars with barley, and one combination of all six cultivars with

barley. Plant numbers per experimental unit were kept constant and the same number of barley plants were harvested in every experimental unit. Barley as a sole crop was one of the treatments. In all there were 1 + 6 + 20 + 1 = 28 treatments. For a randomized complete block design and barley yields for variety g (one variety), one set of response equations is:

Sole crop - variety g

$$Y_{gh0} = \mu + \tau_g + \rho_h + \varepsilon_{gh} .$$

Variety g plus one crop i

$$Y_{ghi1} = \mu + \tau_g + \rho_h + \delta_i + \varepsilon_{ghi} .$$

Variety g plus two crops i and j

$$Y_{ghij2} = \mu + \tau_g + \rho_h + (\delta_i + \delta_j)/2 + \gamma_{ij} + \varepsilon_{ghij} .$$

Variety g plus three crops i, j, and k

$$Y_{ghijk3} = \mu + \tau_g + \rho_h + (\delta_i + \delta_j + \delta_k)/3 + 2(\gamma_{ij} + \gamma_{ik} + \gamma_{jk})/3 + \lambda_{ijk} + \varepsilon_{ghijk} .$$

$$\vdots$$

Variety g plus all cultivars

$$Y_{ghij\cdots v} = \mu + \tau_g + \rho_h + \delta_. + \gamma_. + \lambda_. + \cdots + \pi_{12\cdots v} + \varepsilon_{ghij\cdots} .$$

For the above example, mixtures of barley with two other crops were not included in the experiment. $\mu + \tau_g$ is the mean for barley variety g grown as a sole crop, ρ_h is the h'th block effect, δ_i is a general mixing effect of crop i on barley yields Y_{ghi1}, γ_{ij} is a bi-specific mixing effect of the combination of crops i and j on the yield of barley, λ_{ijk} is a tri-specific effect of the combination of crops i, j, and k on the yield of barley, $\pi_{12\cdots v}$ is a v-specific mixing effect of the combination of all v crops on

the yield of barley, and all the εs are considered to have mean zero and common variance σ^2_ε. The assumption of common variance appears to be a realistic one for this experiment involving only barley yields.

10. Three or More Main Crops – Density Constant

A first step in analyzing data from an intercropping experiment containing mixtures of three or more crops is to obtain statistical analyses for each crop separately. The method of Section 9 may be used for this when appropriate. Response model equations for such experiments designed in a randomized complete blocks design, found useful are:

Sole crop g ($h = 1,2,\cdots,r$; $i = 1,2,\cdots,c_g$):

$$Y_{ghi} = \mu_g + \rho_{gh} + \tau_{gi} + \varepsilon_{ghi} .$$

Mixtures of three in proportions $p_1:p_2:p_3$, $p_1 \geq p_2 \geq p_3$):

Crop 1 yield, i'th line

$$Y_{1hi(jk)} = p_1(\mu_1 + \rho_{1h} + \tau_{1i} + \delta_{1i}) + 2p_2\gamma_{i(j)} + 2p_3\gamma_{i(k)} + 3p_3\pi_{i(jk)} + \varepsilon_{1hi(jk)} .$$

Crop 2 yield, j'th line

$$Y_{2h(i)j(k)} = p_2(\mu_2 + \rho_{2h} + \tau_{2j} + \delta_{2j}) + 2p_2\gamma_{(i)j} + 2p_3\gamma_{j(k)} + 3p_3\pi_{(i)j(k)} + \varepsilon_{2h(i)j(k)} .$$

Crop 3 yield, k'th line

$$Y_{3h(ij)k} = p_3(\mu_3 + \rho_{3h} + \tau_{3k} + \delta_{3k}) + 2p_3(\gamma_{(i)k} + \gamma_{(j)k}) + 3p_3\pi_{(ij)k} + \varepsilon_{3h(ij)k} ,$$

where interaction effects $\gamma_{i(j)}$, $\pi_{i(jk)}$, etc. are defined for equal amounts of material on an area basis, ε_{ghi} have zero mean and common variance

$\sigma^2_{g\varepsilon}$, $\varepsilon_{ghi(jk)}$ have zero mean and common variance $\sigma^2_{g\varepsilon 3} = \sigma^2_{g\varepsilon} p_g$, ρ_{gh} is a block effect for crop g, μ_g is a mean effect for crop g, and the subscripts in parentheses denote the other two crops in a mixture. Crops g, g*, and g' were taken to be 1, 2, and 3, respectively. The i'th line of crop g, the j'th line of crop g*, and the k'th line of crop g' is used. In experiments analyzed to date, only one line of each crop was included but the above equations are written to allow for one to c_g lines of each crop. Also, note that each crop's contribution to an interaction term can be estimated.

The construction of created variables as a linear combination of yields is straightforward from the two crop situation. For crop value, one uses $\Sigma_1^c v_g Y_g$ instead of $v_1 Y_1 + v_2 Y_2$. Or, all values v_g may be made proportional to a base crop value, say v_1; the created relative value will be $\Sigma_1^c Y_g (v_g/v_1)$. For calorie (or protein) value, the created variable $\Sigma_1^c c_g Y_g$ or $\Sigma_1^c Y_g (c_g/c_1)$ would be used. For land use values, the linear combination of yields $\Sigma_1^c Y_{gb}/Y_{gs} = \Sigma_1^c L_g$, or $\Sigma_1^c Y_{gb}(Y_{1s}/Y_{gs}) = Y_{1s}\Sigma_1^c L_i$ would be used for Y_{gb} = yield of crop g in a mixture and Y_{gs} = yield of crop g as a sole crop.

Multivariate discriminant function analyses are not usable (see Federer and Murty, 1984) for analyzing data from intercropping experiments. Multivariate theory needs considerable extension before it can be used. Problems of missing values, comparisons of sole crops with linear combinations of some of the crops, comparisons of different linear combinations, and the practical interpretation of the linear combination appear to make present concepts of multivariate theory unusable for intercropping data. Satisfying mathematical considerations and not practical interpretations is a vacuous solution for an experimenter trying to interpret results from an experiment.

11. Three or More Main Crops – Density Variable

With only two crops in a mixture, the assumption that the sole crop regression of yield on density holds for all densities of the second crop may be tenable in a small region of densities. With more than two crops in a mixture and with varying densities, this assumption may not be appropriate. To illustrate, consider mixtures of three crops gg^*g' for g, g^*, $g' = 1,\cdots,c$ crops at densities d_{ig}, d_{jg^*}, and $d_{kg'}$, for $i = 1,\cdots,c_g$, $j = 1,\cdots,c_{g^*}$, and $k = 1,\cdots,c_{g'}$. The regressions could be obtained for each of the $c_{g^*}c_{g'}$ density combinations and not just the sole crop. These regressions could be compared for homogeneity to ascertain whether the sole crop regression is appropriate for mixtures of three. If the regressions can be considered to be homogeneous or relatively so, the following response model equation for the yield of density combination $(d_{ig}, d_{jg^*}, d_{kg'})$ may be expressed as:

$$Y_{ghi(jk)}(d_{ig},d_{jg^*},d_{kg'}) = \beta_{0g} + \rho_{gh} + \beta_{1g}d_{ig} + \gamma_{i(jk)}(d_{ig},d_{jg^*},d_{kg'}) + \varepsilon_{ghi(jk)}(d_{ig},d_{jg^*},d_{kg'})$$

where $i = 1,\cdots,c_g$, $j = 1,\cdots,c_{g^*}$, and $k = 1,\cdots,c_{g'}$, β_{0g}, ρ_{gh}, and β_{1g} are as defined in Section 4, and $\varepsilon_{ghi(jk)}(d_{ig},d_{jg^*},d_{kg'})$ have zero mean and common variance $\sigma^2_{g\varepsilon}$. The $\gamma_{i(jk)}(d_{ig},d_{jg^*},d_{kg'})$ may be partitioned into an overall effect, an effect of crop g^* at density j, an effect of crop g' at density k, and an interaction effect for the jk'th densities of crops g^* and g'. These effects would relate to the yields of crop g.

12. Modeling Responses for Mixtures of Three or More Crops – Individual Crop Responses Available

Various response models for mixtures of two crops were discussed in Section 5. For mixtures of three of c cultivars, say i, j, and k, the following models are considered plausible for consideration using a RCBD:

Sole crop i

$$Y_{hi} = \mu + \tau_i + \rho_h + \varepsilon_{hi} .$$

Mixture ijk

Crop i yield =

$$Y_{hi(jk)} = (\mu + \rho_h + \tau_i + \delta_i)/3 + 2(\gamma_{i(j)} + \gamma_{i(k)})/3 + \pi_{i(jk)} + \varepsilon_{hi(jk)} .$$

Crop j yield =

$$Y_{h(i)j(k)} = (\mu + \rho_h + \tau_j + \delta_j)/3 + 2(\gamma_{(i)j} + \gamma_{j(k)})/3 + \pi_{(i)j(k)} + \varepsilon_{h(i)j(k)} .$$

Crop k yield =

$$Y_{h(ij)k} = (\mu + \rho_h + \tau_k + \delta_k)/3 + 2(\gamma_{(i)k} + \gamma_{(j)k})/3 + \pi_{(ij)k} + \varepsilon_{h(ij)k} .$$

A simpler form for crop i yield from a mixture of three would be

$$Y_{hi(jk)} = (\mu + \rho_h + \tau_i)/3 + \pi^*_{i(jk)} + \varepsilon_{hi(jk)}$$

where δ_i, $\gamma_{i(j)}$, $\gamma_{i(k)}$, and $\pi_{i(jk)}$ are all combined into an effect $\pi^*_{i(jk)}$. The interpretation of the parameters is the same as described in previous sections. Solutions for

$$\hat{\pi}^*_{i(\cdot\cdot)},\ \hat{\pi}^*_{i(j\cdot)},\ \hat{\pi}^*_{i(\cdot k)},\ \text{and}\ \hat{\pi}^*_{i(jk)},$$

subject to usual restrictions may be obtained when all possible combinations of crops are present. Otherwise, it is recommended that the above simpler form be used.

13. Modeling Responses for Mixtures of Three or More Crops – Individual Crop Responses Not Available

Suppose that sole crops and all possible combinations of three of the crops represent the treatments in a RCBD. Possible response model equations are:

Sole crop

$$Y_{hi} = \mu + \rho_h + \tau_i + \varepsilon_{hi} .$$

Mixture ijk

$$Y_{hijk} = \mu + \rho_h + (\tau_i + \delta_i + \tau_j + \delta_j + \tau_k + \delta_k)/3 + 2(\gamma_{ij} + \gamma_{ik} + \gamma_{jk})/3 + \pi_{ijk} + \varepsilon_{hijk} .$$

If all combinations were not present the model for mixtures may be simplified to:

$$Y_{hijk} = \mu + (\tau_i + \tau_j + \tau_k)/3 + \pi^*_{ijk} + \varepsilon_{hijk}$$

where a sum of general mixing (δ_i), bi-specific mixing (γ_{ij}), and tri-specific (π_{ijk}) effects would be represented in π^*_{ijk}.

Several other models described in Section 6 can be generalized to consider three or more crops in a mixture. When v^3 combinations of lines of three crops or factors are present, a three-factor factorial model may be used. Another response model for sole crops and mixtures of three crops i, j, and k would be:

Sole crop

$$Y_{his} = \mu + \rho_h + \tau_i + \varepsilon_{hi} .$$

Mixture ijk

$$Y_{hijkb} = \mu + \rho_h + \tau_i + \delta_i + \gamma_{ijk} + \varepsilon_{hijk}$$

where γ_{ijk} is an interaction effect within crop line i of component one of the mixture for lines j and k of the second and third components. Alternatively, γ_{ijk} could be an interaction effect within the combination ij. To illustrate, suppose that four lines of a crop, say A, B, C, D, are available, that center row yields only will be obtained, and that the center rows will be bordered on one or both sides by every line. For line A, the center and outside rows would be AAA, AAB, AAC, AAD, BAB, BAC, BAD, CAC, CAD, DAD. The interaction effects λ_{Ajk} would be the deviations of the quantities $\bar{y}_{.ijkb} - \bar{y}_{.i..b}$, and the interaction effects λ_{ABk} would be the difference $\pm \bar{y}_{.ABCb} - \bar{y}_{.ABDb}$.

Martin (1980) states that his model does not extend to a three-factor factorial. A response model for a two-factor factorial in a RCBD would be:

$$Y_{hij} = \mu + \rho_h + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{hij} .$$

Martin's model deals with functions of the $\alpha\beta_{ij}$. A corresponding three-factor factorial response model would be:

$$Y_{hijk} = \mu + \rho_h + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{hijk} .$$

Construction of two-factor responses and using the previous model, $\alpha\beta_{ij}$, $\alpha\gamma_{ik}$, and $\beta\gamma_{jk}$ can all be partitioned. Partitioning of the three-factor interaction $\alpha\beta\gamma_{ijk}$ does not appear to be straightforward. One could

collapse two of the factors into a single category and apply the previous Martin model. The other models discussed in Section 6 can likewise be extended.

14. Spatial, Density, and Intimacy Arrangements for Three or More Crops

For two crops, arrangements have been constructed to have one plant of crop one bordered by zero, one, two, three, and four plants of the second crop an equal number of times. Comparable plans for three or more crops have not been devised to date. As long as all plants of three or more cultivars (crops, lines of a crop, etc.) are randomly intermingled in an experimental unit, no difficulty arises. As soon as cultivars are placed in rows or planted in patterns, spatial patterns must be thoughtfully considered. The following items must be investigated for three crops:

(i) density of crop one,

(ii) density of crop two,

(iii) density of crop three,

(iv) spatial arrangement of crop one,

(v) spatial arrangement of crop two,

(vi) spatial arrangement of crop three,

(vii) intimacy of crops one and two,

(viii) intimacy of crops one and three, and

(ix) intimacy of crops two and three.

When using the Nelder fan or the Okigbo wheel, care must be taken in investigating orientation, density, spatial, and intimacy relations. These designs will be parsimonious of space and should be considered as obtaining preliminary results. More extensive investigation will more than likely be

required in order to determine optimal conditions. The above considerations hold for mixtures of k of v crops.

15. Additional Statistics for Mixtures of Three or More Crops

Many of the statistics described in Section 8 may be extended to consider mixtures of three or more crops. The *total effective area* under three crops as sole crops, in mixtures of two, and in a mixture of three would be:

$$A_1 + A_2 + A_3 + L_{12}A_{m12} + L_{13}A_{m13} + L_{23}A_{m23} + L_{123}A_{m123} ,$$

where A_i = area under sole crop i, A_{mij} = area under mixture of two crops i and j, A_{m123} = area under the mixture of three crops, L_{ij} = land equivalent ratio for mixtures of crops i and j, and L_{123} is a land equivalent ratio for mixtures of the three crops.

A *coefficient of agressivity* for two crops in equal proportions of land area is $L_1 - L_2$. For three crops it would be $L_1 - (L_2 + L_3)/2$ for crop 1, $L_2 - (L_1 + L_3)/2$ for crop 2, and $L_3 - (L_1 + L_2)/2$ for crop 3. Extension to k crops is straightforward. L_i = yield of crop i mixture divided by yield of crop i as a sole crop.

A *competitive ratio index* for two crops in equal proportions of land area is L_1/L_2. For three crops, it would be $2L_1/(L_2 + L_3)$, $2L_2/(L_1 + L_3)$, and $2L_3/(L_1 + L_2)$ for crops 1, 2, and 3, respectively.

A *relative crowding coefficient* for two crops is $L_1L_2/(1 - L_1)(1 - L_2)$. For k crops in a mixture, the coefficient would be $\prod_1^k L_i/(1 - L_i)$.

Graphical representations for linear programming can be made for mixtures of two and three crops, but not for mixtures of four or more crops. However, linear programming techniques allow for k crops in a mixture and as sole crops.

16. Other Mixtures Where Statistical Techniques Are Useful

There are a large number of areas where the ideas and statistical procedures developed for intercropping can be used. For example, consider a survey sampling situation where answers are sought to sensitive, incriminating, and/or embarrassing questions. Direct questioning will not allow the surveyor to obtain this information. Anonymity of response is essential in order to obtain the information. Raghavarao and Federer (1979) have shown how to use the block total response procedure using supplemented and balanced incomplete block designs to obtain sensitive information. The respondent is required to give a total of answers to k of v questions. From the various block totals, estimates for the sample can be obtained without knowing individual responses. This is similar to knowing only the total response for a mixture rather than having the individual mixture component responses.

Other areas where these ideas can be utilized is in applications of drugs, therapies, medicines, recreational programs, physical training programs, educational programs, using sequences of courses and other mixtures, nutritional studies, use of pesticide and herbicide mixtures, and any other area where mixtures of components are involved. Studies in these areas to date have centered on mean comparisons of single or similar components, upon single responses for the mixture, and standard statistical procedures. Modeling aspects and competitive aspects have been ignored. Statistical theory has not provided adequate statistical methodology to do more than what is being done. It is a fruitful area for future research and application.

17. Literature Cited

Federer, W. T. (1979). Statistical designs and response models for mixtures of cultivars. *Agronomy J.* 71:701-706.

Federer, W. T. and H. C. Kirton (1984). Incomplete block and lattice rectangle designs for v = 36 using F-square theory. No. BU-850-M in the Technical Report Series of the Biometrics Unit, Cornell University, September.

Federer, W. T. and B. R. Murty (1984). Uses and limitations of multivariate analyses in analysis of intercropping experiments. No. BU-858-M in the Technical Report Series of the Biometrics Unit, Cornell University, September.

Federer, W. T., J. C. Connigale, J. N. Rutger, and A. Wijesinha (1982). Statistical analyses of yields from uniblends and biblends of eight dry bean cultivars. *Crop Sci.* 22:111-114.

Finlay, K. W. and G. N. Wilkinson (1963). The analysis of adaptation in a plant-breeding programme. *Australian J. Agric. Res.* 14:742-754.

Martin, K. J. (1980). A partition of a two-factor interaction, with an agricultural example. *Applied Statistics* 29(2):149-155.

Mead, R. and J. Riley (1981). A review of statistical ideas relevant to intercropping research (with discussion). *J. Royal Statist. Soc., Series A* 144:462-509.

Nelder, J. A. (1962). New kinds of systematic designs for spacing experiments. *Biometrics* 18:283-307.

Okigbo, B. N. (1978). Personal communication. Intl. Inst. Tropical Agric., Ibadan, Nigeria.

Pearce, S. C. and B. Gilliver (1978). The statistical analysis of data from intercropping experiments. *J. Agric. Sci., Cambridge* 91:625-632.

Pearce, S. C. and B. Gilliver (1979). Graphical assessment of intercropping methods. *J. Agric. Sci., Cambridge* 93:51-58.

Raghavarao, D. and W. T. Federer (1979). Block total response as an alternative to the randomized response method in surveys. *J. Royal Statist. Soc., Series B* 41:40-45.

Snedecor, G. W. and W. G. Cochran (1967). *Statistical Methods*, 6th Edition. Iowa State University Press, Ames, Iowa.

SCORING CASUALTIES FROM FIELD TRIALS

Carl T. Russell
US Army Operational Test and Evaluation Agency
Falls Church, Virginia

Marion R. Bryson
US Army Combat Developments Experimentation Center
Fort Ord, California

ABSTRACT. Real Time Casualty Assessment (RTCA) is often used to "shape the battle" in Army operational tests by simulating attrition in near real time as a function of measured engagement conditions. Based on engagement conditions measured by test instrumentation, a computer obtains a Pk from a table of kill probabilities and draws a random number against that Pk to determine whether the target player lives or dies. Both firing and target players are given near real time feedback concerning the result, and "dead" players are removed from the battle as quickly as possible. As long as the attrition rates used real time are approximately correct, RTCA encourages realistic engagement conditions by generally rewarding smart player actions and penalizing dumb ones. That is, "approximately correct" attrition rates suffice to "shape the battle." However, if test measures of effectiveness involve force losses, attrition rates which are only approximately correct are not good enough. Post-test analysis of the battle typically identifies engagements which either were improperly recorded by test instrumentation or were partially garbled during real time computer processing. Alternatively, the analyst may be asked to estimate what force losses would have been with smaller or larger Pk's for some players. Once the actual engagement conditions are determined post test, an actual or hypothetical Pk (PKA) can be determined and compared to the Pk used real time (PKU). Whenever PKA differs from PKU, the attrition rate used real time was inappropriate and may have started a cascade of misleading real time losses. The analytic goal is to estimate what expected losses would have been if live ordnance (having true Pk=PKA) had been used. "Aliveness analysis" is a computational technique which attempts to meet this goal by crediting kills adjusted for cumulative differences between PKA and PKU. The technique originated at CDEC and was modified for application to the SGT York Follow on Evaluation conducted in April-May 1985. This paper discusses the aliveness analysis technique and illustrates the technique using examples based on this SGT York testing.

REAL TIME CASUALTY ASSESSMENT

RTCA Description. Army use of Real Time Casualty Assessment (RTCA) originated at the Combat Development Experimentation Center (CDEC) in the 1970's and is used extensively in tests conducted on CDEC test ranges at Fort Hunter Liggett, California. RTCA is an instrumented testing technique which shapes the battle by simulating kills in near real time.

Field trials at CDEC are generally two-sided free-play trials up to battalion versus company in size, and they may involve armor, infantry, aviation, air defense, field artillery, mines, or chemical equipment. The trials are conducted under conditions as nearly realistic as possible, and they are highly instrumented for trial control, safety, and data collection. The computerized instrumentation consists of a number of fixed stations (A stations) which poll transponders mounted on players (B stations) and transmit position location and other data back to a central computer (C station) for processing (see Figure 1). Players are instrumented, typically using coded lasers, so that when one player fires at another, the target can be identified. During the play of a test battle, RTCA encourages realistic engagement conditions by generally rewarding smart player actions and penalizing dumb ones. It does this through a three step process (see Figure 2):

- When an engagement occurs (i.e., one player "fires" at another), a coded laser is fired at the target. The B station on the firer tells the computer "I fired" and the type of ammo used. If a target player is paired with the firer, the B station on the target tells the computer "I've been engaged," the code of the laser, and the sensors illuminated.

- The computer receives the engagement information and analyzes it in terms of variables which affect the probability of kill (typically, the nature of the firer, the nature of the target, the range, the ammunition used, firer and target movement, target exposure, and target aspect). Tables of "kill probability" (Pk) determined by pretest modeling are then used to determine the Pk associated with the crucial engagement parameters. This Pk is then used to simulate a "kill" or a "survive" via Monte Carlo. That is, the computer draws a random number against the looked-up Pk, killing the target if the random number is smaller than the Pk.

- This simulated engagement result is fed back to both firer and target in the engagement, usually within a few seconds of the original firing. Dead targets either stop (ground players) or leave the battlefield as soon as possible (air players). Dead players are typically marked by strobes or smoke and their ability to fire at others is disabled.

RTCA Interpretation. Field trials as conducted at CDEC are simulations of actual combat, not reality itself. Representative battle initial conditions are determined prior to the start of testing, and RTCA is used to shape the test battle so that post-test estimates of attrition will provide reasonable predictive insight to actual battle outcomes (see Figure 3). The role of RTCA in this simulation is as a tool to encourage sequences of individual engagement conditions representative of combat under the specified initial conditions.

Figure 1
Position Location Instrumentation

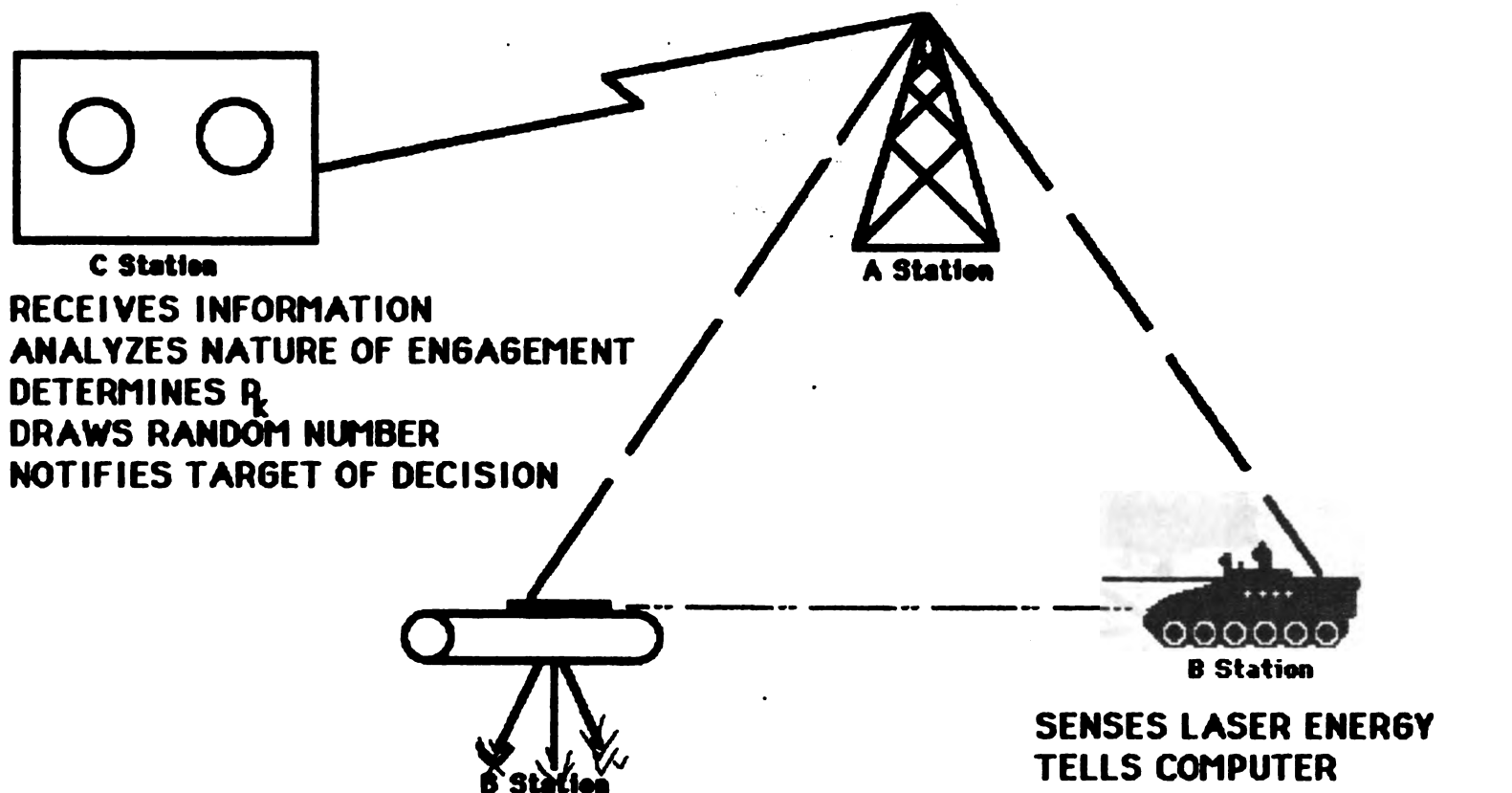


Figure 2
The RTCA Engagement

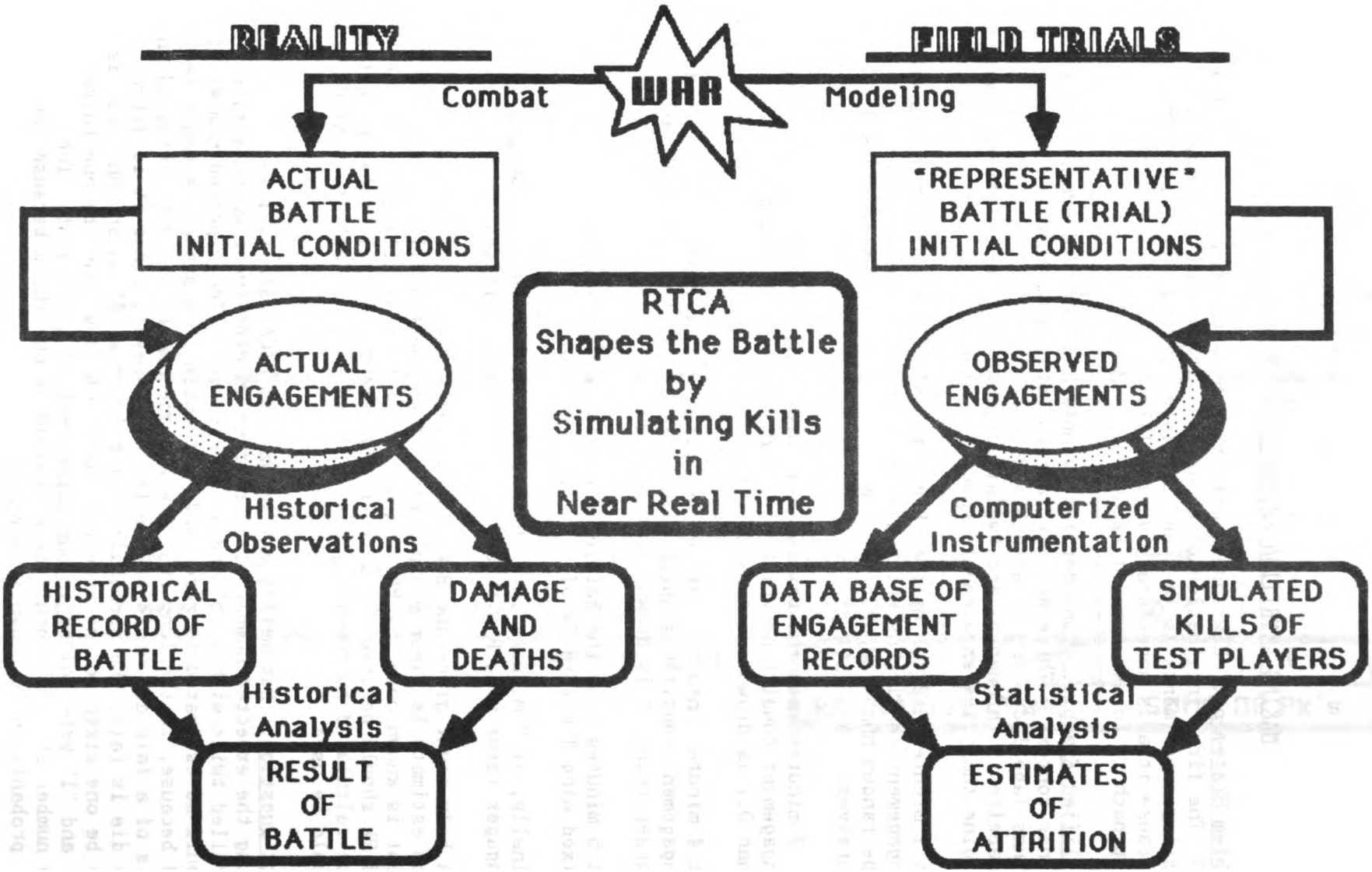


Figure 3
How RTCA Is Interpreted in CDEC Field Trials

MAKING ATTRITION ESTIMATES FROM RTCA DATA

Problem Statement. How should attrition be estimated in the context of RTCA? The first inclination of an analyst accustomed to binomial coin-tossing experiments is to count simulated kills. That approach is wrong because it adds unnecessary variability to attrition estimates. The right approach is to use sums of Pk's.

Simplified Example. Consider for example a simplified example in an air defense context. Three Red aircraft (a rotor wing labelled R and fixed wings labelled S and T) attack a Blue armor force consisting of five tanks (labelled 1 through 5) and two air defense weapons (labelled A and B). Imagine that five engagements ensue (see Table 1):

- At 1 minute into the battle, Red rotor wing R engages tank 1 under engagement conditions which give a Pk of 0.6. The computer draws the random number 0.8635 (or some such) against 0.6, so tank 1 survives.

- At 3 minutes into the battle, Red rotor wing R engages tank 2 under engagement conditions which give a Pk of 0.7. A random number less than 0.7 is drawn, so tank 2 is killed.

- At 4 minutes into the battle, Red fixed wing S engages tank 3 under engagement conditions which give a Pk of 0.1, but a random number larger than 0.1 is drawn, so tank 3 survives.

- At 5 minutes into the battle, Blue air defense weapon B engages fixed wing T with a Pk of 0.8, but T survives.

- Finally, at 7 minutes into the battle, Blue air defense weapon A engages rotor wing R with a Pk of 0.3, but R survives.

The RTCA body count gives one Blue and no Red killed, but a better attrition estimate is clearly available. The expected kill on each engagement is known once the Pk is known, so a partial kill equal to the observed Pk should be credited at each engagement. Overall expected kills should be calculated by summing these credited kills. That is, sums of Pk's should be used.

Trick Question. Estimating attrition from RTCA data is like estimating the expected number of "3's" from an experiment where a fair die is rolled twice with a "5" and a "1" observed. The expected number of "3's" could be estimated as zero, and that estimation procedure would be unbiased because, in the long run, the average number of "3's" obtained in two rolls of a fair die would be one third. However, since it is given that the die is fair, the probability of rolling a "3" on any one roll is known to be one sixth so the expected number of "3's" must be one third. That "5" and "1" were rolled in one experiment is irrelevant. The expected number of "3's" should be estimated as one third because the relevant probability is known in advance.

Table 1
Simplified Example Showing that
Sums of Pk's Should be Used to Estimate Attrition

TIME INTO TRIAL	FIRER	TARGET	Pk OBTAINED		SIMULATED RESULT		CREDITED KILLS	
			vsRED	vsBLUE	RED	BLUE	vsRED	vsBLUE
1 min	Rtr Wng R	Tank 1		0.6		SURVIVE		0.6
3 min	Rtr Wng R	Tank 2		0.7		KILL		0.7
4 min	Fxd Wng S	Tank 3		0.1		SURVIVE		0.1
5 min	Air Def B	Fxd Wng T	0.8		SURVIVE		0.8	
7 min	Air Def A	Rtr Wng R	0.3		SURVIVE		0.3	
		EXPECTED KILLS =			~~?? 0.2?~~	~~?? 1 ??~~	1.1	1.4
					~~RISK~~ BODY COUNT		SUMS OF Pk's	

Crucial Points. In summary, there are three crucial points to be remembered when making attrition estimates from RTCA data.

- The kill probability Pk for each engagement is obtained directly from measured engagement conditions via a a Pk table, not estimated from simulated kills.

- A simulated kill or survive is generated by a draw of a random number against engagement Pk, not observed directly from the engagement.

- RTCA encourages realistic engagements by providing quick feedback to players in terms of simulated kills, but attrition should be estimated using sums of Pk's, not sums of simulated kills.

OVERKILL INHERENT WITH SUMS OF Pk's

Inherent Overkill. There is an inherent difficulty with estimating attrition using sums of Pk's: individual players or groups can be "killed" more than once. This fact has been used as an argument against the aliveness formulas which will be discussed shortly, but the difficulty is actually inherent with simple sums of Pk's. The initial reaction of most analysts to such overkill is to consider it intolerable and attempt to modify the way kills are credited in order to insure that no more than one kill is ever credited against an individual player. There is at least one case where such deflation of overkill is indeed desirable. In general, however, overkill is desirable. The following discussion:

- illustrates how overkill can occur by revisiting the simplified example discussed in the preceding section,

- shows how to deal with one case of undesirable overkill, and

- gives a rather elaborate example, in terms of a hypothetical experiment, which provides a test for any estimation method proposed as an alternative to sums of Pk's.

Revisited Example. If in the simplified example of Table 1, rotor wing R had fired at surviving tank 1 a second time rather than firing at tank 2, the overall attrition estimates should not change. However, 1.3 kills must be credited against tank 1. There is no way around this problem if unbiased estimates of attrition are desired. Overkills are necessary to compensate for only partial kills credited against totally dead players, as originally happened against tank 2 in Table 1: only 0.7 kill was credited but tank 2 was forever 100% dead.

Undesirable Overkill. One situation where overkill is clearly undesirable occurs when one player fires several rounds at another player over a relatively short period of time so that the rounds should be considered only one engagement. Then the engagement Pk should be

calculated using appropriate products of Pk's rather than sums of Pk's. In particular, if n rounds are fired, each with Pk=p, then $1-(1-p)^n$ should be used rather than np as the engagement Pk. As long as np is small, the difference between the two formulas is slight, since $np-[1-(1-p)^n]$ cannot exceed $(np)^2$ (proof: use binomial expansion). However, if np is relatively large (in particular, if np is greater than 1), then the difference is large. For example, if three antitank weapons are fired by the same firer against the same tank during a very short period of time and each firing has Pk=0.7, then np=2.1 but $[1-(1-p)^n]=0.973$. The better answer is clearly 0.973.

Desirable Overkill. If products of Pk's rather than sums of Pk's were used to credit kills, no more than one credited kill could ever be accumulated against a single player. Thus overkill could be avoided if products of Pk's rather than sums of Pk's were always used to credit kills for estimating attrition. The following example shows that crediting kills using products is misleading because it generally underestimates expected attrition. In addition, because it provides a situation with real bullets and real deaths for which the true expected kills can be calculated, the hypothetical experiment in this example provides a test for any estimation method proposed as an alternative to sums of Pk's (see Figure 4).

Hypothetical Example:

- An urn contains 1,000,000(1-p) harmless blanks (painted green) and 1,000,000p absolutely lethal rounds (painted red).
- Draw 100 rounds at random from this urn and load them into a gun (which conveniently holds 100 rounds).
- Now select 100 volunteers and shoot the rounds at them from point blank range.

As long as care is taken to shoot only at live volunteers, 100p volunteers are expected to die. The number of expected kills is the same no matter which of the following methods is used to distribute shots among volunteers:

- Shoot one round at volunteer #1. Shoot one round at volunteer #2. ... Shoot one round at volunteer #100.

- Randomly select (with replacement) a living volunteer. Shoot one round. Repeat the random selection and shooting of one round until all rounds are expended.

- Randomly select (without replacement) a living volunteer. Shoot rounds at this volunteer until all rounds are expended or the volunteer dies. Repeat until all rounds are expended.

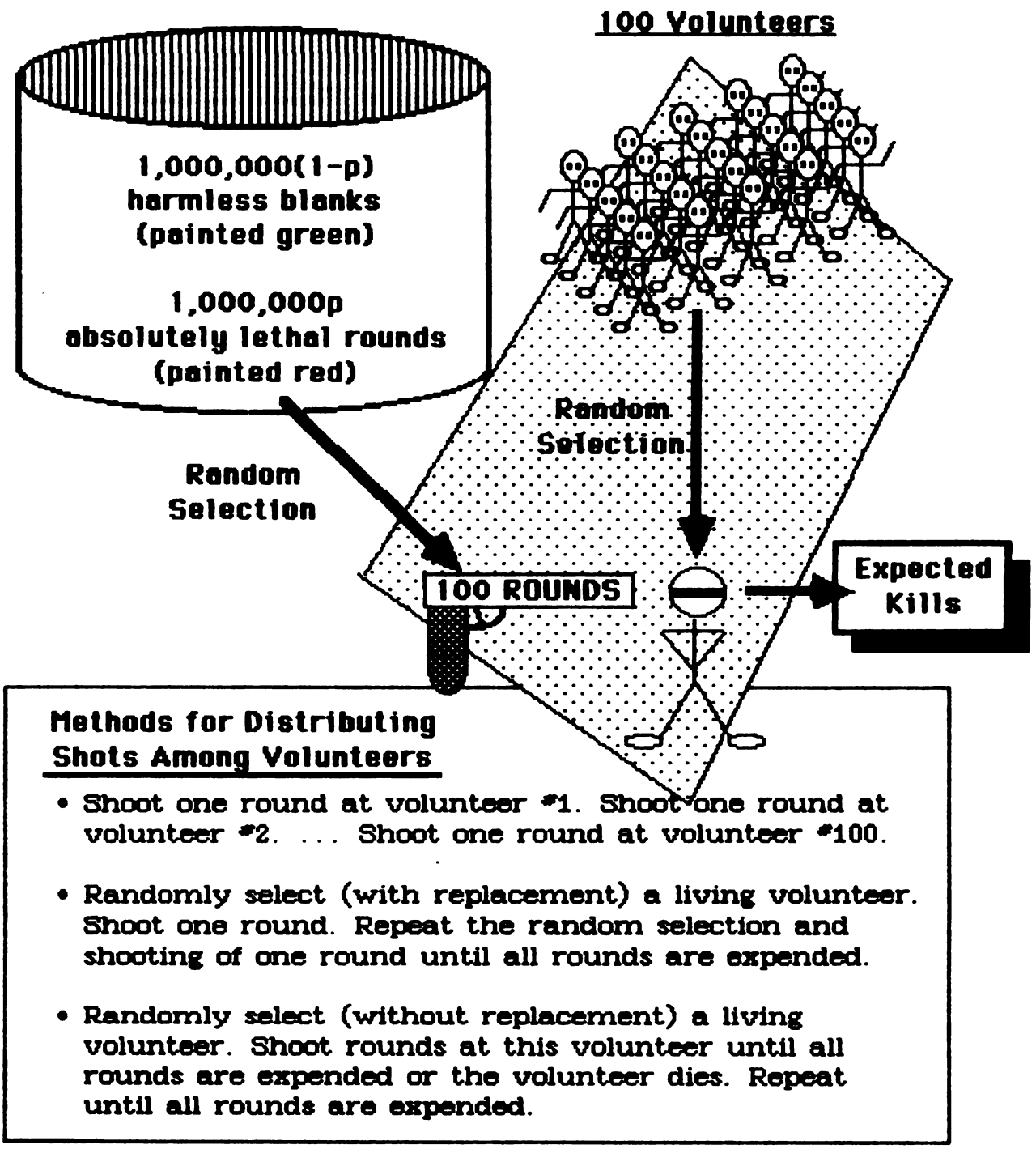


Figure 4
Hypothetical Experiment:
Any Acceptable Method for Estimating Attrition
Should Give 100p Expected Kills

To be convinced that all three methods give the same expected kills, simply consider the number of lethal rounds loaded. If R red rounds (and 100-R green rounds) are loaded into the gun, then exactly R volunteers will die. Thus all three methods give the same number of kills, hence the same expected kills. Moreover, all 100 rounds will be fired with each method. The random variable R has essentially binomial distribution with success probability p and N=100 (actually the distribution is hypergeometric), and the expected value is 100p. Any acceptable method of estimating attrition using Pk's should estimate the true expected kills, and sums of Pk's does. Since p is known, the expected value can be correctly estimated by crediting a partial kill equal to p on each shot and adding credited kills (sums of Pk's) to give 100p. If p is of moderate size, then crediting more than one kill against some volunteers is virtually certain using either the second or third methods: in particular, if p=0.6, overkill would occur against any player shot at more than once, and both the second and third methods virtually guarantee multiple shots against some players. Applying the product formula would avoid such overkill, but would give the wrong estimate for expected kills. In fact, the product formula would generally give different answers for each of the shot distribution methods if $0<p<1$ and $0<R<100$, and except for the first method, the answers themselves would be random:

- For the first method, the product formula would credit a partial kill equal to p against each player, giving the same answer as sum of Pk's, namely, 100p.

- For the second method, the product formula would credit p kill on the first shot against a particular player, credit $p(1-p)$ kill on the second shot, and in general credit $p(1-p)^{(k-1)}$ kill on the kth shot against a particular player. Since $p(1-p)^{(k-1)}<p$, credited kills would be less than 100p unless all players were shot at exactly once (an extremely unlikely occurence unless p is very near 1). Since at least R different players must be fired at a first time, credited kills must be at least Rp. The remaining (100-R) rounds must also be shot and the smallest number of kills would be credited if all those rounds were shot against a single player. Thus the smallest number of kills credited by the second method would be $(R-1)p+[1-(1-p)^{100-R+1}]$, and the actual number of credited kills would vary randomly from this number to 100p.

- For the third method, the smallest number of credited kills would be the same as the second method, but the largest number of credited kills would be strictly less than 100p unless R=99 and the last round left is green. In general the third method would give substantially less credited kills than either other method since at most R+1 players would be shot at.

Overkill Wrap-up. The preceding discussion shows that some overkill must be allowed when estimating attrition in an RTCA experiment. As long as expected kills are to be estimated by crediting partial kills, some players will be removed from play with less than a whole credited kill, so other players must be allowed to accumulate more than a whole credited kill to make up for the shortfall.

WHY ALIVENESS ANALYSIS IS NEEDED AND WHAT IT IS

Why Needed. Until now, this paper has argued that sums of Pk's should be used to estimate attrition from RTCA data. However, this argument was based on the tacit assumption that RTCA works as advertised, assessing all or almost all engagements correctly. Unfortunately, many engagements which should go to real time assessment do not, typically because of instrumentation failure, faulty real-time position-location data or buffer syncronization problems in real-time computer processing. In addition, the Pk's used real time may prove to be incorrect due to software errors, errors in Pk tables, faulty real-time position-location data or simply inability of instrumentation to capture crucial engagement conditions in real time. Moreover, Pk's may change post test either because new data indicates the Pk tables should be modified or because "what-if" analyses of Pk's are desired. In fact, there are effectively two Pk's associated with each RTCA engagement, the Pk used real time (PKU) and the actual Pk (PKA) determined through post-test analysis. Whenever the PKA's are not equal to the PKU's, the attrition rate applied real time was wrong, and too many or too few players were left on the test battlefield. Simply summing the Pk's (that is, the PKA's) could give misleading estimates of attrition if PKA's were frequently unequal to PKU's. Aliveness analysis is an arithmetic adjustment for cumulative differences between PKA's and PKU's which is applied prior to summing Pk's. It is essentially a back of the envelope calculation too big to do on the back of an envelope.

Adjustment Approach. Aliveness analysis makes sensible adjustments to attrition estimates by reducing or increasing credited kills to compensate for cumulative errors in attrition.

- If PKU is less than PKA then too little real time attrition was applied and the subsequent attrition capability of the target should be reduced.

- If PKU is greater than PKA then too little real time attrition was applied and subsequent attrition capability of the target should be increased.

- On the other hand if PKU is equal to PKA then real time attrition was just right and no adjustment should be applied.

Adjustment Formulas. The concept of aliveness analysis was originated at CDEC by M. Bryson. The largest application of aliveness analysis to date was in the analysis of the force-on-force portion of SGT York Follow

on Evaluation I (FOE I), which will be described below. Prior to the start of FOE I the authors of this paper worked together to refine the aliveness methodology and produce specific formulas for performing aliveness analysis. Aliveness analysis adjusts for differences between real time and post test attrition rates by crediting partial kills via "potency" or "aliveness" weightings on live players as follows.

Define a "potency" or "aliveness" factor A for each player, where $A_{initial}=1$ for all players. Track cumulative credited kills by player I versus player J as K(I,J) with $K_{initial}(I,J)=0$ for all player pairs. Then when player I [potency $A_{old}(I)$] engages player J [potency $A_{old}(J)$] with kill probabilities PKA (actual, from post-test analysis or revised table) and PKU (used in RTCA), adjust potency factors and cumulative credited kills as follows:

$$K_{new}(I,J) = K_{old}(I,J) + A_{old}(J)\times[1-(1-PKA)^{A_{old}(I)}]$$

$$A_{new}(I) = A_{old}(I)$$

$$A_{new}(J) = A_{old}(J)\times[(1-PKA)^{A_{old}(I)}]/(1-PKU).$$

Potency of the target, $A_{new}(J)$, is reduced to zero for any engagement which goes to real time assessment and results in a dead target.

Formula Motivation. The underlying motivation for these formulas is straightforward. First, the calculation adjusts potency of surviving players as a ratio of survival probabilities (provided the firer has A=1):

- If a player survives with twice the probability he should have (for example, if PKA=0.6 and PKU=0.2), his potency is halved.
- If a player survives with half the probability he should have (for example, if PKA=0.6 and PKU=0.8), his potency is doubled.

Second, the odd-looking exponential adjustment for the potency of the firer is actually based on a standard statistical formula:

- n firings with Pk=p give total $Pk=1-(1-p)^n$
- a potency n player firing with Pk=p gives total $Pk=1-(1-p)^n$.

In addition, the calculation

- reduces to sums of Pk's when PKA's always equal PKU's
- adjusts in the right direction when firer potency is 1, and
- performs well in practice, as the rest of this paper shows.

APPLICATION OF THE ALIVENESS CALCULATION

Examining Aliveness. The most straightforward way to examine the aliveness calculation is to observe how it performs on actual sequences of

engagements. The remainder of this paper describes the application of aliveness methodology to score casualties in the field trials of SGT York FOE I.

Test Description. The force-on-force portion of FOE I was conducted at the US Army Combat Developments Experimentation Center (CDEC), Fort Hunter Liggett, California, from 2 April to 22 May, 1985. It was a platoon level test conducted to compare capabilities of three different air defense families to provide protection to an armor battalion task force in similar types of missions. The three air defense families were nominally called "SGT York," "Baseline," and "Alternate." All three families had five Stinger missile systems forward and two Chaparral/FLIR missile systems with the battalion trains. What distinguished the families was the large air defense systems deployed forward:

- The SGT York family had four SGT York air defense gun systems forward.

- The Baseline family had four Vulcan air defense gun systems forward.

- The Alternate family had two Vulcan air defense gun systems and two Chaparral/FLIR missile systems forward.

Test Players. Overall, there were typically more than 60 Blue players and more than 30 Red players in each trial. In addition to Blue air defense, the Blue armor task force consisted of roughly 26 Abrams tanks, 13 Bradley fighting vehicles, and 20 other Blue ground forces (M113's, trucks, etc.). The Red air attack force consisted of four fixed wing (two Fitters surrogated by A-7's and two Frogfoots surrogated by A-10's) and four rotor wing (four Hinds or four Havocs, each surrogated by AH-64's). Three surrogate Red ECM aircraft (one fixed wing and two rotor wing stand-off jammers) were present on some trials, and three Blue aircraft (one AH-1S rotor wing and two F-4 fixed wing) were used to investigate possible fratricide. Finally, a small Red armor force (20 T-80 tanks surrogated by M-60's and 8 BMP's surrogated by M113's) permitted a limited armor battle.

Test Criteria. The main mission performance criteria for FOE I addressed the relative proportion of Blue Force losses to Red Air during trials when the three different families of air defense systems were present. That is, for "similar" trials involving each family, it was necessary to estimate Blue Force losses to Red Air, divide by Blue Force size to estimate the proportion lost, and then form appropriate ratios of the proportions. With

Y = Proportion lost in SGT York trials,

B = Proportion lost in Baseline trials, and

A = Proportion lost in Alternate trials,

the required ratio for comparing SGT York versus Baseline was

$$(B-Y)/B = 1 - Y/B$$

while the required ratio for comparing Alternate versus Baseline was

$$(B-A)/B = 1 - A/B.$$

Over 70 trials were attempted during FOE I, and 52 trials (29 York, 12 Baseline, and 11 Alternate) were validated for analysis. Proportion lost was estimated in each trial, and the appropriate ratios were estimated in an analysis of variance framework in order to adjust for differences in trial conditions between families. The criteria values and detailed results are classified, so they will not be discussed fully in this paper.

FOE Problems. In FOE I, there were frequent differences between PKA and PKU. The most common case of inequality between PKA and PKU was when PKA>PKU=0, which typically occured when engagements did not go to real time assessment (that is, no firer-target pairing could be made in real time) but engagement conditions (hence PKA's) were reconstructed through post-test analysis. In fact, in SGT York FOE I, across various firer-target categories, 40 to 50 percent of engagements did not go to real time assessment (see Figure 5), but PKA's were frequently recovered through post-test analysis (indicated by the dotted areas in Figure 5). However, the percent recovered was substantially different for different firer categories because availability of post-test data sources such as video and audio tapes was different for different firer categories. In addition to the cases PKA>PKU=0, there were many cases of 0<PKA<PKU, which occurred because the pre-test tabulations of Pk's for Blue air defense versus Red air were generally too high. Figure 6 shows as an example the extent to which PKA's differed from PKU's for firings by selected Blue air defense firers against selected Red air targets. For these cases, PKA's equalled PKU's less than a third of the time. Almost the only time when PKA's equalled PKU's was when both were zero (dotted portion of PKA=PKU bars in Figure 6).

FOE I Examples. The following three examples are based on engagements in FOE I. Ficticious Pk values and firer-target pairs have been used in order to keep this paper unclassified. However, the actual engagement sequences led to essentially the same aliveness calculations given here. These examples provide convincing evidence that the aliveness calculation performs well in practice.

Example 1. The first example consists of two engagements against a surrogate Frogfoot close air support aircraft. In the first engagement, the actual survival probability (1-PKA=0.62) was 2.14 times what was applied in real time (1-PKU=0.29). That is, in a large number of such engagements, there would be 2.14 times as many survivors as were observed real time. The aliveness calculation increases the potency of Frogfoot-1 to 2.14 and credits 0.38 kill against Frogfoot-1. The second engagement did not go to real time assessment so the real time survival probability

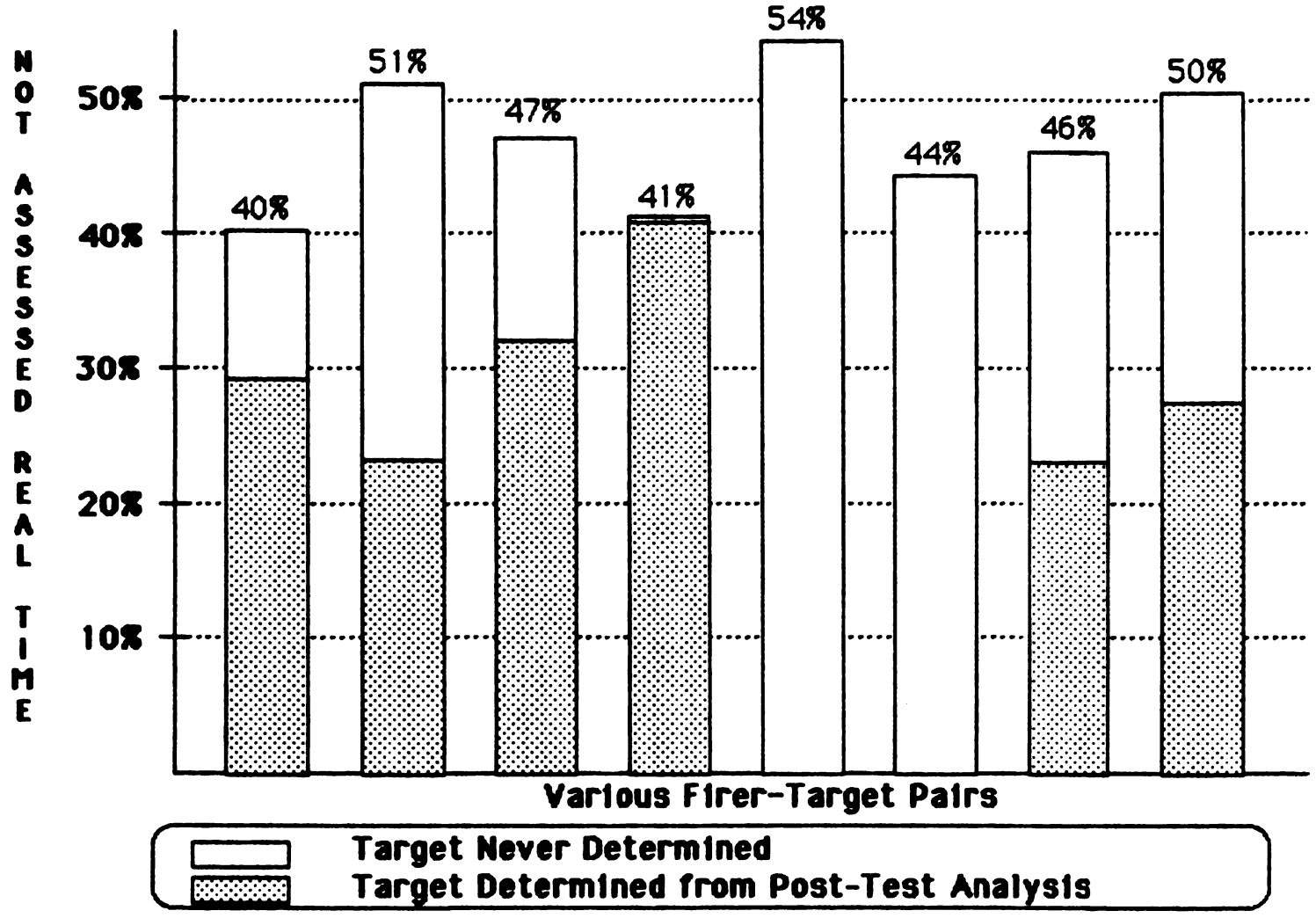


Figure 5
Forty to Fifty Percent of Engagements
Did Not Go to Real Time Assessment

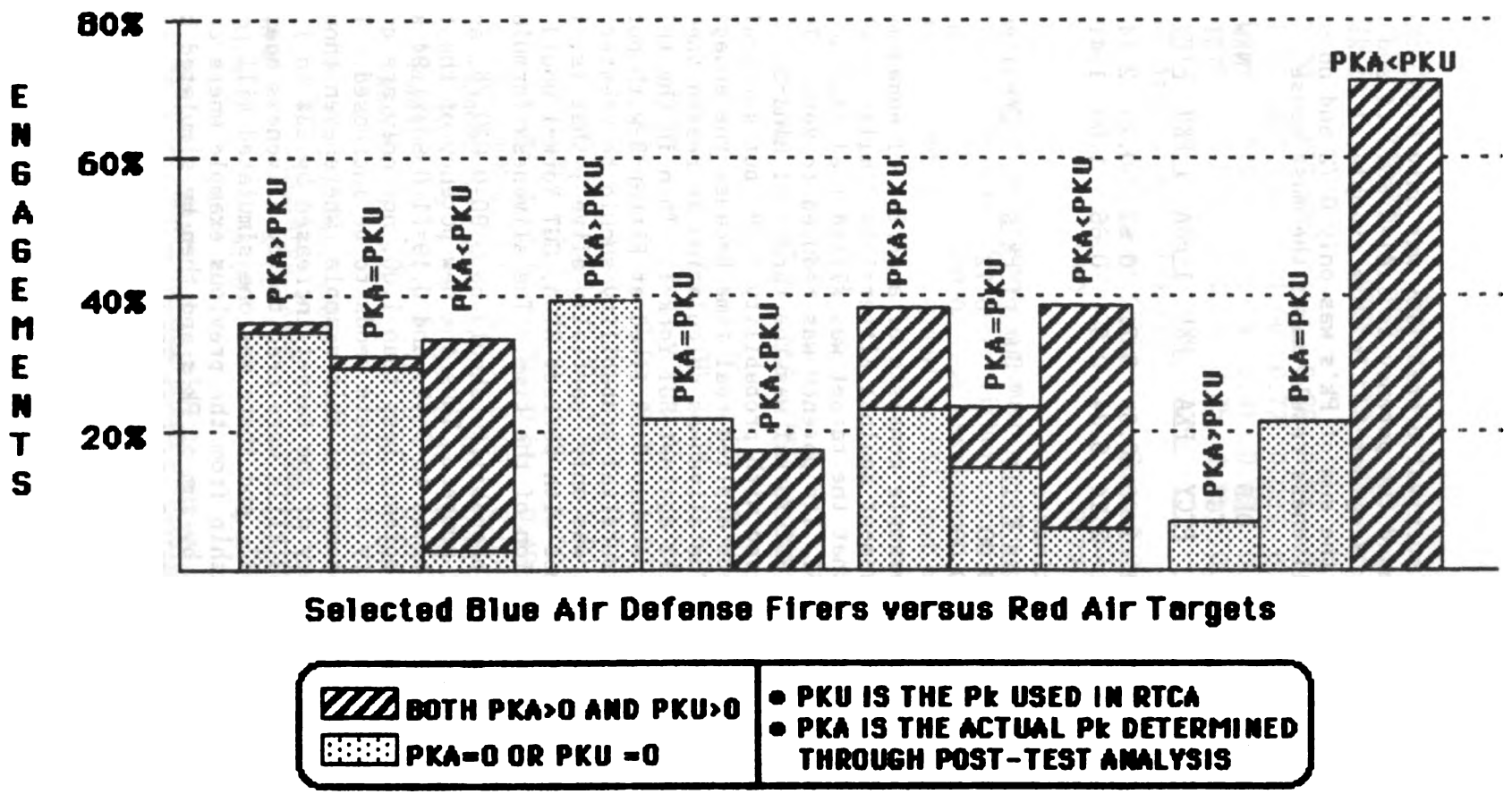


Figure 6
Many Pk Values Were Modified Post Test

was 1.00. Since the actual survival probability was only 0.66, the aliveness calculation decreased the potency to 1.41=0.66x2.14 and credited 0.73=2.14x0.34 kills against Frogfoot-1. In this example, 1.11 kills were credited by aliveness while the sum of Pk's was only 0.72 and no simulated kills were observed. The aliveness result makes the most sense.

FIRER ID	OLD FIR PTCY	TARGET ID	OLD TGT PTCY	PKA	PKU	1-PKA	1-PKU	NEW TGT PTCY	CRTD KILL	SIM KILL
STINGER-4	1.00	FROGFOOT-1	1.00	0.38	0.71	0.62	0.29	2.14	0.38	SURV
STINGER-1	1.00	FROGFOOT-1	2.14	0.34	0.00	0.66	1.00	1.41	0.73	N/A

Summary of Attrition Estimates:

	Simulated Kills	Sum of Pk's	Credited Kills
Against Red	0.00	0.72	1.11
Against Blue	0.00	0.00	0.00

Example 2. The second example consists of a series of engagements by SGT York-1. The first engagement against Fitter-1 is similar to those already considered, except that the target was killed real time so that instead of increasing to 1.61, its potency was reduced to zero. In the second engagement, however, SGT York-1 was the target of Hind-3. SGT York-1 should have survived with probability 0.28, but survival probability 1.00 was in effect applied real time because the engagement did not go to assessment. The aliveness calculation decreased the potency to 0.28 and credited 0.72 kill against SGT York-1. Then in the third engagement, SGT York-1 with potency 0.28 fired at Fitter-3 with potency 1.94. The effective actual survival probability should be greater than 1-PKA=0.69 because the firer was only "partially alive"; that is, if this trial were repeated many times with perfect RTCA, SGT York-1 would only be around to fire a small fraction of the time. The aliveness formula says that the effective survival probability should be $0.90=0.69^{0.28}$, and intuition suggests no better number. Thus the new potency of the target increased by 1.77=0.90/0.51 times to 3.42, and 0.19=(1-0.90)x1.94 kill was credited. Once a firer has potency less than 1.00, not only are credited kills reduced but also potency of targets tends to be increased. This is illustrated by the last engagement of this example, where even though PKA and PKU were the same, potency of the target increased by 23% to 1.23. Overall in this example, 0.53 kills were credited by aliveness against Red while the sum of Pk's was 0.82, and there was one simulated kill. This is exactly the reverse relationship from the previous example where credited kills were largest, followed by sum of Pk's and then by simulated kills. Again, the aliveness result makes the most sense.

FIRER ID	OLD FIR PTCY	TARGET ID	OLD TGT PTCY	PKA	PKU	1-PKA	1-PKU	NEW TGT PTCY	CRTD KILL	SIM KILL
SGT YORK-1	1.00	FITTER-1	1.00	0.26	0.54	0.74	0.46	0.00	0.37	KILL
HIND-3	1.00	SGT YORK-1	1.00	0.72	0.00	0.28	1.00	0.28	0.89	N/A
SGT YORK-1	0.28	FITTER-3	1.94	0.31	0.49	0.69	0.51	3.42	0.06	SURV
SGT YORK-1	0.28	HIND-3	1.00	0.25	0.25	0.75	0.75	1.23	0.05	SURV

Summary of Attrition Estimates:

	Simulated Kills	Sum of Pk's	Credited Kills
Against Red	1.00	0.82	0.82
Against Blue	0.00	0.72	0.72

Example 3. The final example is much more routine, involving a firer with aliveness one, where PKU's were either correct or were zero because the engagement did not go to real time assessment. In all but one instance, the credited kill was equal to PKA, and all three measures of attrition nearly agree.

FIRER ID	OLD FIR PTCY	TARGET ID	OLD TGT PTCY	PKA	PKU	1-PKA	1-PKU	NEW TGT PTCY	CRTD KILL	SIM KILL
HIND-2	1.00	ABRAMS-10	1.00	0.57	0.57	0.43	0.43	0.00	0.57	KILL
HIND-2	1.00	ABRAMS-5	1.00	0.38	0.38	0.62	0.62	1.00	0.38	SURV
HIND-2	1.00	ABRAMS-13	1.00	0.45	0.45	0.55	0.55	0.00	0.45	KILL
HIND-2	1.00	ABRAMS-13	1.00	0.00	0.00	1.00	1.00	0.00	0.00	D/T*
HIND-2	1.00	ABRAMS-7	1.00	0.46	0.00	0.54	1.00	0.54	0.46	N/A
HIND-2	1.00	ABRAMS-5	0.00	0.00	0.00	1.00	1.00	0.00	0.00	D/T*
HIND-2	1.00	UNKNOWN	1.00	0.00	0.00	1.00	1.00	1.00	0.00	N/A
HIND-2	1.00	ABRAMS-14	1.00	0.51	0.51	0.49	0.49	1.00	0.51	SURV
HIND-2	1.00	SGT YORK-4	1.00	0.95	0.00	0.00	1.00	0.05	0.95	N/A
HIND-2	1.00	ABRAMS-16	1.00	0.39	0.00	0.61	1.00	0.61	0.39	N/A
HIND-2	1.00	ABRAMS-16	0.61	0.48	0.48	0.52	0.52	0.00	0.29	KILL
HIND-2	1.00	BRADLEY-10	1.00	0.72	0.72	0.28	0.28	0.00	0.72	KILL
HIND-2	1.00	ABRAMS-16	0.00	0.00	0.00	1.00	1.00	0.00	0.00	D/T*

*D/T=DEAD TARGET

Summary of Attrition Estimates:

	Simulated Kills	Sum of Pk's	Credited Kills
Against Red	0.00	0.00	0.00
Against Blue	4.00	4.91	4.72

OVERALL IMPACT OF ALIVENESS ANALYSIS IN FOE I

Trial-by-trial Summary. Figure 7 displays trial-by-trial estimates for the proportion of blue ground lost to red air by each of the three methods. It shows that results of the aliveness calculation tended to fall between simulated kills and results obtained by sums of Pk's. This occurs because the most common RTCA error was failing to go to real time

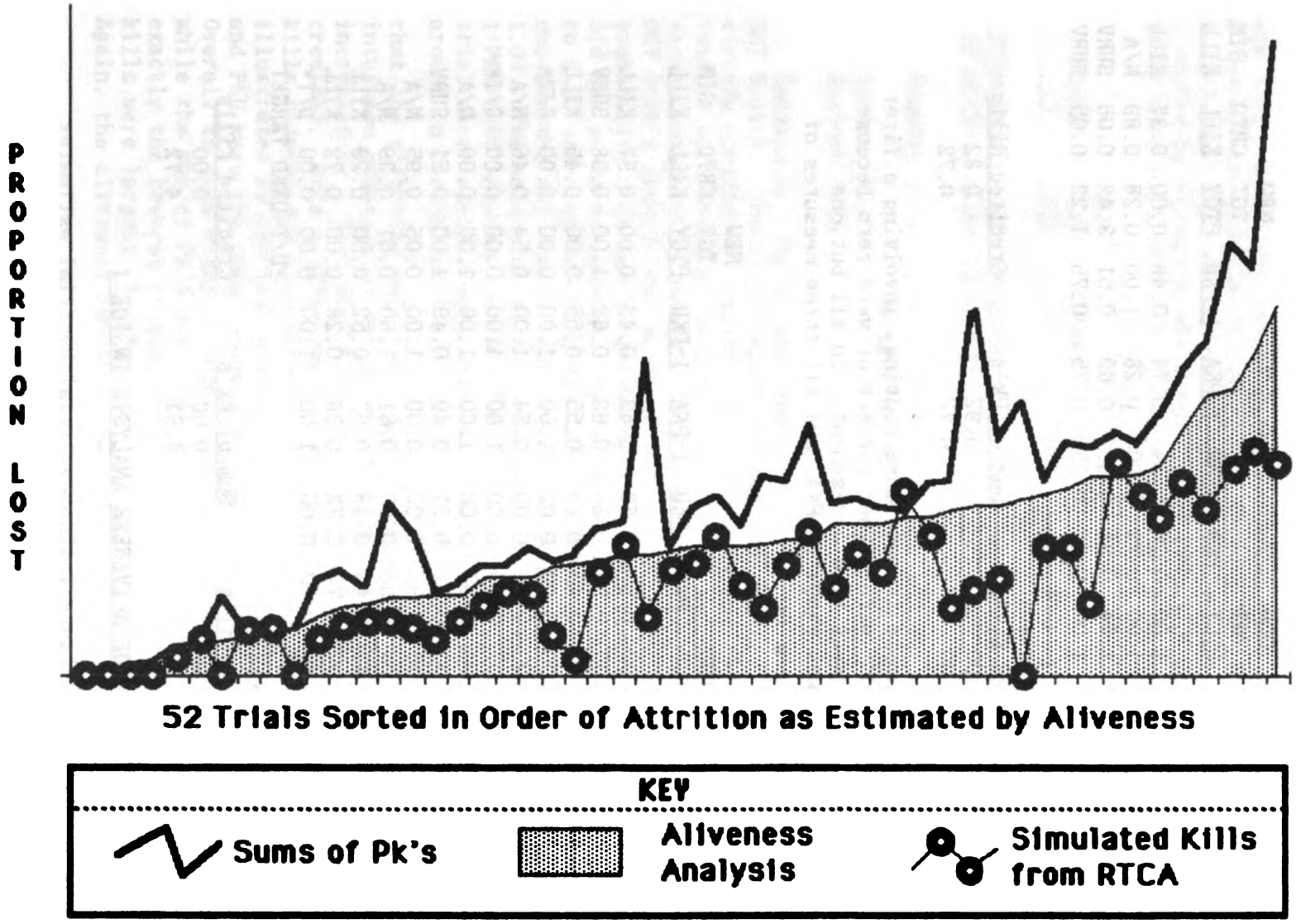


Figure 7
Comparison of Three Methods for Estimating the Proportion Blue Ground Force Lost to Red Air

assessment when a Pk>0 should have been used, which produces no simulated kills and gives potency less than one to survivors. Differences between the attrition estimates were substantial in some trials. (The apparent smoothness of the aliveness curve in Figure 7 compared to the other two curves is due primarily to the order in which trials were sorted for plotting.)

Analysis of Variance Results and Overall Conclusions. For engagements by Red air against Blue ground during FOE I, the tendency for sums of Pk's to be larger than credited kills from aliveness - which in turn tend to be larger than simulated kills - carried over to the attrition estimates obtained from analysis of variance in a general linear models framework. Sums of Pk's were used instead of aliveness analysis to present attrition estimates to decision makers because sums of Pk's are simpler and they gave the same result as aliveness for the crucial SGT York family - the criteria were met. In retrospect, this agreement appears to have been luck. Even though the direction of comparisions between attrition estimates based on sums of Pk's, aliveness, and simulated kills was consistent across air defense families, the relative size of the differences between estimates was not consistent across families. In one case involving Alternate and Vulcan families, a crucial estimate based on aliveness was less than half that obtained from sums of Pk's and made a difference whether or not an important criterion was met. Results from aliveness analysis can be substantially different from analyses based either on simulated kills from RTCA or on unmodified sums of Pk's. Since there can be a real difference between results of the techniques unless RTCA works extremely well, a preferred technique should be chosen. Both simulated kills from RTCA and unmodified sums of Pk's give wrong attrition estimates when PKA's differ from PKU's. Thus aliveness analysis should be the method of choice.

REFERENCES

Adam, John A. (1987). "The Sergeant York gun: a massive misfire," IEEE Spectrum, Volume 24 Number 2 (February, 1987), 28-35. [This article presents a balanced and quite accurate account of the SGT York program.]

Bryson, Marion R. (1984). "Analysis of Opportunities to Engage," Systems Analysis and Modeling in Defense: Developments, Trends and Issues, ed. Reiner K. Huber, New York and London: Plenum Press, 479-474. [This paper, presented at a NATO symposium in 1982, contains an early version of aliveness ideas. Although each author of the current paper has presented later versions of aliveness ideas at other conferences, none have been published.]

Maximum-Likelihood Estimation of the Parameters of a Four-Parameter Class of Probability Distributions

Siegfried H. Lehnigk

U.S. Army Missile Command
AMSMI-RD-RE-OP
Redstone Arsenal, AL 35898-5248

1. Introduction

We shall be concerned with the problem of parameter estimation by means of the likelihood function for a class of four-parameter distributions (hyper-Gamma class) characterized by the probability density function (pdf) class

$$f(x,P) = \begin{cases} \dfrac{\beta}{b\Gamma((1-p)\beta^{-1})}\,\xi^{-p}\exp-\xi^{\beta}\,,\ \xi=(x-s)b^{-1}\,,\ x>s\,, \\ 0\,,\ x<s. \end{cases} \tag{1.1}$$

The parameter vector $P = (s,b,p,\beta)$ has the components s = shift (location), b = scale, p = initial shape, β = terminal shape, with $b > 0$, $p < 1$, $\beta > 0$. In practice we are given a set of absolute frequency data $(x_\nu, f_{a\nu})$ $(\nu = 1,\ldots,m)$, $f_{a1} > 0$, $f_{a\nu} \geq 0$ $(\nu = 2,\ldots, m-1)$, $f_{am} > 0$, $\sum_{\nu=1}^{m} f_{a\nu} = N$ = total number of observations. The shift parameter s ,

therefore, is restricted to $s < x_1$.

The distribution class defined by (1.1) is of wide applicability. As special cases it contains a number of distributions well-known in statistics and statistical physics: Gauss ($p=0$, $\beta=2$), Weibull ($p=1-\beta<1$), exponential ($p=1-\beta=0$), Rayleigh ($p=1-\beta=-1$), Gamma ($p<1$, $\beta=1$), chi-square ($p=(2-\nu)/2$, $\beta=1$), Maxwell ($p=-2$, $\beta=2$), Wien ($p=-3$, $\beta=1$).

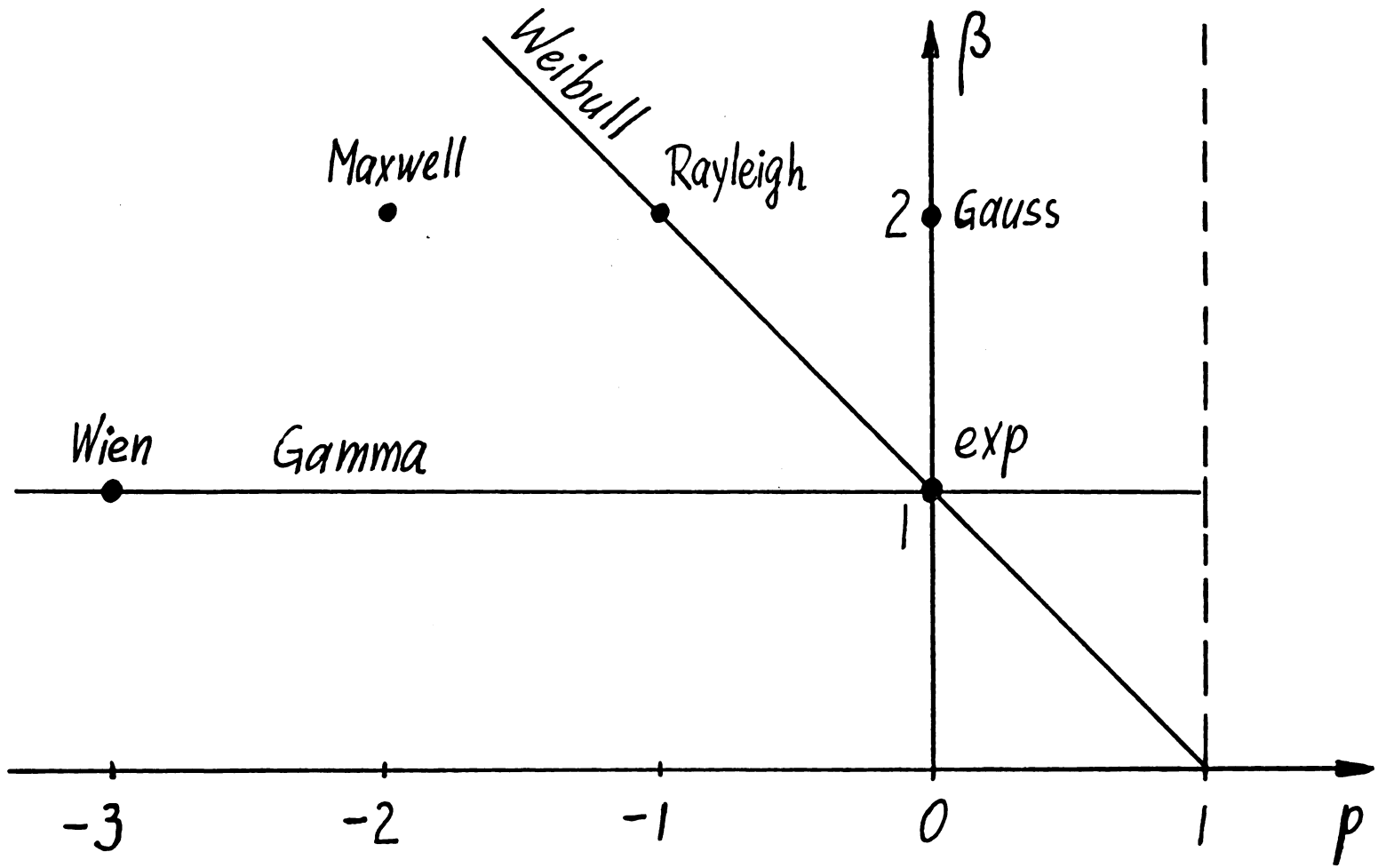
Relative to the most essential parameters p and β, the class (1.1) covers the open quadrant $p < 1$, $\beta > 0$ of the (p,β)-plane. The locations of the special cases just mentioned are shown in the accompanying figure.

It is our objective to show that the class (1.1) is easy to apply in practice. In other words, we shall show that, for a given set of frequency data $(x_\nu, f_{a\nu})$ $(\nu = 1,...,m)$, the four parameters s, b, p, and β can easily be estimated. The likelihood function approach will be used.

2. The Likelihood Function

Let $(x_\nu, f_{a\nu})$ $(\nu=1,...,m)$, $\sum_{\nu=1}^{m} f_{a\nu} = N$, be a set of given frequency data.

We set $\log (x_\nu - s) = \rho_\nu$. Then the likelihood function $L(P)$ for the pdf class (1.1) takes the form



$$L(P) = \beta^N \Gamma^{-N} ((1-p)\beta^{-1})\, b^{-(1-p)N} \left(\prod_{\nu=1}^{m} e^{f_{a\nu} \rho_\nu} \right)^{-p} \exp - \left(b^{-\beta} \sum_{\nu=1}^{m} f_{a\nu} e^{\beta \rho_\nu} \right).$$

Introducing relative frequencies $f_\nu = N^{-1} f_{a\nu}$, we obtain for $R(P) = \log L(P)$ the function

$$R(P) = \log \beta - \log \Gamma ((1-p)\beta^{-1}) - (1-p) \log b - pC - b^{-\beta} B \qquad (2.1)$$

in which

$$B = B(s,\beta) = \sum_{\nu=1}^{m} f_\nu e^{\beta \rho_\nu} > 0 \ , \ C = C(s) = \sum_{\nu=1}^{m} f_\nu \rho_\nu \ .$$

The objective is to maximize the function $R(P)$ given in (2.1) under the constraints $s < x_1$, $b > 0$, $p < 1$, $\beta > 0$.

The partial derivatives of R with respect to s, b, p, and β (in this order) lead to the equations

$$pE + \beta b^{-\beta} F = 0 \ , \qquad (2.2)$$

$$-(1-p)b^{-1} + \beta b^{-\beta-1} B = 0 \ , \qquad (2.3)$$

$$\beta^{-1} \psi ((1-p)\beta^{-1}) + \log b - C = 0 \ , \qquad (2.4)$$

$$\beta^{-1} + (1-p)\beta^{-2} \psi((1-p)\beta^{-1}) + b^{-\beta} B \log b - b^{-\beta} D = 0 \ , \qquad (2.5)$$

3. The Likelihood Equations

The equations (2.2), ... , (2.5) are the four likelihood equations in the four unknowns s, b, p, and β, relative to the logarithmic likelihood function (2.1).

We make the following essential observations.

First, since $E > 0$, $\beta > 0$, $b > 0$, $F > 0$, equation (2.2) shows that, if the shift parameter s is considered as unknown, the initial shape parameter p must be less than zero.

Secondly, equation (2.3) shows that the scale parameter b can be expressed in terms of the parameters s, b, and β ,

$$b^{\beta} = (1-p)^{-1}\beta B . \tag{3.1}$$

Thirdly, using this expression for b^{β} in (2.4) and (2.5), we see that the initial shape parameter p can be eliminated since it can be expressed in terms of s and β ,

$$(1-p)^{-1}\beta = AB^{-1} \quad , \quad p = 1 - \beta A^{-1} B \ , \tag{3.2}$$

where

$$A = A(s,\beta) = \beta(D-BC).$$

Consequently, out of the set of the four equations (2.2, ... , (2.5), we need retain only two, namely (2.2) and (2.4). Eliminating from these b and p by means of (3.1) and (3.2) we arrive at the two equations

$$g(s,\beta) = \psi(A^{-1}B) + \log A - \beta C = 0 , \qquad (3.3)$$

$$h(s,\beta) = (\beta^{-1}A-B)E + F = 0. \qquad (3.4)$$

The solution $(\hat{s},\hat{\beta})$ of these equations and the corresponding numbers

$$\hat{p} = 1 - \hat{\beta}\, A^{-1}(\hat{s},\hat{\beta})\, B(\hat{s},\hat{\beta}) ,$$

$$\hat{b} = \exp\left\{\hat{\beta}^{-1} \log\left[\hat{\beta} B(\hat{s},\hat{\beta})\,(1-\hat{p})^{-1}\right]\right\}$$

obtained from the auxiliary formulas (3.2) and (3.1) give us the desired estimates for the four parameters relative to a given set of frequency data (x_ν, f_ν).

For the numerical solution of the system of equations (3.3) and (3.4), it is convenient to introduce the functions $\tilde{A}$, $\tilde{B}$, and $\tilde{F}$, defined by

$$A = \beta\, e^{\beta\rho_m}\, \tilde{A} \quad , \quad B = e^{\beta\rho_m}\, \tilde{B} \quad , \quad F = e^{\beta\rho_m}\, \tilde{F} \quad .$$

Examples for the solution of equations (3.3) and (3.4) will be given in the paper by Mr. H. P. Dudel.

ON ROTATION IN FACTOR ANALYSIS OF ATMOSPHERIC PARAMETERS

Oskar M. Essenwanger
Aerophysics Branch
Research Directorate
Research, Development, and Engineering Center
U.S. Army Missile Command
Redstone Arsenal, Alabama 35898-5248

ABSTRACT. Many authors of texts and articles on factor analysis recommend rotation of factors after original solutions of unrotated factors. This goal is readily achieved today with the aid of "canned programs" on most of the bigger computer systems. It was noticed, however, that for a system of factor analysis of atmospheric parameters orthogonal rotation was significantly different for only some methods of estimating communalities. Furthermore, oblique rotation differed little from orthogonal rotation.

It will be shown that in cases where the alignment of data in a (rectangular) system are close to the abscissa and ordinate of the system rotation does not contribute much to further alignment. Whenever the original factors display scatter in the diagrams the dispersion is already reduced by an orthogonal rotation. Thus oblique rotation would not bring much improvement.

It will be discussed that the "simplification" of factors by rotation will aid in the diagnosis of the system but does not improve the task of prediction from the system.

1. INTRODUCTION. With the availability of "canned programs" for factor analysis the mathematical difficulties have largely been resolved although one should carefully consider the mathematical background on which these "canned programs" are based. After calculation of the unrotated factors many authors (e.g., Cattel 1952, 1965) recommend simplification by rotation of the systems. The concept of rotation is not supported by some authors dealing with meteorological data. In fact, Buell (1971) finds it completely unnecessary.

In a previous study this author (1986a) deduced that rotation resulted in an alignment of factors although the factors were obtained by different methods of estimating the "communalities". Thus rotation in factor analysis of climatological data seems to serve a useful purpose, reducing the individuality and subjectivity in the decision of estimating the communalities. It was discovered, however, that little difference

between orthogonal and oblique rotation of climatological factor data showed up. Thus the author decided to study rotation in factor analysis of climatological data in more detail.

The analysis disclosed that whenever the original factor analysis displayed little scatter of the factor components plotted into a rectangular coordinate system the rotation did not render significantly different results. Factor components with larger dispersion, however, provide better alignment of data along the axes and provide less scattering after rotation. In the given examples from climatological data of Stuttgart, Germany orthogonal rotation diminished the scatter already to a point where oblique rotation could not contribute to a further reduction.

It can be shown that rotation may simplify the factors and decrease the scatter but does not contribute to improve the ability of prediction utilizing the factor analysis. The prediction error remains the same, whether rotated or unrotated.

2. FACTOR MODEL AND ESTIMATION. The factor model is based on:

$$M_X = M_A M_F + M_\epsilon \qquad (1)$$

where M_X is a data matrix (symmetric), M_A a coefficient matrix and M_F a factor matrix. In the principal components analysis with the number of factors corresponding to the dimension of the data matrix M_ϵ is an error matrix. M_A is also called the factor loading matrix or factor pattern. For diagnostic purposes M_F is not calculated in most cases.

The mathematical solution of eqn. (1) can be formulated:

$$M_X = M_A \phi M^T_A + (\Psi) \qquad (2)$$

which is an eigenvector problem. ϕ is a factor covariance matrix, $\phi = M^T_F M_F$. Ψ is a diagonal matrix if the errors and the factors are uncorrelated. This is generally assumed. In its standard form M_X is a correlation matrix with unity in the diagonal (communalities). In this form the factors are called principal components. In the true factor analysis the assumption is made that not all factors are known. Thus the diagonal element is < 1.0. Several substitutions have been suggested

for the communalities (e.g. Guttman, 1956, or see Essenwanger, 1976, p. 281). In recent times several estimation methods for the communalities as described by Joreskog (1967) have been developed based on statistical principles.

The unweighted least squares (ULSQ) method requires that U is a minimum for:

$$U = (1/2)\ tr\ (M_S - M_X)^2 \qquad (3)$$

where M_S is the correlation matrix with estimates in the diagonal and tr means the trace.

The generalized least squares (GLSQ) method minimizes G for:

$$G = (1/2)\ tr\ (I_n - M_S^{-1} M_X)^2 \qquad (4)$$

and the maximum likelihood (MXLI) method M for:

$$M = tr\ [(M_X^{-1} M_S)] - \ln|(M_X^{-1} M_S)| - n \qquad (5)$$

In addition to the three methods described above a truncation in the number of factors obtained from the principal components analysis could also be used (see Essenwanger, 1986a, b). Since this method maximizes the representation of the variance, the same number of factors as in the other three methods will have a higher percentage of representation of the variance.

Rotation serves to simplify the factors (see Essenwanger, 1976 p. 285, or 1986a). Rotation can be accomplished by a transformation matrix T such as:

$$M_{F_o} = M_A T_1 \qquad (6)$$

where M_{F_o} is the rotated matrix by orthogonal rotation. In various cases simplification is not sufficiently achieved by an orthogonal rotation and an oblique rotation is appropriate. In this case two matrices must be obtained:

$$M_{F_S} = M_A T_2 \qquad (7a)$$

$$M_{F_P} = M_A T_2^{-1} \qquad (7b)$$

where M_{F_S} is named factor structure and M_{F_P} factor pattern matrix. The structure matrix M_{F_S} represents the covariances (correlations) between factors and variables and M_{F_P} can be interpreted as regression coefficients. In the orthogonal rotation the factors remain uncorrelated while in the oblique case the factors are correlated.

3. EXAMPLE OF ORTHOGONAL ROTATION. A simple example is illustrated for an orthogonal rotation of factors. A principal components analysis was performed for the observed data of Frankfurt during January 1946-1956. The correlation matrix between four elements (visibility, i.e. logarithm of visibility, temperature, and windspeed) is given in Table 1. Table 2 exhibits the four factors from this principal components analysis. The (orthogonally) rotated factors are shown in Table 3, Figure 1 depicts one particular phase of this rotation between (modified) factor 1 and 4. In this case the rotation angle calculated by the VARIMAX method (see Kaiser, 1958, or Cattell and Khanna, 1977) was -39^0. The figure illustrates that the four points are much closer to the axes after rotation of the coordinate system. Since the rotation is orthogonal the other factors are not affected.

4. EXAMPLES OF ROTATION AND COMPARISON OF ESTIMATION METHODS. While comparing estimation methods for communalities in factor analysis (Essenwanger, 1986 a, b,) it was noticed that oblique rotation and orthogonal rotation did not differ much for the Stuttgart, Germany climatological data samples. A typical example is exhibited here in Tables 4 and 5. For better readability values ≤ 0.4 were omitted. For the orthogonal rotation the factor loads and for the oblique rotation the structure matrix is shown. It is apparent that orthogonal and oblique rotation differ very little. The rotation procedure, demonstrated in section three, is depicted in Figures 2 and 3. They provide an example from a truncated principal components analysis for the January 1946-1953 data at Stuttgart, Germany. Nine climatological elements (ceiling, cloud amount, visibility, i.e, its logarithm, wind direction and speed, temperature, dewpoint, relative humidity, and pressure) were chosen for the factor analysis. Four factors were retained.

Figure 3 illustrates the reduction of the scatter by orthogonal rotation. In the graph the components of one factor were used as the abscissa and the components of the second as the ordinate. We may assume that further rotation for simplifications is unnecessary if the components pair for the two factors fall within a distance of $\pm$ 0.2 from the axes. For the four factors (54 points for 9 elements) 20 points remain outside this band in the unrotated factors case (Figure 2). After orthogonal rotation only four data points remain outside the band (Figure 3). This leaves little room for improvement by an oblique rotation.

The problem of rotation was further analysed for 12 factor analysis results although only for one station: Stuttgart, Germany. Table 6 discloses the count of data points outside the postulated acceptance band $\pm$ 0.2 around the axes. The left hand part shows the counts for the unrotated and the right hand part the counts after an orthogonal transformation of axes was performed.

Although this count should not be used as the only source for evaluation and interpretation of the merits of an oblique rotation it discloses some interesting facts, however. Apparently the principal components analysis and the unweighted least squares estimations show the greatest dispersion of the 54 component points for the unrotated factors. After rotation only a few points remain outside the "desirable" bounds. A closer scrutiny reveals that oblique rotation may provide an improvement through the alteration of axes only in a few cases (e.g. winter, 12^h) where the orthogonal rotation left between 10 to 14 points outside the $\pm$ 0.2 band. Thus the improvement in simplification may be judged by a count of the number of points falling outside the $\pm$ 0.2 tolerance band. In addition, the distance from the origin (magnitude of vectors) can be included into the judgement criteria.

The closeness between orthogonal and oblique rotation can also be judged by a comparison of the transition matrices T_1 and T_2. For the data of Table 4 and the principal components analysis those two matrices are given in Table 7. A close inspection reveals that the corresponding numerical values differ very little between T_1 and T_2. In addition to the transition matrices the correlation between factors can be examined. In the orthogonal case the factors are uncorrelated and the numerical value is zero. The factor correlation matrix in Table 7 displays that the correlation coefficient, although not precisely zero, is extremely low. In fact, the deviation from zero do not hold up under the scrutiny of a statistical significance test at the 95% level of significance. In this case oblique rotation would not be necessary.

5. PREDICTION FROM FACTOR ANALYSIS: It was illustrated in the previous sections that for climatological data oblique rotation would not add to simplification and diagnosis in factor analysis beyond the achievements by orthogonal rotation. One question remained: Would oblique rotation improve the prediction based on factor analysis? From a theoretical point of view rotation would neither improve nor diminish the results for prediction. This expectation is confirmed by the data presented in Table 8 as follows.

A factor analysis was performed as a pilot for a sample of 15 observation data sets randomly chosen from the winter months December 1946 - February 1948 at Stuttgart. Although the sample is small it reveals the essential facts. The nine elements were chosen as previously used. For every element the prediction was based on four factors whose components were calculated for the 15 observations. Then the differences $\epsilon^2 = \sum (x-x_p)^2/N$ were calculated for every element (x = observed, x_p = predicted). The result for the unrotated case and the oblique rotation is found in Table 8. As expected the two error columns ϵ^2 are identical except for one difference by rounding. This result implies that the goodness of fit for prediction depends only on the number of factors and is independent of rotation. In a previous article (Essenwanger, 1986b) it was pointed out, however, that the (truncated) principal components analysis provided the highest percentage approximation of the total variance. Thus the quality of prediction would depend on the estimation method for the communalities. In this article it was also demonstrated that the dissimilarity of the factors obtained by differences in estimating the communalities virtually vanishes for climatological data after othogonal (and oblique) rotation.

Thus the simplification achieved by rotation leads to the same "climatological factors."

6. CONCLUSION AND SUMMARY: Rotation in factor analysis was studied in detail for climatological data samples. Although the study was limited to one station (Stuttgart, Germany) the result may indicate that orthogonal rotation of factors for climatological data may be sufficient to achieve simplification. Unless simplification is desirable in cases where the factor analysis is utilized as a prediction tool the percentage approximation of the total variance is not improved by rotation. However, rotation serves in aligning the original dissimilar factors to a uniform system of factors in terms of climatology although the individual estimators for the communalities differ.

REFERENCES

Buell, C. E., 1971. Integral Equations Representation for Factor Analysis. J. Atmosph. Sci., 28, 1502-1505.

Cattell, R. B., 1952. Factor Analysis. Harper, New York, pp. 462.

Cattell, R. B., 1965. The Configurative Method for Surer Identification of Personality Dimensions, Notably in Child Study. Psych. Rep. 16, 269-270.

Cattell, R. B. and D. K. Khanna, 1977. Principles and Procedures for Unique Rotation in Factor Analysis. p. 166-202 in "Statistical Methods for Digital Computers," Vol III, edited by Enslein, K. et al.

Essenwanger, O. M., 1976. Applied Statistics in Atmospheric Science Elsevier, Amsterdam, pp. 412.

Essenwanger, O. M., 1986a. A Comparison of Methods for Factor Analysis of Visibility. ARO Report 86-2, p. 39-59. Proceedings of the Thirty-First Conference on the Design of Experiments in Army Research, Development and Testing.

Essenwanger, O. M., 1986b. Comparison of Principal Components and Factor Analysis Methods for Climatological Parameters. Proceedings of the 3rd International Conference on Statistical Climatology, Vienna, 23-27 June 1986.

Guttman, L., 1956 "Best Possible" Systematic Estimates of Communalities. Psychrometrika, 21, 273-285.

Jöreskog, K. G. (1967). Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika, 32, 443-482.

Kaiser, H. F., 1958. The VARIMAX Criterion for Analytical Rotation in Factor Analysis. Psychometrika, 23, p. 187-260.

ACKNOWLEDGEMENT: The author's thanks go to Dr. Dorathy A. Stewart for her critical review of the manuscript. Mrs. Alexa Mims and Mr. Roger Betts deserve the credit for the preparation of the computer programs to obtain the data. Last, not least, Mrs. Gloria McCrary must be thanked for her patience and diligence during the process of establishing the manuscript from the first draft to the final version.

Table 1. Correlation Matrix for Four Meteorological Elements, Frankfurt, Germany, 1946-1956 (Matrix M_X)

	ln(V)	P	T	W
ln(V)	1.0	0.08	0.08	0.42
P	0.08	1.0	0.05	-0.16
T	0.08	0.05	1.0	0.20
W	0.42	-0.16	0.20	1.0

V = Visibility, P = pressure, T = temperature, W = Windspeed

Table 2. Factor Matrix M_A for Correlation Matrix of Table 1. (Principal Components Factors)

	0.76	0.18	-0.45	-0.42
	-0.10	0.95	-0.19	0.24
	0.44	0.31	0.83	-0.15
	0.85	-0.21	-0.04	0.49
λ	1.50	1.07	0.93	0.50 (Eigenvalue)

Table 3. Rotated Factor Matrix (Orthogonal Rotation)

	.96	.22	-.07	.17
	-.12	.98	.06	-.05
	.14	.02	.99	.08
	.36	-.07	.08	.95
λ_1	1.04	1.03	.99	.94

Table 4. ORTHOGONAL (ORT) AND OBLIQUE (OB) ROTATION (STUTTGART, JANUARY 1947-1953) UNIT 0.01

PRINCIPAL COMPONENTS

		ORT	OB	ORT	OB	ORT	OB	ORT	OB
1	CEIL			91	83				
2	CL.AMT			94	95				
3	Ln VIS	79	81						
4	WD	62	66			41	44		
5	WS	73	76						
6	TEMP					89	91		
7	DEWP					97	99		
8	REHU	73	69						
9	PRES							97	98
	VAR	218	223	191	190	217	214	100	100
	$\sum$							726	727

UNWEIGHTED LEAST SQUARES

		ORT	OB	ORT	OB	ORT	OB	ORT	OB
1	CEIL			85	87				
2	CL.AMT			91	92				
3	Ln VIS	76	78						
4	WD	52	55						
5	WS	62	64						
6	TEMP					86	88		
7	DEWP					100	100		
8	REHU	58	56						
9	PRES							99	98
PROD. VAR		171	173	172	171	204	202	99	100
	$\sum$							645	647

Table 5. ORTHOGONAL AND OBLIQUE ROTATION (STUTTGART, JULY 1947-1953) UNIT 0.01

PRINCIPAL COMPONENTS

		ORT	OB	ORT	OB	ORT	OB	ORT	OB
1	CEIL			80	82				
2	CL.AMT			77	80				
3	Ln VIS	69	69						
4	WD			63	61				
5	WS		41	68	64				
6	TEMP	69	68			68	75		
7	DEWP					93	92		
8	REHU	87	86						
9	PRES							97	97
	VAR	208	202	213	215	143	147	101	102
	$\sum$							665	666

UNWEIGHTED LEAST SQUARES

		ORT	OB	ORT	OB	ORT	OB	ORT	OB
1	CEIL			74	77				
2	CL.AMT			91	94				
3	Ln VIS	40	41						
4	WD							70	71
5	WS							67	69
6	TEMP	80	83			57	63		
7	DEWP					99	99		
8	REHU	98	98						
9	PRES							11	11
	PROD. VAR	187	185	145	147	136	137	107	105
	$\sum$							575	574

Table 6. Number of Data Points Outside Tolerance Band (Stuttgart, 1946-1953)

	UNROTATED				ORTH. ROTATION			
SAMPLE DATA	PC	ULSQ	GLSQ	ML	PC	ULSQ	GLSQ	ML
JANUARY	20	25	27	12	4	7	6	8
APRIL	27	34	14	2	8	5	5	1
JULY	21	25	6	7	5	2	3	3
OCTOBER	21	21	6	9	2	1	6	2
SUMMER 00^h	13	26	4	8	9	3	3	2
SUMMER 06^h	21	20	5	6	7	3	3	2
SUMMER 02^h	13	17	5	12	3	8	4	10
SUMMER 18^h	16	23	13	6	5	6	13	6
WINTER 00^h	23	24	24	14	9	6	3	6
WINTER 06^h	22	17	22	13	8	9	6	8
WINTER 12^h	24	22	28	10	13	11	14	10
WINTER 18^h	22	25	14	10	11	5	5	6
$\sum$	243	279	168	109	84	66	71	64
Mean	20.2	23.2	16.0	9.1	7.0	5.5	5.9	5.3

PC = Principal Components Analysis, ULSQ=Unweighted Least Squares Estimators, GLSQ = Generalized Least Squares Estimators, ML = Maximum Likelihood Estimators

Table 7. STUTTGART, JULY (1946-1953)
TRANSFORMATION MATRICES
(PRINC. COMP. METHOD)

ORTHOGONAL				OBLIQUE			
T_1				T_2			
0.67	-0.68	-0.29	0.05	0.65	-0.74	-0.34	0.08
0.66	0.71	-0.17	-0.18	0.68	0.65	-0.19	-0.21
0.34	-0.06	0.94	0.10	0.34	-0.10	0.89	0.11
0.05	0.17	-0.11	0.98	0.04	0.15	-0.10	0.97

FACTOR CORRELATION

1.00	-0.07	-.09	-.01
-0.07	1.00	.06	-.06
-0.09	0.06	1.00	.01
-0.01	-0.06	.01	1.0

Table 8. FACTORS AS PREDICTORS

			UNROTATED		OBLIQUE ROT.
	Mean	σ	ϵ^2	$\sqrt{\epsilon^2}$	ϵ^2
CEILING	$9.2 \bullet 10^3$ ft	$8.2 \bullet 10^3$	5.38	2.31	5.38
CLOUD AMT	0.625	.44	0.009	0.10	.009
Ln VISIBIL.	0.8	0.95	.085	0.29	.086
WIND DIR	276^o	90^o	24.33	4.93	24.33
WIND SP	4.0 kt	3.1 kt	2.22	1.50	2.22
TEMP	24.0^oF	0.6^oF	6.29	2.51	6.29
DEWP	21.0^oF	9.5^oF	2.01	1.41	2.01
REL. HUM.	89.0%	10.3%	30.48	5.52	30.48
PRESSURE	1021.0 mb	6.1 mb	1.15	1.07	1.15
			50.64		50.64

$\epsilon^2 = \Sigma(x-x_p)^2/N$

UNITS IN LAST TWO COLUMNS ARE THE SAME AS IN THE FIRST COLUMN.

FIG 1. EXAMPLE OF ROTATION (ANGLE - 39°)
(FRANKFURT, GERMANY, 1946-1956)

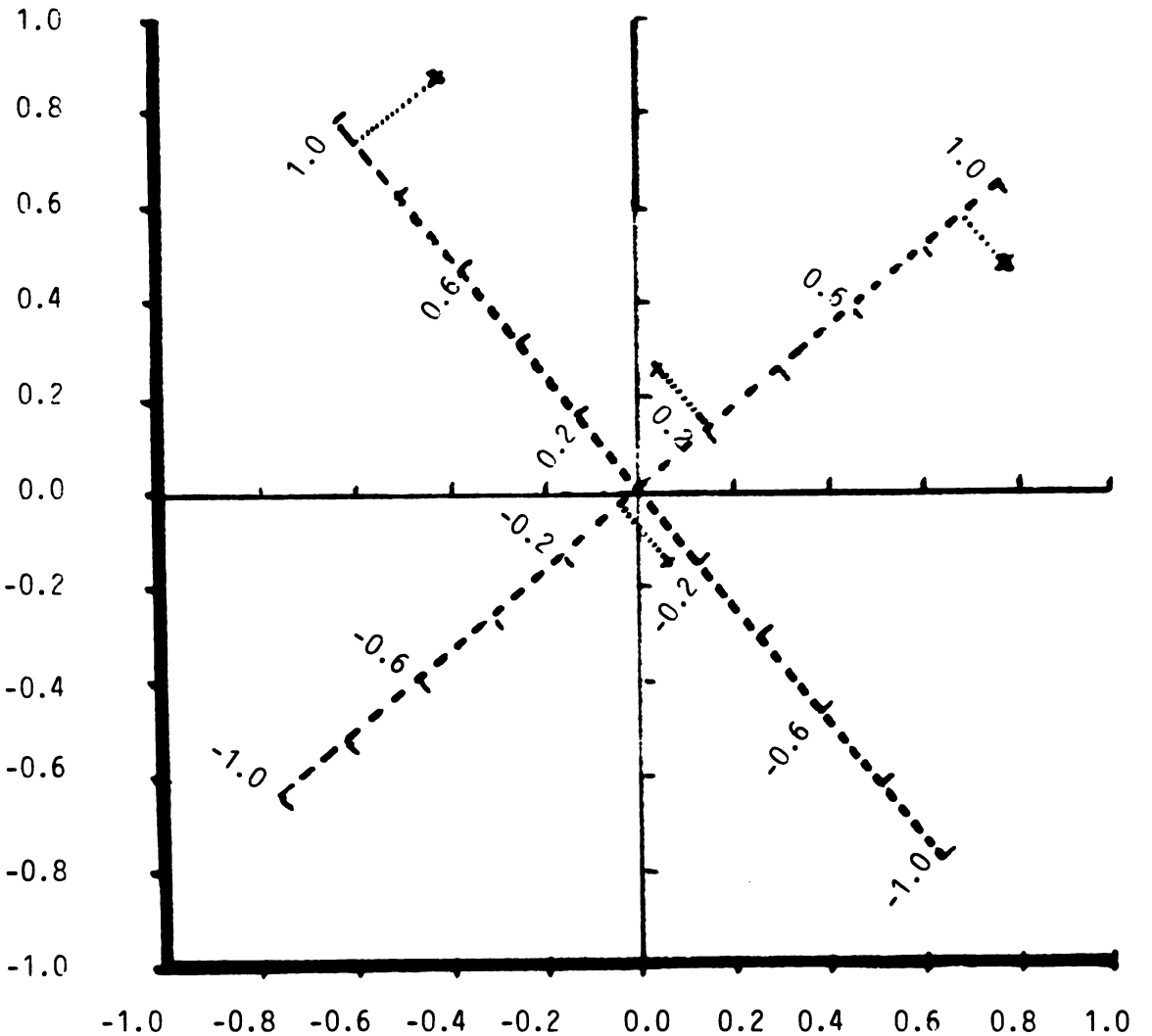


FIG 2. STUTTGART, JANUARY (1946-1953)
PRINCIPAL COMP METHOD
UNROTATED FACTOR LOADS

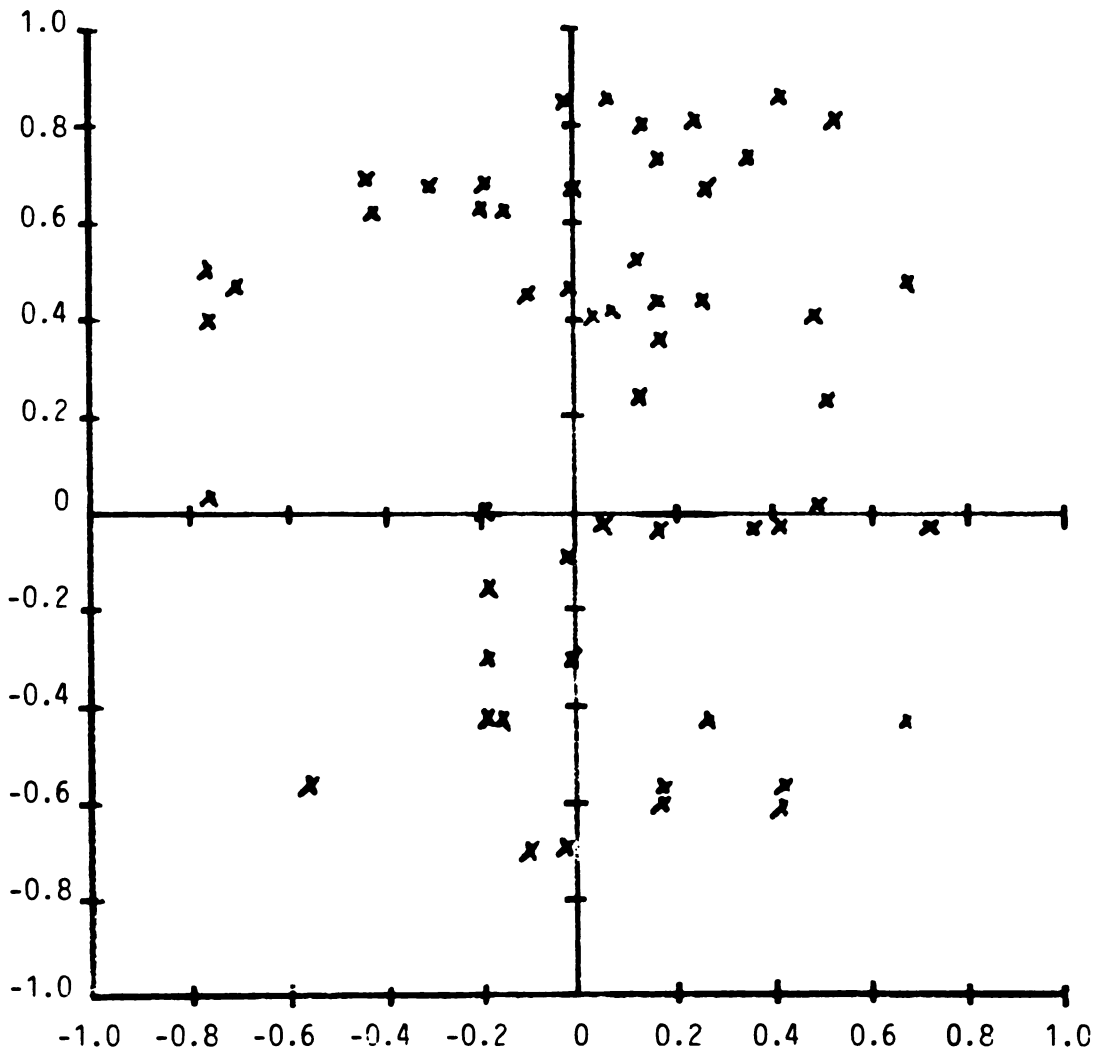
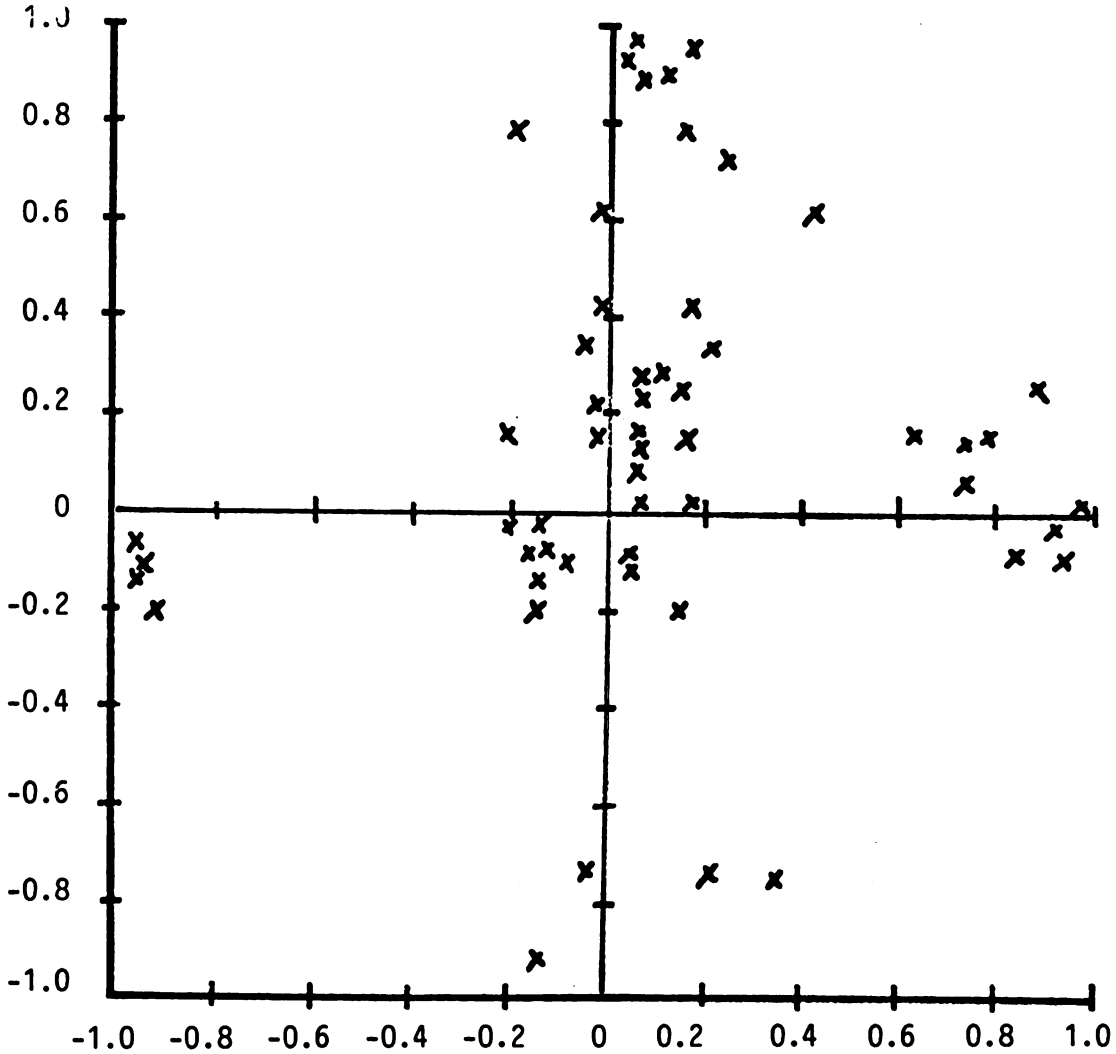


FIG 3. STUTTGART, JANUARY (1946-1953)
PRINCIPAL COMP METHOD
ORTHOGONAL ROTATION
FACTOR LOADS



An Exact Method for One-Sided Tolerance Limits in the Presence of Batch-to-Batch Variation

Mark Vangel

U. S. Army Materials Technology Laboratory
Watertown, Massachusetts 02172-0001

Abstract

Mee and Owen (1983) proposed an improvement on a method of Lemon (1977) for estimating tolerance limits from a balanced one-way ANOVA random effects model. This method uses an approximation of Satterthwaite (1946) to replace a linear combination of two chi-square random variables with a random variable having a chi-square distribution. The tolerance factor is then estimated as a quantile of a noncentral t-distribution. The Mee-Owen procedure is conservative for all values of the population variance ratio.

An alternative approach is to view the tolerance limit problem as a variant of the Behrens-Fisher problem. The work of Welch (1947) and Trickett and Welch (1954) may then be applied to derive an integral equation the solution of which, a function of the ratio of the between batch to the within batch mean squares, provides an exact solution to the problem.

An algorithm is presented for iteratively approximating this function. Neither the existence of a solution nor the convergence of this algorithm are discussed; but numerical evidence is presented which suggests that the proposed solution is, for the purposes of applied statistics, exact for all values of the ratio of between batch to within batch population variances.

Two other topics considered in this paper are an approximation to the tolerance limit based on the Welch-Aspin series solution to the Behrens-Fisher problem and a discussion of the effect pooling and using a single sample procedure has on the coverage probability of the tolerance limit.

An application to determining (.90, .95) lower tolerance limits for composite material strength data in the presence of batch-to-batch variation is discussed. This tolerance limit is referred to as the 'B-basis material property' by aircraft designers and is used to determine the acceptability of a composite material for aircraft applications.

1. Introduction

If a material is manufactured in many large batches and the population of interest consists of all batches, the random effects model may be an appropriate model for measurements made on characteristics of the material.

Let X_{ij} denote the jth of J observations from the ith of I batches. If X_{ij} follows a one-way balanced random-effects model, then

$$X_{ij} = \mu + b_i + e_{ij} \tag{1.1}$$

where μ denotes the population mean, $\mu + b_i$ denotes the mean of the ith batch, and e_{ij} is the error term. The b_i's and the e_{ij}'s are assumed to be independently distributed normal with mean zero and variance σ_b^2 and σ_e^2 respectively. An observation X from this population is thus normally distributed with mean μ and variance

$$\sigma_X^2 = \sigma_b^2 + \sigma_e^2 \tag{1.2}$$

This paper presents techniques for determining one-sided tolerance limits for X based on a random sample of J items from each of I batches. A (β, γ) lower tolerance limit is a random variable T such that a proportion β of the population is covered by the interval $(-\infty, T)$ with probability γ. The methods developed here for lower tolerance limits may be adapted in an obvious way to upper limits.

An important industrial application of tolerance limits is to the characterization and certification of structural materials for aircraft. In order to determine the acceptability of material for aircraft applications, designers use 'material basis properties' which are tolerance limits on the strength of a material as determined from experimental failure data. A (.90, .95) lower tolerance limit is called a 'B-basis' value or 'B-alowable'. The more stringent (.99, .95) limit is referred to as an 'A-basis' value or 'A-allowable'.

There is increasing interest in the use of composite materials as lightweight alternatives to metals for aircraft applications. Composite material properties typically exhibit far more batch-to-batch variability than do

metals; consequently there is a growing need for methods for determining one sided tolerance limits in the presence of batch-to-batch variation.

A modification of a procedure of Lemon (1977) and Mee and Owen (1983) has therefore been adopted for this application by Neal et. al. (1987) and will be included in Mil-handbook-17 (1987), a handbook for the use of composites in aerospace applications. It is hoped that the virtually exact method to be discussed in Section 6 will eventually supersede the Mee-Owen procedure for this application.

2. The Mee-Owen Procedure

let $n = IJ$ denote the sample size. The parameters μ, σ_e^2 and σ_b^2 of the random effects model may be estimated by the pooled mean $\hat{\mu}$, the within batch mean square MS_e and a linear combination of MS_e with the between batch mean square MS_b where:

$$(2.1) \qquad \hat{\mu} = \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}/n,$$

$$(2.2) \qquad MS_b = J \sum_{i=1}^{I} (\hat{\mu} - \bar{X}_i)^2, \qquad \bar{X}_i = \sum_{j=1}^{J} X_{ij}/J,$$

$$(2.3) \qquad MS_e = \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \hat{\mu})^2/n .$$

An unbiased estimator of the population variance σ_X^2 is

$$(2.4) \qquad \hat{\sigma}_X^2 = MS_b/J + (1 - 1/J)\, MS_e .$$

For $0 < \beta < 1$, let K_β be the β quantile of the standard normal distribution, i.e

$$(2.5) \qquad \beta = 1/\sqrt{2\pi} \int_0^{K_\beta} e^{-t^2/2} dt.$$

A (β, γ) lower tolerance limit is a $100\gamma\%$ lower confidence bound on

$$(2.6) \qquad \mu - K_\beta \sigma_X .$$

By analogy with the single sample case (see, for example, Owen (1968)), one seeks an estimator of the form

$$(2.7) \qquad \hat{\mu} - k\hat{\sigma}_X$$

where k is chosen to satisfy

$$(2.8) \qquad P(\hat{\mu} - k\hat{\sigma}_X \leq \mu - K_\beta \sigma_X) = \gamma .$$

Since $\hat{\mu}$ is distributed normal with mean μ and variance

$$(2.9) \qquad \hat{\sigma}^2_\mu = (J\sigma_b^2 + \sigma_e^2)/n$$

one may rewrite (2.8) as

$$(2.10) \qquad P((Z + \sqrt{n}K_\beta B)/(\hat{\sigma}_X/\sigma_X) \leq \sqrt{n}kB) = \gamma$$

where

$$(2.11) \qquad Z = (\hat{\mu} - \mu)/\hat{\sigma}_\mu ,$$

$$(2.12) \qquad B = ((JR + 1)/(R + 1))^{\frac{1}{2}} ,$$

and

$$(2.13) \qquad R = \sigma_b^2 / \sigma_e^2 .$$

The random variable $\hat{\sigma}_\mu^2$ is approximately distributed as the ratio of a chi-square to its degrees of freedom, where the degrees of freedom are given by (Satterthwaite 1946) :

$$(2.14) \qquad f = \frac{(R + 1)^2}{\frac{(R + 1/J)}{I - 1} + \frac{(1 - 1/J)}{n}} .$$

If $T_f^{-1}(\gamma, \delta)$ denotes the inverse of the noncentral t-distribution with f degrees of freedom and noncentrality parameter δ, then

$$(2.15) \quad k \simeq T_f^{-1}(\gamma, \sqrt{nB}K_\beta)/(\sqrt{nB})$$

Unfortunately, the tolerance limit factor k depends on the nuisance parameter R. Mee and Owen suggest replacing R with

$$(2.16) \quad R_\eta \equiv ((MS_b/MS_e)F_\eta - 1)/J$$

where F_η is the 100η percentile of an F random variable with degrees of freedom $I(J-1)$ and $I-1$. R_η is a $100\eta\%$ upper confidence bound estimate on R (Searle, 1971, p.414) and the confidence coefficient may be determined by numerical integration so that

$$(2.17) \quad P(\hat{\mu} - k(R_\eta)\hat{\sigma}_X \leq \mu - K_\beta\sigma_X) \geq \gamma$$

for all I, J and R. These values are reproduced from Mee and Owen (1983) for various combinations of β and γ in Table 1.

For the case of $\beta = .90$ and $\gamma = .95$ some of the conservatism inherent in the above values has been removed by allowing η to vary with I and J. The result of this numerical work is presented in Table 2.

3. An Exact Solution for Known R

If R is known, the tolerance limit factor k is the appropriate quantile of the distribution of

$$(3.1) \quad A = \frac{Z + \delta}{(C_1Y_1 + C_2Y_2)^{\frac{1}{2}}}$$

where Z has a standard normal distribution; Y_i is distributed as a chi-square with n_i degrees of freedom for i=1, 2; and C_1, C_2 and δ are constants with C_1 and C_2 positive. Once this distribution has been determined the

tolerance limit may be obtained exactly.

The density of the linear combination $Y \equiv C_1Y_1+C_2Y_2$ is show in Fleiss (1971) to be

$$f_Y(y) = \frac{\Gamma((n_1+ n_2)/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot$$

$$(3.3) \quad \int_0^1 x^{n_1/2-1}(1 - x)^{n_2/2-1}\chi^2_{n_1+n_2}(y/(C_1x +C_2(1 - x)))dx$$

where $\chi^2_f(\cdot)$ is the chi-square density with f degrees of freedom.

By conditioning on the denominator of (3.1) one sees that

$$F(k) \equiv P(A \le k)$$

$$= \frac{\Gamma((n_1+ n_2)/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \int_0^1 x^{n_1/2-1}(1 - x)^{n_2/2-1}$$

$$(3.4) \quad \int_0^\infty \Phi(kt - \delta)f_Y(t^2/(C_1x+C_2(1 - x)))\, 2t/(C_1x+C_2(1 - x))^{\frac{1}{2}}dxdt$$

$$= \frac{\Gamma((n_1+ n_2)/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot$$

$$\int_0^1 x^{n_1/2-1}(1 - x)^{n_2/2-1}T_{n_1+n_2}(k((n_1+ n_2)(C_1x + C_2(1 - x)))^{\frac{1}{2}}, \delta)dx$$

where $\Phi(\cdot)$ is the standard normal distribution and $T(t, \delta)$ denotes the noncentral t cumulative with f degrees of freedom and noncentrality parameter δ, i.e.

$$(3.5) \quad T_f(t, \delta) = \frac{\sqrt{2\pi}}{\Gamma(f/2)\,2^{f/2-1}} \int_0^\infty u^{f-1}\phi(u)\,\Phi(tu/\sqrt{f} -\delta)\,du$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative respectively.

For the tolerance limit problem, let

(3.6)

$$C_1 = I/(I - 1),\ C_2 = 1/(JR + 1),$$

$$\delta = K_\beta(n(R + 1)/(JR + 1))^{\frac{1}{2}}$$

and

(3.7) $n_1 = I - 1,\ n_2 = I(J - 1)$

where I, J, K_β and R are as in Sections 1 and 2. The value k(R) such that $F(k) = \gamma$ then provides an exact solution to the problem.

Although the above derivation is simple, it is apparently not well known. A much more complicated representation of the distribution of the random variable (3.1) is developed in Ray and Pitman (1961).

4. The Effect of Pooling on the Coverage Probability

The tolerance limit procedure discussed in Section 2 is conservative (i.e. provides a coverage probability greater then the nominal value) when the population variance ratio, R, is small. Mee and Owen therefor suggest that data be pooled and a single sample method be applied when the mean square ratio is less than 1. They then proceed to investigate the conditional behavior of their proposed estimator.

Using the distribution developed in Section 3, one may determine the coverage probability for a single sample procedure applied to pooled data as a function of the variance ratio. This result will be used to determine the unconditional coverage probability of the Mee-Owen method in Section 7.

Let Y_1 and Y_2 be as in (3.2) and let n_1 and n_2 denote the between and within batch degrees of freedom respectively (see 3.7). The pooled variance estimate is

$$(4.1) \quad S^2 = 1/(n - 1) \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \hat{\mu})^2$$

where n denotes the pooled sample size, I the number of batches, J the batch size and $\hat{\mu}$ the grand mean. Partitioning the total mean square and substituting (2.9) for the variance of $\hat{\mu}$, one obtains

$$(4.2) \qquad (S/\sigma_{\mu})^2 = \frac{\sigma_e^2 Y_2 + (J\sigma_b^2 + \sigma_e^2)Y_1}{\sigma_b^2 + \sigma_e^2} \cdot \frac{n}{n-1}$$

where R is the variance ration (2.13).

If k_0 denotes the single sample tolerance limit factor (e.g. Owen, 1968, pp. 446-448), then the coverage probability as a function of R is

$$(4.3) \qquad \hat{\gamma}(R) = P(\hat{\mu} - k_0 S \leq \mu - K_{\beta}\sigma_X)$$

$$= P((Z + K_{\beta}\sigma_X/\hat{\sigma}_{\mu}) / (S/\hat{\sigma}_{\mu}) \leq k_0)$$

with notation as in Section 2. Substituting (4.2) into (4.3) and employing the distribution (3.4), one may readily examine $\hat{\gamma}$ (R) numerically. From the typical plot in Figure 1 it is apparent that the coverage probability obtained may be substantially less than the nominal value even for small values of R. Clearly, criteria which result in the decision to pool must be considered carefully if one is to be assured of a reasonable tolerance limit estimate in the presence of batch-to-batch variation. Alternatively, one might seek an estimator which performs well for all R, eliminating the need to pool altogether. This approach will be taken in Section 6.

5. The Solution for Unknown R: Welch-Aspin series

For unknown variance ratio, the tolerance limit problem is very closely related to the Behrens-Fisher problem. Following the work of Welch (1947) and Trickett and Welch (1954), two forms for a solution are obtained.

A series solution is developed first. While computationally simple, the first order approximation presented here is anticonservative and may only be suitable for many batches.

Alternatively, the tolerance limit factor as a function of the mean square ratio may be obtained approximately as the solution to an integral equation. Although this requires the use of a computer, the method which results appears to give the desired coverage probability - even for small sample sizes.

To simplify the notation in what follows, let S_i^2 be the mean squares and σ_i^2 their expected values for i=1, 2, i.e. :

$$(5.1) \qquad \begin{aligned} S_1^2 &= MS_b & \sigma_1^2 &= J\sigma_b^2 + \sigma_e^2 \\ S_2^2 &= MS_e & \sigma_2^2 &= \sigma_e^2 \,. \end{aligned}$$

Given the two mean squares, the tolerance limit factor may be expressed in terms of the standard normal distribution:

$$\begin{aligned} \gamma &= P(\hat{\mu} - k\hat{\sigma}_X \leq \mu - K_\beta \sigma_X) \\ &= E_{S_1^2,\, S_2^2}(\Phi(k\hat{\sigma}_X/(\sigma_1/\sqrt{n}) - \delta)) \qquad (5.2) \\ &= E_{S_1^2,\, S_2^2}(\Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta)). \end{aligned}$$

The problem is to determine a function $h(S_1^2, S_2^2) = k\hat{\sigma}_X$ so that (5.2) is approximately satisfied for all σ_1^2 and σ_2^2. If tolerance limits on the median are desired, then $\delta = 0$ and the results of Welch (1947) and Aspin (1948) may be used directly. If δ is not zero, the idea behind the Welch-Aspin derivation may still be applied, though the algebra is considerably messier.

Following Welch (1947), one begins by expanding the normal cumulative about (σ_1^2, σ_2^2) and recognizing that the expectation is the moment generating function of the product of two independent chi-squares - with differential operators as the independent variables in the generating functions. If one defines

$$(5.3) \qquad \partial_i \equiv \frac{\partial}{\partial S_i^2} \,\Big|\, S_i^2 = \sigma_i^2$$

then

$$(5.4) \qquad \begin{aligned} &\Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta) \\ &\quad = \prod_{i=1}^{2} e^{(S_i^2 - \sigma_i^2)\partial_i}\ \Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta). \end{aligned}$$

Substituting (5.4) into (5.2) gives

$$(5.5)\quad \gamma = \int_0^{\infty}\int_0^{\infty} \prod_{i=1}^{2} e^{(S_i^2 - \sigma_i^2)\partial_i} \chi^2_{n_i}(S_i^2) \cdot \Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta)\, dS_1^2\, dS_2^2$$

where

$$(5.6)\quad \chi^2_{n_i}(S_i^2)\, dS_i^2 \equiv \frac{1}{\Gamma(n_i/2)2^{n_i/2}} (n_i/\sigma_i)^{n_i/2-1} e^{-n_i S_i^2/(2\sigma_i^2)}\, d(n_i S_i^2/\sigma_i)$$

are the densities of the mean squares S_i^2 and the n_i are their respective degrees of freedom.

In terms of the operator

$$(5.7)\quad \Omega \equiv \prod_{i=1}^{2} (1 - 2\sigma_i^2\partial_i/n_i)^{n_i/2} e^{-\sigma_i^2\partial_i} \simeq 1 + \sum_{i=1}^{2} \sigma_i^4\partial_i^2/n_i$$

the tolerance limit problem can be stated as

$$(5.8)\quad \Omega\Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta) = \gamma$$

The next step is to expand the normal cumulative about K_γ (see (2.5)) in a second Taylor series, giving

$$(5.9)\quad \Phi(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta) = e^{(h(S_1^2, S_2^2)/(\sigma_1/\sqrt{n}) - \delta - K_\gamma)D}\, \Phi(v)$$

where

$$(5.10) \qquad D^r\phi(v) \equiv \frac{d^r}{dv^r}\,\phi(v)\,\Big|_{v=K_\gamma}.$$

Express $h(S_1^2, S_2^2)$ as a series in increasing inverse powers of the degrees of freedom

$$(5.11) \qquad h(S_1^2, S_2^2) = h_0(S_1^2, S_2^2) + h_1(S_1^2, S_2^2) + \ldots$$

where $h_j(S_1^2, S_2^2)$ consists of terms of order j in $1/n_i$ for i=1, 2. One can now, in principle, solve successively for the h_j's. If terms of order greater than zero are considered negligible, then

$$(5.12) \qquad h_0(S_1^2, S_2^2)\,/(\sigma_1/\sqrt{n}) - \delta = K_\gamma$$

which leads to a zeroth order approximation to k:

$$(5.13) \qquad k_0 = h_0(S_1^2, S_2^2)/\hat{\sigma}_X$$
$$= K_\beta + K_\gamma/(I(1 + (J-1)S_2^2/S_1^2))^{\frac{1}{2}}.$$

The next term, $h_1(S_1^2, S_2^2)$, can be shown to be the solution to

$$(5.14) \qquad \gamma = (1 + \textstyle\sum_i \sigma_i^4 \partial_i^2/n_i)\, e^{K_\gamma(S_1/\sigma_1 - 1)D}$$
$$\cdot\, e^{K_\beta((\hat{\sigma}_X - \sigma_X)\,/(\sigma_1/\sqrt{n}))D}$$
$$\cdot\, e^{h_1(S_1^2, S_2^2)\,/(\sigma_1/\sqrt{n})D}\,\Phi(v).$$

After some algebra, the first correction to k_0 is seen to be

$$(5.15) \qquad k_1 = \theta/(4\sqrt{I})\,(K_\gamma(K_\gamma^2 + 1)\,/n_1$$
$$+ 2K_\beta K_\gamma^2\sqrt{I}\,/n_1\theta + K_\beta^2 K_\gamma I\,/n_1\theta^2 + K_\beta\sqrt{I}\,/n_1\theta^3$$
$$+ K_\beta^2 K_\gamma\, I(J-1)^2/(n_2 MSR^2)\theta^2 + K_\beta(J-1)^2\sqrt{I}\,/(n_2 MSR^2)\theta^3)$$

where

$$(5.16) \qquad \theta \equiv (1/(1 + (J - 1)/MSR))^{\frac{1}{2}}$$

and

$$(5.17) \qquad MSR \equiv S_1^2 / S_2^2.$$

The coverage probability for the above approximation as a function of the population variance ratio is plotted in Figure 2 for a (.90, .95) tolerance limit and J = 5. Note that for many batches the series solution performs well, though for few batches it is anticonservative.

6. An Alternative Solution for Unknown R

For small samples, the first order approximation developed above may not be adequate. An alternative approach is to view the problem as an integral equation, following Trickett and Welch (1954). If one defines

$$(6.1) \qquad \tau \equiv 1/(JR + 1)$$

then (3.4) may be written as

$$(6.2) \qquad \gamma = \frac{\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \int_0^1 x^{n_1/2-1} (1 - x)^{n_2/2-1} \cdot T_{n_1+n_2}(k(\tau)((n_1 + n_2)(I/(I - 1)x + \tau(1 - x)))^{\frac{1}{2}}, \delta)dx$$

where

$$(6.3) \qquad \delta = \sqrt{n} K_\beta B = K_\beta((I + (J - 1))^{\frac{1}{2}}\tau$$

and B is as defined in (2.12). The parameter τ may be estimated by the reciprocal of the mean square ratio (4.17):

$$(6.4) \qquad u \equiv 1/MSR = \tau F_{n_2, n_1}$$

where Fn_2, n_1 denotes a random variable with an F distribution with n_2 and n_1 degrees of freedom. The tolerance limit problem reduces to determining a function $\tilde{k}(u)$ such that

$$(6.5)\quad \gamma = P((Z + \delta(\tau))\ /(I/(I - 1)Y_1 + \tau Y_2)^{\frac{1}{2}} \leq \tilde{k}(u))$$
$$= P(Z \leq \tilde{k}((n_1/n_2)(Y_2/Y_1))\ (I/(I - 1)Y_1 + \tau Y_2)^{\frac{1}{2}} - \delta(\tau))$$

where Z, Y_1 and Y_2 are as in Section 3. This is equivalent to the inteqral equation

$$(6.6)\quad \gamma = \frac{\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \int_0^1 x^{n_1/2-1} (1 - x)^{n_2/2-1} \cdot$$
$$T_{n_1+n_2}(\tilde{k}(c)((n_1 + n_2)(I/(I - 1)x + \tau(1 - x)))^{\frac{1}{2}}, \delta(\tau))dx$$

where

$$(6.7)\quad c = n_1(1 - x)\ /(n_2 x)\tau.$$

Using the results of Section 5, one may define

$$(6.8)\quad \tilde{k}(\varepsilon) = \tilde{k}_0(c) + \varepsilon\tilde{k}_1(c)$$

where $\tilde{k}_0(c)$ is the first order approximation from the Welch-Aspin procedure and $\tilde{k}_1(c)$ is an unknown function. If an approximation to $\tilde{k}_1(c)$ can be obtained this approximation may lead to an improved $\tilde{k}_0(c)$.

Letting $V(\cdot)$ represent the functional (6.6), if one expands $V(\cdot)$ in a Taylor series about $\varepsilon = 0$ one obtains the first order approximation

$$(6.9)\quad \gamma = V(\tilde{k}_0(c)) + \varepsilon\tilde{k}_1(c)) \simeq V(\tilde{k}_0(c)) + \varepsilon \left.\frac{dV}{d\varepsilon}\right|_{\varepsilon=0}$$

Since ε is arbitrary, it may be taken to equal one. The approximation may then be written as

$$(6.10)\qquad \gamma \simeq V(\tilde{k}_0) + \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \cdot$$

$$\int_0^1 x^{n_1/2-1}(1-x)^{n_2/2-1}\tilde{k}_1(c)((n_1+n_2)(I/(I-1)x+\tau(1-x)))^{\frac{1}{2}} \cdot$$

$$t_{n_1+n_2}(\tilde{k}_1(c)((n_1+n_2)(I/(I-1)x+\tau(1-x)))^{\frac{1}{2}}, \delta)dx$$

where $t_f(\cdot, \delta)$ denotes the noncentral t density. The noncentral t density with f degrees of freedom and noncentrality parameter δ may be calculated by means of the following formula (Odeh and Owen, 1980, p. 272):

$$(6.11)\qquad t_f(x, \delta) = (f/x)\,(T_{f+2}(((f+2)/f)^{\frac{1}{2}}x, \delta) - T_f(x, \delta)).$$

The first term on the right hand side of (6.10), $V(\tilde{k}_0)$, may be evaluated numerically for given τ since $\tilde{k}_0(c)$ is a known function. The second integral is concentrated about $n_1/(n_1+n_2)$. If $\tilde{k}_1(c)$ is evaluated at this value, the remainder of this integral may also be evaluated numerically. Note that

$$(6.12)\qquad \tilde{k}_1(c)\,\Big|_{x = n_1/(n_1+n_2)} = \tilde{k}_1(\tau)$$

so that, with obvious notation for the two integrals to be evaluated numerically,

$$(6.13)\qquad \gamma \simeq V_0 + \tilde{k}_1(\tau)\, V_1$$

i.e.,

$$(6.14)\qquad \tilde{k}_1(\tau) \simeq (\gamma - V_0)\,/V_1.$$

Since $\tilde{k}(c)$ is the same function of c that $\tilde{k}(\tau)$ is of τ one may use a first approximation $\tilde{k}_0(c)$ to get a new approximation $\tilde{k}_1(c)$ by evaluating (6.11) for

a mesh of τ values. This $\tilde{k}_1$ becomes the $\tilde{k}_0$ for the next iteration. Although it is certainly not obvious that such a procedure will converge, or even that a solution exists, it will be shown below that this algorithm appears to provide a solution to the tolerance limit problem that is (for practical purposes) exact.

7. Discussion

The situation of primary interest to the aircraft industry, (.90, .95) lower tolerance limits, is the only case yet examined in detail. Four methods have been presented in this paper: the Mee-Owen method (Section 2), a modified Mee-Owen method (Section 2), a method based on the Welch-Aspin series (Section 5) and a method based on the solution of an integral equation (Section 6). The coverage probability functions corresponding to these methods are numbered 1-4 in Figure 3 for five batches each of size five.

The integral equation solution is for most practical purposes an exact solution to the problem. The Mee-Owen method has the disadvantage of being substantially conservative when the variance ratio is small.

Only a modest reduction in this conservative has resulted from the modification of the confidence level of the variance ratio estimate (Section 2, Table 2).

The Welch-Aspin series solution is clearly not suitable for as few as five batches, as discussed in Section 5. However, it is easy to compute and provides an adequate starting function for the iterative solution of the integral equation (6.11).

From the rescaled plot of the coverage probability function for the integral equation solution (Figure 4) it can be seen that for $R > 1$ the actual coverage probability differs from .95 by no more than $\pm$ .00005. This small difference can be attributed to roundoff error. For $R < 1$. however, the difference in the actual and nominal coverage probability indicates that the convergence is not uniform. The convergence of the successive approximations to the tolerance limit factor needs to be more thoroughly examined, though the practical gain from such an investigation may be slight.

8. Example

The data in Table 3 are a pseudo-random sample of 25 from a normal distribution with mean 50 and standard deviation 10. These data have been arbitrar-

ily grouped into five batches of five. By fitting a one-way random effects model to these data one obtains (2.1 - 2.4) :

$$\hat{\mu} = 48.30$$

$$(8.1) \qquad MS_b = 89.88 \qquad \hat{\sigma}_X^2 = 144.9$$

$$MS_e = 158.6 \ .$$

A (.90, .95) lower tolerance limit is of the form

$$(8.2) \qquad T = \hat{\mu} - K\hat{\sigma}_X \ .$$

For the method of Mee and Owen (1983) K = 1.90. If the Mee-Owen method is modified as suggested in in Section 2, then K only decreases to 1.89. The series solution of Section 5 gives K = 1.78 and the integral equation of Section 6 results in K = 1.83. The tolerance limit estimates are, respectively, 25.42, 25.54, 26.82 and 26.29. These values may be compared with the tolerance limit estimate for the pooled data, which is 26.00.

9. Conclusion

One-sided tolerance limits for random effects models is a topic of considerable importance in engineering statistics. The purpose of this paper has been to consider this tolerance limit problem from the point of view of the Welch-Aspin interpretation of the Behrens-Fisher problem. This approach leads to what will very likely prove to be a solution which, for the purposes of applied statistics, is exact. Some numerical work remains to be done, leading to the preparation of tables to be presented in a subsequent publication.

10. References

Aspin, A. A. (1948), "An Examination and Further Development of a Formula Arising in the Problem of Comparing Two Mean Values", Biometrika, 35, 88-96.

"Composite Materials for Aircraft and Aerospace Applications", Mil Handbook 17, in preparation.

Fleiss, J. L. (1971) "On the Distribution of a Linear Combination of Independent Chi Squares", Journal of the American Statistical Association, 66, 142-144.

Lemon, G. H. (1977) "Factors for One-Sided Tolerance Limits for Balanced One-Way ANOVA Random-Effects Model", Journal of the American Statistical Association, 72, 676-680.

Mee, R. W. and Owen, D. B. (1983) "Improved Factors for One-Sided Tolerance Limits for Balanced One-Way ANOVA Random Model", Journal of the American Statistical Association, 78, 901-905.

Neal, D., Vangel, M. and Todt, F. (1987), "Determination of Statistically Based Composite Material Properties", in Engineered Materials Handbook, ed. Cyril A. Dostal, American Society of Metals Press, Metals Park, OH.

Odeh, R. E. and Owen, D. B. (1980), Tables of Normal Tolerance Limits, Sampling Plans and Screening, Marcel Dekker.

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

Ray, W. D. and Pitman A. E. N. T. (1961), "An Exact Distribution of the Fisher-Behrens-Welch Statistic for Testing the Difference Between the Means of Two Normal Populations with Unknown Variances", Journal of the Royal Statistical Society, 23, 377-84.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimate of Variance Components", Bimmetrics Bulletin, 2, 110-114.

Searle, S. R. (1971), Linear Models, John Wiley and Sons, N.Y.

Trickett, W. H. and Welch, B. L. (1954), "On the Comparison of Two Means: Further Discussion of Iterative Methods for Calculating Tables", Biometrika, 41, 361-374.

Welch, B. L. (1947), "The Generalization of Student's Problem When Several Different Population Variances are Involved", Biometrika, 34, 28-35.

Table 1

η Values for (β, γ) Tolerance Limits (Mee and Owen, 1983, p.90)

		γ		
		.90	.95	.99
β	.90	.78	.85	.94
	.95	.79	.86	.95
	.99	.81	.875	.96

Table 2

η Values for (.90, .95) Tolerance Limits for the Mee-Owen Method.

ROWS: Number of batches

COLUMNS: Batch size

	3	4	5	6	7	8	9	10
3	.63	.69	.73	.75	.76	.77	.78	.79
4	.75	.78	.80	.81	.82	.82	.83	.83
5	.80	.82	.83	.83	.83	.84	.84	.84
6	.82	.83	.83	.84	.84	.84	.84	.84
7	.82	.83	.83	.84	.84	.84	.84	.84
8	.82	.83	.83	.84	.84	.84	.84	.84
9	.82	.83	.83	.84	.84	.84	.84	.84
10	.82	.83	.83	.83	.84	.84	.84	.84

Table 3
Example Data

Batch 1	2	3	4	5
59.45	38.46	30.58	55.65	60.41
40.70	43.24	29.15	50.68	64.45
24.67	66.82	46.29	67.62	36.57
30.60	51.95	63.85	42.02	59.76
52.51	38.50	51.71	41.09	40.84

Figure 1

Coverage Probability of a (.90, .95) Tolerance Limit for Pooled Data as a Function of the Variance Ratio

I = J = 5

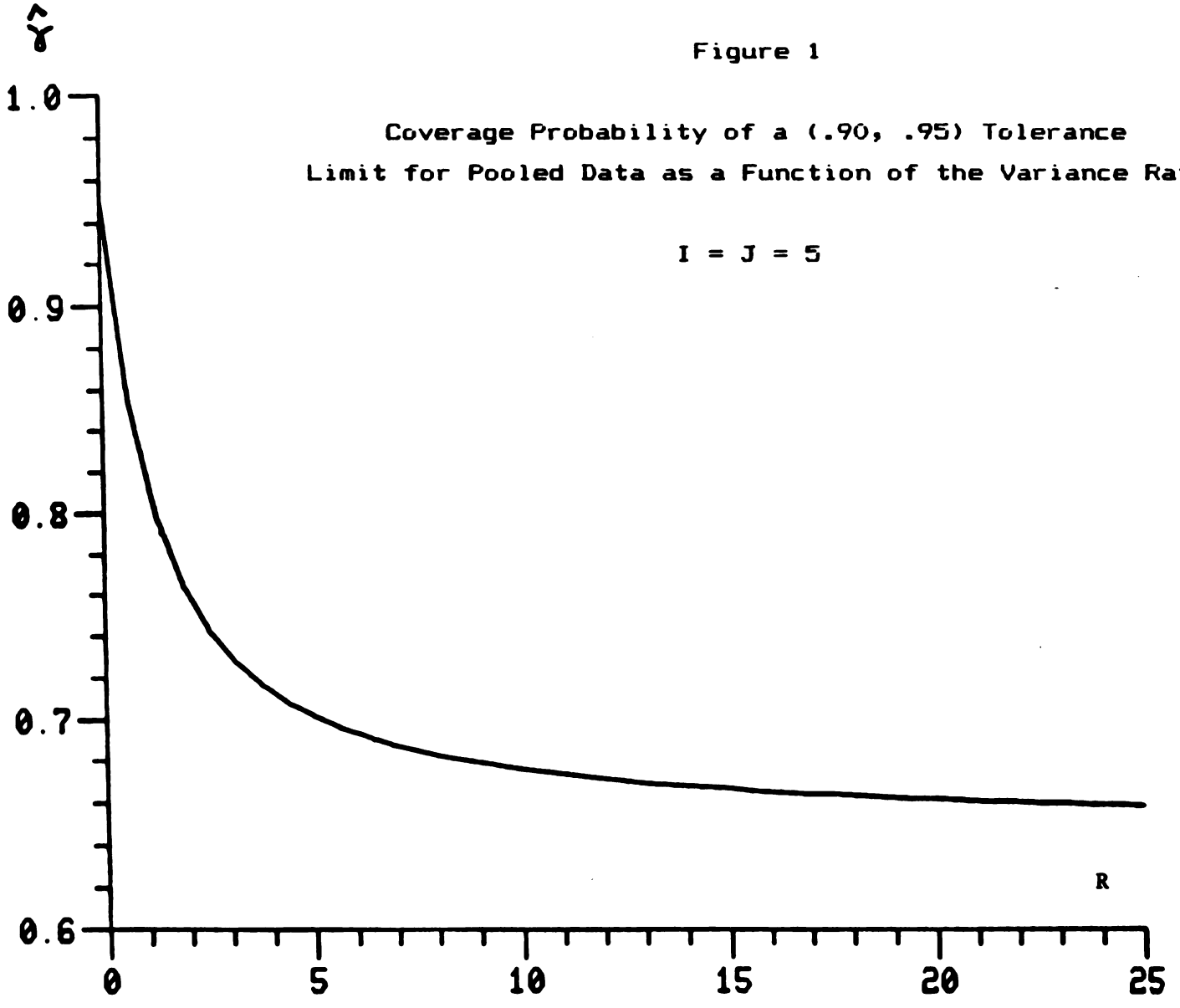


Figure 2

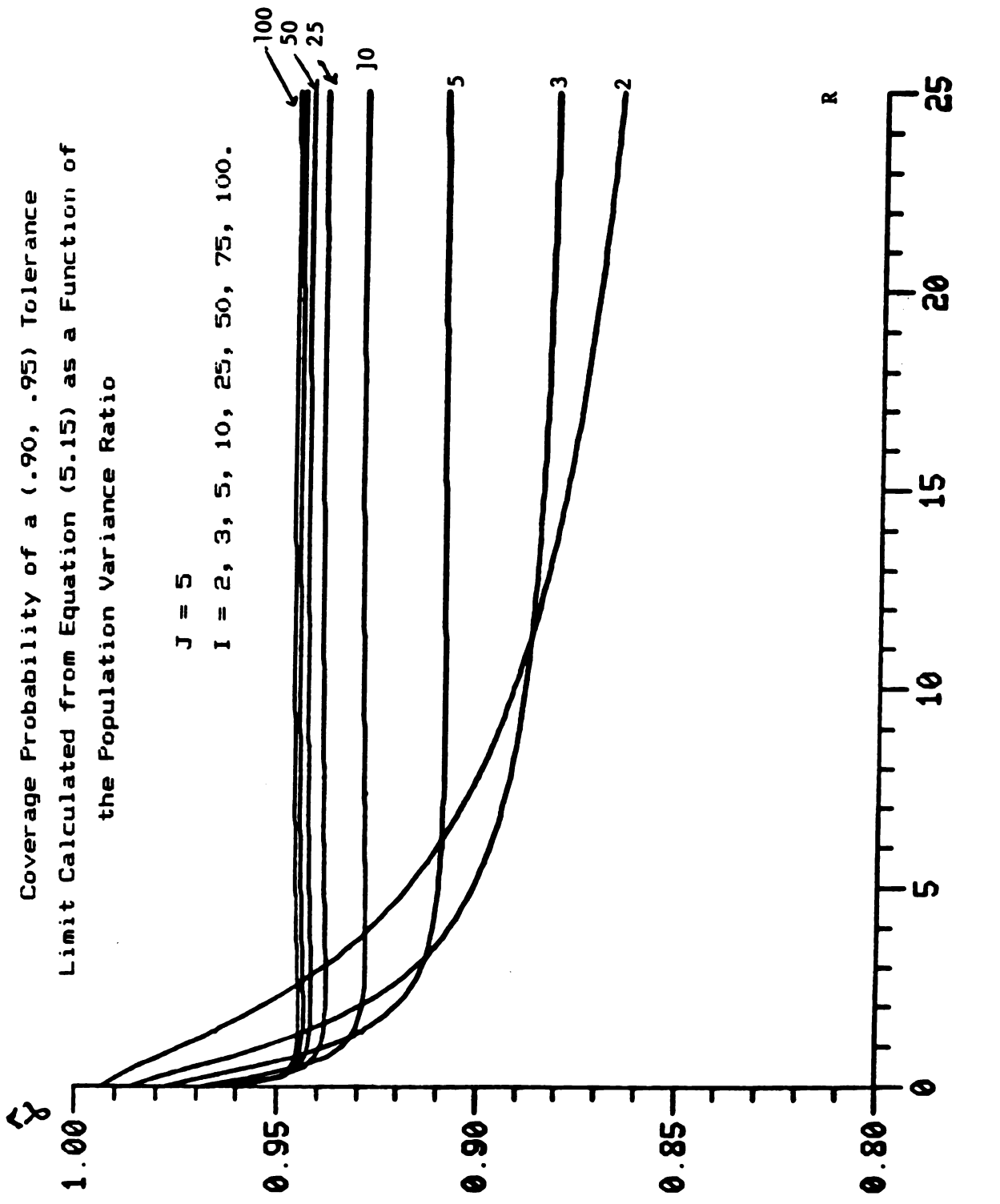


Figure 3

Coverage Probability of a (.90, .95) Tolerance Limit Calculated by Four Methods as a Function of the Population Variance Ratio

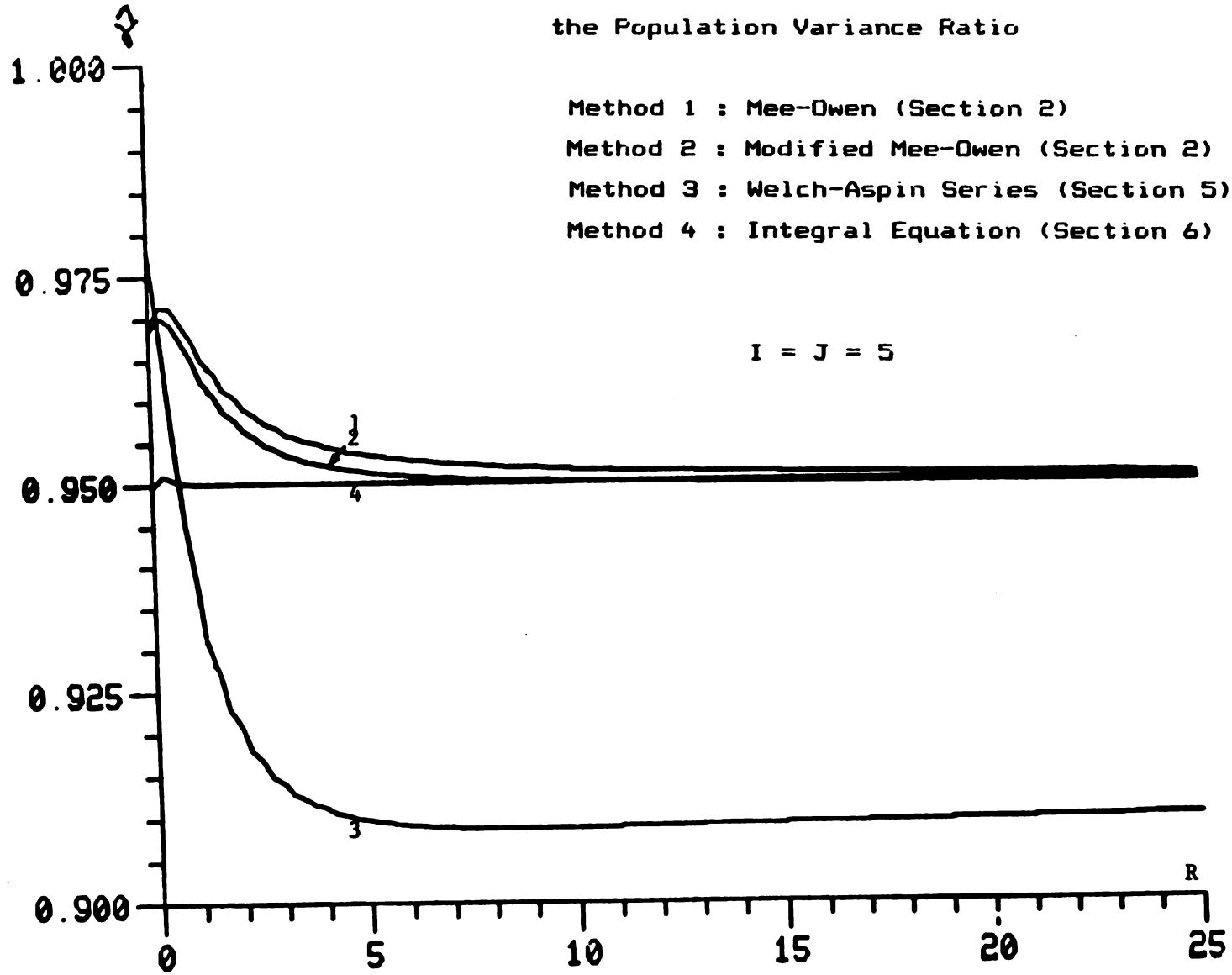
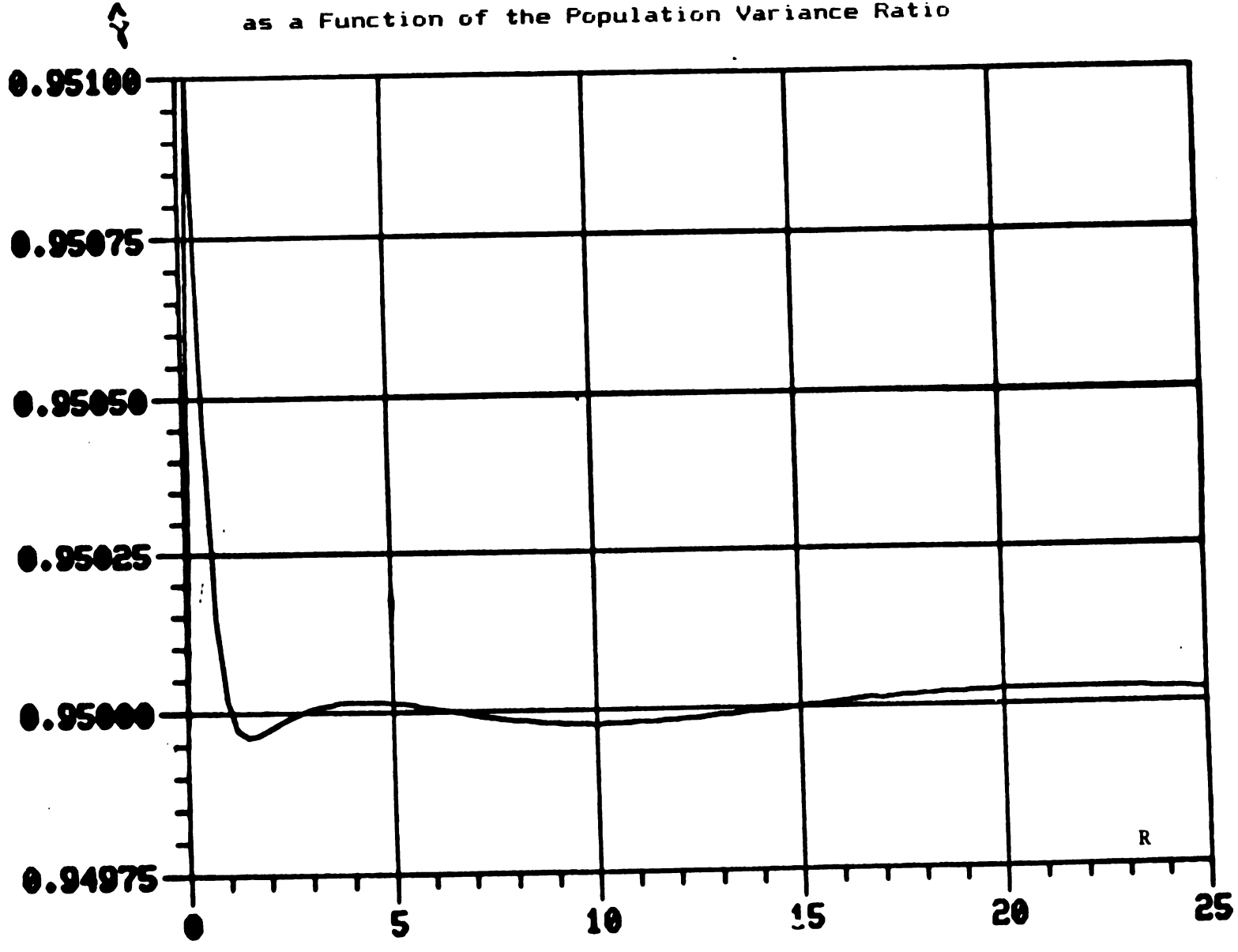


Figure 4

Coverage Probability of a (.90, .95) Tolerance
Limit Based on Equation (6.10) (Figure 3, Method 4)
as a Function of the Population Variance Ratio



THE DISTRIBUTION OF THE NUMBER OF EMPTY CELLS IN A GENERALIZED RANDOM ALLOCATION SCHEME

Bernard Harris[1]
Morris Marden[2]
C.J. Park[3]

ABSTRACT

n balls are randomly distributed in N cells, so that no cell may contain more than one ball. This process is repeated m times. In addition, balls may disappear; such disappearances are independent and identically Bernoulli distributed. Conditions are given under which the number of empty cells has an asymptotically ($N\to\infty$) standard normal distribution.

[1]University of Wisconsin, Madison
[2]University of Wisconsin, Milwaukee
[3]San Diego State University

1. INTRODUCTION

The distribution of the number of empty cells in the following random allocation process is considered. Let n, N be positive integers with $n \leq N$. Assume that n balls are randomly distributed into N cells, so that no cell may contain more than one ball. Then, the probability that each of n specified cells will be occupied is $\binom{N}{n}^{-1}$. This process is repeated m times, so that there are $\binom{N}{n}^{m}$ random allocations of nm balls among the N cells. In addition, for each ball, let p, $0 \leq p \leq 1$, be the probability that the ball will not "disappear" from the cell. The "disappearances" are assumed to be stochastically independent for each ball; thus the disappearances constitute a sequence of nm Bernoulli trials.

Several special cases of this problem have previously been considered . In particular, $p = 1$, $n = 1$ is the classical occupancy problem, see [2],[3],[10]. The case $p = 1$, n arbitrary has been discussed in [4] and [7]. The case $0 < p < 1$, $n = 1$ is treated in C. J. Park [5].

In this paper, we obtain the probability distribution and moments of the number of empty cells. In section 3, we show that the number of empty cells may be represented as a sum of independent Bernoulli random variables. This representation permits us to determine conditions on m, n, p, N such that the number of empty cells is asymptotically normally distributed.

This random allocation process may be viewed as a filing or storage process. Objects are randomly assigned to files or storage bins. From time to time, objects may be missing or have disappeared.

2. THE PROBABILITY DISTRIBUTION AND THE MOMENTS OF THE NUMBER OF EMPTY CELLS

Let m,n,N be positive integers with $n \le N$. m sets, each consisting of n balls, are distributed into N cells at random so that no cell can contain more than one ball from the same set. As each set is distributed, the balls that have been placed during the preceding distributions are left in the cells. Thus, at the end of the process, cells may contain as many as m balls. In addition, each ball may "disappear" with common probability $1 - p$, $0 \le p \le 1$. These disappearances are stochastically independent and thus constitute a sequence of mn Bernoulli trials.

Let $P_{m,n,N,p}(j)$ be the probability that exactly j of the N cells are empty.

We now establish the following theorem.

Theorem 1.

$$P_{m,n,N,p}(j) = \binom{N}{n}^{-m}\binom{N}{j}\sum_{\ell=0}^{N-j}(-1)^{\ell}\binom{N-j}{\ell}\cdot$$

$$\Big[\sum_{i=0}^{j+\ell}(1-p)^{i}\binom{N-j-\ell}{n-i}\binom{j+\ell}{i}\Big]^{m}, \qquad 0 \le j \le N . \tag{1}$$

Proof. Let A_{ν} be the event that the νth cell is empty, $\nu = 1,2,\ldots,N$. Then,

$$P(A_\nu) = \binom{N}{n}^{-m}\Big[\sum_{i=0}^{1}\binom{N-1}{n-i}(1-p)^i\Big]^m . \tag{2}$$

For $1 \le \nu_1 < \nu_2 \le N$,

$$P(A_{\nu_1} \cap A_{\nu_2}) = \binom{N}{n}^{-m}\Big[\sum_{i=0}^{2}\binom{N-2}{n-i}\binom{2}{i}(1-p)^i\Big]^m . \tag{3}$$

Thus, for $1 \le \nu_1 < \nu_2 < \cdots < \nu_k \le N$,

$$P(A_{\nu_1} \cap A_{\nu_2} \cap \cdots \cap A_{\nu_k}) = \binom{N}{n}^{-m}\Big[\sum_{i=0}^{k}\binom{N-k}{n-i}\binom{k}{i}(1-p)^i\Big]^m . \tag{4}$$

Thus, using the inclusion-exclusion method, the probability that exactly j cells are empty is

$$P_{m,n,N,p}(j) = \binom{N}{n}^{-m}\sum_{r=j}^{N}\binom{N}{r}(-1)^{r-j}\binom{r}{j}\Big[\sum_{i=0}^{r}\binom{N-r}{n-i}\binom{r}{i}(1-p)^i\Big]^m . \tag{5}$$

We can write (5) in the form (1) by letting $r = j + \ell$.

We now determine the factorial moments of S, the number of empty cells.

Theorem 2. The νth factorial moment of S,

$$E(S^{(\nu)}) = \binom{N}{n}^{-m} N^{(\nu)}\Big[\sum_{j=0}^{\nu}(1-p)^j\binom{N-\nu}{n-j}\binom{\nu}{j}\Big]^m . \tag{6}$$

Proof. From J. Riordan [9], p. 53, from (4), it follows immediately that

$$E(S^{(\nu)}) = \binom{N}{\nu}\nu!\binom{N}{n}^{-m}\Big[\sum_{i=0}^{\nu}\binom{N-\nu}{n-i}\binom{\nu}{i}(1-p)^i\Big]^m . \tag{7}$$

We thus obtain the following.

Corollary. $$E(S) = N(1 - \frac{pn}{N})^m , \tag{8}$$

$$\sigma_S^2 = N(N-1)\Big[\frac{(N-n)(N-n-1)}{N(N-1)} + 2(1-p)\frac{n(N-n)}{N(N-1)} + (1-p)^2\frac{n(n-1)}{N(N-1)}\Big]^m$$

$$+ N(1 - \frac{pn}{N})^m\Big[1 - N(1 - \frac{pn}{N})^m\Big] . \tag{9}$$

Proof. From (7)

$$E(S) = N\binom{N}{n}^{-m}\Big(\binom{N-1}{n} + \binom{N-1}{n-1}(1-p)\Big)^m = N(1 - \frac{pn}{N})^m .$$

Since

$$\sigma_S^2 = E(S^{(2)}) + E(S) - (E(S))^2 ,$$

the conclusion follows readily from (6), after some elementary calculations.

For some purposes, the following equivalent forms of (9) will prove useful.

$$\sigma_S^2 = N(N-1)\left[1 - \frac{np(2(N-1)-p(n-1))}{N(N-1)}\right]^m + N\left(1-\frac{pn}{N}\right)^m\left(1-N\left(1-\frac{pn}{N}\right)^m\right) \quad (10)$$

and

$$\sigma_S^2 = N^2\left(1-\frac{pn}{N}\right)^{2m}\left\{\left[1-\frac{np^2(N-n)}{(N-1)(N-pn)^2}\right]^m - 1\right\}$$

$$+ N\left(1-\frac{pn}{N}\right)^m\left\{1-\left(1-\frac{pn}{N}\right)^m\left[1-\frac{np^2(N-n)}{(N-1)(N-pn)^2}\right]^m\right\}. \quad (11)$$

From Theorem 2, we readily obtain the following.

Theorem 3. The factorial moment generating function of S is given by

$$\phi_m(t) = E(1+t)^S = \sum_{r=0}^{N}\binom{N}{r}t^r\binom{N}{n}^{-m}\left(\sum_{j=0}^{r}(1-p)^j\binom{N-r}{n-j}\binom{r}{j}\right)^m. \quad (12)$$

Note that $\phi_m(t)$ is a polynomial in t of degree N. This fact is exploited in the next section, where the asymptotic distribution of S is obtained. In particular,

$$\phi_0(t) = (1+t)^N \quad (13)$$

and

$$\phi_1(t) = (1+t)^{N-n}(1 + (1-p)t)^n . \qquad (14)$$

We now investigate the asymptotic distribution properties of the number of empty cells.

3. THE ASYMPTOTIC DISTRIBUTION OF THE NUMBER OF EMPTY CELLS

In this section, we determine conditions under which the number of empty cells (when suitably normalized) has an asymptotically normal distribution. In order to establish this, a number of preliminary results are required.

Lemma 1. Let N,n,r be non-negative integers, $r \le n \le N$. Then

$$\sum_{\nu=\alpha}^{r} \binom{r}{\nu}\binom{\nu}{\alpha}\binom{N-r}{n-\nu} = \binom{r}{\alpha}\binom{N-\alpha}{n-\alpha} . \tag{15}$$

Proof. Since $\binom{\nu}{\alpha} = 0$ whenever $\nu < \alpha$, we can write

$$\sum_{\nu=\alpha}^{r} \binom{r}{\nu}\binom{\nu}{\alpha}\binom{N-r}{n-\nu} = \sum_{\nu=0}^{r} \binom{r}{\nu}\binom{\nu}{\alpha}\binom{N-r}{n-\nu} .$$

To obtain the conclusion, note that

$$\sum_{x=0}^{r} \frac{\binom{x}{\alpha}\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = E\{X^{(\alpha)}\}/\alpha! ,$$

where X has the hypergeometric distribution. From B. Harris [1], p. 105,

$$\sum_{x=0}^{r} \frac{\binom{x}{\alpha}\binom{n}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \frac{r^{(\alpha)}n^{(\alpha)}}{N^{(\alpha)}\alpha!} .$$

The conclusion follows immediately.

Lemma 2.

$$\sum_{j=0}^{n} \frac{(-1)^j \binom{n}{j}\binom{r}{j} p^j}{\binom{N}{j}} = \sum_{\nu=0}^{r} \frac{(1-p)^\nu \binom{N-r}{n-\nu}\binom{r}{\nu}}{\binom{N}{n}} \quad . \tag{16}$$

Proof. The right-hand side of (16) may be written

$$\sum_{\nu=0}^{r} \frac{\binom{N-r}{n-\nu}\binom{r}{\nu}}{\binom{N}{n}} \sum_{j=0}^{\nu} \binom{\nu}{j}(-1)^j p^j = \frac{\sum_{j=0}^{r}(-1)^j p^j \sum_{\nu=j}^{r} \binom{\nu}{j}\binom{N-r}{n-\nu}\binom{r}{\nu}}{\binom{N}{n}}$$

Thus, the coefficient of p^j is

$$(-1)^j \sum_{\nu=j}^{r} \binom{\nu}{j}\binom{N-r}{n-\nu}\binom{r}{\nu}/\binom{N}{n}.$$

From Lemma 1,

$$(-1)^j \sum_{\nu=j}^{r} \binom{\nu}{j}\binom{N-r}{n-\nu}\binom{r}{\nu}/\binom{N}{n} = (-1)^j \binom{r}{j}\binom{N-j}{n-j}/\binom{N}{n} ,$$

from which the conclusion follows immediately. Employing the above lemmas, we can now establish the following theorem.

Theorem 3. The factorial moment generating function of the number of empty cells $\phi_m(t)$ (12) satisfies the following differential-difference equation,

$$\phi_{m+1}(t) = \left(\sum_{j=0}^{n} \frac{(-1)^j \binom{n}{j} (pt)^j D^j}{N^{(j)}} \right) \phi_m(t), \quad m = 0,1,\ldots, \quad (17)$$

where $D^j = \frac{d^j}{dt^j}$.

Proof. For $m = 0$, $\phi_0(t) = (1+t)^N$; hence

$$\left(\sum_{j=0}^{n} \frac{(-1)^j \binom{n}{j} (pt)^j}{N^{(j)}} D^j \right) (1+t)^N = \sum_{j=0}^{n} \frac{(-1)^j \binom{n}{j} (pt)^j}{N^{(j)}} N^{(j)} \cdot [(1+t)^{N-n}(1+t)^{n-j}]$$

$$= (1+t)^{N-n} \sum_{j=0}^{n} (-1)^j \binom{n}{j} (pt)^j (1+t)^{n-j}$$

$$= (1+t)^{N-n}(1+t-pt)^n ,$$

in agreement with (14).

Assume that (17) holds for $m = 1,2,\ldots,k$. Then, from (12),

$$\left(\sum_{j=0}^{n}\frac{(-1)^j\binom{n}{j}(pt)^j\,D^j}{N^{(j)}}\right)\sum_{r=0}^{N}\frac{\binom{N}{r}t^r}{\binom{N}{n}^k}\left(\sum_{\alpha=0}^{r}(1-p)^\alpha\binom{N-r}{n-\alpha}\binom{r}{\alpha}\right)^k$$

$$=\sum_{j=0}^{n}\frac{(-1)^j\binom{n}{j}(pt)^j}{N^{(j)}}\sum_{r=0}^{N}\binom{N}{r}r^{(j)}t^{r-j}\left(\sum_{\alpha=0}^{r}\frac{(1-p)^\alpha\binom{N-r}{n-\alpha}\binom{r}{\alpha}}{\binom{N}{n}}\right)^k,$$

$$=\sum_{r=0}^{N}t^r\sum_{j=0}^{n}\frac{(-1)^j\binom{n}{j}\binom{r}{j}}{\binom{N}{j}}p^j\left(\sum_{\alpha=0}^{r}\frac{(1-p)^\alpha\binom{N-r}{n-\alpha}\binom{r}{\alpha}}{\binom{N}{n}}\right)^k.$$

The conclusion now follows from Lemma 2.

Let

$$T(f(t))=\left(\sum_{j=0}^{n}\frac{(-1)^j\binom{n}{j}(pt)^j\,D^j}{N^{(j)}}\right)f(t),\quad 0<p<1. \tag{18}$$

Then, from Theorem 3, we have that

$$\phi_{m+1}(t)=T(\phi_m(t)),\quad \phi_0(t)=(1+t)^N. \tag{19}$$

Lemma 3. Extend the domain of T to the complex plane, letting $z=x+iy; x,y$ real. Let

$$\psi(z)=\prod_{\alpha=1}^{N}(z-z_\alpha)$$

and

$$\psi_1(z) = T(\psi(z)) = c_1 \prod_{\alpha=1}^{N} (z-z_\alpha^{(1)}). \tag{20}$$

If the zeros of $\psi(z)$ are real and satisfy

$$-b \le x_\alpha \le -a, \quad a,b \ge 0 ,$$

then the zeros of $\psi_1(z)$ are real and satisfy

$$-\frac{b}{(1-p)} \le x_\alpha^{(1)} \le -a . \tag{21}$$

Proof. Let

$$C_\gamma = \{z:|z+(c-i\gamma)| \le [(c-a)^2+\gamma^2]^{1/2}, \ c=\tfrac{1}{2}(a+b)\}. \tag{22}$$

Clearly $-a$ and $-b$ are on the boundary of the circular region C_γ. Consequently all zeros of $\psi(z)$ are in C_γ. Let z^* be a zero of $\psi(z)$. Let

$$\psi_1^*(z_1^{(1)},z_2^{(1)},\ldots,z_N^{(1)}) = c_1(z^*-z_1^{(1)})(z^*-z_2^{(1)}) \cdots (z^*-z_N^{(1)}).$$

That is, $\psi_1^*(z_1^{(1)},z_2^{(1)},\ldots,z_N^{(1)}) = T(\psi(z^*))$ is a linear symmetric function of $z_1^{(1)},z_2^{(1)},\ldots,z_N^{(1)}$). Thus, the conditions of Walsh's theorem (M. Marden [5], p. 62) are satisfied. Thus, if $z_1^{(0)},z_2^{(0)},\ldots,z_N^{(0)}$ are points in C_γ, then there is at least one point ζ in C_γ such that

$$T[(z^* - \zeta)^N] = 0 ,$$

that is, one can set $z_1^{(1)} = \zeta$, $z_2^{(1)} = \zeta,\ldots,z_N^{(1)} = \zeta$ and preserve the value 0. From (18),

$$T[(z^* - \zeta)^N] = (z^* - \zeta)^{N-n}(z^* - \zeta - pz^*)^n = 0 .$$

Thus either $z^* = \zeta$ and therefore z^* is in C_γ or $z^* = \zeta(1-p)^{-1}$ and z^* is in

$$B_{p,\gamma} = \{z:|z+(c-i\gamma)(1-p)^{-1}| \le [(c-a)^2+\gamma^2]^{1/2}(1-p)^{-1} . \qquad (23)$$

However, γ is real and arbitrary. Hence it is clear that

$$C = \bigcap_{-\infty<\gamma<\infty} C_\gamma = \{z:z \text{ real}, -b \le x \le -a\} \qquad (24)$$

and

$$B_p = \bigcap_{-\infty<\gamma<\infty} B_{p,\gamma} = \{z : z \text{ real}, -b(1-p)^{-1} \le x \le -a(1-p)^{-1}\}. \quad (25)$$

Consequently, $C \cup B_p$ is contained in the interval (21), proving the lemma.

We now establish the following theorem.

Theorem 4. Let

$$\phi_m(t) = \sum_{r=0}^{N} \binom{N}{r} t^r \binom{N}{n}^{-m} \left(\sum_{j=0}^{r} (1-p)^j \binom{N-r}{n-j} \binom{r}{j} \right)^m .$$

Let $t_1^{(m)}, t_2^{(m)}, \ldots, t_N^{(m)}$ be the zeros (not necessarily distinct) of $\phi_m(t)$. Then $t_j^{(m)}$, $j = 1,2,\ldots,N$ are all real and

$$-(1-p)^{-m} \le t_j^{(m)} \le -1, \quad j = 1,2,\ldots,N; m = 0,1,\ldots .$$

Proof. From (19),

$$\phi_{m+1}(t) = T(\phi_m(t)), \quad m = 0,1,\ldots,$$

and from (13),

$$\phi_0(t) = (1+t)^N .$$

The zeros of $\phi_0(t)$ are $t_1^{(0)} = t_2^{(0)} = \ldots = t_N^{(0)} = -1$. The zeros of $\phi_1(t)$ are $t_1^{(1)} = -1, \ldots, t_{N-n}^{(1)} = -1$, $t_{N-n+1}^{(1)} = -1/(1-p), \ldots, t_N^{(1)} = -1/(1-p)$. Now apply Lemma 3 with $\psi(z) = \phi_1(z)$ obtaining $a = 1$, $b = (1-p)^{-1}$. Then, the zeros of $\phi_2(t)$ are real and satisfy

$$-(1-p)^{-2} \le t_j^{(2)} \le -1, \quad j = 1,2,\ldots,N .$$

It then follows readily by induction that the zeros of $\phi_k(t)$ are real and satisfy

$$-(1-p)^{-k} \le t_j^{(k)} \le -1, \quad j = 1,2,\ldots,N, \quad k = 2,3,\ldots .$$

Theorem 5. For $1 \le n \le N$, $0 \le p \le 1$, $m \ge 1$, S has a representation as the sum of N mutually independent Bernoulli random variables. That is, there exist mutually independent Bernoulli random variables, $Y_j = Y_j(N,m,p,n)$, $j = 1,2,\ldots,N$, such that

$$S = \sum_{j=1}^{N} Y_j \tag{26}$$

and

$$P\{Y_j = 1\} = \gamma_j = 1 - P\{Y_j = 0\} . \tag{27}$$

Proof. Let Y be a Bernoulli random variable with $P\{Y = 1\} = \tau$.

Then the factorial moment generating function of Y is

$$E_Y\{(1+t)^Y\} = (1+\tau t).$$

If

$$W = \sum_{j=1}^{N} Y_j ,$$

where $Y_1, Y_2, \ldots, Y_N$ are mutually independent Bernoulli random variables with $P\{Y_j = 1\} = \tau_j$, then the factorial moment generating function of W is

$$\xi(t) = E_W\{(1+t)^W\} = \prod_{j=1}^{N} E_{Y_j}\{(1+t)^{Y_j}\} = \prod_{j=1}^{N} (1+\tau_j t), \qquad (28)$$

where $0 \le \tau_j \le 1$, $j = 1,2,\ldots,N$. From Theorem 4, the factorial moment generating fraction of S may be written

$$\phi_m(t) = (1-p)^{nm} \prod_{j=1}^{N} (t-t_j^{(m)}), \quad m = 0,1,\ldots, \qquad (29)$$

where $t_j^{(m)}$ are real and $-(1-p)^{-n} \le t_j^{(m)} \le -1$, $j = 1,2,\ldots,N$.

Since every polynomial of degree N with real roots has a unique representation of the form

$$f(x) = c(x-x_1)(x-x_2)\cdots(x-x_N), \quad x_1 \le x_2 \le \cdots \le x_N ,$$

the representation follows by setting $\tau_j = -(t_j^{(m)})^{-1}$ and noting that $\xi(0) = \phi_m(0) = 1$.

Let $\kappa_\ell = \kappa_\ell(n,N,m,p)$ be the cumulants of S and let $\kappa_{[\nu]}$ be the factorial cumulants of S. That is,

$$\log \phi_m(t) = \sum_{\nu=1}^{\infty} \kappa_{[\nu]} t^\nu/\nu! \quad .$$

Then

$$\kappa_\ell = \sum_{j=1}^{\ell} \beta_{j,\ell} \kappa_{[j]}, \qquad \ell \ge 2,$$

where $\beta_{j,\ell}$ are the Stirling numbers of the second kind.

Then, as $N \to \infty$,

$$V = (S - E(S))/\sigma_S$$

is asymptotically distributed by the standard normal distribution (mean 0, variance unity), whenever

$$\kappa_\ell/\kappa_2^{\ell/2} \to 0, \qquad \ell > 2 .$$

From (29),

$$\log \phi_m(t) = nm \log(1-p) + \sum_{i=1}^{N} \log(t-t_j^{(m)}) = \sum_{i=1}^{N} \log(1+\tau_i t)$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{\infty} \frac{(\tau_i t)^k}{k} (-1)^k .$$

Thus,

$$\frac{\kappa_{[\nu]}}{\nu!} = \sum_{i=1}^{N} \frac{(-1)^\nu}{\nu} \tau_i^\nu, \quad 0 < \tau_i \le 1,$$

and

$$|\kappa_{[\nu]}|/\nu! \le \frac{1}{\nu} \sum_{i=1}^{N} |\tau_i^\nu| \le N/\nu .$$

Then

$$|\sum_{j=1}^{\ell} \beta_{j,\ell} \kappa_{[j]}| \le c_\ell N, \tag{30}$$

since the $\beta_{j,\ell}$ do not depend on N,n,m, or p .

We now establish the following theorem.

Theorem 6. $V = (S-E(S))/\sigma_S$ has an asymptotically standard normal distribution as $N \to \infty$, whenever any of the following conditions are satisfied.

1. $\frac{mpn}{N} \to 0$, $p \to p^* \neq 1$ and $\frac{mnp}{N^{2/3}} \to \infty$;

2. $\frac{mpn}{N} \to 0$, $(1-p) \to 0$ so that for some $c > 0$,

$$(1-p) = c\left(\frac{mnp}{N}\right)^{\rho} + o\left(\left(\frac{mnp}{N}\right)^{\rho}\right), \quad 0 < \rho < 1, \text{ and}$$

$$\frac{mnp}{N^{\left(1-\frac{1}{3(\rho+1)}\right)}} \to \infty ;$$

3. $\frac{mnp}{N} \to 0$, $(1-p) = c\left(\frac{mnp}{N}\right)^{\rho} + o\left(\left(\frac{mnp}{N}\right)^{\rho}\right)$, $\rho \geq 1$, and

$$\frac{mnp}{N^{5/6}} \to \infty ;$$

4. $\frac{mnp}{N} \to r > 0$;

5. $\frac{mnp}{N} \to \infty$ and $\frac{3mpn}{N} - \log N \to -\infty$.

Proof. From (11), we can write, for $\alpha \to 0$,

$$\kappa_2 = N(e^{-\alpha})(1-e^{-\alpha} - \alpha p e^{-\alpha}) + O(np\alpha) + O(p^2\alpha^2)$$

where $\alpha = \frac{mnp}{N}$. Then, as $\alpha \to 0$,

$$\kappa_2 = N(1-\alpha+\alpha^2/2)(\alpha - \frac{\alpha^2}{2} - \alpha p + \alpha^2 p) + O(N\alpha^3) + O(mn\alpha).$$

Then, if $p \to p^* \neq 1$,

$$\kappa_2 = N\alpha(1-p) + O(N\alpha^2)$$

and

$$\frac{\kappa_2^{3/2}}{N} \to \infty \quad \text{whenever} \quad \frac{mnp}{N^{2/3}} \to \infty .$$

Similarly, if $(1-p) = c(\frac{mnp}{N})^{\rho} + C((\frac{mnp}{N})^{\rho})$, $0 < \rho < 1$, $c > 0$, then

$$\kappa_2 = N\alpha(1-p) + o(N\alpha(1-p))$$

and

$$\frac{\kappa_2^{3/2}}{N} \to \infty \quad \text{whenever} \quad \frac{mnp}{N^{\left(1-\frac{1}{3(\rho+1)}\right)}} \to \infty \ .$$

The conclusion is obvious whenever $\frac{mnp}{N} \to r > 0$.

If $\alpha \to \infty$ as $N \to \infty$, then

$$\kappa_2 = Ne^{-\alpha} + O(Ne^{-2\alpha})$$

and

$$\frac{\kappa_2^{3/2}}{N} \to \infty \quad \text{whenever} \quad 3\alpha - \log N \to -\infty.$$

REFERENCES

1. Harris, B. (1966). Theory of Probability, Addison-Wesley Publishing Company, Reading, Mass.

2. Harris, B. and Park, C. J. (1971). "A note on the asymptotic normality of the distribution of the number of empty cells in occupancy problems," Ann. Inst. Statist. Math., 23, 507-513.

3. Holst, Lars (1977). "Some asymptotic results for occupancy problems," Ann. Probability, 5, 1028-1035.

4. Kolchin, V. F., Sevast'yanov, B. A. and Chistyakov, V. P. (1978). Random Allocations. V. H. Winstons & Sons, Washington, DC.

5. Marden, M. (1966). Geometry of Polynomials. Second Edition, Mathematical Surveys, No. 3, American Mathematical Society, Providence, R.I.

6. Park, C. J. (1972). "A note on the classical occupancy problem," Ann. Math. Statist., 43, 1698-1701.

7. Park, C. J. (1981). "On the distribution of the number of unobserved elements when m-samples of size n are drawn from a finite population population," Comm. Statist., A-Theory Methods, 10, 371-383.

8. Renyi, A. (1962). "Three new proofs and a generalization of a theorem of Irving Weiss," Magyar Tnd. Akad. Math. Kutató. Int. Közl. A, 7, 203-214.

9. Riordan, J. (1958). An Introduction to Combinatorial Analysis, John Wiley and Sons, Inc., New York, N.Y.

10. Sevast'yanov, B. A. and Chistyakov, V. P. (1964). "Asymptotic normality in the classical ball problem," Theory of Probability and Its Applications, 9, 198-211.

APPLICATION OF EXPERIMENTAL DESIGN TO THE EVALUATION OF EXPERT OPINION

Franklin E. Womack and Carl B. Bates

US Army Concepts Analysis Agency

Bethesda, Maryland 20814-2797

ABSTRACT. Expert opinion can be a valuable source of information to tap in the building of a systems model. At the US Army Concepts Analysis Agency (CAA), the computer model FORCEM (Force Evaluation Model) is used to evaluate the theater-level combat system. FORCEM is built and maintained by a group of CAA analysts. The command and control part of FORCEM is a logical surrogate for the field commander at various levels of combat (i.e., theater, army, corps, or division). A simulated war is conducted by exercising FORCEM. The command and control part of FORCEM is allowed to perceive information about the state of the war through a perception data base. Using the information from the data base, it applies decision rules for the further conduct of the war. In order to validate these decision rules and make enhancements to the present model, 81 students at the Army War College, Carlisle, Pennsylvania, participated in an information gathering experiment. Several decisions from the command and control part of FORCEM were presented to these experts in the form of a structured experimental design. Information from the perception data base served as factors for the experimental design, and responses were solicited from these experts. This paper discusses the experimental design employed and the statistical analysis performed.

Comments by panelists Drs. Kaye Basford and W. T. Federer are at the end of this artical.

1. INTRODUCTION. The US Army Concepts Analysis Agency developed the Force Evaluation Model (FORCEM) during the period 1982-1985. FORCEM is a fully automated computer simulation of a conventional theater campaign treating combat, combat support, and combat service support in a theater of operations. The model is used in studies of the capabilities of current combat forces; requirements for support forces; and requirements for personnel, supplies, and major items of equipment.

FORCEM is a time-sequenced model; each cycle represents 12 hours of simulated time. At the beginning of each cycle, intelligence and communications determine a set of perceived data for each headquarters unit. Based on these data, command and control (C^2) decisions are made. Then the activities of the cycle are represented: combat movement and combat service support.

Command and control representation depends on a perception data base and decision algorithms. The decision algorithms are built into the model and involve a set of input threshold parameters. This paper addresses a study of the C^2 decision algorithms.

2. PROBLEM DESCRIPTION. The purpose of the study was to verify or enhance the C^2 decision algorithms of FORCEM. The decision algorithms considered are identified in Table 1. Each decision algorithm was examined and the factors within the algorithm were identified. Naturally, some factors are contained in more than one algorithm. The 12 unique factors involved in the 8 decision algorithms are listed in Table 2.

Table 1. FORCEM Decisions Considered

Number	Decision
1	Assignment of New Corps
2	Assignment of New Division
3	Assignment of New Field Artillery Battalion
4	Designation of Posture of Online Corps
5	Specification of Priority to Corps for CAS
6	Specification of Priority to Corps for CSS
7	Specification of Priority to Division for CAS
8	Specification of Priority to Division for CSS

Table 2. Decision Factors

Symbol	Factor
A	Has Reserve Corps
B	Corps Engagement Status
C	Corps Force Ratio
D	Location of Objective of Corps/Posture of Corps
E	Echelon to Which Corps Assigned/Has Reserve Corps
F	Echelon to Which Corps Assigned
G	Ratio of Corps Artillery Battalions to Divisions
H	Location of Objective of Corps
I	Posture of Parent Army
J	Division Equipment Status
K	Division Force Ratio
L	Echelon to Which Division Assigned

The levels of each of the factors are given in Table 3. Factor D is actually a combination of two factors (Objective Location and Posture); however, all factor-level combinations of the two factors did not exist. Only the six combinations shown in Table 4 existed.

Table 3. Levels of Decision Factors

Factor	Level 1	2	3	4	5	6
A	No res	Has res				
B	No res	Engaged				
C	1:3	1:1	3:1			
D	Rear/ Withdr	Reached/ Delay	Reached/ Defend	Fwd/ Delay	Fwd/ Defend	Fwd/ Attack
E	Reserv	Onln/Yes	Onln/No			
F	Reserv	Online				
G	1.00	0.25				
H	Rear	Reached	Forward			
I	Delay	Defend	Attack			
J	No	Yes				
K	1:3	1:1	3:1			
L	First	Second	Third			

Table 4. Possible Location/Posture Combinations

Objective location	Posture			
	Withdraw	Delay	Defend	Attack
Rear	1			
Reached		2	3	
Forward		4	5	6

All the factors within each experiment were completely crossed within all other factors of the experiment. Consequently, all designs were factorial designs. The factors and the number of cells are shown in Table 5. The sizes of the experiments range from 12 to 108 cells.

Table 5. The Eight Experiments

Decision number	Factors	Number of levels	Number of cells
1	AxBxCxD	2x2x3x6	72
2	BxCxDxE	2x3x6x3	108
3	BxDxFxG	2x6x2x2	48
4	CxHxI	3x3x3	27
5	BxCxF	2x3x2	12
6	BxCxF	2x3x2	12
7	JxKxL	2x3x3	18
8	JxKxL	2x3x3	18

For each experiment, a questionnaire was developed that described the scenarios defined by the factor-level combinations (cells). Subjects (military officers) were asked to assign a criticality index (from 0 to 100) except for decision number 4. For decision 4, subject's response was one of the four postures--Withdraw, Delay, Defend, or Attack.

3. TEST METHOD. The approach was to use students at the US Army War College as subjects, use computerized questionnaires for each of the eight decisions, and collect data from the Army "experts" concerning the criticality of each of the scenarios of each of the eight decisions.

To test the feasibility of the planned approach, a pilot test was conducted inhouse. Decision number 1, which involves factors A, B, C, and D, was selected for the pilot test. Nine senior officers of the US Army Concepts Analysis Agency were selected as subjects. In the pilot test, only the high and low levels (1:3 and 3:1) were used for factor C (corps force ratio). Five to ten practice questions (Figure 1) were given before the 48 questions of the 2x2x2x6 design were given to allow for any learning effect. To assess the subject effect, Subjects (S) were treated as a random factor (factors A, B, C, and D were fixed). Five of the cells were replicated to provide an estimate of within error variance.

YOU WILL BE ASKED TO RESPOND BY ENTERING A NUMBER BETWEEN 0 and 100
based on the following scale of how critical
you think it is for the newly arrived CORPS
to be assigned to reserve status behind the
ONLINE CORPS. After entering a number hit 'XMIT'.

0 20 40 60 80 100

NOT CRITICAL | SLIGHTLY CRITICAL | MODERATELY CRITICAL | VERY CRITICAL | EXTREMELY CRITICAL

PLEASE HIT 'XMIT' NOW TO PROCEED

PAUSE 00000

WARMUP NUMBER # 1

1. There is currently at least one CORPS assigned in reserve behind the ONLINE CORPS.

2. The ONLINE CORPS is currently engaged.

3. The location of the parent Army's Objective Phase Line is now located at the present position of the ONLINE CORPS' current forward phase line.

4. Assuming all divisions currently assigned to the ONLINE CORPS are in place, the current posture of the ONLINE CORPS is defend.

5. Assuming all divisions currently assigned to ONLINE CORPS are in place, the friendly-to-enemy combat worth force ratio is currently perceived to be FRIEND:ENEMY (1:3)

PLEASE RESPOND BY ENTERING A NUMBER BETWEEN 0 AND 100
based on the aforementioned scale of how critical
you think it is for the newly arrived CORPS
to be assigned to reserve status behind the
ONLINE CORPS. After entering a number hit 'XMIT'.

PLEASE ENTER THE NUMBER NOW.

Figure 1. Sample Question

An analysis of variance (ANOVA) was performed on the data. The ANOVA model was

$$y = \mu + A + B + C + D + S + AB + AC + \ldots + DS + \vdots + ABCDS + R,$$

where y represents criticality index; μ is a true but unknown constant; A, B, C, D, and S are as defined above; and R is residual. All effects involving S were tested over MS(R), and all fixed effects were tested over their corresponding interaction with S. That is, the F-ratio for testing the factorial effect of factor A is MS(A)/MS(AS), and the F-ratio for testing the AB interaction effect is MS(AB)/MS(ABS). Some of the Subject variance components were statistically significant; however, the four fixed effects factors accounted for over 60 percent of the total variability.

The ANOVA results were used to give a hypothesized "significant" model for fitting. Dummy variables were used for the qualitative factors and regression analysis was used to develop a prediction equation. This prediction equation provided the model to be compared with the current FORCEM algorithm for the particular decision. The comparison is shown in Table 6, which contains the regression model predicted values, the 48 cell means, and the current algorithm priority. The first and the forty-eighth priorities of all three priorities agree. Also, the first six to seven and the last five of the regression model and cell mean priorities agree.

Table 6. Comparison of Models

Regression model predicted value and priority		Cell mean critical index and priority		Present FORCEM priority
1	99.4	1	96.2	1
2	92.6	2	91.4	13
3	79.8	3	79.3	25
4	72.9	4	77.8	37
5	64.6	5	67.1	3
6	57.8	6	65.0	15
7	54.4	9	54.5	5
8	53.8	7	62.1	7
9	53.6	8	58.4	2
10	47.5	12	43.4	17
11	47.0	14	43.3	19
12	46.8	10	47.8	14
13	45.4	11	44.5	9
14	45.0	13	43.4	27
15	41.8	15	40.0	26
16	38.6	17	35.8	21
17	38.1	21	32.4	39
18	37.5	18	35.6	11
19	35.0	22	31.1	38
20	34.8	16	37.7	29
21	34.2	25	30.0	31
22	33.8	19	34.4	4
23	31.1	20	32.5	8
24	30.7	24	30.5	23
25	27.9	27	26.7	41
26	27.3	30	24.7	43
27	26.9	28	25.3	16
28	26.1	26	28.4	6
29	25.8	23	30.6	33
30	25.4	31	22.3	10
31	24.3	29	25.1	20
32	22.0	37	17.7	28
33	21.3	33	19.0	12
34	19.2	40	14.7	32
35	19.3	36	18.0	18
36	18.9	32	20.1	45
37	18.5	38	17.7	22
38	17.9	34	18.1	35
39	15.1	35	18.1	40
40	14.5	44	12.8	24
41	14.3	41	14.6	30
42	13.5	42	13.0	34
43	12.5	39	17.2	44
44	11.0	43	13.0	47
45	9.5	45	10.7	36
46	7.5	47	8.5	42
47	6.7	46	8.8	46
48	2.7	48	5.5	48

The regression model equation was considered to be an adequate fit of the data for the intended purpose. Consequently, the pilot test was considered successful, despite the fact that the results of the developed models were inconsistent with the algorithm priorities. The decision was made to proceed with the project as planned.

4. DATA COLLECTION. A questionnaire was computerized for each of the eight decisions in Table 1 and administered to a group of students (Subjects) from the US Army War College. The experiments were administered on four afternoons during December 1985 and January 1986. Each afternoon consisted of two 2-hour sessions with approximately 10 subjects. The allocation of subjects to experiments is shown in Table 7. Experiments 1, 2, 3, 7, and 8 were administered to 20 subjects, experiments 5 and 6 were administered to 21 subjects, and experiment 4 was administered to all 81 subjects.

Table 7. Allocation of Subjects to Experiments

Group (session)	Decision 1	2	3	4	5	6	7	8	Number of subjects
1	X			X		X			10
2	X			X			X		10
3		X		X	X				10
4		X		X					10
5			X	X				X	10
6			X	X					10
7				X			X	X	10
8				X	X	X			11

5. ANALYSIS. In addition it was recognized that the present FORCEM decision could be written as a linear equation. In decision #1, one online corps among several candidates must be chosen by the theater headquarters to receive a newly arrived corps in reserve. The factors used to make this decision are A, B, C and D discussed in Table 2 above. Each candidate has a specific set of four values associated with it. Each such value corresponds to a particular level of one of the factors as discussed in Table 3 above. For each candidate, the equation $Y = 55 - 36 \cdot A + 18 \cdot B - C + 3 \cdot D$ is evaluated using the four values associated with it. The Y value so calculated is the priority for the candidate. The candidate corps whose priority is larger than all of the other candidates is chosen to receive

the newly arrived corps in reserve. If two or more candidates tie with the largest priority, no decision can be made based on these factors. In this case, each of the four values for these candidates would be equal. This would imply their equivalence in relation to the four factors considered.

A more fruitful analysis could be obtained if the subjects' responses could be transformed to a response similar to the priority value assigned to each online corps by the present FORCEM algorithm. One transformation that showed definite promise was the rank transformation. The rank transformation used consisted of ranking each subject's response from 1 to N, where N is the number of cells in the particular design (N = 72 for Decision #1). The smallest criticality index, usually a zero, was assigned the value 1 and the largest criticality index, say 100, was assigned a 72. Where several responses of the subject had the same value (i.e., ties), the average rank was used. The rank transformation did not seem to affect the overall results obtained in the original cell means model, and had the added advantage of being directly testable against the present linear model of the FORCEM algorithm. Using the ranked responses of the military experts and estimating coefficients of the same linear model of the FORCEM algorithm, the equation $Y = 62.4 - 15 \cdot A + 3.94 \cdot B - 12.1 \cdot C + 4.27 \cdot D$ was obtained. However this model lacked fit and a better model was obtained by adding terms related to the significant cross products of the cell mean model, $Y = 50.0 - 9.57 \cdot A + 10.4 \cdot B - 14.2 \cdot C + 17.1 \cdot D - 6.17 \cdot D^2 + 0.698 \cdot D^3 - 4.28 \cdot A \cdot B + 0.588 \cdot C \cdot D$. Testing the null hypothesis of no difference between this model and the model of the FORCEM algorithm, one obtains a

calculated F ratio of 215.3. This is much larger than F(9,1368,.95) = 1.92. This implies that the hypothesis of no difference between the models must be rejected.

Decision 4, Designation of Posture of Online Corps, was treated differently from the other decisions because it involves an ordered categorical response variable. The response variable is posture. The subject was required to choose the most appropriate posture for a given set of input factors. The choice was attack, defend, delay, or withdraw. The factors gave a structure on which to base the experiment; however, each cell was analyzed independently of the other cells. The factors C, H, and I are described in Table 2, and the levels are shown in Table 3. In FORCEM, a definite posture must be assigned to a corps given a set of factor levels. A posture assignment is unique for a given set of factor levels and is given to each corps possessing a particular factor-level combination during a run of FORCEM. In the real world posture assignment would probably be stochastic rather than deterministic. An approach to dealing with this statistically is to test each cell with a simple statistical hypothesis test. For each of the 27 cells, the null hypothesis for the cell is that less than half of the expert population chooses any one of the postures. The alternate hypothesis, the statement desired for the cell, is that more than half of the expert population chooses one common posture; i.e., a "majority" posture. The test takes the form of H_0: $p \leq 0.5$ and H_A: $p > 0.5$. The random variable X_i ($i = 1$ to 81, for sample of 81 expert subjects) takes on the value 1 when a subject picks the posture with largest number of responses (i.e., the "favored" posture) in

the cell under consideration; the probability that $X_i = 1$ is p. The random variable X_i takes on the value 0 if the subject picks any other posture; the probability that $X_i = 0$ is (1 - p). If there is a tie for the favored posture, the test cannot logically result in a rejection of the null hypothesis. Assuming there is a favored posture, a test must be constructed to decide whether (1) to reject the null hypothesis or (2) not to reject the null hypothesis because of insufficient evidence to the contrary. The appropriate distribution is the distribution of the sum of the random variables X_i. This is the binomial distribution with parameters N = 81 and p = 0.5. A critical region must then be determined for which the null hypothesis is rejected when in fact true with no more than a stated probability. This probability is referred to as alpha, the significance level of the test. For the case under consideration, (N = 81), alpha = 0.05, the critical region corresponds to a count of responses of K = 48. For alpha = 0.01, K = 52. On this basis, the count for each of the 27 cells is tested in the hope of rejecting the null hypothesis. Table 8 displays the results of the test. The favored posture is designated in the cell for the given levels of the factors C, H, and I. The number of subjects of the total of 81 choosing the posture is indicated in parentheses. Double asterisks (**) indicate that the null hypothesis can be rejected at the alpha = 0.01-level of significance, and a single asterisk (*) indicates that the null hypothesis can be rejected at the alpha = 0.05 level. For the remaining cells (those without asterisks), there is insufficient evidence to reject the null hypothesis; indeed, as noted in the table, for some cells there is no favored posture.

Table 8. Decision 4 "Favored" Postures

Posture of parent army	Force ratio	Location of objective: Rear	Reached	Forward
Delay	1:3	Delay (51)*	Defend (42)	Defend (56)**
	1:1	Delay (47)	Defend (46)	Defend (51)*
	3:1	Delay (31)#	Defend (39)#	Attack (58)**
Defend	1:3	Defend (37)#	Defend (72)**	Defend (68)**
	1:1	Defend (47)	Defend (70)**	Defend (52)**
	3:1	Defend (38)#	Defend (53)**	Attack (70)**
Attack	1:3	Defend (59)**	Defend (61)**	Defend (51)*
	1:1	Defend (53)**	Defend (49)*	Attack (43)
	3:1	Attack (52)**	Attack (67)**	Attack (81)**

Key:

**: significant at alpha = 0.01

*: significant at alpha = 0.05

#: no majority posture

6. SUMMARY. Concerning the seven experiments having criticality as the response variable, the smaller experiments appeared more successful than the larger experiments. Subjects' responses suggest that the scale 0 to 100 is too large a scale. Most subjects assigned values by 10s--10, 20, 30, ...; some assigned values by 5s--5, 10, 15, ...; and very few assigned by unity such as 17, 43, or 83. Large experiments may exceed the differentiality of subjects. There was also evidence that all subjects were not on the same scale. Some tended to use the lower portion, some the center, and some the upper portion of the scale. Hetrogeneity was also a problem. This also seemed more severe with the larger experiments.

Concerning the experiment with the discrete response, it was not felt that the analysis employed was the most appropriate. Time did not permit further study and research of the problem.

Finally, if subjects employed are indeed experts, the statistical methods of experimental design, analysis of variance, and regression analysis have potential for verification of algorithms of simulation models.

COMMENTS BY PANELISTS DR. KAYE BASFORD AND PROFESSOR W. T. FEDERER ON THE FOLLOWING ARTICAL

Application of experimental design to the evaluation of expert opinion by

Franklin E. Womack
and Carl B. Bates
U.S. Army Concepts Analysis Agency

Kaye Basford: The authors are attempting to validate decision rules and make enhancements to the present model based on responses from 81 students at the Army War College. This appears to be a different population from the one used to originally specify the model. Thus different answers could be a result of the differing populations rather than just a larger sample from the same population.

W.T. Federer: The U.S. Army Concepts Analysis Agency (CAA) has a computer model FORCEM wherein the command and control part can be used by a field commander at various levels of combat. Expert opinion is a valuable component of such a systems model. Using FORCEM a field commander can make decisions about the future conduct of a war. In order to further improve FORCEM, 81 students from the Army War College participated in an information gathering system. To this writer, it would appear that the use of FORCEM would be costwise efficient if commanders simulated a war rather than actually field testing everything. It is realized that final decisions from any simulation should be field tested but considerable insight can be gained from simulations and at a relatively low cost.

A number of experiments were conducted using a factorial treatment design and groups of 20 (21 in one case) students in an experiment. The response variable was a criticality score (zero to 100) except for one response variable. The writers used an effect by subjects interaction as an error mean square for each effect. Why weren't some interactions with subjects pooled to increase degrees of freedom in an error term? Why wasn't an analysis performed on the eight decisions and eight group sessions in Table 7? Also, the regression model used needs more explanation. Presumably, this is a main effects regression model with the eight decisions as the eight independent variables in the regression equation. If the interactions are small compared to main effects, it would be expected that the agreement between predicted values from regression and cell means would be good (see Table 6).

ANALYSIS OF AN INCOMPLETE BLOCK DESIGN WITH MISSING CELLS

Wendy A. Winner
Jill H. Smith

Director,
U. S. Army Ballistic Research Laboratory
ATTN: SLCBR-SE-D
Aberdeen Proving Ground, MD 21005-5066

Abstract

The Ballistic Research Laboratory (BRL) conducted an interactive Firepower Control Experiment from 2 thru 20 December 1985 to acquire knowledge on how military personnel make tactical fire control decisions for field artillery, and, for the first time, automatically collected data on the digital communications between the field artillery battery Fire Direction Center (FA btry FDC) and simulated 155mm howitzer firing units. This later portion of the experiment, the Battery Fire Direction Center (Btry FDC) portion, was designed to test 3 levels of the number of howitzers per battery, 3 levels of simultaneous missions, and 2 levels of fire mission control ratios with each other. The intended design was three replications of a 3 x 3 x 2 factorial with the linear Howitzer x Mission interaction blocked by day. Unforeseen software problems precluded the completion of the design for this controlled laboratory experiment. As a result, informative data was only collected for twelve of the eighteen treatment combinations of a single replication. At the conference, expert advice was solicited on the appropriate method of analysis and the appropriate conclusions to draw from the analysis on data collected from this experiment.†

Comments by panelists Drs. Kaye Basford and W. T. Federer are at the end of this artical.

I. Introduction

The Ballistic Research Laboratory (BRL) conducted an interactive Firepower Control Experiment from 2 thru 20 December 1985 to acquire knowledge on Fire Direction Officers' (FDOs') tactical fire control decisions, and, for the first time, automatically collected data on the digital communications between the field artillery battery Fire Direction Center (FA btry FDC) and simulated 155mm howitzer firing units. The objectives of this experiment were (1) to collect data on the FDOs' decisions on the type and volume of howitzer fire for selected targets, and (2) to characterize the net utilization between the Battery Computer System (BCS) at the btry FDC and the Gun Display Units (GDUs) at each howitzer of the firing btry. These two objectives were achieved by conducting a controlled laboratory experiment that simultaneously focused on these two independent objectives, i.e., portions, of the Firepower Control Experiment.

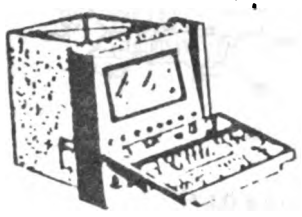
II. Test Configuration and Design

To run the portions together, the BRL integrated a commonly shared database, uniquely developed BRL simulators, and a combination of tactical and commercial computer equipment interfaced by a BRL "Bit Box", i.e., a modem between GDU protocol and standard, commercial computer RS232 protocol. Officers from the U.S. Army Field Artillery School, Fort Sill, OK, participated as FDOs and BCS operators while BRL's interactive simulators emulated forward observers (i.e., the Multiple Forward Observer SCEnario simulator, MFOSCE), the Tactical Fire Direction System (TACFIRE) battalion FDC (i.e., the Fire Direction Simulator, FDS), and the firing btry (i.e., the Gun Display Unit Simulator, GUNSIM). **Figure 1** outlines these major components of the laboratory setup, and **Figure 2** depicts their field counterparts.

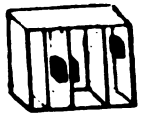
Six different test cells containing sixty targets each were developed from a Scenario Oriented Recurring Evaluation System Europe-I, Sequence 2A (SCORES 2A) division slice. Each test cell was developed to contain an identical mixture of twenty different target types. The sixty targets in each test cell were randomized, and the six test cells were used to produce a total of eighteen scenario test cells. All targets in each test cell were forwarded to the FDO for selection of a type and volume of fire. Twelve preidentified targets of the sixty targets were sent to the BCS operator as fire missions, i.e., targets to be fired on with the specified type and volume of fire. It was hypothesized that the BCS would require an hour of testing to fire *all* twelve fire missions and that an additional forty-eight targets would be needed to "load" the FDO for an hour of testing. In designing the experiment, it was implicitly assumed that the FDOs' decisions on targets being forwarded to the BCS for simulated firing would not affect the btry FDC portion's measures of performance.

The *factors* for the FDO portion of the experiment were (1) FDO, (2) target type and subtype, (3) target size, (4) type of fire mission control, and (5) the initial ammunition load (**Figure 3**). The factors for the btry FDC portion of the experiment were (1) the number of simultaneous fire missions at the BCS, (2) the number of howitzers in the btry, and (3) the fire mission control ratios (**Figure 4**). The levels of each of these

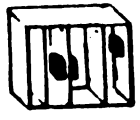
FIREPOWER CONTROL EXPERIMENT



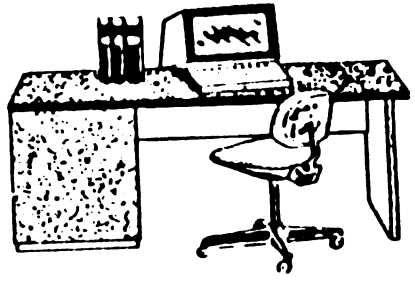
Battery Computer Unit



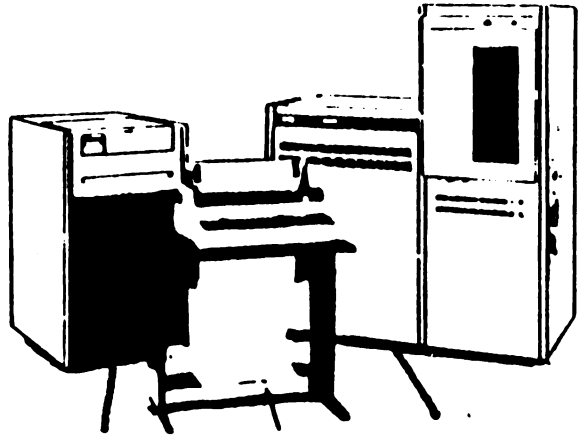
Bit Box



GDU Bit Box



FDO Terminal

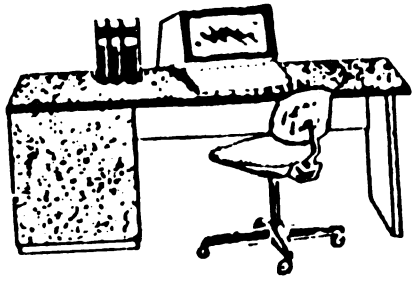


Interactive Computer System

- Ether-Net & Comm Simulator
- Master Control & Monitoring Software
- Data Logging

Target Lists
Target Acquisition Node Simulator
Battalion FDC Simulator
Gun Display Unit Simulator
FDO Display Software

Interactive Simulators & Special Software



Master Control & Data Display

FIGURE 1. Major Components of the Firepower Control Experiment.

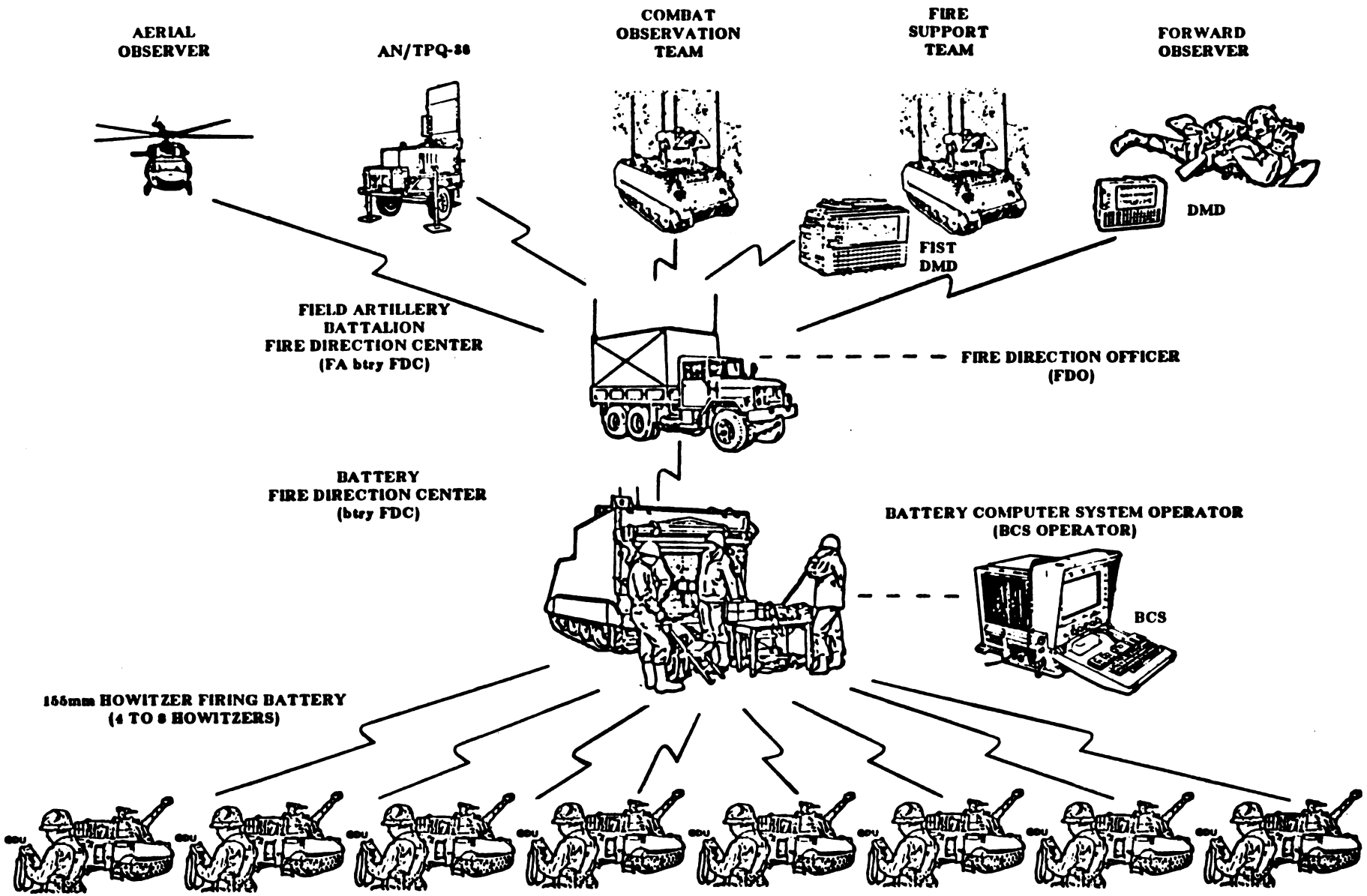


FIGURE 2. Field Configuration of the Firepower Control Experiment.

- **FDO**

  3 levels, i.e., 3 FDOs

- **TARGET TYPE AND SUBTYPE**

  10 levels, i.e., 10 different target descriptions

- **TARGET SIZE**

  2 levels, i.e., 2 sizes per target type and subtype

- **TYPE OF FIRE MISSION CONTROL**

  2 levels, i.e., adjust fire and fire-for-effect

- **INITIAL AMMUNITION LOAD**

  3 levels, i.e., 100%, 80%, or 60% of a basic load

Figure 3. Factors and Levels for the Fire Direction Officer Portion of the Firepower Control Experiment

- **NUMBER OF HOWITZERS IN A BATTERY**
  - 4 HOWITZERS
  - 6 HOWITZERS
  - 8 HOWITZERS
- **NUMBER OF SIMULTANEOUS MISSIONS**
  - 1 MISSION
  - 2 SIMULTANEOUS MISSIONS
  - 3 SIMULTANEOUS MISSIONS
- **CONTROL RATIO OF THE FIRE MISSIONS**
  - 2 ADJUST FIRE : 1 FIRE-FOR-EFFECT
  - 1 ADJUST FIRE : 2 FIRE-FOR-EFFECT

Figure 4. Factors and Levels for the Battery Fire Direction Center Portion of the Firepower Control Experiment

factors were selected as factors the BRL was interested in testing. *First*, for example, the BCS is only designed to handle up to 3 fire missions at time. *Second*, each BCS currently handles 6 howitzers in the field and future alternative considerations may have the BCS handle 4 or 8 howitzers. *Third*, fire mission control refers to how fire is directed on the target. For *all* Adjust Fire (AF) missions being sent to the BCS operator, a default of two "adjustments" (consisting of a total of two High Explosive rounds) were fired to better locate the target's position before the expenditure of the btry volleys, i.e., the Fire-for-Effect (FFE) portion of the fire mission. In the case of FFE missions, the observer has accurately located the target, and it is unnecessary to "adjust" before firing the btry volleys.

During the first week of testing, the BCS operator noticed anomalous behavior of the firing btry simulator, GUNSIM, in comparison to the actual tactical equipment. While GUNSIM was modified, the FDO portion of the experiment was run. As a result, these unexpected software problems precluded the completion of the design for the btry FDC portion of the Firepower Control Experiment. The remainder of this paper will focus on the appropriate method of analysis for the data collected and computed from this portion of the experiment.

III. Battery FDC Portion of the Experiment

1. Design Matrix and Measures of Performance

The intended design was three replications of a 3 x 3 x 2 factorial with the linear Howitzer x Mission interaction blocked by day (**Figure 5**). The purpose was to measure the effect of these factors and their interactions on the btry fire direction (FD) net's message traffic. Two different responses were computed to measure message traffic. The first, *net utilization*, is computed by dividing the total transmission time by the total time required to complete the simulated firing of the twelve fire missions associated with a treatment combination. The significance of the btry FD net's usage in the battlefield is that higher net usage increases the enemy's opportunity to detect the locations of the btry FDC and the 155mm howitzers. Presumably, detection would lead to enemy destruction of these important assets. The second, the *average number of messages per minute*, is computed by dividing the total number of messages for a particular treatment by the total time required. This indicates the number of times the tactical equipment must be turned on and off to transmit and receive messages. Both of these measures are important indicators of btry FD net usage when radios (rather than wire) will link the FA btry FDC and future semi-autonomous howitzer systems.

As previously mentioned, mid-experiment software problems did not permit the completion of this design. Subsequently, data collected under experimentally controlled conditions was only available for twelve of the eighteen treatment combinations of a single replication of this design, specifically, days 2 and 3 of the design matrix in **Figure 5**. This paper will focus on the analysis of the average number of messages per minute for these twelve treatment combinations.

Day	Hour	Number of Howitzers Per Battery	Number of Missions	AF:FFE Ratio
1	1	4	2	2:1
	2	4	2	1:2
	3	6	1	1:2
	4	6	1	2:1
	5	8	3	1:2
	6	8	3	2:1
2	1	8	1	1:2
	2	8	1	2:1
	3	4	3	1:2
	4	4	3	2:1
	5	6	2	2:1
	6	6	2	1:2
3	1	6	3	2:1
	2	6	3	1:2
	3	8	2	1:2
	4	8	2	2:1
	5	4	1	2:1
	6	4	1	1:2

Figure 5. Design Matrix for Each Replication of the Battery FDC Portion of the Firepower Control Experiment

2. Average Number of Messages Per Minute

Six different fixed format messages can be transmitted on the btry FD net, and each of these messages has a different purpose and fixed message format. Message types A, B, C and D correspond to messages associated with instructions from the BCS operator to the btry personnel via the GDU located at every howitzer, and message types E and F are the messages associated with polling between the BCS and GDUs (**Table 1**), i.e., requests and responses for the firing status of each howitzer. Before the body of each of these messages, a preamble (i.e., a continuous 1200 hertz sine wave) is transmitted for a specific time to allow the transmitting and receiving equipment to reach operating conditions. The minimum specification preamble for BCS and GDU messages, i.e., 250 milliseconds, was used for all message preambles on the btry FD net.[1] During the experiment, all message preambles on the btry FD net were fixed at 250 milliseconds.

Table 2 presents the average number of messages per minute for the twelve treatment combinations from the experiment. From scanning this table, the average number of messages for the treatment combination 6 howitzers, 3 simultaneous missions, and a 2:1 AF:FFE control ratio is low compared to the surrounding treatment combinations. A detailed investigation revealed that the busy BCS operator failed to act on the first several transmissions of a critical mission message, and for another mission, the operator sent an erroneous message creating approximately 4 minutes of net silence. The net result of these actions that the BCS operator was essentially only actively working with 2 of the 3 mission buffers. From this data, the BRL wanted to determine if the number of howitzers, simultaneous missions, control ratios or their interactions had a significant effect on this measure of performance, and solicited expert advice on the following proposed method of analysis and on suggestions for alternative methods of analysis.

IV. Analysis of Battery FDC Portion of the Experiment

1. Proposed: Cell Mean Estimation Procedure

The cell means model equation for the btry FDC experiment is

$$y_{ijkn} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijkn}$$

where

y_{ijkn} observation for control ratio level i, simultaneous mission level j, howitzer

[1] "External Interface Specification for Computer, Gun Direction CP-1317( )/GYK-29 Part of the Computer System, Gun Direction AN/GYK-29( ) (V)," United Technologies Corporation--Norden Division, Specification No. EL-CP-2678B-TF, 31 October 1981, 3.14.3.16, p. 55.

Table 1. Message Formats and Message Lengths

Message Code	Message	Transmitted by	Transmitted to	Length (in characters)
A	Common Data	BCS	GDUs	43
B	Common Special	BCS	GDUs	25
C	Individual Gun Order	BCS	GDU	27
D	Control	BCS	GDUs	13
E	Request	BCS	GDU	11
F	Response	GDU	BCS	11

Table 2. Average Number of Messages Per Minute Transmitted on the Battery Fire Direction Net

SIMULTANEOUS MISSION(S)	CONTROL RATIO (AF:FFE)	NUMBER OF HOWITZERS PER BATTERY		
		4	6	8
1	2:1	24.51		26.80
	1:2	25.06		25.89
2	2:1		34.99 *	36.10
	1:2		40.96	40.93
3	2:1	41.19	29.71 **	
	1:2	42.73	42.88	

*Based on 10 targets, not 12.

**FDO missed SHOT message and incorrectly sent an MTO message.

	level k, and observation n
μ	overall mean
α_i	effect of control ratio level i, i=1,2
β_j	effect of simultaneous mission level j, j=1,2,3
γ_k	effect of howitzer level k, k=1,2,3
$(\beta\gamma)_{jk}$	effect of blocking
$(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\alpha\beta\gamma)_{ijk}$	two- and three-way interactions
e_{ijkn}	error for observation y_{ijkn} which is distributed independently and normally with mean 0 and variance σ^2, i.e., N(0, σ^2)

This model is overparameterized for the btry FDC experiment since observations are missing for six cells. It was recommended[‡] that a cell mean estimation procedure using the basic linear model could be employed to estimate the six missing cells.[2]

Using this procedure, estimates for the missing cell means can only be made if the model is constrained by assuming one or more interactions are zero. The application of constraints, however, may not relate all missing cell means to the other observed cell means, and will yield one of two types of models: (1) *connected* models where all means of the missing cells are linearly estimable and any linear hypothesis on the cell means can be tested; and (2) *unconnected* models where not all missing cell means are estimable and the hypotheses of interest may still be tested. If one can justify the constraints necessary to produce a connected design and the constraints are valid, then stronger conclusions can be drawn; however, constraints should not be applied to just produce a connected design. Hocking also notes that there are varying degrees of connectedness, and the application of additional constraints increases the precision of missing cell mean estimates.

For the btry FDC experiment, the first reasonable constraint would be to assume that there is no three-way interaction Howitzer x Mission x Control Ratio, i.e., $(\alpha\beta\gamma)_{ijk} = 0$. Based on this assumption, the missing cell means, μ_{ijk}s, can be estimated by the following equation:

---

‡ Jock O. Grynovicki, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD.

2 Hocking, Ronald R., *The Analysis of Linear Models*, Monterey, CA: Brooks/Cole Publishing Company, 1985.

$$\mu_{ijk} - \mu_{i'jk} - \mu_{ij'k} + \mu_{i'j'k} = \mu_{ijk'} - \mu_{i'jk'} - \mu_{ij'k'} + \mu_{i'j'k'}$$

where

i,i'	$= 1,2;\ i \neq i'$	for the control ratio levels
j,j'	$= 1,2,3;\ j \neq j'$	for the simultaneous mission levels
k,k'	$= 1,2,3;\ k \neq k'$	for the howitzer levels

However, this constraint yields an unconnected model with none of the six missing cells being linearly estimable. Based on the design assumptions, another reasonable assumption would be that there is no Howitzer x Mission interaction, i.e., $(\beta\gamma)_{jk} = 0$, in addition to no Howitzer x Mission X Control Ratio interaction. Based on these two assumptions, the missing cell means can be estimated by the following formula:

$$\mu_{ijk} - \mu_{ij'k} = \mu_{ijk'} - \mu_{ij'k'}$$

These constraints provide estimates for the 6 missing cell means, and the associated single effective constraint is

$$\mu_{111} - \mu_{113} - \mu_{122} + \mu_{123} - \mu_{131} + \mu_{132} = 0\ .$$

Using these constraints, the missing cell means can be related to the observed cell means as follows:

$$\mu_{112} = \mu_{113} - \mu_{123} + \mu_{122} = 26.80 - 36.10 + 34.99 = 25.69\ ,$$

$$\mu_{212} = \mu_{213} - \mu_{223} + \mu_{222} = 25.89 - 40.93 + 40.96 = 25.92\ ,$$

$$\mu_{121} = \mu_{123} - \mu_{113} + \mu_{111} = 36.10 - 26.80 + 24.51 = 33.81\ ,$$

$$\mu_{221} = \mu_{223} - \mu_{213} + \mu_{211} = 40.93 - 25.89 + 25.06 = 40.10\ ,$$

$$\mu_{133} = \mu_{131} - \mu_{111} + \mu_{113} = 41.19 - 24.51 + 26.80 = 43.48\ ,$$

$$\mu_{233} = \mu_{231} - \mu_{211} + \mu_{213} = 42.73 - 25.06 + 25.89 = 43.56\ .$$

Table 3 provides the estimates for the 6 missing treatment combinations along with the 12 treatment combinations from the experiment. By using the values in this table, an analysis of variance (ANOVA) was performed and is provided in **Table 4**.

Table 3. Observed and Estimated Average Number of Messages Per Minute on the Battery Fire Direction Net

SIMULTANEOUS MISSION(S)	CONTROL RATIO (AF:FFE)	NUMBER OF HOWITZERS PER BATTERY		
		4	6	8
1	2:1	24.51	25.69†	26.80
	1:2	25.06	25.92†	25.89
2	2:1	33.81†	34.99 *	36.10
	1:2	40.10†	40.96	40.93
3	2:1	41.19	29.71 **	43.48†
	1:2	42.73	42.88	43.56†

†Estimated.

*Based on 10 targets.

**FDO missed SHOT message and incorrectly sent an MTO message.

Table 4. ANOVA on the Effect of the Factors on the Average Number of Messages Per Minute

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	F RATIO
Howitzers	2	22.8559	11.4280	0.32
Missions	2	760.0711	380.0356	10.63
Control Ratio	1	55.6864	55.6864	1.56
Howitzers X Control Ratio	2	21.0985	10.5493	0.30
Missions X Control Ratio	2	29.4570	14.7285	0.41
Pooled Error	2	71.4755	35.7378	
Total	11	960.6444		

$F_{2,2,\alpha=0.05} = 19.00$
$F_{1,2,\alpha=0.05} = 200$

One should note that the Howitzer x Mission x Control Ratio and the Howitzer x Mission interactions were pooled for error since it was assumed that these interactions were not significant in the cell mean estimation procedure. From **Table 4**, one concludes that none of the main effects or two-way interactions are significant at $\alpha = 0.05$ based on the assumptions of no Howitzer x Mission x Control Ratio and Howitzer x Mission interactions. If either of these assumptions are incorrect then the pooled error is biased; using a biased error value lowers the F ratios and can result in factors or their interactions being statistically insignificant.

A consequence of using this cell mean estimation procedure is that one-third of an unreplicated design was estimated based on two assumptions. The resulting ANOVA failed to detect any significant main effects or interactions despite seemingly differences between certain levels of the factors. Additional experimentation would be required to test if the assumptions associated with the cell mean estimation procedure were justified and more confidently determine the conclusions of no significant main effects or interactions. In lieu of additional testing and this cell mean estimation procedure, the panel recommended paired t tests.

2. Suggested: Paired t Tests

Based on the panel's suggestions, paired t tests were performed on the data to test for a significant difference between the means, i.e., the average number of messages, of the levels of each factor assuming no interactions. The null hypothesis, H_0, for each test was that there was no difference between the means of two levels of a factor versus the alternative hypothesis, H_1, that the mean for a given level exceeded another. This one sided alternative hypothesis was not rejected only if the difference between the means was significantly greater than zero.

An overall paired t test was computed for the difference between the 1:2 and 2:1 AF:FFE control ratio levels under the same howitzer and mission levels, i.e., 6 paired differences. H_0 was not rejected since the computed t statistic at a significance level of $\alpha = 0.05$ was close to but did not exceed the tabled t value $t_{5\ df} = 2.015$. This was a bit surprising since only "one GDU's worth" of messages are requested and transmitted for each "adjustment", and each "adjustment" requires "one round's worth" of time. Thus, one would expect the average number of messages per minute for a 2:1 AF:FFE control ratio to be lower than a 1:2 AF:FFE control ratio.

In addition to this paired test, two other paired t tests were computed; one with the pairs by howitzer level and the other with the pairs by mission level. In computing these tests, two difference pairs were obtained for each howitzer and mission level by computing the difference across a specific control ratio level. H_0 was rejected if the computed t statistic exceeded the tabled t value at a significance level of $\alpha = 0.05$, i.e., $t_{1\ df} = 6.314$. Only two of the six null hypotheses could not be accepted at a significance level of $\alpha = 0.05$. *First*, H_0 was not accepted between 1 and 3 simultaneous missions for 4 howitzers. This supports the expectation that as the number of simultaneous missions increases more missions are handled in a shorter time, i.e., the average number of messages per minute increases. However, no significant difference was

detected at $\alpha = 0.05$ between 2 and 3 simultaneous missions for 6 howitzers, or between 1 and 2 simultaneous missions for 8 howitzers. *Second*, H_0 was not accepted between 6 and 8 howitzers handling 2 simultaneous missions. This also supports the expectation that as the number of howitzers increases more howitzers are sending and receiving messages in essentially the same amount of time, i.e., increasing the average number of messages per minute. Similarly, no significant difference was detected at $\alpha = 0.05$ between 4 and 8 howitzers handling 1 mission, or between 4 and 6 howitzers handling 3 simultaneous missions.

V. Conclusions

The data collected from the btry FDC portion of the Firepower Control Experiment suggests that different procedures should be considered to reduce btry FD net usage when radios will link the FA btry FDC and future semi-autonomous howitzer systems. The paired t tests on the average number of messages per minute detected a significant difference at $\alpha = 0.05$ between 1 and 3 simultaneous missions for 4 howitzers, and 6 and 8 howitzers handling 2 simultaneous missions. Although this supports our initial design assumptions that the number of howitzers and simultaneous missions significantly affect the usage of the btry FD net, it also clearly points out that completing the intended design could have produced more confident conclusions.

COMMENTS BY PANELISTS DR. KAYE BASFORD AND PROFESSOR W. T. FEDERER ON THE FOLLOWING ARTICAL

Analysis of an Incomplete block design of experiments by

Wendy A. Winner
and Jill H. Smith
U.S. Army Ballistic Research Laboratory

Kaye Basford: Because full data were not collected on the original designed experiment, I suggest that it be analysed in a much simpler way. For instance, simple t tests or non-parametric tests could be used to compare fire mission control ratios over all howitzer and mission levels. Although not giving the detail of the planned analysis, it should allow some information to be obtained from the data collected.

W.T. Federer: The resulting design is a two-thirds fraction of a 2×3^2 factorial of the following nature:

	a_0 1:2			a_1 2:1		
	b_i			b_i		
c_j	1	2	3	1	2	3
4	x		x	x		x
6		x	x		x	x
8	x	x		x	x	

where x denotes combination present and blank denotes combination absent. In the above fraction main effects will be estimable as well as 12-(1+1+2+2) = 6 degrees of freedom for interactions. These 6 degrees of freedom are A x B (2 d.f.), A x C (2 d.f.) B (linear) x C (linear) (1 d.f.), and A x B (linear)

x C (linear) (1 d.f.). Unless it were known that one or more of the degrees of freedom for interaction represented experimental error, no error mean square would be available for testing the significance of the effects. For no available error mean square, it is suggested that use be made of Cuthbert Daniel's half normal plot procedure to ascertain which of the eleven treatment simple degrees of freedom sums of squares were alike and which were different. If the smaller contrasts responding similarly could be considered as possible candidates for no treatment effects; then an error term can be obtained using Cuthbert Daniel's procedure (see e.g. S.A. Krane (1963) "Half normal plots for multi-level factorial experiments", Proc. Eighth Conf. Design Expt. Army Res. Dev. Testing, pp 261-285).

A HEURISTIC APPROACH TO POST - HOC COMPARISONS FOR SIGNIFICANT INTERACTIONS - A SIMPLIFIED NOTATION

Eugene Dutoit
U.S. Army Infantry School
Fort Benning, Georgia

ABSTRACT:

The omnibus F ratio test used in analysis of variance is used to determine if any of the main or interaction effects are statistically significant. Customarily, various techniques are used for performing post-hoc comparisons on the statistically significant main effects. The purpose of this paper will be to present a heuristic approach for post-hoc procedures on the significant interaction effects. These procedures will use the conventional graphical methods to show the overall interaction effect and then apply conservative methods to detect the significant components of the overall interaction. The paper will develop graphical and notational method for decomposing a complex interaction into its significant components for further analysis. Examples will be given for a two-way design with variables at two and more levels.

ACKNOWLEDGEMENT: The author wishes to thank Dr. John Tukey for his suggestion to use a Bonferroni contrast in addition to the Scheffe method. The Bonferroni method will be calculated for each of the examples in this paper and the results compared with those obtained by using Scheffe contrasts.

SECTION 1 (A TWO-WAY ANOVA PROBLEM)

Consider the following two way ANOVA problem obtained from Ostle. The dependent variable is the yield in soy beans (bushels/acre). The raw data and the resulting ANOVA are presented in the table below.

TABLE I

TWO WAY ANOVA

Date of Planting	EARLY				LATE			
Fertilizer	Cl	Aero	Na	K	Cl	Aero	Na	K
	29	29	28	29	30	33	30	33
	37	29	27	28	32	31	33	32
	33	31	26	28	32	31	33	32
	33	29	29	32	31	34	34	29
$\bar{X}$ =	33	29.5	27.5	29.25	31.25	32.25	32.5	31.5

Source	DF	SS	MS	F	
Day of Planting	1	34.031	34.031	10.44	*(Sig)
Fertilizer	3	20.594	6.865	2.11	(Not Sig)
Interaction	3	47.344	15.781	4.84	**(Sig)
Error	24	78.250	3.260		

* $F_{1, 24}(.05) = 4.26$

** $F_{3, 24}(.05) = 3.01$

The usual Scheffe contrast ($\hat{\psi}$) is formed:

$$\hat{\psi} = \Sigma\, AiXi, \text{ where } \Sigma\, Ai = 0. \quad (1)$$

The critical difference (CD) is calculated as

$$CD = (S)(SE_{\psi}), \text{ where} \quad (2)$$

$$S = [(\text{number of treatment levels} - 1)\, F(\text{critical}, \alpha)]^{1/2} \quad (3)$$

For a simple pairwise contrast between two means:

$$SE_{\hat{\psi}} = \left[\frac{(2)\ (MS_{error})}{N\ group} \right]^{1/2} \qquad (4)$$

The contrast is statistically significant if

$$|\hat{\psi}| > CD \qquad (5)$$

The above procedure is applied to the data in this 2 way ANOVA against the main effects of day of planting and fertilizer type.

Day of Planting Effect:

Average yield for early planting (Early) = 29.81 bushels/acre

Average yield for late planting (Late) = 31.88 bushels/acre

The contrast ($\hat{\psi}$) is:

$\hat{\psi} = \bar{X}_{late} - \bar{X}_{early} = 31.88 - 29.81 = 2.07$ bushels/acre

$S = [(\#treat\ levels - 1)\ F\ (critical)]^{\frac{1}{2}} = [(1)(4.26)]^{\frac{1}{2}} = 2.06$

$$SE_{\hat{\psi}} = \left[\frac{(2)(MS_{error})}{N\ group} \right]^{1/2} = \left[\frac{(2)(3.26)}{16} \right]^{1/2} = .64$$

$CD = (S)(SE_{\hat{\psi}}) = 1.32$

Since $\{|\hat{\psi}| = 2.07\} > \{CD = 1.32\}$; the contrast is significant at the 5% level of significance. Of course, the ANOVA table results already indicated this effect. The Scheffe calculation was presented to illustrate/review the procedure.

Fertilizer Effect

Average yield for chlorine (Cl) = 32.125 bushels/acre
Average yield for aero (Aero) = 30.875 bushels/acre
Average yield for sodium (Na) = 30.000 bushels/acre
Average yield for potassium (K) = 30.375 bushels/acre

There are $({}_4C_2) = 6$ possible pairwise contrasts. These are given in the table below:

Table 2
Pairwise differences $|\hat{\psi}|$ for four Fertilizers
(bushels / acres)

	Cl	Aero	Na	K
Cl	-	1.25	2.125	1.75
Aero		-	.875	.50
Na			-	.375
K				-

$$S = [(\#\text{treat levels} - 1)\, F_{\text{critical},\alpha}]^{\frac{1}{2}} = [(3)(3.01)]^{\frac{1}{2}} = 3.005$$

$$SE_{\hat{\psi}} = \left[\frac{(2)(\text{MS error})}{N \text{ group}}\right]^{1/2} = \left[\frac{(2)(3.26)}{8}\right]^{1/2} = .90$$

$$CD = (S)(SE_{\hat{\psi}}) = 2.705$$

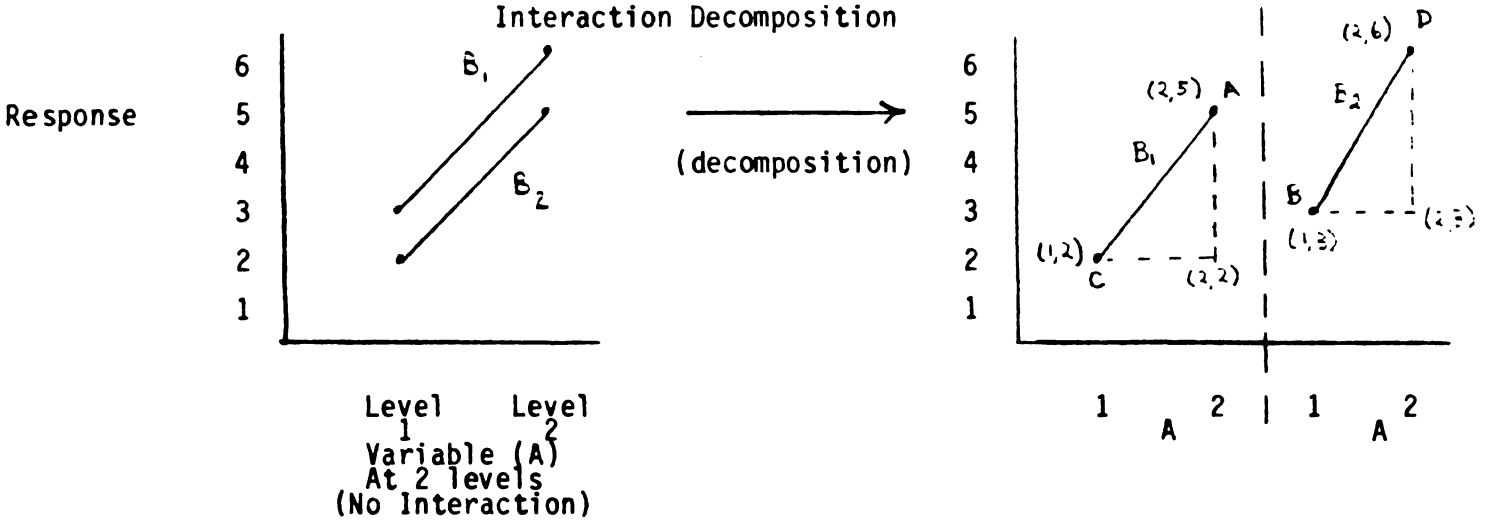
Note that no value of $|\hat{\psi}|$ in table 2 above is greater than the CD. The ANOVA table furnished the same information as the above calculation. Now let us examine the significant interaction effect as shown in table 1.

Interaction Effect

This section will develop a way to examine the interaction effects. Consider the diagram below (Figure 1). In this example, there are two factors A and B, each factor at two levels. The parallel lines indicate there is no interaction. The total interaction can be decomposed into two seperate graphs for each level of factor B. The decomposition just makes it visually easier to calculate the slopes for each of the two lines.

Figure 1

Interaction Decomposition



Paying attention to the right side of the arrow in the above figure, the slopes for B_1 and B_2 respectively are:

$$\text{Slope} = \frac{\Delta Y}{\Delta X} = \frac{5-2}{2-1} = \frac{6-3}{2-1} = 3$$

or alternatively

$$\text{Slope} = A - C = D - B$$

This expression can be written as

$$(A + B) - (C + D) = 0$$

This identity forms the basis for writing the contrast ($\hat{\psi}$). If no interaction exists, then the contrast can be written as:

$$\hat{\psi} = (A + B) - (C + D) = 0 \qquad (6)$$

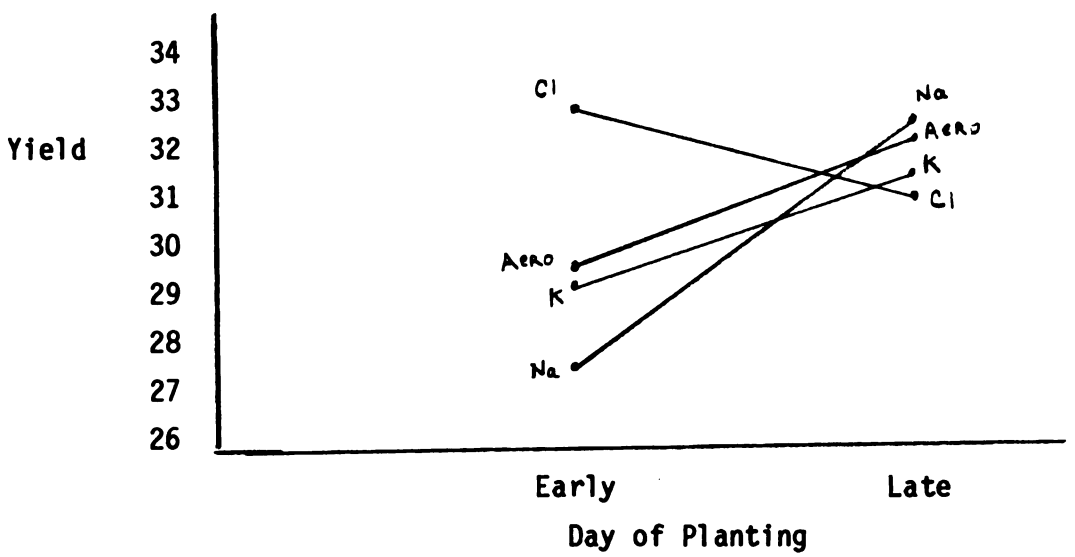
This contrast will be used to examine the significant interaction term in Table 1. The arithmetic means for the day of planting and fertilizer interactions are given below in Table 2.

Table 2
Interaction Data
(Arithmetic Means)
(Yield; Avg Bushels per Acre)

Early Cl	=	33	Late Cl	=	31.25
Early Aero	=	29.5	Late Aero	=	32.25
Early Na	=	27.5	Late Na	=	32.50
Early K	=	29.25	Late K	=	31.50

The above interaction effect can be plotted in the usual way. This is shown below in Figure 2.

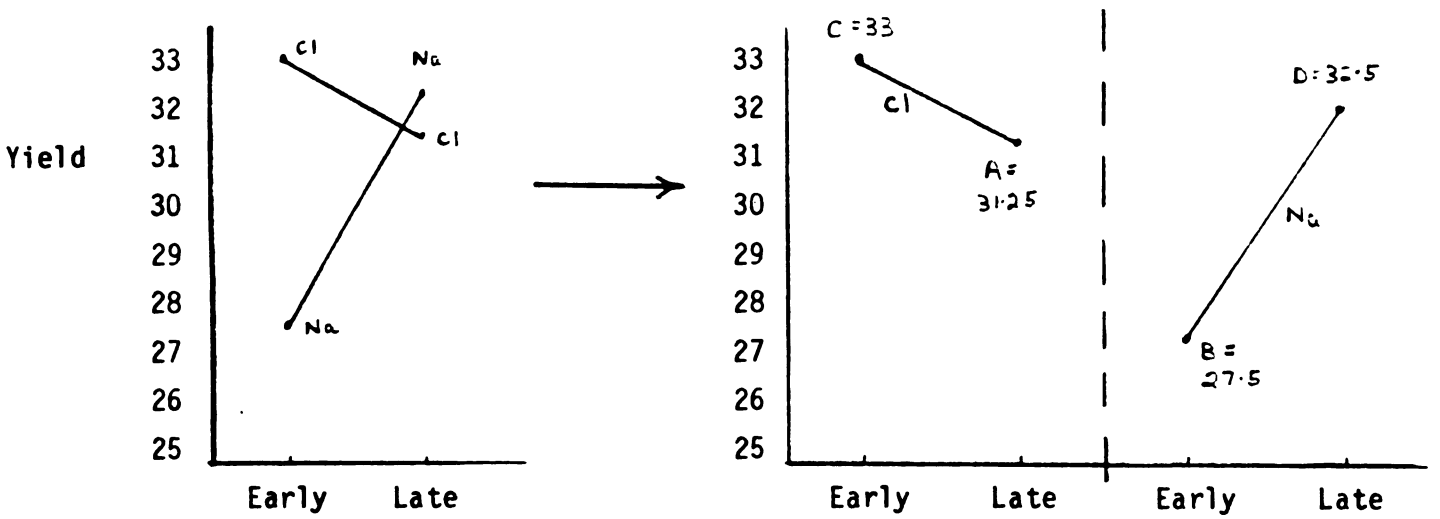
Figure 2
The Total Interaction



Examination of figure 2 suggests that the significant interaction shown in Table 1 is driven by the effect of the chlorine fertilizer as it interacts with the other three fertilizers. Figure 3 below gives the interaction decomposition (using the notation of Figure 1) of the Na, Cl component of the total interaction.

Figure 3

The Na, Cl Component



Using equation (6), the interaction component contrast can be calculated:

$$\hat{\psi} = (\bar{X}_A + \bar{X}_B) - (\bar{X}_C + \bar{X}_D)$$

$$\hat{\psi} = (31.25 + 27.5) - (33 + 32.5)$$

$$|\hat{\psi}| = 6.75$$

In order to determine if this component of the total interaction is statistically significant (is $|\hat{\psi}|$ significantly greater than zero?), the Scheffe critical difference (CD) will be calculated using the methods reflected in equations (2) through (5).

$$SE^2_{\hat{\psi}} = MS_E \Sigma a_i^2/n_i \ .$$

$$SE^2_{\hat{\psi}} = MS_E[1/n_A + 1/n_B + 1/n_C + 1/n_D] \ .$$

In this case all n_s are equal.

$$SE^2_{\hat{\psi}} = MS_E[4/n]$$

Therefore

$$SE^2_{\hat{\psi}} = (3.26)(4/4)$$

$$SE_{\hat{\psi}} = 1.81 \ .$$

The Other Component:

$$S^2 = \underbrace{(df_{t_1})\,(df_{t_2})}_{\text{df for interaction}} \quad \underbrace{F_{(df_{t_1})\,(df_{t_2}),\,(df_{Error}),\,\alpha}}_{\text{critical value}}$$

$$S^2 = (df_{\text{interaction}})\ F_{(\text{critical},\alpha)}$$

In this case

$$S^2 = (3)\ F_{3,\,24\ (.05)} = (3)\ (3.01) = 9.03$$

$$S = 3.00$$

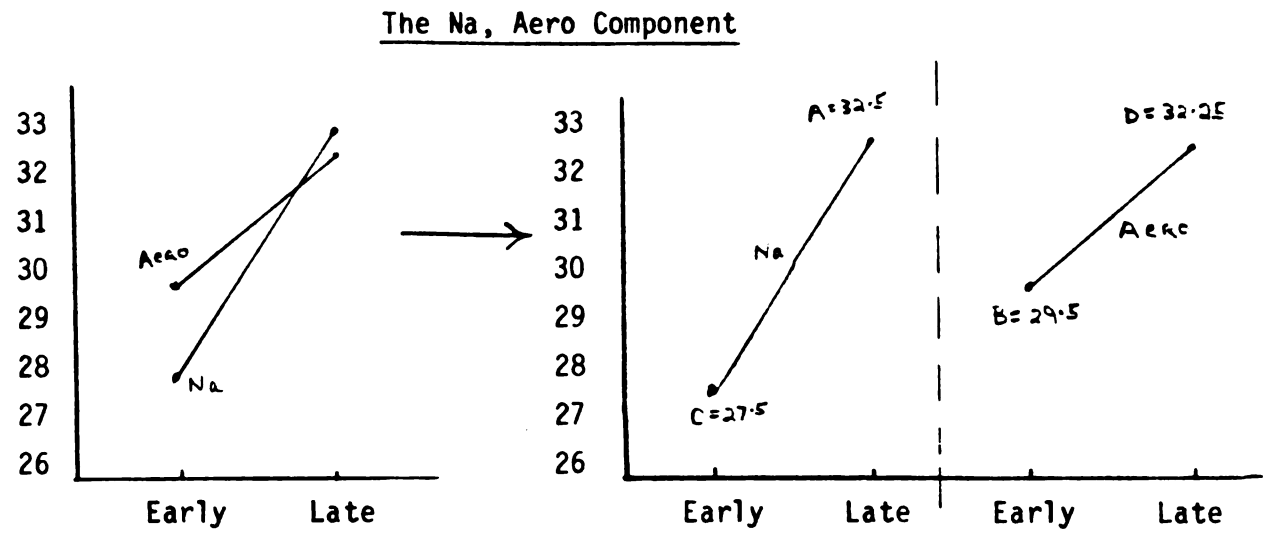
The Critical Difference

$$CD = (SE_{\hat{\psi}})\ (S) = (1.81)\ (3) = 5.43$$

Since $\{|\hat{\psi}| = 6.75\} > \{CD = 5.43\}$ the day of planting, Na/Cl interaction component of the total interaction is significant.

Figure 4 shows the interaction decomposition of the Na, Aero component of the total interaction.

Figure 4



The lines are not exactly parallel but are the differences in slopes statistically significant?

The contrast is:

$\hat{\psi} = (\bar{X}_A + \bar{X}_B) - (\bar{X}_C + \bar{X}_D) = (32.5 + 29.5) - (27.5 + 32.25)$

$\hat{\psi} = 2.25$

The value for the critical difference (CD) is still 5.43.

Since $\{\hat{\psi} = 2.25\} < \{CD = 5.43\}$; the day of planting, Na/Aero component of the total interaction is not significant.

In this example problem there are $({}_4C_2)$ or 6 pairwise components that make up the total interaction. The results of these six components are summarized in Table 3.

Table 3

Component Summary α = .05 CD = 5.43		
Component	$\lvert\hat{\psi}\rvert$	Results
Na/Cl	6.75	Sig
Na/Aero	2.25	NS
Na/K	2.75	NS
Cl/Aero	4.50	NS
Cl/K	4.00	NS
Aero/K	.5	NS

It should be noted that although only one of the components of the total interaction was found to be statistically significant (α = .05), the chlorine fertilizer effect was involved with the largest values of $\hat{\psi}$.

The results of the Bonferroni method will now be compared to the results obtained from the Scheffe method used so far in this paper. The values for the contrast ($\hat{\psi}$) and the SE are calculated the same way as for the Scheffe methods. The $\lvert\hat{\psi}\rvert$ is significant if:

$$\lvert\hat{\psi}\rvert > (t_{\alpha/2p}, \nu)(SE_{\hat{\psi}}) \qquad (7)$$

where: p is the number of components or contrasts examined in the total interaction.

ν is the degrees of freedom for the error term.

$t_{\alpha/2p}$ is then obtained from tables of the critical values for the Bonferroni t (Milliken and Johnson).

The Bonferroni critical difference (BCD) is calculated as:

$$BCD = (t_{\alpha/2p}, \nu)(SE_{\hat{\psi}}) \quad . \tag{7A}$$

In this example:

$\alpha = .05$

$p = 6$ possible contrasts

$\nu = 24$

therefore $(t_{.05/2p}, \nu = 24) = 2.88$

and $SE_{\hat{\psi}} = 1.81$.

therefore the BCD = 5.21.

Referring to Table 3, it is apparent that this 4% decrease in critical difference (5.43 versus 5.21) does not change the decision regarding the significant component of the total interaction for this particular example.

Section 2 (A Three-Way ANOVA Problem).

The following example will expand the discussion of section 1 to a three way ANOVA. The dependent variable is the time (seconds) required by a blind rat to run a maze. The independent variables are:

1) When the rat was blinded (early or late in life).

2) Intelligence (bright, mixed, dull).

3) Movement (free (F) or restrained (R))

The data and the resulting ANOVA tables are shown below:

Table 4

Three Way ANOVA

	Early Blinded						Late Blinded					
	Bright		Mixed		Dull		Bright		Mixed		Dull	
	F	R	F	R	F	R	F	R	F	R	F	R
	27	55	130	140	55	132	90	105	61	65	140	142
	45	81	120	150	76	96	120	110	82	80	99	96
$\bar{X}$ =	36	68	125	145	65.5	114	105	107.5	71.5	72.5	119.5	119

Source	df	MS	F
Time of Blindness (B)	1	287.04	.83 NS
Intelligence (I)	2	1652.05	4.76 (Sig 5%)
Environment (E)	1	1785.38	5.15 (Sig 5%)
B x I	2	7638.79	22.02 (Sig 5%)*
B x E	1	1584.37	4.57 NS
I x E	2	91.12	.26 NS
B x I x E	2	115.88	.33 NS
Error	12	346.88	
	23		

$F_{2,\ 23(.05)} = 3.89$

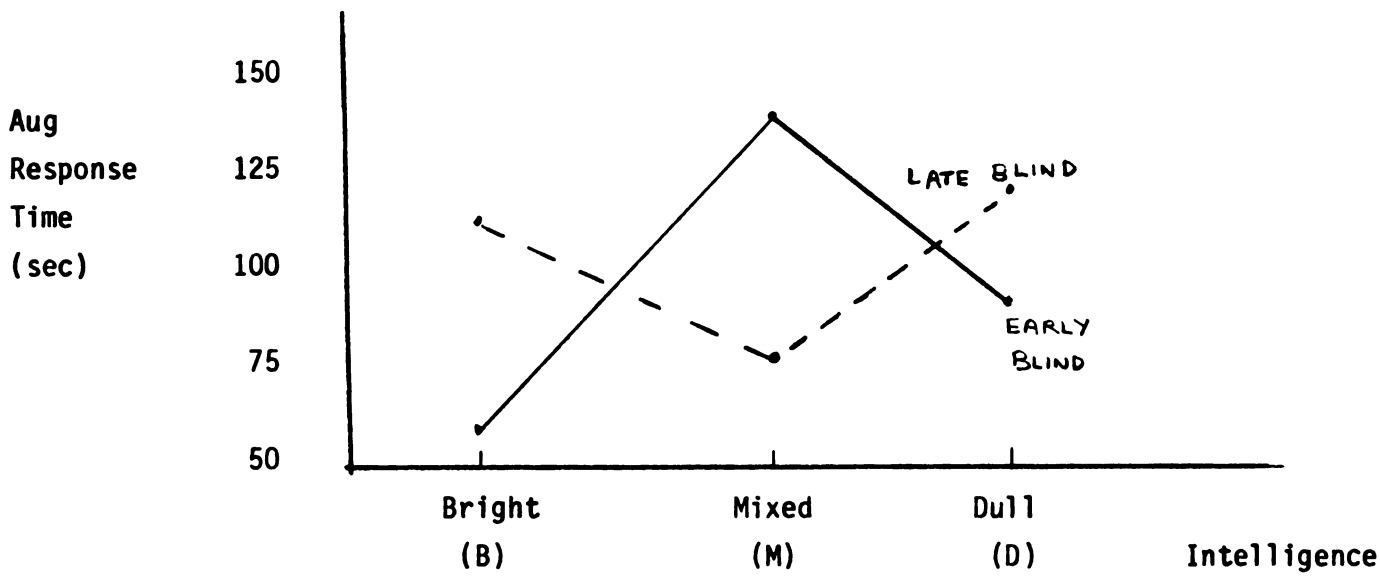
At this point only the significant interaction (time of blindness, intelligence) will be examined in detail. The data for this particular interaction are given in Table 5.

Table 5

Interaction Data (Arithmetic Means) (Dependent variable; time in seconds)	
Early Blind; Bright = 52 sec	Late Blind; Bright = 106.25 sec
Early Blind; Mixed = 135 sec	Late Blind; Mixed = 72 sec
Early Blind; Dull = 89.75 sec	Late Blind; Dull = 119.25 sec

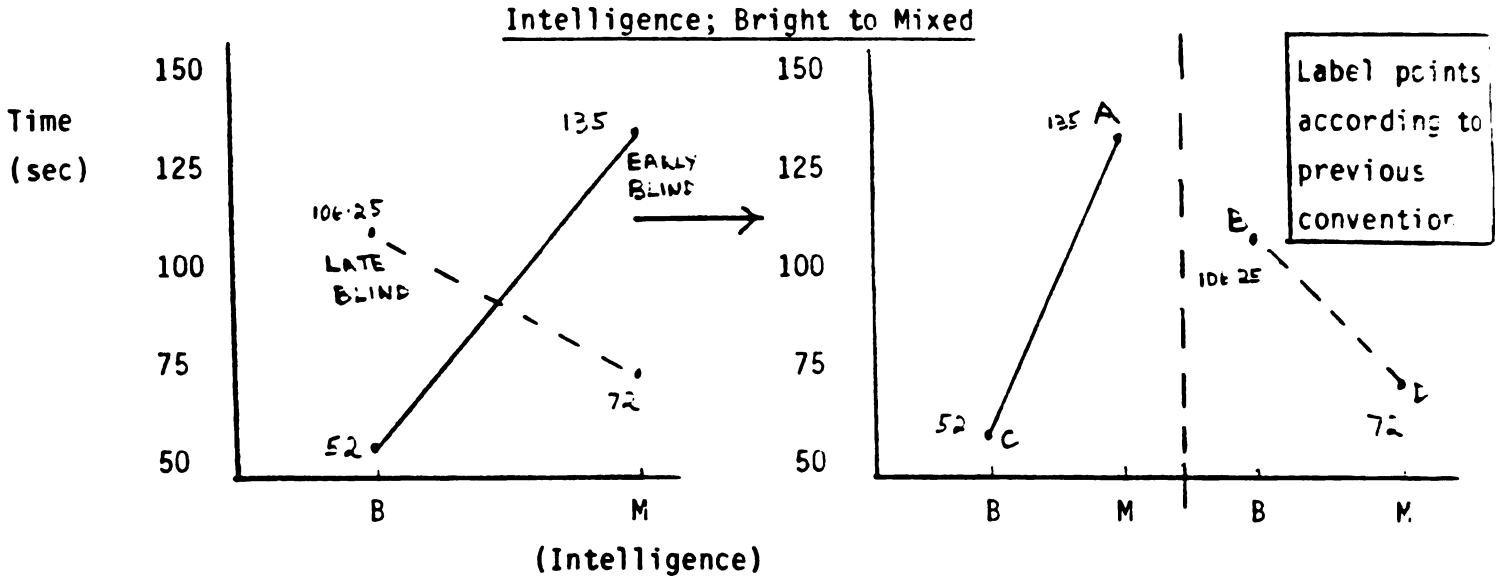
The plot of the interaction is given as figure 5.

Figure 5
The Total Interaction



The total interaction will be decomposed in the same manner as shown in section 1. Thereare($_3C_2$) or 3 components to examine. The components of {Bright to Mixed Intelligence} and of {Mixed to Dull Intelligence} appear to be significant. The third component {Bright to Dull} is probably not significant.

Figure 6 shows the bright to mixed intelligence component.



By labeling the points according to the previous convention and using equation (6), the interaction component contrast can be calculated:

$\hat{\psi} = (\bar{X}_A + \bar{X}_B) - (\bar{X}_C + \bar{X}_D) = (135 + 106.25) - (52 + 72)$

$\hat{\psi} = 117.25$

In this case

$$SE^2_{\hat{\psi}} = M_E \Sigma a_i^2/n_i$$

or $SE^2_{\hat{\psi}} = MS_E(4/n)$ which is the same as in section 1. The number of observations for each cell (n) is 4 therefore:

$SE^2_{\hat{\psi}} = 346.88(^4/_4) = 346.88$

$SE_{\hat{\psi}} = 18.62$

The other component (S^2) is:

$S^2 = (df_{interaction})\, F_\alpha = (2)\,(3.89) = 7.78$

$S = 2.79$

The (CD) is:

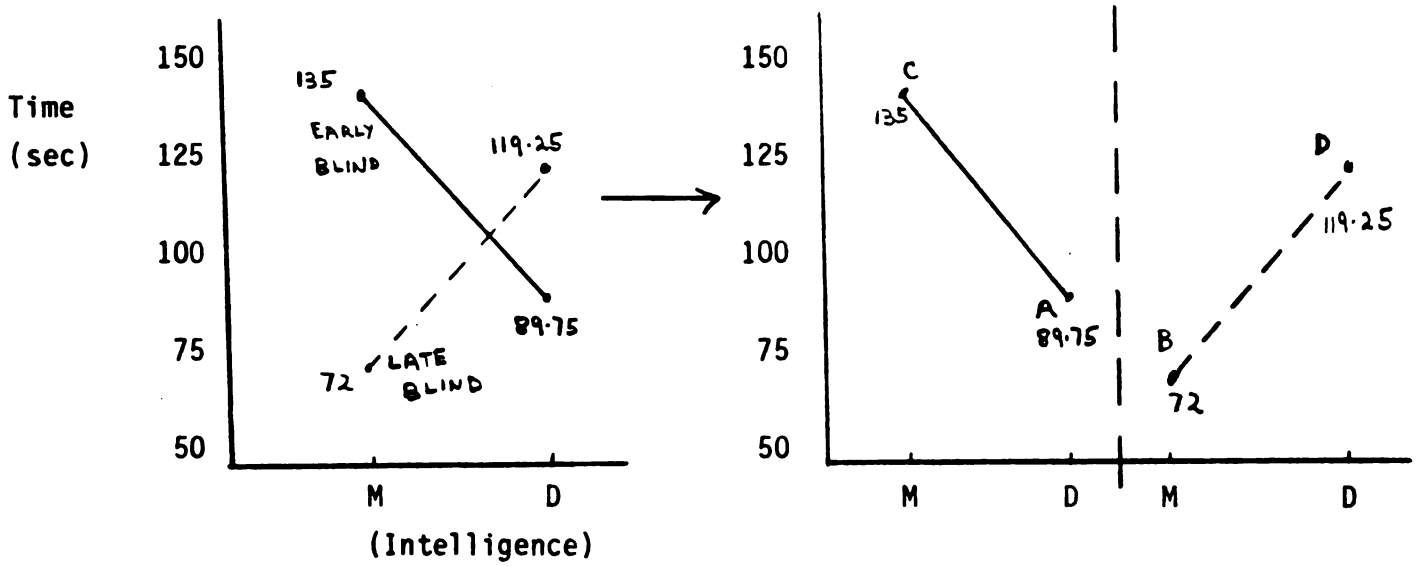
$CD = (S)\,(SE_{\hat{\psi}}) = 51.95$

since $\{|\hat{\psi}| = 117.25\} > \{CD = 5.95\}$, this component of the interaction [Intelligence; bright to mixed] is statistically significant.

Figure 7 gives the mixed to dull intelligence component.

Figure 7

Intelligence; mixed to dull



The contrast for this component is:

$$\hat{\psi} = (\bar{X}_A + \bar{X}_B) - (\bar{X}_C + \bar{X}_D) = (89.75 + 72) - (135 + 119.25)$$

$$\hat{\psi} = -92.5$$

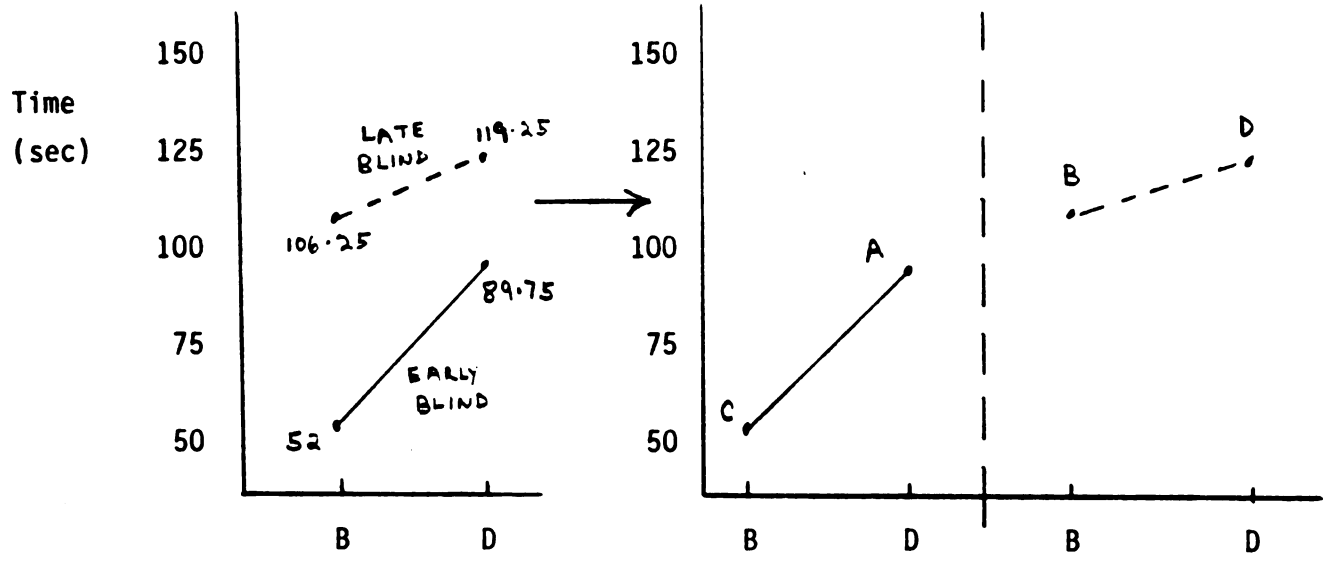
The CD is still equal to 51.95

Since $\{|\hat{\psi}| = 92.5\} > \{CD = 51.95\}$, the mixed to dull component of the interaction is statistically significant. This was expected.

Finally, Figure 8 shows the bright to dull component.

Figure 8

Intelligence; Bright to Dull



$$\hat{\psi} = (\bar{X}_A + \bar{X}_B) - (\bar{X}_C + \bar{X}_D) = (89.75 + 106.25) - (52 + 119.25)$$

$$\hat{\psi} = 24.75$$

Note that $\{\hat{\psi} = 24.75\} < \{CD = 51.95\}$, therefore this component of the interaction is not statistically significant.

In this problem, 2 out of 3 total pairwise interaction components were significant at a 5% level of significance.

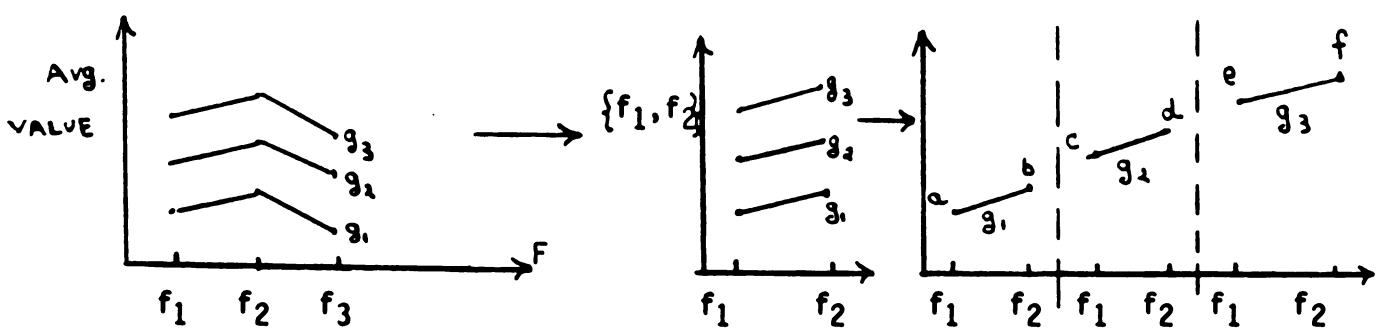
The Bonferroni method will be applied to this problem. SE is the same value (18.62), P is equal to 3 and n is 12. Therefore the Bonferroni t ($\alpha = .05$) is equal to 2.78. The Bonferroni critical difference (BCD) is (18.62) (2.78) or 51.76. This is only slightly less than the Scheffe CD of 51.95. There are no differences in the decision regarding significant components between the two methods for this particular example.

Section 3 (Interaction Where Both Factors (F, G) Have More Than Two Levels)

Consider the case where factor F has three equally spaced levels $\{f_1, f_2, f_3\}$ and factor G has levels $\{g_1, g_2, g_3\}$. Consider Figure 9 below which shows the decomposition of the total interaction in the interval for factor F in $\{f_1, f_2\}$.

Figure 9

Two Factors at Three Levels



Given that no interaction exists; is the "chaining notation" discussed in sections 1 and 2 of this paper true in this situation? Following the notation/convention discussed earlier:

$$\hat{\psi} = (\bar{X}_b + \bar{X}_c) - (\bar{X}_d + \bar{X}_a) + (\bar{X}_d + \bar{X}_e) - (\bar{X}_f + \bar{X}_c) = 0 \qquad (8)$$

The demonstration is simple. Given that no interaction exists, then $(\bar{X}_b - \bar{X}_a) = (\bar{X}_d - \bar{X}_c) = (\bar{X}_f - \bar{X}_e)$.

Equation (8) can be written without brackets:

$$\hat{\psi} = \bar{X}_b + \bar{X}_c - \bar{X}_d - \bar{X}_a + \bar{X}_d + \bar{X}_e - \bar{X}_f - \bar{X}_c$$

Re-arranging terms:

$$\hat{\psi} = \bar{X}_b - \bar{X}_a - \bar{X}_d + \bar{X}_c + \bar{X}_d - \bar{X}_c - \bar{X}_f + \bar{X}_e$$

Inserting brackets

$$\hat{\psi} = (\bar{X}_b - \bar{X}_a) - (\bar{X}_d - \bar{X}_c) + (\bar{X}_d - \bar{X}_c) - (\bar{X}_f - \bar{X}_e)$$

Since all terms in brackets are equal to each other, it follows that:
$\hat{\psi} = 0$ if no interaction is present.
Note that the $\Sigma a_i = 0$ for the contrast.
In this case:

$$SE^2_{\hat{\psi}} = MS_{E\Sigma a_i^2/n_i} .$$

If n_i are equal to n, then

$$SE^2_{\hat{\psi}} = \frac{MS_E(8)}{n}$$

In the case of the Scheffe methodology:

$$S^2 = (df_{interaction})\, F_\alpha = (2 \times 2)\, F_\alpha = 4F_\alpha$$

or,

$$S = 2\sqrt{F_\alpha}$$

These values for (SE , S) could be the same for examining interaction components in the interval $\{f_2, f_3\}$ and $\{f_1, f_3\}$.

Summary

This notation (chaining) can be applied for two factor interactions where the factors are at any number of levels. Each linear component of the total interaction can be examined to determine which component(s) contributed to the overall significant interaction. As the examples presented in this paper show, a statistically significant interaction (per the omnibus F test) does not imply that all components of the total interaction are statistically significant. The post-hoc analysis of the interaction should lead to improved insights about the data just as these methods aid in the analysis of the main effects. Both Scheffe and Bonferroni methods were applied to the example data. No differences were made in the decision concerning which components of the interaction were significant and the differences between the "critical differences" were small. It should be noted that these comparisons are based on just two examples using a two-way ANOVA. The differences may become more apparent for more complex designs.

REFERENCES

1. Milliken, G. and Johnson, D. Analysis of Messy Data (Notes), 1983.
2. Ostle, B. Statistics In Research (Second Edition), Iowa State University Press, 1963.

STATISTICAL EVALUATION OF DESERT INDIVIDUAL CAMOUFLAGE COVERS (ICC) BY GROUND OBSERVERS

George Anitole and Ronald L. Johnson
U.S. Army Belvoir, Research, Development and Engineering Center
Fort Belvoir, Virginia 22060

Christopher J. Neubert
U.S. Army Engineer School
Fort Belvoir, Virginia 22060

ABSTRACT

The ICC is a personal camouflage net for soldiers which will be useful for patrols, snipers, and ambush situations. This study determined whether the ICC should have large or small Hogan incisions, and what color(s) best blended with the desert backgrounds. Ten U.S. Marines and two civilians subjectively evaluated seventy-four ICCs (thirty-seven different colors half large and half small Hogan incisions) at five desert sites. The ICCs were ranked in groups of six, selecting four at a time, to reduce the number to the final six colors with associated incisions. The final six were subjected to paired comparison rankings which overcomes the problem of inconsistency of judgements given by the same observer. The data was analyzed statistically to determine preferred color with associated incision, establish confidence limits, and color grouping for each site and across all sites.

1.0 SECTION 1 - INTRODUCTION

The Countersurveillance and Deception Division was tasked by FORSCOM in early 1986 to develop the individual camouflage cover (ICC) for desert, woodland, and snow environments. The ICC is a small cloth cover, 5' x 7', which will weigh about 10-14 ounces, and be able to fit into a battle dress uniform pocket when not being used. It will deny the detection of a prone soldier in an ambush situation, or when on a surveillance, long-range patrol situation. The purpose of this study was twofold. The task first was to determine if a small or large Hogan garnish incision was best. The second task was to determine the best desert color to accompany the incision. Five sites were selected in the desert southwest, and the ICCs were evaluated by ground observers as to how well they blended with the desert backgrounds.

2.0 SECTION 2 - PROCEDURE

2.1 Test ICCs.

There were a total of thirty-seven variations of desert colors for this study. The nucleus of these colors was taken from the Saudi Arabian net palette study. These original colors were tested in the deserts of Saudi Arabia[2/] and the U.S. desert southwest. Additional colors were obtained through modification. Each of thirty-seven colors were painted on seventy-four vinyl-coated sheets, 5' x 7', which were then incised with either the small or large Hogan incision. Thus, there was a total of seventy-four vinyl-coated ICCs - thirty-seven small Hogans and thirty-seven large Hogans.

2.2 Test Sites.

Five sites were used to evaluate the ICCs. Two of the sites were in the Yuma, Arizona area, two at Anza Borrego State Park, California, and one at Jean Lake, near Las Vegas, Nevada. Both sites at Anza Borrego State Park were sandy with small stones. Vegetation was very sparce. Yuma site #1 was very sandy with some vegetation, while Yuma site #2 was on Ogilby Road and was rocky with very sparce vegetation. The Jean Lake site contained moderate vegetation with rocks, and was located on a hillside.

2.3 Test Subjects.

The test subjects consisted of ten enlisted U.S. Marine Corps personnel from Camp Pendleton, California, and two civilians from the Belvoir Research, Development, and Engineering Center, Fort Belvoir, Virginia. All personnel had corrected 20/20 vision and normal color vision. No observations were made with sunglasses.

2.4 Data Generation.

The seventy-four Hogan incised ICCs were randomly assigned to groups of six each. The four that best blended with the desert environment, in terms of color and texture, were selected and put aside for additional evaluations. This process continued until the original seventy-four ICCs were reduced to the six best. The best six ICCs were then shown in all possible pairs - fifteen, with the best ICC for each pair chosen for ability to blend with the desert. The number of times the individual ICC was judged to be the best was tabulated and subjected to data analysis.

3.0 SECTION 3 - RESULTS

The ICCs were evaluated at each of the five sites to determine which colors best blended with the desert environment. Section 2.4 describes how the best six ICCs were selected for each site. Table 1 shows the top six colors for each of the five sites.

TABLE 1
Summary of the Best Six Desert ICCs for Each Site

	Site				
Colors	Yuma Site 1	Yuma Site 2	Jean Lake	Anza Borrego Site 1	Anza Borrego Site 2
P6-S			X		
W-S	X	X		X	
XI-S	X	X			X
XI-L	X	X	X		
12-S				X	
21-S			X		X
21-L	X	X			X
26-S	X		X	X	X
26-L		X		X	
33-S	X	X	X	X	X
33-L				X	X
37-S			X		

NOTE: The L is large Hogan incision, while S is small Hogan incision. Net 33-S is the only color to make the best six colors for all five sites.

The results of each site for the above six best nets will not be included, because they would be too voluminous to present in these proceedings. This data is available upon request from the U.S. Army Belvoir Research, Development and Engineering Center, ATTN: STRBE-JDS, Fort Belvoir, VA 22060. When averaging the final best six ICCs across all five sites, a total of twelve ICCs made the best list. Some nets such as 37-S made the final six ICCs for only one site. A value of zero was added for each cell block when the ICC did not make the final six for that particular site. Tables 2-4 contain the statistics for the twelve ICCs. Figure 1 is the graphic display of Table 2. Table 5 describes the final twelve ICC nets as to color and incision.

TABLE 2

Descriptive Data for Final ICCs (Color Blend) with Desert Background, Across All Sites

COLOR	N	MEAN	STANDARD ERROR	95% CONFIDENCE INTERVAL LOWER LIM	UPPER LIM
P6-S	59	0.1864	0.6010	0.0298	0.3431
W-S	59	1.4237	1.6422	0.9957	1.8517
XI-S	59	1.5932	1.5550	1.1879	1.9985
XI-L	59	1.6780	1.8795	1.1881	2.1678
12-S	59	0.1017	0.6616	0.0000	0.2741
21-S	59	0.9153	1.3808	0.5554	1.2751
21-L	59	0.9831	1.2931	0.6460	1.3201
26-S	59	2.8983	1.8541	2.4151	3.3816
26-L	59	1.2712	1.7304	0.8202	1.7222
33-S	59	2.7119	1.4026	2.3463	3.0774
33-L	59	0.6610	1.1539	0.3603	0.9618
37-S	59	0.5763	1.2206	0.2581	0.8944

Note that the higher the mean value, the better the ICC blended with the desert environments.

TABLE 3

Analysis of Variance for Final ICCs (Color Blend) with Desert Background, Across All Sites

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F-TEST	SIG LEVEL
Color	11	508.5466	46.2315	22.8823	0.0000*
Error	696	1406.2034	2.0204		
Total	707	1914.7500			

* Significant at α less than .001 level.

This table indicates that there are significant differences in the ability of the final ICCs to blend with the desert backgrounds. Table 4 identifies which ICCs are significantly different from each other.

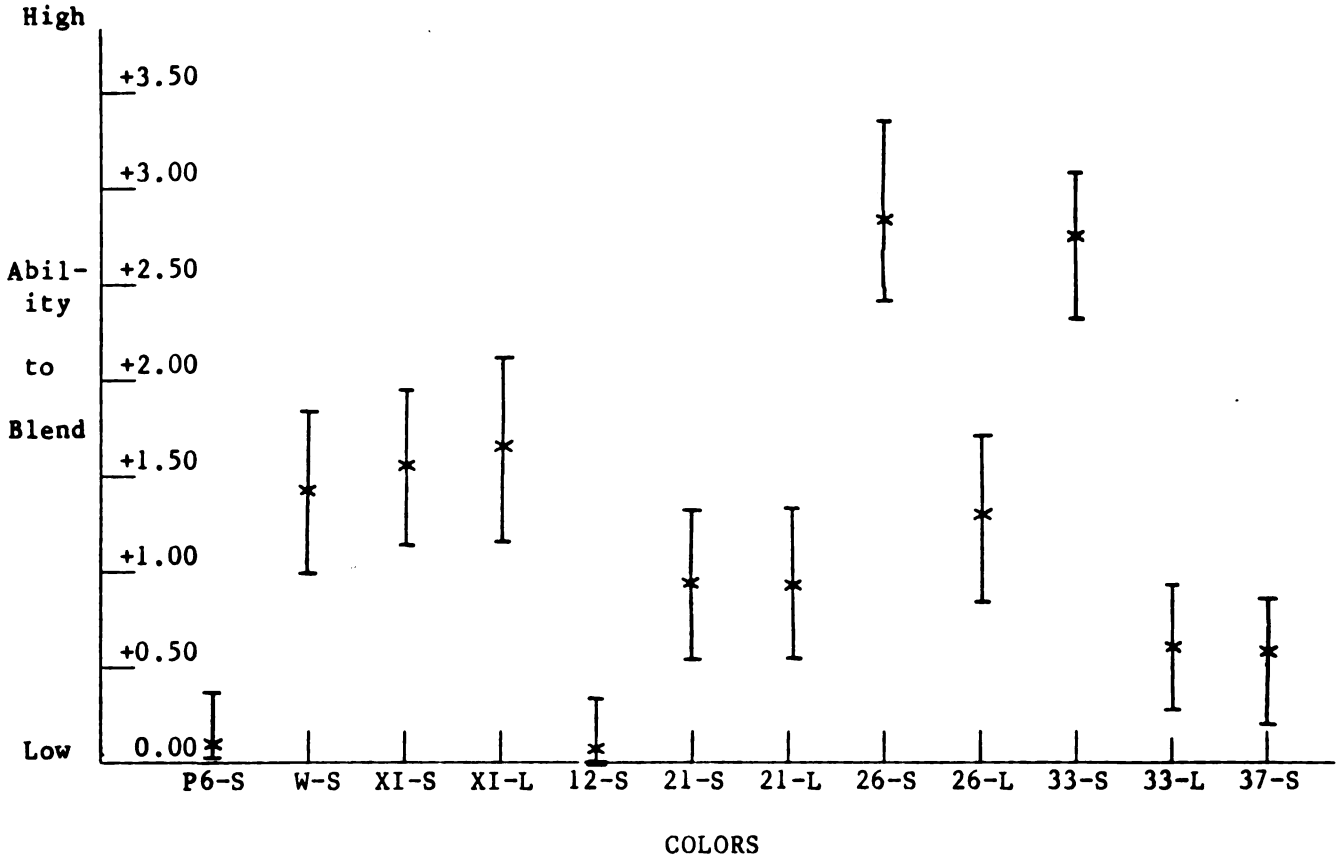


Figure 1. Ability of the Final ICCs to Blend with Desert Background, Averaged Across All Sites.

TABLE 4

Individual Comparisons, Identifying Which of the Final ICC Colors Differed Significantly from Each Other, Averaged Across Sites

COLOR P6-S		AND COLOR W-S		
COMPARISON =		-1.23729	SUM OF SQUARES =	45.16102
F =	22.352	SIGNIFICANCE LEVEL =	0.00000	***
COLOR P6-S		AND COLOR XI-S		
COMPARISON =		-1.40678	SUM OF SQUARES =	58.38136
F =	28.896	SIGNIFICANCE LEVEL =	0.00000	***
COLOR P6-S		AND COLOR XI-L		
COMPARISON =		-1.49153	SUM OF SQUARES =	65.62712
F =	32.482	SIGNIFICANCE LEVEL =	0.00000	***
COLOR P6-S		AND COLOR 12-S		
COMPARISON =		0.08475	SUM OF SQUARES =	0.21186
F =	0.105	SIGNIFICANCE LEVEL =	1.00000	

TABLE 4 (Cont)

COLOR P6-S AND COLOR 21-S
COMPARISON = -0.72881 SUM OF SQUARES = 15.66949
F = 7.756 SIGNIFICANCE LEVEL = 0.00543 **

COLOR P6-S AND COLOR 21-L
COMPARISON = -0.79661 SUM OF SQUARES = 18.72034
F = 9.266 SIGNIFICANCE LEVEL = 0.00238 **

COLOR P6-S AND COLOR 26-S
COMPARISON = -2.71186 SUM OF SQUARES = 216.94915
F = 107.379 SIGNIFICANCE LEVEL = 0.00000 ***

COLOR P6-S AND COLOR 26-L
COMPARISON = -1.08475 SUM OF SQUARES = 34.71186
F = 17.181 SIGNIFICANCE LEVEL = 0.00004 ***

COLOR P6-S AND COLOR 33-S
COMPARISON = -2.52542 SUM OF SQUARES = 188.14407
F = 93.122 SIGNIFICANCE LEVEL = 0.00000 ***

COLOR P6-S AND COLOR 33-L
COMPARISON = -0.47458 SUM OF SQUARES = 6.64407
F = 3.288 SIGNIFICANCE LEVEL = 0.06998

COLOR P6-S AND COLOR 37-S
COMPARISON = -0.38983 SUM OF SQUARES = 4.48305
F = 2.219 SIGNIFICANCE LEVEL = 0.13656

COLOR W-S AND COLOR XI-S
COMPARISON = -0.16949 SUM OF SQUARES = 0.84746
F = 0.419 SIGNIFICANCE LEVEL = 0.51732

COLOR W-S AND COLOR XI-L
COMPARISON = -0.25424 SUM OF SQUARES = 1.90678
F = 0.944 SIGNIFICANCE LEVEL = 0.33148

COLOR W-S AND COLOR 12-S
COMPARISON = 1.32203 SUM OF SQUARES = 51.55932
F = 25.519 SIGNIFICANCE LEVEL = 0.00000 ***

COLOR W-S AND COLOR 21-S
COMPARISON = 0.50847 SUM OF SQUARES = 7.62712
F = 3.775 SIGNIFICANCE LEVEL = 0.05222

COLOR W-S AND COLOR 21-L
COMPARISON = 0.44068 SUM OF SQUARES = 5.72881
F = 2.835 SIGNIFICANCE LEVEL = 0.09243

COLOR W-S AND COLOR 26-S
COMPARISON = -1.47458 SUM OF SQUARES = 64.14407
F = 31.748 SIGNIFICANCE LEVEL = 0.00000 ***

TABLE 4 (Cont)

COLOR W-S		AND COLOR 26-L		
COMPARISON =	0.15254	SUM OF SQUARES =		0.68644
F =	0.340	SIGNIFICANCE LEVEL =	0.56006	
COLOR W-S		AND COLOR 33-S		
COMPARISON =	-1.28814	SUM OF SQUARES =		48.94915
F =	24.227	SIGNIFICANCE LEVEL =	0.00000	***
COLOR W-S		AND COLOR 33-L		
COMPARISON =	0.76271	SUM OF SQUARES =		17.16102
F =	8.494	SIGNIFICANCE LEVEL =	0.00362	**
COLOR W-S		AND COLOR 37-S		
COMPARISON =	0.84746	SUM OF SQUARES =		21.18644
F =	10.486	SIGNIFICANCE LEVEL =	0.00123	**
COLOR XI-S		AND COLOR XI-L		
COMPARISON =	-0.08475	SUM OF SQUARES =		0.21186
F =	0.105	SIGNIFICANCE LEVEL =	1.00000	
COLOR XI-S		AND COLOR 12-S		
COMPARISON =	1.49153	SUM OF SQUARES =		65.62712
F =	32.482	SIGNIFICANCE LEVEL =	0.00000	***
COLOR XI-S		AND COLOR 21-S		
COMPARISON =	0.67797	SUM OF SQUARES =		13.55932
F =	6.711	SIGNIFICANCE LEVEL =	0.00968	**
COLOR XI-S		AND COLOR 21-L		
COMPARISON =	0.61017	SUM OF SQUARES =		10.98305
F =	5.436	SIGNIFICANCE LEVEL =	0.01987	*
COLOR XI-S		AND COLOR 26-S		
COMPARISON =	-1.30508	SUM OF SQUARES =		50.24576
F =	24.869	SIGNIFICANCE LEVEL =	0.00000	***
COLOR XI-S		AND COLOR 26-L		
COMPARISON =	0.32203	SUM OF SQUARES =		3.05932
F =	1.514	SIGNIFICANCE LEVEL =	0.21870	
COLOR XI-S		AND COLOR 33-S		
COMPARISON =	-1.11864	SUM OF SQUARES =		36.91525
F =	18.271	SIGNIFICANCE LEVEL =	0.00002	***
COLOR XI-S		AND COLOR 33-L		
COMPARISON =	0.93220	SUM OF SQUARES =		25.63559
F =	12.688	SIGNIFICANCE LEVEL =	0.00038	***
COLOR XI-S		AND COLOR 37-S		
COMPARISON =	1.01695	SUM OF SQUARES =		30.50847
F =	15.100	SIGNIFICANCE LEVEL =	0.00011	***

TABLE 4 (Cont)

COLOR XI-L		AND COLOR 12-S		
COMPARISON =	1.57627	SUM OF SQUARES =		73.29661
F =	36.278	SIGNIFICANCE LEVEL =	0.00000	***
COLOR XI-L		AND COLOR 21-S		
COMPARISON =	0.76271	SUM OF SQUARES =		17.16102
F =	8.494	SIGNIFICANCE LEVEL =	0.00362	**
COLOR XI-L		AND COLOR 21-L		
COMPARISON =	0.69492	SUM OF SQUARES =		14.24576
F =	7.051	SIGNIFICANCE LEVEL =	0.00801	**
COLOR XI-L		AND COLOR 26-S		
COMPARISON =	-1.22034	SUM OF SQUARES =		43.93220
F =	21.744	SIGNIFICANCE LEVEL =	0.00000	***
COLOR XI-L		AND COLOR 26-L		
COMPARISON =	0.40678	SUM OF SQUARES =		4.88136
F =	2.416	SIGNIFICANCE LEVEL =	0.12032	
COLOR XI-L		AND COLOR 33-S		
COMPARISON =	-1.03390	SUM OF SQUARES =		31.53390
F =	15.608	SIGNIFICANCE LEVEL =	0.00008	***
COLOR XI-L		AND COLOR 33-L		
COMPARISON =	1.01695	SUM OF SQUARES =		30.50847
F =	15.100	SIGNIFICANCE LEVEL =	0.00011	***
COLOR XI-L		AND COLOR 37-S		
COMPARISON =	1.10169	SUM OF SQUARES =		35.80508
F =	17.722	SIGNIFICANCE LEVEL =	0.00003	***
COLOR 12-S		AND COLOR 21-S		
COMPARISON =	-0.81356	SUM OF SQUARES =		19.52542
F =	9.664	SIGNIFICANCE LEVEL =	0.00192	**
COLOR 12-S		AND COLOR 21-L		
COMPARISON =	-0.88136	SUM OF SQUARES =		22.91525
F =	11.342	SIGNIFICANCE LEVEL =	0.00078	***
COLOR 12-S		AND COLOR 26-S		
COMPARISON =	-2.79661	SUM OF SQUARES =		230.72034
F =	114.195	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 12-S		AND COLOR 26-L		
COMPARISON =	-1.16949	SUM OF SQUARES =		40.34746
F =	19.970	SIGNIFICANCE LEVEL =	0.00001	***
COLOR 12-S		AND COLOR 33-S		
COMPARISON =	-2.61017	SUM OF SQUARES =		200.98305
F =	99.477	SIGNIFICANCE LEVEL =	0.00000	***

TABLE 4 (Cont)

COLOR 12-S		AND COLOR 33-L			
COMPARISON =		-0.55932	SUM OF SQUARES =		9.22881
F =	4.568	SIGNIFICANCE LEVEL =		0.03275	*
COLOR 12-S		AND COLOR 37-S			
COMPARISON =		-0.47458	SUM OF SQUARES =		6.64407
F =	3.288	SIGNIFICANCE LEVEL =		0.06998	
COLOR 21-S		AND COLOR 21-L			
COMPARISON =		-0.06780	SUM OF SQUARES =		0.13559
F =	0.067	SIGNIFICANCE LEVEL =		1.00000	
COLOR 21-S		AND COLOR 26-S			
COMPARISON =		-1.98305	SUM OF SQUARES =		116.00847
F =	57.418	SIGNIFICANCE LEVEL =		0.00000	***
COLOR 21-S		AND COLOR 26-L			
COMPARISON =		-0.35593	SUM OF SQUARES =		3.73729
F =	1.850	SIGNIFICANCE LEVEL =		0.17403	
COLOR 21-S		AND COLOR 33-S			
COMPARISON =		-1.79661	SUM OF SQUARES =		95.22034
F =	47.129	SIGNIFICANCE LEVEL =		0.00000	***
COLOR 21-S		AND COLOR 33-L			
COMPARISON =		0.25424	SUM OF SQUARES =		1.90678
F =	0.944	SIGNIFICANCE LEVEL =		0.33148	
COLOR 21-S		AND COLOR 37-S			
COMPARISON =		0.33898	SUM OF SQUARES =		3.38983
F =	1.678	SIGNIFICANCE LEVEL =		0.19543	
COLOR 21-L		AND COLOR 26-S			
COMPARISON =		-1.91525	SUM OF SQUARES =		108.21186
F =	53.559	SIGNIFICANCE LEVEL =		0.00000	***
COLOR 21-L		AND COLOR 26-L			
COMPARISON =		-0.28814	SUM OF SQUARES =		2.44915
F =	1.212	SIGNIFICANCE LEVEL =		0.27108	
COLOR 21-L		AND COLOR 33-S			
COMPARISON =		-1.72881	SUM OF SQUARES =		88.16949
F =	43.639	SIGNIFICANCE LEVEL =		0.00000	***
COLOR 21-L		AND COLOR 33-L			
COMPARISON =		0.32203	SUM OF SQUARES =		3.05932
F =	1.514	SIGNIFICANCE LEVEL =		0.21870	
COLOR 21-L		AND COLOR 37-S			
COMPARISON =		0.40678	SUM OF SQUARES =		4.88136
F =	2.416	SIGNIFICANCE LEVEL =		0.12032	

TABLE 4 (Cont)

COLOR 26-S		AND COLOR 26-L		
COMPARISON =	1.62712	SUM OF SQUARES =		78.10169
F =	38.656	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 26-S		AND COLOR 33-S		
COMPARISON =	0.18644	SUM OF SQUARES =		1.02542
F =	0.508	SIGNIFICANCE LEVEL =	0.47633	
COLOR 26-S		AND COLOR 33-L		
COMPARISON =	2.23729	SUM OF SQUARES =		147.66102
F =	73.085	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 26-S		AND COLOR 37-S		
COMPARISON =	2.32203	SUM OF SQUARES =		159.05932
F =	78.726	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 26-L		AND COLOR 33-S		
COMPARISON =	-1.44068	SUM OF SQUARES =		61.22881
F =	30.305	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 26-L		AND COLOR 33-L		
COMPARISON =	0.61017	SUM OF SQUARES =		10.98305
F =	5.436	SIGNIFICANCE LEVEL =	0.01987	*
COLOR 26-L		AND COLOR 37-S		
COMPARISON =	0.69492	SUM OF SQUARES =		14.24576
F =	7.051	SIGNIFICANCE LEVEL =	0.00801	**
COLOR 33-S		AND COLOR 33-L		
COMPARISON =	2.05085	SUM OF SQUARES =		124.07627
F =	61.412	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 33-S		AND COLOR 37-S		
COMPARISON =	2.13559	SUM OF SQUARES =		134.54237
F =	66.592	SIGNIFICANCE LEVEL =	0.00000	***
COLOR 33-L		AND COLOR 37-S		
COMPARISON =	0.08475	SUM OF SQUARES =		0.21186
F =	0.105	SIGNIFICANCE LEVEL =	1.00000	

The following ICCs differed significantly from each other: P6-S vs. W-S, P6-S vs. XI-S, P6-S vs. XI-L, P6-S vs. 21-S, P6-S vs. 21-L, P6-S vs. 26-S, P6-S vs. 26-L, P6-S vs. 33-S, W-S vs. 12-S, W-S vs. 26-S, W-S vs. 33-S, W-S vs. 33-L, W-S vs. 37-S, XI-S vs. 12-S, XI-S vs. 21-S, XI-S vs. 21-L, XI-S vs. 26-S, XI-S vs. 33-S, XI-S vs. 33-L, XI-S vs. 37-S, XI-L vs. 12-S, XI-L vs. 21-S, XI-L vs. 21-L, XI-L vs. 26-S, XI-L vs. 33-S, XI-L vs. 33-L, XI-L vs. 37-S, 12-S vs. 21-S, 12-S vs. 21-L, 12-S vs. 26-S, 12-S vs. 26-L, 12-S vs. 33-S, 12-S vs. 33-L, 21-S vs. 26-S, 21-S vs. 33-S, 21-L vs. 26-S, 21-L vs. 33-S, 26-S vs. 26-L, 26-S vs. 33-L, 26-S vs. 37-S, 26-L vs. 33-S, 26-L vs. 33-L, 26-L vs. 37-S, 33-S vs. 33-L, and 33-S vs. 37-S.

* Significant at α less than .05 level.
** Significant at α less than .01 level.
*** Significant at α less than .001 level.

TABLE 5

Physical Description of the Final Twelve ICCs

COLOR/INCISION	DESCRIPTION
P6-S	Black spots on tan color 26, color XI on reverse side.
W-S	A fifty-fifty mixture of Saudi Arabian color 8 and 7 in both sides of the net.
XI-S	Standard tan color on both sides of the net.
XI-L	Same color as XI-S, only this ICC has large incisions.
12-S	New color on both sides of net.
21-S	Color XI on one side of the net, new color 33 on the other side.
21-L	Same color as 21, only this ICC has large incisions.
26-S	New color on both sides of net.
26-L	Same color as 26, only this ICC has large incisions.
33-S	New color on both sides of net.
33-L	Same color as 33, only this ICC has large incisions.
37-S	Color XI on one side of the net, with color W on the other side.

Note that S is small Hogan incisions, while L is large Hogan incisions.

4.0 SECTION 4 - DISCUSSION

All the colors were on the gray or tan scale, with the tan colors rated as having the most ability to blend with the desert background. Table 1 shows that the pattern ICC net P6-S was the only multi-color to make the final twelve ICCs, and it along with net 12-S was judged by the ground observers as having the least ability to blend with the desert background when averaged across all five sites. Net 33-S was the only net to make the final six for all sites. ICC 26-S was a final net for all sites, except for Yuma site #2. These nets did not significantly differ from each other ($\alpha = 0.476$), with net 33-S having a preference rating of 3.07 to 3.38 for net 26-S. The Yuma site #2 area was very rocky, while the other sites were very sandy. The test team has seen deserts in Egypt and Saudi Arabia, and these deserts were very sandy. Therefore, net 26-S appears to be the best ICC for general desert use. This color was among the best six at Yuma site #2, only it had large Hogan inci-

sions (26-L). The texture of the rocks is larger and more rough in appearance than that of sand. It appears that the texture of the rocks was the driving force in the selection of 26-L rather than 26-S. Four of the top five ICCs, 26-S, 33-S, XI-S and W-S, were small incisions. The only exception is ICC XI-L. Except for very rocky deserts, the small incision blends best with the texture of the desert floor. Desert color paint studies[2,3,4] have shown that the desert southwest is a darker more gray desert than those seen in Saudi Arabia and Egypt. Additional deserts of interest in the Middle East should be photographed and soil samples studied before a final decision is made for the colors 26 and 33.

5.0 SECTION 5 - SUMMARY AND CONCLUSIONS

A total of thirty-seven colors were painted on seventy-four vinyl-coated sheets 5' x 7'. Each color was given either the small or large Hogan incision. These ICCs were then taken to five sites in the desert southwest and evaluated as to their ability to blend with the desert background in terms of color and texture. Ten enlisted U.S. Marine Corps personnel from Camp Pendleton, California, and two civilians from the Belvoir Research, Development and Engineering Center, Fort Belvoir, Virginia, served as ground observers. The seventy-four ICCs were randomly assigned to groups of six each. The four ICCs that best blended with the desert environment were selected and put aside for additional evaluation which continued until the best six for each site remained. These best six ICCs were then viewed on all possible pairs (15), with the best selected for each pair in their ability to match the desert floor. The number of times the individual ICC was judged to be best was tabulated and subjected to data analysis. The following conclusions were drawn:

a. Colors 26 and 36 were the most effective in blending with the desert.

b. Color 26 was selected for initial ICC production.

c. The small Hogan incision (S) is more effective than the large Hogan incision (L) except for very rocky terrain.

d. The U.S. desert southwest is darker and more gray than the sites seen in the Middle East, making additional work on the two colors necessary before final color selection.

REFERENCES

1. Natrella, Mary G., Experimental Statistics, National Bureau of Standards Handbook 91, U.S. Department of Commerce, Washington, D.C., 1966

2. Anitole, George and Johnson, Ronald L., Saudi Arabian National Guard Camouflage Net Development Program, U.S. Army Belvoir Research and Development Center, Fort Belvoir, Virginia, February 24, 1984

3. Anitole, George and Johnson, Ronald L., Statistical Evaluation of Desert Paint Colors, Image Interpretation, U.S. Army Belvoir Research, Development and Engineering Center, Fort Belvoir, Virginia, September 1985

4. Anitole, George and Johnson, Ronald L., Statistical Evaluation of Desert Paint Colors, Ground Observers, U.S. Army Belvoir Research, Development and Engineering Center, Fort Belvoir, Virginia, December 1985.

The Combinatorics of Message Filtering

Terence M. Cronin
US Army Signals Warfare Center, Warrenton, Virginia

Topic: Computational Aspects of Event Recognition Under Conditions of Sparse Reporting, Uncertainty, and Information Decay

[**Background**: The general problem of filtering a stack of documents is arguably *context-sensitive*; i.e., an individual document cannot be prioritized independently of semantic knowledge about the current environment. In pursuing this line of thought, an attempt is being made to *recognize* background events which change dynamically in time, with the ultimate motivation being to assess the import of any given message with respect to the *time-criticality* of the most recent set of events.

[**Abstract**: Given a set of message traffic and an exhaustive menu of possible events, select the event which is best explained by the message data. This problem involves a reasoning process known as *abduction*, as differentiated from the processes of deduction and induction. An argument is made that the recognition of events from message data is a diagnosis problem. In the medical world, disorders are diagnosed from observation of symptoms. In the case of electronic troubleshooting, failure of a whole circuit may be explained by failure of single components or sets of components. In the general sense, an event may be diagnosed by careful observation of the constituent phenomena which comprise the event. With respect to battlefield situation assessment, both the manifestations for events and the events themselves change dynamically as more message traffic enters the system, since the decay of one event is accompanied by the emergence of another over time. This paper develops a formal theory of machine-assisted event recognition, but also casts an eye on the feasibility of implementation. Treated with some rigor are the combinatorics associated with such new formalisms as *suspecting* an event; *confirming* an event; computing the *threat* of an event; *revoking* a stale event; introducing two levels of relaxation into statistical testing; recovering from fundamental forms of string error; and the number of feasible ways to filter a stream of n messages.

Fundamentals: Definitions and Concepts.

A ***message*** **m** is a feature vector together with a string of text: $\mathbf{m} = \{x_1, x_2, \ldots, x_n, x_{string}\}$. The feature vector is a set of sensor-measurable observable attributes of a manmade object. The string represents natural language which may have been generated by one of two communicators: either an individual who in some way controls the manmade object, or an outside observer describing the interaction of the object with the world.

The ***timeliness*** of a message m_i is the time t_i at which its feature vector was created (time at sensor detection). A message m_i is said to be more ***timely*** than message m_j iff $t_i > t_j$.

A ***map y*** is a spatially organized representation of a section of the world upon which the manmade objects referenced in messages move about.

A ***constituent phenomenon*** **g** is a logical function of message data conjoined with map data. If the expression $g(x_i,y)$ evaluates to true, then $g(x_i,y) = 1$; otherwise $g(x_i,y) = 0$.

An ***event e*** is a set of ***constituent phenomena***: $e = \{g_1, g_2, \ldots, g_k\}$.

The ***message set*** M is the set of all messages: $\mathbf{M} = \{m_i \mid i = 1,n\}$.

The ***event space*** E is the set of all events: $\mathbf{E} = \{e_i \mid i = 1,m\}$.

Becoming Suspicious, Amassing Support, and Confirming Event Hypotheses.

The problem of recognizing an event by aggregating the truth or falsity of its message-derived constituent phenomena is being treated as a diagnosis problem. Message-driven event recognition must avail itself of a reasoning process known as *abduction* (as contrasted with *induction* and *deduction)*, in which the event which best explains the message data is selected as the most likely hypothesis, even when the message data is incomplete, subject to some error, and describe temporally transient phenomena. This form of automated reasoning is still very much a research issue, with several disjoint efforts seemingly offering potential leverage. An abductive inferencing mechanism is being explored for a medical domain, by assembling those hypotheses which are best explained by a set of data [J1]. There has been promising work recently in the areas of justification-based and assumption-based truth maintenance systems [D1, D2, M1]. These techniques achieve truth maintenance by detecting inconsistency, followed respectively by dependency-directed backtracking, or by gathering the most general context which preserves consistency. Yet another interesting line of research is a minimal covering set theory approach [N1, P1], which attempts to diagnose medical disorders by constructing the least set of symptoms which point to each disorder. However, the computational feasibility of this technique is questionable, since derivation of the minimal covering set belongs to the class of NP-complete problems [G1].

The foundation of a new theory of event recognition emerges if one unifies the disciplines of truth maintenance systems with minimal covering sets. If a dimension is added to accomodate other than temporally static situations, then the theory permits recognizing events from their manifestations, when both the events and their manifestations may be changing dynamically in time. A crucial underpinning of the theory is that the emergence of a new event is inversely proportional to the decay of an older event, since the same observable primitive resources are involved. Also assumed as axiomatic is the concept that full credence in an event is well nigh impossible, due to the non-systematic way in which evidence accrues, together with the difficulty in retracting an assertion once it is assigned a probability of one [K3]. Therefore, the theory must be capable of confirming events when only *partial* support is manifested. It will be seen that this becomes feasible if one is permitted to revoke support for phenomena which have already emerged and sustained under both spatial and temporal constraints.

A message m is said to ***support*** an event e if and only if there exists some feature x_i of m, some constituent phenomenon g_k of e, such that $g_k(x_i)$ evaluates to true. If such is the case, we also say that phenomenon g_k is ***supported*** by m. Any *unsupported* phenomenon of e is called a ***virtual*** phenomenon.

Entropy is a measure of information ***not*** available to make a decision about an event's feasibility. In this context, entropy is synonymous with ***uncertainty***.

A ***phenomenon entropy function*** f_p assigns to each constituent phenomenon g_k of an event e_i some integer value n_k based on the relative utility of g_k: $f_p(g_k) = n_k$. Phenomenon g_k is said to have entropy value n_k. Small values are assigned to constituent phenomena which are of *minimal* use in the decidibility of event e_i.

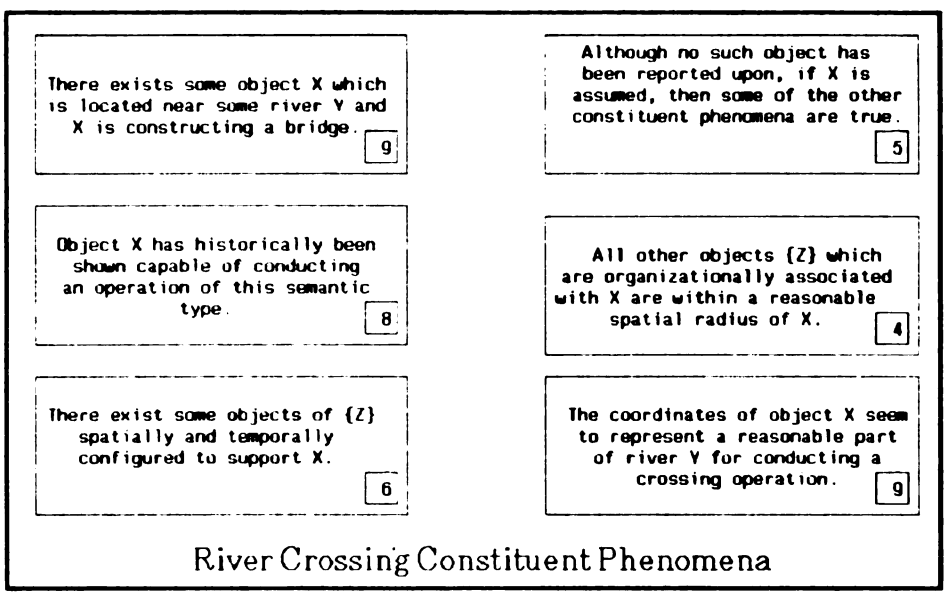


Figure 1. Illustration of Constituent Phenomena and Respective Entropy Values for a Hypothetical Event. Note the Subjunctive Voice of the Upper Right Phenomenon.

The ***total entropy*** T_e of an event e_i is the sum of its phenomenon entropy values: $T_e = \Sigma f_p(g_k), k = 1,n$.

The ***instantiated entropy*** I_e of an event e_i is the sum of the entropy values associated with the currently supported phenomena of e_i.

The ***suspicion-ratio*** for an event e_i is the quotient of the instantiated entropy of e_i with the total entropy of e_i.

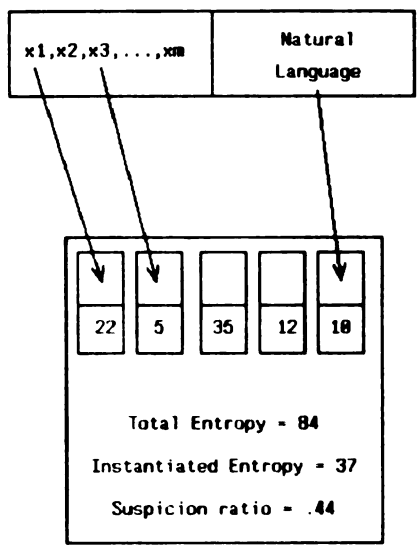


Figure 2. A single message supporting some constituent phenomena of a single event, with total and instantiated entropy values illustrated, together with the instantaneous suspicion-ratio.

The ***suspicion accumulator*** $s_n(e_i)$ for an event e_i is a temporal sequence of suspicion-ratios, updated whenever a new message is processed.

The ***volume-ratio*** for an event e_i is the quotient of the number of messages which support e_i with the total number of messages contained within a time frame of interest.

The ***volume accumulator*** $v_n(e_i)$ for an event e_i is a temporal sequence of volume-ratios, updated whenever a new message is processed.

The ***suspicion-volume accumulator*** $sv_n(e_k)$ for an event e_k is the sequence defined by the point-by-point multiply of the suspicion accumulator with the

volume accumulator: $sv_n(e_k) = \{s_i(e_k) * v_i(e_k) \mid i = 1, n\}$, where n is the number of messages processed during the time frame of interest.

An event e_i warrants ***suspicion-arousal*** if its suspicion-ratio exceeds a specified necessity condition, or if its suspicion accumulator sequence becomes monotonically increasing.

Example. The figure below depicts an event template which contains a total entropy of 27 information units, within a framework of 10 constituent phenomena. Suppose the criterion for suspicion-arousal is that the instantiated entropy be greater than 6 information units. There are $2^{10} = 1024$ ways of logically conjuncting the 10 constituent phenomena. An Interlisp search routine was implemented to identify those which fail to trigger suspicion-arousal. Result: 78 cases fail to satisfy the criterion.

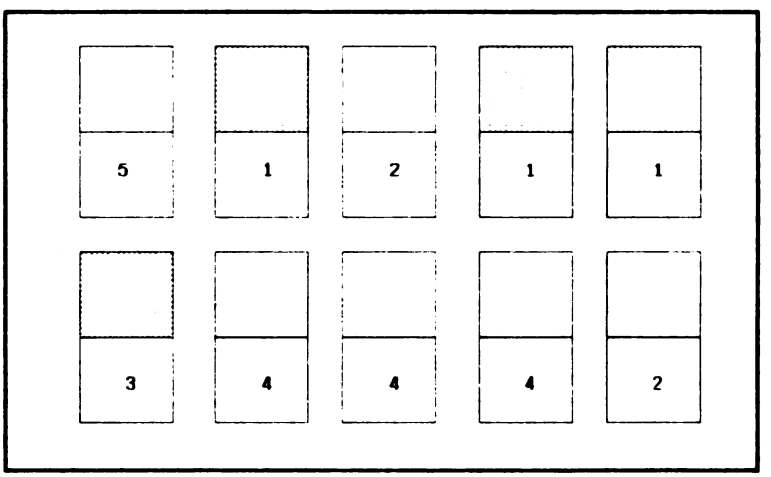
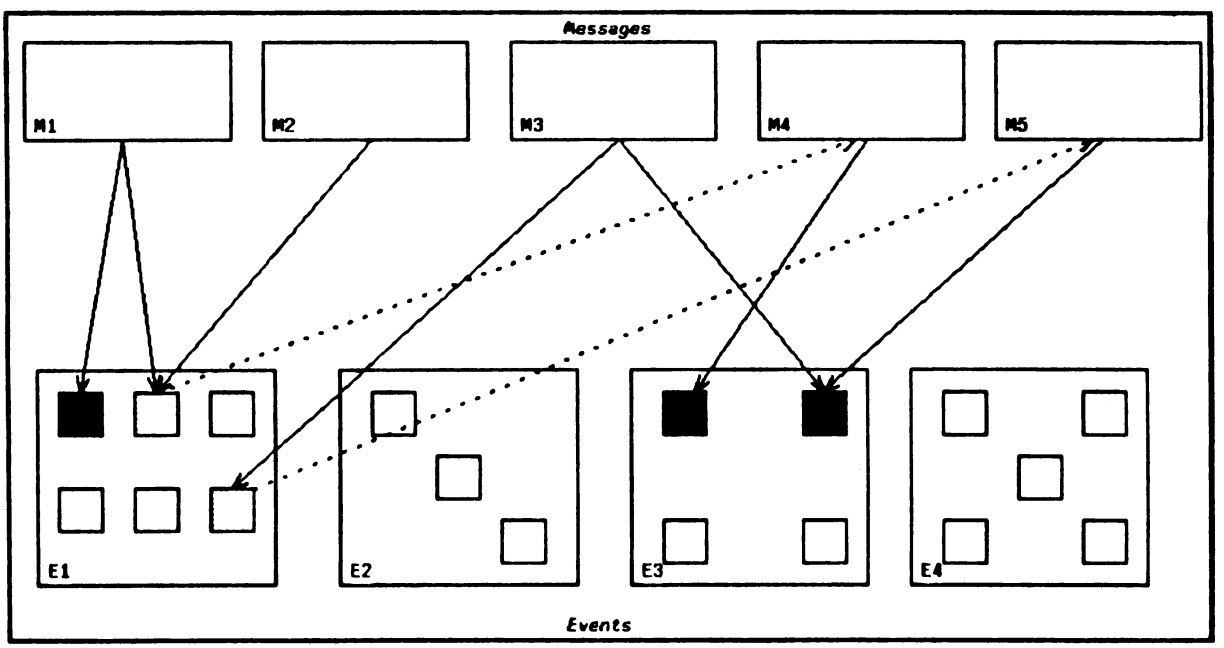


Figure 3. One of 78 Event Template Configurations (out of 1024) which Fails to Trigger Suspicion-arousal Under the Specified Constraint.

A ***temporal cusp*** is defined to be a point in time when the suspicion-volume accumulator sequence for one event becomes monotonically increasing (decreasing), while concurrently the suspicion-volume accumulator for another event becomes monotonically decreasing (increasing).

An event e_i warrants ***suspicion-confirmation*** if its suspicion-ratio exceeds a specified sufficiency condition, or a temporal cusp favorable to e_i is detected and all other events have less instantiated entropy than e_i.

Exercise. Consider the message stream below together with support arcs pointing to events (the dashed lines represent phenomenon revocation, which is defined in the next section, but for the purpose of this example causes cancellation of a support arc). Compute the suspicion-accumulator, volume-accumulator, and suspicion-volume accumulator sequences for events E1 and E3. Also identify any messages which cause a temporal cusp.



Solution.

time	s_n(E1)	v_n(E1)	s_n(E3)	v_n(E3)
M1	{.33}	{1.0}	{0.0}	{0.0}
M2	{.33,.33}	{1.0,1.0}	{0.0,0.0}	{0.0,0.0}
M3	{.33,.33,.50}	{1.0,1.0,1.0}	{0.0,0.0,.25}	{0.0,0.0,.33}
M4	{.33,.33,.50,.33}	{1.0,1.0,1.0,.75}	{0.0,0.0,.25,.50}	{0.0,0.0,.33,.50}
M5	{.33,.33,.50,.33,.17}	{1.0,1.0,1.0,.75,.60}	{0.0,0.0,.25,.50,.50}	{0.0,0.0,.33,.50,.60}
sv_n	E1: {.33,.33,.50,.25,.10}		E3: {0.0,0.0,.08,.25,.30}	

Figure 4. An illustration of the suspicion accumulator and volume accumulator sequences for two events. Also shown are the suspicion-volume accumulators for each. The dashed lines indicate strong-sense constituent phenomenon revocation (defined below). Message M4 causes a temporal cusp, together with suspicion-confirmation of E3 (assuming that the instantiated entropy for E1 is diminutive when compared to that of E3).

Note that this theory of event recognition relies upon *monotonic* conditions induced by the conservation of resources shared by events evolving in time, and by so doing abstains from decision based *on numerical thresholds*. In the example above, suspicion about the existence of E3 was confirmed with only half its constituent phenomena instantiated by message evidence, and with an instantaneous suspicion-volume accumulator value of only .25!

A potentially powerful technique to abduce an event from message data is the occasional use of the *subjunctive* voice when attempting to logically instantiate the constituent phenomena of an event. It may be the case that several constituent phenomena become true if the truth of just one primitive clause is (for the time being) assumed, even though the message data has not yet corroborated the primitive clause. Refer back to Figure 1 for an instance of the explicit use of the subjunctive voice.

Discounting Events which have already Emerged, Sustained, and Decayed in Time.

Much attention has been paid in the literature to deciding when an event is supported by evidence. Equally important is determining when an event no longer warrants having its constituent phenomena maintained because of the decay of information over time. Currently automated systems are frequently incompetent when event probabilities reach a plateau. When such is the case, a computer process should be capable of deciding whether the event is continuing to progress, or has already sustained and decayed. When an event becomes obsolete, automated techniques are required to revoke its constituent phenomena so that the computer's belief in the event is retracted, or at least discounted. The following section describes a set of computational techniques designed to solve problems in this area.

A message m_j is said to be **revocation-provocative in the weak sense with respect to event e_k** iff $\exists$ some message m_i less timely than m_j; $x_s \in m_i, m_j$; $g_t \in e_k$; $g_t(x_s|m_i) = 1$, and $g_t(x_s|m_j) = 0$. See Figure 5.

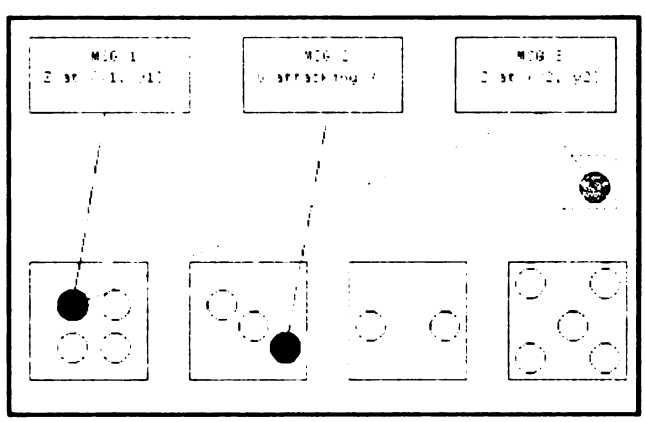


Figure 5. Weak-sense Phenomenon Revocation.

Discussion. Weak-sense phenomenon revocation may provide the rudiments for automated *non-monotonic reasoning*. Before one may accomodate the unanticipated, one must be capable of suspending belief in a previous state of the world by reasoning in the following way:

a) Some object has obtained new spatial and temporal coordinates which negate belief in an earlier set of coordinates which were accountable by some event;

b) No other explicitly modeled event contains constituent phenomena capable of explaining the new coordinates.

Once an automaton demonstrates a weak-sense phenomenon revocation capability, its next logical step would be to generate a new event to explain the coordinates of the errant object. There is currently no technology available to perform this process, and it is not likely that there will be for some time, since a leap of this magnitude is intrinsically linked to *data-driven templating*, and *learning by discovery*.

A message m_j is said to be **revocation-provocative in the strong sense with respect to event e_k** iff ∃ some event e_l different from event e_k, ∃ some messsage m_i less timely than m_j; $x_s \in m_i, m_j$; $g_t \in e_k$; $g_u \in e_l$; with $g_t(x_s|m_i) = 1$, $g_t(x_s|m_j) = 0$, and $g_u(x_s|m_j) = 1$.

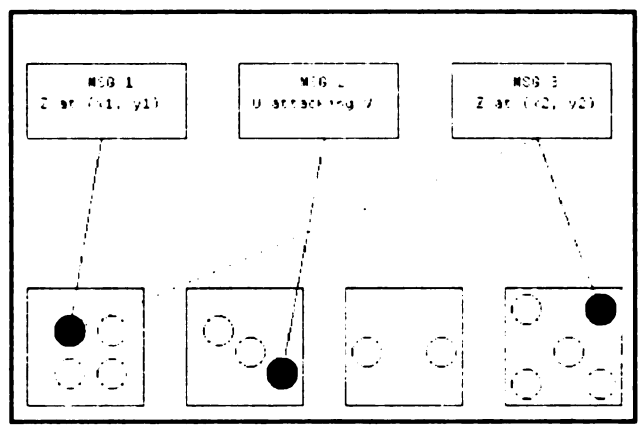


Figure 6. Strong-sense Phenomenon Revocation.

An event e_i becomes ***stale*** if both its suspicion and volume accumulator sequences become *strictly* monotonically decreasing.

An event e_i warrants having its attributes ***revoked*** (i.e., its constituent attributes g_i set to 0) under two conditions:

i. A temporal cusp unfavorable to e_i is detected.

ii. e_i is determined to be stale.

Event Stasis Induced by External Phenomena.

Under certain conditions the constituent phenomena of an event may become inert for protracted periods of time. In such a situation, the event is said to be undergoing ***stasis***. Stasis is caused by the existence of *external* phenomena (*not* associated with the event) which tend to force spatial immobility upon the objects which (logically conjoined with a map) define the constituent phenomena of the event.

The ***stasis factor*** of an event is defined to be the tendency for the constituent phenomena of an event to remain inert. The stasis factor is computed by a two-step process:

1. Construct the stasis matrix as follows: for each constituent phenomenon (whether instantiated or virtual) belonging to the event, assign a probabilistic estimate representing the certainty that there exist external phenomena committed to any of the following:

   a. Prolonging the constituent state.

   b. Transitioning the constituent object(s) from the current state to one recently visited.

2. Average across all probabilities derived at step 1.

Stasis as used here is a state of the world induced by countermeasures, and is functionally akin to the result obtained by applying a minimax criterion utilized by game theorists.

Threat Computation is a Nonsimplistic, Data-driven Process.

The threat of an event cannot be derived by isolating the event from its environment. An event which unto itself seems threatening may in fact be quite innoculous given that sufficient countermeasures are brought into play. Other factors which must be utilized in the derivation of threat include both the nature of preceding events and the potential impact of follow-on events. This section describes a computational technique to derive the threat of an event based on both the support for an event and the countermeasures at hand to thwart the event.

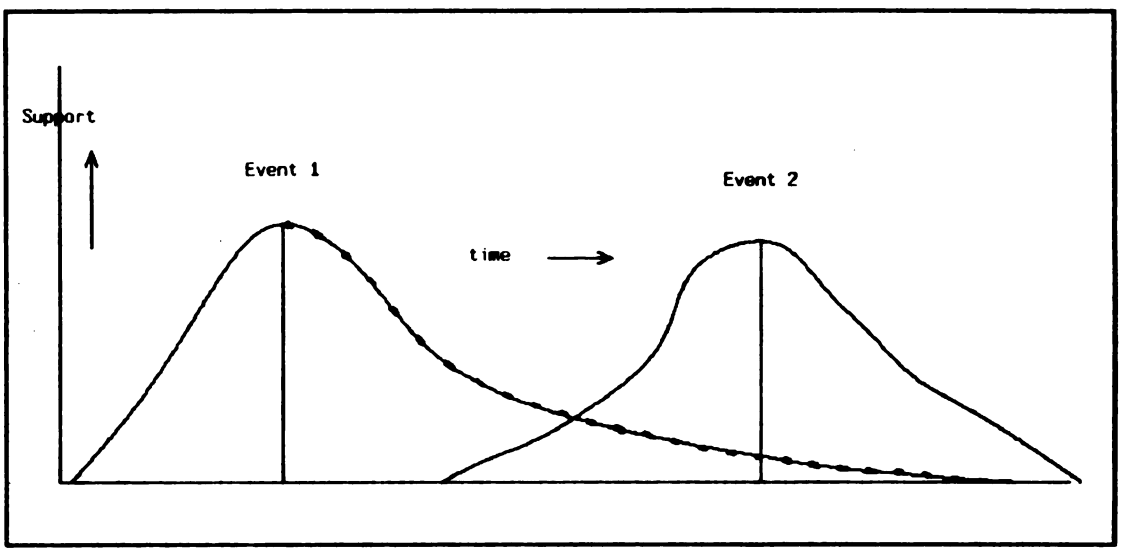


Figure 7. An Event Emerging During the Decay of its Predecessor (dashed area indicates the region of constituent phenomena revocation for the first event).

Assume that an event E1 has already transpired, and that another event E2 may be emerging. Since the same primitive resources will be utilized in event E2 as were used in event E1, we expect to see the computed subjective probability of event E2 rise at the same time that the computed probability of event E1 starts to fall (see Figure 7). Symbolically, we represent this as P(E2|E1), read "the probability that E2 is emerging given that E1 is decaying". Earlier research focused on developing a data-driven technique which lends itself to modeling the unsystematic skewness of events for which message data is providing asynchronous clues, and against which countermeasures may be progressing [C2]. The distribution of choice is the *Weibull* distribution, which has density function:

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}} \qquad \text{for } t > 0, \quad [5.1]$$
$$= 0 \qquad \text{elsewhere,}$$
$$\text{for } \alpha, \beta > 0.$$

If this function is differentiated with respect to time and set to zero, the critical value of t is:

$$t = \left(\frac{\beta-1}{\alpha\beta}\right)^{1/\beta} \qquad [5.2]$$

This expression is significant because it predicts at what value of time (in terms of α and β) the distribution will peak. All that remains is to couch the probability of an emerging event together with the countermeasures available to thwart the event in terms of α and β.

Computation of the Probability of Emerging Events.

Let P(E2|E1) be the probability that event E2 is emerging given that event E1 is decaying. This probability is equal to the quantity obtained by normalizing the suspicion-volume accumulator for E2 with respect to those for all other events in the event set. This quantity is also known as the ***evidence for event E2 with respect to the reference class E1***, or simply as the ***evidence for event E2***.

Computation of the Probability of Countermeasures to an Emerging Event.

Let P(C|E2) be the probability that countermeasures are available to thwart E2, given that E2 is emerging. This probability is computed by noting the real and virtual constituent phenomena of E2, setting up the stasis matrix for E2, and computing the stasis factor across all phenomena for E2.

Make the following substitutions for α and β in equation 5.2:

$$\alpha = 1 / (1 - P(C|E2)) \qquad [5.3]$$

$$\beta = 1 / P(E2|E1) \qquad [5.4]$$

The resultant critical value is:

$$t = [(1 - P(E2|E1))(1 - P(C|E2))]^{P(E2|E1)} \qquad [5.5]$$

Define the ***threat*** T_e of an event to be equal to this critical value. Note that threat is a function of probability-valued functions, and is mapped to the interval [0,1].

Discussion of the computational implications of the threat expression: A close look at equation [5.5] reveals that the derived threat is polynomially related to both the support for other events [called the *plausibility* of the event under the Dempster-Shafer formalism], and to the lack of countermeasures at hand to thwart the event. However, threat is exponentially related to the direct support for the event.

The ***threat*** of a message is defined to be precisely equivalent to the maximal threat of the list of events whose constituent phenomena are supported by the message. Let E_i be the event in the event set with the maximum instantaneous suspicion ratio. The message threat is directly proportional to both the evidence for E_i and the stasis factor of E_i.

Filtering Operations on Message Streams, and the Equivalence of Priority with Threat.

As messages enter a processing center for analysis, the sheer volume of traffic can rapidly generate a backlog which begs attention. It is reasonable to seek automated assistance in ordering the queue based on the *priority* of the messages, so that the most time-critical items are presented first. Queuing based solely on either message time of arrival or message timeliness is inappropriate because the threat of events for which the messages provide evidence must be brought into play. Regrettably, threat is a context-sensitive process, and must be painfully derived by abduction of events from the message data. The following section develops two theorems which show respectively: a) the number of ways to *order* a stream of n messages; b) the number of *feasible filtering solutions* on a stream of n messages.

A ***message stream*** is a queue of messages ordered chronologically by time of arrival in the queue.

A ***time-ordered queue*** is a message queue sorted by timeliness of the individual messages.

A ***filtering*** of a message stream m is a permutation based on ordering m as a monotonically decreasing function of threat.

A ***coarse threat quantization scheme*** on a message stream m of n messages is a partition of m into k threat classes such that every message contained in m is assigned to exactly one of the k classes.

Theorem 1. Number of Possible Ways to Order a Stream of n Messages.

There are n! ways to order a message stream of length n.

Proof. Since a filtering is a permutation on n objects, there are n! ways to order a message stream.

Definition. A ***feasible filtering solution*** is a filtering in which every message is correctly assigned to a threat class by a coarse threat quantization scheme.

Theorem 2. Number of Feasible Filtering Solutions on a Stream of n Messages.

Let P be a coarse threat quantization scheme on a message stream of length n into threat classes $\{C_1, C_2, ..., C_k\}$. Let $|C_i|$ denote the order of class C_i. Then the ***number of feasible filtering solutions*** is equal to $\Pi\, |C_i|!$, $i = 1,k$; with $\Sigma\, |C_i| = n$.

Proof. Let C_i be an arbitrary threat class. Then the number of ways to order $|C_i|$ messages within the class is $|C_i|!$. Over all k threat classes, the number of possible orderings is $|C_1|! * |C_2|! * ... |C_k|! = \Pi\, |C_i|!$, $i = 1,k$.

Example. Below is a diagram showing 9 messages coarsely quantized into 5 levels of threat. Theorem 1 asserts 9! = 362,880 *possible* orderings on this message stream. Theorem 2 says that this number can be reduced to 1! * 1! * 3! * 3! * 1! = 36 *feasible* filterings.

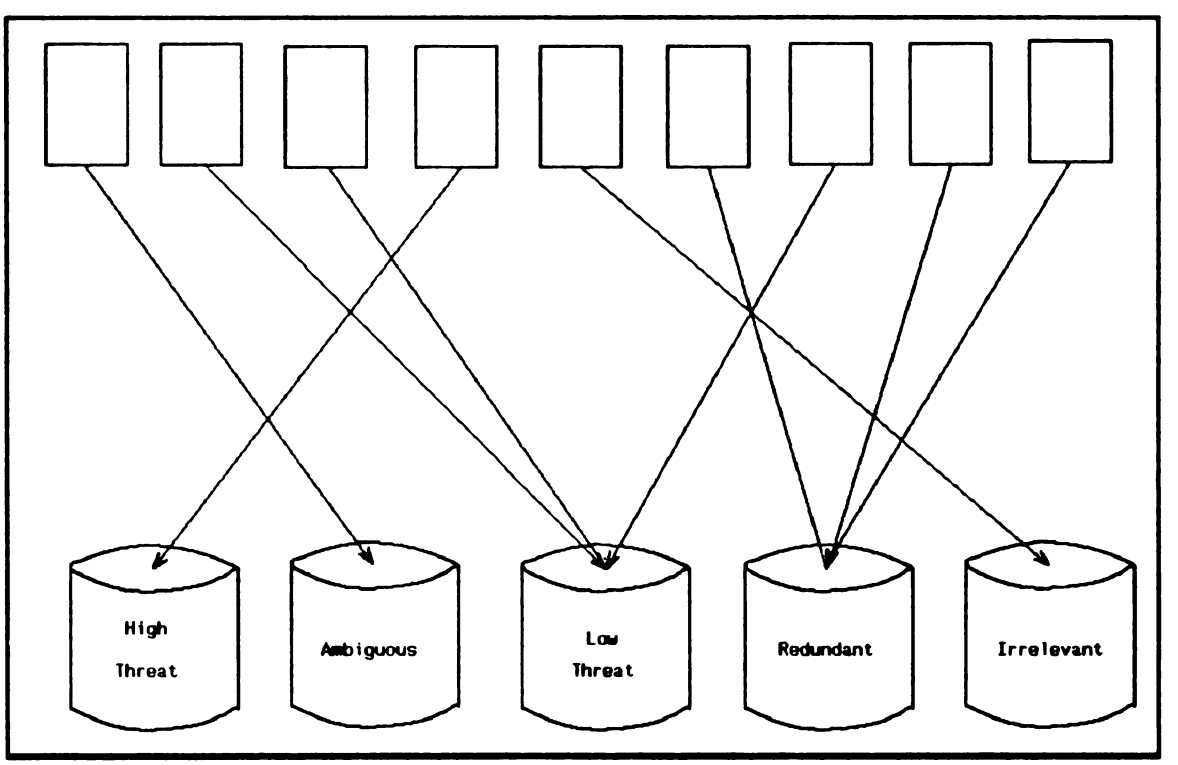


Figure 8. A Coarse Threat Quantization Scheme on 9 Messages. Application of Theorem 2 Yields 36 Feasible Filtering Solutions on this Message Stream.

Message Processing Sources of Error and the Potential for Recovery.

Messages may contain two distinct data structures: a statistical feature vector, and an excerpt of language uttered by a human being who in some way interacts with the object characterized by the feature vector. Machine processing of messages therefore involves comparing and contrasting feature vector data, together with natural language processing. These two types of reasoning processes are sufficiently diverse that mainstream technology thrusts in each area have been pursued in parallel for several decades, with one thrust being in the statistical pattern recognition arena, and the other in computational linguistics. Both technologies continue to produce new research, and each suffers from its own peculiar form of error. It is instructive to play the devil's advocate and construct a taxonomic error tree, which graphically portrays the ways in which an automated message processing system may be fooled, either by errors in the message data, or by faulty reasoning about the data:

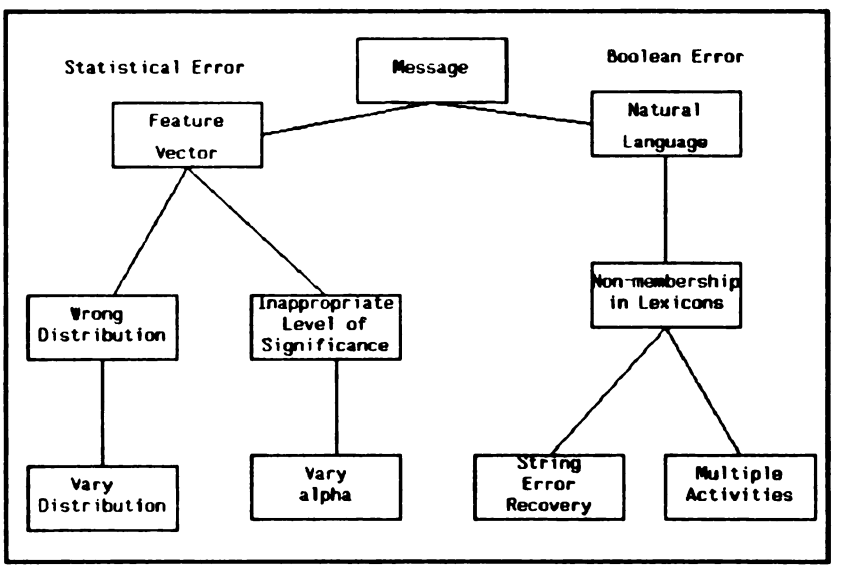


Figure 9. Message Processing Error Forms and Recovery Techniques.

Feature vector error is generally attributable to measurement error of the sensor which gave rise to the feature values, but can also occur during statistical testing because of faulty modeling. Due to the limited sensitivity of the sensor working within the constraints of terrain and other sources of interference, any particular attribute value must be characterized by an error ellipse probable (EEP),

with the semi-major axis along the perceived line-of-bearing, and the semi-minor axis sweeping across the arc described by the angular resolution of the direction-finding capability of the sensor.

However, there is also an error associated with modeling the statistical distributions of the geodynamic objects which the sensors are attempting to measure. It may be that an object's location is inappropriately characterized by a normal distribution, whereas if the probable direction of movement is known *a priori*, it may behoove the system modeler to utilize some distribution which is conveniently skewed in the direction of the motion. Conversely, if it is known that an object is currently stationary, it is advisable to ensure that the distribution used for modeling possesses a bell shape.

Yet another source of error when performing statistical tests with feature vector data is the problem associated with hardwiring a statistical level of significance to a particular test. A Boolean decision is made about the null hypothesis based on the outcome of this test. For example, it may be the case that a test of means fails at the .95 confidence level, and therefore the null hypothesis is rejected out of hand. A less biased approach would be not to make a deterministic decision about the truth of the null hypothesis, but rather post an indication of how *well* the test was passed, or what level of significance would *guarantee* that the test is passed.

The worst-case branching factor of introducing two levels of relaxation into statistical testing is m X n, where m is the number of distributions used to model phenomena, and n is the number of levels of significance over which the tests are conducted. Knowledge-based statistical testing permits an intelligent ordering of the tests, so that the most likely distribution (based on *data-driven* knowledge about the phenomenon) is selected to be checked first. For example, a check for a moving object's location may pass a chi-square test of means at the .95 confidence level, yet not pass a Gaussian test until the level of significance is dropped to a .50 level. The more powerful the search knowledge, the less costly the relaxation process. When the data is well-modeled and sensor measurement error is at a minimum, a cost of 1 is enjoyed, since the appropriate distribution is selected immediately, and the highest confidence level test of means (for the given distribution) is passed.

Because any statistical test of a null hypothesis will be passed (no matter what the distribution) if the confidence level is sufficiently low, it is not prudent from a decision-theoretic standpoint to use a depth-first search during the two levels of relaxation. Instead, it makes sense to start with a high confidence level, breadth-wise test across an intelligently ordered menu of distributions for acceptance of the null hypothesis, and then decrement to a lower confidence level if all tests are failed.

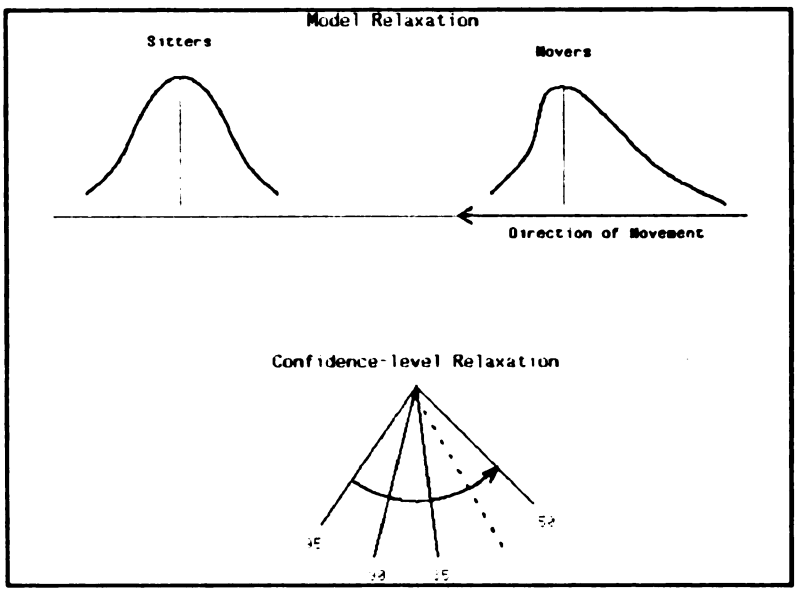


Figure 10. Relaxing Both Distribution Constraints Together With Levels of Significance to Enhance Statistical Testing.

Natural language processing, independent of the particular grammar used, also is subject to different forms of error. The problems of ambiguous words, anaphora, ellipsis, and prepositional phrase attachment are four areas which continue to produce thesis-quality research. Parsers work with sets, whether they be sets of parts of speech, sets of case frames for verbs, or sets of semantic primitives. In some form or another, whether syntactic or semantic, all possible words and actions are partitioned into cells (lexicons), each of which represents some generalized concept about a grammar, or more generally about the world. Error in the most fundamental sense can occur in two ways, just as in statistical testing: a string may fail to be inserted into its proper cell; or it may be inserted into an improper cell.

If a string of natural language fails the set membership test for any lexicon during processing, and the string is in fact appropriate to the target domain, then one of two alternative hypotheses may be true: either the system designer failed to install the string into the appropriate lexicon during the knowledge engineering phase, or the string may be mispelled. In the former case, an intelligent natural language parser may be able to use context to deduce the grammatical class of the string (e.g., it is frequently possible to guess that a test string is a location). If on the other hand the string is misspelled, it may be computationally feasible to recover if the

error is not too serious. The following table enumerates the number of ways that generic types of string error may occur during transmission, followed by the number of strings which a machine must brute-force generate to guarantee recovery.

Some Preliminary Results Regarding n-character[1] String Error Recovery

Type Error	Possibilities[2]	Recovery Combinatorics[3]
transposition	$n - 1$	$n - 1$
k-extra-letters	$(n + k)C_k * 26^k$	$(n + k)C_k$
k-dropped-letters	nC_k	$nC_k * 26^k$
k-wrong-letters	25^k	$nC_k * 25^k$
k1-drops-and-k2-adds[4]	$nC_{k1} * (n - k1 + k2)C_{k2} * 26^{k2}$	$(n - k1 + k2)C_{k2} * (n + k2)C_{k1} * 26^{k1}$
k2-adds-and-k1-drops[4]	$(n + k2)C_{k2} * 26^{k2} * (n + k2)C_{k1}$	$(n + k2)C_{k1} * 26^{k1} * (n + k2)C_{k2}$

[1] Assuming for didactic reasons that a character is a member of the English alphabet

[2] The number of ways that the error can happen in the world.

[3] The number of strings which a machine must generate to guarantee recovery.

[4] The processes of dropping and adding are obviously not commutative.

Implications of the String Error Combinatorial Expressions.

All string error can be explained in terms of linear combinations of added or dropped characters. From the above table, it can be seen that guaranteed recovery from errors of the type indicated in the last four rows requires an algorithm of *exponential complexity*, since an exponent appears in the recovery combinatorics column. It has been shown elsewhere [G1] that for fixed source and destination strings and a finite number of operations, that the destination string can be derived from the source string in polynomial time, given that characters are corrected one at a time rather than in groups of k.

Conclusions.

Due to context-sensitivity, the topic of message filtering cannot be broached without addressing the more fundamental problem of *recognizing events* pointed to by message evidence. To this end, a formal theory of event recognition is being developed, complete with a treatment of the computational aspects of implementation. Formal definitions have been developed for such concepts as *constituent phenomena, suspicion-arousal, suspicion-confirmation,* weak and strong-sense *event revocation*, *event stasis*, and the *threat* of an event. Combinatorial expressions have been derived for the number of *feasible* ways to filter a stream of n messages; the branching factor introduced by permitting both distribution-level and confidence-level relaxation during statistical tests of means; and the number of machine-generated strings necessary to *guarantee* recovery from generic forms of string error encountered during natural language processing.

Future Directions of the Research.

Work shall continue on developing a coherent theory to explain message-driven event recognition, with the ultimate goal being to filter a stream of messages which are providing clues to the events. Although the work thus far has striven to explain how a human decision maker suspects and confirms hypotheses while handicapped with sparse data, the theory remains flawed because it is incomplete. New work shall focus on an epistemology of reasoning with the constituent phenomena which comprise an event. Currently driving the research is the realization that a human problem solver frequently tests the truth of an unsupported clause belonging to a constituent phenomenon by posing it in the *subjunctive* voice, because by so doing the truth of a significant portion of the other constituent phenomena may be induced, especially when they were for all intents and purposes already true but for the lone dissension.

Implementation Issues.

The objects characterized by feature vector data in many applications may be represented by a taxonomic hierarchy of semantic activities. To limit search, a message router has been developed in Interlisp to extract the list of possible activities alluded to by a message. The generic Conceptual Structures Representation Language

(CSRL) developed by Ohio State University [B1] is being utilized as a rapid prototyping tool to further process the message by invoking the set of specific functional parsers pointed to by the router list. The natural language system must of necessity be Type C, which means that the beliefs and intentions of the communicators are taken into account [H2]. As such, recent research on planning [G3, K1, L1] is being investigated to bolster the Type C NLP knowledge base, and to enhance the control of the parsers. Since an event is defined in terms of constituent phenomena, which are themselves defined in terms of a map, spatial representation of the objects is crucial. There has been some commendable work undertaken to represent the *relative* positions of objects described with natural language [H1], but much remains to be done, especially in bringing such a spatial configuration together with the *absolute* description conveyed by a map. A companion document is in preparation to describe the implementation which is currently underway.

Bibliography

[B1] Bylander, Tom, and Sanjay Mittal, *"CSRL: A Language for Classificatory Problem Solving and Uncertainty Handling"*, AI Magazine, Vol. 7, No. 3, 1986.

[C1] Charniak, Eugene, *"The Bayesian Basis of Common Sense Medical Diagnosis"*, 1983 AAAI Proceedings, pp. 70-73.

[C2] Cronin, Terence M., *"Symbolic Filtering of Message Streams Using Artificial Intelligence Techniques"*, Proceedings of the 1984 US Army Science Conference, June 1984.

[D1] de Kleer, Johan, *" Problem Solving with the ATMS"*, Artificial Intelligence 28, 1986, pp. 197-224.

[D2] Doyle, Jon, *"Some Theories of Reasoned Assumptions: An Essay in Rational Psychology"*, Department of Computer Science CS-83-125, Carnegie-Mellon University, Pittsburgh PA, 1983.

[G1] Garey, Michael R., and David S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman and Company, New York, 1979.

[G2] Genesereth, Michael R., Matthew L. Ginsberg, and Jeffrey S. Rosenschein, *"Cooperation without Communication"*, 1986 AAAI Proceedings, pp. 51-57.

[G3] Georgeff, Michael, *"A Theory of Action for MultiAgent Planning"*, 1984 AAAI Proceedings, pp. 121-125.

[G4] Georgeff, Michael P., *"The Representation of Events in Multiagent Domains"*, 1986 AAAI Proceedings, pp. 70-75.

[H1] Hagert, Goran, *"What's in a mental model? On conceptual models in reasoning with spatial descriptions"*, 1985 IJCAI Proceedings, pp. 274-277.

[H2] Hendrix, G.G., and E.D. Sacerdoti, *"Natural Language Processing: The Field in Perspective"*, Byte, September 1981, pp. 304-352.

[J1] Josephson, John, B. Chandrasekaran, and Jack Smith, *"Abduction by Classification and Assembly"*, Ohio State technical report, Columbus OH, 1985.

[K1] Kautz, Henry A., and James Allen, *"Generalized Plan Recognition"*, 1986 AAAI Proceedings, pp. 33-37.

[K2] Khan, Naseem A., and Ramesh Jain, *"Uncertainty Management in a Distributed Knowledge Based System"*, 1985 IJCAI Proceedings, pp. 318-320.

[K3] Kyburg, Henry, *" Full Belief"*, University of Rochester technical report, Rochester NY, 1986.

[L1] Litman, Diane J., *" Understanding Plan Ellipsis"*, 1986 AAAI Proceedings, pp. 619-624.

[M1] Morris, Paul H., and Robert A. Nado, *"Representing Actions with an Assumption-Based Truth Maintenance System"*, 1986 AAAI Proceedings, pp. 13-17.

[N1] Nau, Dana, James Reggia, and Pearl Y. Wang, *"Knowledge-based Problem-Solving Without Production Rules"*, Proceedings Trends and Applications, IEEE Computer, Society Press, 1983, pp. 105-108.

[N2] Nordhausen, Bernd, *"Conceptual Clustering Using Relational Information"*, 1986 AAAI Proceedings, 508-512.

[P1] Peng, Yun, and James A. Reggia, *"Plausibility of Diagnostic Hypotheses: The Nature of Simplicity"*, 1986 AAAI Proceedings, pp. 140-145.

[W1] Waltz, David, *"Phenomenologically Plausible Parsing"*, 1984 AAAI Proceedings, pp. 335-339.

[W2] Winslett, Marianne, *"Is Belief Revision Harder than You Thought?"*, 1986 AAAI Proceedings, pp. 421-427.

Use of the P-Value and a Q-Value in Rejection Criteria

Paul H. Thrasher
Plans and Quality Assurance Directorate
White Sands Missile Range, New Mexico 88002

ABSTRACT

The p-value in a hypothesis test, which is a well established and useful although not universally used statistic, may be supplemented with q-values. Each q-value, just like each possibly designed value of the Type-II risk universally denoted by β, corresponds to a possible value of the tested parameter. The algorithm for calculating q-values is the same as for calculating β's; the inputs that yield β's include the Type-I risk, which is universally denoted by α, and a planned number of measurements (i.e., planned sample size). The corresponding inputs that yield q-values are the p-value and the actual number of measurements (i.e., available sample size). Thus, the q-values contain post-test Type-II risk information in the same manner that the p-value contains post-test information about the Type-I risk.

By using a q-value which corresponds to a particular unacceptable value of the tested parameter, different criteria can be established for the rejection of the null hypothesis. Three alternate criteria imply rejection if (1) (q-value/β), (2) (q-value/β)/(p-value/α), or (3) (q-value/p-value) is greater than unity. The use of any of these three would be a radical departure from the traditional rejection when 1/(p-value/α) is greater than unity. The (q-value/p-value) criterion is independent of α, β, and the planned sample size because both the p-value and q-value depend only on the results of experimental measurements. All three of these alternate criteria

---

Comments by panelists Drs. Kaye Basford and W. T. Federer are at the end of this artical.

lead to trends in critical region size which differ from the trend resulting from the traditional criteria. Replacement of the traditional rejection criterion, with one of the proposed alternate criteria or a decision procedure incorporating rational from the alternate criteria, could significantly influence government and contractor relations and the products or services involved.

1. Introduction. Hypothesis testing is a widely used procedure for designing and conducting experiments to evaluate a parameter against a standard. In government-contractor relations, the government sets the standard. The acceptability of the contractor's product or service is often determined by a hypothesis test.

a. The basic procedure is to:

(1) Formulate a null hypothesis, H_0, relating a parameter, θ, to a standard, θ_0, and

(2) Reject H_0 only if there is sufficient experimental evidence that the assumption is unlikely.

The null hypothesis in government-contractor relations is usually the assumption that the product or service meets the specification. The traditional basis for rejection of H_0, stated in terms of a statistic which is increasingly being reported and interpreted, is that the p-value is too small.

b. The p-value is defined as the probability of an additional experimental result as unlikely as the data. It is a function of two properties of the data:

(1) The used sample size, n_u, which is the actual number of measurements and

(2) Either the measurements or their ranks.

It is also a function of a third factor:

(3) The distribution of all possible measurements under the assumption that H_0 is true.

The traditional rejection criterion, written in a slightly obscure manner which is in the same format as alternate rejection criteria proposed below, is

$$\frac{1}{p \,/\, \alpha} \quad > \quad 1$$

where p is the p-value and α is a predetermined probability of the Type-I risk or the contractor's risk. This is the risk that the contractor's product or service meets the standard but will be rejected by the hypothesis test. Rejection when the p-value is too small is justified by an insistance that the contractor will take a reasonable risk.

c. The p-value provides one aspect of post-test information. Statistics called q-values described the other viewpoint. For the introduction of q-values, see "Proposed Additional Inferential Information During and After Hypothesis Testing", Procedings of the Thirtieth Conference on the Design of Experiments in Army Research, Development, and Testing, Paul H. Thrasher, 1984.

d. Before data is taken, the Type-I risk is supplemented by the Type-II or government's risk. This risk, denoted by β, is the probability of incorrectly failing to reject H_0. It is the companion risk to α, since α is the probability of incorrectly rejecting H_0. Since there are many values of θ for which H_0 is false and the alternate hypothesis denoted by H_a is true, there are many β's. Each β a function of:

(1) α and

(2) The planned sample size, n_p, which is the planned number of measurements.

The β's differ from one another because each is also a function of

(3) A specific value of θ which is not equal to or better than θ_o. If one of these unacceptable parameters, denoted by θ_u, is of particular interest, then it is meaningful to concentrate on one β.

e. After the data is taken, the p-value is supplemented by q-values. The same algorithm used to calculate β from α, n_p, θ_u, H_o, and H_a may be used to find a q-value. A q-value calculation differs from a β calculation in that

(1) The p-value, instead of the original value of α, and

(2) n_u, whether or not this is equal to n_p,

are used in the algorithm. Use of

(3) The same value of the parameter θ_u and the same hypotheses, that were used in the calculation of β, permits direct comparison between β and a q-value. A q-value tends to be greater than a planned value of β if either $n_u < n_p$ or the p-value $< \alpha$. Similarly, making $n_u > n_p$ or obtaining data whose p-value $> \alpha$ tends to yield a q-value smaller than the original value of β.

2. Alternate Rejection Criteria. The traditional rejection criterion is well established. It is not however the only rational decision technique.

a. Instead of requiring the contractor's risk not be too low, one alternate is to require that the government's risk not be too high. This argument replaces the traditional rejection criterion,

$$\frac{1}{p \,/\, \alpha} > 1,$$

with the first alternate rejection criterion:

$$\frac{q}{\beta} > 1.$$

The use of this alternate rejection criteria naturally requires that an unacceptable paramenter, θ_u, must be set along with the standard, θ_o. This first alternate criteria shifts the emphasis completely away from the Type-I risk to the Type-II risk.

b. A second alternate rejection criterion which considers both the Type-I and Type-II risks is to reject if

$$\frac{q / \beta}{p / \alpha} > 1$$

This result may be obtained by multiplying the traditional and the first proposed alternate criteria.

c. A third alternate rejection criterion which concentrates entirely on the post-test information, by considering only the p-value and a q-value while ignoring the specific values α and β, is

$$\frac{q}{p} > R .$$

When the limiting ratio of the post-test Type-II risk to the Type-I risk, R, is set equal to one, rejection occurs under this criterion if the government's risk exceeds the contractor's risks. Other values of R may be used to design a test with other relative emphasis on the government's and contractor's risks. This criterion considers the ratio, $R = \beta/\alpha$, instead of considering α and β separately.

d. The traditional rejection criterion and the three alternate rejection criteria introduced above have contradictory and incomplete attributes.

No single criterion will provide a panacea for all situations. For example, the third alternate criterion may be appropriate when large values of α and β fortuitously cause no large financial or logistic difficulties (e.g., when the contractor can easily rework rejected items and the government can feasibly replace items not functioning properly). If either α or β must be small however, the third alternate may be inappropriate (i.e., setting R may not provide the desired values of α or β). In this case, the second alternate may be desired or perhaps a simultaneous application of the traditional and first alternative criteria may be warranted. Satisfying the second alternate criterion does not gurarantee that the traditional and first alternate criteria are simultaneously satisfied. All of the criteria must be scrutinized individually. Each, or each combination, must be justified or discarded on the basis of its own characteristics. Only one, or one combination, can be used in any particular hypothesis test.

3. Critical Regions In One Example. The critical regions, defined as intervals in which data implies rejection of H_0, may be found for any situation in which traditional hypothesis testing is done. The specific situation used in this section is one used in the previously referenced presentation at the Thirtieth Conference on the Design of Experiments. Basically, this situation has a standard, σ_0^2, and as unacceptable level, σ_u^2, for the variance, σ^2, of a random variable which is assumed normal. The Chi-squared distribution then describes $(n-1)\, s^2/\sigma^2$ where s^2 is the sample variance.

a. This example yields the critical regions plotted in figures 1 through 13. For this example at least, the trends in the critical regions of the proposed alternate criteria are significantly different than those of the

traditional criterion. Two very evident trends are seen by looking at the lower ends of the critical regions which are called the critical points. For the traditional criterion with reasonably low values of α, the critical points,

(1) All correspond to measurements better than the standard and

(2) Decrease as the sample size increases.

b. For some situations involving the alternate criteria, the critical points

(1) Correspond to measurments worse than the standard and

(2) Increase as the used sample size increases.

c. Both of these properties are naturally disturbing to hypothesis testers who are used to the normal criterion with low values of α. However, both actually occur in the traditional criterion when the value of α is made large enough.

d. The figures describing $q/\beta > 1$ and $(q/\beta) / (p/\alpha) > 1$ are much more complicated than those describing $1 / (p/\alpha) > 1$. However, the figure describing $q/p > 1$ is as simple as the figures describing $1 / (p/\alpha) > 1$. This occurs because both p and q are independent of α, n_p, and β.

4. Generalizations, Extensions, and Applications.

a. In figures 1 through 13, there is an inversion of trends between the traditional criterion and any choice of alternate criterion. This appears to be a general property for this particular hypothesis test. Much theoretical and simulation work needs to be done, however, before extending this statement to other hypothesis tests.

b. If any of the alternate criteria are applied in government-contractor relations, significant changes will occur in the way business is done. It would be entirely possible, for example, that contractors would be in the governments's present position of wanting an increased sample size. Using the traditional criteria, the government is vulnerable when $n_u < n_p$; using an alternate criteria, the contractor may feel vulnerable when this change in the planned number of tested items occurs. This inversion is certainly significant. It could lead to a significant decrease in cost and/or increase in quality of products or services that the government procures from contractors.

c. A secondary benefit from using any of the alternate criteria is that the government would be forced to specify an unacceptable parameter as well as a standard. This requirement would yield an improvement in management.

d. The choice of a criterion or perhaps a set of simultaneous criteria for any situation must consider the costs of production, testing, and use of the product or service. This consideration will undoubtable be complicated and many faceted. The measurement of cost may not even be straightforward. (e.g., dollars, time, lives, and military success may be competing measures of cost.) Nevertheless, the total cost should be minimized by a selection from the possible rejection criteria.

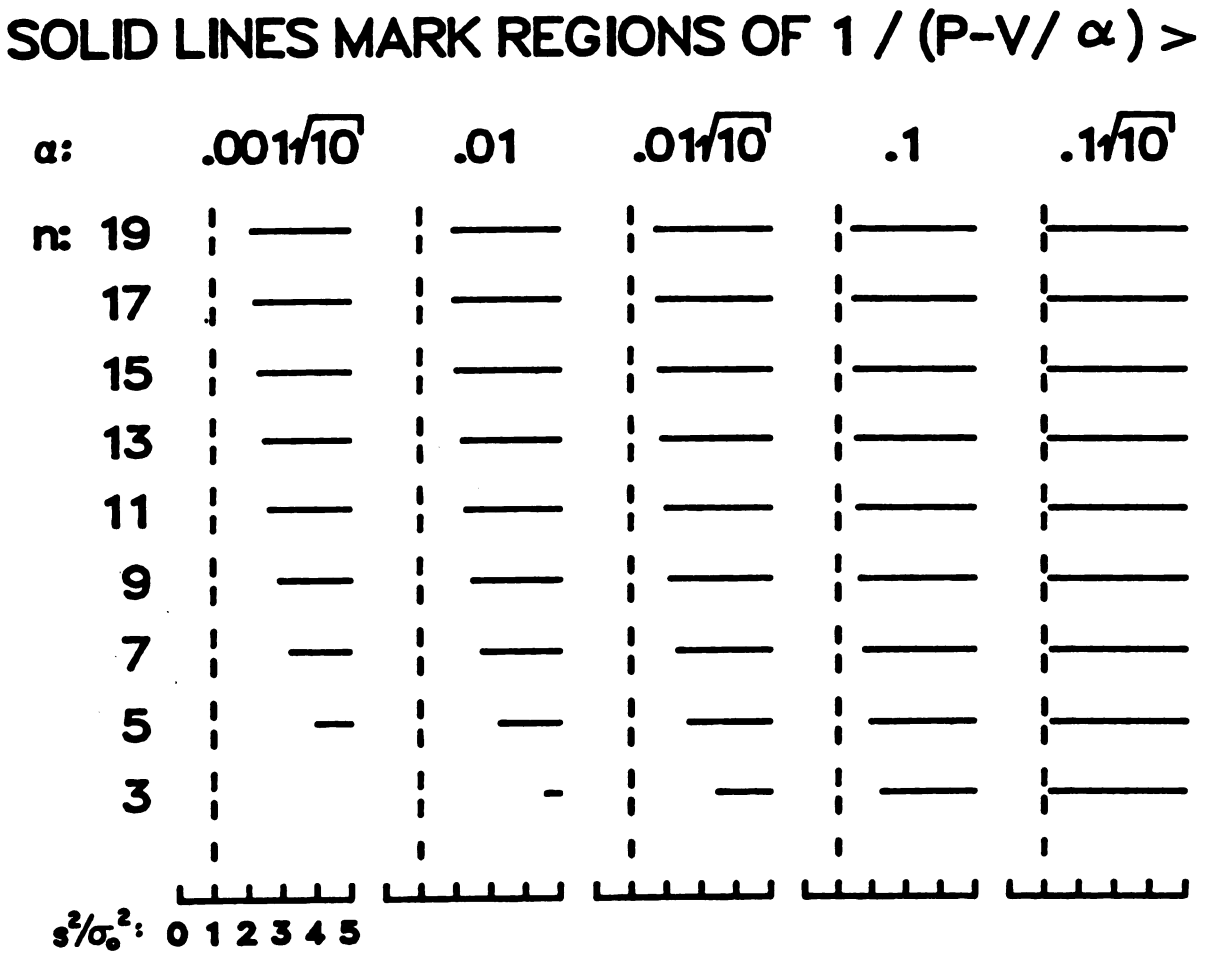


Figure 1. "Normal" relationship between number of measurements, n, and minimum sample variance, s^2_{min}, which implies rejection of the null hypothesis that the variance is at least as small as a standard, σ^2_0. This "normal" decrease in s^2_{min} with an increase in n results from using a rejection criterion based on the ratio of attained to planned Type-I risks, p-value/α, for commonly used values of α.

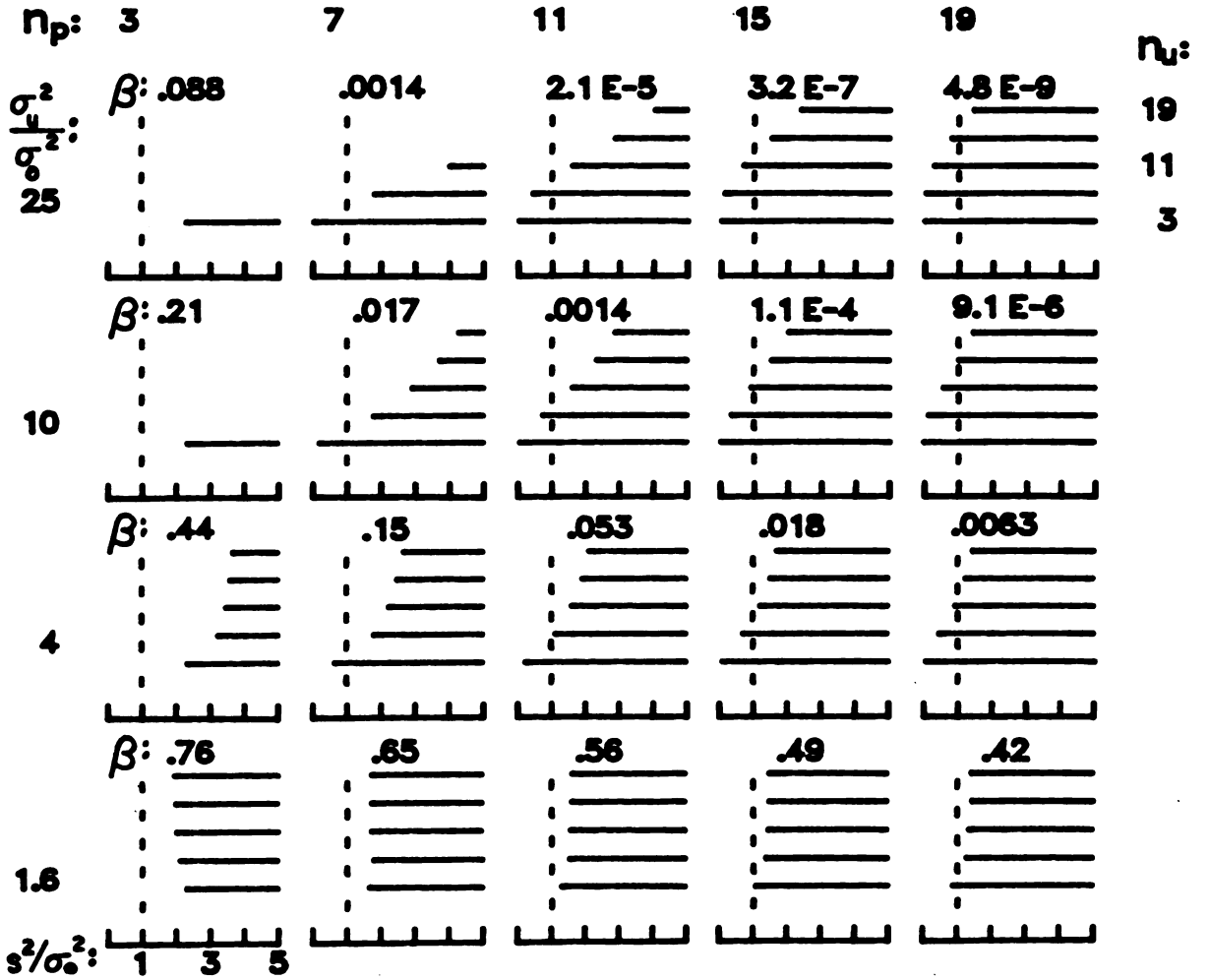


Figure 2. "Inverted relationship between s^2_{min} and used number of measurements, n_u, resulting from alternate criterion based on ratio of attained to planned Type-II risks, q-value/β. This "inverted" increase in s^2_{min} with an increase in n_u occurs for $\alpha = 0.01$ and several combinations of planned number of measurements, n_p, and discrimination ratios of unacceptable to standard variances, σ_u^2 / σ_o^2.

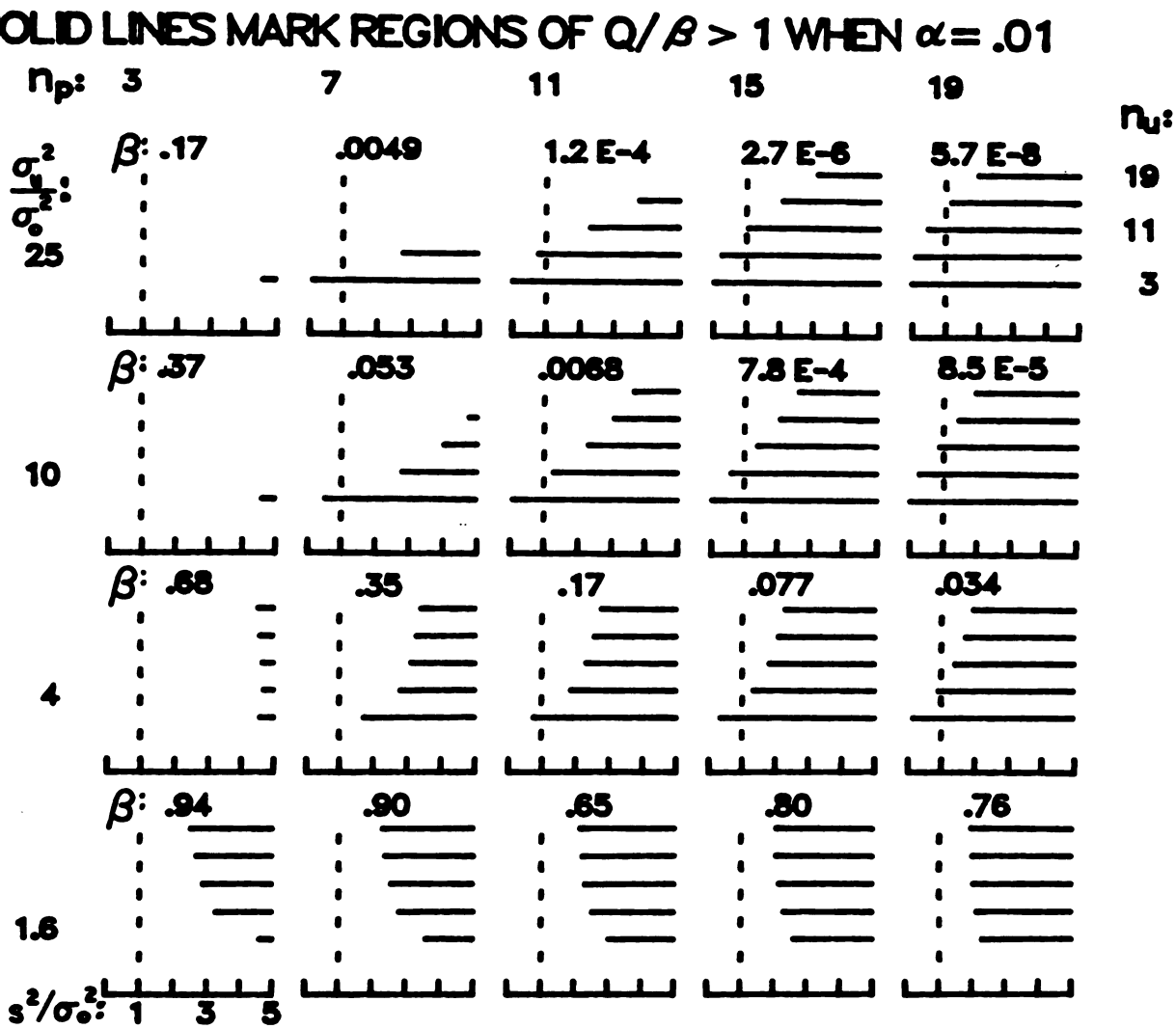


Figure 3. "Inverted" s^2_{min} - n_u relationship for high σ_u^2 / σ_o^2 and "normal" s^2_{min} - n_u relationship for low σ_u^2 / σ_o^2. This split between "inverted" and "normal" relationships occurs for $\alpha = 0.01$ when the q-value/β criterion is used.

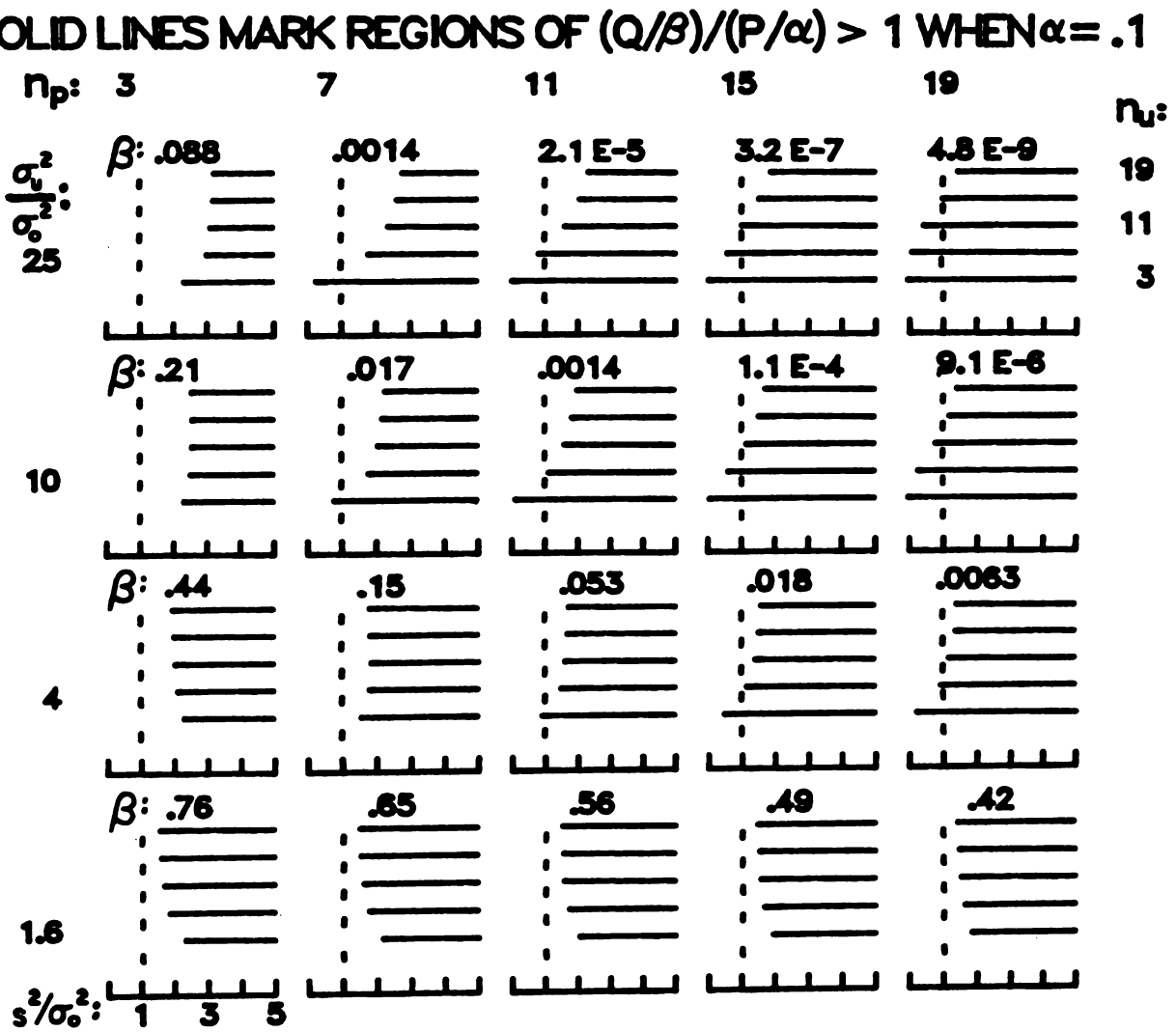


Figure 4. "Inverted" s^2_{min} - n_u relationship for high σ_u^2 / σ_o^2 and large n_p and "normal" s^2_{min} - n_u relationship for low σ_u^2 / σ_o^2 and small n_p. The "inversion" predominates when the criterion is based on (q-value/β)/(p-value/α) and α = 0.1.

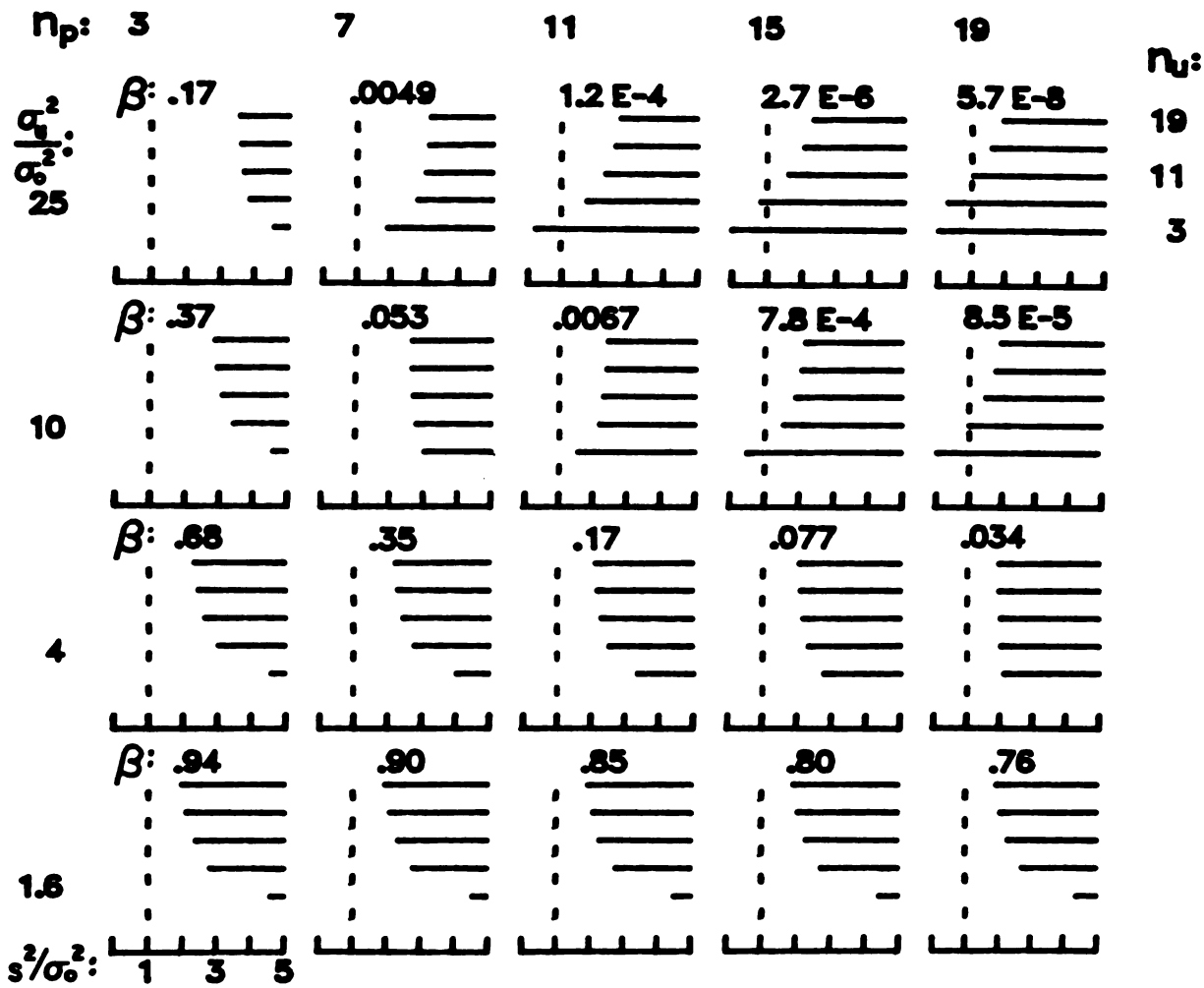


Figure 5. "Normal" s^2_{min} - n_u relationship for low σ_u^2 / σ_o^2 and small n_u and "inverted" s^2_{min} - n_u relationship for high σ_u^2 / σ_o^2 and large n_u. The "inversion" is dominated by "normalcy" for $\alpha = 0.01$ when the (q-value/β/(p-value/α) criterion is used.

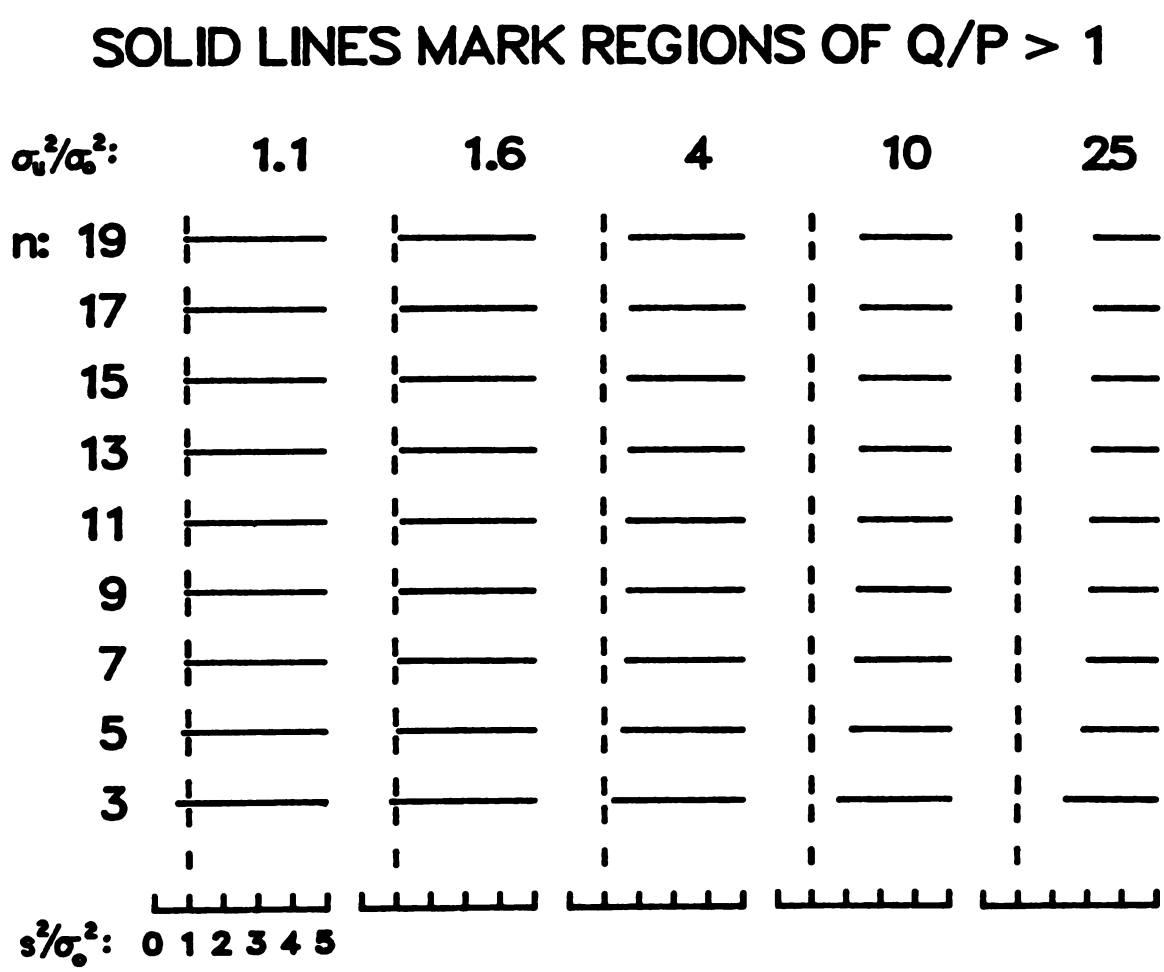


Figure 6. "Independence" of s^2_{min} and n for low σ^2_u / σ^2_o and large n. This "almost constancy " occurs when the criterion is based on the ratio of the attained Type-II to attained Type-I risks, q-value/p-value, which is independent of α, β, and n_p.

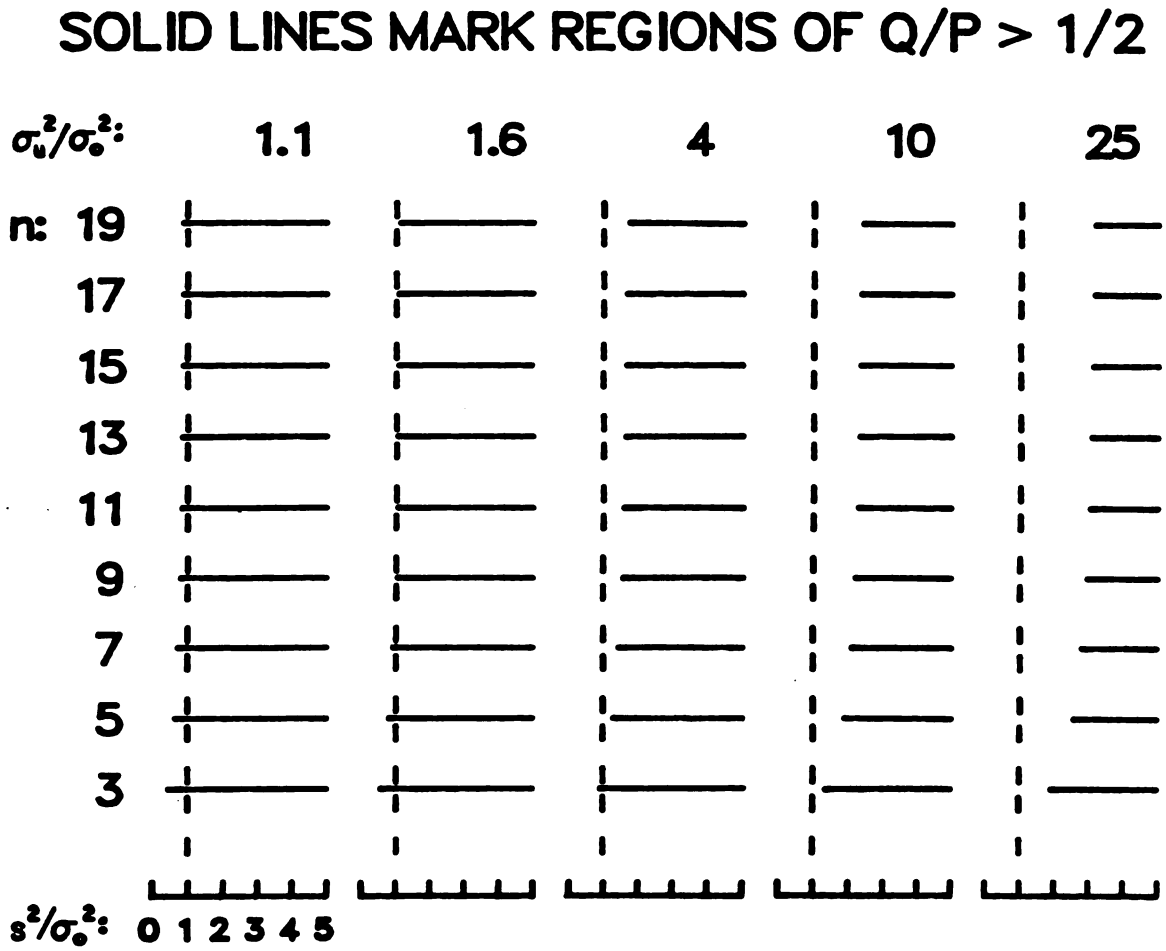


Figure 7. "Diminshed independence" of s^2_{min} and n for low σ_u^2 / σ_o^2 and large n. This "not quite constancy" occurs when the q-value/p-value criterion uses 1/2 instead of 1 as the standard. Rejection is possible from point estimates of s^2 less than σ_o^2 when both σ_u^2 / σ_o^2 and n are low.

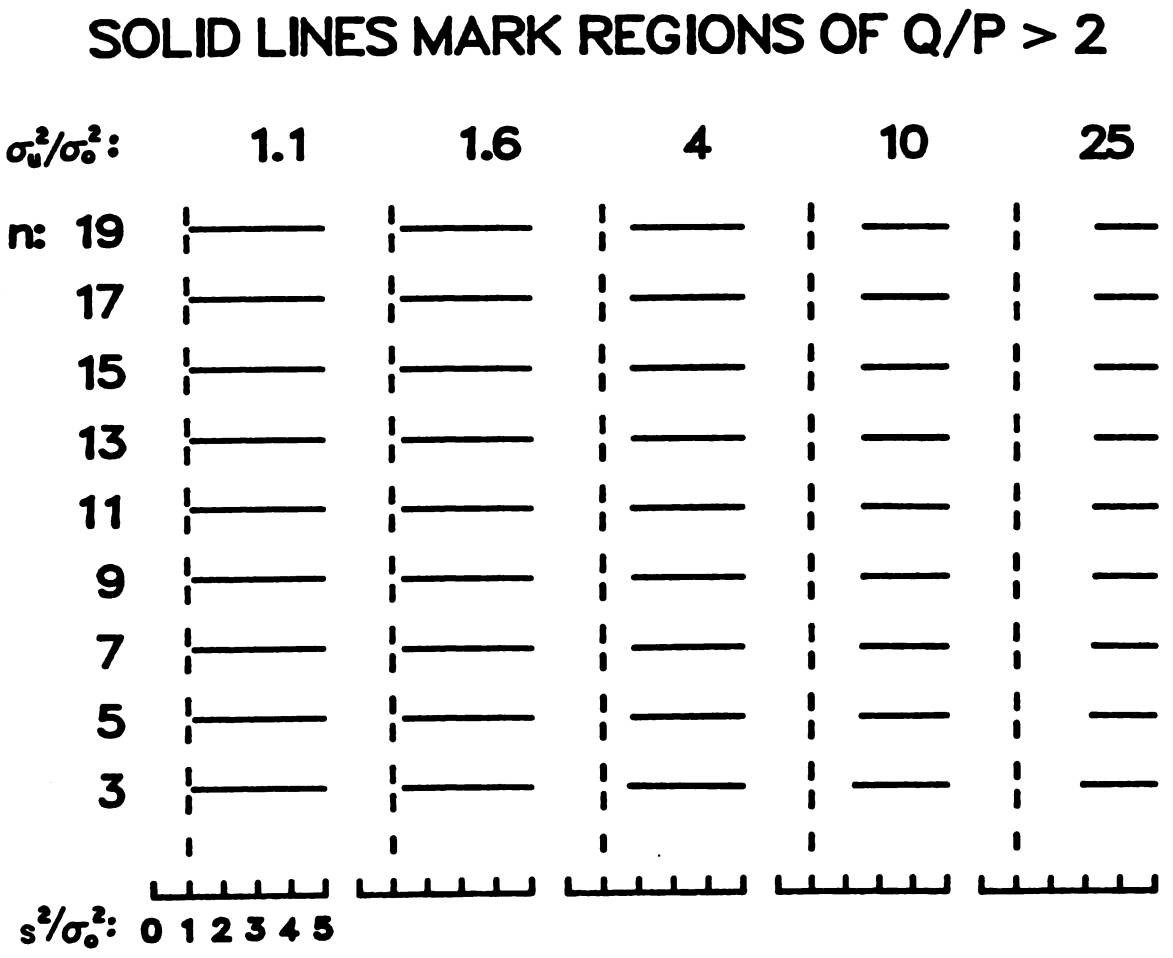


Figure 8. "Enhanced independence" of s^2_{min} and n for low σ^2_u / σ^2_o and large n. This "almost constancy" is closer to actual independence when the q-value/p-value criterion uses 2 instead of 1 as the standard.

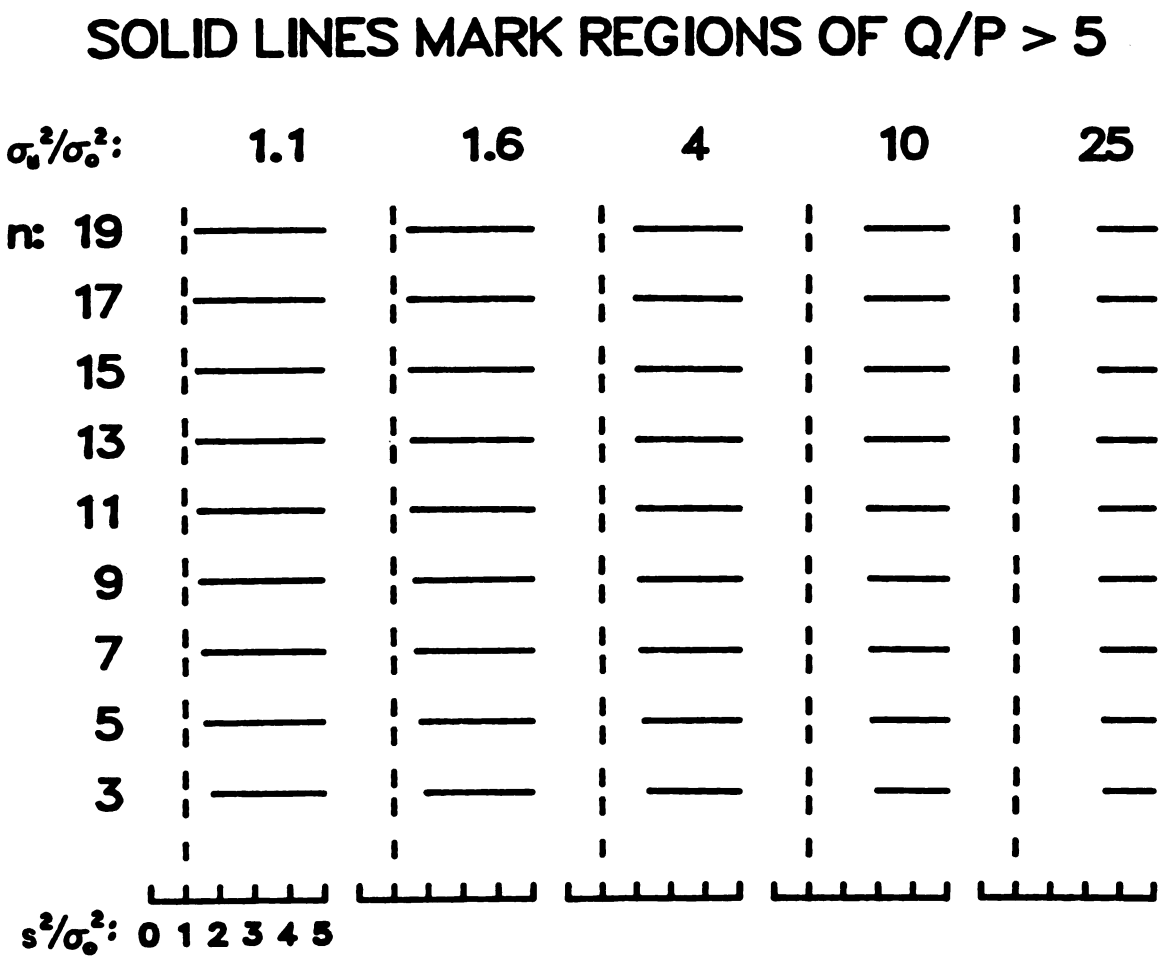


Figure 9. "Diminished independence" of s^2_{min} and n for low σ^2_u / σ^2_o and large N. "Slight normalcy" as opposed to "very slight inversion" results when the q-value/p-value criterion uses 5 instead of 2 as the standard.

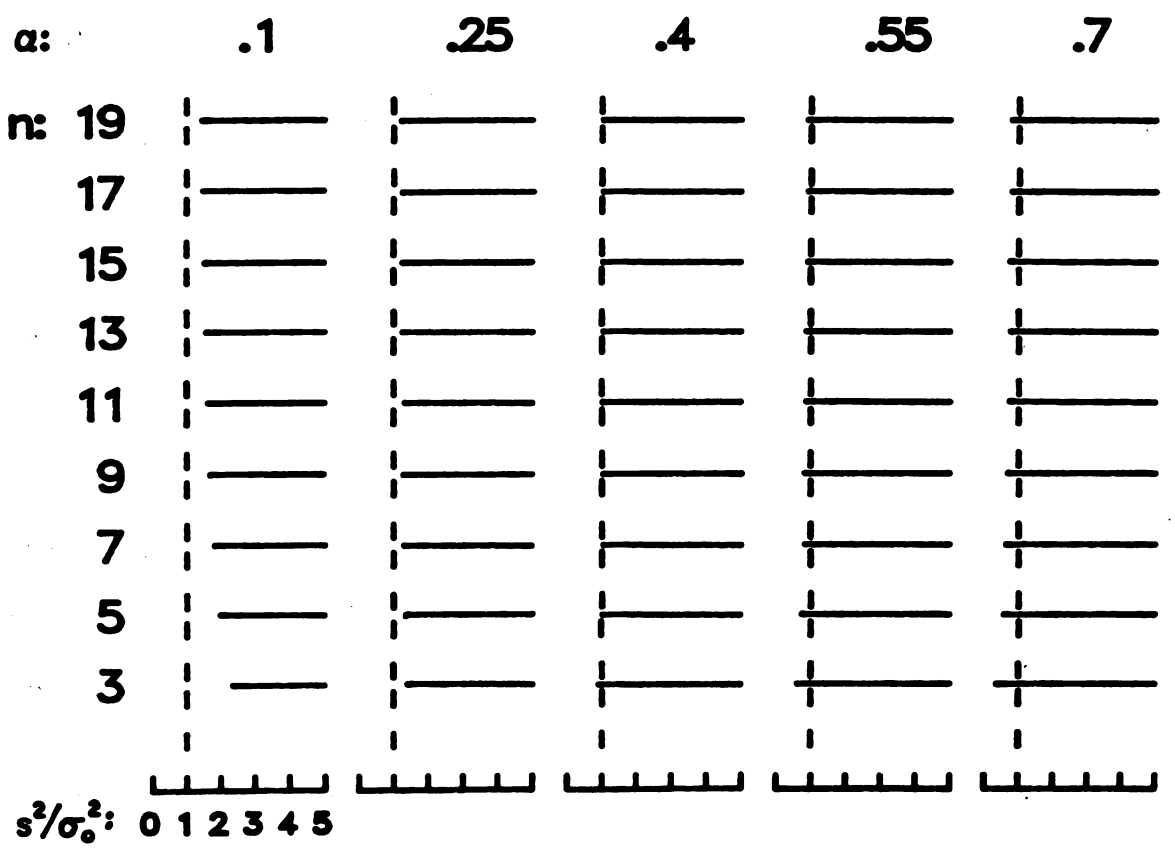


Figure 10. "Inverted" s^2_{min} - n relationship for large α with the p-value/α criterion. The "normalcy" is normal in this traditional criterion only if α is a usually used small number. Also, using large numbers for α results in the possibility of rejection from point estimates of s^2 less than σ_0^2.

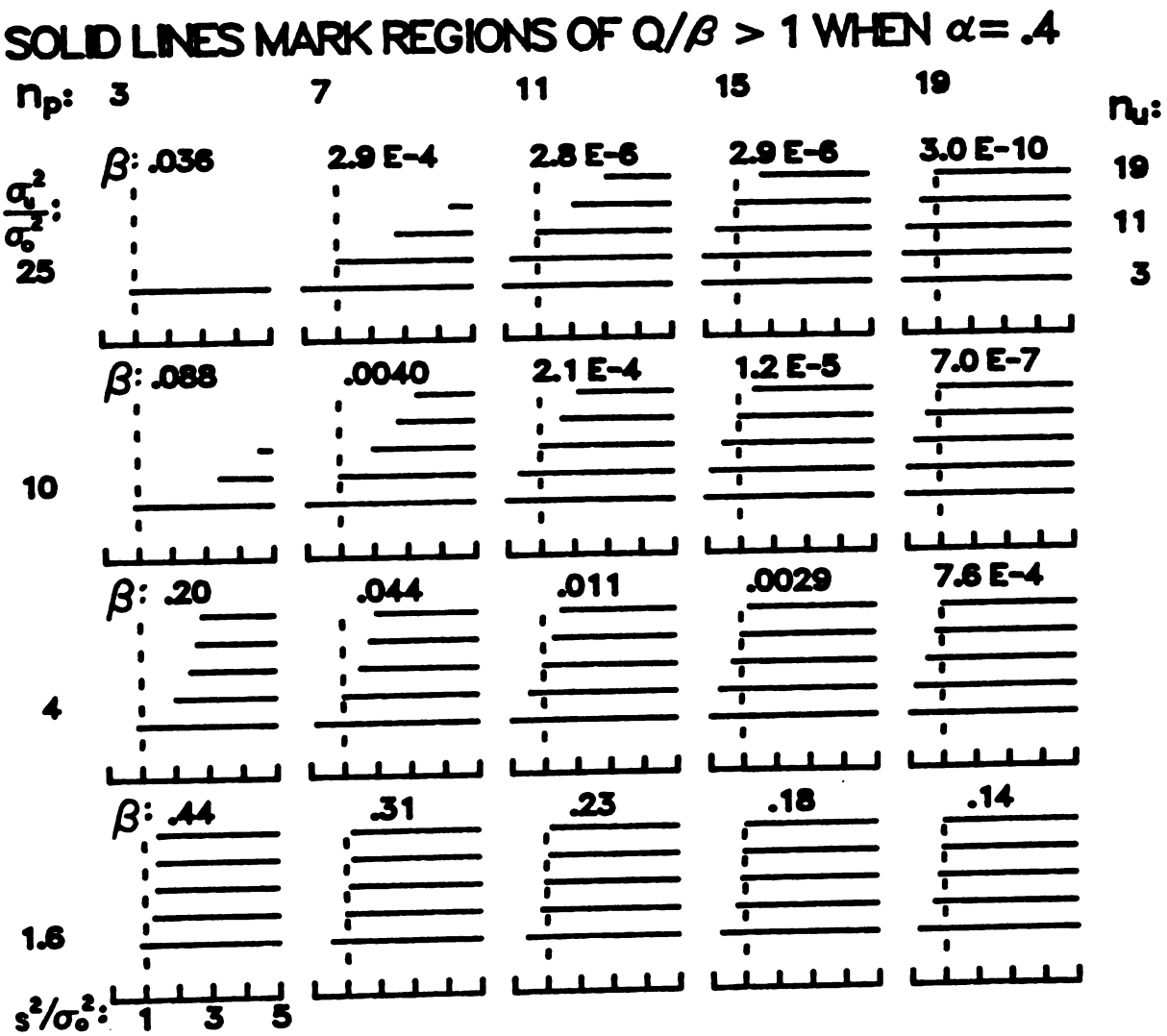


Figure 11. "Inverted" s^2_{min} - n_u relationship with q-value/β criterion and $\alpha = 0.4$. This "inversion" strongly predominates when a high α is used in this criterion. Also, high α and low β (or high n_p) allow a possibility of rejection if the point estimate of s^2 is less than σ_o^2.

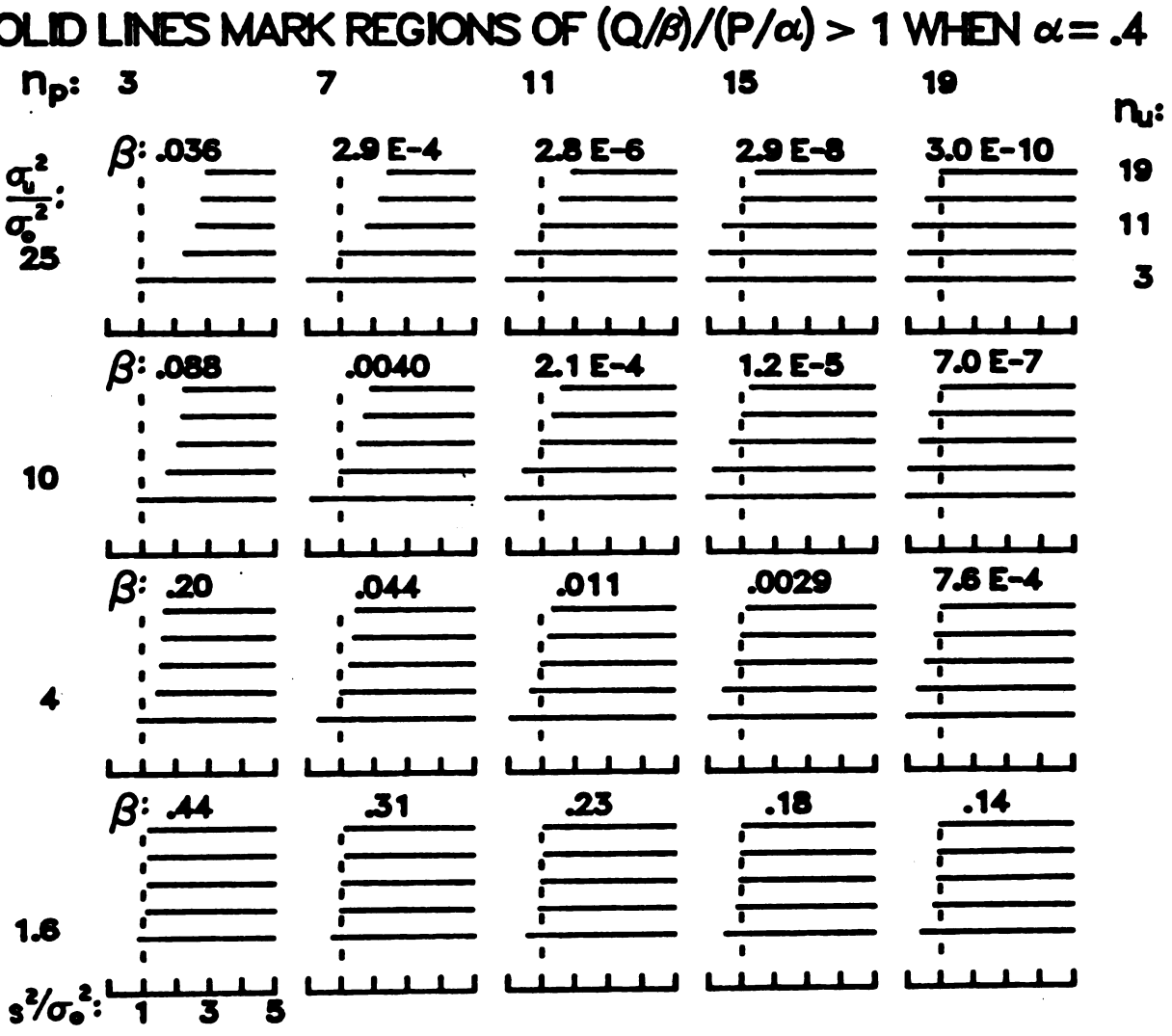


Figure 12. "Inverted" s^2_{min} - n_u relationship with (q-value/β)/(p-value/α) criterion and α = 0.4. This "inversion" strongly predominates when a high α is used in this criterion. Again, high α and low β (or high n_p) allow a possibility of rejection if the point estimate of s^2 is less than σ_0^2.

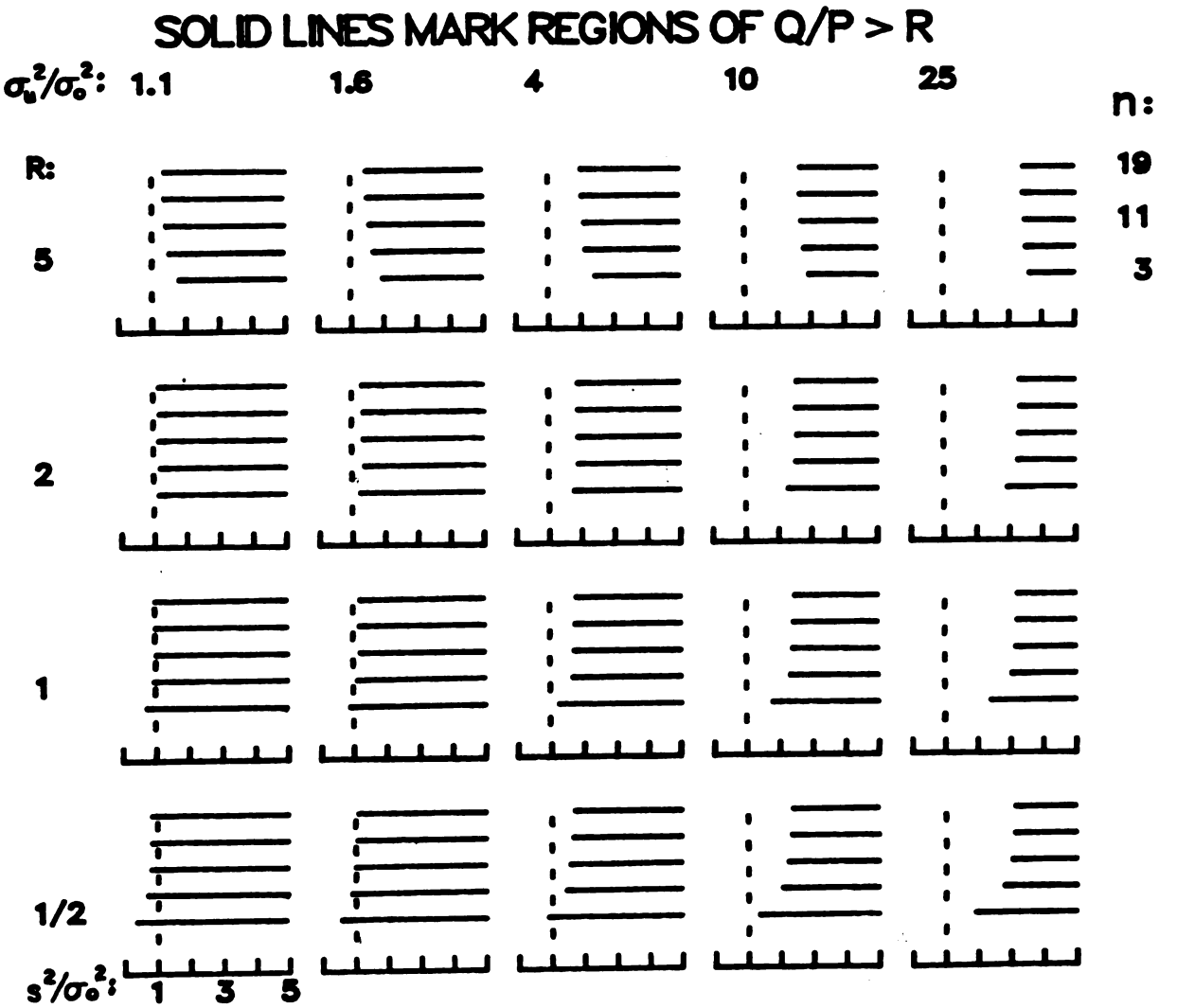


Figure 13. Combined illustration of s^2_{min} - n relationships when q-value/p-value alternate criterion is used with different standards, R. This complilation of figures 6 through 10 shows more "independence" between s^2_{min} and n (and also α, β, and n_p) than the traditional p-value/α criterion or the alternate q-value/β or (q-value/β)/p-value/α) criteria. Also, the possibility of rejection when $s^2 < \sigma_o^2$ occurs only when R, n, and σ_u^2 / σ_o^2 are small.

COMMENTS BY PANELISTS DR. KAYE BASFORD AND PROFESSOR W. T. FEDERER ON THE FOLLOWING ARTICAL

Use of the P-value and a Q-value in rejection criteria by

Paul H. Thrasher

U.S. Army White Sands Missile Range.

This paper was not presented but MSI received a copy of the paper subsequently.

Kaye Basford: This is a very interesting paper in which it is suggested that the usual type I error (α) used to accept or reject the null hypothesis be supported by some information on the Type II error (β). Hence instead of a decision being made solely on the p-value, the q-value would also be used.

Dr. Thrasher suggested three alternative criteria for rejection of the null hypothes and studied their behavior for one particular test. The general properties of these decision criteria need to be investigated fcr hypothesis tests with different underlying distributions. Only then could a recommendation he made on the desirability and feasibility of introducing such a criterian. This is a challenging research project which I hope will be taken up in the near future.

INCORPORATING FUZZY SET THEORY INTO STATISTICAL HYPOTHESIS TESTING

William E. Baker
Probability and Statistics Branch
Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT

In many instances the data used in statistical hypothesis testing may be vague or imprecise and, as such, may suggest results that are incorrect. Rank tests, in particular, seem susceptible, since the original data, once ranked, has no further influence on the testing procedure no matter how closely they are grouped. A possible solution is to treat the ranks as fuzzy integers represented by membership functions that indicate the degree to which each rank assumes each integer value. In this paper, a method is suggested for obtaining these membership functions; and the concept is incorporated into an existing rank test. An application of this fuzzy hypothesis-testing procedure is provided.

I. INTRODUCTION

Suppose we have the following set of data:

$$\{-0.888,\ 0.200,\ -1.000,\ -0.417,\ -0.052,\ 0.186,\ 0.067,\ -0.467,\ -0.623,\ -0.181\}\ . \qquad (1)$$

By considering their absolute values, we obtain a set S consisting of ordered pairs,

$$S = \{(1, -0.052),\ (2, 0.067),\ (3,-0.181),\ (4, 0.186),\ (5, 0.200),\ (6, -0.417), \qquad (2)$$
$$(7, -0.467),\ (8, -0.623),\ (9, -0.888),\ (10, -1.000)\}\ ,$$

where the first member of each ordered pair is the ranking (smallest to largest) of the absolute value of the second member of the ordered pair. This type of data is often used in rank tests, nonparametric hypothesis tests which generally examine the mean or median of a distribution or the equality of means or medians of several distributions. Rank tests are sometimes eschewed because once the ranking has been established, the data are treated as though they were equally spaced; and potentially-valuable information concerning the proximity of the data points is discarded. In the preceding example, note that some of the rankings may be tenuous; for example, ranks 3 and 4 could easily have been permuted had the numbers to which they correspond been inaccurate in the third decimal place. Therefore, the degree of accuracy in the ranks is directly related to the degree of accuracy of the original data; and this can sometimes be a problem.

In many applications, the available data may be vague or imprecise, due to a variety of reasons which may include improper calibration of equipment and subjectivity of the experimenter. This, of course, can lead to imprecise ranking of the data and possibly an incorrect conclusion from the resulting hypothesis test. Such data, as well as their ranks, can be represented by fuzzy numbers[1] - a relatively new concept in which a number is described by a central value along with a spread about that value. When applied to ranks, this technique may overcome the previously-mentioned problem inherent in rank tests; and in certain situations this representation will allow for a more realistic approach to hypothesis testing.

II. FUZZY RANKS APPLIED TO THE WILCOXON SIGNED-RANKS TEST

A. Wilcoxon Signed-Ranks Test

The Wilcoxon signed-ranks test is a nonparametric hypothesis test which is generally used to test for equal medians of two distributions. The data consist of paired observations (x_i, y_i) from the two distributions. The differences between the observations, $D_i = x_i - y_i$, are then calculated; and their absolute values are assigned a rank R_i from smallest to largest. Finally, R_i is multiplied by -1 if D_i is negative. The sum of the ranks of the positive differences, $T = \sum R_i,\ R_i > 0$, is the test statistic. If the two distributions have the same median, we would expect about one-half of the D_i's

[1] Zadeh, L.A., "Fuzzy Sets," Information and Control, Vol. 8, 1965.

to be positive. Very high or very low values of T indicate that numbers from the first distribution are consistently higher or consistently lower than those from the second distribution and, therefore, will cause rejection of the null hypothesis of equal medians. The theory behind the test along with tables containing various quantiles of T are provided by Conover[2].

For each ordered pair of the set S, we can consider the second value to be D_i and the first value to be the R_i associated with it. Taking the sum of the R_i's associated with the positive D_i's, we find that T = 2+4+5 = 11. Probability levels for the Wilcoxon signed-ranks test for a sample of size 10 are given in Table 1. Referring to this table, we find that our value of T indicates that there is insufficient evidence for rejecting the hypothesis of equal medians at a 10% level of significance. In this case the probability of T being less than or equal to 11 is 0.0527; and since we are performing a two-sided test (examining T to see if its value is either too low or too high), we double that figure to get the critical level of the test. Had the value of T been 10 or less, rejection of the null hypothesis would have been warranted.

TABLE 1. Probability Levels for the Wilcoxon Signed-Ranks Test Statistic with a Sample Size of 10. *

T	P	T	P	T	P	T	P
0	.0010	7	.0186	14	.0967	21	.2783
1	.0020	8	.0244	15	.1162	22	.3125
2	.0029	9	.0322	16	.1377	23	.3477
3	.0049	10	.0420	17	.1611	24	.3848
4	.0068	11	.0527	18	.1875	25	.4229
5	.0098	12	.0654	19	.2158	26	.4609
6	.0137	13	.0801	20	.2461	27	.5000

T = sum of positive ranks

P = probability that the sum of positive ranks will be less than or equal to T under the null hypothesis

* Since the distribution of T is symmetrical, only one-half of the distribution is tabulated.

B. Fuzzy Ranks

Let $R = \{r_1, r_2, ..., r_n\}$ be a set of elements and Q be a subset of R. Then we can define the characteristic function μ_Q: R→{0, 1} such that

---

[2] Conover, W.J., Practical Nonparametric Statistics, John Wiley and Sons, Inc., 1971.

$$\mu_Q(r_i) = \begin{cases} 1 \text{ if } r_i \in Q \\ 0 \text{ if } r_i \notin Q \end{cases}. \tag{3}$$

If, however, R is the set of men and Q is taken to be the set of old men, there may be some vagueness about the membership of certain r_i in Q. Is a 50-year-old man a member of Q? I used to think so; but now that I'm older, I'm not quite so sure. Suppose we let μ_Q take on values other than 0 and 1; in particular, any value between 0 and 1 so that μ_Q: R→[0, 1].

In this case Q is called a fuzzy subset of R and μ_Q is called the membership function of Q. Each r_i has associated with it a value $\mu_Q(r_i)$ representing a degree of membership in Q. The closer this value is to one, the more completely the associated r_i is a member of Q. Numerical data can be represented by equating R with the set of real numbers, in which case Q is called a fuzzy number.

In this application we will examine fuzzy numbers and, in particular, fuzzy integers since we are concerned with ranks. A fuzzy number will be represented by a membership function quantifying the degree to which it takes on any specific value. Figure 1 shows a membership function μ for "fuzzy six". This function assumes its maximum value at six, $\mu(6) = 1$; the closer any number is to six, the higher its degree of membership in "fuzzy six". When we examine fuzzy ranks, the membership functions will be discrete, since our interest will be only in the degree of membership for integer values.

This membership function is not unique; rather, it is subjective - determined by the user and based on his perception of the vagueness of the data. In order to fully utilize this methodology, the Extension Principle[3] permits definition of a mathematical operation f on two fuzzy numbers. It states that if X is a fuzzy number with membership function $\mu_X(x)$ and Y is a fuzzy number with membership function $\mu_Y(y)$, then Z = f (X,Y) is a fuzzy number with membership function

$$\mu_Z(z) = \max_{\substack{x,y \\ f(x,y)=z}} \min [\mu_X(x), \mu_Y(y)]. \tag{4}$$

---

[3] Zadeh, L.A., "The Concept of a Linguistic Variable and its Application to Approximate Reasoning I, II, III," Information Sciences, Vols. 8, 9, 1975.

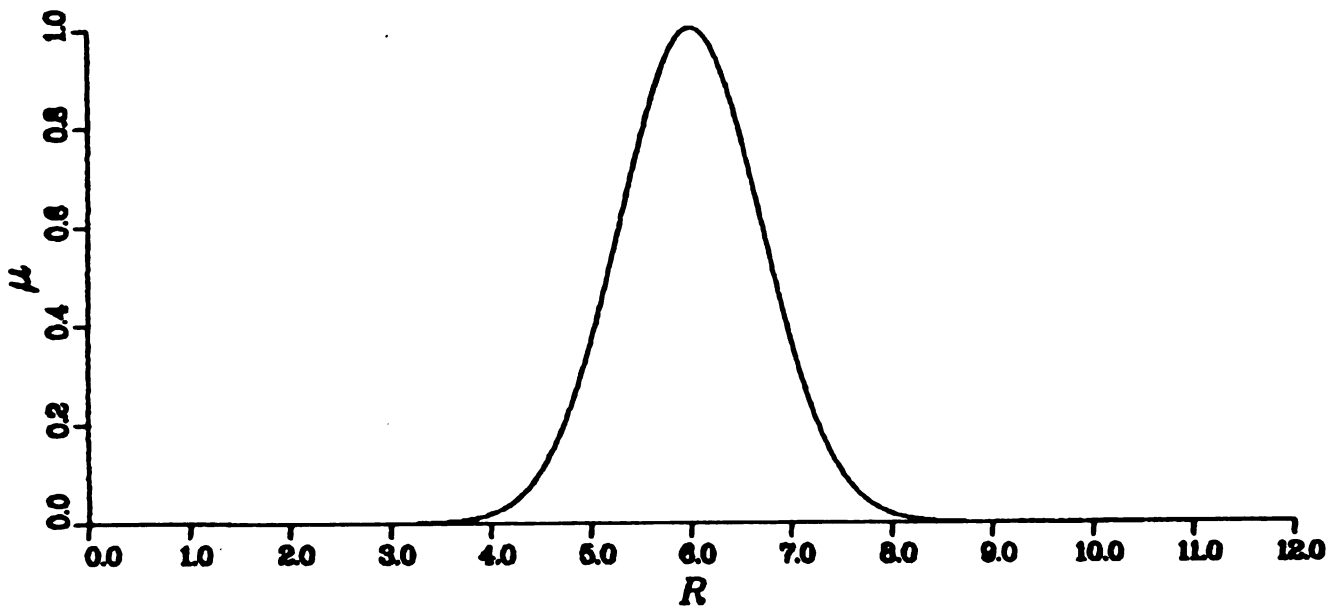


Figure 1. Membership Function of Fuzzy Six.

Figure 2 shows some membership functions established for the absolute value of three of the members of the original data set (-0.181, 0.186, 0.200). Recall that the set S contained ordered pairs of the form (I_X, X) where X was a number from the original data set and I_X was the rank associated with the absolute value of X. The shapes of these membership functions are symmetric and triangular with a spread equal to ten percent of the largest value in the data set (remember that these are modeling decisions). Hence, the membership value of "fuzzy 0.181" is non-zero from 0.081 to 0.281 and has its zenith at 0.181.

We can define a membership function for the first member of each ordered pair - the rank denoted by I_X - as follows:

$$\mu_{I_X}(I_Y) = \max_{z \in R} \min [\mu_X(z), \mu_Y(z)] . \quad (5)$$

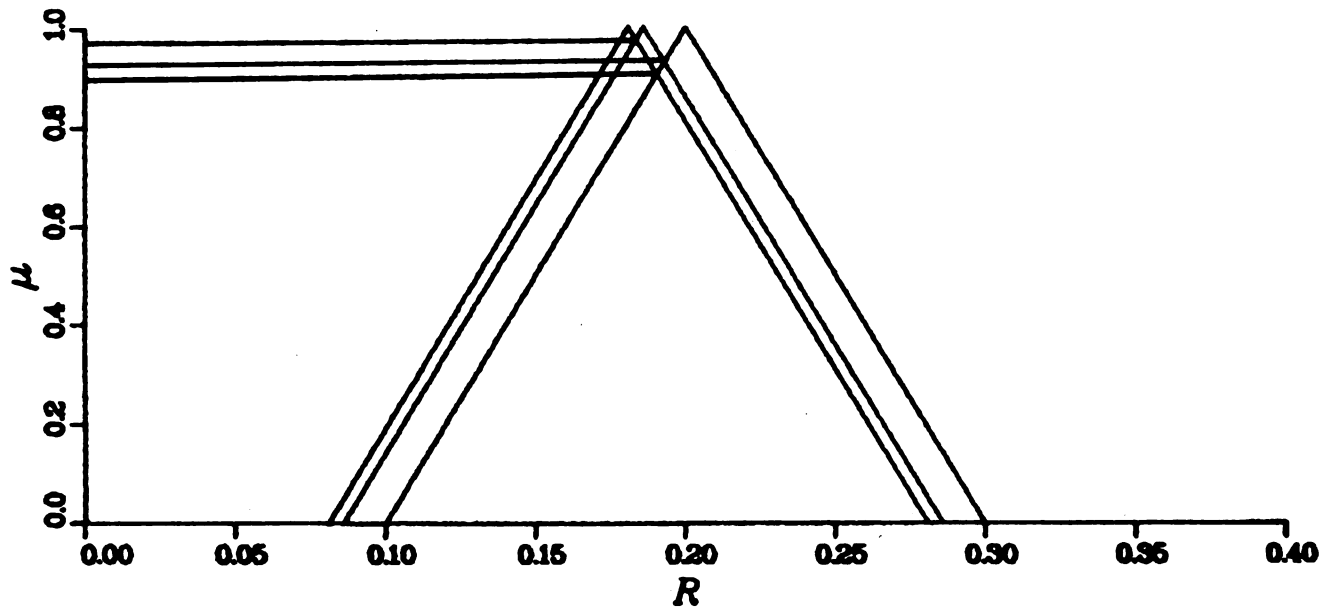


Figure 2. Membership Functions of a Portion of the Original Data Set.

This equation provides the membership value for I_Y in "fuzzy rank I_X". Thus, in Figure 2, the top horizontal line intersects the ordinate at a point equal to $\mu_3(4)$, the middle horizontal line intersects the ordinate at a point equal to $\mu_4(5)$, and the bottom horizontal line intersects the ordinate at a point equal to $\mu_3(5)$. This definition of the membership function for the fuzzy ranks produces the following properties:

$$\mu_{I_X}(I_X) = 1, \tag{6}$$

$$\mu_{I_X}(I_Y) = 0 \text{ if } \mu_X(x) \text{ and } \mu_Y(y) \text{ do not intersect, and} \tag{7}$$

$$\mu_{I_X}(I_Y) = \mu_{I_Y}(I_X)\,. \tag{8}$$

Figure 3 shows the membership functions for the entire set of original data. The ordinate values of their points of intersection are listed in Table 2. These, of course, are the values of $\mu_{I_X}(I_Y)$ shown in Equation 4 and define the membership functions of the fuzzy ranks of the data, such functions being discrete since the ranks can take on only integer values. Note that the table is symmetric, a result of Equation 7.

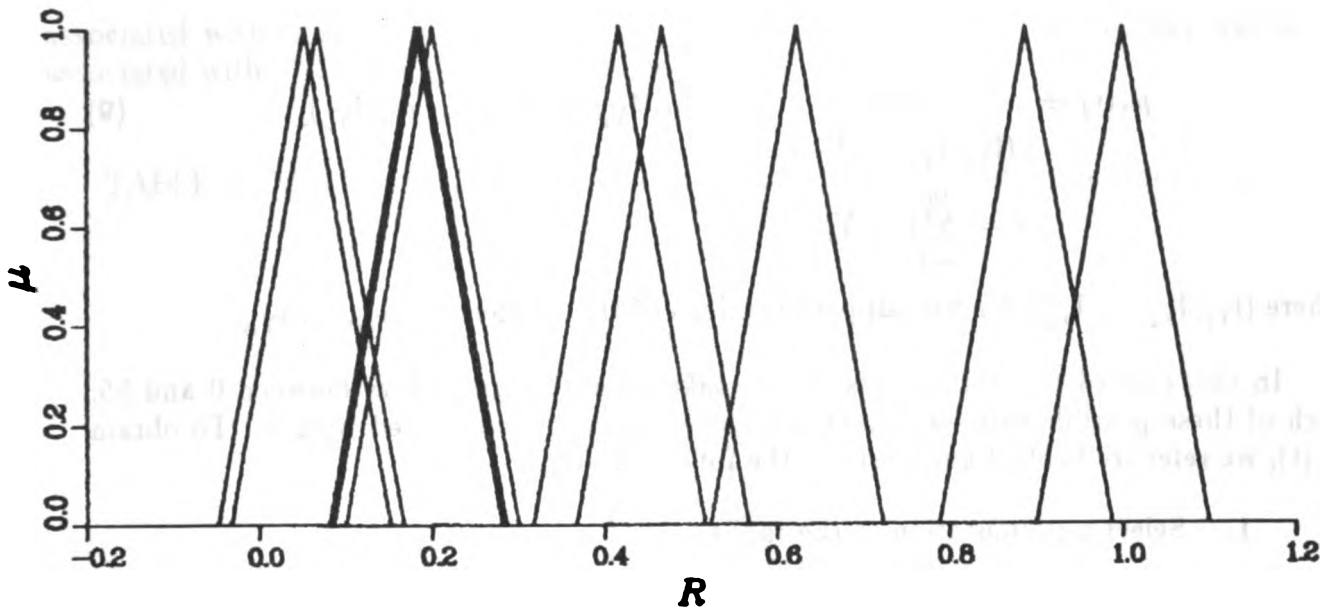


Figure 3. Membership Functions of the Original Data Set.

TABLE 2. Membership Functions Associated with the Fuzzy Ranks for the Original Data Set.

Ranked Data Points	1	2	3	4	5	6	7	8	9	10
1	1.00	0.93	0.36	0.33	0.26	0.00	0.00	0.00	0.00	0.00
2	0.93	1.00	0.43	0.41	0.34	0.00	0.00	0.00	0.00	0.00
3	0.36	0.43	1.00	0.98	0.91	0.00	0.00	0.00	0.00	0.00
4	0.33	0.41	0.98	1.00	0.93	0.00	0.00	0.00	0.00	0.00
5	0.26	0.34	0.91	0.93	1.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.00	0.75	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.75	1.00	0.22	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.22	1.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.44
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	1.00

C. Incorporating Fuzzy Ranks into the Wilcoxon Signed-Ranks Test

Once the membership functions of the ranks are established, it is necessary to calculate the value of T, the sum of the positive ranks. T will be the sum of fuzzy integers and, as such, will be a fuzzy integer itself. To determine its membership function, we refer to the Extension Principle and determine that

$$\mu_T(t) = \max_{\substack{(I_{Y_1}, I_{Y_2}, \ldots, I_{Y_{10}}) \\ t = \sum_{i=1}^{10} I_{Y_i},\ Y_i > 0}} \min\left[\mu_1(I_{Y_1}), \mu_2(I_{Y_2}), \ldots, \mu_{10}(I_{Y_{10}})\right], \tag{9}$$

where $(I_{Y_1}, I_{Y_2}, \ldots, I_{Y_{10}})$ denotes all permutations of the integers $I_{Y_1}, I_{Y_2}, \ldots, I_{Y_{10}}$.

In this case of ten data points, T can take on all integer values between 0 and 55; each of these possible sums will have a membership value associated with it. To obtain $\mu_T(t)$, we refer to Table 2 and perform the following steps:

1. Select a permutation of the ranks.

2. From Table 2 determine the minimum membership value of the ranks in their respective positions for this particular permutation.

3. If that minimum membership value is greater than zero, determine the sum of the positions of the positive ranks for this particular permutation.

4. If the membership value is greater than the membership value currently associated with that sum, replace with the new membership value.

We continue with this sequence of operations until all the permutations have been exhausted, at which time we have associated with every possible value of T a membership value which is the maximum over all permutations of the minimums for each individual permutation.

Using our set of ordered pairs, S, we can provide an example of the sequence above:

1. Suppose our selected permutation is 5 1 3 2 4 7 6 8 10 9.

2. Referring to Table 2, we can see that the membership value of rank 5 in the first position is 0.26, the membership value of rank 1 in the second position is 0.93, the membership value of rank 3 in the third position is 1.00, and so forth. If any one of these is equal to zero, then the minimum is equal to zero, and we skip steps three and four. For this particular permutation, the minimum membership value is 0.26.

3. The sum of the positions of the positive ranks for this particular permutation is equal to ten (first plus fourth plus fifth).

4. If 0.26 is greater than the current membership value associated with a sum of ten, then replace it.

When we have examined all possible permutations, the membership function associated with the sum of positive ranks, T, is shown in Table 3. Membership values associated with $T \leq 5$ and $T \geq 13$ are all equal to zero.

TABLE 3. Membership Function Associated with the Sum of Positive Ranks for the Original Data Set (Non-zero Values).

T	$\mu(T)$
6	0.330
7	0.355
8	0.905
9	0.925
10	0.975
11	1.000
12	0.430

Of course, examining all permutations can be very time consuming. This particular case required 201 seconds of central processor unit (CPU) time on a CDC 7600 computer. However, because of the large number of membership values that were equal to zero (see Table 2), many of the permutations could be ignored, since resulting minimums would be equal to zero and would not affect subsequent maximums. By taking advantage of this information to modify the permutation subroutine, I was able to reduce the CPU requirement to 43 seconds. Even with this kind of reduction, it is difficult to exceed a sample size of twelve without incorporating other shortcuts. One very effective method is to segment the data set, particularly if there is a datum point which is crisp rather than fuzzy; that is, its membership value at all but one position is equal to zero. Using this characteristic, I was able to handle a sample size of 32 in a later application of this work.

III. INTERPRETING RESULTS

When the data were considered non-fuzzy, we saw that there was insufficient evidence for rejecting the hypothesis of equal medians. We could have provided a critical level as defined by Conover; in doing so, we would have concluded that the null hypothesis could have been rejected at a significance level of 10.54% (see Table 1 and recall that we are performing a two-sided test).

Treating the data as fuzzy numbers provides a fuzzy result for T with a membership function described in Table 3. This allows for several methods of interpretation. Observing that $\mu(T) = 1$ (its maximum) when T=11, we might state that there is insufficient evidence for rejecting the null hypothesis at the $\alpha = .10$ level. Thus, the classical (non-fuzzy) signed-ranks test emerges as a special case. Alternatively, knowing that T=10 was the threshold for rejection, we might state that the null hypothesis can be rejected at the $\alpha = .10$ level with a membership value of 0.975. Since we recognize the data as imprecise, perhaps the best alternative is to accept the imprecision inherent in the resulting test statistic and make the decision as to whether or not to reject the null hypothesis based on the entire membership function. In our example, the membership value exceeds 0.900 for T=8 through T=11. Therefore, none of these values should be disregarded when analyzing the data; they all became viable candidates for T when the model took into account the proximity of the data points. The nature of any particular application should assist in making the final decision less subjective. Our example represents a situation in which the null hypothesis of equal medians would not have been rejected based on the original data set but may be rejected when the data, imprecise in nature, are treated as fuzzy numbers.

V. SUMMARY

Hypothesis testing is an important and useful tool for data analysis. When the data are vague or imprecise, an additional source of error is introduced and may result in an incorrect decision whether or not to reject the null hypothesis. Treating the data as fuzzy numbers allows us to model the uncertainty; and manipulating the data using fuzzy arithmetic allows us to carry the uncertainty through to the final results, at which point a more informed decision can be made.

Rank Tests are a class of hypothesis tests which are especially susceptible to the problems of imprecise data since the data, once ranked, have no further influence regardless of how closely they might be grouped. The Wilcoxon signed-ranks test is one example; and it was this particular hypothesis test that was applied to some data assumed to be vague in nature. The data were represented as fuzzy numbers, and the test statistic was calculated using fuzzy arithmetic. This provided a final result which was itself a fuzzy number, and several methods of interpreting this result were discussed.

I found computer time to be a major problem with incorporating fuzzy data into rank tests. In this case I needed to examine all possible permutations of rankings for all the data. For 10 data points the problem is not too bad; but if the data set is expanded to 30 points, then even with newer, faster computers some special techniques must be applied. In most cases one should be able to segment the data set, so that groups of ten or less can be examined and the results combined. This should make fuzzy hypothesis testing feasible as well as reasonable -- an even more important and more useful tool for the statistician!

A Central Limit Theorem for Fuzzy Random Variables

Steven B. BOSWELL
Department of Biostatistics
Harvard University School of Public Health
and
Department of Radiology
Harvard Medical School

Malcolm S. TAYLOR
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Abstract

Fuzzy random variables have been proposed to treat situations in which both random behavior and fuzzy perception must be considered. A definition of independence is given for fuzzy random variables, as well as a notion of fuzzy Gaussian random variables. It is shown that a sum or mean of independent fuzzy random variables converges in the limit to a fuzzy Gaussian random variable, thus providing a fuzzy analogue of the central limit theorem of classical probability theory.

This paper will appear in the journal Fuzzy Sets and Systems.

An Application of a Fuzzy Random Variable to Vulnerability Modeling

Steven B. BOSWELL
Department of Biostatistics
Harvard University School of Public Health
and
Department of Radiology
Harvard Medical School

Malcolm S. TAYLOR
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Abstract

Fuzzy sets are useful as a modeling tool in situations which have an ingredient of uncertainty or vagueness, as distinct from randomness. One class of problems fitting this description arises in vulnerability analysis. An application of a fuzzy random variable to enhance a vulnerability model currently in use is discussed.

1. Introduction

Kwakernaak, in a seminal paper [6] introduced the notion of a fuzzy random variable as a random variable whose values are not real but fuzzy numbers. Expectation and probabilities relating to a fuzzy random variable are developed as images of a fuzzy set, representing the fuzzy random variable, under appropriate mappings. A natural development of the theory is to examine fuzzy analogues of classical probability laws.

Toward this end, Kruse [5] and Miyakoshi and Shimbo [8] report on a strong law of large numbers. Stein and Talati [13], following Nahmias [9], develop a theory specifically for convex fuzzy random variables. Boswell and Taylor [2] provide a fuzzy analogue of the central limit theorem for fuzzy random variables admitting a moment generating function extension. Puri and Ralescu [11] outline a theory similar to Kwakernaak's and derive a dominated convergence theorem.

Application of these potentially powerful concepts has yet to evolve. Schlegel, Shear and Taylor [12] cite areas of vagueness in vulnerability modeling and suggest fuzzy sets as a potential modeling tool. The implementation of one such suggestion using a fuzzy random variable is the topic of this paper.

2. Fuzzy Random Variables

Kwakernaak [6] defines a fuzzy set $\mathbf{f}$ as a triple $\mathbf{f} = (A, t, p)$ consisting of a basic set A, a logical proposition p which can be applied to every member of the basic set, and a function t which assigns to every member $x \in A$ a truth value $t(p(x))$ indicating the appropriateness of the proposition p as applied to x. Most authors suppress the proposition p notation, since it is implicit in the organizing principle of the fuzzy set, and compose the proposition and truth value into a membership function $\mu: A \to [0, 1]$ which acts on the basic set, $\mu(x) = t(p(x))$. Thus $\mathbf{f}$ would be written $\mathbf{f} = (A, \mu)$; we shall adopt this convention.

An α-level set corresponding to a given fuzzy set $\mathbf{f} = (A, \mu)$ is an ordinary non-fuzzy set, denoted

$$L_\alpha(\mathbf{f}) = \{x \in A \mid \mu(x) \geq \alpha\} . \tag{2.1}$$

A fuzzy number is a fuzzy set having the real line $\mathbf{R}$ as its basic set. The fuzzy number $\mathbf{f}$, or its membership function μ, is said to be unimodal if for every $\alpha \in (0, 1]$, $L_\alpha(\mathbf{f})$ is convex. We shall be concerned with a collection C of fuzzy numbers defined as follows: a fuzzy number $\mathbf{f} = (\mathbf{R}, \mu)$ belongs to C if its membership function μ satisfies

(i) μ is upper semicontinuous,

(ii) for some $x \in \mathbf{R}$, $\mu(x) = 1$,

and

(iii) for all $\alpha > 0$, $L_\alpha(\mathbf{f})$ is bounded.

The set of membership functions satisfying (i) - (iii) will be called S.

Fuzzy random variables are constructed as a means of modeling phenomena which could properly be described by ordinary real random variables defined on a probability space (Ω, F, P), but which are partially obscured by fuzzy perception of the real line. In particular, if U_0 is the underlying random variable and ω is the outcome of a random experiment, the exact value $U_0(\omega)$ is unobservable; instead, it is assumed that a fuzzy number $\mathbf{f} = (\mathbf{R}, X_\omega)$ is known which characterizes the result $U_0(\omega)$. The mapping $X{:}\Omega \to S$ given by $X(\omega) = X_\omega$ supplies a membership function for each random outcome, and is called a fuzzy perception function. To the observer who must perceive random outcomes via X, the identity of U_0 is lost, and in fact there may be many reconstructions of U_0 which are amenable to fuzzy perception. By the standard operations of fuzzy logic [4], X generates a valuation function which applies to random variables as entities. Namely, if U is an F-measurable random variable, then

$$\mu_x(U) = \inf_{\omega \in \Omega} X_\omega(U(\omega)) \tag{2.2}$$

is the valuation of its suitability as a reconstruction of U_0.

Kwakernaak's development of the basic set of random variables to serve as candidates for reconstruction is rather involved. In [2] we make some simplifying assumptions which are sufficient for our application. Briefly, we admit as a basic set U_F, the set of all F-measurable random variables on Ω, and enforce partial retention of the structure of (Ω, F, P) through the requirement that for all $\alpha \in (0, 1]$ the functions

$$U_\alpha^*(\omega) = \inf \{x \in \mathbf{R} \mid X_\omega(x) \geq \alpha\}$$

and (2.3)

$$U_\alpha^{**}(\omega) = \sup \{x \in \mathbf{R} \mid X_\omega(x) \geq \alpha\}$$

are measurable with respect to (Ω, F). The sigma algebra generated by the random variables U_α^*, $\alpha \in (0, 1]$ and U_α^{**}, $\alpha \in (0, 1]$ is denoted by $\sigma(X)$, and χ denotes the set of all $\sigma(X)$-measurable random variables on Ω.

Letting U_F be the collection of all F-measurable random variables on Ω, the ***fuzzy random variable*** induced by X is defined as

$$\mathbf{X} = (U_F, \mu_x).$$

Some properties of a fuzzy random variable may be obtained directly by the extension principle [14]. For example, the expectation of a fuzzy random variable $\mathbf{X}$ is a fuzzy number

$$\mathbf{EX} = (\mathbf{R}, \mu_{\mathbf{EX}})$$

with membership function

$$\mu_{EX}(x) = \sup_{U \in U_F:\, EU = x} \inf_{\omega \in \Omega} X_\omega (U(\omega))$$

$$= \sup_{U \in U_F:\, EU = x} \mu_x (U), \quad x \in \mathbf{R}. \tag{2.4}$$

In (2.4), E denotes the usual mathematical expectation.

A fuzzy random variable **X** is called unimodal if for each $\omega \in \Omega$, the membership function X_ω is unimodal. Kwakernaak shows that if **X** is unimodal the basic set U_F may be restricted to χ, the set of all $\sigma(X)$-measurable random variables on Ω.

Theorem (2.1). If **X** is unimodal, then

$$\mu_{EX}(x) = \sup_{U \in \chi:\, EU = x} \inf_{\omega \in \Omega} X_\omega(U(\omega)), \quad x \in \mathbf{R}. \tag{2.5}$$

3. Vulnerability Modeling

This is an account of an application of a fuzzy random variable to an important problem in vulnerability modeling. Succinctly, vulnerability modeling is an attempt to characterize the interaction between a target (armored vehicle, aircraft, bunker, ...) and a munition (kinetic energy penetrator, shaped charge, explosive device, ...) and to assess quantitatively the resulting damage sustained (inflicted) within the target-munition combination.

Experimental testing required to provide data pertinent to vulnerability modeling is destructive, and the data base upon which these models are built may be modest, or in the case of conceptual systems, nonexistent. Furthermore, while certain damage-related measurements (velocity of impact, depth of penetration, component function) may be determined in an unambiguous manner, many others (structural deformation, fracturing, component degradation-of-function) may not. The composition of quantitative measurements and qualitative information into a cohesive assessment of damage remains at the core of difficulty in vulnerability modeling. We will consider a particular vulnerability model [10] currently in use and demonstrate the applicability of fuzzy sets to its enhancement.

Figure 1 represents a data summary of an encounter between an armored vehicle and a kinetic energy penetrator. A rectangular grid of 10x10 cm cells has been superimposed on the profile of the armored vehicle, and within each cell of the grid, the probability that the vehicle will be rendered inoperable (killed) should it sustain an impact within that cell, is listed. Within the bold rectangle, for example, the probability-of-kill, given a hit in cell i, $P_{k|h_i}$, is estimated to be .19.

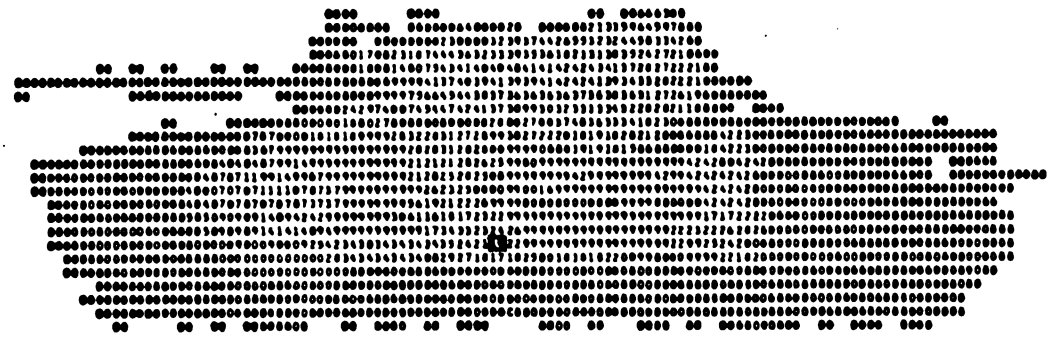


Fig. 1. Data summary of an encounter between an armored vehicle and a kinetic energy penetrator.

Consider the conditional probability-of-kill, $P_{k|h_i}$, for an arbitrary cell i. This number is produced by a computer simulation [10] involving a blend of geometry, probability, heuristics, and archival information about similar systems. The cell probabilities are combined by an averaging over all cells to produce a single value, P_k, representing probability-of-kill for this particular configuration of vehicle vs. kinetic energy penetrator. If an aim-point on the vehicle is designated, then a weighted average of the cell $P_{k|h_i}$s is calculated, the weights provided by a bivariate probability density located at the aim-point.

The overall probability-of-kill estimate P_k is subject to criticism as a value which conveys little useful information, and none about the variability inherent in the estimate. The magnitude of the computer simulation prohibits repeated runs to provide an empirical or bootstrap estimate of the distribution of overall probability-of-kill.

While randomness is clearly present in the experimental data collected, an even greater source of uncertainty lies in the procedure producing the cell $P_{k|h_i}$s, and suggests the incorporation of fuzziness as a modeling artifice. We consider the data in Figure 1 as representing the sample space Ω of an experiment which has been discretized by the overlaid grid. The experiment consists of firing at the tank, and a random variable U provides for each impact location a corresponding probability-of-kill $P_{k|h_i}$. We replace the cell $P_{k|h_i}$ value with a fuzzy number whose membership function is illustrated in Figure 2. The width of the interval on which $\mu(x)$ takes the value one is chosen to be $P_{k|h_i}(1 - P_{k|h_i})$, the variance of the Bernoulli distribution modeling the individual cell probabilities. We have thus defined a fuzzy random variable **X**, whose expectation we seek.

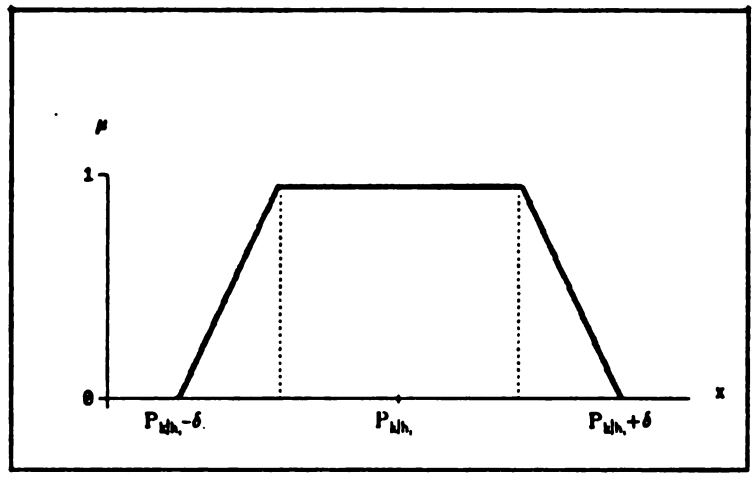


Fig. 2. Membership function for fuzzy $P_{klh_{ij}}$ where $\delta = P_{klh_{ij}}(1 - P_{klh_{ij}})$.

Since the membership functions are unimodal, the expectation of **X** is a fuzzy number **EX** with membership function

$$\mu_{EX}(x) = \sup_{U \in \chi:\, EU = x} \inf_{\omega \in \Omega} X_\omega(U(\omega)), \quad x \in \mathbf{R}. \tag{3.1}$$

This expression can be evaluated using the α-level sets (2.1). Given the family of level sets $L_\alpha(\cdot)$, the membership function μ_{EX} may be recovered with the aid of the formula

$$\mu(x) = \sup \{\alpha \in [0, 1] \mid x \in L_\alpha(\cdot)\}, \quad x \in \mathbf{R}. \tag{3.2}$$

For the simple membership function of Figure 2, this computation can be simplified using procedures detailed by Dubois and Prade [3] or Bonnisone [1]. Applying these procedures to the data in Figure 1 we obtain for **EX** the membership function shown in Figure 3.

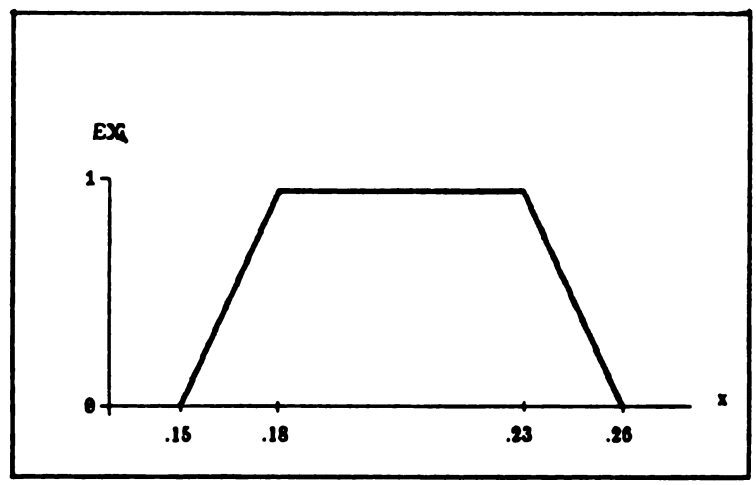


Fig. 3. Membership function for expectation EX of fuzzy random variable X.

4. Conclusion

The form chosen for the cell $P_{k|h_i}$ membership functions (symmetric, normal, convex) leads to a membership function for **EX** that is similar to the constituent $\mu(x)$ in Figure 2. The interpretation of the resultant μ_{EX} is that estimates of overall P_k in the interval [.18, .23] are, within our framework of uncertainty, wholly plausible. The impulse to take the level set $L_{.90}(\cdot)$, say, and consider it a 90% confidence interval for P_k must be resisted; there are no probability statements carried by the α-level sets. We have, however, modeled the uncertainty of the cell $P_{k|h_i}$ estimates in the overall probability-of-kill estimate P_k in a direct way, and distinguished between randomness and uncertainty in the vulnerability model. We also have the framework in place to consider $P_{k|h_i}$ membership functions far more intricate than the one shown in Figure 2. This is a significant methodological improvement.

References

[1] P. Bonissone, A fuzzy sets based linguistic approach: Theory and applications, Approximate Reasoning in Decision Analysis, M. Gupta and E. Sanchez (eds.), North-Holland (1982), 329-339.

[2] S. Boswell and M. Taylor, A central limit theorem for fuzzy random variables, *Fuzzy Sets and Systems* (to appear).

[3] D. Dubois and H. Prade, Operations on fuzzy numbers, *International Journal of Systems Science* 9(6) (1978), 613-626.

[4] B. Gaines, Stochastic and fuzzy logics, *Electron. Letters* 11(9) (1975), 188-189.

[5] R. Kruse, The strong law of large numbers for fuzzy random variables, *Information Sciences* 28 (1982), 233-241.

[6] H. Kwakernaak, Fuzzy random variables - I. Definitions and theorems, *Information Sciences* 15 (1978), 1-29.

[7] H. Kwakernaak, Fuzzy random variables - II. Algorithms and examples for the discrete case, *Information Sciences* 17 (1979), 253-278.

[8] M. Miyakoshi and M. Shimbo, A strong law of large numbers for fuzzy random variables, *Fuzzy Sets and Systems* 12 (1984), 133-142.

[9] S. Nahmias, Fuzzy variables, *Fuzzy Sets and Systems* 1 (1978), 97-110.

[10] C. Nail, Vulnerability analysis for surface targets (VAST)-Revision I, Computer Sciences Corp Report CSC-TR-82-5740 (1982).

[11] M. Puri and D. Ralescu, Fuzzy random variables, *Journal of Mathematical Analysis and Applications* 114 (1986), 409-422.

[12] P. Schlegel, R. Shear and M. Taylor, A fuzzy set approach to vulnerability analysis, Ballistic Research Laboratory Technical Report BRL-TR-2697 (1985).

[13] W. Stein and K. Talati, Convex fuzzy random variables, *Fuzzy Sets and Systems* 6 (1981), 271-283.

[14] L. Zadeh, The concept of a linguistic variable and its application to approximate reasoning II, *Information Sciences* 8 (1975), 301-357.

PROBLEMS ENCOUNTERED IN FITTING A LARGE NUMBER OF SHORT TIME SERIES

Franklin E. Womack and Elizabeth N. Abbe
US Army Concepts Analysis Agency
Bethesda, Maryland 20814-2797

ABSTRACT. The US Army Concepts Analysis Agency has become increasingly involved in various forecasting projects. Most of the projects have some common characteristics. Typically each series has less than 100 observations and often less than 50 observations. Box-Jenkins[2] suggests that more than 100 observations are preferable and that one uses experience and past information to yield preliminary models where fewer than 50 observations are available. Usually the Agency analysts are not extremely familiar with the systems and processes which generates these series. Each project commonly has a set of series which consists of many elements. The number of elements in a set can range from 600 to 1,000 individual time series. The identified model form of each individual time series in the set varies greatly. Often only white noise is present. On the other hand some of the series will exhibit seasonal behavior. Many of the series are nonstationary and have potential interventions. Many of the series take on only a discrete set of values such as the set of positive integers from zero to ten.

1. INTRODUCTION. One project requires a forecast of the quantities of various commodities shipped over various routes. The forecast of potential loads would be helpful in scheduling limited transportation facilities. This project involved about 400 individual series each one describing the history of a particular commodity; for example, coal over a particular route, say port of New York to Europe. Another project involved forecasting the number of separations from the US Army of enlisted grades E-5 and E-6 for about 300 different military occupational skills. These series are often influenced by policy changes. Table 1 illustrates the types of forecasting projects and their requirements for two recent projects.

Table 2 compares the results of several different forecasting techniques giving a "best" forecast, as described further in this paper, for a selected sampling of time series from these two projects.

Table 1. Series Characteristics

Characteristic	Project	
	1	2
Length of series	84	38
Length of fit	72	34,35, 36,37
Forecast horizon	12	1
Total number of series to evaluate	400	579

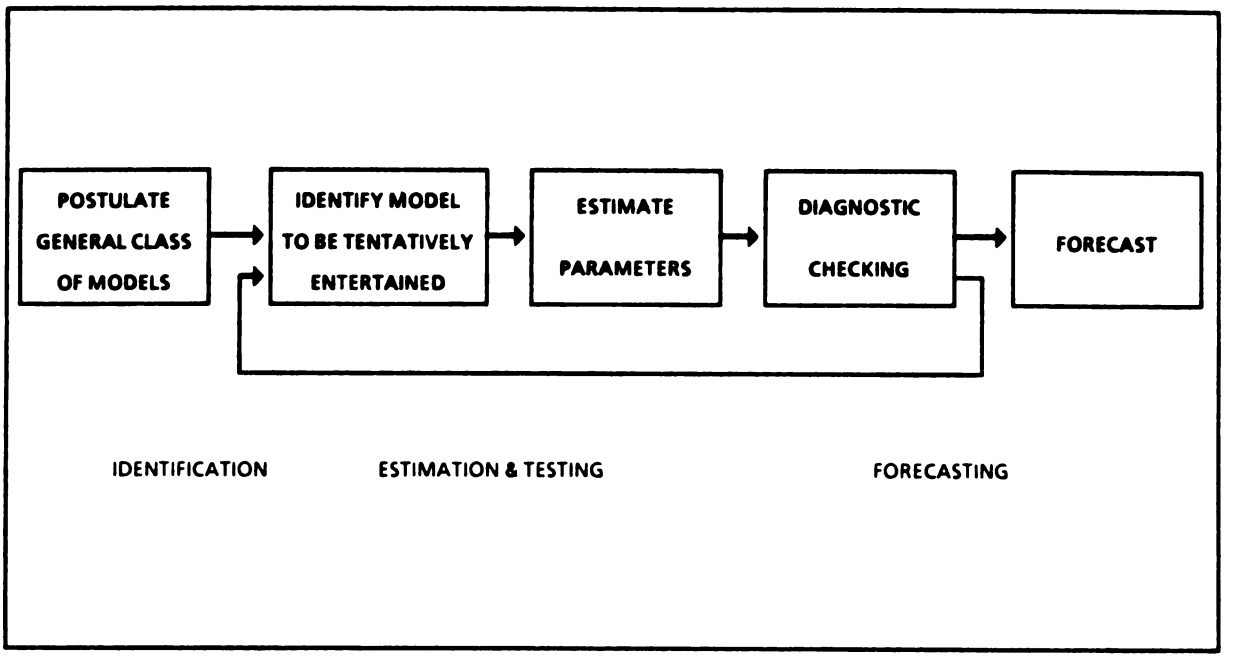
Table 2. Comparison of Forecasts

Forecast	Project	
	1	2
Box-Jenkins	34	14
Winter/Gardner	27	39
Ties	0	10
Unknown	5	0
Total	66	63

2. PROBLEMS. Many problems arise in the analysis of time series. However, the literature is limited on methods to handle short series. The Agency is often confronted with both a methodological and a procedural problem. The methodological problem is largely a result of the inherent instability of model form and values of estimated parameters for short time series. The procedural problem is usually imposed by study sponsors who require a process which will act in a production mode by incorporating new observations into the forecast as they become available.

Since no particular methodologies are suggested as superior for short series we have adopted an all inclusive policy. The foremost technique for evaluating long time series is the Box-Jenkins process (Table 3).

Table 3. Box-Jenkins Process



The Box-Jenkins process is a three state process consisting of two iterative stages; (1) identification, followed by (2) estimation and testing, and finally (3) a forecasting stage. The task of identifying 400 individual series by evaluating the sample autocorrelation and sample partial autocorrelation functions can be monumental. This is especially true when typically not only the original series must be examined, but several other series due to nonstationarity and the consideration simultaneously of seasonal as well as nonseasonal model forms. Automated identification is an essential consideration in these Agency projects. Therefore we have employed two pieces of software which render automatic identification. These software and their attributes are described in Table 4.[1][5]

Table 4. Automatic Identification Routines for Box-Jenkins Methods Employed

(1) AUTOBOX

AUTOMATIC IDENTIFICATION
INTERVENTION DETECTION (UP TO 5)
BOX-COX TRANSFORMATIONS (SQUARE ROOT, NATURAL LOG, RECIPROCAL SQUARE ROOT, AND RECIPROCAL)

(2) ARIMAID

AUTOMATIC IDENTIFICATION
USES AKAIKE'S INFORMATION CRITERION
MANUAL TRANSFORMATIONS POSSIBLE

In addition to the Box-Jenkins technique several techniques of modeling an exponentially weighted moving average have been employed. Two such techniques are the Holt/Winters[6] Model and the Gardner-Mckenzie[3][4] Model. The update forms of these models are shown in Table 5.

Table 5. Multiplicative Seasonal Models Update Sequences

WINTERS - HOLT

$$e_t = X_t - \hat{X}_{t-1}(1)$$

$$S_t = S_{t-1} + T_{t-1} + \alpha e_t / I_{t-p}$$

$$T_t = T_{t-1} + \alpha\gamma e_t / I_{t-p}$$

$$I_t = I_{t-p} + \delta(1-\alpha) e_t / S_t$$

$$\hat{X}_t(1) = (S_t + T_t) I_{t-p+1}$$

GARDNER NONLINEAR

$$e_t = X_t - \hat{X}_{t-1}(1)$$

$$S_t = S_{t-1} + \phi T_{t-1} + \alpha(2-\alpha) e_t / I_{t-p}$$

$$T_t = \phi T_{t-1} + \alpha(\alpha - \phi + 1) e_t / I_{t-p}$$

$$I_t = I_{t-p} + \delta[1-\alpha(2-\alpha)] e_t / S_t$$

$$\hat{X}_t(1) = (S_t + \phi T_t) I_{t-p+1}$$

WHERE

- X_t = OBSERVED VALUE TIME t
- e_t = FORECAST ERROR AT TIME t
- S_t = LEVEL (MEAN) AT TIME t
- I_t = SEASONAL INDEX AT TIME t
- γ = TREND SMOOTHING PARAMETER
- ϕ = TREND MODIFICATION PARAMETER
- P = NUMBER OF PERIODS IN A CYCLE
- $\hat{X}_t(1)$ = ONE-STEP AHEAD FORECAST AT TIME t
- T_t = TREND AT TIME t
- α = LEVEL SMOOTHING PARAMETER
- δ = SEASONAL INDEX SMOOTHING PARAMETER

The local implementation of the Gardner Mckenzie technique is described in Table 6.

Table 6. Gardner Model Procedure

1. FIT LINEAR REGRESSION TO RAW OR TRANSFORMED DATA TO ESTABLISH BEGINNING SLOPE AND INTERCEPT, So AND To.
2. FOR NONSEASONAL MODEL, ESTIMATE α, γ, AND ϕ BY A GRID-SEARCH METHOD TO MINIMIZE MEAN SQUARE ERROR.
3. FOR SEASONAL MODEL, ESTIMATE δ BY HOLDING α, γ, AND ϕ FIXED AND DOING A GRID SEARCH FOR δ TO MINIMIZE MEAN SQUARE ERROR. CHOOSE INITIAL SEASONAL INDICES

   I_{1-p} TO I_0 SUCH THAT $I_{j-p} = P \cdot \bar{X}_{.j} / \sum_{j=1}^{p} \bar{X}_{.j}$

   WHERE X_{ij} = OBSERVATION i IN PERIOD j
   n_j = NUMBER OF OBSERVATIONS IN PERIOD j
   p = NUMBER OF PERIODS

The estimation process for the exponentially weighted techniques requires the development of parameter values which are traditionally chosen in an arbitrary ad hoc fashion.

In order to develop a reiterative process and also to some extent to ameliorate the stability problems of short series, a sequential technique was employed. This process is described in Table 7.

Table 7. Algorithm Followed in Project #2

(1) RESERVE SOME N* OF THE LAST OBSERVATIONS .
(2) ITERATE N* TIMES CALCULATION OF PARAMETERS FOR THE SEVERAL MODEL TYPES
(3) USE N-N* OBSERVATIONS ON FIRST ITERATION
(4) CALCULATE ONE-STEP AHEAD FORECAST AT FIRST ITERATION
(5) ON SECOND AND SUCCESSIVE ITERATIONS ADD AN ADDITIONAL OBSERVATION UNTIL N-1 OBSERVATIONS ARE INCLUDED, CALCULATE PARAMETERS FOR THE SEVERAL MODEL TYPES
(6) CALCULATE ONE-STEP AHEAD FORECAST AT EACH ITERATION
(7) CALCULATE ROOT MEAN SQUARE ERROR OVER ITERATED ONE-STEP AHEAD FORECASTS FOR EACH MODEL TYPE
(8) FIT ALL N OBSERVATIONS BY METHOD YIELDING MINIMUM ROOT MEAN SQUARE ERROR ON THE N* RESERVED OBSERVATIONS

At each new time point all techniques are simultaneously applied to the time series. For a certain period of the recent past, in our case the last four points are evaluated to determine a "best" technique. In one project which involved 38 quarterly observations, the last 4 observations formed this recent past and are referred to as the reserve set. Four successive steps involving 34, 35, 36, and 37 observations in each series were modeled by each of the several techniques. At each step, for each of the techniques, the one-step ahead forecast error was calculated by subtracting the technique's forecast from the true observation. A very rough measure of selecting the best technique to forecast the 39th point in this case could be made by selecting that technique which yielded the minimum mean square error on the reserve set (i.e., times 35, 36, 37, and 38).

3. Examples. Table 8 gives a frequency chart for 63 typical series selected from the project; each series in the project involved 38 quarterly observations described above. Each row specifies the model form type identified by the automatic identification software for the Box-Jenkins technique described above. Each column identifies one of the techniques employed in the local process described above. Each element of the table gives the number of series among the 63 which were identified as a particular ARIMA form and were "best" modeled by a particular technique. Figure 1 illustrates a model which was identified as autoregressive nonseasonal, and this form was selected as "best" on the reserve set (i.e., minimum mean square error over time points 35, 36, 37, and 38). Figure 2 illustrates a series identified as a nonstationary seasonal moving average, and this form was selected as "best" on the reserve set. Figure 3 illustrates a series identified as nonseasonal autoregressive, but a Winters/Holt Model was selected as the "best" on the reserve set. Figure 4 illustrates a series identified as white noise, but a model of the Gardner Multiplicative Nonseasonal Nonlinear Trend type was selected as the "best" on the reserve set. Figure 5 illustrates a series identified as white noise, but a model of the Winters/Holt Multiplicative Seasonal was selected as the "best" on the reserve set. From this sampling of typical time series it is evident that the all inclusive process picked model forms from a variety of techniques. It is amazing that even series identified as white noise by examination of their sample autocorrelation and partial autocorrelation can sometimes be better approximated over the reserve set by an exponentially weighted moving average model.

Figure 6 illustrates the preponderance of these 63 series which have only a small set of values. In this sampling 26 out of 63 (about 41 percent) of the series take on values from zero to ten.

Table 8. Model Frequency

IDENTIFIED AS:	MINIMUM MSE OVER RESERVE SET: BOX-JENKINS MODELS	GARDNER MODEL	TRANSFORMED GARDNER	WINTERS HOLT MODEL	AVERAGE OF LAST 4 QUARTERS	TIES	IDENTIFICATION TOTALS
WHITE NOISE	4	3	1	8	1	8	25
AR	4		1	7	2	3	17
MA	1				1		2
SEASONAL MA	2	1		1	1	3	8
MIXED NONSEASONAL AR AND SEASONAL MA PARAMETERS	3	1	1	4		2	11
MINIMUM MSE OVER RESERVE SET TOTALS	14	5	3	20	5	16	63

4. SUMMARY. It is difficult to obtain guidance from the literature on how to evaluate short time series. Even though the sample statistics from which time series forms are identified become very unstable for short series, there is an operational need for forecasting the short series. This paper has described attempts to cope with a real problem in the face of little guidance. The purpose of this paper is to solicit further guidance on the subject. Specifically we would like to ask several questions;

a. What forecasting method(s) are recommended for situations involving hundreds of "short" series with little time to accomplish?

b. Why does not Box-Jenkins make a better showing with respect to the exponentially weighted moving average models?

c. How do you recommend comparing forecasts from several techniques?

d. Are there any special techniques for treating series which take on only a small set of values such as the integers zero to ten?.

REFERENCES

(1) Autobox, The User's Guide, Auto Forecasting System, Inc., (1986)

(2) Box, G. E. P., and G. M. Jenkins (1976), Time Series Analysis; Forecasting and Control, 2nd ed., Holden-Day, San Francisco

(3) Gardner, E. S. and McKenzie, E., "Forecasting Trends in Time Series," Manage Sci., 31, 1237-1246

(4) Gardner, E.S., "Exponential Smoothing; The State of the Art," Journal of Forecasting, 4, 1-38

(5) Kang, C. A., Bedworth, D. D., Rollier, D. A., "Automatic Identification of Autoregressive Integrated Moving Average Time Series," IIE Transactions, 14, 156-165

(6) Winters, P. R., (1960) "Forecasting Sales by Exponentially Weighted Moving Averages," Manage Sci., 6, 324-342

IDENTIFIED AS AUTOREGRESSIVE NONSEASONAL
MODEL YIELDING LEAST MSE ON THE RESERVE SET
AT t-38, $(1-.784B+.320B^3)(Z_t-52.53)=a_t$

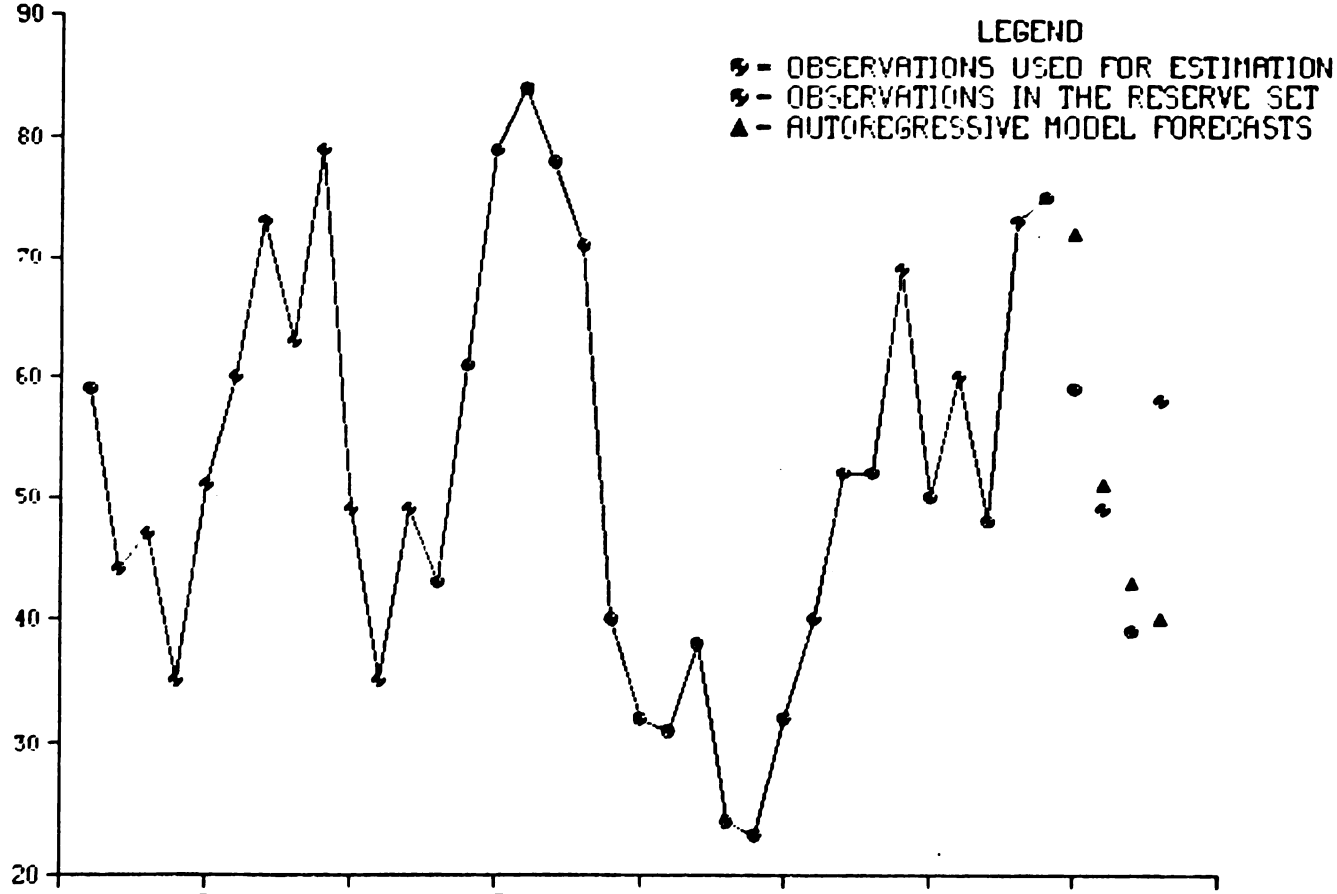


FIGURE 1

IDENTIFIED AS MOVING AVERAGE SEASONAL WITH INTERVENTION
MODEL YIELDING LEAST MSE ON THE RESERVE SET
AT t-38, $\nabla Z_t = (1-.405B^3)(1-.752B^8)a_t$

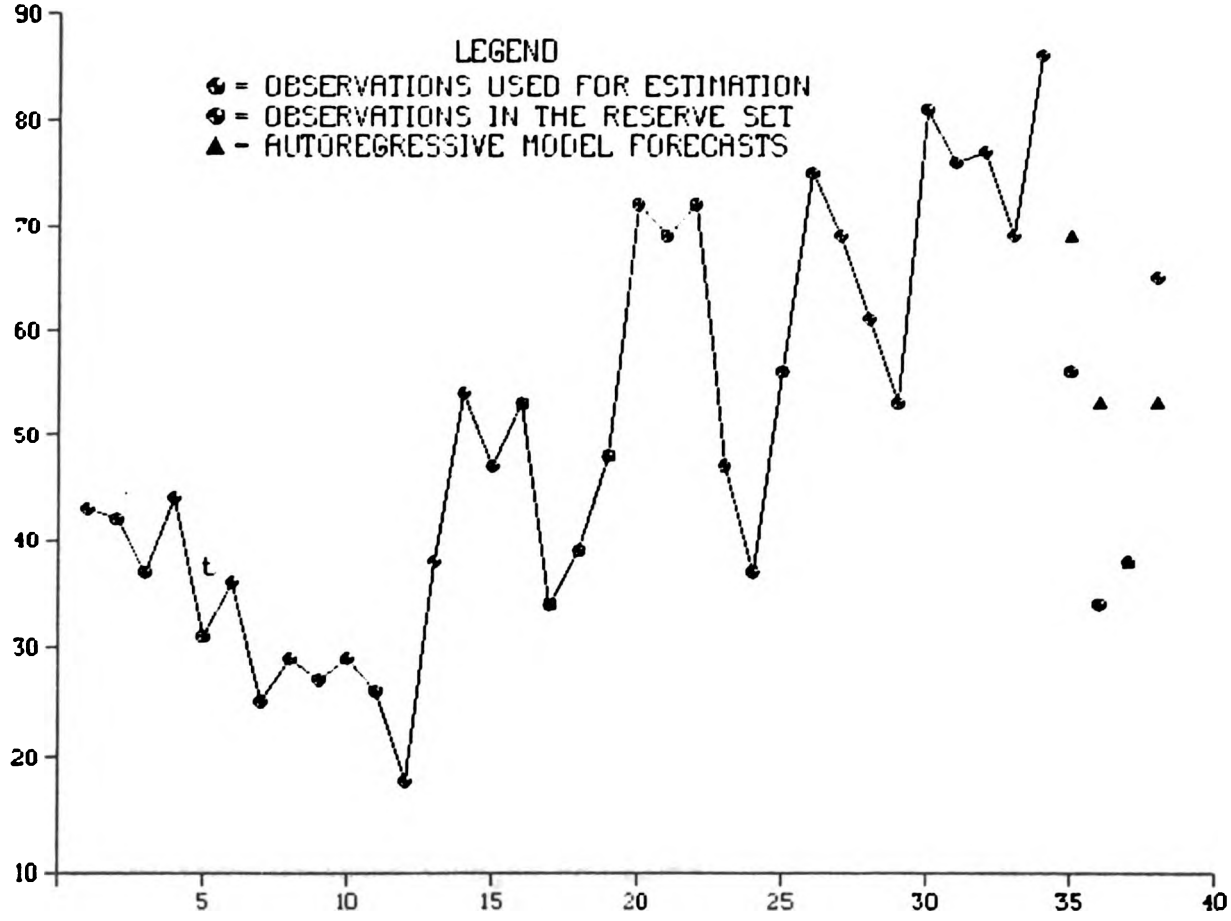


FIGURE 2

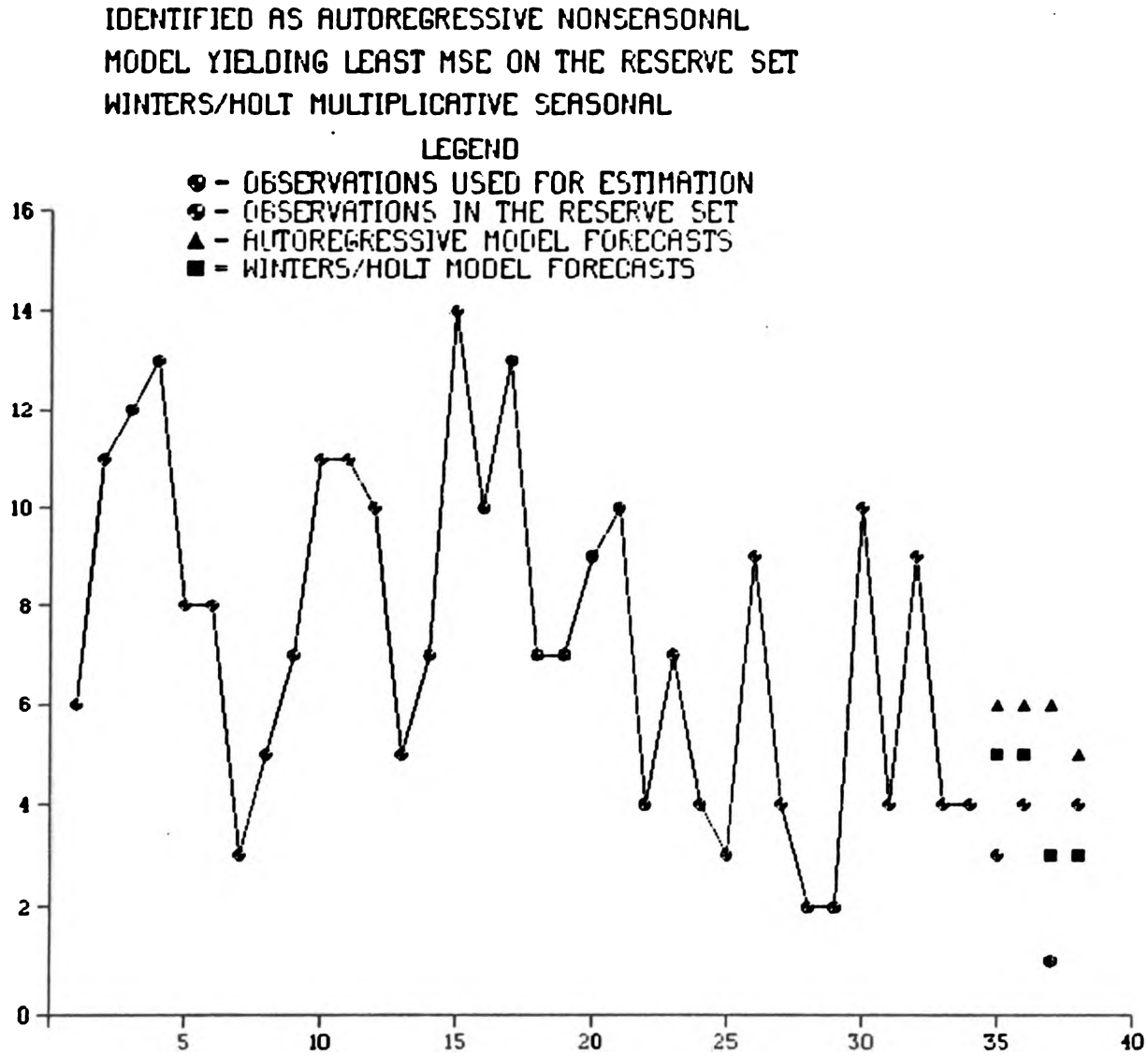


FIGURE 3

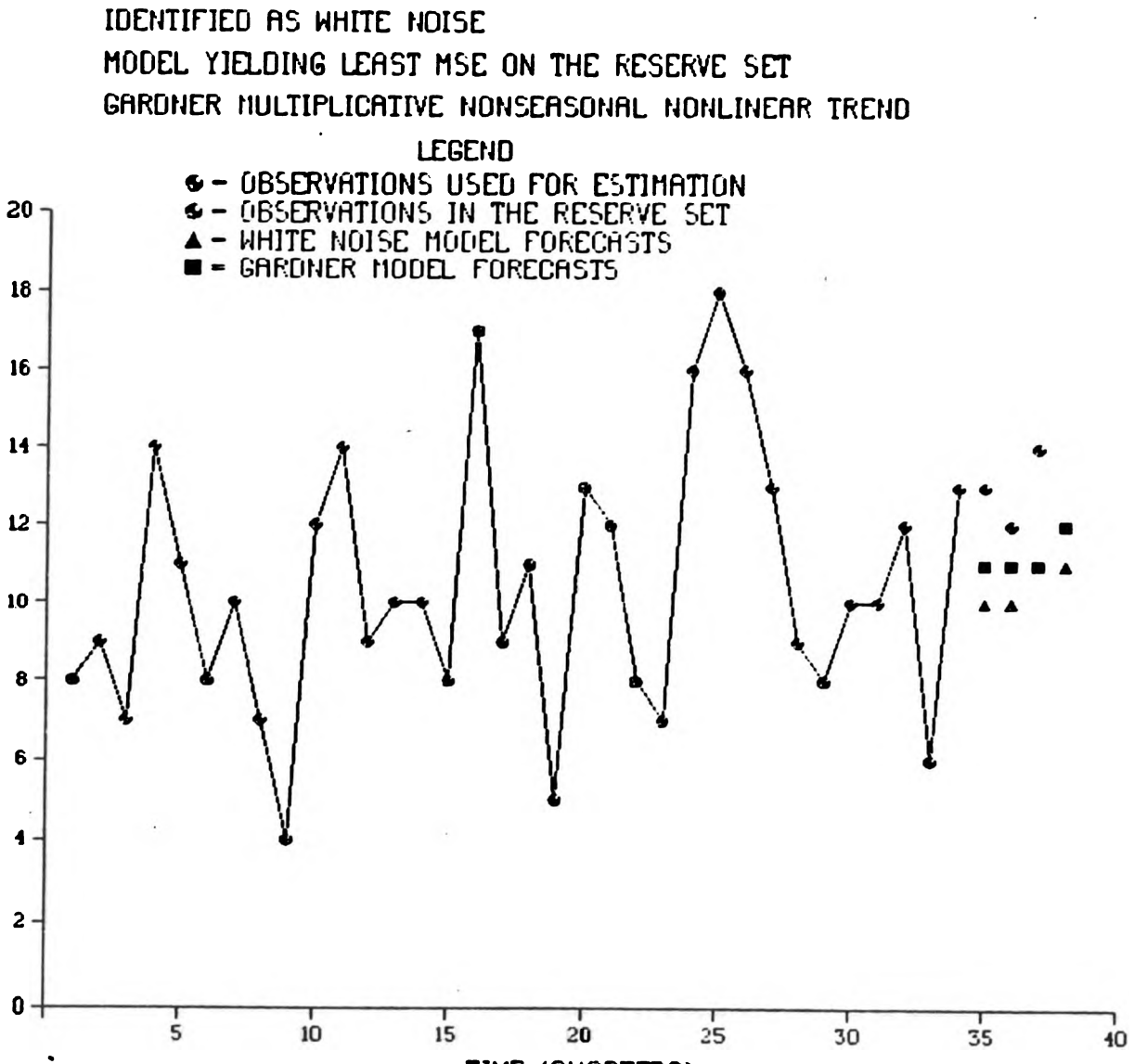


FIGURE 4

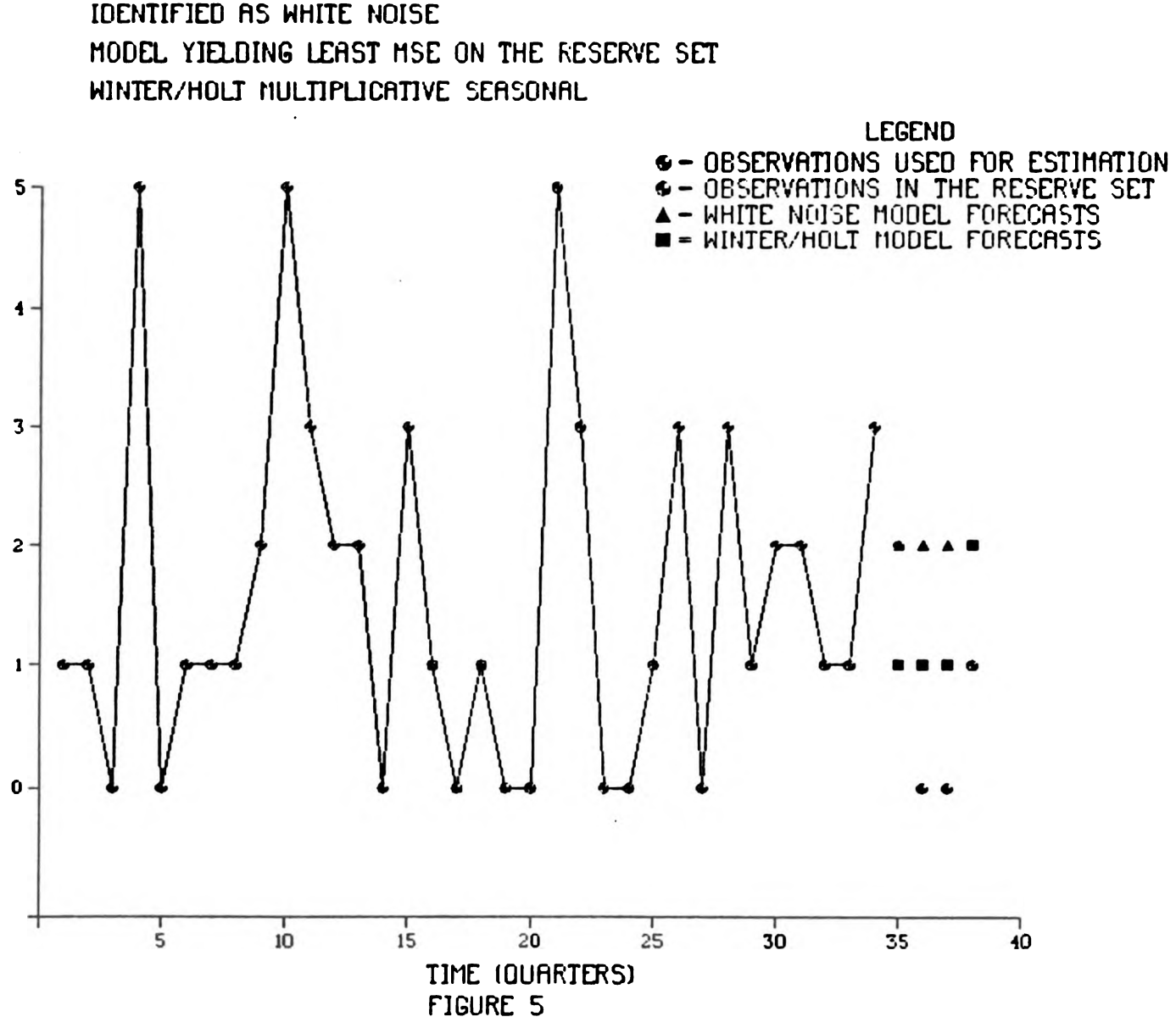


FIGURE 5

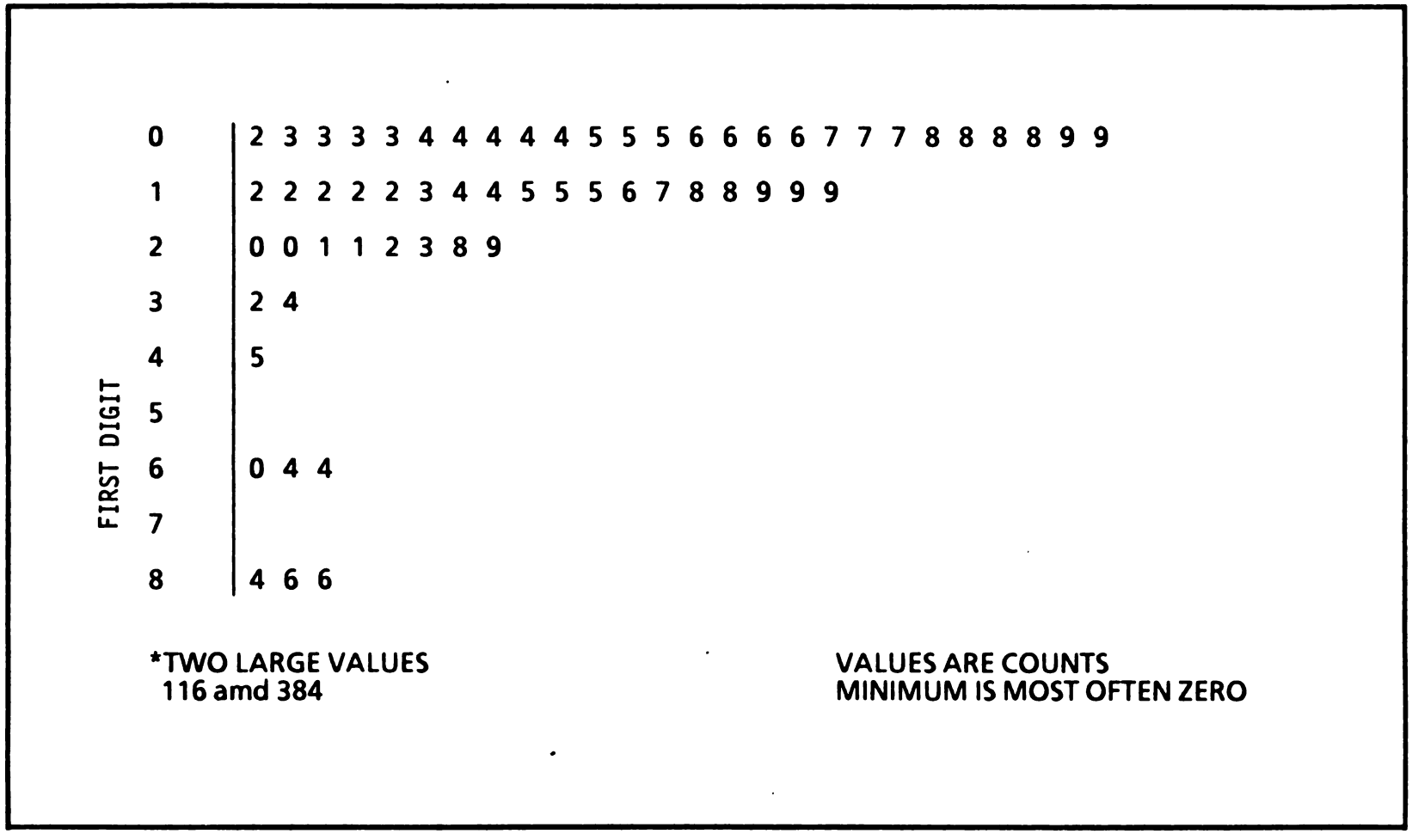


Figure 6. Stem-and-Leaf Display of Maximum Values Observed in 63 Time Series

STUDY ON THE FEASIBILITY OF GENERATING "PREDICTIVE ANALYSIS MODEL" by UTILIZING THE ARMY'S EXISTING DATA SOURCE

Li Pi Su
Logistics Engineers/Readiness Division
Materiel Readiness Support Activity (MRSA)
U.S. ARMY

ABSTRACT: This is a preliminary report on the feasibility of a predictive model for an Army data source.

The following seven tasks will be taken to determine the feasibility of a predictive model for a certain Army data source:

1. Clarification of the description of the data.

2. Classification.

3. Determination of the effects that influence the data.

4. Stratification of the data by location, mission usage, etc.

5. Examination of the quality of the data.

6. Adjustment of data. Effects, time period, outliers, etc.

7. Selection of a predicting model.

I. INTRODUCTION:

a. An adequate data source is important for obtaining reliable results from statistical analysis. However, if the data source is inadequate, the choice of analytical techniques selected to perform an analysis can improve the validity of the results and thus increase the accuracy of the prediction. The existing Army data collection methodology is not fully compatible with known predictive techniques. It is difficult to analyze the existing data statistically and to obtain useful and valid information such as: safety, reliability, readiness, cost, mean time between failures, mean time between replacement, or maintenance cost of certain systems. It is even more difficult to use the presently collected data for predicting any of the above information with a high confidence level.

b. This is a preliminary report. The report addresses some of the ways in which current Army data sources may be used in the application of a predictive technique and of some of the

techniques that could be utilized for conducting predictive analysis, given adequate data.

c. In section II, some of the applicable prediction techniques are presented. The basic requirements for a data source to be compatible with predicting techniques are discussed in section III. Then the problematic areas for both Army data sources and predicting techniques are stated in Section IV. In Section V, the approaches to be taken to determine the feasibility of generating a "Predictive Analysis Model" for an Army data source are discussed. Some of the possible applications for a predictive analysis technique are provided in Section VI.

II. Predicting Techniques and Fitting Criteria. The predicting (or forecasting) techniques discussed here are the known scientific ones. The structures of these scientific predictions can be determined by statistical and mathematical methods. Although each technique is somewhat unique in its predicting capability, in practice, it has been found that a very large group of data can be fitted with a "reasonable confidence-level" by one of the following basic models or one of their combinations: constant mean, linear trend, linear regression, autoregression, moving average, seasonal and periodic models, and exponential and non-linear models. A brief description of each of these techniques is given below. (See Gilchrist)

A. The constant mean model. This technique is of the form

$$x_t = \mu + \varepsilon_t , \qquad t = 1,2,3,...$$

where μ is the constant mean of all x_t's and ε_t is one of a sequence of independent random variables with zero expectation, i.e. $E(\varepsilon_t)=0$, and constant variance σ^2. Fitting criteria: zero mean error, reasonable small confidence interval. This method deals with a set of data fitted approximating to the global constant mean.

B. Linear trend model. This technique is of the form

$$x_t = \alpha + \beta t + \varepsilon_t , \qquad t = 1,2,3, ...$$

where α has the expectation of x_0 , β is a constant slope and ε_t is a sequence of independent random variables with zero expectation, i.e. $E(\varepsilon_t)=0$, and variance of ε_t , $Var(\varepsilon_t)= \sigma^2$.

Use the least square method for the fitting criteria.
The method deals with the data structure showing a linear trend with a random variation added.

NOTE: Both techniques A&B use only the past values of the variable being forecasted, the future values, and thus these approaches are limited to obtain the best forecasts because it

fails to use the influence information contained in other variables.

C. Regression model. The general form of this technique is

$$y_t = \sum_{i=o}^{k} \beta_i x_{ti} + \varepsilon_t , \qquad t = 1,2,3, \ldots$$

where β_i's, i=0,1,...,k are constants, and x_{ti}'s are variables related to y_t, and ε_t is a random variation. When the data structure shows a seasonal trend, some of x_{ti}'s, may be replaced by harmonic terms. The criteria for fitting is the mean square forecasting error.

NOTE: Random variables in techniques A,B and C, discussed above are simply added "errors" which were added tc a strictly deterministic function. Technique C will also make use of other information that is related to the one being forecasted so that a better forecasting product could be obtained.

D. Autoregressive model. This technique is of the form

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t , \qquad t = 1,2,3, \ldots$$

Where ϕ_i's, i=1,...p are constants estimated from given data, and ε_t's are identically distributed with zero mean and constant variance, i.e. $\mu(\varepsilon_t)=0$, and $Var(\varepsilon_t)= \sigma^2$. It is usually denoted by AR(p), where p is the order of this autoregressive model and is a positive integer. For fitting criteria see page 81 of Pankratz.

E. Moving average model. This technique is on the form

$$x_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} , \qquad t = 1,2,3, \ldots$$

Where ϕ_j's, j=1,...,q are constants estimated from given data, and ε_t's are identically distributed with zero mean and constant variance, i.e. $\mu(\varepsilon_t)=0$, and $Var(\varepsilon_t)= \sigma^2$. It is usually denoted by MA(q), where q is the order of this moving average model and usually is a finite positive integer. For fitting criteria see page 81 of Pankratz.

NOTE: Techniques D & E are stochastic models and have the random variables play the dominate part in determining the structure of the models. Box and Jenkins (1970) mixed autoregression and moving average models into one model that will improve the forecasting. This integrated model, usually called Box-Jenkins model, is denoted by ARMA(p,q). Moreover, some non-stationary models may become stationary by replacing the x_t's by differences of x_t's. The d-th difference is obtained by taking differences for the d-th time from x_t's. The integrated autoregressive - moving average model, denoted by ARIMA(p,d,q) is a result of combining d-th differencing process and ARMA(p,q). (See Box and Jenkins for mathematical forms.)

F. Seasonal and periodic models. These techniques do not have a unique form. They deal with the data to display a repeating pattern at a certain period. Some methods often used to deal with seasonal data are: Seasonal index methods, Fourier methods (see Gilchrist), and stochastic (or integrated autoregression - moving average) methods (see Box-Jenkins, Gilchrist, Jenkins and Pankratz.)

G. Exponential and non-linear models. In many situations, e.g. growth or decay, the data can only be fitted to provide a reliable forecast by exponential, logarithmic parabola, or their modified curves.

Remarks:

1. In the situation where the data structure is highly stable and the chosen model is the truth about the underlying structures of the data, the model is called a"global model". In other situations where the data structure is not stable overall but it is stable in the short run, and this instability may not affect the techniques ability for forecasting over a short period, the model for forecasting over a short period is called "local model". There are no differences in the mathematical or statistical formulation of these two types of models. The only difference is in the way in which the models are used.

2. Most predictions involve forecasting more than one variable. If the variables are independent then each variable is predicting as a univariate forecasting. If the variables have some correlation, then multivariate forecasting should be used. The techniques for multivariate forecasting are studied by various Time Series Analysts. (See Hannan, Jones and Robinson.)

III. Basic Requirements for a data source to be compatible with predictive tecniques.

It is easy to see that the quality of forecasting can not be any better than the quality of data available for analysis. But it is nearly impossible to define the quality of data. Generally, the data should provide the information to meet the following requirements (see Gilchrist):

A. The data should provide directly relevant information.

B. The data should provide reliable information.

C. The data should continually and promptly provide new information.

Thè directly relevant and reliable information will help obtain forecasting results with higher accuracy, and the new information is essential for validation of the model. In practical forecasting, there are many occasions in which the data sets do not satisfy these three conditions. In this event, the following processes may help to improve the forecast.

A. The data should be obtained by the forecasters themselves.

a. The data should be examined to see how well the three requirements are met.

b. The data should also include information about the external environment.

B. Where the data were not obtained by the forecasters, the forecasters must use the robust and exploratory methods to cope with the sets of data in an informal way that will provide the forecasters with data structure, and an extensive reportoire of methods for the detailed study of the data.

IV. Problematic areas of existing Army data collections and statistical Predictive Techniques.

A. Data Sources.

In general, Army data collected are affected by the following conditions that are outside the controls of the data collector or analysts:

i. Fleet changes in aircraft configurations.

ii. Periods when the fleet was grounded.

iii. Changes in usage rates which provide for more or less exposure to replacement.

iv. The data are not fully compatible with classical statistical analysis or predictive analysis techniques.

v. There are multiple National Stock Numbers (NSN) and part numbers (NP), manufacture lots, etc. for an individual generic piece on some systems. (e.g. parts in aircraft)

vi. The location changes in fleet employment have been shown to influence part replacement rates.

vii. There are dynamic changes in maintenance procedures: inspection intervals, inspection activity, repair levels, part rework procedures, etc.

viii. Some data collections programs do not contain the relevant information about the external environment.

ix. In most cases the forecasters are not involved in data collecting plan nor process.

B. Statistical Predictive Techniques.

a. Risk/Confidence Levels: The statistical predictions

always involve some uncertainty and the desired confidence level is not always attainable.

b. Compatibility with data: Only data with a stable structure can be forecasted with low risk and high confidence levels.

c. Predictive Capabilities: The confidence level decreases as the number of lead predictions increase.

d. Data requirements: In order to obtain the validity of statistical analysis the three minimum requirements mentioned in Section III should be met.

V. Approaches. Since the basic requirements for a data set to be compatible with the predictive methodologies and problematic areas are identified, the forecaster must select an approach to filter out the unwanted information and to obtain a forecast model with high reliability. The following tasks must be accomplished to obtain a more valid analytical output of a predictive model with an acceptable confidence level.

Task 1. Clarification of the description of the data. The purpose of this task is to select from a given data set the most relevant information needed for prediction. If the forecaster was not directly involved in collecting the data, then intensive interviews with data collectors and/or field visits should be done to understand how the data were initially obtained and the standard method used. This will assist the forecaster in the selection of the forecasting model and in the interpretation of the forecasting results.

Task 2. Classification. There are many kinds of weapon and equipment systems. The scenario and environment in which each of these systems may be used could have significant impact on the collection and structuring of the data, the selection of the forecasting techniques, and in the interpretation of the forecast results. Therefore it is essential for the forecaster to classify the commodity, i.e., aircraft, missiles, etc.,of the systems for which forecasting efforts are to be applied and to identify the prediction rationale prior to the initiation of a forecasting exercise.

Task 3. Determine the effects that influence the raw data. There are certain effects that are known to influence the raw data. Some of these have already been mentioned in Section IV. A suitable adjustment should be made for the effects to improve the forecasting accuracy.

Task 4. Stratification. Many data collections contain vast amounts of information obtained from different locations, missions, usages, climates, etc. Each group contains relevant information for its particular predicting purpose.

Task 5. Examine the quality of the data. Use the three minimum requirements in Section III to determine the quality of the stratified data and whether any adjustments need to be made.

Task 6. The adjustment of data. Most Army data are not collected for the purpose of forecasting. The raw data may need to be modified to allow or disallow for some features before the date can be used in the prediction process.

a. Adjusting for known influencing causes. Some of the known causes have already been mentioned in Section IV. The ways of dealing with these vary greatly.

b. Adjusting for time period. Almost all forecasting methods assume that data are input at fixed intervals. If this is not so, then some adjustment has to be made to produce a new data set of the required intervals. For example, adjust monthly data set into quarterly or yearly data set, or adjust yearly data set into monthly, bi-monthly, quarterly, etc. data set.

c. Adjusting data by transformations. Some data sets with nonstationary variance may be transformed into a stationary one by natural logarithms. Some data sets with a nonstationary mean may be transformed into a data set with a stationary mean by applying a differencing procedure.

d. Adjustment for outliers. Most practical forecasting systems contain quality control procedures that will pick out values which are in some sense extreme (by engineering judgement, or out of some standard deviations from the mean error, etc.), and are called outliers. In an operational forecasting system, it is advisable to replace the outliers with some other suitable, but less extreme, values so that the leading forecasts will not be influenced by outliers.

Task 7. Selection of a predicting model. Having completed the above six tasks for a data set, there remains to examine the three minimum requirements of Section III again before constructing a model for its forecasting. In general, the forecaster begins to fit the simplist model with the reduced data. If the model does not fit well, according to its criteria, then the next model should then be tried. It is often that a forecaster can not obtain a model with the confidence level desired. Some experienced judgement must be made before the final model selection, or sometimes a forecaster may use a model and then modify it as new data come in.

VI. Applications: The ability to predict (forecast) a given operational parameter of a system is one of the most important elements of logistic support and managerial decision. The predicting analysis techniques can help project operational readiness, dependability, safety and hence, the probability of mission success of a system. These techniques also can assist in the development of a mathematical model for provision planning of a system, manpower maintenance planning, logistic support

planning, etc. Moreover, the result of the predicting analysis can be used to assess the engineering design specification and influence engineering designs or changes.

Summary:

There are numerous sources of data collection in the Army community. It is worthwhile to investigate which of these data collections may be used to predict safety, reliability, cost, etc. There are also several known forecasting techniques that are available, and the forecaster must use discretion in the preparation of the data to be used with each technique, as well as the selection of the technique.

References

G. E. P. Box and G. M. Jenkins. Time Series Analysis: Forecasting and Control. 2nd ed. San Francisco, Holden-Day, 1976.

W. Gilchrist. Statistical Forecasting. John Wiley and Sons, New York, 1976.

E. J. Hannan. Multiple Time Series. John Wiley and Sons, New York, 1970.

G. M. Jenkins. Pratical Experiences with Modelling and Forecasting Time Series. 1976 Annual Conference. North-Holland Publishing Company, N.Y. 1979.

R. H. Jones. Exponential Smoothing for Multivariate Time Series. J. Roy Statist, Soc. B 28,241-251, 1966.

A. Pankratz. Forecasting With Univariate Box-Jenkins Models. John Wiley and Son, N.Y. 1983.

E. Robinson. Multichannel Time Series Analysis with Digital Computer Programs. Holden-Day, Inc. San Francisco, 1967.

QUANTILE STATISTICAL DATA ANALYSIS

Emanuel Parzen[1]
Department of Statistics
Texas A&M University

Abstract.

This paper presents some reasons why theoretical and sample quantile functions should be routinely used by contemporary statistical data analysts. Quantile methods are introduced in the context of the exponential distribution as a fit to the historically important life table data of Graunt (1661). Section titles are: history of statistics and contemporary textbooks; quantile concepts; identification quantile function; identification quantile box plot; tail classification of probability laws; goodness of fit plots; IQQ plot; cumulative weighted spacings function $D(u)$; quantile simulation and distribution of extreme values; comparison quantile function; nonparametric estimation of probability density; conclusion.

1. History of Statistics and Contemporary Textbooks.

A central problem of statistical data analysis [that was formulated by 19th century pioneers such as Quetelet (1796-1874) and Galton (1822-1911)] is identifying distributions that fit the data. In *The History of Statistics*, Stigler (1986) writes (p. 268) that these pioneers emphasized the use of normal curves to fit data; they 'proposed that the conformity of the data to this characteristic [normal] curve was to be a sort of test of the appropriateness of classifying the data together in one group; or rather the nonappearance of this curve was indicative that the data should not be treated together.'

By 1875 Galton 'had devised a different way of displaying the data. He ordered the data in increasing order and, effectively, graphed the data values versus the ranks.' Galton used the name 'ogive' for the theoretical form of this curve for a normal distribution; Stigler writes 'we now call it the inverse normal cumulative distribution function'. I call this ideal graph a quantile function of the normal distribution; the graph of ordered data values, denoted $X(j;n)$, versus $(j-.5)/n$ or $j/(n+1)$, is called the sample quantile function, denoted $Q^{\sim}(u), 0 < u < 1$.

This paper presents some reasons why theoretical and sample quantile functions should be routinely used by contemporary statistical data analysts. They can be used to not only test the fit (or lack of fit) of a normal distribution to data, but also to describe other general families of distributions and to identify which distributions fit the data.

Textbooks with titles such as *Introduction to Contemporary Statistical Methods* omit many important topics that are actually useful in the theory and practice of statistical data analysis. On my list of important topics (for which I always look in the index and usually fail to find) are: uniform distribution, exponential distribution, order statistics, extreme values, quantile function. Traditional introductory textbooks describe methods based on mean and variance. To qualify as 'contemporary' a textbook adds the following topics: box plot, fences, stem and leaf plot, trimmed and Winsorised sample. In my opinion quantile function interpretations are needed for these topics to acquire beauty and utility that will excite students; however how to do this is not explicitly discussed in this paper.

We introduce the ideas of quantile-based statistical data modeling in the context of the exponential distribution. Let X be a continuous random variable with distribution function $F(x) = \Pr[X \leq x]$ and probability density function $f(x) = F'(x)$.

[1]Research Sponsored by the U. S. Army Research Office Project DAAL03-87-K-0003.

We call $F(x)$ an exponential distribution with parameter λ if

$$1 - F(x) = \exp(-\lambda x), x > 0, f(x) = \lambda \exp(-\lambda x), x > 0$$

Its mean μ equals $1/\lambda$, since (for a non-negative random variable)

$$\mu = \int_0^\infty x f(x) dx = \int_0^\infty (1 - F(x)) dx = \int_0^\infty \exp(-\lambda x) dx.$$

The standard exponential distribution is the exponential distribution with mean 1.

2. Quantile Concepts.

The QUANTILE FUNCTION $Q(u), 0 < u < 1$, is the inverse $x = F^{-1}(u)$ of the distribution function $u = F(x)$. To find $x = Q(u)$ one solves $u = F(x)$.

For an exponential distribution, one obtains $x = Q(u)$ by solving $1 - u = \exp(-\lambda x)$; therefore

$$Q(u) = (1/\lambda) \log(1 - u)^{-1} = \mu(-\log(1 - u))$$

The mean μ of a distribution F or random variable X can be computed from the quantile function Q:

$$\mu = \int_0^1 Q(u) du.$$

The MEDIAN and QUARTILES of a distribution F or random variable X are defined to be

$$Q(.5), Q(.25), Q(.75),$$

the values of $Q(u)$ at $u = .5, .25, .75$. We define QUARTILE DEVIATION DQ by $DQ = 2(Q(.75) - Q(.25))$.

For an exponential distribution, $Q(.5) = \mu \log 2 = .69\mu$; $Q(.25) = \mu \log(4/3) = .29\mu$; $Q(.75) = \mu \log 4 = 1.39\mu$. The interquartile range $Q(.75) - Q(.25) = 1.1\mu$; quartile deviation $DQ = 2(Q(.75) - Q(.25)) = 2.2\mu$.

Two important quantile concepts are $q(u) = Q'(u)$, QUANTILE DENSITY FUNCTION, and $fQ(u) = f(Q(u))$, DENSITY QUANTILE FUNCTION. For F continuous, $F(Q(u)) = u$ and $fQ(u)q(u) = 1$. For a standard exponential distribution, $fQ(u) = 1 - u$.

Two important universal measures of scale of a distribution are DQ and $1/f(median) = 1/fQ(.5) = q(.5)$. They approximately equal each other because DQ is a numerical derivative of $Q(u)$ at $u = .5$.

How do we apply these concepts to determine distributions that fit data? Given data (sample) compute a sample quantile function denoted $Q^\sim(u)$. The sample distribution function is defined by $F^\sim(x)$ =fraction of sample$\leq x$; the sample quantile function $Q^\sim(u)$ is the inverse of $F^\sim(u)$. In terms of the order statistics $X(1;n) \leq \ldots \leq X(n;n)$ of a sample

$$Q^\sim(u) = X(j;n) \text{for} (j-1)/n < u \leq j/n.$$

One usually adopts a continuous version of the sample quantile function defined by linear interpolation between its values

$$Q^\sim((j - .5)/n) = X(j;n), j = 1, \ldots, n.$$

When true mean $\mu = 18$, and the distribution is exponential, $Q(.5) = 12.4, Q(.25) = 5.2, Q(.75) = 25$. If similar values hold for the sample analogues of population parameters (denoted by adding a tilde ($\sim$) to the population notation) one suspects, and conjectures, that an exponential distribution fits.

Table 1. GRAUNT'S LIFE TABLE (1661). OBSERVED PROPORTION AND CUMULATIVE PROPORTION IN VARIOUS INTERVALS OF OBSERVED VALUES OF AGE AT TIME OF DEATH (IN LONDON 1534).

Index j	Age Interval $Q^-(u(j-1)) - Q^-(u(j))$	Proportion $p(j)$	Cumulative proportion $u(j)$
1	0 – 6	.36	$.36 = F^-(6)$
2	6 – 16	.24	$.60 = F^-(16)$
3	16 – 26	.15	$.75 = F^-(26)$
4	26 – 36	.09	$.84 = F^-(36)$
5	36 – 46	.06	$.90 = F^-(46)$
6	46 – 56	.04	$.94 = F^-(56)$
7	56 – 66	.03	$.97 = F^-(66)$
8	66 – 76	.02	$.99 = F^-(76)$
9	76 – 86	.01	$1.00 = F^-(86)$

Table 2. GRAUNT'S LIFE TABLE SAMPLE QUANTILE FUNCTION.

j	0	1	2	3	4	5	6	7	8	9=k
$u(j)$	0.	.36	.60	.75	.84	.90	.94	.97	.99	1.00
$Q^-(u(j))$	0	6	16	26	36	46	56	66	76	86

For an illustrative example we consider Graunt's Life Table data (that should be familiar to all students of statistics). It was published in 1661 by John Graunt, in an attempt to analyze data dealing with age at time of death in London. The original data was collected by Thomas Cromwell in 1534 from Church of England records of births and deaths. Graunt is credited with starting modern statistics by creating Table 1. Brilliant lectures by James R. Thompson of Rice University brought this important data set to my attention.

From Graunt's life table (Table 1) one computes sample mean $\mu^- = 18.22$ (in words, the average age at death was approximately 18 years), $Q^-(.25) = 4.2, Q^-(.5) = 11.8$ (median age at death was approximately 12 years), $Q^-(.75) = 26, DQ = 43.6$. These are found by interpolating the values of the sample quantile function in Table 2.

To compute sample mean (from grouped data) we use formulas

$$\mu^- = \sum_{j=1}^{k} .5(Q^-(u(j-1)) + Q^-(u(j)))(u(j) - u(j-1))$$

$$= \sum_{j=1}^{k} (Q^-(u(j)) - Q^-(u(j-1))(1 - .5(u(j-1) + u(j)))$$

The second formula can be interpreted using the fact that $1 - u$ is the standard exponential density quantile.

It does not seem to be customary in the literature to discuss which distributions fit the data that one is analyzing (here Graunt's life table). Techniques are discussed in this paper which can guide the statistical data analyst to identify and test standard parametric distributions (such as the exponential distribution) as a smooth distribution that fits the sample. We discuss the respective roles: (i) $F^-(x)$, sample distribution function, (ii) $Q^-(u)$, sample quantile function, (iii) $F^\wedge(x)$, smooth distribution estimated from data (for Graunt life table, an exponential distribution with mean 18.22), (iv) $Q^\wedge(u)$, smooth quantile function, (v) $D^-(u) = F^\wedge(Q^-(u))$, comparison quantile function, (vi) $D^-(u)$, cumulative weighted spacings, tests constancy of ratio of derivatives $Q^{-\prime}(u)/Q^{-\prime}(u)$, (vii) $QI(u)$, identification quantile function. The statistician's problem

is to develop a framework which explains how and why to use these functions to develop graphical and numerical diagnostics which guide us to identify distributions (such as the normal or exponential) that fit the data.

3. Identification Quantile Function.

The median, which we henceforth denote $MQ = Q(.5)$, is a universal measure of location. It is superior to the mean by the criterion of being more robust (resistant to outliers in the data whose presence will in fact be detected by the identification quantile function). But we recommend the median not because of its robustness but because it forms one of the tools of quantile based methods of statistical data analysis.

Statisticians who favor (or at least teach) mean and standard deviation as measures of location and scale use them to standardise the data by subtracting the mean and dividing by the standard deviation. The quantile based analogy to standardisation is to transform the random variable X to

$$XI = (X - MQ)/DQ$$

whose quantile function is

$$QI(u) = (Q(u) - MQ)/DQ$$

We call $QI(u)$ the Identification Quantile Function. Our motivation for introducing this function is that it is approximately equal to the unitised quantile function

$$Q1(u) = (Q(u) - MQ)/Q'(.5) = fQ(.5)(Q(u) - MQ).$$

which has value 0 and slope 1 at $u = .5$. The probability density $f(x)$ corresponding to the unitised quantile function has been normalised so that $f(\text{median}) = 1$. The unitised normal probability density is $f(x) = \exp(-\pi x^2)$.

Universal measures of location and scale are MQ and DQ. Diagnostic measures of skewness are

$$QI(.25), QI(.75), QIM = .5(QI(.25) + QI(.75)), -.25/QI(.25), .25/QI(.75);$$

note that always $QI(.75) - QI(.25) = .5$. Diagnostic measures of (left and right) tail behavior are $QI(.01)$ and $QI(.99)$. A combined measure of tail behavior (useful for probability density estimation) is $QI(.99) - QI(.01)$, called the identification quantile range.

4. Identification Quantile Box Plot.

An identification quantile box plot is a plot consisting of a box from $QI(.25)$ to $QI(.75)$ with a midline at $QI(.5) = 0$ and a cross at QIM. Fences are defined to be $\max(-1, QI(0))$ and $\min(1, QI(1))$. Lines are drawn from identification quartiles to fences. Data values outside the fences are considered outliers or out-and- outliers, depending on whether they are interpreted as representing long tails or blunders. One also indicates the location of (sample mean-MQ)/DQ. The values of identification quartiles and fences are recorded on the plot.

5. Tail Classification of Probability Laws.

Representations of the density quantile function behavior as u tends to 0 or 1 is used to provide a quantitative index of tail behavior which we call the tail exponent. It is used to qualitatively classify tail behavior in three types, called short, medium, and long. Medium tails are further classified in three groups: medium-short, medium-medium, medium-long; a good summary of these concepts introduced by Parzen (1979) is given by Schuster (1984).

These five groups reduce to three groups (short, medium, long) when expressed in terms of hazard rate functions (decreasing, constant, increasing). The right and left hazard functions are respectively defined by

$$h_1(x) = f(x)/(1 - F(x)), h_0(x) = f(x)/F(x).$$

The right and left hasard quantile functions are defined

$$h_1Q(u) = fQ(u)/(1-u), h_0Q(u) = fQ(u)/u.$$

Our classifications of tail behavior can be empirically related to the behavior of the identification quantile function as u tends to 0 or 1. The left tail is classified: $0 > QI(.01) > -.5$, short tail; $-.5 > QI(u) > -1$, medium-short; $-1 > QI(u)$, medium-long and long tail. The right tail is classified short, medium short, or long according as $QI(.99) < .5, .5 < QI(.99) < 1, 1 < QI(.99)$.

For Graunt's Life Table, $QI(.25) = -.17, QI(.75) = .33, QIM = .5QI(.75) + QI(.25) = .08, QI(.01) = -.27, QI(.99) = 1.47$. Experience with typical values of these diagnostic measures for various standard frequently encountered distributions leads one to conjecture that the sample distribution function $F^{\sim}(x)$ of the data in Table 1 is fit by an exponential distribution $F^{\wedge}(x)$ with a suitable estimated mean $\mu^{\sim}$.

6. Goodness of Fit Plots.

To evaluate the fit of a model described by $F^{\wedge}(x)$ or $Q^{\wedge}(u)$ to data described by $F^{\sim}(x)$ or $Q^{\sim}(u)$ one has a bewildering number of options. The theory of goodness of fit tests is concerned with the theoretical study of the many test statistics available, and offers little practical guidance on which methods to use in practice. This extensive literature can only be briefly illustrated in this paper, with emphasis on graphical comparisons.

One can compare plots: (1) $F^{\wedge}(x)$ and $F^{\sim}(x)$ vs. x, on the same graph; (2) $Q^{\sim}(u)$ vs. $Q^{\wedge}(u)$, called Q-Q plot; (3) $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$ vs. u, called D-uniform plot (it is equivalent to a plot of $F^{\wedge}(x)$ vs. $F^{\sim}(x)$ called a P-P plot). We recommend variants of the last method. One can interpret $D^{\sim}(u)$ as sample quantile function of the transformed random variable $U^{\wedge} = F^{\wedge}(X)$. The goodness of fit problem is transformed to tests of fit of $U^{\wedge}$ by a uniform [0,1] distribution and by estimation of the true quantile function, denoted $D(u)$, of $U^{\wedge}$. We call $D^{\sim}(u)$, $0 < u < 1$, a sample comparison quantile function.

When $F^{\wedge}$ is exponential, $D^{\sim}(u) = 1 - \exp(-Q^{\sim}(u)/\mu^{\wedge})$. Its values for Graunt's life data is given in Table 3. Figure 2 presents a IQQ plot as a test of fit of Graunt's life table by an exponential distribution. Figures 3-6 present plots on same graph of sample and smooth distributions. The combinations are $F^{\sim}(x)$ and $F^{\wedge}(x)$ vs. x (Figure 3), $Q^{\sim}(u)$ and $Q^{\wedge}(u)$ vs. u (Figure 4), $Q^{\sim}(u)$ vs. $Q^{\wedge}(u)$, a Q-Q plot (Figure 5), and $F^{\wedge}(x)$ vs $F^{\sim}(x)$, a P-P plot (Figure 6) which also plots $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$. Figures 6 and 7 present $D(u)$ plots as tests of fit of Gaunt's life table by an exponential distribution; $D^{\sim}(u)$ = cumulative weighted spacings in Figure 7.

7. IQQ (Identification quantile - quantile) Plot.

To test whether a sample is normal or exponential, one tests the hypothesis $Q(u) = \mu + \sigma Q_0(u)$ by a scatter plot of $(Q_0(u(j)), Q^{\sim}(u(j)))$ at suitable values $u(j), j = 1, \ldots, k$, in the interval $0 < u < 1$. This plot, called a Q-Q plot, is judged visually for linearity.

We prefer to use what we call a IQQ plot; it is a scatter diagram of $(Q_0I(u(j)), Q^{\sim}I(u(j)))$ with a grid of lines which may make it easier to judge visually for linearity. A IQQ plot for Graunt's life table is given in Figure 2.

8. Cumulative Weighted Spacings Function $D(u)$.

Users of QQ and IQQ plots report that they are difficult to interpret. I propose that one should prefer plots that are graphs of functions such as various functions $D(u), 0 < u < 1$, which can be defined to measure the 'distance' between two distributions.

To compare $Q(u)$ with $\mu + \sigma Q_0(u)$ we recommend comparing their derivatives (equal to $q(u)$ and $\sigma q_0(u)$ respectively). Since σ is unknown we test for constancy the ratio $q(u)/q_0(u) = q(u)f_0Q_0(u)$; equivalently test the deviation from 1 of

$$d(u) = q(u)f_0Q_0(u)/\sigma_0,$$

$$\sigma_0 = \int_0^1 q(t) f_0 Q_0(t) dt.$$

We call $d(u)$ a weighted spacings function, since spacings $X(k;n) - X(k-1;n)$ are the building blocks of estimators of $q(u)$.

One approach to testing $d(u)$ is to estimate and test the deviation (from the uniform function $D_0(u) = u$) of the cumulative weighted spacings function

$$D(u) = \int_0^u d(t) dt$$

The sample analogue of $d(u)$ and $D(u)$ to test exponentiality is: for $u(j-1) < u < u(j)$, $d\tilde{}(u) = d\tilde{}(j)$,

$$d\tilde{}(j) = (Q\tilde{}(u(j)) - Q\tilde{}(u(j-1)))(1 - .5(u(j-1) + u(j)))/\mu\tilde{};$$

$D\tilde{}(u)$ linearly interpolates its values $D\tilde{}(u(j)) = d\tilde{}(1) + \ldots + d\tilde{}(j)$. Note that $\sigma_0\tilde{} = \mu\tilde{}$.

Table 3. GRAUNT'S LIFE TABLE $Q\hat{}$, $Q\tilde{}$, $F\tilde{}$, $F\hat{}(Q\tilde{}) = D\tilde{}$ FOR FITTED EXPONENTIAL $F\hat{}(x) = 1 - \exp(-x/\mu\tilde{}), \mu\tilde{} = 18.2$, $D\tilde{}(u)$ CUMULATIVE EXPONENTIAL WEIGHT SPACINGS (CUMWTSPAC).

j	$Q\hat{}(u(j))$	$Q\tilde{}(u(j))$	$F\tilde{}Q\tilde{}(u(j))$	$F\hat{}Q\tilde{}(u(j))$	$D\tilde{}(u(j))$CUMWTSPAC
0	.09	0	.00	.00	.00
1	8.13	6	.36	.28	.27
2	16.69	16	.60	.58	.56
3	25.26	26	.75	.76	.73
4	33.39	36	.84	.86	.85
5	41.95	46	.90	.91	.92
6	51.26	56	.94	.95	.96
7	63.89	66	.97	.97	.986
8	83.91	76	.99	.98	.997
9	96.54	86	1.00	.99	1.00

Figures 6 and 7 show how we plot $D\tilde{}(u)$ for comparison with $D_0(u) = u$. In addition to the graphical diagnostic of the plot, there are many numerical diagnostics that can be performed.

9. Quantile Simulation and Distribution of Extreme Values.

A general distribution function $F(x), -\infty < x < \infty$, is a non- decreasing function continuous from the right. Its quantile function (or inverse distribution function), defined by

$$Q(u) = \inf\{x : F(x) \geq u\},$$

is a non-decreasing function continuous from the left. It is an inverse under inequality; for any x and u

$$F(x) \geq u \text{ if and only if } x \geq Q(u).$$

An important property of quantile functions is a formula for functions of random variables. THEOREM. Assume g is non-decreasing and continuous from the left. Then $Y = g(X)$ has quantile function

$$Q_Y(u) = g(Q_X(u)).$$

One can represent X in terms of a uniform [0,1] random variable U by $X = Q(U)$ since $Q(U)$ has quantile function $Q(Q_U(u)) = Q(u)$.

When F is continuous, one can transform X to U, a uniform [0,1] random variable, by $U = F(X)$ since $F(X)$ has quantile function $F(Q(u)) = u$.

A random sample $X(1), \ldots, X(n)$ of X can be simulated by generating a random sample $U(1), \ldots, U(n)$ of U, and forming $X(j) = Q(U(j))$. This process, illustrated in Figure 8 for the normal and Cauchy distributions, demonstrates that the quantile function provides a powerful graphical representation of a distribution because of the following equivalence: (1) a random sample of X, (2) observing Q(u), quantile function of X, at a random sample of points on the unit interval. To compare two distributions, such as the normal or Cauchy, one way is to plot (as in Figure 8) graphs of their identification quantile functions plotted on the same scale (the longer tailed one will have to be truncated at a suitable value).

The representation of X in terms of U by $X = Q(U)$ provides a quantile approach to the distribution theory of order statistics and extreme values. Let $X(1;n) < \ldots < X(n;n)$ be the order statistics of a random sample $X(1), \ldots, X(n)$. The kth order statistic $X(k;n)$ has the same distribution as $Q(U(k;n))$ where $U(k;n)$ is the kth order statistic of a random sample from uniform [0,1].

10. Comparison Quantile Function.

A quantile based concept that unifies parameter estimation and goodness of fit hypothesis testing procedures is the comparison quantile function $D(u) = F(G^{-1}(u))$ which compares two distribution functions $F(x)$ and $G(x)$. The comparison quantile density is

$$d(u) = D'(u) = f(G^{-1}(u))/g(G^{-1}(u))$$

The Kullback information divergence can be evaluated by

$$I(G;F) = -\int_{-\infty}^{\infty} (\log(f(x)/g(x))g(x)dx = \int_0^1 -\log d(u)du$$

The graph of $d(u)$ provides insight into the rejection method of simulation. One seeks to generate a sample $X(1), \ldots, X(m)$ from F as an acceptable subset of a sample $Y(1), \ldots, Y(n)$ from $G(x)$. THEOREM. Assume that $D(0) = 0$ and there is a constant c such that $d(u) \leq c$ for all u. Generate two independent uniform [0,1] random variables $U(1)$ and $U(2)$. Acceptance and rejection rule: If

$$U(2) \leq d(U(1))/c,$$

then accept $Y = G^{-1}(U(1))$ as an observed value of X. Otherwise reject Y. (Continue by generating two more uniform [0,1] random variables). The probability of acceptance is 1/c.

The relation between two distributions F and G is best understood by a plot of $u_2 = d(u_1)$.

This plot can be used to graphically describe the rejection rule of simulation and to prove it. Verify that the area under the curve from $u_1 = 0$ to $u_1 = G(x)$ equals $D(G(x)) = F(x)$; the event that $U(1) \leq G(x)$ and $U(2) \leq d(U(1))/c$ has probability $F(x)/c$; the event that $X \leq x$ can be shown to have probability $F(x)$.

11. Nonparametric Estimation of Probability Density.

To identify distributions that fit data, one can use parametric models such as the location-scale parameter model $Q(u) = \mu + \sigma Q_0(u)$, or one can nonparametrically form estimators $f\hat{}(x)$ of the probability density function (see Silverman (1986)). We consider only the kernel estimator

$$f\hat{}(x) = (1/n)\sum_{j=1}^{n}(1/h)K((x - X(j))/h)$$

where $K(x)$ is a probability density function and h is a bandwidth to be selected.

For K we recommend (Parzen (1962)) the 'Parzen window' which is the probability density of the sum of four uniforms

$$K(x) = \begin{cases} (4/3) - 8x^2 + 8x^3, & 0 < x < .5 \\ (8/3)(1-x)^3, & .5 < x < 1 \\ 0, & 1 < x \\ K(-x), & x < 0 \end{cases}$$

As a first choice to consider for h, by adapting Silverman (1986), p. 47, we recommend

$$hopt = K(0)DQn^{-.2}$$

To accept or reject the goodness of the value of h chosen we judge the deviation from uniformity of the comparison quantile function $D\tilde{}(u) = F\hat{}(Q\tilde{}(u))$. We evaluate this function at $u = (j-.5)/n$ by $F\hat{}(X(j;n))$. Other choices of hopt are multiples of hopt based on diagnostics of the tail behavior of the distribution, given by $QI(.99) - QI(.01)$. The deviation of $D\hat{}(u)$ from uniformity is used to guide the search for the best value of h for the data being analyzed.

The details of this procedure for choosing a kernel probability density estimator cannot be given in this paper. It is best explained by examples of the quality of nonparametric probability density estimators to which it leads for famous data sets (Buffalo snowfall, Yellowstone geyser eruption times) which are used as test cases for density estimation methods (compare Silverman (1986)).

12. Conclusion.

The process of analysing a univariate sample can be viewed as fitting a smooth distribution $F\hat{}(x)$ to a sample distribution $F\tilde{}(x)$. The process of comparing $F\hat{}$ and $F\tilde{}$ requires a knowledge of the theory and practice of quantile functions. *'In order to get to the fruit of the tree you have to go out on a limb'* is a proverb that statisticians may take as an omen that they should explore the quantile limb which is always lurking.

References

GRAUNT, JOHN (1665) *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*, 3rd. ed. London: John Martyn and James Allestry (1st ed. 1662).

PARZEN, EMANUEL (1962) On Estimation of a Probability Density Function and Mode, *Ann. Math. Stat.*, 33, 1065-1076.

PARZEN, EMANUEL (1979) Nonparametric Statitical Data Modeling, *J. Amer. Statist. Assoc.*, 74, 105-131.

SCHUSTER, EUGENE F. (1984) Classification of Probability Laws by Tail Behavior, *J. Amer. Statist. Assoc.*, 79, 936-940.

SILVERMAN, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

STIGLER, STEPHEN M. (1986) *The History of Statistics*, Cambridge: Harvard University Press.

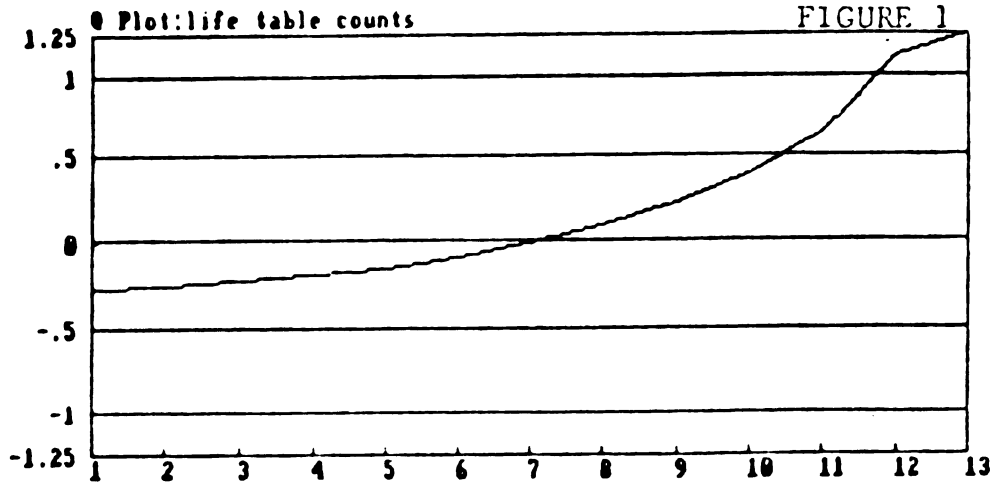


Figure 1. Identification quantile plot of Graunt life table.

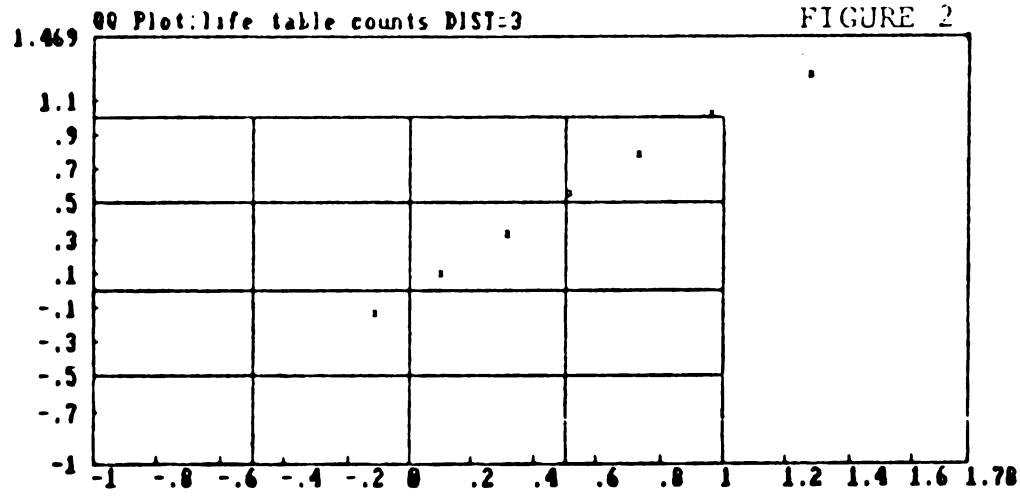


Figure 2. Identification quantile-quantile (IQQ) plot of Graunt life table vs. exponential distribution.

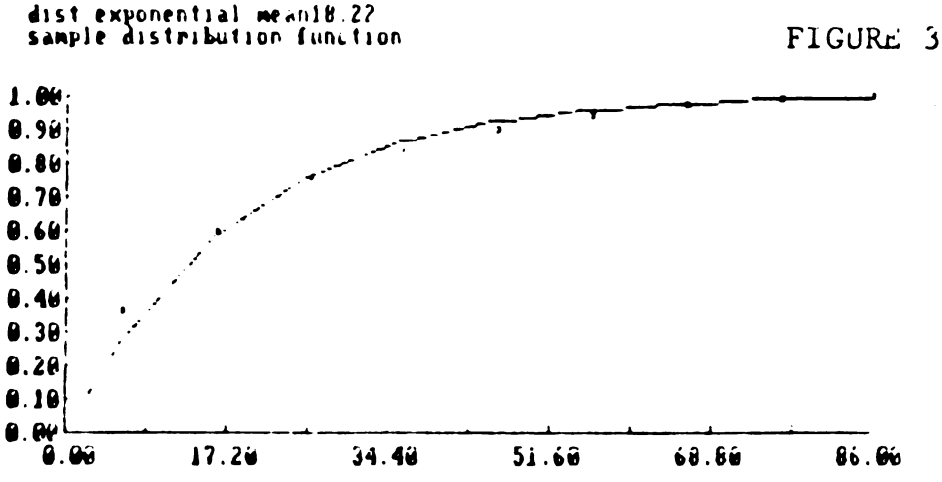


Figure 3. $F^{\sim}(x)$, Graunt life table sample distribution (dot) and $F^{\sim}(x)$, exponential mean 18.22 distribution (solid).

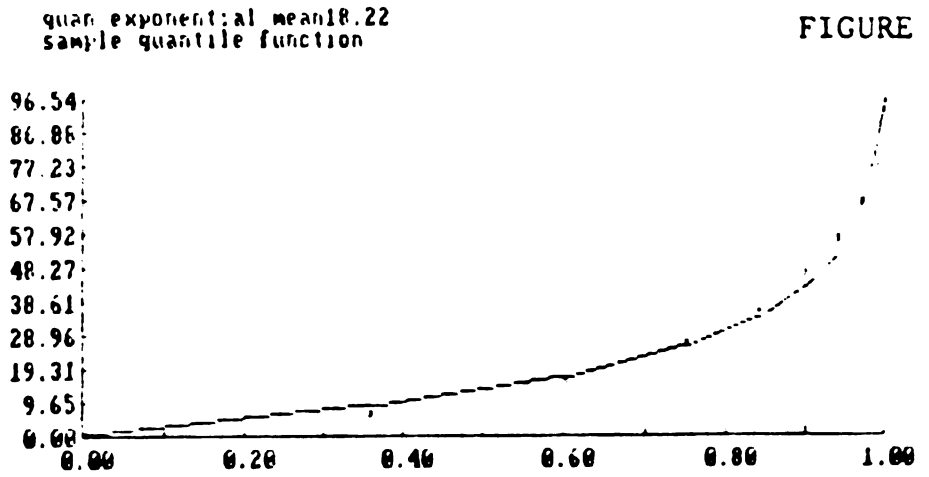


FIGURE 4

Figure 4. $Q^{\sim}(u)$, Graunt life table sample quantile (dot), and $Q^{\wedge}(u)$, exponential mean 18.22 quantile (solid).

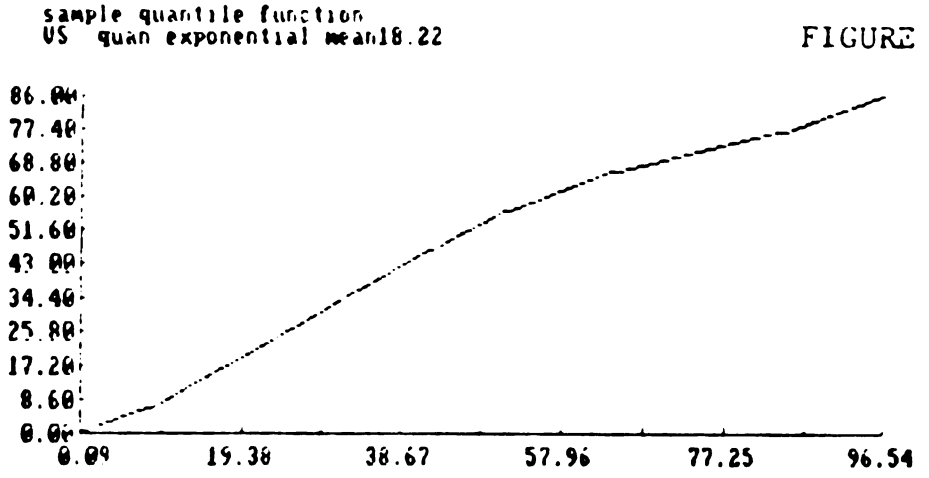


FIGURE 5

Figure 5. Q-Q plot of $Q^{\sim}(u)$ vs $Q^{\wedge}(u)$.

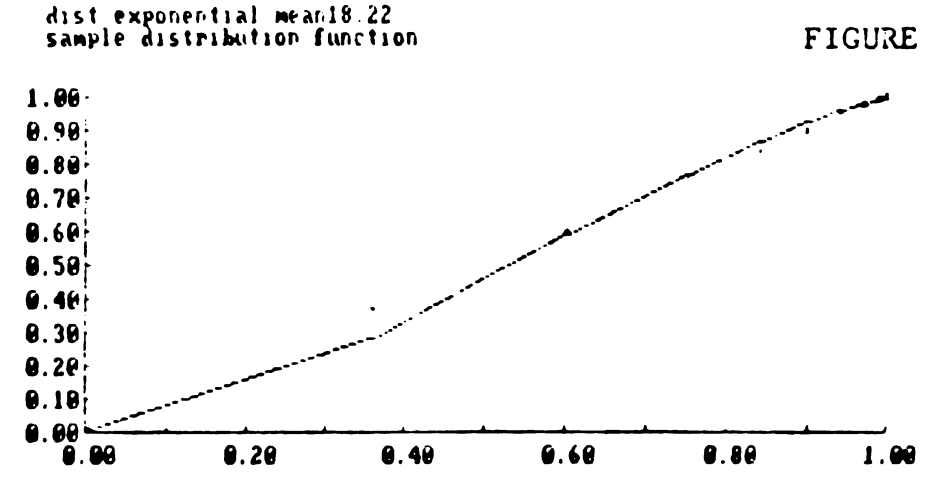


FIGURE 6

Figure 6. P-P plot of $F^{\wedge}(x)$ vs $F^{\sim}(x)$, same as plot of $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$, $D_0(u) = u$ is also plotted (dots).

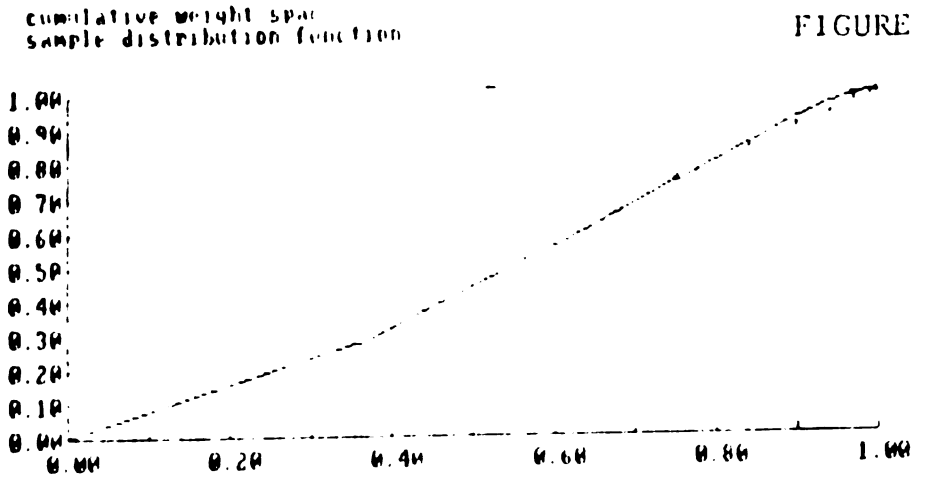


Figure 7. $D^{-}(u)$, cumulative exponential weight spacings (solid); $D_0(u) = u$ (dot).

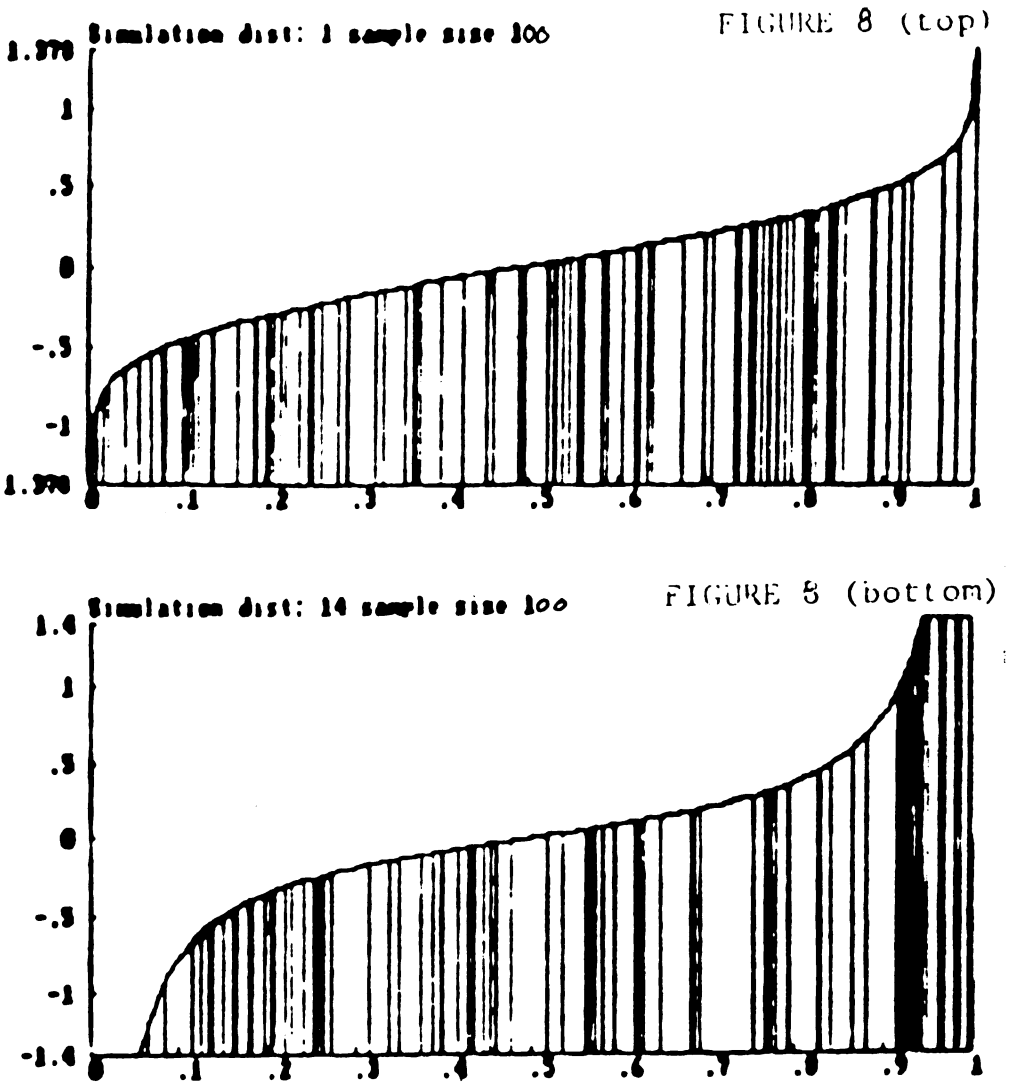
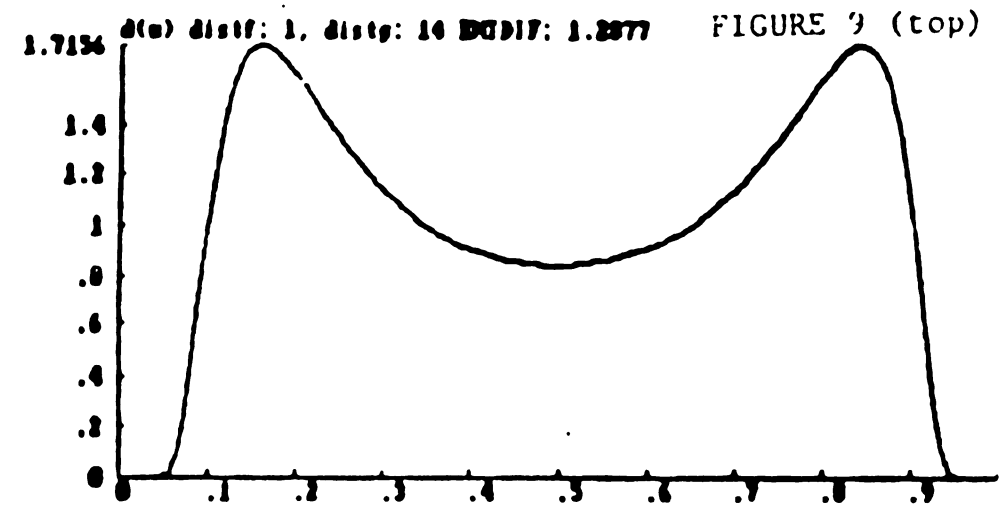


Figure 8. Random sample from normal (top) and Cauchy (bottom) represented as values of quantile function $Q(u)$ at random sample from uniform [0,1].



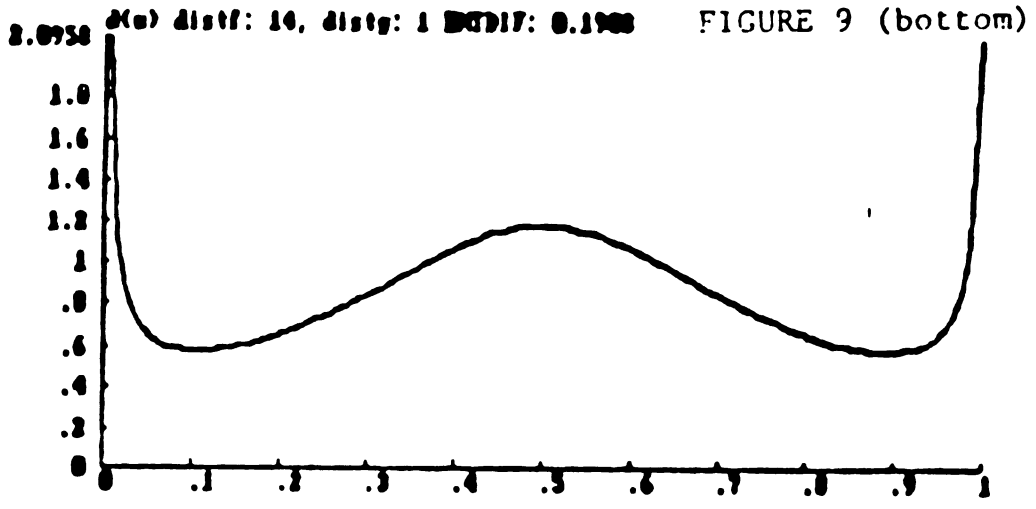


Figure 9. Comparison quantile density $d(u) = D'(u)$, $D(u) = GF^{-1}(u)$, F normal, G Cauchy, $d(u)$ bounded (top), F Cauchy, G normal $d(u)$ unbounded (below).

A COMPARISION OF TWO SENSITIVITY TESTING PROCEDURES WITH IMPLICATIONS FOR SAMPLE SIZE DETERMINATION

Barry A. Bolt
US Army Ballistic Research Laboratory

Henry B. Tingey
University of Delaware

ABSTRACT

Wu (1985) proposed an efficient class of sequential designs for estimating response distribution quantiles in a sensitivity test enviroment. Here, a good performer within that class of designs, logit-MLE, is compared to a Delayed Robbins-Monro procedure in which the final quantile estimate is obtained via maximum likelihood. Their similar Monte-Carlo performance under many test conditions is discussed. Implications for sample size determination when estimating the median and 3rd quartile are briefly considered.

I. INTRODUCTION

A sensitivity test is a destructive test in which a level of stimulus is applied to an experimental unit and a binary response is observed. The binary response is commonly referred to as a success or failure under the level of stimulus chosen.

Each member of the population from which experimental units are sampled is assumed to have a critical stimulus. For a given experimental unit, a stimulus applied at or above the critical stimulus level would necessarily result in a success. A stimulus below this critical level would result in a failure. Critical stimulus is a continuous random variable which is not directly observable, but rather only through success or failure is it observed. A success or failure constitutes only partial or indirect information, as it indicates only whether the stimulus level chosen was at or above, or below the critical stimulus for that unit.

In a sensitivity test, an adequate characterization is desired of some region or quantile of the response distribution - the distribution function of the random variable, critical stimulus. Such a characterization lends insight, in a probabilistic sense, to the sensitivity of the population to various levels of stimulus. To this end extensive literature exists, some of which is contained in our list of references.

Much of the work contained in these references pertains to sequential design and estimation. In many applications, data are expensive to collect and are gathered most cost effectively in a sequential manner. This is the case in large-caliber-munition testing for the Army. In this sequential setting our dual objective is first, to choose a good design and estimation procedure among those available; second, to briefly consider sample size determination for estimating distribution quantiles at specified levels of precision and under a variety of test conditions.

II. DESIGN AND ESTIMATION

The proposal of a new class of sequential designs and a detailed comparison of the new class to existing procedures is given by Wu (1985). Under varied test conditions a comparison of these procedures, some of which are modified, is given by Bodt and Tingey (1987). Drawing from these two studies, only the Delayed Robbins-Monro with maximum likelihood estimation and the logit-MLE will be considered as candidate procedures.

The Delayed Robbins-Monro (DRM) is a modification of the Stochastic Approximation Method of Robbins and Monro (1951). Denote the nth level of stimulus as x_n, the nth response as y_n and the quantile of interest as L_p. Let $y_n = 1$ signify a success and $y_n = 0$ signify a failure. Then referencing the work of Kesten (1958), Cochran and Davis (1964), Davis (1969) the next design points for a DRM-c design are given by,

$$x_{n+1} = x_n - c\,(y_n - p) \tag{1}$$

where c is an appropriately chosen constant according to the variance of the population.

Data is collected in this manner until a reversal occurs. Reversal is the occurrence of a (success, failure) or (failure, success) in succession. Subsequent design points are chosen according to the usual Stochastic Approximation Method by,

$$x_{n+1} = x_n - \frac{c}{n-k+1}(y_n - p) \tag{2}$$

where k is the first sample number corresponding to the first reversal.

The primary advantage to delaying the reduction in step size until the first reversal is evident in the common situation where a reasonable guess for the quantile location is not available. The design refrains from attempted convergence until some indication (reversal) of being in the desired region is present. This convention makes the most sense if the quantile of interest is the median but will be used here for the .75 quantile as well. Davis (1969) shows the DRM to be a good performer.

The logit-MLE is one application of Wu's general technique found by him to be effective in estimating the quantiles of the distribution. The next design point is taken to be the desired quantile's maximum likelihood estimate based on all of the data gathered up until that point. The maximum likelihood procedure assumes a logistic model, hence the name logit-MLE. Silvapulle (1981) shows that the unique existence of this maximum likelihood estimate is guaranteed by a zone of "mixed" results; the necessary and sufficient condition for which can be expressed,

$$(x^1_{min}, x^1_{max}) \cap (x^0_{min}, x^0_{max}) \neq \phi \tag{3}$$

where x^1_{min} is the minimum level of stimulus at which a success was observed. In the first few tests there is reasonable likelihood that this condition will not be satisfied. Furthermore, use of maximum likelihood estimation on only a few sensitivity data points often results in poor estimates. What is needed is another data collection procedure to be used until the logit-MLE can be applied. In this study, the Delayed Robbins-Monro was used until condition (3) was satisfied and more stable maximum likelihood estimates were likely.

An algorithm for this procedure is to collect data as per DRM-c until condition (3) is satisfied or sample point six has been reached, which ever comes later. After which time the next design point, x_{n+1}, is taken to be the logit-MLE with restrictions imposed by the following equations.

If d_n is the solution of

$$\hat{L}_p = x_n - \frac{d_n}{n-k+1}(y_n - p) \tag{4}$$

where $\hat{L}_p$ is the logit-MLE for the p^{th} quantile based on n observations, then

$$x_{n+1} = x_n - \frac{d^*}{n-k+1}(y_n - p) \tag{5}$$

where

$$d^* = \max[\delta, \min(d_n, d)] \quad d > \delta > 0. \tag{6}$$

Here δ is fixed at .01.

This restriction prohibits the procedure from varying wildly in its choice of the next design point. Henceforth, the above procedure will be denoted MLE(c,d). For a more detailed discussion see Wu (1985).

Preparing to compare the two procedures, DRM-c and MLE(c,d), we note that when no prior knowledge of the distribution is available Wu (1985) finds MLE(c,d) to be better than the RM type designs he examined. In addition, Bodt and Tingey (1987) show in a Monte-Carlo study that MLE(c,d) specifically out performs DRM-c under a variety of practical test conditions such as restricted sample sizes ($\leq$ 15), stimulus noise, varied response distributions, and varied combinations in the selection of the initial design point and the constant c. Based on this work we will make one final modification of the estimation technique when using DRM-c. To motivate that change we will first take a brief digression.

One goal of this experimentation is to precisely estimate a quantile of the critical stimulus distribution. Since the advent of sequential procedures in this setting, much of the attention has been placed on asymptotic convergence properties. It is true that many of these procedures for collecting data also serve to consistently estimate. In addition, designs such as DRM-c are nonparametric so no restrictive model assumptions need be made regarding the shape of the response distribution. For these reasons some experimenters have ceased to separate design and estimation when considering this problem. Consistent with the experiment goal mentioned, the performance of various combinations of design and estimation procedures are examined, Bodt and Tingey (1987). In the restricted sample size environment we found that if data were collected using DRM-c and estimation was carried out via maximum likelihood, the results were as good or better than for any other design and estimation scheme studied. This result was true under the variety of practical test conditions mentioned previously.

Thus the promised modification is that when using DRM-c, data is collected as that procedure dictates; but final estimation is accomplished using the same logit-MLE technique as per Wu's procedure. We will continue to refer to this combined design and estimation scheme as DRM-c.

III. A SIMULATION STUDY

Before making sample size determinations, we wished to first compare DRM-c and MLE(c,d) under practical test conditions and sample sizes which are not unduly restricted. This comparison was performed in a Monte-Carlo study under the crossed conditions listed in Figure 1. For this part of the study the .5 quantile was estimated. The measure of precision was $(MSE)^{1/2}$. The number of iterations performed was 500 per treatment combination.

FACTORS	LEVELS
Response Distribution	normal, Cauchy, exponential, uniform
Initial design point	median, median - 3
Design & Estimation	MLE (10,30), MLE (20,30) DRM-10, DRM-20
Sample size	10 to 50 by 5

Figure 1. Factors Included in the Design.

Four response distributions representing a variety of shapes were chosen. Each had median equal to zero. Three were given a standard deviation of unity. The quartiles of the Cauchy were made to match those of the normal distribution. The purpose in considering c = 10, 20 for DRM-c is that we wished to compare DRM-c to MLE(c,d) under suboptimal conditions for the data collection aspect of DRM-c while maintaining good conditions for MLE(c,d). Based on Wu's findings, d=30 should yield good results for MLE(c,d). Through results of Chung (1954) and Hodges and Lehmann (1955) the optimum choice of c for the usual Stochastic Approximation Method is $(F'(.5))^{-1}$ where F is the response distribution. For the response distributions chosen, these values of $(F'(.5))^{-1}$ range between 2 and 2.5. Thus the chosen values 10, 20 are much removed from the optimum and will act to slow convergence relative to the optimum. It is in such an environment, suboptimal values of c or limited prior knowledge of the response distribution, where MLE(c,d) was shown to be superior to RM type designs.

The results of the simulation comparison are efficiently represented in graphical form. In Figure 2 we are examining the relative magnitude of $(MSE)^{1/2}$ for nine sample sizes. The true response distribution was normal, and the initial design point was zero as indicated by the arrow. To obtain the DRM-10 and MLE(10,30) points for each sample size, the same random number sequence was used for both in each iteration so that any difference in the quality of the design points chosen was a function of the design under these particular conditions. Unless otherwise noted, the procedures yield estimates which are, for practical purposes, unbiased.

Given that the response distribution standard deviation is unity, the mild fluctuations between procedures illustrated here are considered negligible. Similar results hold true for the uniform and Cauchy distributions. See Figures 3-4.

The disparity in precision among the three distributions for small sample sizes is believed to be caused by the different response distribution shapes. The reasons for this belief are given in the following discussion. Since the disparity is most noticeable between the Cauchy and the other two, the discussion will focus on the effect of the heavy tails of the Cauchy distribution.

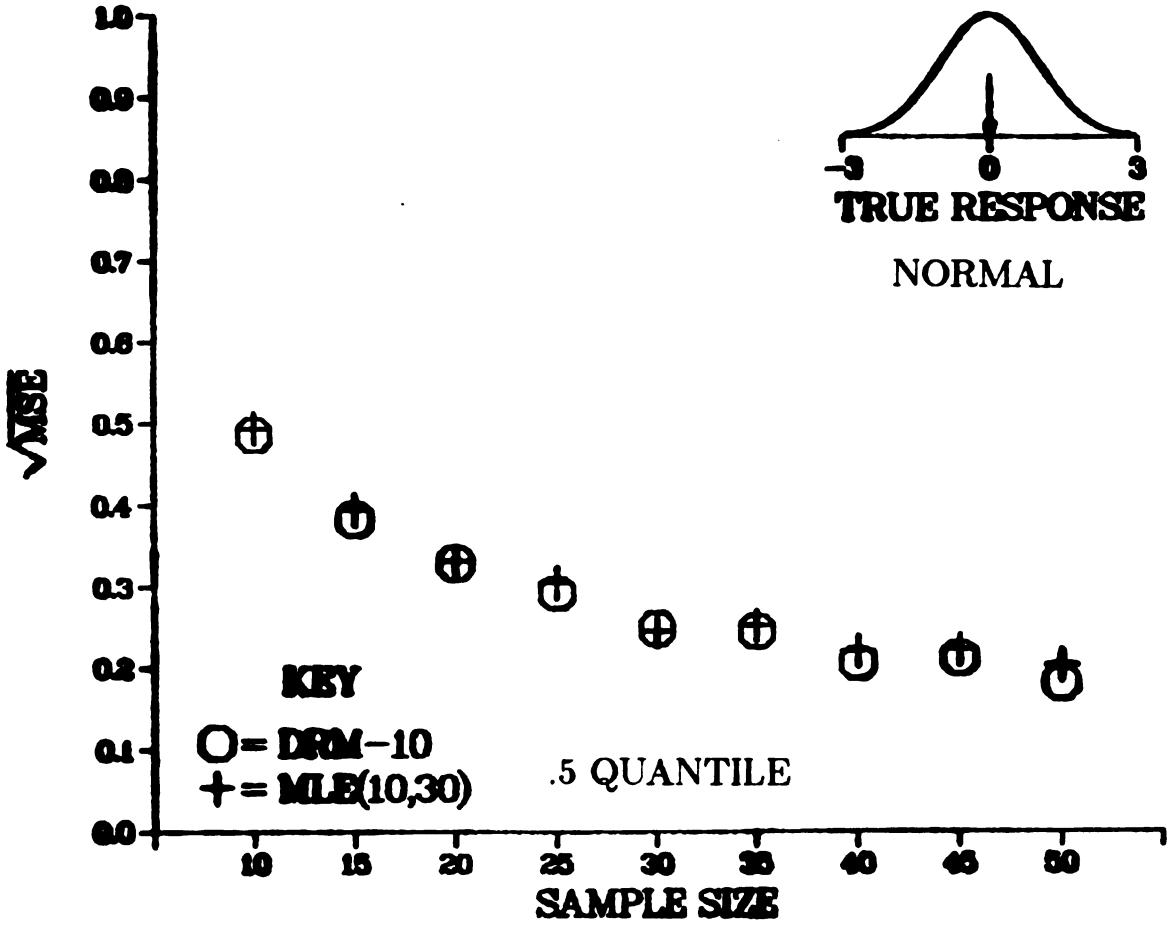


Figure 2. Precision Under a Normal Response Distribution.

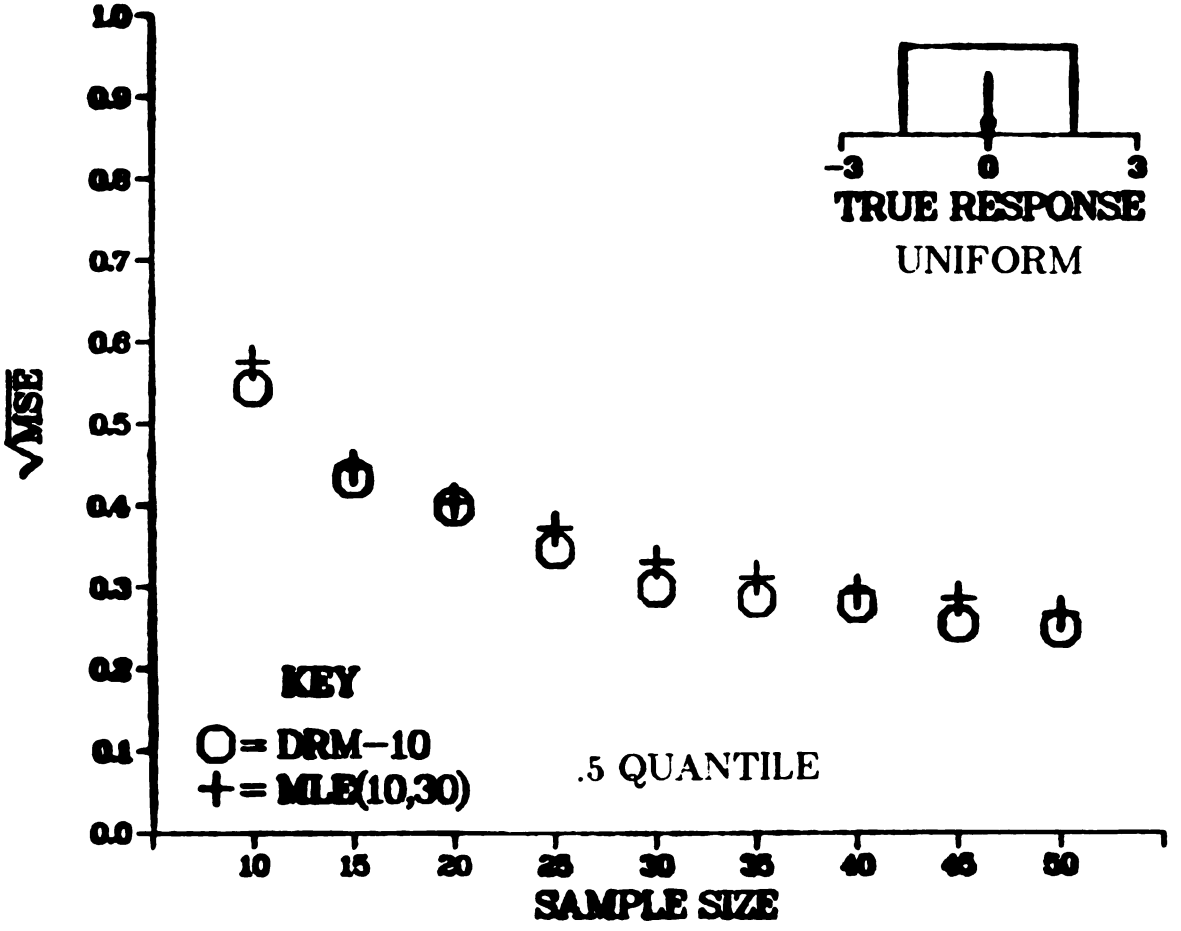


Figure 3. Precision Under a Uniform Response Distribution.

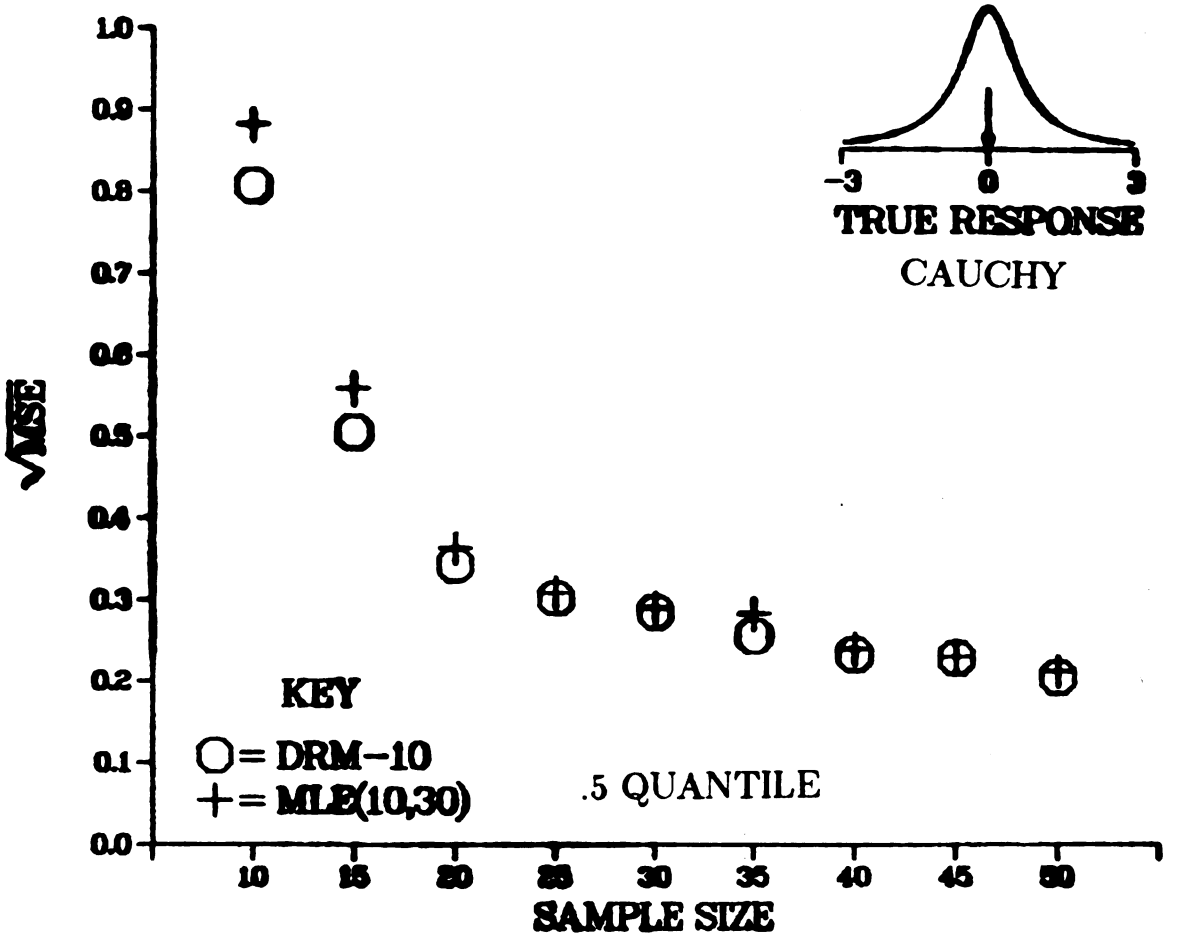


Figure 4. Precision Under a Cauchy Response Distribution.

First, consider the estimation of the median of the response distribution. Define a wrong decision as moving away from the median to collect the next design point. Wrong decisions inflate the variance of the maximum likelihood estimate. This follows because when the design steps away from the median it causes the collection of data holding less information regarding its location. Banerjee (1980) shows this rigorously for a normal response distribution. Additionally, when a wrong decision occurs a larger value of $\frac{c}{n-k+1}$ (consistent with small samples) will result in data collection farther from the median than for a small value of that quantity. In an extreme case the design may begin errantly sampling in the tail of the response distribution and take several steps to return to levels of stimulus more likely to yield useful information. Second, if presently sampling in the tail of the response distribution, the Cauchy distribution is more likely to cause a wrong decision than the other two. If x_n is currently below the true median a wrong decision occurs with probability $F(x_n)$. Thus, for a fixed x_n in the tail area of the Cauchy distribution $F(x_n)$ is large relative to corresponding probabilities as evaluated for the normal and uniform distributions. Third, if few samples are used the importance of the informational content of those samples is accentuated thus leading to the disparity mentioned.

In Figures 5-8 all 500 iterations are represented in histogram form. The observations are estimates of the median by DRM-10 or MLE(10,30) under the Cauchy response distribution for a sample size of 15 or 35. The arrow indicates the true median. As expected after viewing Figure 4, no substantive difference exists among the empirical densities.

In Figures 9-10 the exponential response distribution is considered, with results similar to the previous three, in terms of relative precision. However, as Figure 10 illustrates, the estimates produced by either method are biased with DRM-10 arguably more biased than MLE(10,30). V_{50}, zero in this case, denotes the median of the response distribution associated with critical velocity. Velocity is a common stimulus in Army testing. Each point on Figure 10 represents the average of 500 estimates of V_{50}. The reasons for the bias are similar to the reasons for precision disparity mentioned earlier. Although not displayed, similar results hold true for comparison of DRM-20 to MLE(20,30) and under the condition of the initial design point equaling the median - 3.

In estimating the median, the results are clear. There is virtually no difference in precision between the two procedures for a variety of response distribution forms. In general the designs must be judged equivalent in their ability to gather pertinent data for the estimation of the median, since the estimation is accomplished using maximum likelihood with a logistic model for each and the random number sequences were identical for each. The only studied exception was that MLE(c,d) produced slightly less biased estimates than did DRM-c for the exponential response distribution. In this case it appears that MLE(c,d) gathered data in a slightly more efficient manner.

Their general equivalence is important to consider when choosing a design. Extending the comparison to computational ease, DRM-c is easier to employ than is MLE(c,d) in many practical settings. Prior to each test DRM-c requires of the field experimenter

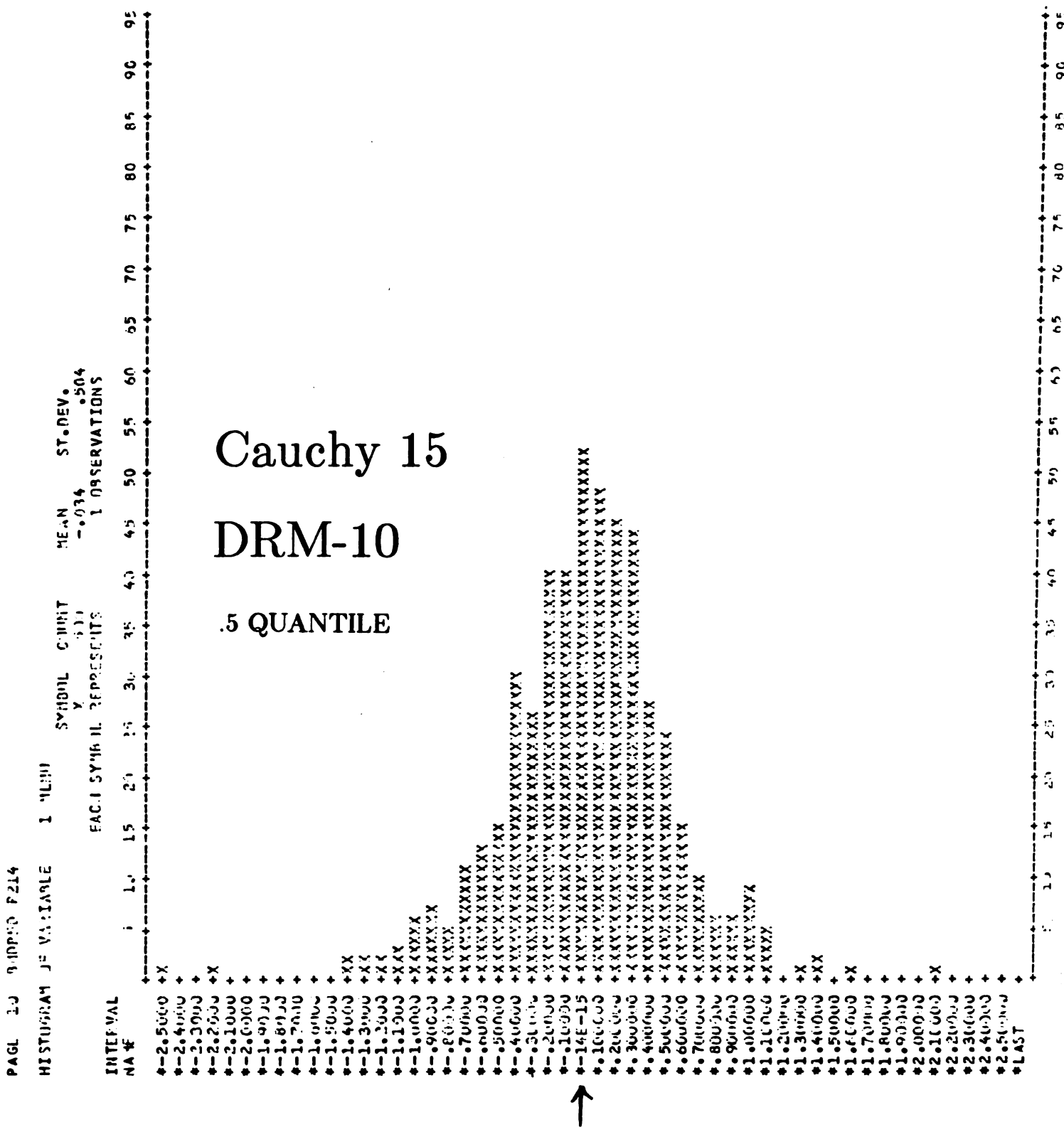


Figure 5. Empirical Density of the Estimator DRM-10.

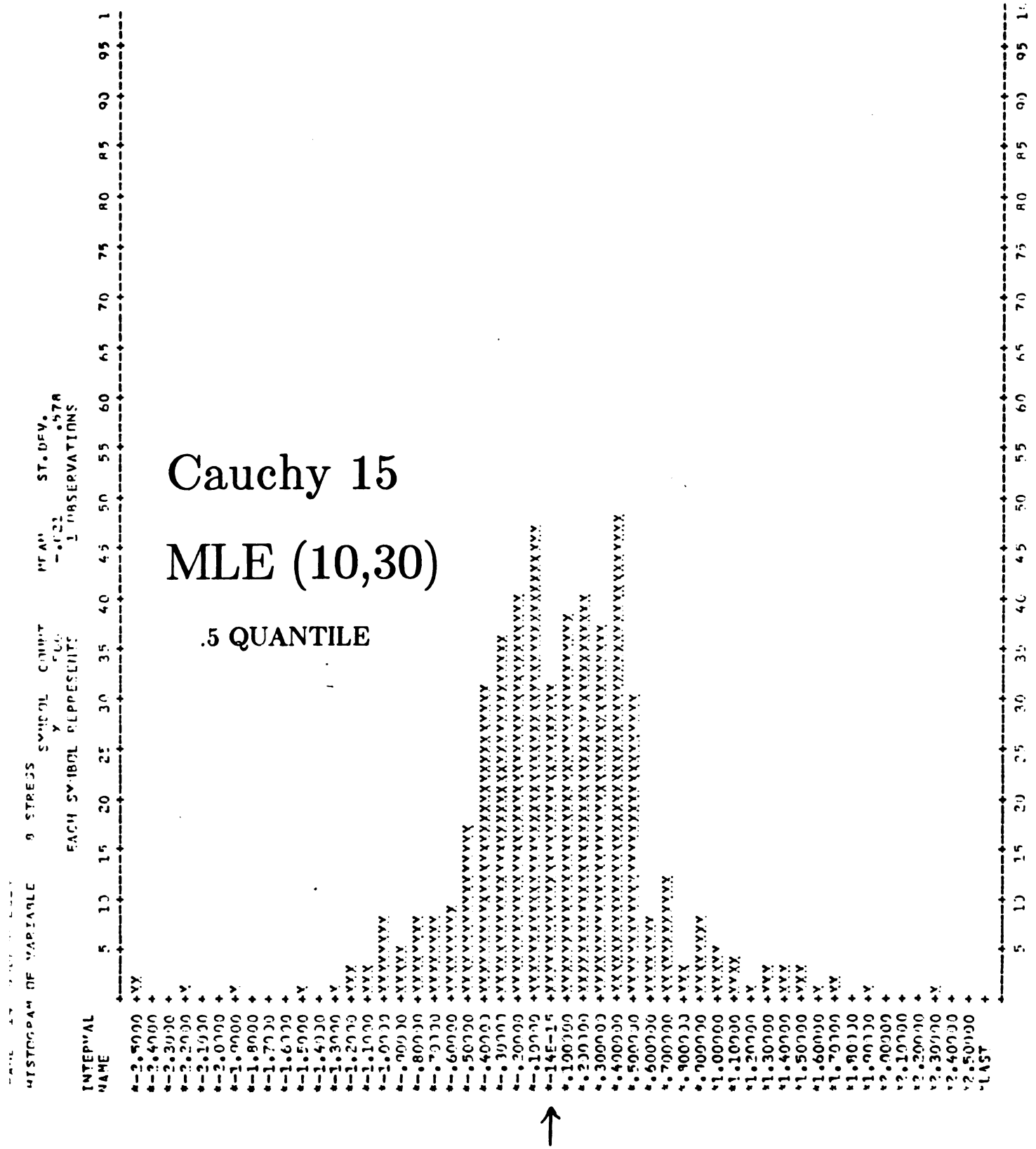


Figure 6. Empirical Density of the Estimator MLE(10,30).

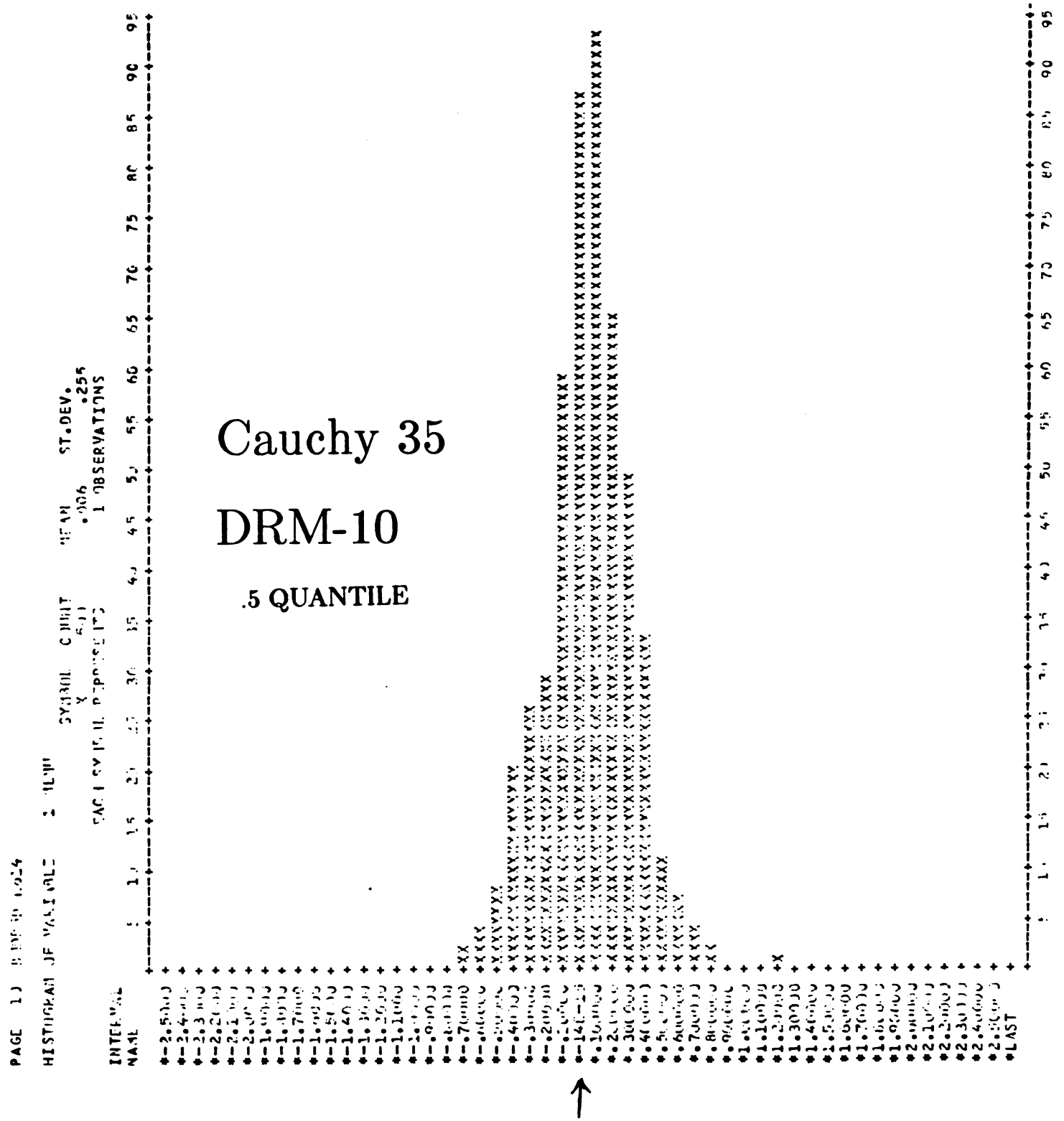


Figure 7. Empirical Density of the Estimator DRM-10.

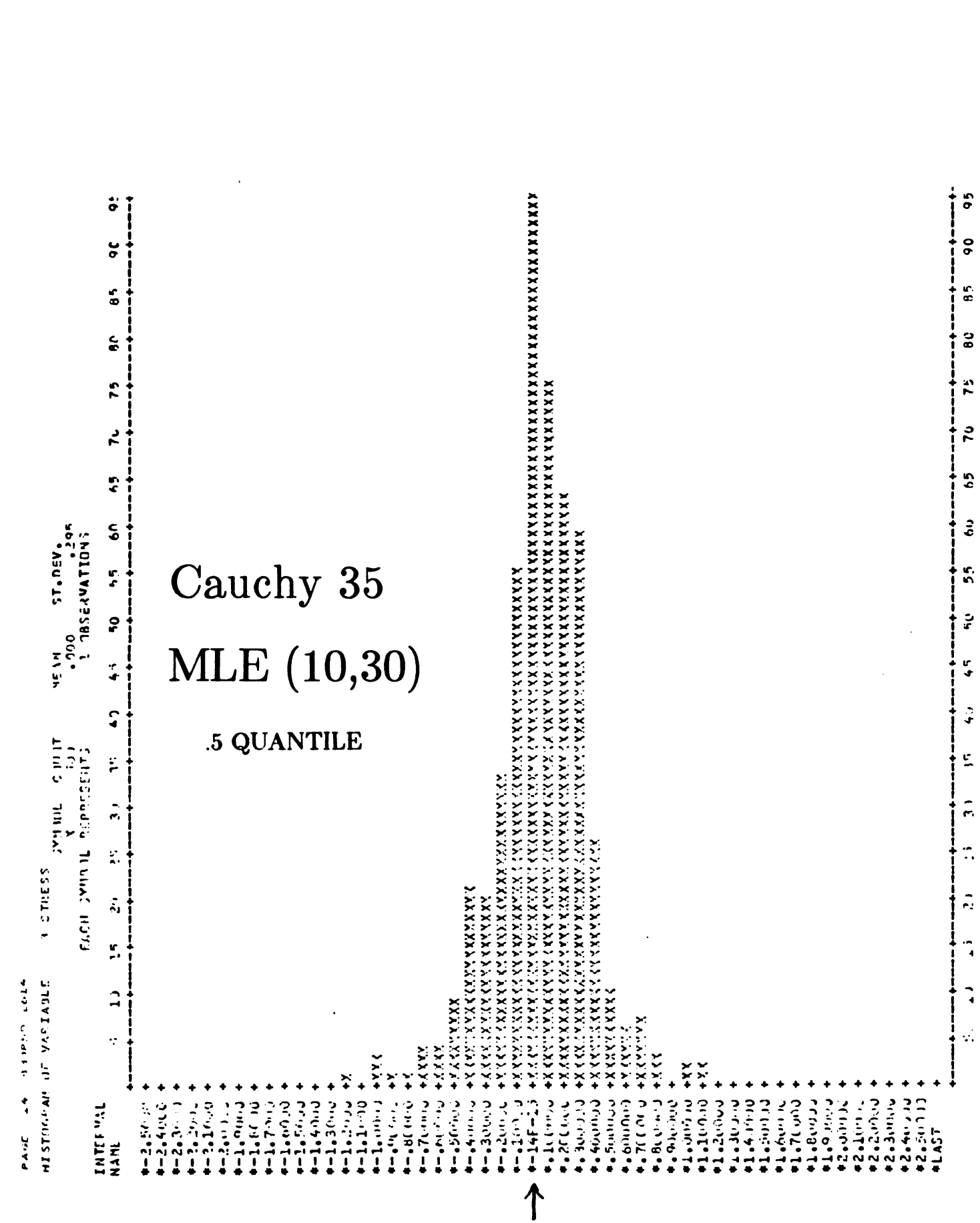


Figure 8. Empirical Density of the Estimator MLE(10,30).

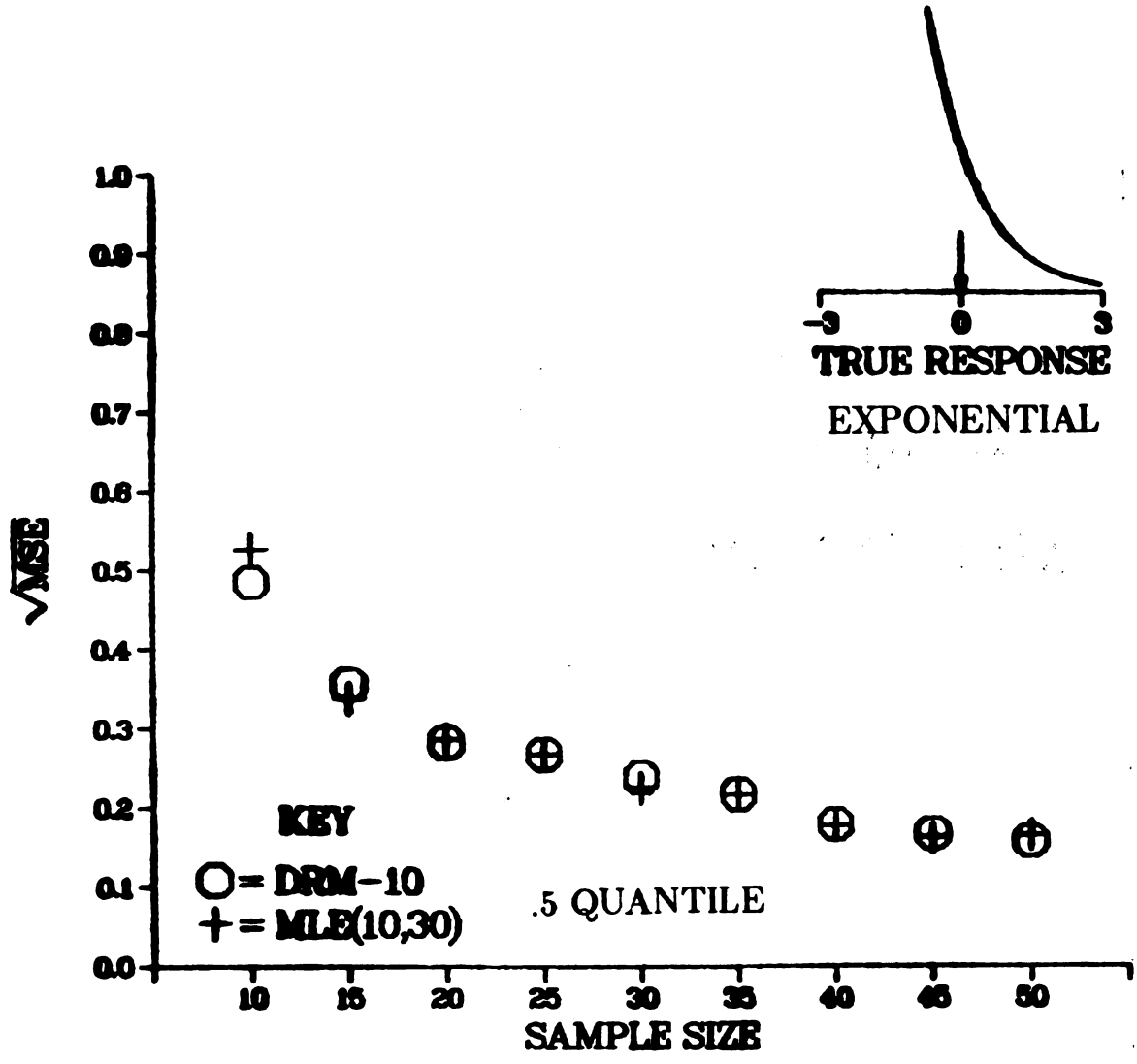


Figure 9. Precision Under an Exponential Response Distribution.

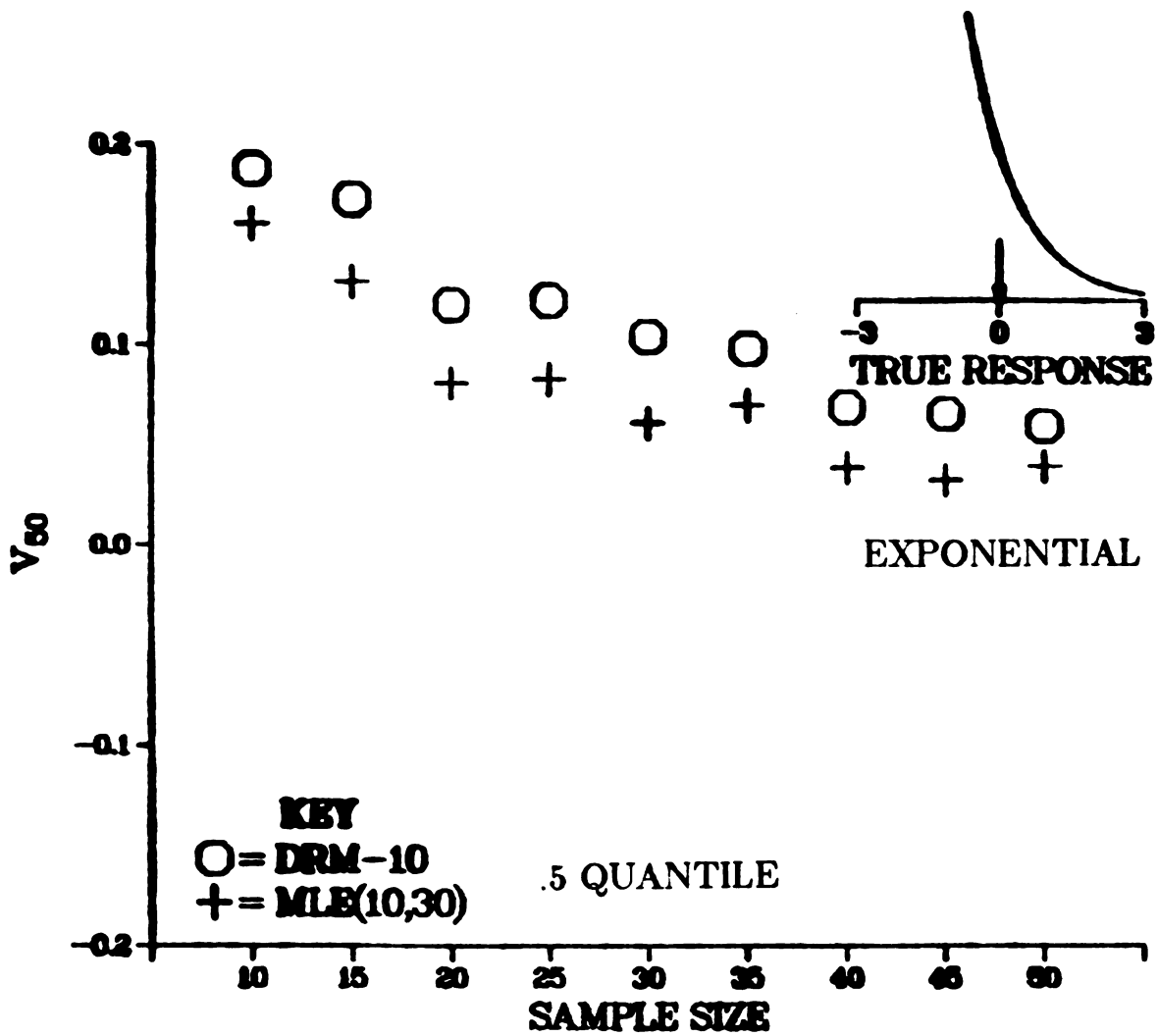


Figure 10. Average of the Estimates Under an Exponential Response Distribution.

only the solution of a single equation with one unknown. This is a computation that certainly could be performed by hand. Not until the data are collected is it necessary to iteratively solve for the maximum likelihood estimates in a more time consuming effort. Assuming that the conditions for maximum likelihood estimation have been met, use of MLE(c,d) would require that prior to each sample taken, an iterative solution for the estimates be performed. The potential difficulties in a field test are obvious.

At this point we mention two additional facts regarding MLE(c,d). First, there is a suggested provision for delaying implementation of maximum likelihood estimation, Wu (1985). We denote this MLE(c,d,*l*), where *l* is a lag delay after the unique existence criterion is satisfied; after which maximum likelihood estimation is to be employed. The intent of this provision is to delay use of maximum likelihood estimation until it is likely that the estimates will be more stable. This is why we used maximum likelihood no sooner than in the selection of the 7th design point. We mention this for completeness because *l* could be chosen to be variable so that maximum likelihood estimation was delayed until the last data point was gathered; in which case MLE(c,d,*l*) reduces to DRM-c as defined in this study.

Second, although an iterative solution is necessary to solve for the maximum likelihood estimates, Wu (1985) does suggest an approximation which would eliminate the need for an iterative solution. The approximation is valid if design points are close to the quantile being estimated. Caution is warranted when using this approximation in a small sample test environment with no prior knowledge of the response distribution. There, closeness of the design points to the quantile of interest cannot be assured.

Thus far only the median has been considered. It is certainly possible, and in many situations more desirable, to estimate quantiles other than the median. The median is the quantile commonly used for inference primarily because it is the easiest to estimate. We also compared the two procedures for estimating the 3rd quartile. In practice, for estimating quantiles beyond the first or third quartile, specific extreme value designs may be more practical.

Figure 11 shows the precision of the two procedures when estimating the 3rd quartile of the normal distribution. Once again, any differences between the two methods appears negligible. The procedures appear to be biased in estimating the 3rd quartile for small sample sizes. In Figure 12 the ordinate is now averaged estimates of the 3rd quartile. The arrow represents the true quantile value, .675.

Figures 13-14 concern estimation of the 3rd quartile of the Cauchy response distribution. Remember that the normal and Cauchy response distributions were chosen to have the same quartiles. Thus by comparison we see that the precision of the methods is much worse for the heavier tailed distribution. It does appear that MLE(10,30) tends to be more precise and less biased than DRM-10 for larger sample sizes.

Figures 15-16 constitute our cursory look at sample size determination. Our approach was to indicate the best and worst precision for each method for the different sample sizes. The extreme precisions were extracted from the performance of the procedures under the four response distributions. Initial design point selection and

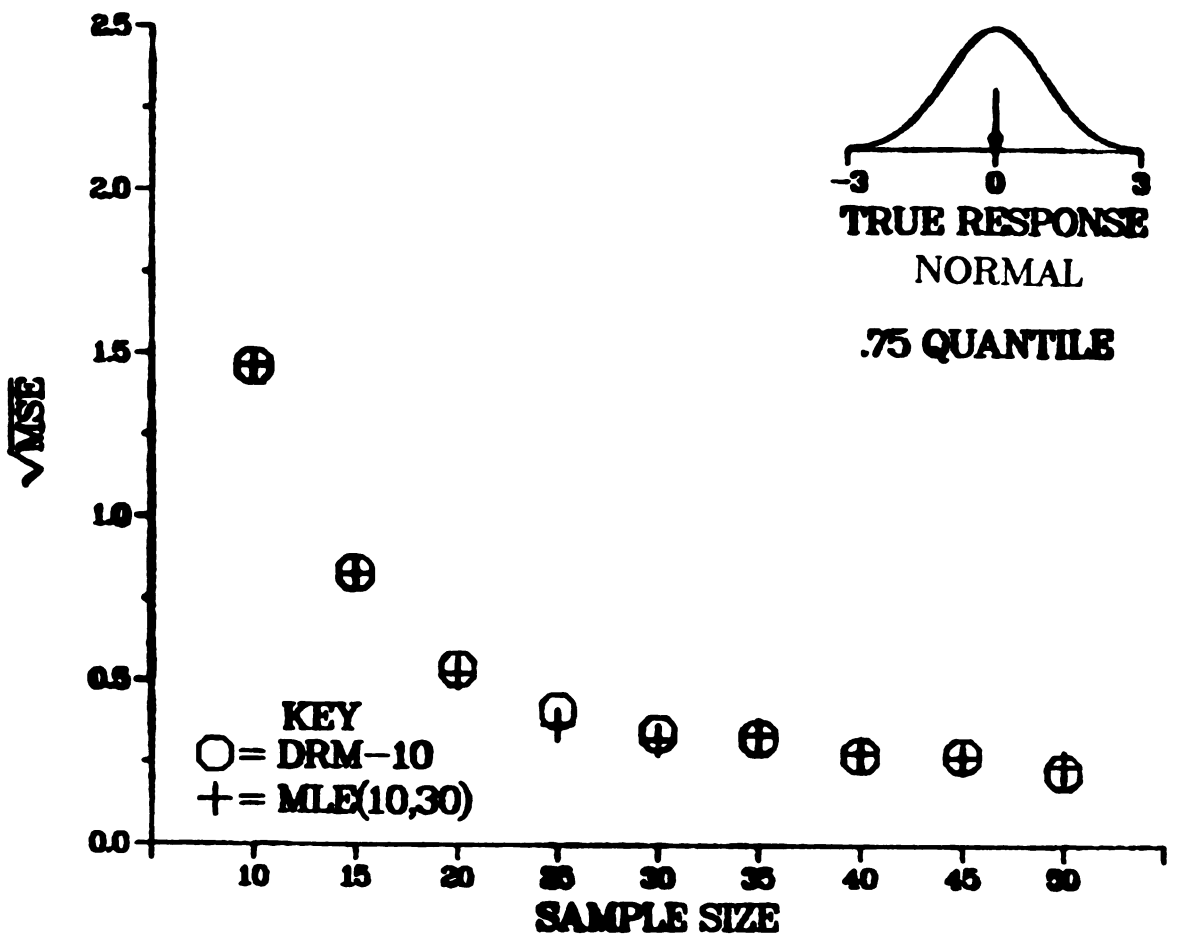


Figure 11. Precision Under a Normal Response Distribution.

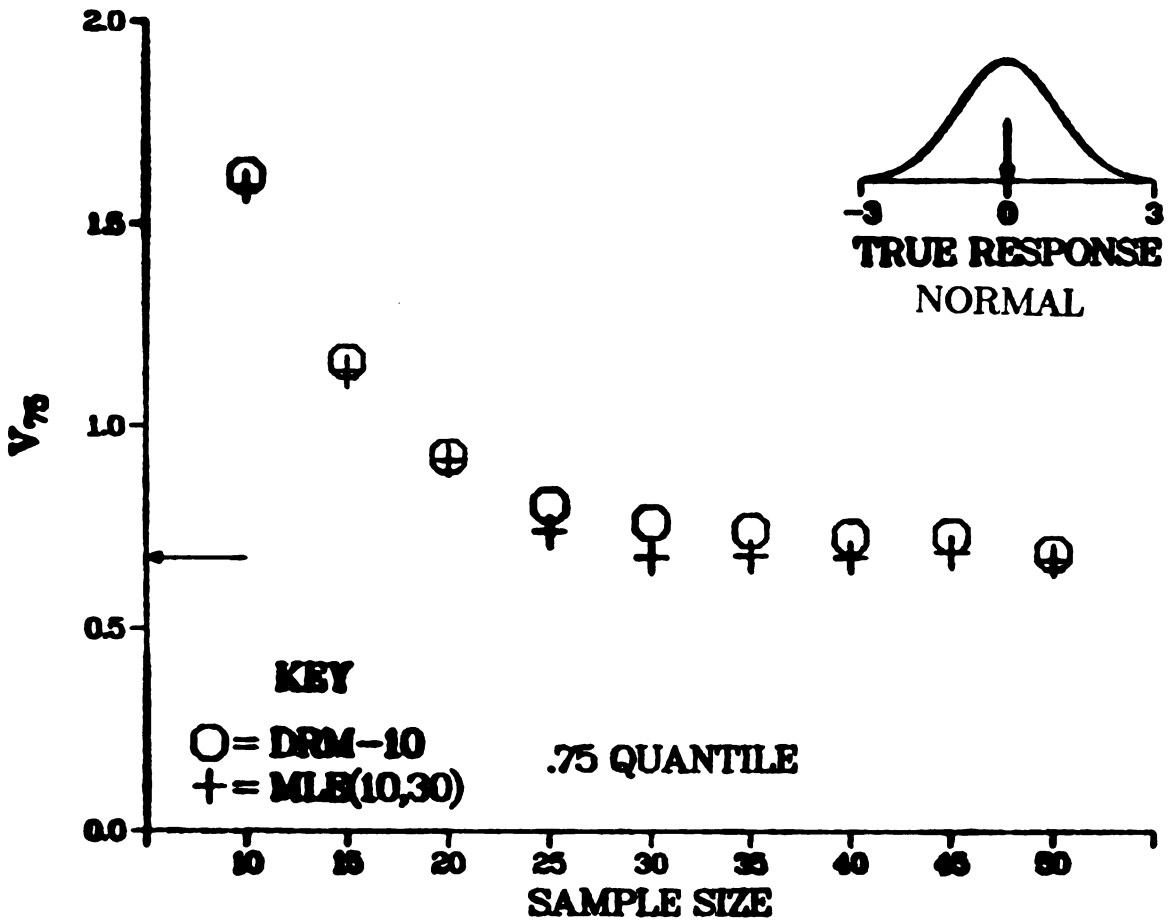


Figure 12. Average of the Estimates Under a Normal Response Distribution.

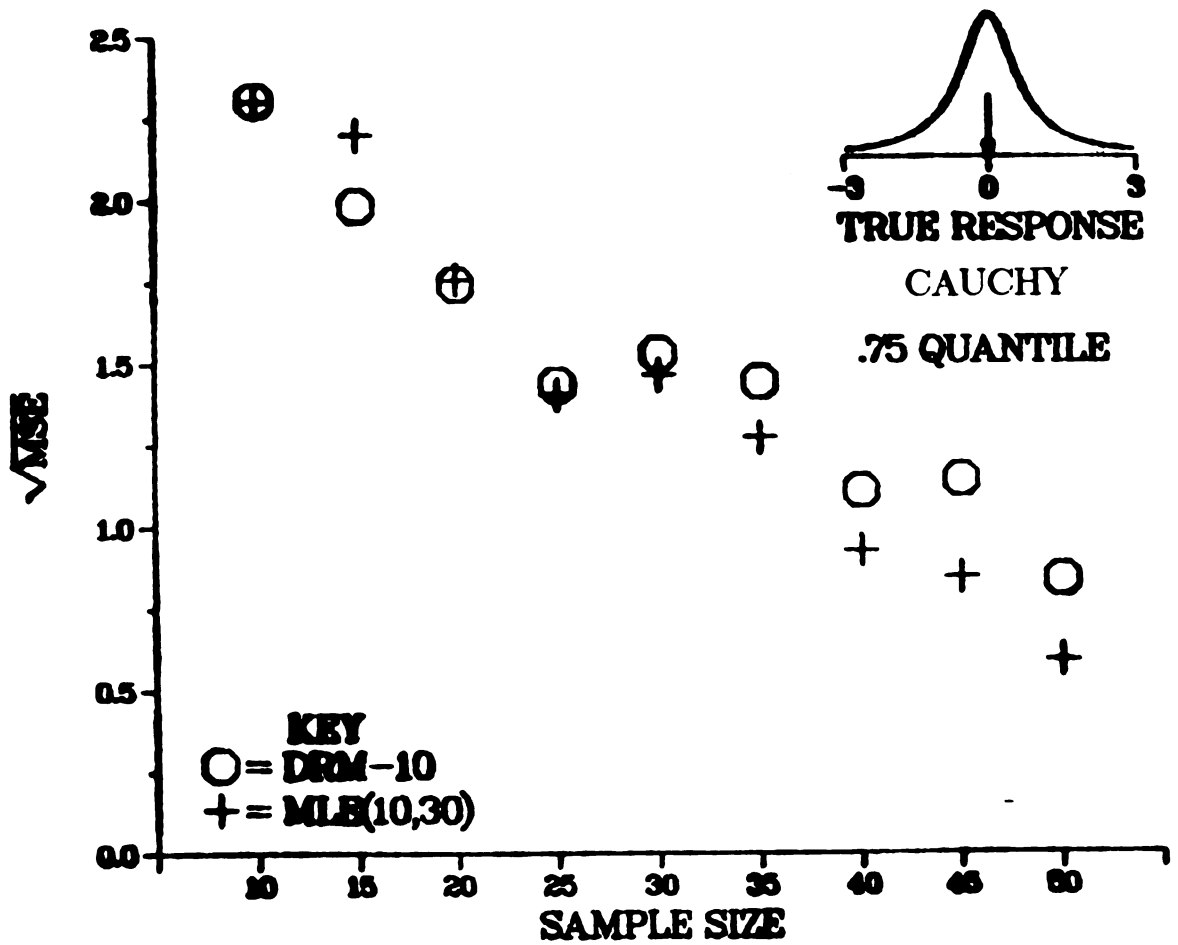


Figure 13. Precision Under a Cauchy Response Distribution.

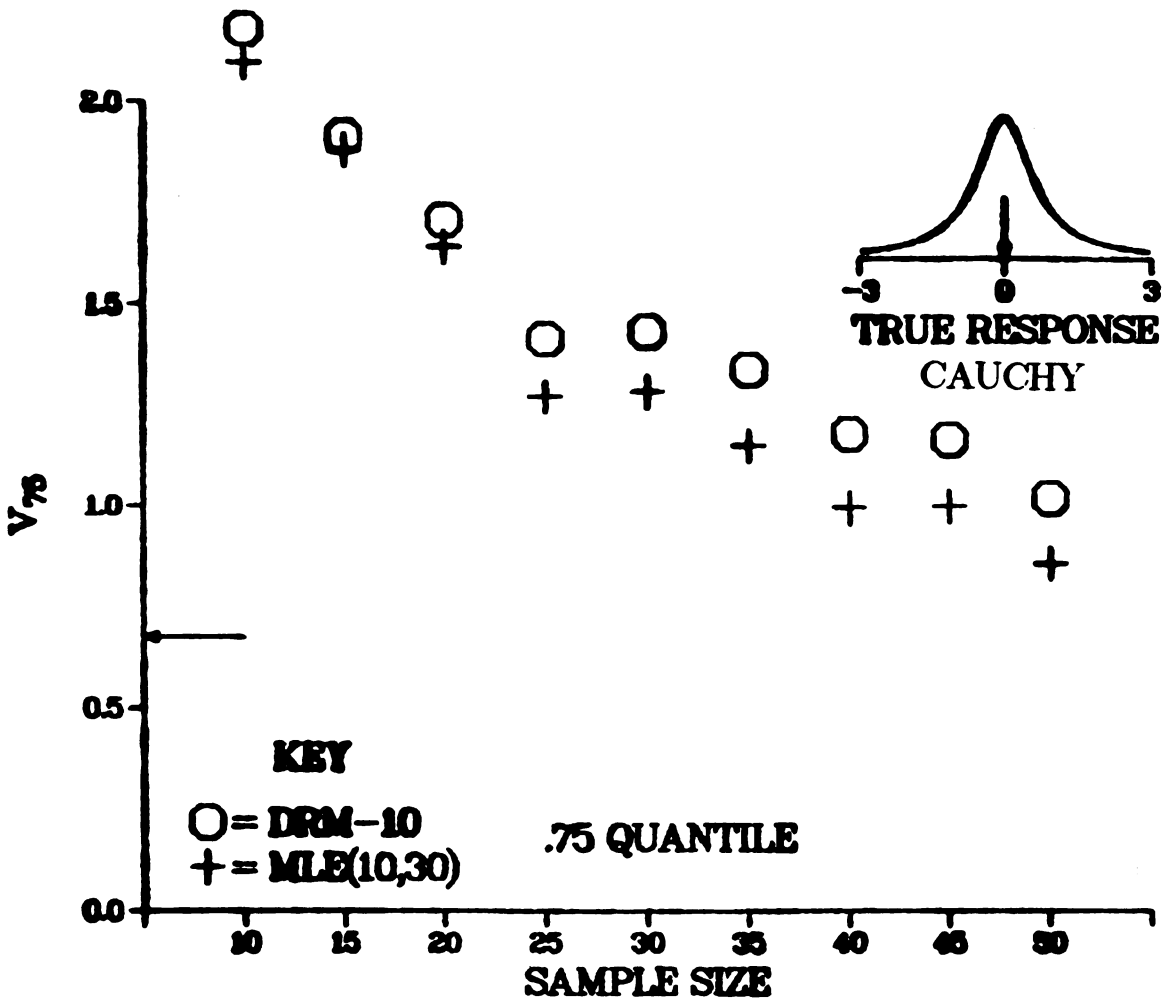


Figure 14. Average of the Estimates Under a Cauchy Response Distribution.

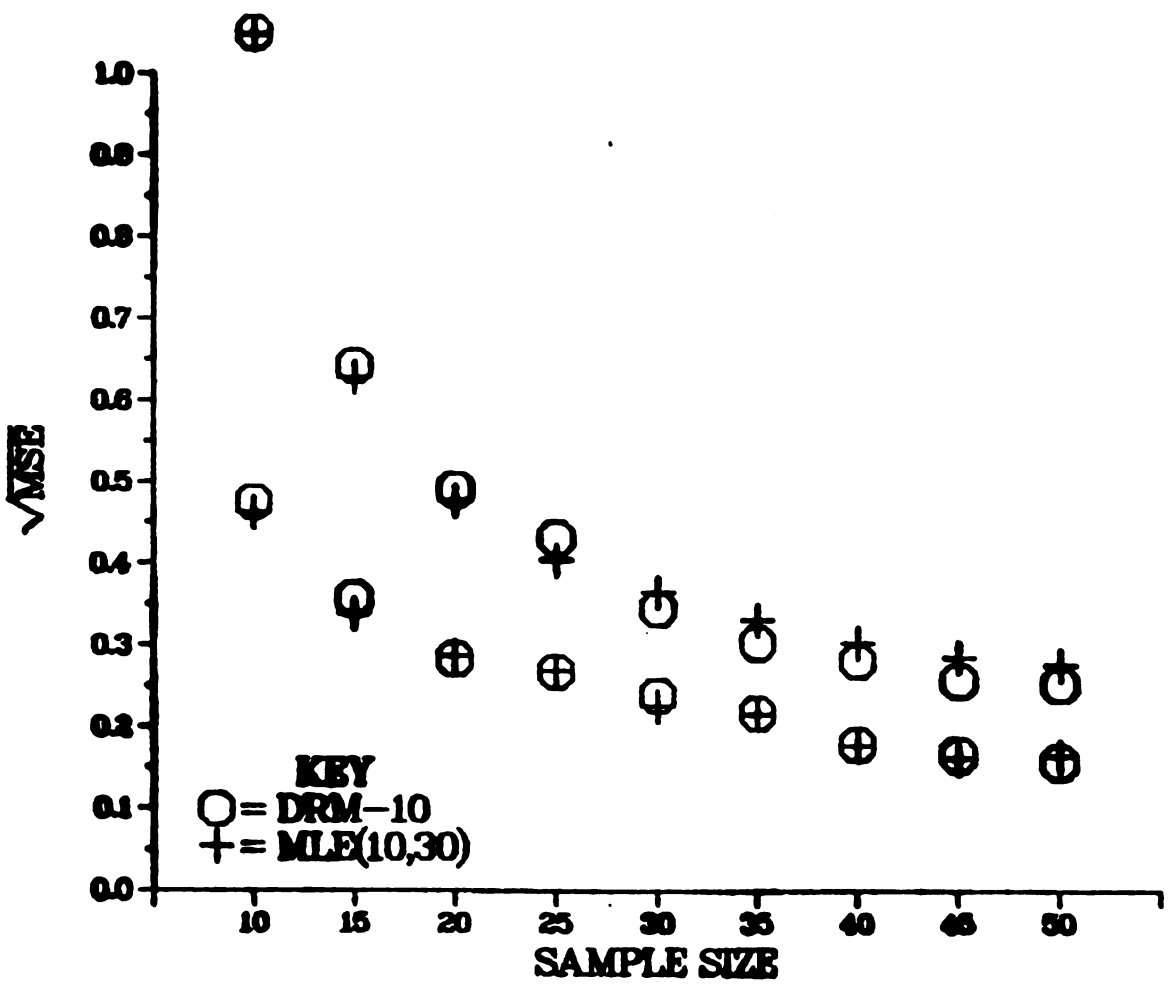


Figure 15. Range of Precision Over Practical Test Conditions - By Sample Size.

magnitude of the constant c were not considered, since we are primarily interested in larger samples where they have no noticeable effect. Another common problem, stimulus noise, was not considered. Bodt and Tingey (1987) show maximum likelihood estimation under a similar normal response to be insensitive to stimulus noise.

Figure 15 clearly indicates the range of precision gained from additional samples. Figure 16 compares sample sizes necessary for estimating the median and third quartile of the response distributions with approximately the same precision. Note the cost effectiveness of using the median if it can serve as a reasonable point for inference in the application at hand.

Normal Response		Cauchy Response	
V_{50}	V_{75}	V_{50}	V_{75}
10	20 to 25	10	45
15	25	15	> 50
20	35		
25	35 to 40		
30	45 to 50		
35	45 to 50		
40	> 50		

Figure 16. Sample Sizes Required for Estimating the Median and Third Quartile with Approximately the Same Precision.

IV. SUMMARY

Under suboptimal conditions for the stochastic approximation method, DRM-c's ability to collect data pertinent to the estimation of the median of the response distribution, was comparable to that of MLE(c,d). This was true over a variety of response distribution shapes and sample sizes. For the estimation of the 3rd quartile they were again comparable when the response distribution was normal. If the response followed a Cauchy distribution, MLE(c,d) was slightly superior to DRM-c. The experimenter is encouraged to take into account these findings when planning a sensitivity test.

LIST OF REFERENCES

Anbar, D. (1978), "A Stochastic Newton-Raphson Method," Journal of Statistical Planning and Inference, Vol. 2, pp 153-163.

Banerjee, K.S. (1980), "On the Efficiency of Sensitivity Experiments Analyzed by the Maximum Likelihood Estimate Procedure Under the Cumulative Normal Response," Ballistic Research Laboratory Technical Report, ARBRL-TR-02269.

Bodt, B.A., Tingey, H.B. (1987), "Design and Estimation in Small Sample Quantal Response Problems - A Monte-Carlo Study," Ballistic Research Laboratory Technical Report, in publication.

Brownlee, K.A., J.L. Hodges, Jr., and Murray Rosenblatt (1953), "The Up-and-Down Method with Small Samples," Journal of the American Statistical Association, Vol. 48, pp 262-277.

Chung, K.L. (1954), "On a Stochastic Approximation Method," Annals of Mathematical Statistics, Vol. 25, pp 463-483.

Cochran, W.G. and Davis, M. (1964), "Stochastic Approximation to the Median Effective Dose in Bioassay," Stochastic Models in Medicine and Biology, J. Garland ed., pp. 281-300, Madison: University of Wisconsin Press.

Cox, D. (1970), Analysis of Binary Data, London: Methoen.

Davis, M. (1971), "Comparison of Sequential Bioassays in Small Samples," Journal of the Royal Statistical Society, Ser. B, Vol 33, pp. 78-87.

DiDonato, A.R. and M.P. Jarnagin, Jr. (1972), "Use of the Maximum Likelihood Method Under Quantal Responses for Estimating the Parameters of a Normal Distribution and its Application to an Armor Penetration Problem, Naval Weapons Laboratory Technical Report, TR-2846.

Dixon, W.J. and Mood, H. M. (1948), "A Method for Obtaining and Analyzing Sensitivity Data," Journal of the American Statistical Association, Vol. 43, pp. 109-126.

Dixon, W.J. (1965), "The Up-and-Down Method for Small Samples," Journal of the American Statistical Association, Vol. 60, pp. 967-978.

Golub, A. and Grubbs, F.E. (1956), "Analysis of Sensitivity Experiments when the Levels of Stimulus Cannot be Controlled," Annals of Mathematical Statistics, Vol. 57, pp. 257-265.

Hampton, L.D. (1967), "Monte Carlo Investigations of Small Sample Bruceton Tests," Naval Ordnance Laboratory Technical Report, NOLTR-66-117.

Hodges, J.L. and Lehmann, E.L. (1955), "Two Approximations to the Robbins-Monro Process," Proceedings 3rd Berkely Symposium, Vol. 1, pp. 95-104.

Langlie, H.J. (1962), "A Reliability Test Method for 'One-Shot' Items," Aeronutronic Publication No. U-1792.

McKaig, A.E. and Thomas, J. (1983), "Maximum Likelihood Program for Sequential Testing Documentation," Ballistic Research Laboratory Technical Report, ARBRL-TR-02481.

Robbins, H and Monro, S. (1951), "A Stochastic Approximation Method," Annals of Mathematical Statistics, Vol. 22, pp. 400-407.

Rothman, D., Alexander, M.J., Zimmerman, J.M. (1965), "The Design and Analysis of Sensitivity Experiments," NASA CR-62026, Vol. 1.

Rubinstein, R.Y. (1981), Simulation and the Monte Carlo Method, New York, John Wiley & Sons Inc..

Silvapulle, M.J. (1981), "On the Existence of Maximum Likelihood Estimators for the Binomial Response Model," Journal of the Royal Statistical Society, Vol. 43, pp. 310-313.

Wetherill, G.B. (1963), "Sequential Estimation of Quantal Response Curves," Journal of the Royal Statistical Society, Vol. 25 (1963), pp 1-48.

Wu, C.F. Jeff (1985), "Efficient Sequential Designs with Binary Data," Journal of the American Statistical Association, Vol. 8, pp 974-984.

TESTS FOR CONSISTENCY OF A CLASS OF VULNERABILITY MODELS

DAVID W. WEBB

Experimental Design and Analysis Branch
System Engineering and Concepts Analysis Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

The author wishes to thank Dr. J. Richard Moore for his helpful insight and comments.

TESTS FOR CONSISTENCY OF VULNERABILITY MODELS

This paper studies the problem of confirming a set of estimated probabilities of kill for a small number of independent, but not identically distributed, Bernoulli outcomes. The problem originates from vulnerability studies on tanks in which kill probabilities of individual components are desired. The cost of resources makes it unfeasible to obtain these probability estimates by repeated field testing. Therefore, computer simulation is used to get the desired estimates. Researchers then want to test the accuracy of these computer generated values. Again, the economics of live firing sometimes allows for the firing of only one round at the tank. The question becomes "Can the kill probabilities obtained through simulation be confirmed by the results of a single round field test?"

Denote the simulation estimates by the vector $[p_1^o, p_2^o, \ldots, p_k^o]$ where p_i^o is the probability of kill for the i^{th} tank component of interest. If we assume that the components are independent, we may rewrite the above question in the form of a hypothesis test.

$$H_o: p_1 = p_1^o, p_2 = p_2^o, \ldots, p_k = p_k^o, \quad \text{vs.}$$

$$H_a: p_i \neq p_i^o, \quad \text{for some } i.$$

Note that when $p_1^o = p_2^o = \ldots = p_k^o = p^o$, this is the k-trial binomial case, $B(k,p^o)$, and the null hypothesis is simply $H_o: p = p^o$. We seek a test for the more generalized case of unequal p_i's. As may be expected, the small size of k will present problems with power. Also, it should be pointed out that the alternative hypotheses only says that at least one inequality exists. In practice though, it will usually take several gross differences between the hypothesized and actual vectors for any test to reject H_o. Therefore the tests to be explored will not be able to validate the hypothesized probabilities; they will be able to check for consistency between the simulation estimates and field test results as a whole.

Suppose we observe a set of k independent 0 or 1 outcomes (representing survive or kill), denoted by the row vector $A = [a_1, a_2, \ldots, a_k]$. For example, if $k = 5$, we may have $A = [0, 1, 0, 0, 1]$. There are 2^k possible outcome vectors. The probability of observing outcome vector A under the null hypothesis is given by the density function

$$P(A) = {p_1^o}^{a_1} \cdot (1 - p_1^o)^{(1-a_1)} \cdot {p_2^o}^{a_2} (1 - p_2^o)^{(1-a_2)} \cdot \ldots \cdot {p_k^o}^{a_k} \cdot (1 - p_k^o)^{(1-a_k)}$$

$$= \prod_{i=1}^{k} {p_i^o}^{a_i} \cdot (1 - p_i^o)^{(1-a_i)}.$$

A test of the null hypothesis needs some way of ordering the 2^k possible outcome vectors. We will examine three test procedures characterized by their ordering schemes.

Test One

This test rejects the null hypothesis if the observed vector is among some predefined critical set of "rarest" outcomes. The outcome set is ordered by the density function in increasing magnitude, and each outcome is numbered so that A_1 is the least likely to occur and A_{2^k} is the most likely. Define a "cumulative function" B, whereby

$$B_i = \begin{cases} P(A_1) & i = 1 \\ B_{i-1} + P(A_i) & i = 2, 3, \ldots, 2^k . \end{cases}$$

Choosing a "c" such that

$$c = \max \{ j \mid B_j < \alpha \text{ and } P(A_j) \neq P(A_{j+1}) \},$$

then the set

$$A_{RR} = \{A_1, A_2, \ldots, A_c\}$$

represents the c rarest outcomes and defines the rejection region for test of H_o with a $(\alpha)100\%$ level of significance. The "test statistic" is the observed vector A; if it is in A_{RR}, then H_o is rejected.

Test Two

This test is based upon the number of kills observed. The underlying notion is that under the hypothesized model, a certain number of kills is expected. Letting K(A) be the number of observed kills, then the expected value of K(A) under the null hypothesis is

$$\begin{aligned} E[K(A)] &= p_1^o + p_2^o + \ldots + p_k^o \\ &= \sum_{i=1}^{k} p_i^o . \end{aligned}$$

If the observed K(A) is much smaller than this value, then perhaps the simulation overestimated the kill probabilities and H_o should be rejected. On the other hand, if the observed K(A) is much larger, then H_o should be rejected since the kill probabilities may be underestimated.

To perform this test, we begin by calculating P(A) and K(A) for all 2^k outcomes. The outcomes are then ordered by increasing magnitude by the number of kills and numbered so that

$$K(A_1) \leq K(A_2) \leq \ldots \leq K(A_{2^k}).$$

(The order among outcomes with the same K(A) is irrelevant.) Similarly to Test One, the "cumulative function" is calculated. Since rejecting H_o may be the result of too large or too small a value of K(A), a two-tailed test is used. Critical values c_1 and c_2 are selected so that the actual alpha level

$$P[K(A) \leq c_1] + P[K(A) \geq c_2] \qquad (*)$$

is maximized but still less than or equal to α. The rejection region for Test Two is

$$K(A_{RR}) = \{0, 1, \ldots, c_1\} \cup \{c_2, c_2 + 1, \ldots, k\}.$$

The simulation generated estimates will be rejected as inconsistent with the field test if $K(A) \in K(A_{RR})$.

Test Three

This test examines the number of "correct responses", where a correct response is defined as:

$$\gamma_i = \begin{cases} 1, & \text{if } a_i = 0 \text{ when } p_i^\circ < .5, \text{ or } a_i = 1 \text{ when } p_i^\circ > .5 \\ .5, & \text{if } p_i^\circ = .5 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, a correct response of 1 is given if the more likely outcome (kill or survive) is observed. On the other hand a correct response of 0 means that the less likely event occurred. As somewhat of a tiebreaking policy, if the hypothesized probability is .5, then the correct response value is .5, no matter which outcome is actually observed.

The test statistic is $C(A) = \gamma_1 + \gamma_2 + \cdots + \gamma_k$

$$= \sum_{i=1}^{k} \gamma_i .$$

The expected value of the test statistic is

$$E[C(A)] = C_L + k^*/2 + C_U,$$

where

$$C_L = \sum_j (1 - p_j^\circ) \qquad \text{for all } p_j^\circ < .5$$

$$C_U = \sum_j p_j^\circ \qquad \text{for all } p_j^\circ > .5$$

$$k^* = \text{"number of } p_j^\circ \text{ equal to .5"} .$$

The test procedure begins by calculating P(A) and C(A) for all possible outcomes. The outcomes are arranged by increasing magnitude by the number of correct responses without regard for ties so that

---

*

$$P[K(A) \leq c_1] = B_i, \quad \text{for } i = \max \{j \mid K(A_j) < K(A_{j+1}) \text{ and } K(A_j) = c_1\}$$
$$P[K(A) \geq c_2] = 1 - B_i, \text{ for } i = \min \{j \mid K(A_j) > K(A_{j-1}) \text{ and } K(A_j) = c_2\}$$

$$C(A_1) \leq C(A_2) \leq \ldots \leq C(A_{2^k}).$$

The cumulative density is computed as usual. Observing a value of C(A) much smaller *or* larger than the expected value leads us to believe that H_o is false. Therefore, a two-tailed test is desired, and the critical values c_1 and c_2 are chosen to maximize

$$P[C(A) \leq c_1] + P[C(A) \geq c_2] \leq \alpha,$$

the actual alpha level, where

$$P[C(A) \leq c_1] = B_i, \qquad \text{for } i = \max\{j \mid C(A_j) < C(A_{j+1}) \text{ and } C(A_j) = c_1\}$$

$$P[C(A) \geq c_2] = 1 - B_i, \qquad \text{for } i = \min\{j \mid C(A_j) > C(A_{j-1}) \text{ and } C(A_j) = c_2\}.$$

Since the rejection region is $C_{RR} = \{0, 1, \ldots, c_1\} \cup \{c_2, c_2 + 1, \ldots, k\}$, we will reject H_o at the α level of significance if $C(A) \in C_{RR}$.

Properties of the Tests.

To study the three test procedures, 2000 pairs of k-dimensional probability vectors were randomly generated for k=6,...,10. The first vector of a pair (P_o, P_a) was considered the hypothesized probability vector, and the second was the alternative probability vector. The level of significance was set at $\alpha = .05$. The power of the three tests was computed for each pair (P_o, P_a).

Figures 1 through 3 show a graphical way of comparing power between any two tests, A and B. Each point represents a pair (P_o, P_a). Its coordinates (x,y) are the power of Tests A and B, respectively. If Test A is more powerful than Test B, then we expect to see a graph similar to Figure 1. If the opposite is true, the graph will be similar to Figure 2. But if both have approximately the same power then Figure 3 is the proper scatterplot.

Comparison of the three tests based upon the 2000 randomly generated vectors is shown in Figures 4-8. Several observations can be made from these graphs.

1. For most pairs of vectors, and for k=6,7,8,9,10, Test 1 has greater power than either of the other two tests.

2. Median power increases with k for Tests 1 and 3 (see Figure 9).

3. Median power remains fairly constant for all k with Test 2. The median power of Test 2 is not much greater than the alpha level. This indicates how poor a procedure the test is.

When comparing the power of all three tests for each point, it was occasionally found that the superior test was either Test 2 or Test 3. For example,

$\vec{H}_o$ = [.71 .23 .10 .09 .15 .67 .50 .93]

$\vec{H}_a$ = [.80 .36 .47 .34 .36 .27 .94 .95]

TEST	1	2	3
Rej. region	187 least likely	{0,1,7,8}	{.5,1.5,2.5,3.5}
Exact alpha	.0499	.0495	.0495
Power	.3631	.7728	.7728

This leads us to ask what can be determined from (P_o, P_a) about the power of the tests, if anything? One possible relationship studied was the power versus the distance between P_o and P_a in k-space, i.e. Δ (P_o, P_a) where

$$\Delta (P_o, P_a) = \sqrt{\sum_{i=1}^{n} (P_i^o - P_i^a)^2}$$

$$= \sqrt{(P_1^o - P_1^a)^2 + \ldots + (P_k^o - P_k^a)^2}$$

Figures 10-14 show scatterplots of this relationship for each test and sample size. The correlations between power and Δ (P_o, P_a) are shown in Figure 15.

The problem with looking at the relationship between P_o and P_a, however, is that in practice we do not know what P_a is. It will not be very helpful to know that the choice of best test for a given P_o, is dependent upon the choice of P_a. We should look for a best test given P_o only. This is the topic of ongoing research.

SUMMARY

The problem is most complicated by the fact that we must judge the entire set of computer generated estimates on a single fired shot. While we admit that Test 1 was not able to detect some greatly differing alternative set of probabilities, it was in general the best of the three test procedures. The reasons become obvious when we closely examine the other two.

Test 2 does not take into consideration the order in which the a_i's appear. For example, let our hypothesized set of probabilities be P_o = [.01, .02, .03, .97, .98, .99]. For the observed outcome vectors A_1 = [0, 0, 0, 1, 1, 1] and A_2 = [1, 1, 1, 0, 0, 0], we compute $P(A_1)$ = .8857 and $P(A_2)$ = .000000000036. However for both outcomes we compute $K(A_1) = K(A_2) = 3$, the expected value of the test statistic under H_o. Therefore we would not reject H_o in either case. Not only does Test 2 not reject H_o given A_2 (when it obviously should), but it has managed to equate the most likely and least likely outcomes.

Test 3 does consider the order of the observed outcomes, however it does not incorporate the magnitude of the p_i^o's. To see how this is dangerous, let P_{o1} = [.53, .52, .51, .49, .48, .47], P_{o2} = [.47, .48, .49, .51, .52, .53], and A = [1, 1, 1, 0, 0, 0]. Under H_{o1}, P(A)=.019756 and C(A)=6, while under H_{o2}, P(A)=.012220 and C(A)=0. The test has exaggerated the difference between the two probability vectors, despite their being nearly equal.

Test 1 is the best of the three candidate procedures because it simply tries to create the largest possible rejection region. Imagine trying to fill a fishbowl with as many marbles as possible when the marbles are different sizes. Since we do not want to take up space with larger marbles, we fill the fishbowl one marble at a time starting with the smallest, then the second smallest, and so on until the bowl is full. In a similar fashion, this is how the rejection region for Test 1 is formed, thus resulting in a most powerful test.

Further research into this problem will look at other possible tests and easier implementation of Test 1.

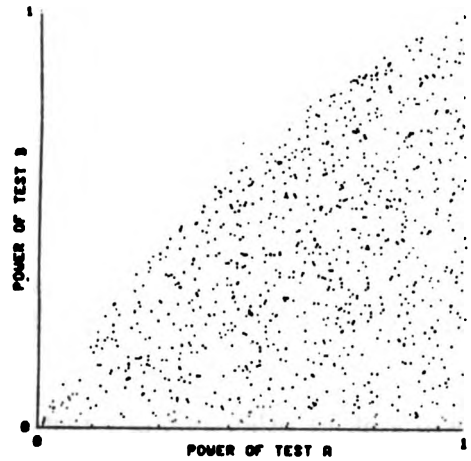


Figure 1.

Test A More Powerful Than Test B.

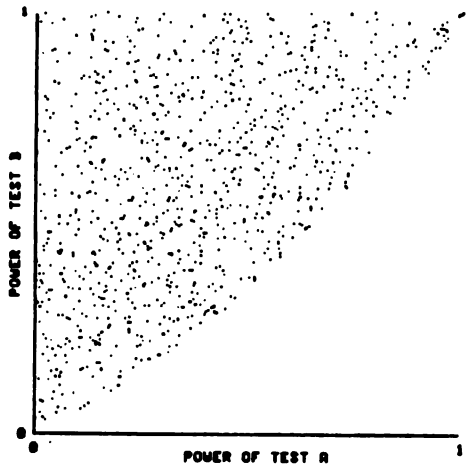


Figure 2.

Test B More Powerful Than Test A.

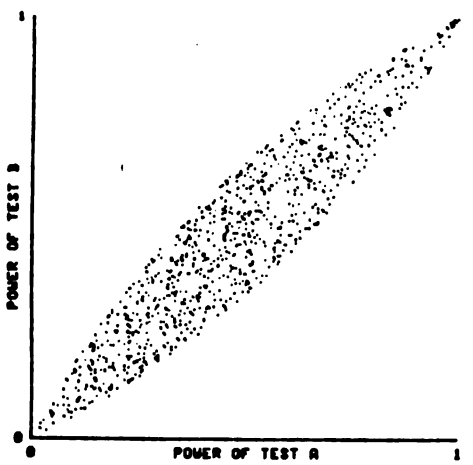
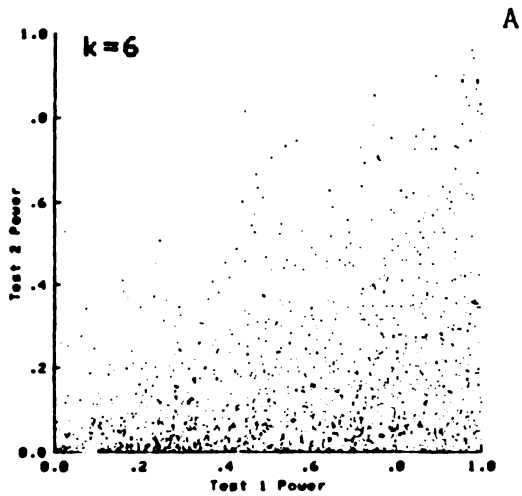


Figure 3.

Test A and Test B With Similar Power.



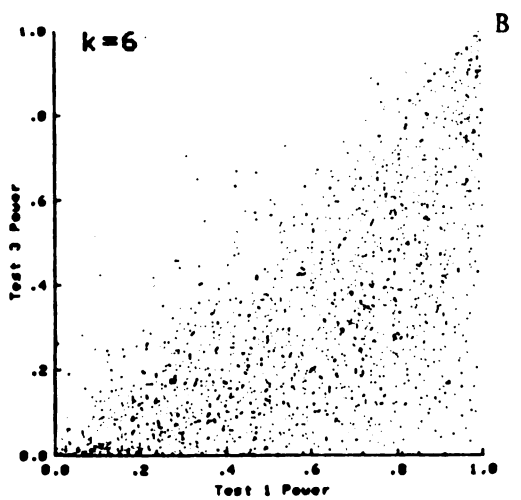
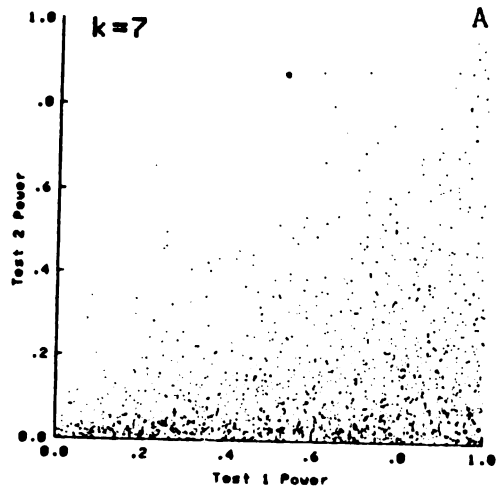


Figure 4.

Test 1 vs. Tests 2 and 3 (K=6).



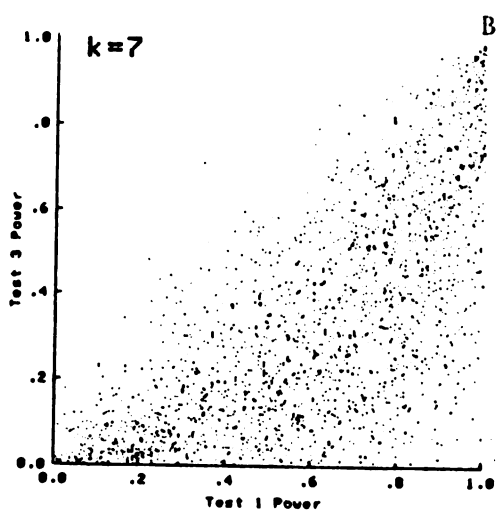


Figure 5.

Test 1 vs. Tests 2 and 3 (K=7).

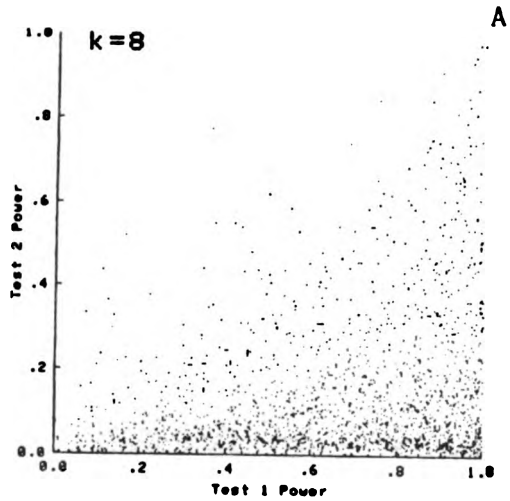


Figure 6.

Test 1 vs. Tests 2 and 3 (K=8).

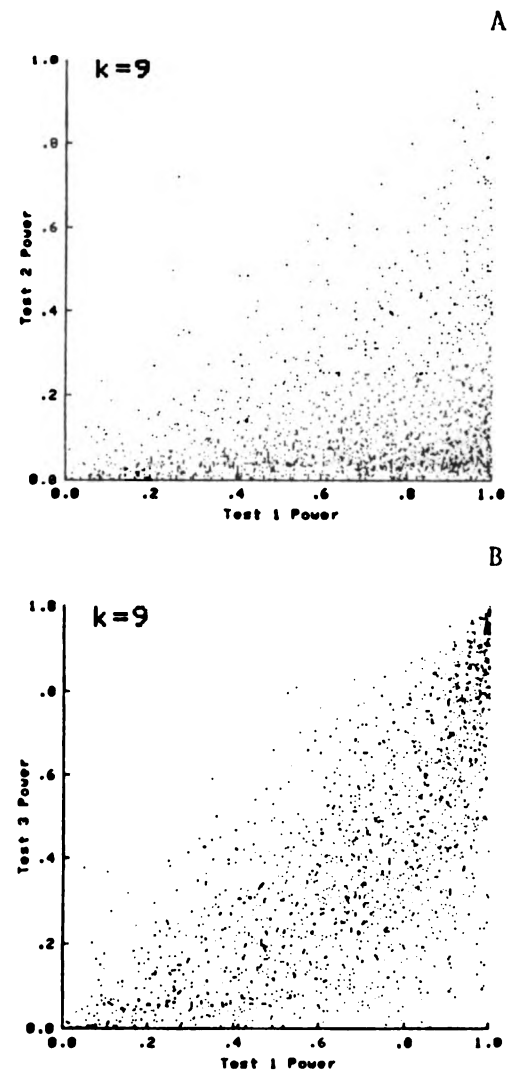


Figure 7.

Test 1 vs. Tests 2 and 3 (K=9).

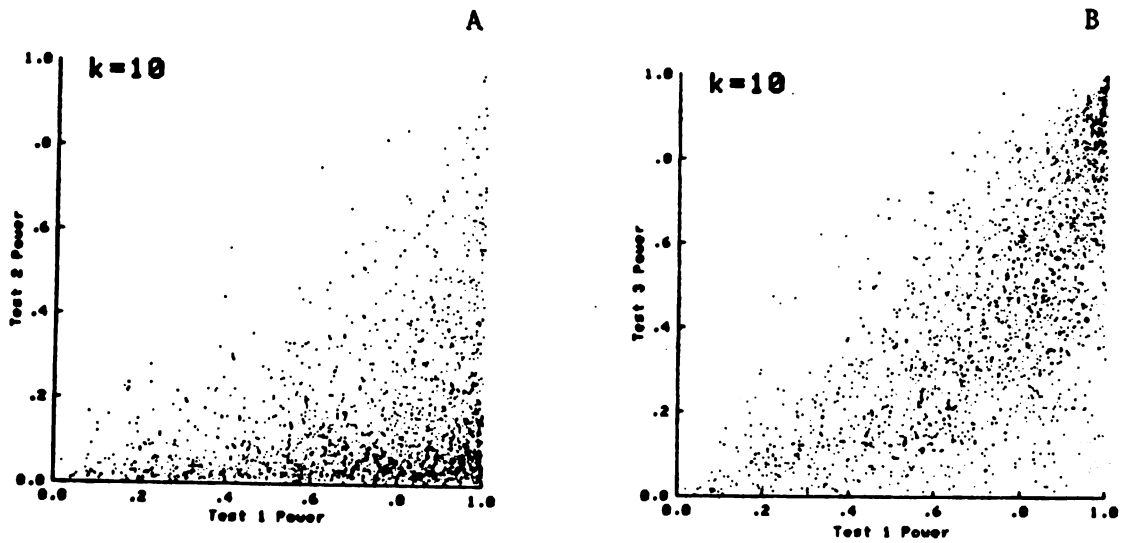


Figure 8. Test 1 vs. Tests 2 and 3 (K=10).

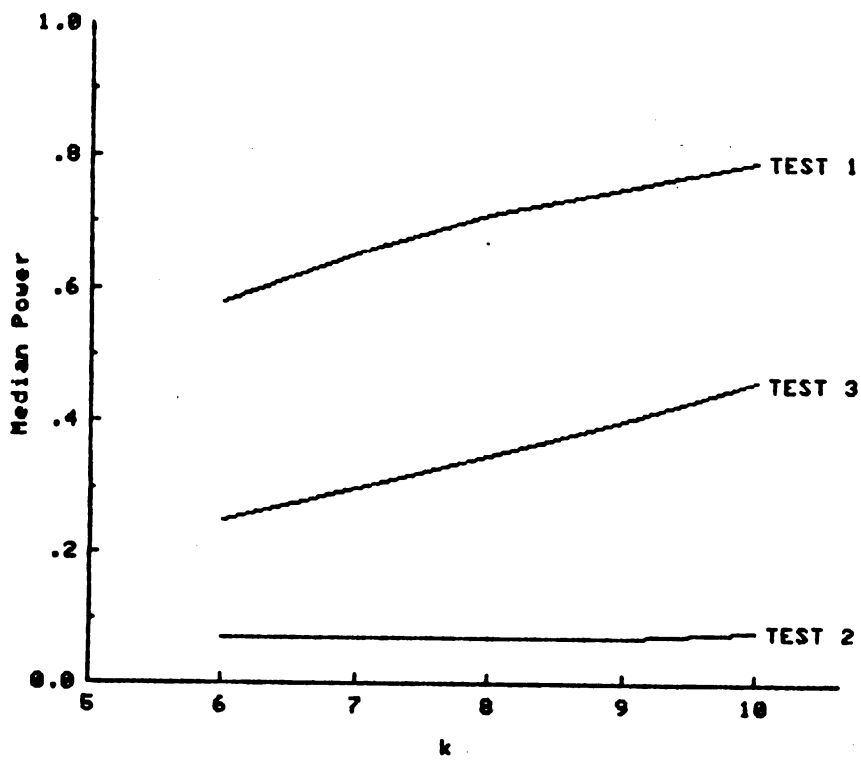
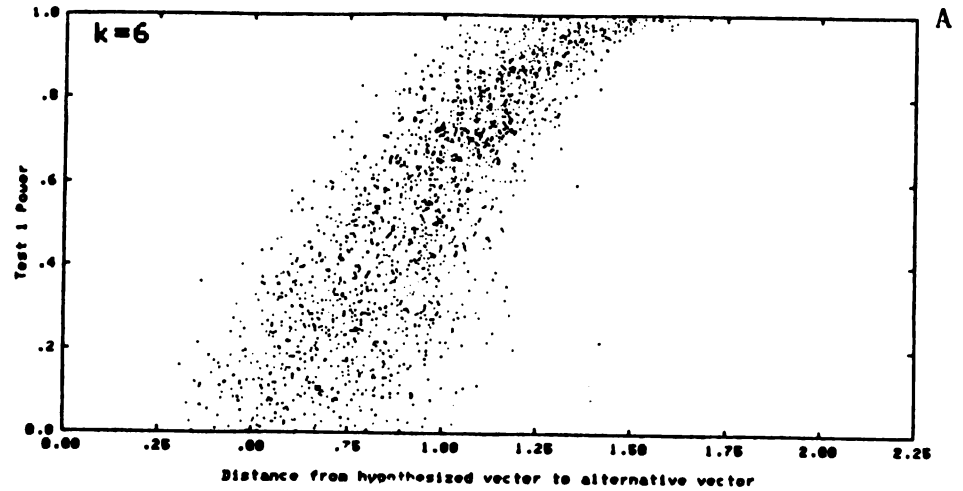
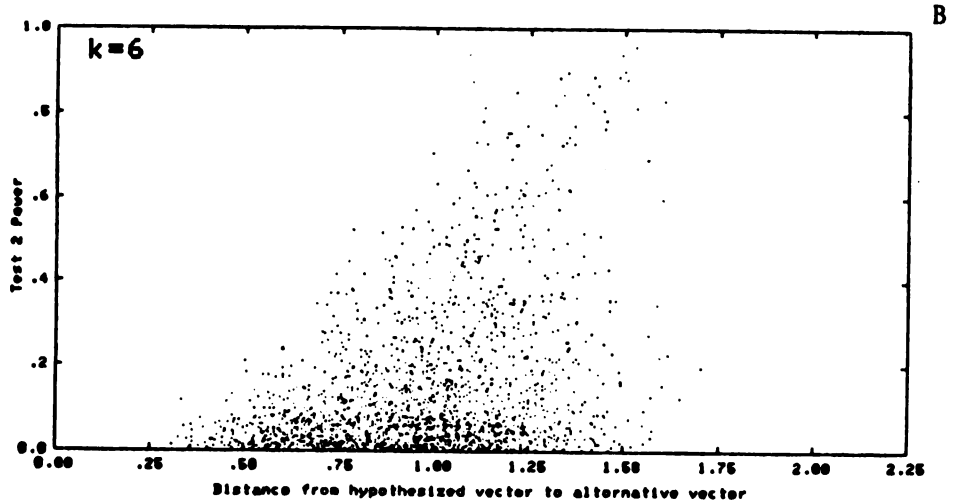


Figure 9. Power vs. Sample Size for the Three Tests.





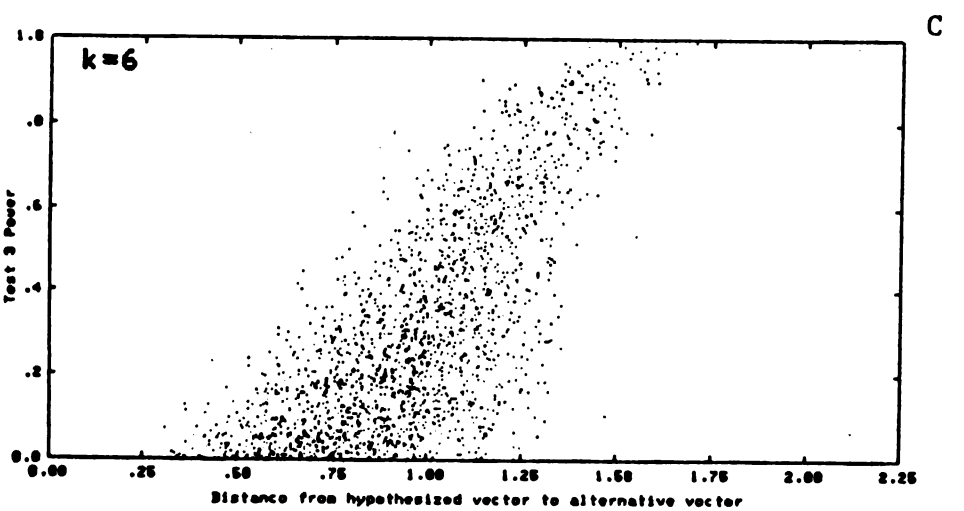


Figure 10. Power vs. Δ (P_o, P_a) (K=6).

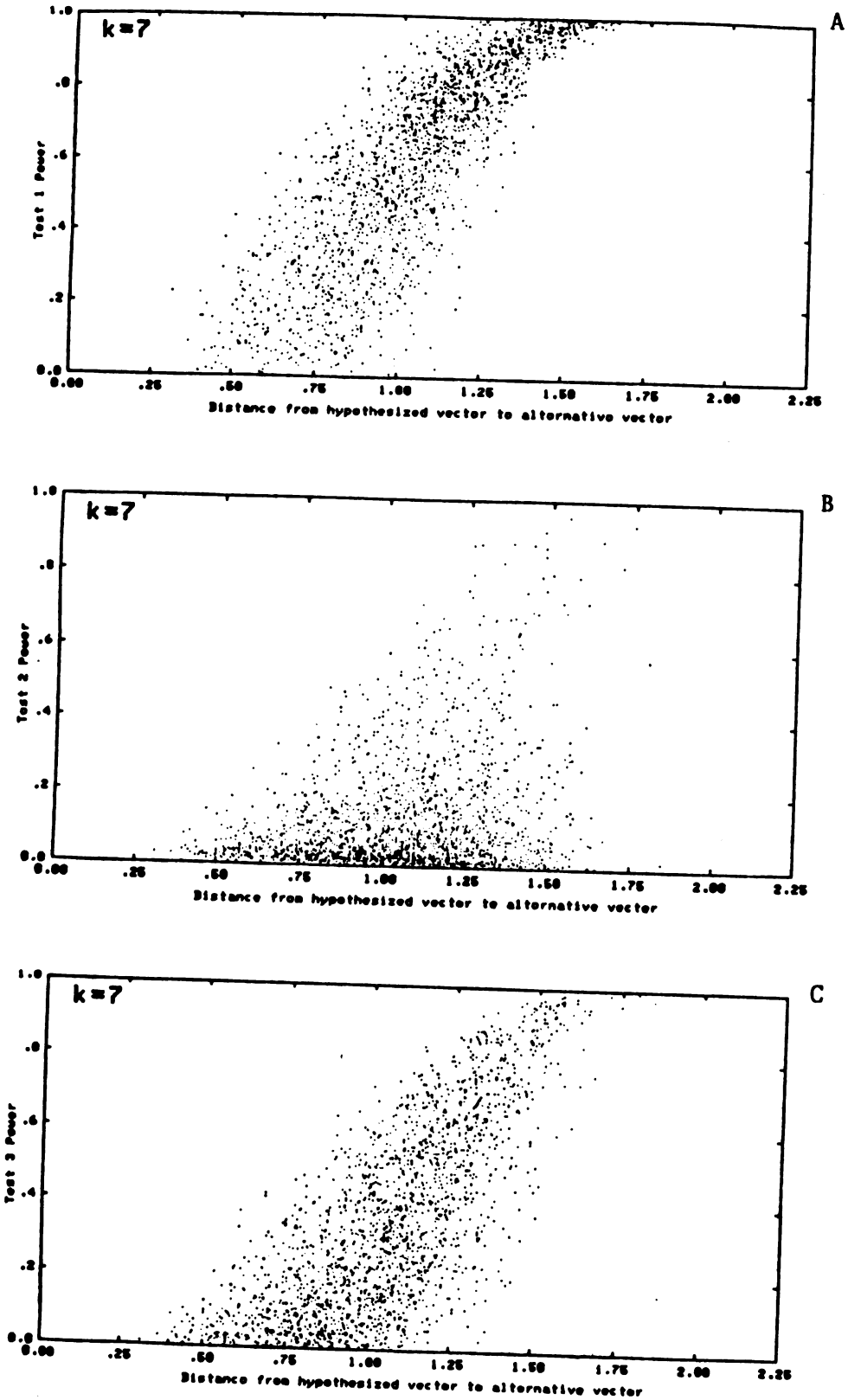
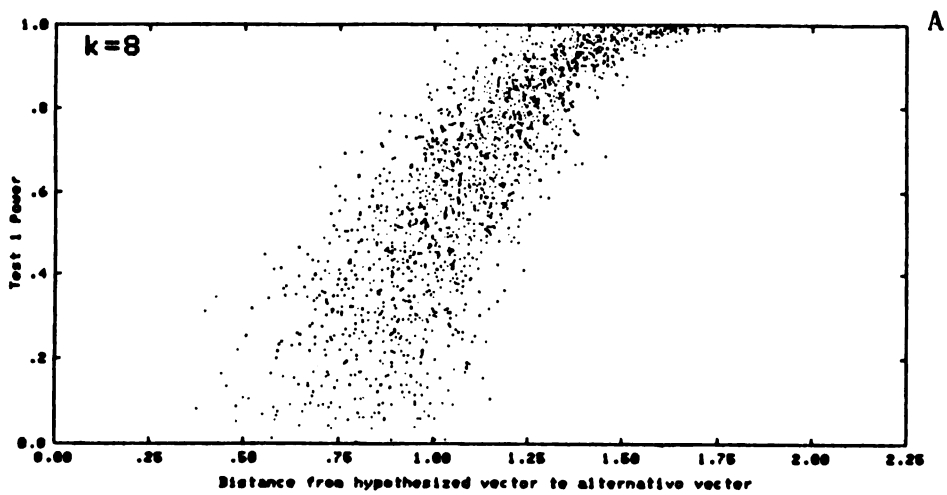
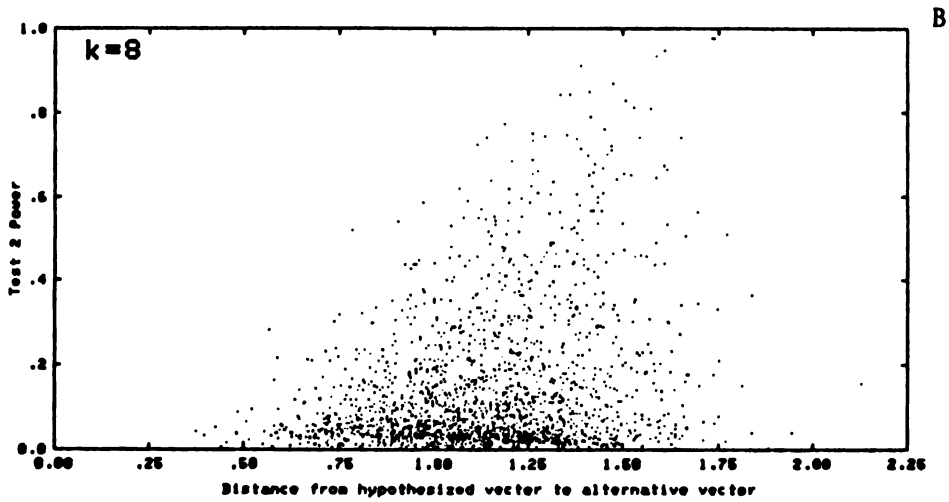


Figure 11. Power vs. Δ (P_o, P_a) (K=7).





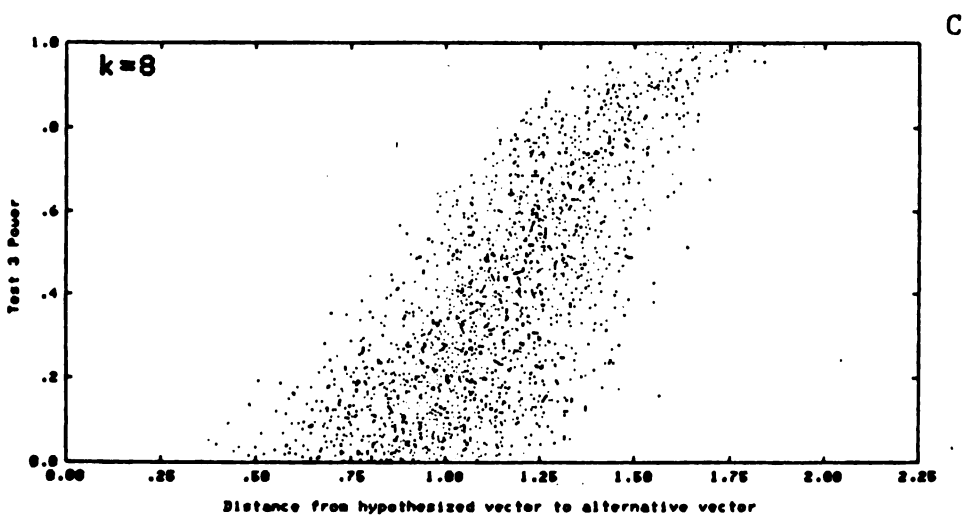


Figure 12. Power vs. Δ (P_o, P_a) (K=8).

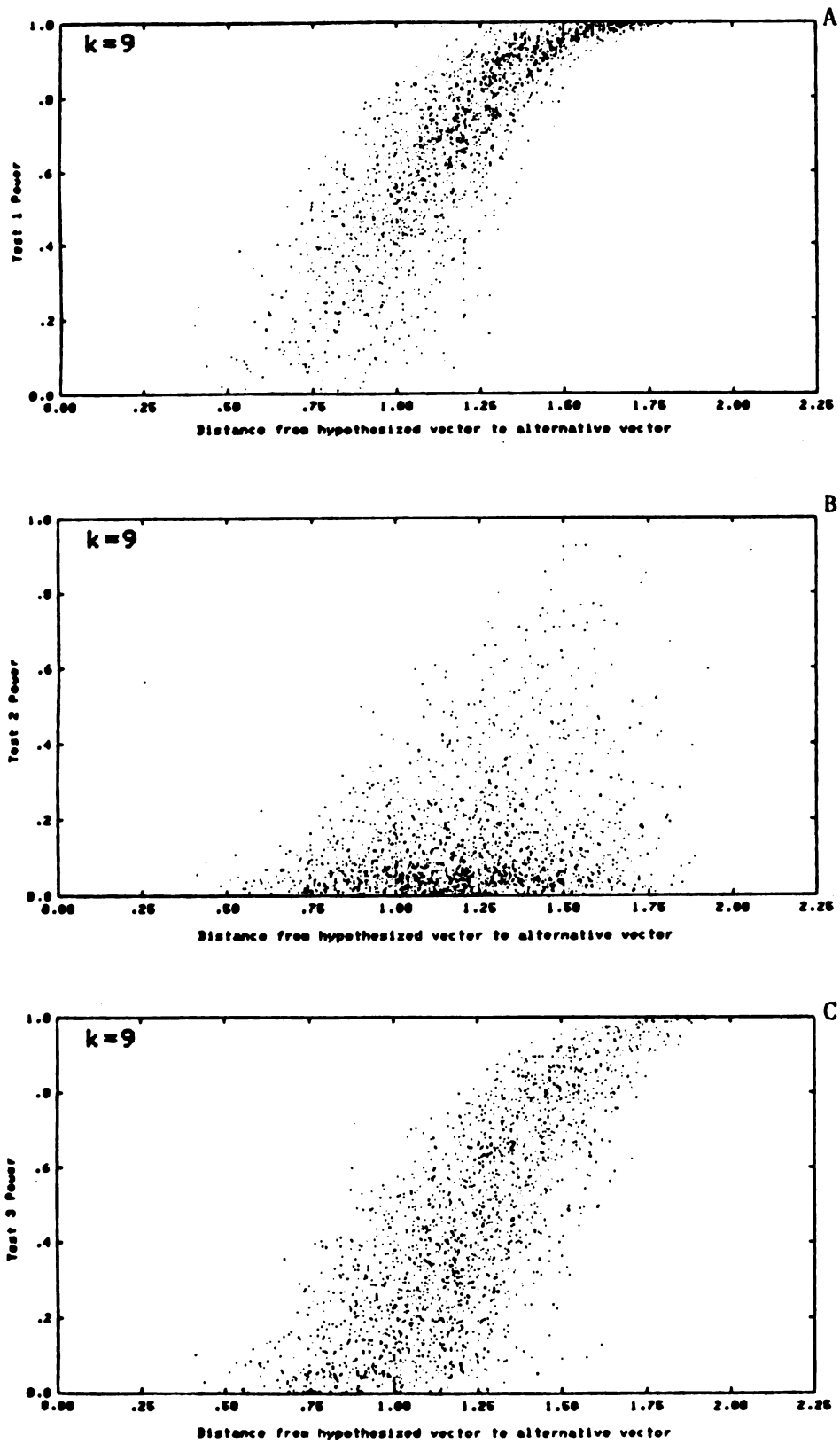
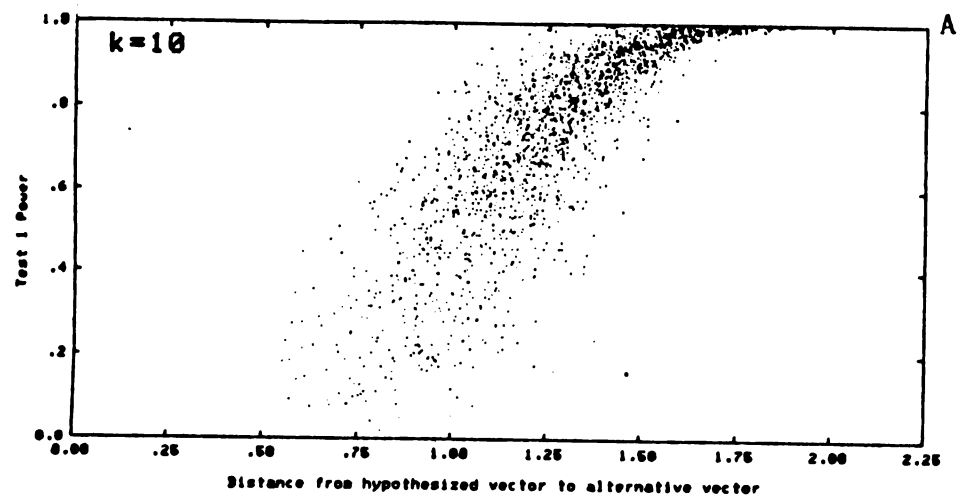
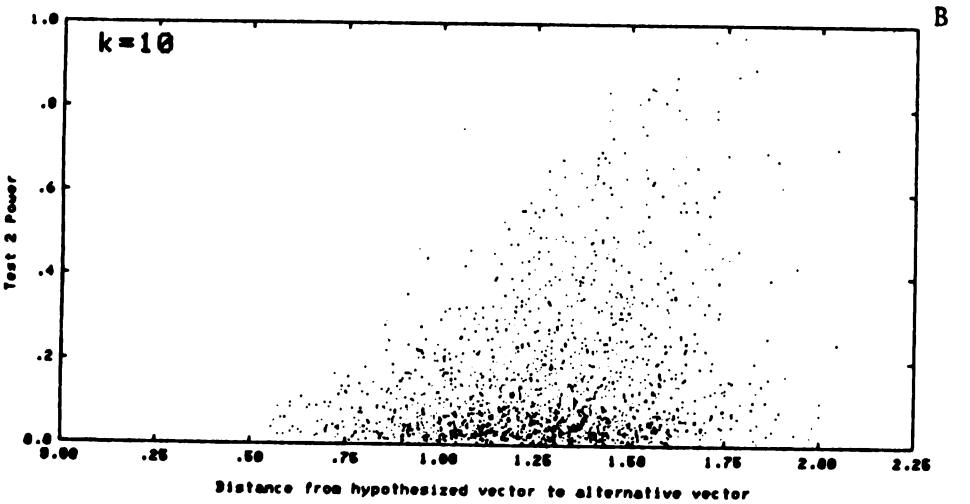


Figure 13. Power vs. Δ (P_o, P_a) (K=9).





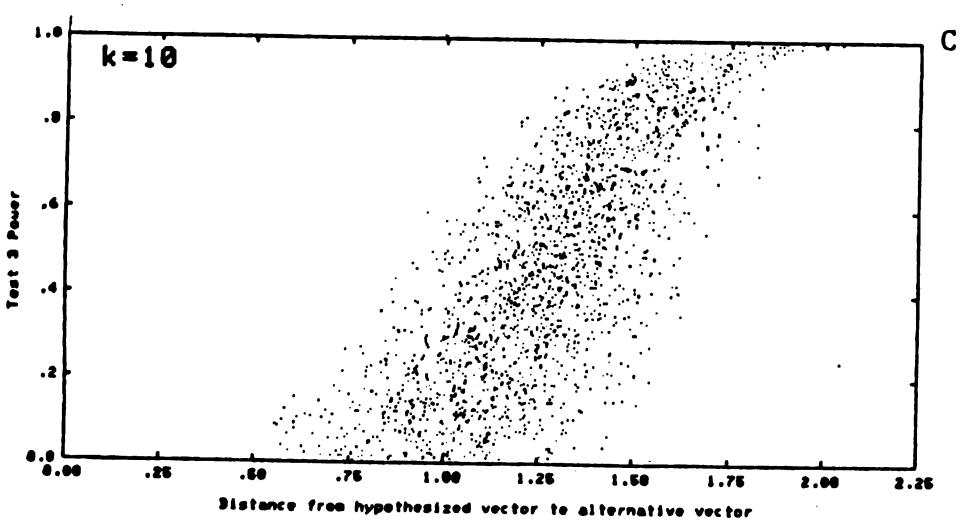


Figure 14. Power vs. Δ (P_o, P_a) (K=10).

Test	k				
	6	7	8	9	10
1	.824	.835	.829	.837	.825
2	.345	.301	.264	.284	.290
3	.736	.748	.751	.789	.780

Figure 15. Correlations Between Power and $\Delta (P_o, P_a)$.

NONPARAMETRIC SMALL SAMPLE TOLERANCE LIMITS

Donald M. Neal, Mark G. Vangel, and John Reardon

Materials Technology Laboratory
Watertown, MA 02172-0001

ABSTRACT

Results from this clinical study have identified the Hanson-Koopmans[1] method as the most desirable nonparametric small sample lower tolerance limit estimator in the range where conventional nonparametric procedures are not defined. The Monte Carlo studies indicated that this method worked well for sample sizes from 2 to 28. The authors' initial effort using a linear function of the first four order statistics was reasonably effective for sample sizes greater than 15.

Other efforts to obtain a solution to the problem include extension of the quantile sign test, a scheme involving a reduction factor for the first order value, and a smooth nonparametric quantile estimator. These methods were not satisfactory due to either instability of first ordered value when sample sizes are small, or the inability to provide proper coverage rate for N < 28.

INTRODUCTION

The inability to obtain exactly the same structural properties from all specimens obtained from a manufactured material results in a relatively large variability in strength measurements when a large number of specimens are considered. In the case of designing an aircraft structure, it is required to design such that a maximum stress value exists in critical locations, and these values do not exceed the minimum guaranteed material properties (strength). Obtaining minimum strength values will reduce the possibility of some production components containing weaker material than that from the laboratory

test element. This guaranteed minimum strength value is defined as the design allowable (basis value) by aircraft design engineers.

Usually, the measured value is considered acceptable in estimating the population parameters for predicting population percentiles. In the case of the design engineer, it is advisable to have a prediction which will determine the accuracy of the percentile estimate at a high degree of statistical confidence. This is the correct interpretation of a basis value. For example, certain military standards, e.g., MIL-HDBK-5[2] require material property data to be presented on an A or B allowable basis. The allowables represent a value determined from a specified probability of survival with a 95 percent confidence in the assertion. The survival probabilities are .99 for the A allowable, and .90 for the B allowable.

MTL is involved in the development of the statistics chapter for the MIL-17 Handbook[3] on composite material in aircraft structural design. The chapter will include methods for determining the design allowable values. The inability to identify the statistical model from limited or multi-modal data motivated the authors to find a non-parametric model which will provide a correct tolerance bound (P=.95) on the quantile values (P=.10). The conventional nonparametric method using the quantile sign test[4] provides a solution if there are at least 28 values in the sample. Unfortunately, the model needed is one for sample sizes less than 28.

This paper presents the results of a clinical paper presented at the ARO sponsored Thirty-Sixth Annual Design of Experiments Conference on methods for obtaining an accurate measure of the above mentioned design allowables involving small sample nonparametric modeling. It should be noted that there are difficulties in extreme quantile modeling techniques involving determination of tolerance bounds for the quantile values in the allowable computation. Brieman, Stone and Gins[5] have discussed the difficulties existing in model identification when very small tail probabilities are required. This is the result of parameter estimates that usually are obtained from data in the central portion of the distribution, where most failures occur, leaving the

tail region limited in representation. This is unfortunate, since the relatively small amount of data in the tail region is of prime importance to the allowable computation. The nonparametric scheme can model the lower ordered values of the distribution. The Hanson-Koopmans[1] model is recommended as a solution to the nonparametric small sample tolerance limit when considering the various alternative solutions, the application of the method does not result in overly conservative estimates of the allowable values. Other methods were attempted, including an extension of the quantile sign test, linear function of first four order statistic (authors' proposed method), a smooth nonparametric quantile estimator[6], and an adaptive scheme involving simulation procedures for obtaining ratio of the first order value to the allowable value. None of the above methods were acceptable for the sample size requirements of $2 \leq n \leq 28$, due to either computational problems or inability to provide minimum coverage of 95%.

QUANTILE ESTIMATE - SAMPLE SIZE

The importance of determining a tolerance limit on the quantile values is graphically displayed in Figures 1a and 1b. The standard normal distribution function is plotted for sample sizes of 50 and 10, using 25 sets of data. In figure 1a, N = 50 the amount of spread in quantile for the 10 percentile values in .80. Figure 1b shows a spread of 2.4 for the same percentile. This example shows the importance of having large sample sizes, or otherwise providing a tolerance limit on the quantile estimate.

Often in structural design, a criteria requires material property values to be larger than the design stress in order to define the margin of safety. Determining a property value from 10 material strength tests in order to obtain 90% reliability estimates could result in nonconservative values and possible structural failure. Obtaining a lower 95% confidence bound on the reliability estimate can provide the necessary assurance.

DEFINITION OF THE B-BASIS VALUE

The B-basis value is a random variable where an observed basis value (design allowable) from a sample will be less than the 10 percentile of the population with a probability of .95. In figures 2a and 2b, a graphical display is shown for the basis value probability density function (N(0,1)) for sample sizes of n = 10 and 50. The dotted vertical lines represent the location for the 10 percentile ($X_{.10}$) of the population and the probability (basis value < $X_{.10}$) = 95 for the basis value probability density function. The graphical display of the basis value density functions show much less dispersion for n = 50 than for n = 10. Small sample sizes will result in more conservative estimates of the basis values.

QUANTILE SIGN TEST (Conventional Nonparametric Analysis)

The quantile sign test[4] is introduced in the text as a procedure that provides an accurate B-basis value for n > 28. The authors initially attempted to extend this method for n $\leq$ 28 using various procedures related to the first ordered value without success.

The analysis involves considering, for example, $q_{.10}$ as a quantile of a distribution, then the values < $q_{.10}$ are binomial random variables with n trials and probability of .10. If $X_{(r)}$ is the rth ordered value in the sample, the B-basis value is equal to $X_{(r)}$ where r $\geq$ 1 is the largest integer solution to

$$\sum_{w=r}^{n} \binom{n}{w} (.10)^{w} (.90)^{n-w} \geq .95 \qquad (1)$$

where $\binom{n}{w} = n!/w!(n-w)!$

and n = sample size

See Table I for computing values for r given n.

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

MEAN = 0.0

STANDARD DEVIATION = 1.0

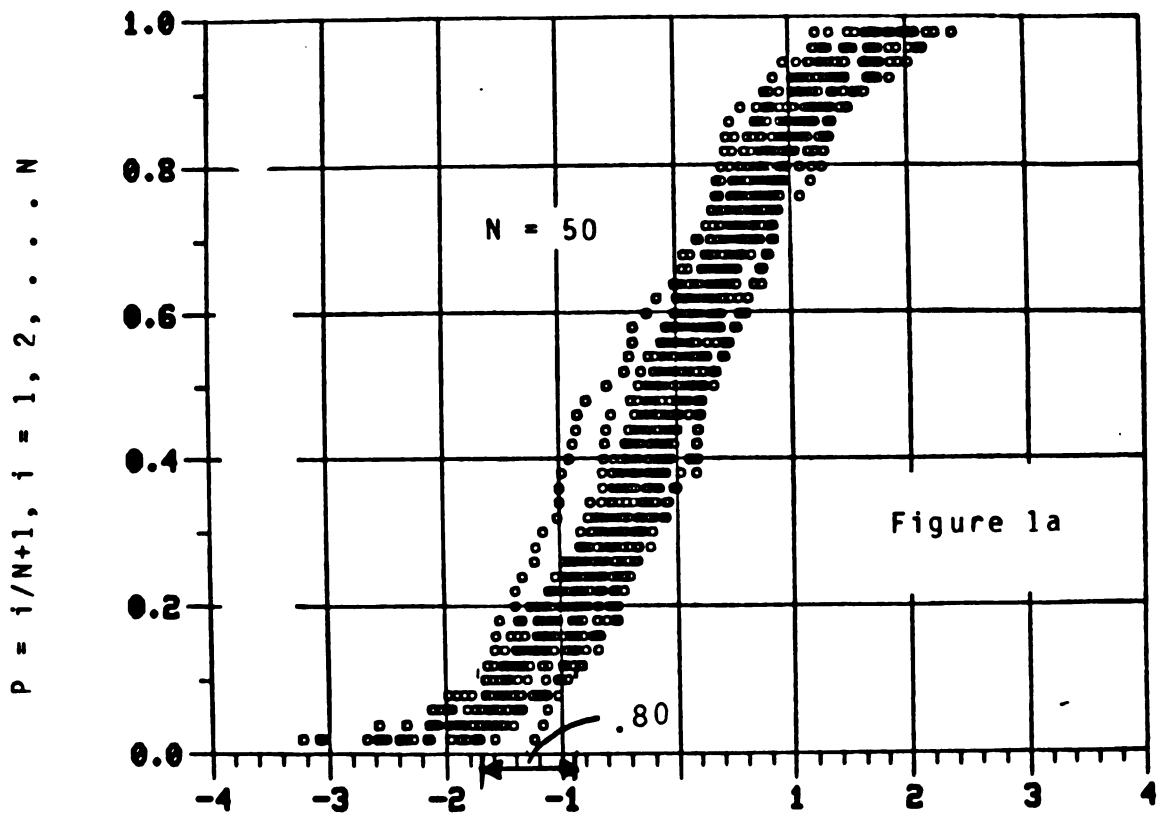


Figure 1a

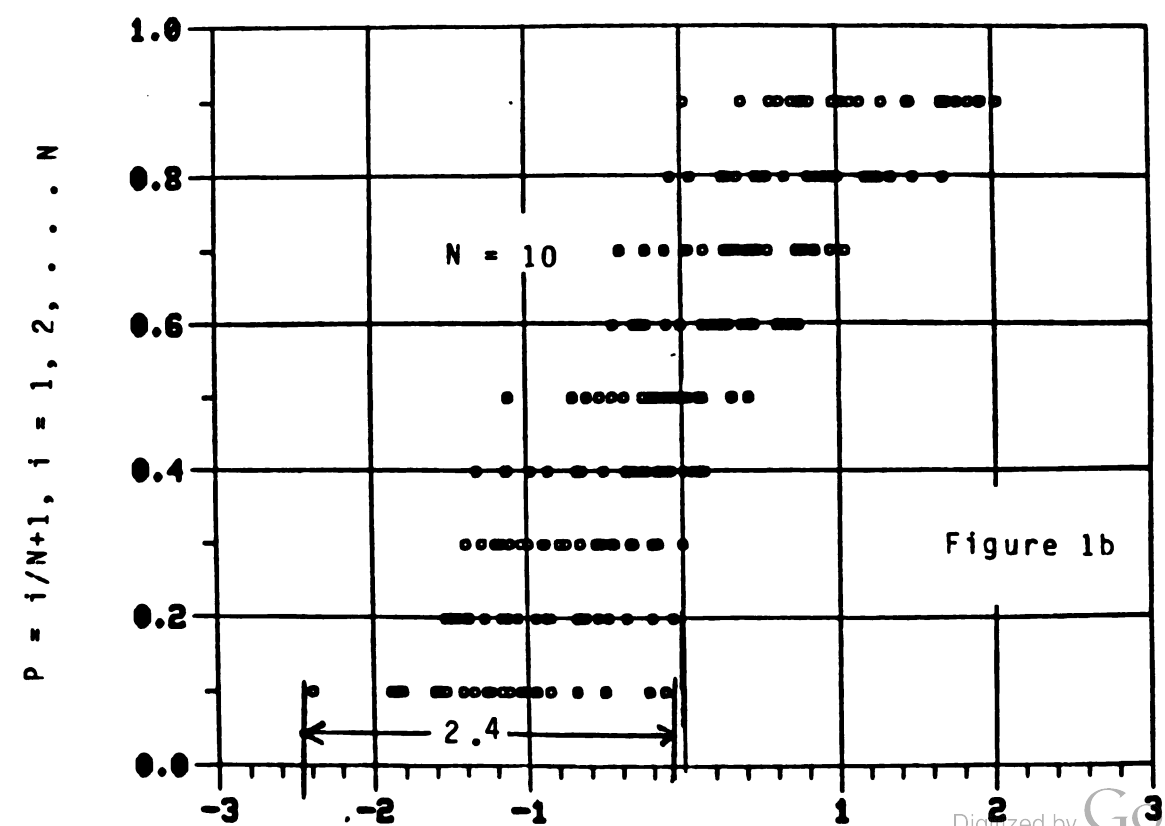


Figure 1b

NORMAL PROBABILITY DENSITY FUNCTION

MEAN = 0.0

STANDARD DEVIATION = 1.0

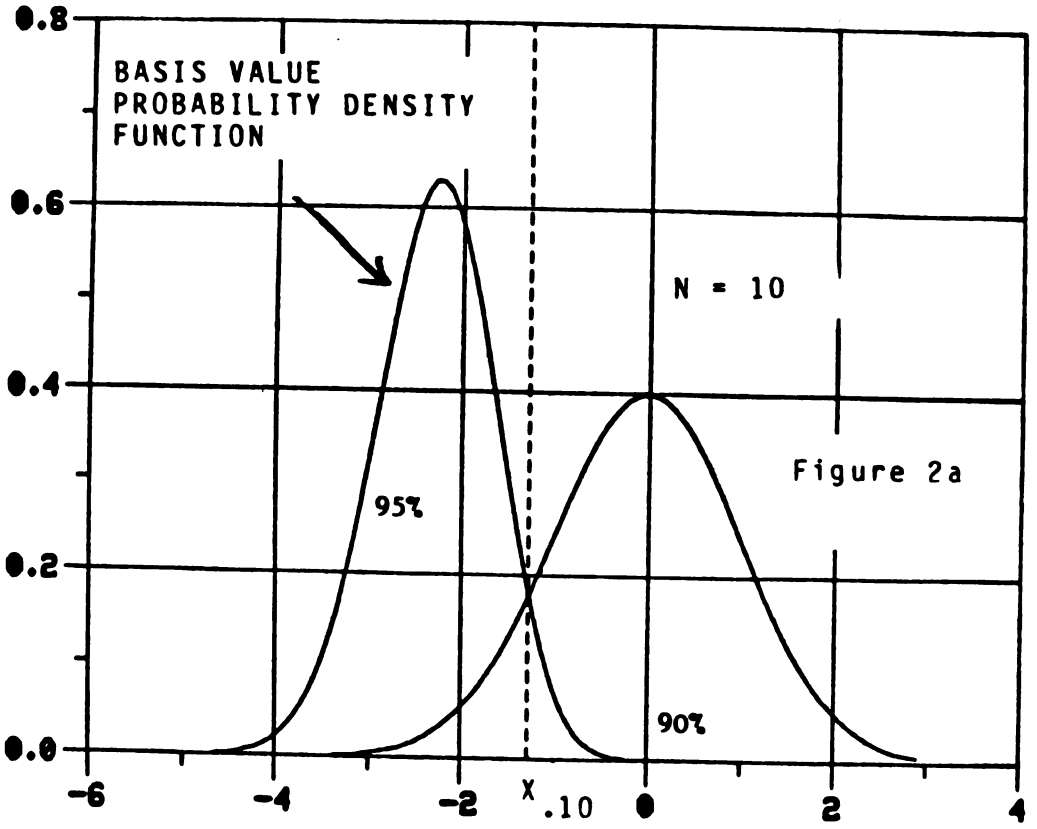


Figure 2a

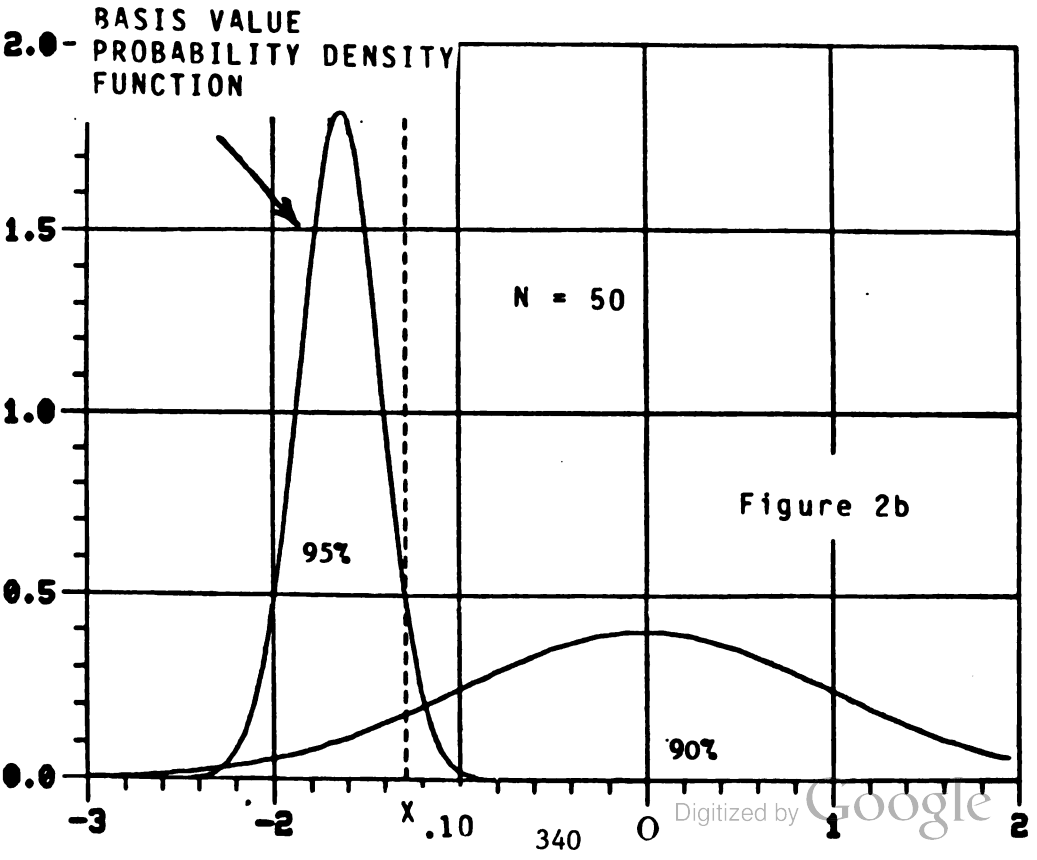


Figure 2b

TABLE I

Ranks (r) of Observations (n) for determining
B-Basis Values for an Unknown Distribution

n	r	n	r	n	r
<29	-	129	8	227	16
29	1	142	9	239	17
46	2	154	10	251	18
61	3	167	11	263	19
76	4	179	12	275	20
89	5	191	13	298	22
103	6	203	14	321	24
116	7	215	15	345	26

EXTENDED QUANTILE SIGN TEST

The extended quantile sign test was developed in order to obtain the B-basis values for $n < 29$ using nonparametric procedures.

Let n be a fixed value less than 29. Calculate the probability values $P_1, P_2, \ldots, P_k$ as follows: $1 \leq j \leq k$, and P_j is the solution of

$$.05 = (1 - P_j)^n + \binom{n}{1}(1 - P_j)^{n-1} P_j + \binom{n}{2}(1-P_j)^{n-2} \quad (2)$$

$$+ \ldots \binom{n}{j}(1 - P_j)^{n-j} P_j{}^j$$

where $k << n$.

Example: If .n = 15, let k = 3, then P = .181, P_2 = .280 and P_3 = .364, with corresponding order statistics $X_{(1)}$, $X_{(2)}$, and $X_{(3)}$, (see Figure 3).

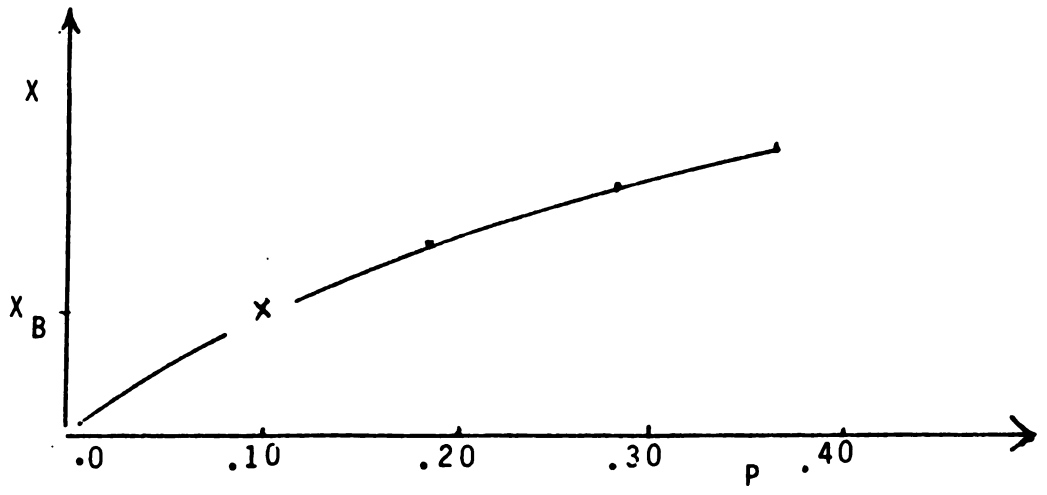


FIGURE 3

The following interpolation models were applied in order to obtain X_B for P = .10:

$$X = A \text{ Log } (BP + 1), \tag{3}$$

$$X = A \text{ Log } (PB + C) \text{ given } P = .0001 \text{ and } X_0 = 0.0, \tag{4}$$

$$\text{and } X = AP^2 + BP + C \text{ given } P = .0001, X_0 = 0.0. \tag{5}$$

The model in equation (3) failed to provide acceptable interpolation results because of its inability to represent (P_0, X_0), (P_1, X_1), (P_2, X_2), and (P_3, X_3) effectively.

The models from equations (4) and (5) provided adequate interpolation results for P = .10. The computation procedures for obtaining each B-basis value from sample requires either linear (Equation 5) or nonlinear (Equation 4) regression models. These are not simple computational methods when compared to conventional quantile sign test application. The authors applied a simulation process with a given n value and $N(0,\sigma)$ models to approximate probability density function for

$$F = \frac{\text{B-basis}}{X_{(1)}} \tag{6}$$

The purpose was to obtain a reduction factor F for the first ordered value in order to obtain the basis value for given sample size n. This is a similar computational procedure, as in the conventional nonparametric method.

A schematic of the results are shown below:

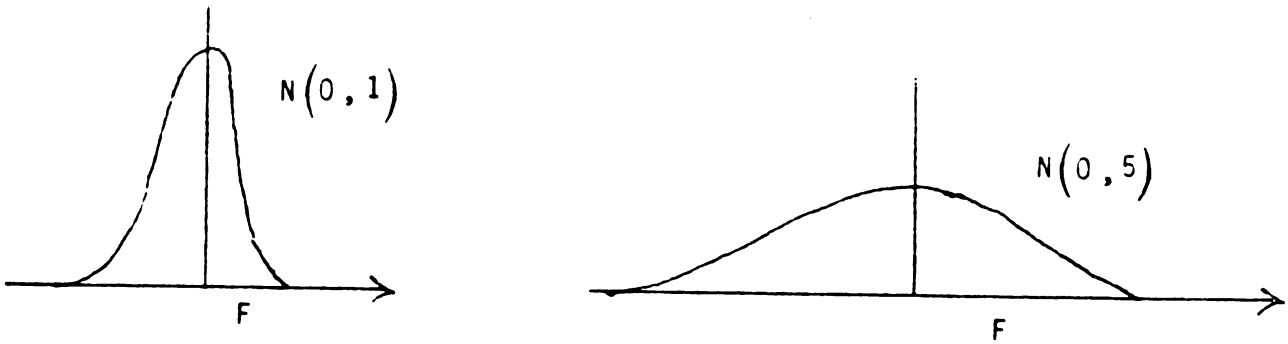


FIGURE 4

The substantial spread in the N(0,5) case for F was primarily the result of unstable first ordered values for the simulation. The authors have also rejected the above approach since it requires individual regression modeling for each sample if an accurate basis number is to be obtained.

B-BASIS VALUES FROM FIRST FOUR ORDERED STATISTICS

This method involves determining a linear relationship between the weighted gaps of the first four ordered values ($X_{(1)}$, $X_{(2)}$, $X_{(3)}$, and $X_{(4)}$), and the standard deviation s of a normal population N(20,S). Random selection of R = 5000 samples of size n were obtained for selected integer values $1 \leq S \leq 5$. A 95% tolerance limit is obtained for the 10 percentile of the $N_i(20,S)$ distributions. This value $\left[X_{.10}\right]_{.95}$ represents the value where 95% of all values are $< X_{.10}$ (10 percentile of distribution) from the random sample of size n. This $\left[X_{.10}\right]_{.95}$ value approximates the B-basis value for $N_i(20,S)$.

The following linear relationships are applied in the analysis:

$$y_1(s) = X_4 + X_3 + X_2 - 3X_1$$
from weighted gaps $X_2 - X_1$, $X_3 - X_1$, $X_4 - X_1$ (7)
of the order statistics

and $$y_2(s) = \frac{X_4 + X_3 + X_2 + X_1}{4} - [X_{.10}].95 \quad (8)$$

where X_i's are expected values from the order statistics.

The linear relationships in equations 7 and 8 are represented graphically in Figure 5

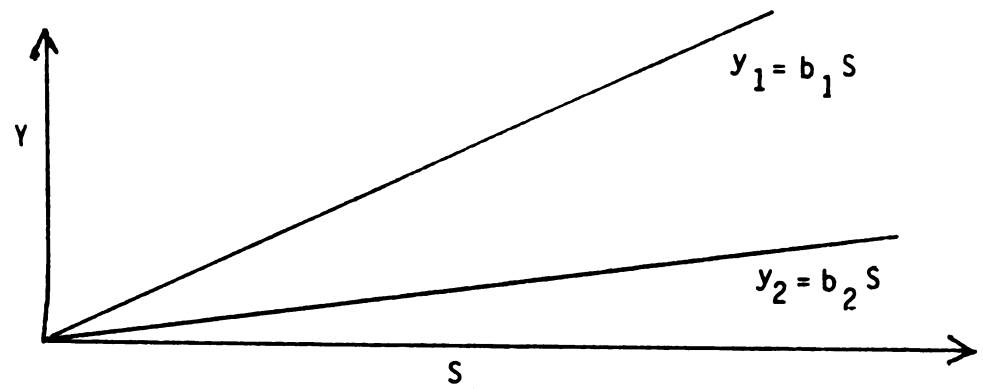


FIGURE 5

The B-basis value can now be defined as,

$$\text{B value} = (3S^* + .25)X_1 - (S^* - .25)(X_4 + X_3 + X_2), \quad (9)$$
where $S^* = b_2/b_1$.

The above method was very effective for $n > 15$, and provided reasonable coverage rates, as indicated in the Monte Carlo study using Weibull and normal distribution.

HANSON-KOOPMANS TOLERANCE LIMITS

The Hanson-Koopmans[1] method provided the most satisfactory results for obtaining small sample nonparametric B-basis values. The method involves applying the following equation:

$$\text{B value} = X_{(k+j+1)} - C_n(X_{(k+j+1)} - X_{(k+1)}) \qquad (10)$$

where $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$ are the ordered statistic values, and C_n is obtained from solution of

$$.95 = \frac{n!}{(n-j-k-1)!(j-1)!k!} \left| \int_0^p \int_0^v + \int_p^1 \int_0^{p^{1/C_n} v^{(C_n-1)/C_n}} \right| (11)$$

$$\cdot w^k (v-w)^{j-1} (1-v)^{n-k-j-1} dw dv$$

where $j \geq 1$ and $k \geq 0$. The region of integration by the lines $w=0$ for $0 \leq v \leq 1$, $v=1$ for $0 \leq w \leq 1$, the line $w=v$ for $0 \leq v \leq p$ and by the curve

$$w = p^{1/C_n} v^{(C_n-1)/C_n}$$

for $p \leq v \leq 1$.

The C_n values for equation 10 are tabulated in Table II where $j = 3$ and $k = 0$ to obtain basis values for $3 < n \leq 30$, as in the following manner:

$$\text{B value} = X_4 - C_n(X_4 - X_1).$$

Using the first and fourth ordered values provided acceptable results, although another combination could have provided a better approximation to the desired coverage rate of 95%. In the case where $3 \geq n > 1$, the following equations were obtained:

$$\text{B value} = X_2 - C_n(X_2 - X_1),$$
$$\text{where } C_2 = 35.2 \text{ and } C_3 = 28.8,$$
$$\text{for } n = 2 \text{ and } 3 \text{ respectively.}$$

A Comparison: Hanson-Koopmans (HK) vs. Linear Order Statistic Model (LOSM).

TABLE II

n	C_n	n	C_n	n	C_n	n	C_n
4	4.505	11	2.635	18	1.745	25	1.198
5	4.101	12	2.474	19	1.651	26	1.136
6	3.765	13	2.327	20	1.564	27	1.078
7	3.478	14	2.192	21	1.482	28	1.027
8	3.229	15	2.068	22	1.404	29	.971
9	3.009	16	1.952	23	1.331	30	.811
10	2.812	17	1.845	24	1.263		

A Monte Carlo study was completed involving a comparison of coverage rates obtained from HK and LOSM, where the Weibull $W(\alpha,\beta)$ and Normal $N(\mu,\sigma)$ were the selected probability density functions. In the simulation, the confidence coefficient (coverage rate) was obtained from determining the percentage of replications for which the B value was less than the actual 10th percentile of the distributions. 5,000 replications were used in the experiment. A minimum percent of 95 is required. Percentage slightly greater than 95 is also desirable.

In Tables III and IV, the coverage rate's percent is tabulated for both the Linear Order Statistic Method and the Hanson-Koopmans Method where the normal and Weibull models are used in the simulation process. In Table III, a range of standard deviations are considered in order to examine for the effects of dispersion in the data. LOSM results show poor coverage rates when n = 10, and acceptable coverage rates for n=15 and $10 \geq \sigma$. The Hanson-Koopmans results show universal acceptance except for marginal acceptability for n = 15. The authors also obtained results for n = 14, 15, 16, 17, 18, and 25 for the HK method. In all cases, coverage rates of at least .95 were obtained, indicating that the lowest values are for n = 15. A different set of ordered values could possibly increase coverage values for n = 15. Results from the table indicate an optimization process could be developed where a set of ordered values would be determined to provide the

minimum acceptable coverage (.95) depending on sample size n. This would prevent the over conservativism shown in the tables (e.g. .99, .98, .97 coverage rate).

In Table IV, a range of α values (shape parameter) for Weibull functions, are used for examining effects of dispersion. Again, the LOSM results show poor coverage rates when n = 10. The case when N = 15 shows reason values since .93 is the lowest value.

The Hanson-Koopmans results are similar to those shown in Table III. The minimum values are at n = 15, which also occurred when the normal model was used in the simulation process.

It can be inferred from the above results that the HK method is a desirable nonparametric procedure for obtaining B-basis values when $n \leq 28$. It is not clear why the reduction in coverage to .94 exists for n = 15, while n = 2 and n = 30 have a coverage rate of .99. Ideally, coverages of .95 for all n and dispersion parameters would be desirable to prevent overly conservative estimates of basis values.

TABLE III

Confidence Coefficient (%), $N(\mu, \sigma)$, $\mu = 50$,

	LINEAR ORDER STATISTIC METHOD		HANSON-KOOPMANS METHOD			
σ	n=10	n=15	n=5	n=10	n=15	n=30
2	.60	.99	.99	.97	.95	.99
6	.76	.98	.99	.97	.94	.99
10	.78	.94	.99	.98	.94	.99
14	.78	.92	.99	.98	.94	.99
30	.80	.90	.99	.97	.94	.99

TABLE IV

Confidence Coefficient (%), $W(\alpha, \beta)$, β=50

	LINEAR ORDER STATISTIC METHOD		HANSON-KOOPMANS METHOD			
α	n=10	n=15	n=5	n=10	n=15	n=30
2	.82	.93	.99	.98	.96	.99
6	.78	.93	.99	.97	.95	.99
10	.77	.96	.98	.97	.94	.99
14	.77	.98	.98	.97	.94	.99
30	.74	.99	.98	.97	.94	.99

CONCLUSIONS

The Hanson-Koopmans nonparametric small sample tolerance limit model provided the most desirable solution to obtaining B-basis values. The authors method, LOSM, provided an acceptable method if $n \geq 15$. For small sample sizes, results were excessively non-conservative.

Methods involving factors of the first order statistic resulted in overly conservative or non-conservative B-value estimates, depending on the dispersion of data and the sample size. The extended quantile sign test failed to provide either a computationally simple solution to obtaining basis values, or a factor associated with first ordered value in calculated B-basis value. The need for repeated application of non-linear regression to each sample, when factors were not available, reduces its value as an engineer's statistical method. The conventional quantile sign test was not applicable for $n < 29$, although it is an acceptable procedure otherwise.

REFERENCES

1. D. L. Hanson and L. H. Koopmans, "Tolerance Limits for the Class of Distributions with Increasing Hazard Rates," Ann. Math. Stat. 35: 1964.

2. MIL-HDBK-5C, Military Standardization Handbook, "Metallic Materials and Elements for Aerospace Vehicle Structures," Naval Publications, 5801 Tabor Avenue, Philadelphia, PA 19120.

3. MIL-HDBK-17B, Military Standardization Handbook, "Composite Materials for Aircraft and Aerospace Applications," Materials Technology Laboratory, Watertown, MA 02172-0001.

4. Conover, W. J., Practical Nonparametric Statistics, John Wiley and Sons, NY, pg. 111: 1980.

5. Brieman, L., C. Stone, and J. Gins, "Further Developments of New Methods for Estimating Tail Probabilities and Extreme Value DIstributions," Technical Report No. TSD-PH-A243-1, Technology Services Corporation, Santa Monica, CA: 1981.

6. Padgett, W. J., "A Nonparametric Quantile Estimator: Computation," Technical Report No. 117 G2G05-9, Department of Statistics, University of South Carolina, Columbia, SC 29208.

ACKNOWLEDGEMENTS

The authors are indebted to Professor Bernard Harris of the University of Wisconsin for presenting this paper at the clinical session of the Thirty-Sixth Design of Experiment Conference at Monterey, California in October of 1986. The authors would also like to thank Betty Landry of the Materials Technology Laboratory for preparing this manuscript.

A SECOND LOOK AT THE PERVERSITY OF MISSING POINTS IN THE 2^4 DESIGN

Carl T. Russell
US Army Operational Test and Evaluation Agency
Falls Church, Virginia

ABSTRACT. At the 1982 Design of Experiments Conference, the author presented a Clinical Paper entitled *The Perversity of Missing Points in the 2^4 Design.* That paper tried to characterize what points could be deleted from the 2^4 design without losing the resolution V property (that is, main effects and 2-factor interactions are estimable). That paper used brute force (computer plus sweat) methods to investigate numerous special cases and formulate some promising conjectures, but no general conclusions were reached. G.E.P. Box was the primary discussant on the paper, and he suggested using a matrix trick to reduce the dimensionality of the problem from eleven to five. The current paper shows how notation from group theory and graph theory can be used to exploit Box's suggestion to prove the conjectures of the original paper. In particular, even if five of the sixteen points are deleted at random from a 2^4 design, the probability is almost 0.7 that the resulting design is still resolution V—that is, all eleven parameters are estimable from the remaining eleven data points. Unfortunately, the method used does not appear to generalize to larger designs of greater interest.

1. INTRODUCTION. Execution of a military field test seldom proceeds exactly as planned, and rather large amounts of missing data are common. In fact, two other papers given at this Design of Experiments Conference dealt with aspects of the problem. Winner and Smith described a situation where a large portion of the planned experiment captured no data; Bryson and Russell presented a method for adjusting attrition estimates from "Real Time Casualty Assessment" based on changed estimates of kill probabilities which were "missing" when the real time casualty assessments were made. In 1982, I approached the problem from a different angle by studying what happens when points are arbitrarily deleted from a factorial design (Russell, 1983a). This study was motivated by the observation that most field tests of military materiel are designed in a factorial framework and conducted in blocks of time and/or space. The blocks could in theory be constructed from appropriately chosen fractional factorials to reduce the potential bias due to confounding

which is common in much traditional field test design (see Russell—1981, 1982, 1983b, and Section V of this paper). Before such a design approach can be prudently implemented in expensive field tests, however, an understanding of its robustness to substantial data loss is needed.

In the summer of 1982, I began this study by looking at the simplest interesting factorial design, the 2^4 design. The study asked two questions:

(1) Characterization problem—what points can be deleted from the 2^4 design without losing estimability of the mean, main effects and 2-factor interactions (resolution V property)?

(2) Structural problem—when the remaining design is resolution V, what is the structure of the least squares estimates obtained?

The problem turned out to be much harder than I anticipated, and it grew into a Clinical Paper presented at the 1982 Design of Experiments Conference (Russell, 1983a). That paper used brute force methods to beat a portion of the structural problem to death using a computer and to make some promising conjectures for the characterization problem. G.E.P. Box was the primary discussant, and he made a suggestion for the characterization problem which enabled me to prove the original conjectures and quantify the likelihood that random deletions of points from the 2^4 design would destroy the resolution V property. This paper presents the results growing out of Professor Box's suggestion. Unfortunately, the methods used do not appear to generalize to larger designs of greater interest.

Why write this paper if the results essentially represent a dead end? First, it closes the loop from a Clinical Session where as an Army statistician, I received useful assistance on an important problem which enabled me to proceed further than I otherwise could have. Second, even though the methods of this paper do not appear to generalize, they are mathematically appealing, they took me a good part of the 1982-83 winter to derive, and they enable quantitative results which make me more optimistic that some fractional factorial blocking approaches may be quite robust against random data loss. Third, this paper re-emphasizes an important problem which needs and deserves further work by statisticians.

II. PRELIMINARIES. A standard notation for the four factors and sixteen points in the 2^4 design is the following.

		Lo B		Hi B	
		Lo C	Hi C	Lo C	Hi C
Lo D	Lo A	(1)	c	d	cd
	Hi A	a	ac	ad	acd
Hi D	Lo A	b	bc	bd	bcd
	Hi A	ab	abc	abd	abcd

(1)

The full model can be written (in slightly unusual order) as

$$\begin{aligned} Y_{ijk\ell} = {} & \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ & + (\alpha\beta)_{ij} + \alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} \\ & + (\beta\gamma\delta)_{jk\ell} + (\alpha\gamma\delta)_{ik\ell} + (\alpha\beta\delta)_{ij\ell} + (\alpha\beta\gamma)_{ijk} \\ & + (\alpha\beta\gamma\delta)_{ijk\ell} + \varepsilon_{ijk\ell}, \end{aligned} \tag{2}$$

where the subscripts can be removed by the usual side conditions:

$$\alpha_0+\alpha_1=0, \text{ that is } \alpha_i=\pm\alpha \tag{3}$$

$$\vdots \qquad\qquad \vdots$$

$$\delta_0+\delta_1=0, \text{ that is } \delta_i=\pm\delta$$

$$(\alpha\beta)_{00}+(\alpha\beta)_{01}=(\alpha\beta)_{10}+(\alpha\beta)_{11}=0, \text{ that is } (\alpha\beta)_{ij}=\pm(\alpha\beta)$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$(\gamma\delta)_{00}+(\gamma\delta)_{01}=(\gamma\delta)_{10}+(\gamma\delta)_{11}=0, \text{ that is } (\gamma\delta)_{k\ell}=\pm(\gamma\delta)$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$(\alpha\beta\gamma\delta)_{0000}+(\alpha\beta\gamma\delta)_{0001} = \ldots = 0, \text{ ie, } (\alpha\beta\gamma\delta)_{ijk\ell}=\pm(\alpha\beta\gamma\delta).$$

In matrix form, these *normal equations* are

$$Y = X\cdot\beta + \varepsilon, \tag{4}$$

where the design matrix, X, has rows corresponding to design points and columns corresponding to parameters. The reduced main effects and 2-factor interactions model is:

$$Y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} + \varepsilon_{ijk\ell}, \quad (5)$$

which reduces the number of columns of X from 16 to 11. Figure 1 shows the design matrix for the full model, (2), partitioned to accentuate the missing parameters in model (5), and assuming the unsubscripted parameters from (3) are used. Deleting points from the design corresponds to deleting rows from X. The normal equations, (4), have a least squares solution iff $X'\cdot X$ is nonsingular, in which case the solution is

$$\hat{\beta} = (X'\cdot X)^{-1}\cdot X'\cdot Y. \quad (6)$$

Since there are 11 parameters in the reduced model, (5), solving the characterization problem via (6) for 5 or less missing points requires checking a matrix of dimensions at least 11×11 for singularity.

To reduce the dimensionality of the characterization problem, Box suggested partitioning the design matrix, X, by making the *m missing points* correspond to the last m rows and the *p parameters of interest* correspond to the first p columns:

$$X = \begin{array}{c} \begin{array}{cc} p & n-p \end{array} \\ \left[\begin{array}{c|c} X_1 & X_3 \\ \hline X_2 & X_4 \end{array}\right] \begin{array}{c} n-m \\ m \end{array} \end{array} \quad (7)$$

By assuming orthogonality of X and expanding the orthorgonality relationships $X'\cdot X = nI_m = X\cdot X'$ in matrix form, Box proved the following lemma via an eigenvalue argument.

Lemma. $X_1'\cdot X_1$ is nonsingular iff $X_4'\cdot X_4$ is nonsingular.

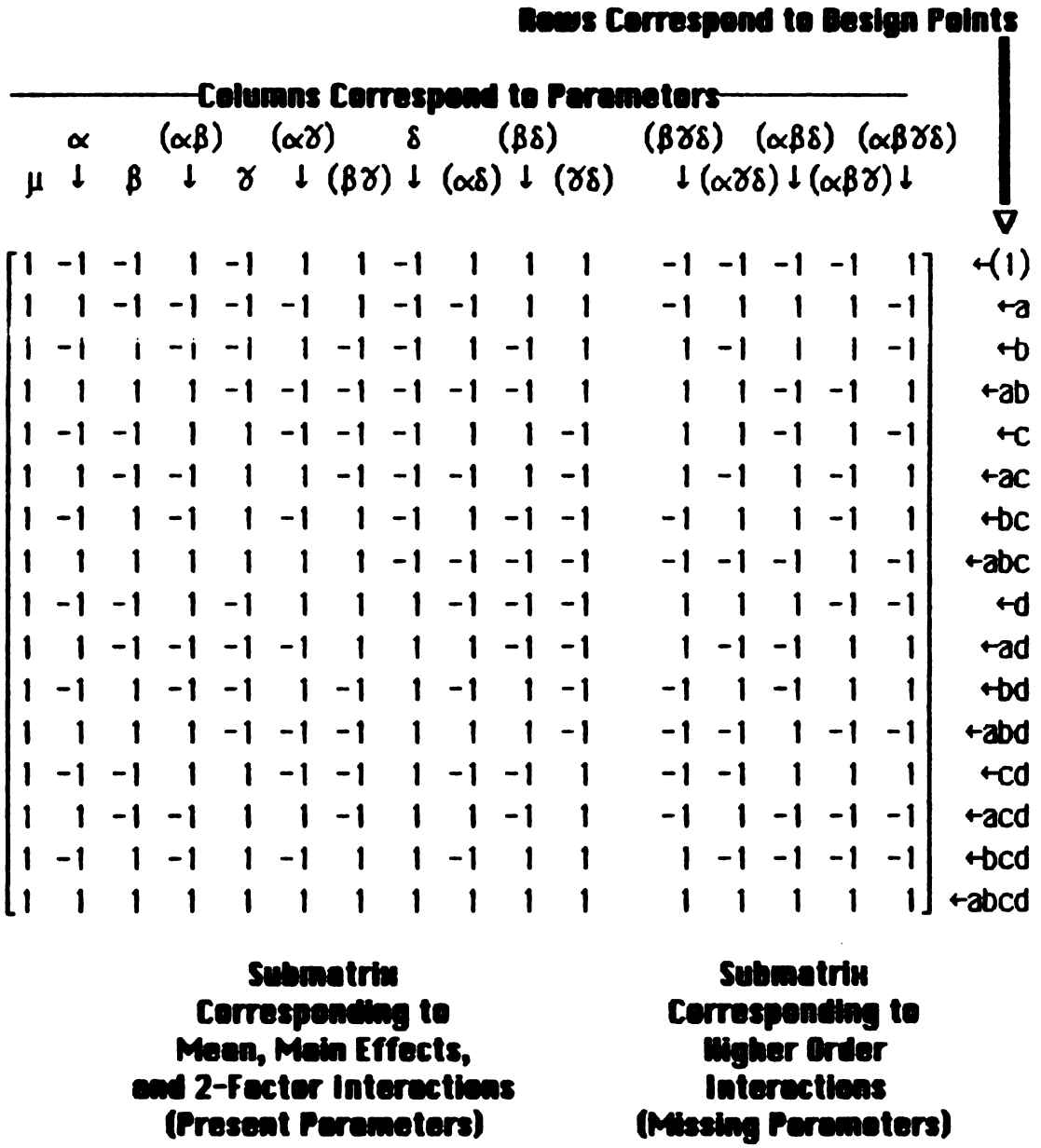


Figure 1. The Design Matrix, X, for the 2^4 Design, Partitioned to Show Present and Missing Parameters in the Main Effects and 2-Factor Interactions Model.

In the case of the 2^4 design, the lemma reduces the dimensionality of the characterization problem from eleven or more to five or less. For example, with five missing points (m=5) and p=11 (the number of parameters in (5)), the 5-dimensional square matrix $X_4^{\cdot}X_4$ could be examined for singularity instead of the 11-dimensional square matrix $X_1^{\cdot}X_1$, and the problem gets easier with less than five missing points.

The rows of the full 2^4 design corresponding to 3- and 4-factor interactions (that is, the rows of the missing parameter submatrix in Figure 1) represent the vertices of the 5-dimensional hypercube which have an even number of minus signs. By Lemma 1, the characterization problem is reduced to characterizing the subsets of these even vertices which are linearly dependent. All the vertices of the 5-dimensional hypercube form a group G_5 under coordinatewise multiplication, and the even vertices form a subgroup E_5. The subgroup E_5 is isomorphic to the quotient group obtained by identifying opposite vertices (ie, those with all signs switched) in G_5. By mnemonically relabeling the columns of missing parameters in Figure 1,

$(\beta\gamma\delta) \rightarrow \mathbf{A}$, since α is missing from $(\beta\gamma\delta)$,
$(\alpha\gamma\delta) \rightarrow \mathbf{B}$, since β is missing from $(\alpha\gamma\delta)$,
$(\alpha\beta\delta) \rightarrow \mathbf{C}$, since γ is missing from $(\alpha\beta\delta)$,
$(\alpha\beta\gamma) \rightarrow \mathbf{D}$, since δ is missing from $(\alpha\beta\gamma)$,
$(\alpha\beta\gamma\delta) \rightarrow \mathbf{E}$, since ε is missing from $(\alpha\beta\gamma\delta)$,

and letting a letter **A**, **B**, **C**, **D** or **E** appear in a vertex label iff the sign of the respective coordinate is positive, the quotient group becomes $E_5 \approx G_5/\{\mathbf{I}, \mathbf{ABCDE}\}$. Figure 2 gives the new labeling of the missing parameter submatrix from Figure 1 in terms of **A**, **B**, **C**, **D**, and **E** together with the respective cosets. The ±1's of Figure 1 have been replaced by simply +'s and −'s in Figure 2, and one- or two-letter design point labels are underlined to indicate that they will be used as standard coset labels. (Group theory is used here only for limited notational convenience: not much algebra is exploited. Likewise, the graph theory introduced in the next section is used simply as a bookkeeping tool. Better exploitation of these mathematical objects might lead to more general results.)

OLD POINT LABELS[1]	PARAMETER LABELS[2] (βγδ) A	(αγδ) B	(αβδ) C	(αβγ) D	(αβγδ) E	NEW POINT LABELS[3] (COSETS[4])
(1)	-	-	-	-	+	E ≈ ABCD
a	-	+	+	+	-	BCD ≈ AE
b	+	-	+	+	-	ACD ≈ BE
ab	+	+	-	-	+	ABE ≈ CD
c	+	+	-	+	-	ABD ≈ CE
ac	+	-	+	-	+	ACE ≈ BD
bc	-	+	+	-	+	BCE ≈ AD
abc	-	-	-	+	-	D ≈ ABCE
d	+	+	+	-	-	ABC ≈ DE
ad	+	-	-	+	+	ADE ≈ BC
bd	-	+	-	+	+	BDE ≈ AC
abd	-	-	+	-	-	C ≈ ABDE
cd	-	-	+	+	+	CDE ≈ AB
acd	-	+	-	-	-	B ≈ ACDE
bcd	+	-	-	-	-	A ≈ BCDE
abcd	+	+	+	+	+	ABCDE ≈ 1

Tabulated Symbols Are the Signs of Entries in the Missing Parameters Submatrix of Figure 1.

[1] The usual way of labeling design points, or treatments, in 2^n experimental designs—see (1) in text.

[2] Greek letters in parentheses are the old parameter labels, outlined capital letters are the new parameter labels.

[3] An outlined letter appears in the new treatment label (first column) iff "+" appears in the respective column.

[4] Cosets obtained by identifying opposite vertices of the five dimensional hypercube. Underlined labels (one- or two-letter combinations) are used as standard coset labels.

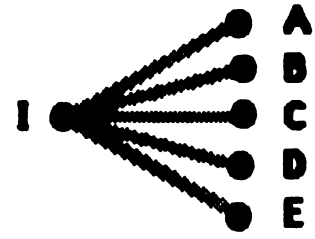
Figure 2. The Matrix of Missing Parameters, Showing New Labels.

III. MAIN RESULT. The geometric underpinnings of the reformulated problem suggest a geometric approach to its solution. The problem is to characterize linearly dependent subsets of the vertices of a hypercube. Clearly, opposite vertices of a hypercube are pairwise linearly dependent (hence interchangeable in linearly dependent subsets), so the identification of opposite vertices in cosets simply gets rid of a trivial nuisance. Once opposite vertices are identified in the 5-dimensional hypercube, defining an edge between two new vertices if there was an edge between pairs of old vertices is natural. In fact, it is useful to define a *quotient graph* on the new vertices as follows.

> **Definition**. The *quotient graph*, G_5 of G_5 is the graph whose vertices, $V(G_5)$, are the elements of the quotient group E_5 and whose edges, $E(G_5)$, connect any pair of vertices in $V(G_5)$ whose cosets were connected in G_5.

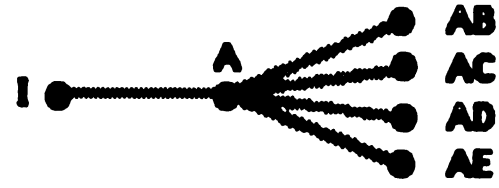
In the notation of Figure 2, $V(G_5)$ consists of the identity, single letters, and double letters, and $E(G_5)$ consists of edges connecting:

- **I** with each of the single letters **A**, **B**, **C**, **D**, and **E**.



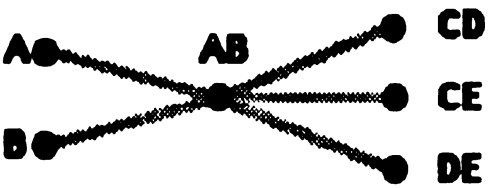
- Each of **A**, **B**, **C**, **D**, and **E** with **I** and with each double letter containing it.

Example:



- Each double letter with all single letters contained in it and with each other double letter disjoint from it.

Example:

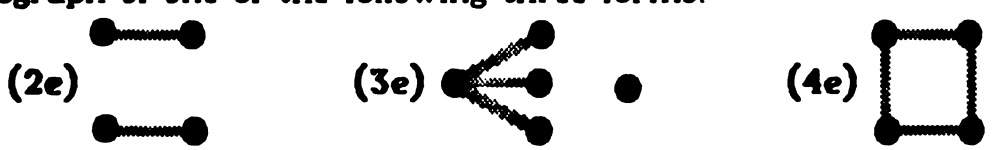


The usual definition of a subgraph is:

Definition. A *subgraph* $\mathcal{S}$ of a graph $\mathcal{G}$ defined by the vertices $V(\mathcal{S})$ is the graph such that $V(\mathcal{S}) \subseteq V(\mathcal{G})$ and $E(\mathcal{S})$ consists of the edges of $\mathcal{G}$ connecting vertices of $\mathcal{S}$.

With this definition, the following theorem will be proved.

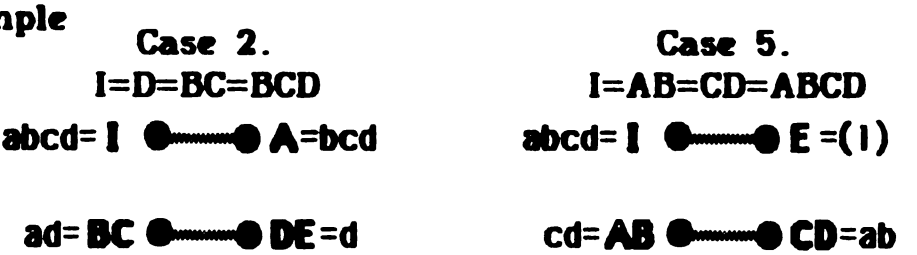
Theorem (Main Result). Let $\mathcal{S}$ be a subgraph of $\mathcal{G}_5$ defined by $V(\mathcal{S})$ where $V(\mathcal{S})$ has 5 elements. Then the elements of E_5 corresponding to $V(\mathcal{S})$ are linearly dependent iff $\mathcal{S}$ contains a subgraph of one of the following three forms.



Proof of Sufficiency. It is easy to show that each of the three types of subgraph give linear dependence. All are closely related to P.W.M. John's three-quarter replicates (John 1971, pages 161-163), which formed the basis for much of my earlier paper (see especially Table 1, page 504, of Russell 1983a, denoted by "T1" below).

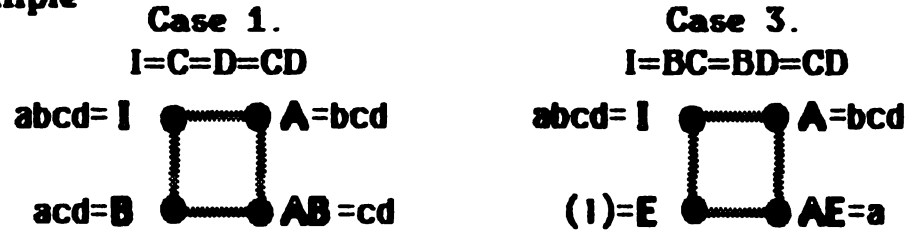
The 2-edge type, (2e) — with 4 vertices — corresponds to deleting the quarter replicates of cases 2 and 5 in T1 (defining contrasts similar to I=D=BC=BCD and I=AB=CD=ABCD).

Example



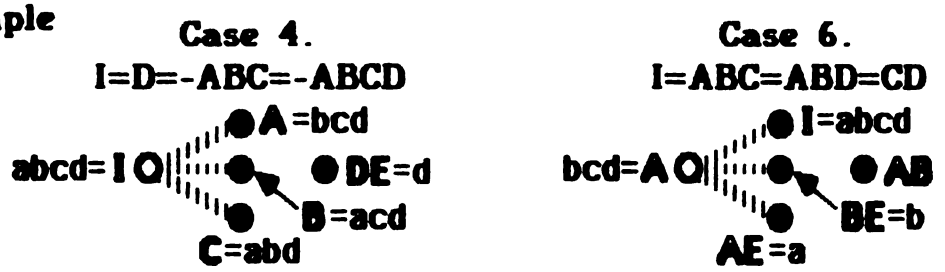
The 4-edge type, (4e) — with 4 vertices — corresponds to deleting the quarter replicates of cases 1 and 3 in T1 (defining contrasts similar to I=C=D=CD and I=BC=BD=CD).

Example



The 3-edge type, (3e) — with 5 vertices — corresponds to deleting an appropriate additional point from the quarter replicates of cases 1 and 3 in T1 (defining contrasts similar to I=D=-ABC=-ABCD and I=ABC=ABD=CD).

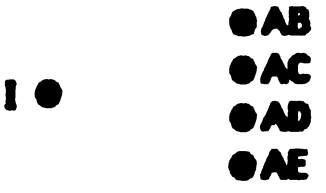
Example



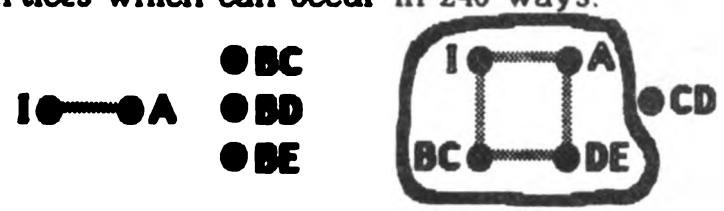
These correspondences between the graphs and three-quarter replicates show that the main result actually establishes the conjecture on page 522 of Russell, 1983a.

Proof of Necessity. To prove necessity assume without loss of generality that I is a vertex of the subgraph and that I has the maximum number of incident vertices. The proof considers cases based on the number of vertices at I, and as a byproduct used in later extensions of the theorem, counts the number of possible graphs of each type which result in linear dependence.

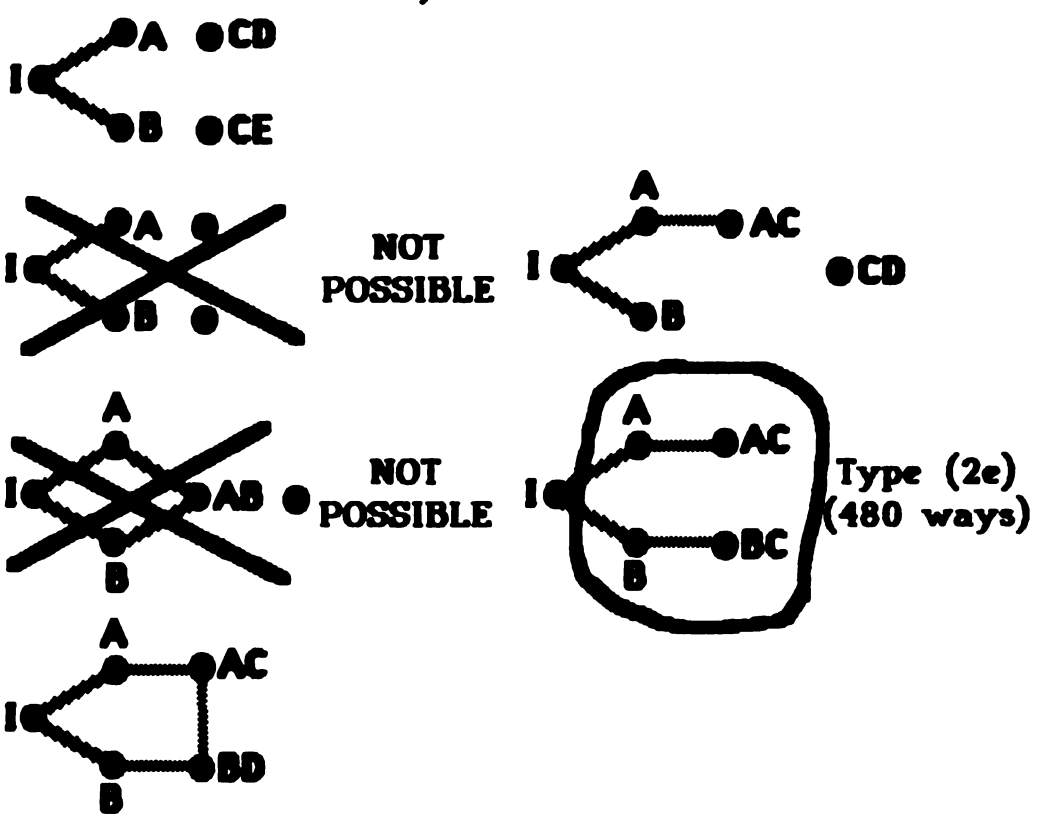
Case 0-Edge. If there are no edges incident at I, then the only subgraph has independent vertices.



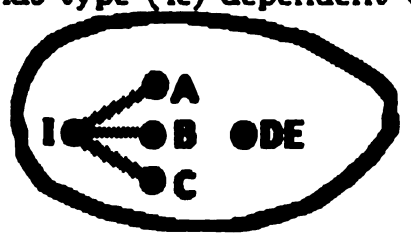
Case 1-Edge. If there is one edge incident at I, then there are two types of subgraph, one of which has type (2e) dependent vertices which can occur in 240 ways.



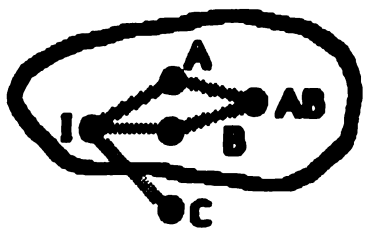
Case 2-Edge. If there are two edges incident at I, then there are six conceptual types of subgraph, two of which are not realizable, and one of which has type (2e) dependent vertices which can occur in 480 ways.



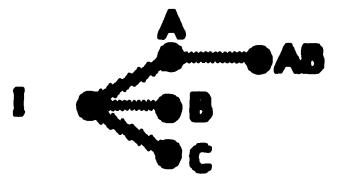
Case 3-Edge. If there are three edges incident at $\mathbf{I}$, then there are three conceptual types of subgraph. One has type (3e) dependent vertices which can occur in 160 ways. The other has type (4e) dependent vertices which can occur in 480 ways.



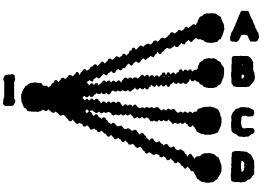
Type (3e)
(160 ways)



Type (4e)
(480 ways)

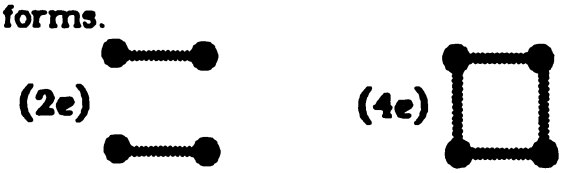


Case 4-Edge. If there are four edges incident at $\mathbf{I}$, then the only subgraph has independent vertices.



IV. CONSEQUENCES OF MAIN RESULT. It follows immediately form the main result that cases (2e) and (4e) represent the only ways four points can be deleted from the 2^4 design and fail to leave a design of resolution V and that if less than four points are deleted then the remaining design is of resolution V.

Corollary. Let $\mathcal{S}$ be a subgraph of G_5 defined by $V(\mathcal{S})$ where $V(\mathcal{S})$ has 4 elements. Then the elements of E_5 corresponding to $V(\mathcal{S})$ are linearly dependent iff $\mathcal{S}$ contains a subgraph of one of the following two forms.



Corollary. If less than 4 points are deleted at random from the 2^4 design, then the remaining design is of resolution V.

Totalling numbers of linearly dependent subgraphs identified in the proof of the main result (with a recount in the 4-point case) gives the following rather surprising quantitative results. They establish the conjecture that most designs obtained by deleting four or five points at random from the 2^4 design are of resolution V (Russell, 1983a, page 522).

Extension. If 5 points are deleted at random from the 2^4 design, then the probability that the resolution V property is lost is

$$\frac{240+489+160+480}{\binom{16}{5}} = \frac{1360}{4368} = 0.31.$$

Extension. If 4 points are deleted at random from the 2^4 design, then the probability that the resolution V property is lost is

$$\frac{160+40}{\binom{16}{4}} = \frac{200}{1820} = 0.11.$$

V. GENERALIZATION OF MAIN RESULT. The methods of this paper do not appear to generalize in a useful manner to 2^n designs with $n>4$. They might extend to the case where just n- and (n-1)-factor interactions are ignored in the 2^n design, but that situation is of little interest. Reduction of dimensionality is already a big problem in the resolution V case with the 2^5 design: there are as many excluded parameters as present parameters (16 parameters), and the 16-dimensional hypercube looks very complicated. The present methods would require looking at 32 of the 65,536 vertices of the 16-dimensional hypercube to study loss of the resolution V property. The case which originally interested me in this problem was even larger, namely, a resolution V quarter replicate of the 2^8 design. I felt and still feel that such a design should be relatively insensitive to data loss, but the methods of this paper don't seem to provide a good way to look at missing points in such designs. It is possible that extending the geometric, graph theoretic approach might be easier than it appears. Or a usable general characterization of linearly dependent vertices in an n-dimensional hypercube may be known. Alternatively, there might be a purely algebraic approach to the problem which would yield general results.

In any case, the general problem still needs and deserves more work. As an exercise to see how far one might push 2^{n-k} fractional factorial designs in a field test framework, I designed in 1983 a

hypothetical operational test for a communications jammer using formal experimental design methods (Russell, 1983b). The resulting design examined 32 factors, each nominally at 2 levels, in such a way that 62 effects (including carefully chosen interactions) would be estimable. The design had 512= 2^9 points and in fact was a 2^{-23} = 1/8,388,608 fraction of a 2^{32} design (4,294,976,296 points) run in 64 blocks of size 8. The design would have required 8 days to run and could have been easily extended in 8-day increments to a "full factorial" test 2^{19} = 254,288 times as long and lasting over 11,000 years. If one were really to try to run such a design, however, the risk associated with missing points shouldn't be too serious because there are many more points than parameters, and the results of this paper concerning the 2^4 design suggest that the risk could be quite small. But even at a cost of only $1,000 per data point, actually running such a design would cost more than a half million dollars. Large field tests cost many times more. The statistician's risk in proposing even substantially more modest designs (such as that in Russell, 1982) would be much less if there were better theoretical understanding of robustness to data loss.

VI. REFERENCES.

John, Peter W.M. (1971). *Statistical Design and Analysis of Experiments*. New York: Macmillan.

Russell, Carl T. (1981). "The Potential Utility of Crossing a Fractional Factorial with a Full Factorial in the Design of Field Tests." Raleigh: US Army Research Office Report No. 81-2, *Proceedings of the Twenty-Sixth Conference on Design of Experiments in Army Research Development and Testing*, 335-346.

Russell, Carl T. (1982). "Selling a Complicated Experimental Design to the Field Test Operator." Raleigh: US Army Research Office Report No. 82-2, *Proceedings of the Twenty-Seventh Conference on Design of Experiments in Army Research Development and Testing*, 293-308.

Russell, Carl T. (1983a). "The Perversity of Missing Points in the 2^4 Design." Raleigh: US Army Research Office Report No. 81-2, *Proceedings of the Twenty-Eighth Conference on Design of Experiments in Army Research Development and Testing*, 499-523.

Russell, Carl T. (1983b). "Using Efficient Experimental Design to Accomodate a Wider Variety of Test Conditions within Resource Constraints." Falls Church, VA: US Army Operational Test and Evaluation Agency, *Proceedings of the Twenty-Second Annual Army Operations Research Symposium*, 4-354–4-379.

A METHOD FOR THE STATISTICAL ANALYSIS OF THE STRESS-STRAIN PROPERTIES OF EARTH MATERIALS

G. Y. Baladi and B. Rohani
Geomechanics Division, Structures Laboratory
U.S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi

ABSTRACT. Stress-strain properties of earth materials under various test boundary conditions, such as uniaxial strain, hydrostatic compression, and triaxial shear, are required for conducting a two-dimensional (2D) analysis of explosive-induced ground shock. Such properties are random and often contain artificial instrumentation-induced noise. The randomness is primarily due to spatial variation of the soil properties, biases associated with field sampling disturbance, and errors in laboratory testing equipment and procedures and must be accounted for in ground shock analysis. This necessitates the use of 2D probabilistic wave propagation computer codes as opposed to deterministic procedures. To use the stress-strain properties for such probabilistic calculations, one must first eliminate (or reduce) the spurious instrumentation-induced noise in the "raw" data and then statistically quantify the "smoothed" data. The outcome of the statistical quantification is the representation of the stress-strain data in terms of the expected response, its variance, and the associated correlation coefficients. The paper discusses a methodology for smoothing the raw stress-strain data and the subsequent statistical analysis. Application of the methodology is demonstrated for Nellis Baseline sand.

I. INTRODUCTION: The ground shock calculation techniques currently used to predict the states of stress and ground motions induced in earth masses by explosive detonations are deterministic tools. That is, the input parameters (media constitutive properties and surface airblast loadings) are specified as single-valued deterministic quantities or functions. In actuality, however, both the constitutive properties of earth materials and the characteristics of the airblast pulses are dispersed random variables. The randomness of these input variables indicates that resulting stresses and ground motions are also random variables. Therefore, ground shock problems should be analyzed probabilistically. The purpose of the probabilistic analysis is to obtain a quantitative understanding of how the variabilities or uncertainties in the input parameters for a particular problem affect the dispersion of the output quantities or parameters. To use the stress-strain properties for such probabilistic analysis, one must first eliminate (or reduce) the spurious instrumentation-induced noise in the "raw" data and then statistically quantify the "smoothed" data. The outcome of the statistical quantification is the representation of the stress-strain data in terms of the expected response, its variance and the associated correlation coefficients.

The paper presents the development of a computerized methodology for statistically analyzing a set of random stress-strain data. This includes (1) a procedure for eliminating the spurious noise in the raw data due to instrumentation without affecting the actual physical response of the material and (2) a procedure for statistically analyzing the random behavior of the "smoothed" data. The outcome of these procedures is a representation of the stress-strain data in terms of the expected response, its variance, and the

---

Comments by panelists Drs. Kaye Basford and W. T. Federer are at the end of this artical.

correlation coefficients. Application of the methodology is demonstrated for Nellis Baseline sand.

II. DATA SMOOTHING PROCEDURE. The laboratory stress-strain data often contain artificial noise due to instrumentation which must be filtered out before the data can be used. Therefore, a technique to smooth the measured data without changing the actual physical response of the material is needed. Such a procedure has been developed by Baladi and Barnes (Reference 1) and is based on the concept of a marching mean square. If the measured value of the i^{th} data point is expressed as $y_m(X_i)$, the corresponding smoothed response $y_s(X_i)$ can be expressed as

$$y_s(X_i) = \sqrt{\frac{1}{n-1} \sum_{k=i-\frac{n-1}{2}}^{k=i+\frac{n-1}{2}} y_m^2(X_k)} \qquad (1)$$

where $n - 1$ is the window over which the marching mean square is taken (i.e., $\frac{n-1}{2}$ is the number of data points to the left and to the right of the i^{th} data). Note that n has to be an odd number equal to or greater than 3.

Equation 1 was applied to smooth the raw data from uniaxial strain (Figure 1) and triaxial compression (Figure 2) tests for Nellis Baseline sand. As shown in these figures, the results of these tests are quite noisy. The value of n used to smooth these data was 5. Several passes had to be made in order to obtain a satisfactory set of smoothed stress-strain relations. The final set is shown in Figures 3 and 4, and it is noted that the overall character of the stress-strain relation is not altered as a result of the smoothing process (for example, compare Figures 1 and 3).

III. STATISTICAL ANALYSIS OF SMOOTHED STRESS-STRAIN DATA. In this section, a generic procedure is outlined for statistical analysis of nonlinear stress-strain data. Consider a set of curves relating the random variables y and x (Figure 5). The objective of the statistical analysis is to determine the mean curve with its one-standard-deviation bounds relating the random variables y and x. This can be accomplished by applying standard statistical procedures to the slope of the random curves in Figure 5. The following steps should be taken to conduct the statistical analysis:

(1) For a given set of n curves, divide the x-axis into μ number of equal increments Δx (Figure 5).

(2) For the i^{th} increment, determine the slope of the j^{th} curve denoted by Ω_{ij}

$$\Omega_{ij} = \frac{\Delta y_j}{\Delta x_i}, \quad j=1,2,\ldots,n \qquad (2)$$

(3) Determine the expected value and the standard deviation of the slope at the i^{th} increment for all the curves according to the following expressions:

$$\bar{\Omega}_i = E(\Omega_i) = \frac{1}{n} \sum_{j=1}^{j=n} \Omega_{ij} \tag{3}$$

$$\sigma(\Omega_i) = \sqrt{\frac{1}{n-1} \sum_{j=1}^{j=n} (\Omega_{ij} - \bar{\Omega}_i)^2} \tag{4}$$

(4) Next, compute the mean and the standard deviation of y. To accomplish this, the covariance and the correlation coefficient matrices of the slopes $\text{cov}(\Omega_k,\Omega_m)$ and ρ_{km}, respectively, can be first calculated from the following relations:

$$\text{cov}(\Omega_k,\Omega_m) = E[(\Omega_k - \bar{\Omega}_k)(\Omega_m - \bar{\Omega}_m)] = \frac{1}{n-1} \sum_{j=1}^{j=n} (\Omega_{kj} - \bar{\Omega}_k)(\Omega_{mj} - \bar{\Omega}_m) \tag{5}$$

$$\rho_{km} = \frac{\text{cov}(\Omega_k,\Omega_m)}{\sqrt{E[(\Omega_k - \bar{\Omega}_k)^2]\, E[(\Omega_m - \bar{\Omega}_m)^2]}} \tag{6}$$

in which

$$E\,(\Omega_k - \bar{\Omega}_k)^2 = \frac{1}{n-1} \sum_{j=1}^{j=n} (\Omega_{kj} - \bar{\Omega}_k)^2 \tag{7}$$

where $k = 1,2,\ldots i \ldots \mu$ and $m = 1,2,\ldots i \ldots \mu$.

Finally, the mean value and standard deviation of y at the i^{th} increment become

$$\bar{y}_i = \sum_{\ell=1}^{\ell=i} E(\Omega_\ell)\, \Delta x_\ell \tag{8}$$

$$\sigma(y_i) = \sqrt{\sum_{m=1}^{m=i} \sum_{k=1}^{k=i} \rho_{km}\, \sigma(\Omega_k)\, \Delta x_k\, \sigma(\Omega_m)\, \Delta x_m} \tag{9}$$

Equations 8 and 9 were applied to the smoothed stress-strain data for Nellis Baseline sand presented in Figures 3 and 4. The resulting curves are shown in Figures 6 and 7. Each figure contains the mean response with its one-standard-deviation bounds.

IV. ACKNOWLEDGEMENT. The work reported herein was conducted at the U.S. Army Engineer Waterways Experiment Station under the sponsorship of the Defense Nuclear Agency (DNA) under Task Code RSRB, Work Unit 00040: "Probabilistic Constitutive Model." The permission from DNA and the Office, Chief of Engineers, to publish this paper is gratefully acknowledged.

V. REFERENCE.

Baladi, G. Y., and Barnes, D. E. 1983. "An Objective Waveform Comparison Technique," Technical Report SL-83-4, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

VERTICAL STRESS, MPa

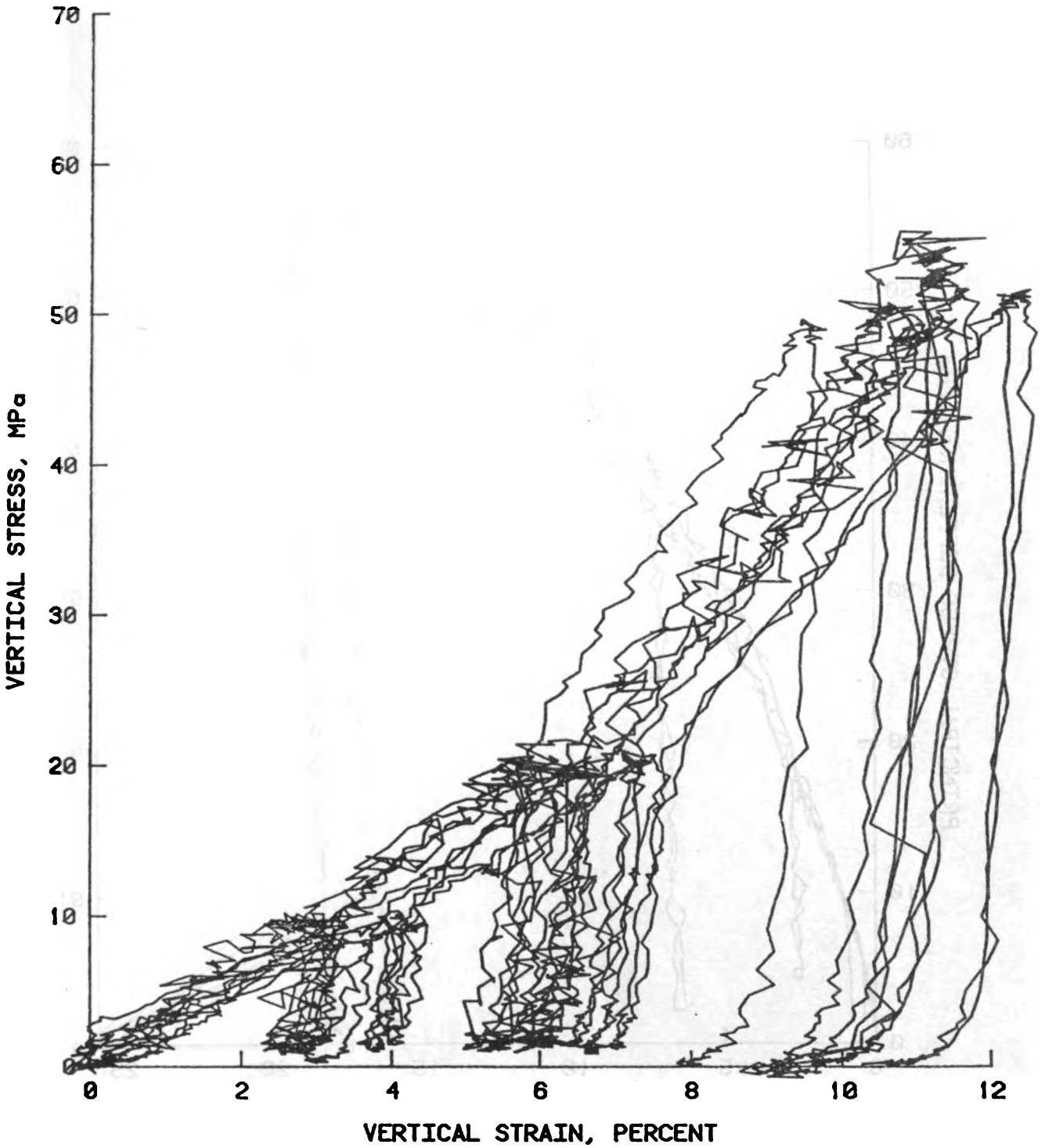


Figure 1. Uniaxial strain test results for Nellis Baseline sand.

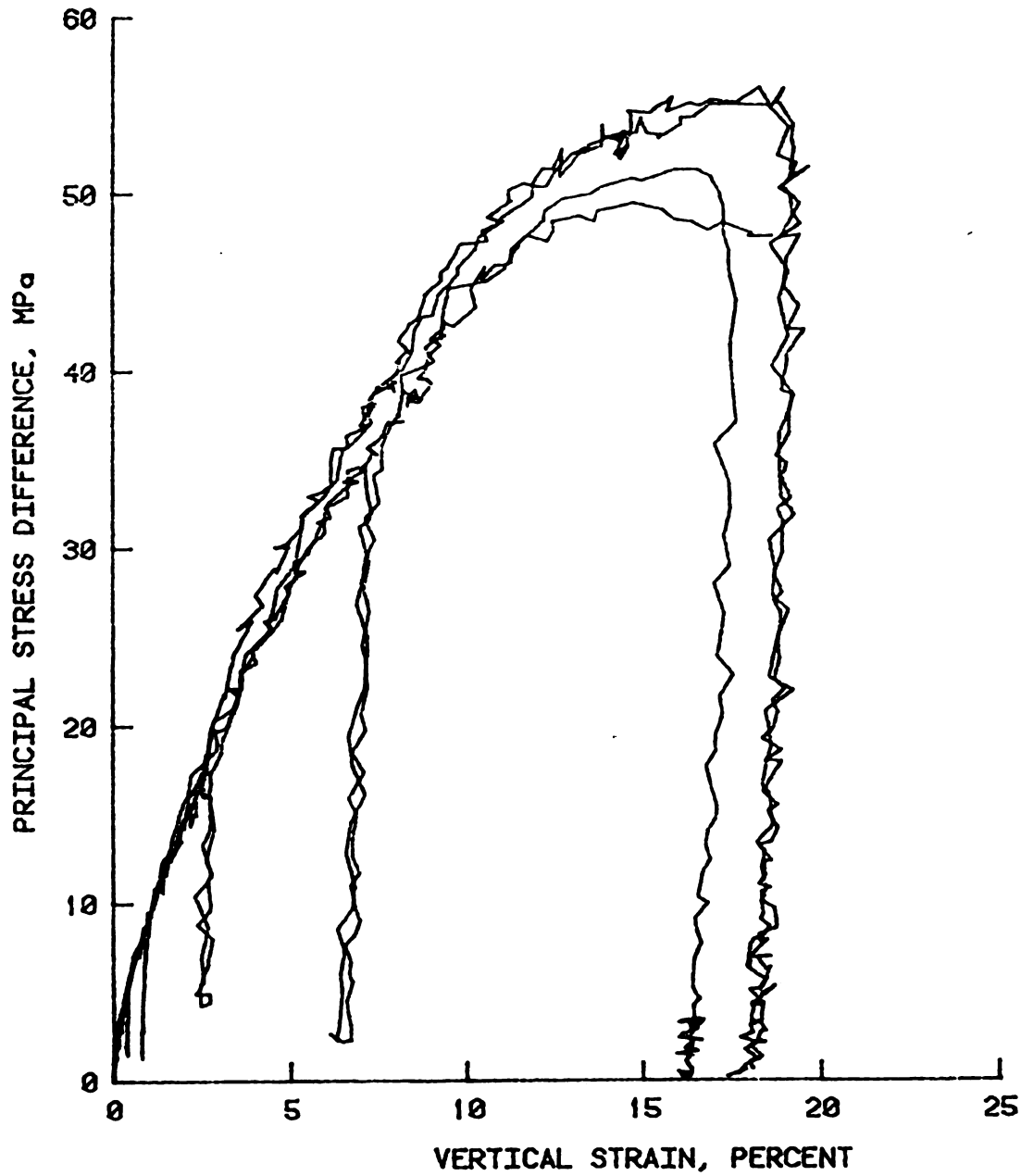


Figure 2. Triaxial compression test results (σ_r = 20 MPa) for Nellis Baseline sand.

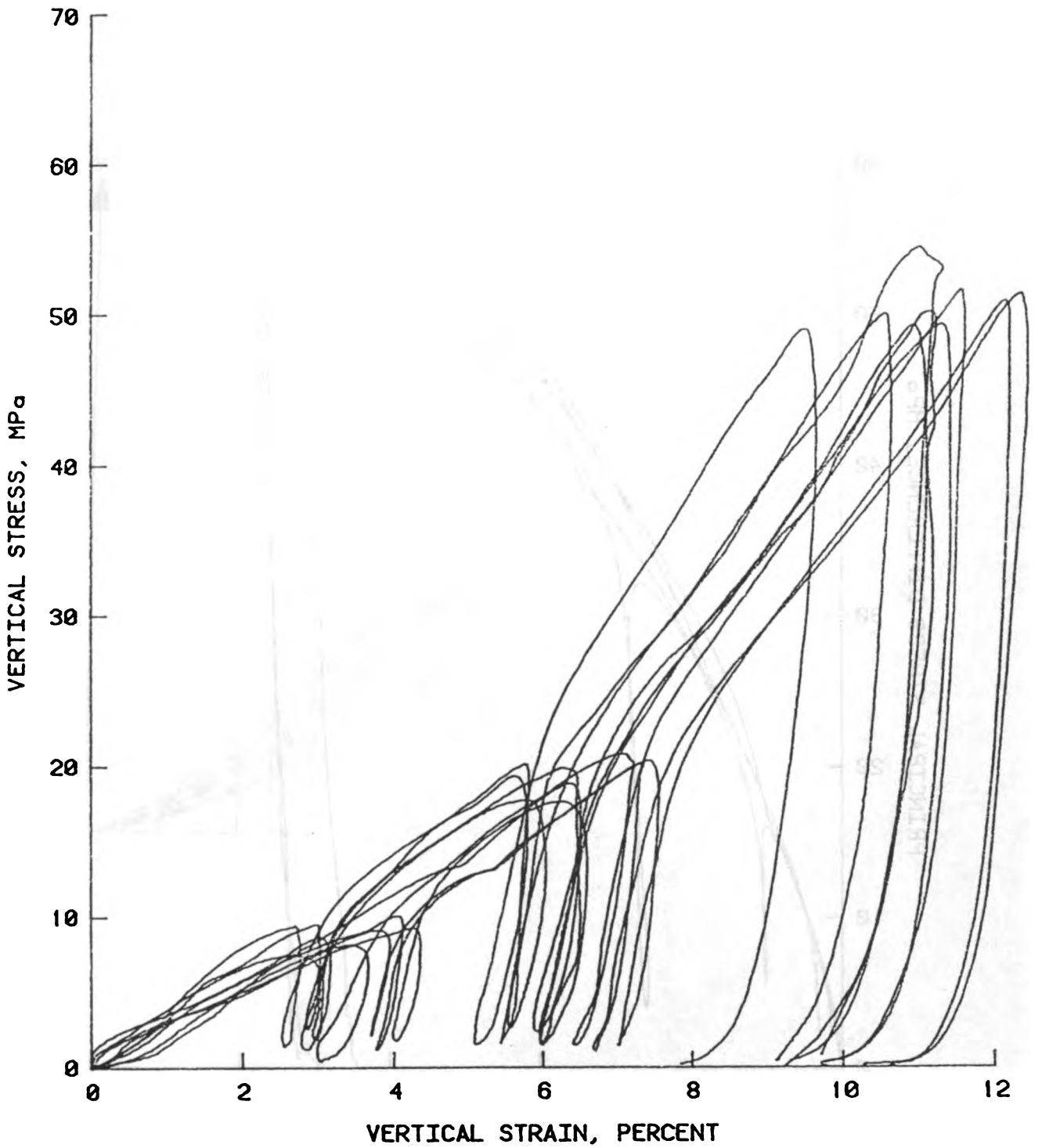


Figure 3. Smoothed uniaxial strain test results for Nellis Baseline sand.

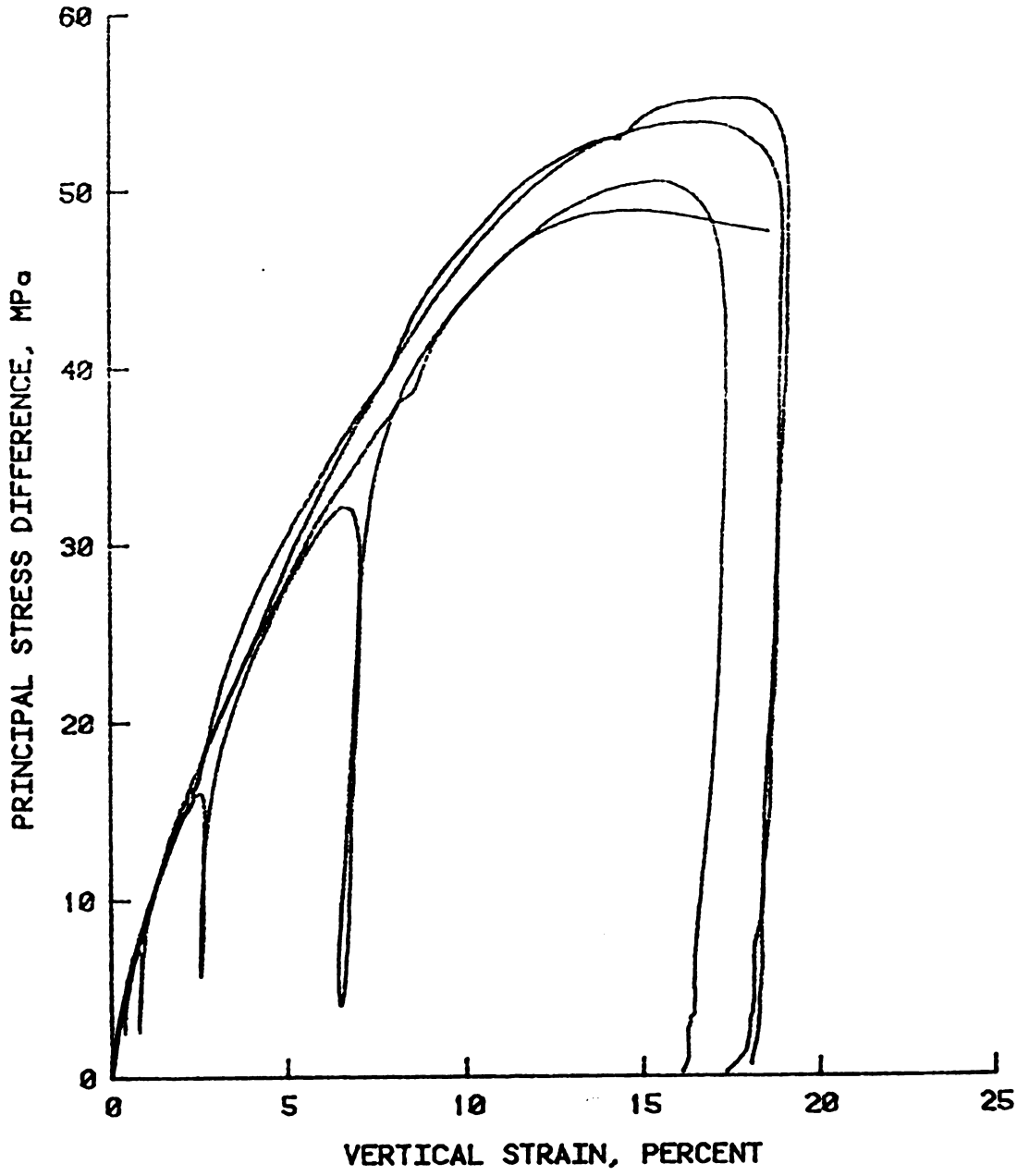


Figure 4. Smoothed triaxial compression test results (σ_r = 20 MPa) for Nellis Baseline sand.

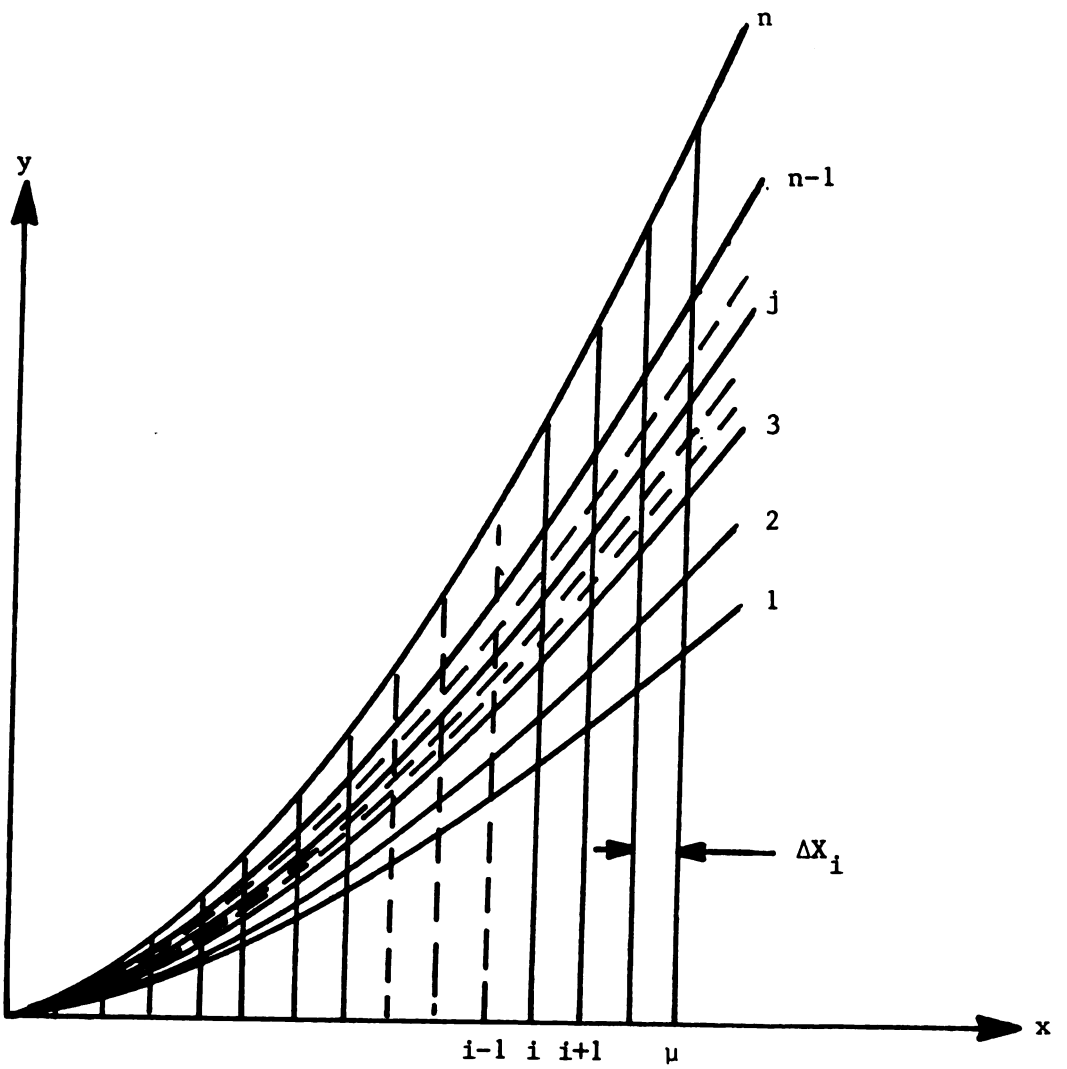


Figure 5. General curves relating the random variables y and x .

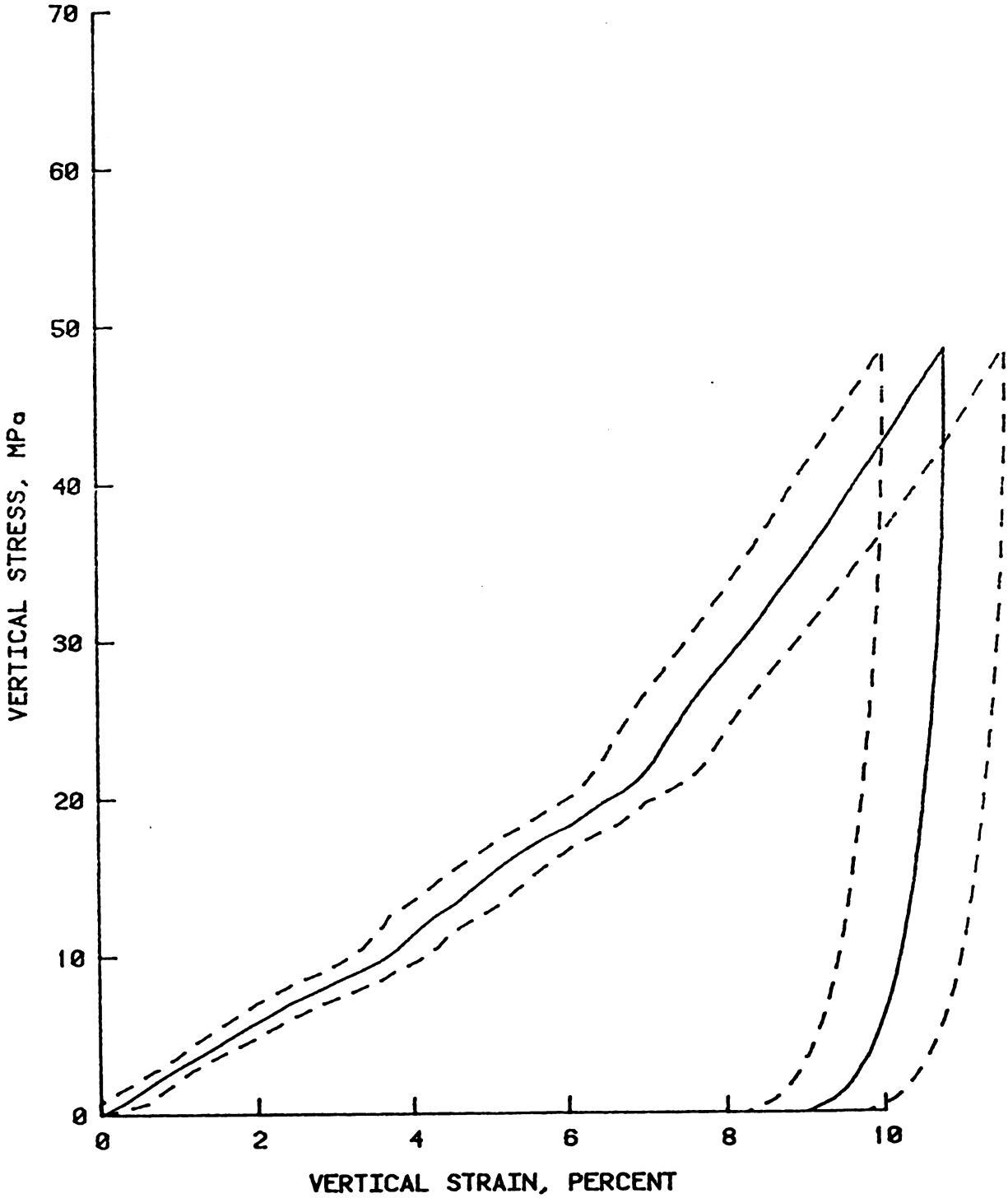


Figure 6. Uniaxial strain test results for Nellis Baseline sand; mean response with its one-standard-deviation bounds.

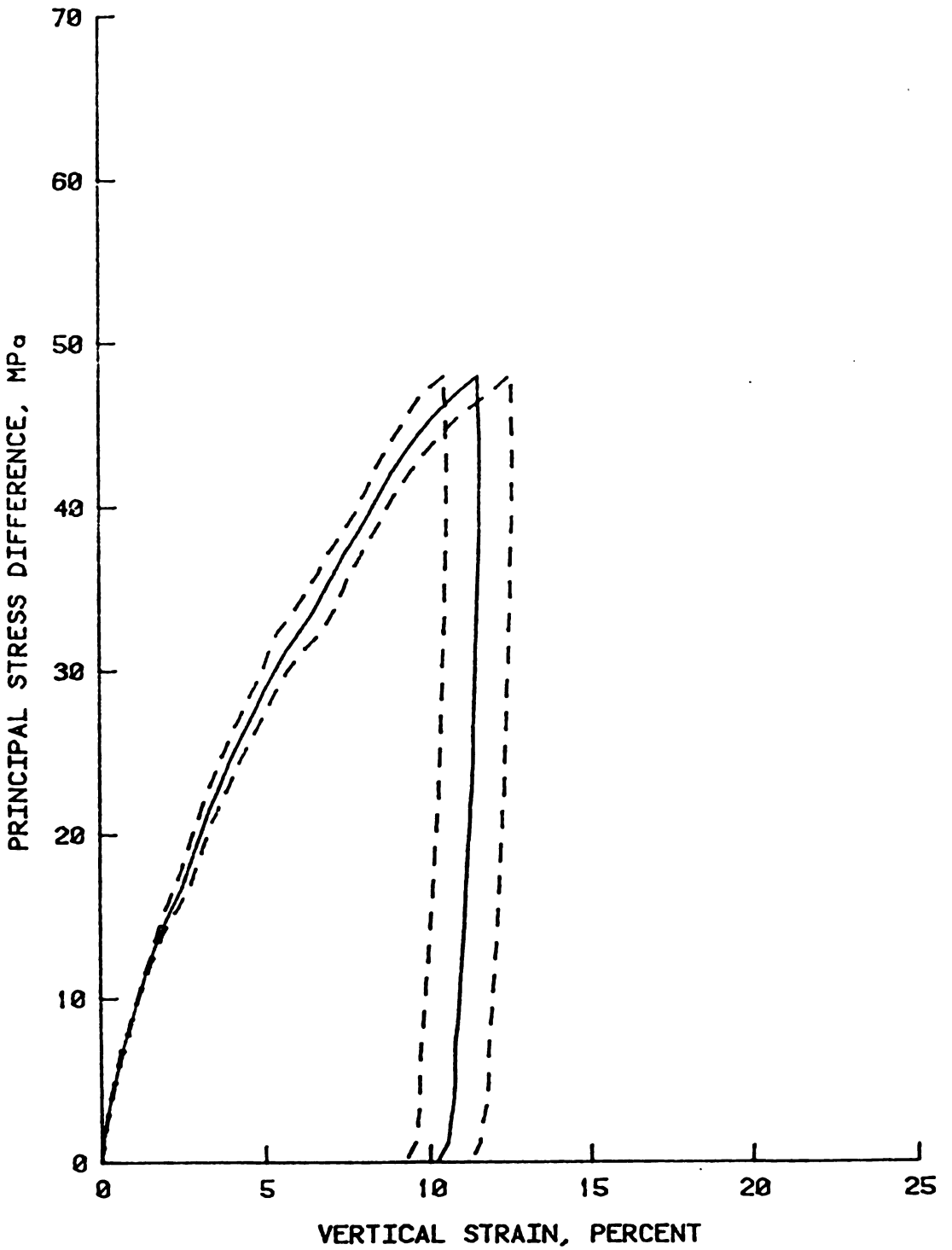


Figure 7. Triaxial compression test results (σ_r = 20 MPa) for Nellis Baseline sand; mean response with its one-standard-deviation bounds.

COMMENTS BY PANELISTS DR. KAYE BASFORD AND PROFESSOR W. T. FEDERER ON THE FOLLOWING ARTICAL

A Method for the Statistical Analysis of the Stress-Strain Properties of Earth Materials by

G.Y. Baladi
and B. Rohani.
U.S. Army Engineer Waterways Experiment Station

W.T. Federer: Before any comments in depth could possibly be made, a copy of the paper and discussions with the experimenter, would be required. From listening to the lecture and pondering on the topic, it would appear that an explosion creates a spherical shock wave effect with only the radius of the sphere being a random variable. The above would follow for one medium such as air or water. However, when a second medium is encountered the radius of a sphere changes. That is, the sphere for air is not the same as for water.

The real problem had to do with an air burst's effect on underground structures. The heterogeneity of the soil meant that several media were being encountered. This adds considerably to the complexity of the problem. It would appear that concentrating on the radii and confidence intervals for radii in various media would simplify the problem. If the bursts were directional, then other regular figures such as an ellipsoid would need to be considered. The shape of the burst would determine which measurement should be used. Hence, more emphasis on the shape of bursts in various media should be made. The statistical problem is usually simplified when the model structure is completely specified. Then, for a given number of media (e.g. sand, clay, loom, rocks, etc.) in a given proportion, confidence intervals could be constructed.

SOME APPLICATIONS OF BAYESIAN IMAGE ANALYSIS

Stuart Geman [1]
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912/USA

The various tasks of image processing, such as removing blur, finding boundaries, and detecting objects, have traditionally been approached on a case-by-case basis. The result is a spectrum of ad hoc techniques. The author and his colleagues are trying to develop a coherent mathematical foundation that will support a variety of these tasks, ranging from problems in "low level vision", such as noise removal, to problems in "high level vision", such as scene segmentation and analysis. The framework is *Bayesian:* probabilistic image models are constructed. These are probability distributions jointly on picture element grey-levels, locations of edge elements, placements and types of textures, and other image attributes as may be appropriate in a particular application. Markov random fields (equivalently, Gibbs distributions) are especially apt and convenient for representing real-world prior knowledge about these attributes. The end product of the formulation is a *posterior distribution,* on the uncorrupted grey-levels, locations of edges, texture labels, and so-on, given an observed and possibly degraded picture. Image restoration and analysis amount to the identification of the mode (or sometimes the mean) of this posterior distribution.

The approach is implemented in four steps. Each step will be discussed in detail, highlighting the important theoretical issues. These steps are:

1. Construction of a prior distribution. The result is a probability distribution, $\pi(\vec{x})$, where the components of $\vec{x}$ represent picture element grey levels, locations and orientations of edges, types and locations of textures, labels and locations of objects, and other image attributes relevant to the image processing task. The dimensionality is very high, in the order of 10^5 or 10^6. This prior distribution is a Markov random field, and is constructed to be consistent with prior information about such things as the spatial smoothness of the image intensity levels, the tendency of textures to appear in homogeneous patches, and so-on. This construction is greatly facilitated by the equivalence between Markov random fields and Gibbs distributions; the Gibbs representation is well-suited for accommodating the various types of prior knowledge in a consistent manner.

2. Modelling of the Degradation Mechansim. The *observation,* $\vec{y}$, is some degradation of the ideal image, $\vec{x}$. The degradation may, for example, involve an attenuated Radon transform, as in tomography, or a blur and noise process, as in satellite or infrared imaging. Or, it may simply be a projection, as in the problem of boundary finding or object identification: we model the degradation as "hiding" the boundary locations or the object labels. Modelling the degradation amounts to specifying the conditional distribution, $\pi(\vec{y}|\vec{x})$, on the observable process, $\vec{y}$, given the ideal (and unknown) image $\vec{x}$.

3. Identificaiton of the Posterior Distribution. This is simply a matter of applying Bayes' rule to $\pi(\vec{x})$ and $\pi(\vec{y}|\vec{x})$ to derive $\pi(\vec{x}|\vec{y})$, the posterior distribtuion on the ideal image given the observable process $\vec{y}$.

4. Identification of the Mode or Mean of the Posterior Distribution. This cor-

[1] Research partially supported by Army Research Office contract DAAG29-83-K-0116, National Science Foundation grand DMS-8352087, and the General Motors Corporation.

responds to image restoration and analysis. If, for example, $\vec{x}$ involves such "high-level" attributes as texture and object labels, then identifying the mode of $\pi(\vec{x}|\vec{y})$ corresponds to choosing the most likely *interpretation,* in the sense of texture and objects identification, given the observed process $\vec{y}$. The posterior mean is computed by a hgihly parallel algorithm called *stochastic relaxation.* This is a Monte Carlo technique that yields an ergodic Markov process, $\vec{x}(t)$, with equilibrium distribution $\pi(\vec{x}|\vec{y})$. The mode can be found by a variation called *simulated annealing,* which can be shown to converge (weakly) to a global maximum of $\pi(\vec{x}|\vec{y})$.

The utility of the approach has been demonstrated by the results of experiments with real scenes. These illustrate: (1) boundary detection; (2) texture segmentation and labeling; and (3) single photon emisiion tomography. Details of these experiments, together with theoretical results on parameter estimation for the prior, and on convergence of stochastic relaxation and simulating annealing, can be found in the following references. These contain, as well, discussions of the many contributions made by by other authors to the Markov random field/Bayesian framework for image analysis.

REFERENCES

1. D. Geman, S. Geman, and C. Graffigne, "Locating texture and object boundaries," *Pattern Recognition Theory and Application,* Ed. P. Devijver, NATO ASI series, Springer-Verlag, Heidelberg, 1986.

2. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern Anal. Machine Intell., 6, 721-741, 1984.

3. S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," *Proceedings of the International Congress of Mathematicians 1986,* Ed. A.M. Gleason, American Mathematical Society, Providence, 1987.

4. S. Geman and D.E. McClure, "Bayesian image analysis: an application to single photon emission tomography," *1985 Proceedings of the American Statistical Association. Statistical Computing Section,* 1985.

5. B. Gidas, "Non-stationary Markov chains and convergence of the annealing algorithm," J. Stat. Phys., 39, 73-131, 1985.

8. B. Gidas, "Parameter estimation for Gibbs distributions," Division of Applied Mathematics, Brown University, 1986. (preprint)

8. B. Gidas, "Convergence of maximum-likelihood and pseudo-likelihood estimators for Gibbs distributions," *Proceedings of the Workshop on Stochastic Differential Systems with Applications in Electrical/Computer Engineering, Control Theory, and Operations Research,* Institute for Mathematics and its Applications, University of Minnesota, Springer-Verlag, New York, 1987. (to appear)

9. U. Grenander, *Lectures in Pattern Theory,* Volumes I, II, III, Springer-Verlag, New York, 1976.

10. U. Grenander, "Tutorial in pattern theory," Division of Applied Mathematics Brown University, 1983.

An Algorithm For Diagnosis of System Failure*

Robert L. Launer
U. S. Army Research Office.

1. Background.

In this note, the optimal diagnosis of system failure is considered. Suppose that there is a system of n components $C_1, C_2, \ldots, C_n$, and that this system becomes inoperable or fails when any one of the components fails. In order that this problem be well posed, the term "component" may also represent a subsystem of units operating in parallel so that subsystem failure occurs when all of the components in that subsystem fail.

The problem considered here is that of finding the failed component or components in the least possible time, or cost when a system failure occurs. The testing will be conducted one component at a time, initially. The more general case will be considered later. Since the testing sequence will be based on probabilistic information, the component reliabilities (or equivalently the failure rates) and the average time (or cost) to test each of the components are assumed to be known.

Wong [3] considered this problem of finding a (single) malfunctioning component "such that the expected test time is optimal in the sence of Bellman's principle of Optimality." The main result of that paper is that "the minimum number of test points required for conclusive detection of system failure is equal to the total number of terminal test points; this set of points constitutes the optimal choice." No algorithm for sequencing the components for achieving optimality is presented in this paper. It is pointed out, however, that the "optimal strategy .. proceeds with the most unreliable and the least test time component .. as the first component to be tested; next in the sequence .. is the next most unreliable and costly component. ..between the last two components, an optimal strategy always chooses the one having a smaller test time regardless of their reliability data."

In the present paper, a precise sequencing algorithm is developed and presented. The problem is also generalized by considering multiple failures, subsystem testing, and the idea of allowing a time for testing a component that has failed which differs from the time to test when it has not failed. The overall goal is to develop sequences which minimize the expected value of the testing time or cost for the several testing situations considered.

---

*The author of this paper presented it at the 31st Conference on the Design of Experiments.

2. One Component at a Time Testing.

Let $R_i(t)$ be the reliability function of the i-th component at standard use conditions. Let T_i and T_i' represent the time to test the i-th component when it is operable or failed, respectively. It will be assumed that the components fail independently of one another. If the reliability functions are continuous, then the probability of more than one failure occuring at time t is zero. This, of course, excludes catastrophic failures from externally imposed destructive forces or other common-cause failures. Nevertheless, multiple failures will also be discussed.

The probabilities of component failure given system failure are obtained as follows. The probability of component i surviving until time t is $R_i(t)$. This may be computed explicitly from $R_i(t) = \exp(-\int_0^t h(u)du)$. Let S_t represent the event that the system does not survive beyond time t, and C_i the corresponding event for component C_i. Then the following equality involving conditional probabilities holds:

$$P[S|C_i]\ P[C_i] = P[C_i|S]\ P[S]$$

From the previous assumption about component failures, the system fails when any component fails so that $P[S|C_i] = 1$, and

$$P[C_i|S] = P[C_i] \ / \ P[S]$$

This is the conditional probability of the failure of component i given system failure (at time t). Let this probability be denoted by p_i. Then from the previous assumptions it follows that

$$p_i = (1-R_i) \prod_{j\neq i} R_j \ / \sum_{k=1}^{n} (1-R_k) \prod_{j\neq k} (1-R_j) \qquad (1)$$

Suppose that the system has failed and that the components are tested one at a time in the order 1, 2, 3, ... until the defective component is found at which time testing is terminated. The initial ordering of the components is arbitrary. The expected test time, E, is then

$$E = T_1' p_1 + \sum_{k=2}^{n} \{[\sum_{i=1}^{k-1} T_i + T_k'] \prod_{i=1}^{k-1} (1-p_i) p_i\} + (\sum_{i=1}^{n} T_i)(\prod_{i=1}^{n} (1-p_i)) \qquad (2)$$

Let E' represent the expected test time when the order of the k-th and the (k+1)-th components are interchanged and all others remain the same. The difference E' - E is easily seen to be,

$$E'-E = p_k p_{k+1} \prod_{i=1}^{k-1} (1-p_i)[(T_{k+1}/p_{k+1})-(T_k/p_k)-(T_{k+1}-T'_{k+1})+(T_k-T'_k)] \qquad (3)$$

The expected testing time is decreased by this permutation if $E'-E$ is negative. Using a finite induction argument, then the optimal ordering is found by computing the n quantities,

$$G_k = (T_k/p_k)-(T_k-T'_k) \qquad (4)$$

for each component and order the G_k beginning with the smallest and ending with the largest. Examination of the first and last terms in E indicate that the ordering scheme (4) also applies to these terms.

The optimal expected test time is obtained from (2) with the terms arranged in the optimal order, but without including the last term since it would be unnecessary to test the "last" component if the other n-1 were tested and found to be operative.

Notice that if the terms T_k and T'_k are equal, then the G_k are easily seen to correspond to the intuitive feeling that the components with shorter testing time and higher failure probabilities should be tested first generally.

3 Multiple Failures With One At A Time Testing

The case of multiple failures is considerably more complicated than the single failure case. There is first of all the problem of determining the multivariate failure law, which would yield the conditional failure failure probabilities corresponding to (1) in the simpler case. The derivation of this set of probabilities should be based on the physics of the particular situation. In the absence of specific information one might use compound probabilities.

Another complicating factor is how testing and repair is to be conducted. If all of the failed components are to identified before any repair begins, then exhaustive testing would be implemented in which case the testing sequence is irrelevant. If, however, testing proceeds one component at a time until a failed one is found, followed by immediate repair of that component with further testing following the repair only if it is required, then the testing sequence is important. The following development treats the latter case.

Assume for the moment that the system in question is known to contain exactly $m < n$ failed components. The expected repair time for this case can be written explicitly. It can be analyzed similarly to (2). The result has

been worked out for the cases m=2 and m=3 and may be described in the following way. The testing order of the first m components does not effect the testing time. The remaining n-m components should be tested in the order dictated by (4). The general case was not worked out because of the inordinate amount of algebra involved. The lower order cases indicate no surprises for the higher order ones.

The point of this discussion is that if, unknown to the tester, the system contains more than one failure, the procedure given by (4) will still result in an optimal or near optimal sequence if continued testing is indicated by system malfunction after the first failed component has been found and repaired. Naturally, system "turn-on" after repair could induce a failure among the previously tested components. Without appropriate data or probabilistic information about this phenomenon, no definitive guidance can be given about optimal or reasonable strategies to protect against it.

4. Subsystem Testing

It seems reasonable to ask what further saving in testing time can be realized by simultaneously testing components in groups if that is possible. For example, if half of the components in a system could be tested together in a reasonable period of time, followed by testing smaller subgroups or single components when appropriate, it would appear that the expected testing time could be further reduced, especially if only one component has failed.

Assume that the system in question yields a natural decomposition into M subsystems or modules $M_1, M_2, \ldots M_m$. Module k consists of n(k) components, and its reliability is given by Q_k. The average time to test module k as a single entity (that is, exclusive of any component testing) is U_k if it is operational and U_k' if not, while the corresponding average times for component j of the k-th module are T_j^k and T'^k_j. The probability that module k has failed given system failure, P_k, is

$$P_k = (1-Q_k) \prod_{j \neq k} Q_j \Big/ \sum_{i=1}^{M} (1-Q_i) \prod_{j \neq i} Q_j \qquad (5)$$

The probability that component j has caused module k to fail is given by (1) where the p_i, T_i and T_i' are restricted to the components of module k.

Corresponding to the previous testing set-up, it will be assumed that testing proceeds one module at a time until the failed module is discovered.

Then the individual components are tested one at a time until the failed component is found.

Let $\underline{U}_m$ represent the vector $(U_1, U_2, .., U_{m-1}, U_m')$, $\underline{T}_j^m$ represent the vector $(T_1^m, T_2^m, .., T_{j-1}^m, T_j^m)$ and $\underline{1}_k$ the k-vector each component of which is a 1. Note that the transpose of a matrix or vector will be denoted by a superscript T. Further, let $\underline{M}_m$ represent the event that, given system failure, modules 1 through m-1 were found to be operational and module m was diagnosed as failed. Let $\underline{C}_j^m$ represent the event that, given failure of module m, components 1,2,..j-1 were found not to have failed and component j was diagnosed as failed.

Then for an arbitrary ordering of modules and components, the expected testing time E is

$$E = \sum_{m=1}^{M} P[\underline{M}_m] \sum_{j=1}^{n(m)} [\underline{1}_m^T \cdot \underline{U}_m + \underline{1}_j^T \cdot \underline{T}_j^m] \; P[\underline{C}_j^m | \underline{M}_m] \qquad (6)$$

If the m-th and (m+1)-th modules are interchanged, and the quantity $E-E'$ is computed as was done in section 2, the minimizing algorithm is obtained. That is, the quantities H_j are obtained:

$$H_j = E_j + (U_j' - U_j) + U_j / P_j \qquad (7)$$

where E_j represents the average time to complete the one at a time testing of the components in module j. The algorithm indicates that optimization of modular testing depends on the optimization of the component-wise testing within each module, but both optimizations are obtained independently of the other.

The optimizing algorithm is therefore to order the components in each module according to (1) within the module, and then to compute the H_j for each module, j=1,2,...,M. The testing proceeds by diagnosing the module corresponding to the smallest value of the H_j, followed by the module corresponding to the next smallest value of H_j, and so on to the module corresponding to the largest value last.

It is useful to ask when component-wise testing within a given module is more efficient than modular testing for that module. A good rule of thumb is to use that method which requires the lesser overall average testing time. This leads to the following algorithm. If the following inequality (8) holds

then use modular testing.

$$P_j U_j' \leq (1-P_j)\left(\sum_{i=1}^{n(j)} T_i - U_j\right) \qquad (8)$$

It should be pointed out that there are several loose ends related to this discussion which should be kept in mind. First, the given algorithms are optimal in the sence of lowest expected testing time under certain restricted conditions. The most general test situation would allow for an unrestricted mix of any combination of single components and subsystems, whether these are natural subsystems or not. There are 2^n-1 such possible subsets to consider in various combinations. This would require a prohibitively large amount of computer time for even a moderately small system.

Another possible area of further exploration involves the computation of the component reliability functions at every new failure. Unless the failure rates are unusually well behaved, such as all constant failure rates, the quantities G_i and H_i must be recomputed at each failure. With constant failure rates for example, the crossings of the reliability functions could be computed once and for all yielding a set of $(n(n+1)/2)$ time zones of consideration.

Finally, there is the question of data and prior or partial information. The R_i, T_i and so forth, might not be known exactly. Moreover, if other prior information is available, it certainly should be incorporated into the analysis.

REFERENCES

[1]. Barlow, R. E. , and F. Proschan, Statistical Theory of Reliability and Life Testing: Probability Models; 2nd Edition; Holt, Rinehart & Winston, 1975

[2]. Boggs, Paul T. and Robert L. Launer, Time-Optimal Rejection Sequencing; Transactions of the Twenty-Fourth Conference of Army Mathematicians

[3]. Wong, James T., An Optimal Diagnostic Strategy for Finding Malfunctioning Components in Systems; NASA Tech Memo 84335 (USAAVRADCOM 83-A-7), Mar 1983

INDIVIDUAL VERSUS GROUP SAMPLING*

Paul A. Roediger

John G. Mardo

U.S. Army Armament, Munitions and Chemical Command

Dover, New Jersey 07801

ABSTRACT: Lot acceptance based on INDIVIDUAL sampling has been widely used during the past decade. Recently it was recommended that this practice be discontinued and future sampling be done on a GROUP basis. The need for specific conversion guidance and procedures was thereby created A model assuming the family of negative log gamma distributions on incoming INDIVIDUAL quality rates has been developed for the purpose of selecting the GROUP plan most comparable to a given INDIVIDUAL plan. In addition to the model details, examples are presented and a previously published alternative is discussed.

1.0 INTRODUCTION

In lot-by-lot attributes sampling inspection, product is divided into inspection lots and random samples are drawn from each. We assume there are m quality characteristics each having a well-defined attribute requirement, i.e., a requirement which is either met or is not. A unit not in conformance with the j-th requirement is called a j-type defective. A non conforming unit

*The authors of this paper presented it at the 31st Conference on the Design of Experiments.

with respect to one or more requirements is called a defective. Two sampling modes, one based on defectives, called GROUP sampling, and the other based on the m defective types, called INDIVIDUAL sampling, are described. Both are permitted in [1], MIL-STD-105D.

Let d be the number of defectives obtained in a sample of N units. We say that sampling is done in the GROUP mode when the decision rule to accept or reject the lot is based only on d, without further regard to defective types therein. In practice GROUP sampling is implemented by the following

RULE G: ACCEPT LOT IF $d \leq C$, OTHERWISE REJECT.

The numbers N and C are called the "sample size" and "acceptance number" of the GROUP plan. Such plans are denoted by (N,C). Note, the GROUP criterion ignores underlying defective types entirely. For now, "reject" stands for any course of action taken on lots not accepted.

Let p be the true lot fraction defective and q=1-p. Then, the GROUP probability of acceptance (PA_G), in binomial form, of lots of quality q is

$$(1.1) \qquad PA_G(q) = OC(q;N,C) = \sum_{i=0}^{C} \binom{N}{i} q^{N-i}(1-q)^i \ .$$

OC, for convenience treated as a function of q instead of p, is called the Operating Characteristic (OC) curve of (N,C).

The second type of sampling, used in many current U.S. Army commodity specifications, is INDIVIDUAL sampling. Let d_j be the number of j-type defectives found in a sample of n units. In this mode, lot acceptance is based only on the d_j's and is typically invoked via the following

RULE_I: ACCEPT LOT IF EACH $d_j \leq c$, $j=1,2,\ldots,m$, OTHERWISE REJECT.

The numbers n and c are called the "sample size" and "acceptance number" of the INDIVIDUAL plan, which is denoted by $(n,c)^m$. Let p_j be the true j-type defective rate and $q_j = 1-p_j$. The INDIVIDUAL probability of acceptance (PA_I) of lots with quality profile $\tilde{Q} = (q_1, q_2, \ldots, q_m)$ is

$$PA_I(\tilde{Q}) = \prod_{j=1}^{m} OC(q_j; n, c) \quad . \tag{1.2}$$

Note, PA_I is not a function of the one parameter q, as is PA_G, but is instead a product of GROUP-like OC curve terms.

For a given profile $\tilde{Q}$, the overall lot quality, assuming independence among the m defective types, is given by

$$q = \prod_{j=1}^{m} q_j \quad . \tag{1.3}$$

This equation establishes the basic connection between the two sampling approaches, relating q, the GROUP quality of (1.1), with the q_j's, the INDIVIDUAL qualities of (1.2).

The following conversion problem is considered:

PROBLEM "P": GIVEN THE $(n,c)^m$ INDIVIDUAL PLAN, FIND THE "BEST" (N,C) GROUP PLAN REPLACEMENT.

The inverse problem is apparently more complicated, but, in principle, can be back-solved by iteratively solving a converging sequence of problems of the type posed.

The two approaches share a curious history. GROUP sampling, once the authorized method, was eventually replaced by the INDIVIDUAL method. This development was an outgrowth of a computer revolution that helped promote a component oriented approach to system reliability. Subsequent years have seen more than just a balancing of this trend; indeed, a steady return to a more integrated "systems" approach has ensued. With it, interest in GROUP sampling has grown, to the point that direction was recently given in [4] to discontinue the use of INDIVIDUAL sampling altogether. Unfortunately, a sound conversion rationale does not exist. Several sets of tables prescribe GROUP acceptable

quality levels (AQL's) for various m. Most, however, are either statistically unfounded, rely on untenable assumptions or can not be generalized; as such, they have no real bearing upon our problem. The tables of [2] were more disappointing in that more was promised. They are considered below in greater detail.

2.0 BOUNDARY CURVES

Consider the range of $PA_I(\tilde{Q})$ values obtained by varying $\tilde{Q}$, keeping q, as defined in (1.3), constant. The bounds on $PA_I(\tilde{Q}|q)$, calculated in Appendix 1, are given by

$$(2.1) \qquad OC(q;n,c) \leq PA_I(\tilde{Q}|q) \leq \{OC(q^{1/m};n,c)\}^m .$$

We call these bounds L(q) and U(q) respectively. The minimum L(q) is attained when

$$q_j = \begin{cases} q, & \text{if } j=k \\ 1, & \text{if } j\neq k \end{cases} , \text{ for each } k=1,2,\ldots,m ,$$

representing, at the one extreme, profiles where all but one of the incoming defective rates are zero.

The maximum U(q) is attained when

$$q_1 = q_2 = \ldots = q_m = q^{1/m} ,$$

representing the other extreme where all incoming defective rates are equal.

As q is allowed to vary on [0,1], the bounds of (2.1) produce an envelope that contains all possible $PA_I(\tilde{Q}|q)$ values. A sample envelope resulting from the $(125,1)^{14}$ INDIVIDUAL plan is depicted in Figure 1. The importance of the envelope is stated in the following :

CONCLUSION: A CANDIDATE GROUP OC CURVE MUST BE CONTAINED WITHIN THE ENVELOPE. THEREFORE, $L(q) \leq OC(q;N,C) \leq U(q)$.

The envelope collapses if and only if m=1 or c=0. In both cases, INDIVIDUAL and GROUP sampling are identical, provided of course that (N,C)=(n,c).

3.0 MODEL REQUIREMENTS

Where exactly within the envelope should the "best" GROUP OC curve be located? To help guide us, a model is proposed that relies on probability distributions used as weighting functions. The model has two desirable properties: it is general, taking into account important aspects of the problem, yet tractable, allowing computations to be carried out and simulated in terms of known statistical quantities.

The following are utilized as part of the model:
the Beta probability density function (pdf),

(3.1) $$bet(x;a,b) = x^{a-1}(1-x)^{b-1}/B(a,b) ,$$

and the Negative Log Gamma pdf.

(3.2) $$nlg(x;a,b) = a^b x^{a-1}\{\ln(1/x)\}^{b-1}/\Gamma(b) ,$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a-b)$,

and $z>0$, $a>0$, $b>0$ and $0<x\leq 1$.

The cumulative distribution functions (cdf's) obtained by integrating bet(t;a,b) and nlg(t;a,b) with respect to t, $t\in[0,x]$, are denoted by BET(x;a,b) and NLG(x;a,b), respectively.

The negative log gamma density closely resembles the more familiar beta. In fact, (3.2) is obtained from (3.1) by replacing (1-x) with ln(1/x), two nearly equal terms for x close to 1, and adjusting the constant term to normalize the integral. For fixed "a", this family of densities has the special feature of being closed under multiplication, i.e., if $U_i \sim nlg(x;a,b_i)$, i=1,2, then the product $U_1U_2 \sim nlg(x;a,b_1+b_2)$. The family is also a rich one, taking on a wide variety of shapes including the "U", "L" and "J" shaped, uniform and uni-modal densities. Its name is derived from the fact that Y is negative log gamma distributed, if and only if, -lnY is gamma distributed. References [5] and [6] contain more details about this distribution.

4.0 THE MODEL

The model specifies weighting pdf's on the q_i's

(4.1) $f_i(q_i) = nlg(q_i;a,b_i)$, where $a>0$ and $b_i>0$, $i=1,2,\ldots,m$.

In order to randomly generate vectors $\tilde{Q}$ given q, the conditional cdf of an arbitrary q_k, given q and possibly some, or all, of the other q_j's must be determined. The desired cdf's, developed in Appendix 2, are, for $k=1,2,\ldots,m-1$,

(4.2) $$G_k(q_k \mid q, q_1, q_2, \ldots, q_{k-1}) = BET(T_k(q_k); S_{k+1}, b_k) ,$$

where $$T_k(q_k) = \ln(q_k/P_k)/\ln(1/P_k) \quad , \quad P_k \leq q_k \leq 1 ,$$

$$S_j = \sum_{i=j}^{m} b_i \quad \text{and} \quad P_1 = q, \quad P_j = q / \prod_{i<j} q_i \quad , \quad 1 \leq j \leq m.$$

Equation (4.2) provides the basis for the following procedure: for $k=1,2,\ldots,m-1$, take

(4.3) $$q_k = (P_k)^{1-BET^{-1}(R_k;S_{k+1},b_k)} ,$$

where the R_k's are random numbers generated from a uniform distribution on [0,1]. Once the first (m-1) q_k's are generated, q_m is simply P_m.

Implementation of procedure (4.3) allows us to study the distribution of $PA_I(\tilde{Q}|q)$ via Monte Carlo simulation methods.

5.0 PARAMETER SELECTION

The pdf on q resulting from (1.3) and (4.1) is

$$f(q) = nlg(q;a,S_1) \quad . \tag{5.1}$$

An interpretation of "a" is found by taking the expected value of (5.1), giving E(q)=a/(a+1), so that a=E(q)/[1-E(q)], the odds of randomly picking an effective unit when quality is at its average. Consequently, in most practical applications, "a" will be quite large. Note, however, (4.2) and (4.3) are independent of this parameter.

Of particular interest to us are the "J" shaped pdf's that result when a>1 and b=1. Then, (5.1) defines a one-parameter family which is deemed sufficiently rich for the purpose of assigning appropriate weights to q. Having no prior information, the b_i's are assumed to be equal. Since they sum to b=1, $b_i=1/m$ for i=1,2,...,m, yielding J-shaped weights on the q_i's which are asymptotic at $q_i=1$.

6.0 CONVERSION STRATEGY

A three step approach to solving problem "P", fully computerized and documented in [8], will now be described.

STEP 1: Monte Carlo simulation-

(a) Choose K, the number of distinct q's at which to simulate $PA_I(\tilde{Q}|q)$. We take K=19.

(b) Select an appropriate q-interval $[u_1, u_K]$. We utilize the criteria $U(u_1)<.10$ and $L(u_K)>.95$, ensuring that $PA_I(\tilde{Q}|q=u_1)<.10$ and $PA_I(\tilde{Q}|q=u_K)>.95$.

(c) Define the equi-spaced intermediate points $u_{i+1}=u_i+(u_K-u_1)/(K-1)$, for $i=1,2,\ldots,K-2$.

(d) Generate Isum random vectors $\tilde{Q}|q=u_1$, per (4.3). Our simulations utilize Isum=1000 repetitions.

(e) Obtain the empirical density of $PA_I(\tilde{Q}|q=u_1)$.

(f) Compute the 50^{th} percentile, and call it y_1 (y_i if $q=u_i$). Other percentiles, namely .0,.1,.2,.3,.4,.6,.7,.8,.9 and 1.0, are computed and processed as are the medians. However, we do not consider the resulting GROUP plans to be as useful simply because the risks to producer and consumer are unbalanced.

(g) Repeat (d) thru (f) using $u_2, u_3, \ldots, u_K$ instead of u_1.

(h) Obtain $\{(u_i, y_i)\}$, $i=1,2,\ldots,K$.

(i) If the y_i's are increasing, proceed directly to step 2.

If not, and this case has never occurred, either increase the the number of trials Isum, decrease the number of points K, or open up the interval $[u_1, u_K]$.

STEP 2: Interpolation-

(a) Linearly connect the points (u_i, y_i), i=1,2,...,K.

Call this increasing piecewise linear function y=f(q).

(b) Use inverse interpolation to find six unique q values, $\bar{u}_1$ thru $\bar{u}_6$, corresponding to $\bar{y}_1=f(\bar{u}_1)$ thru $\bar{y}_6=f(\bar{u}_6)$, where $\bar{y}_1=.1$, $\bar{y}_2=.3$, $\bar{y}_3=.5$, $\bar{y}_4=.7$, $\bar{y}_5=.9$ and $\bar{y}_6=.95$.

STEP 3: Find the "best" (N,C) approximation-

(a) Define a range [Cmin,Cmax] for C.

We take Cmin=max(0,c-5), Cmax=Cmin+10 and begin the search with C=Cmin.

(b) Permissible (N,C) are required to satisfy

$$OC(\bar{u}_1;N,C)<\bar{y}_1+e_1 \text{ and}$$

$$OC(\bar{u}_6;N,C)>\bar{y}_6-e_6 .$$

where e_1, e_6 are two small positive constants. The use of perturbed values ($e_1, e_6 \neq 0$) helps ensure that the "best" N,C combination is not eliminated at the start of the search. In the terminology of Hald ([3], pp 25), (N,C) is said to be "stronger" than a plan whose OC curve passes thru the two points $(\bar{u}_1, \bar{y}_1+e_1)$ and

$(\bar{u}_6, \bar{y}_6 - e_6)$. We utilize $e_1 = e_6 = .2$.

(c) Find the largest interval I(C) such that the conditions of (b) are met for all $N \in I(C)$. Approximate formulae developed in [3], pp 51, are used to determine the exact interval. If I(C) is empty, proceed directly to (e).

(d) Find the N that minimizes $Del(N,C) = \sum_{i=1}^{5} \left| OC(\bar{u}_i; N, C) - \bar{y}_i \right|$,

for $N \in I(C)$. Call it N_C. Note, Del does not depend on $\bar{u}_6$.

(e) Repeat (b) thru (d) for C=Cmin+1,...,Cmax.

(f) Obtain a final set of candidate plans

$\{(N_C, C) \mid C \in [Cmin; Cmax],\ I(C) \neq \phi\}$

(g) Find the C that minimizes $Del(N_C, C)$,

for $C \in [Cmin, Cmax]$, $I(C) \neq \phi$. Call it C^*.

(h) Obtain the "best" GROUP plan (N,C), namely (N_C^*, C^*).

7.0 EXAMPLES AND DISCUSSION

The above three step procedure will be designated method B (for "best"). The "best" (N,C) will be called the B-plan.

Several examples provide a setting for our discussion of method B.

First, consider P_1 : Given $(n,c)^m = (125,1)^{14}$,

Find the "best" (N,C).

Obtain its B-plan : (N,C) = (94,1).

A partial summary of simulation data (steps 1h and 2b) along with approximating B-plan OC curve values are presented in Table 1.

Table 1. B-plan for P_1

q	Median $PA_I(\tilde{Q}\|q)$	B-plan OC(q;94,1)	q	Median $PA_I(\tilde{Q}\|q)$	B-plan OC(q;94,1)
.9525	.04728	.058647	.9800	.43265	.436909
.9550	.06110	.071625	*.9822	.50000	.499582
.9575	.07410	.087243	.9825	.50934	.508711
.9600	.09319	.105966	.9850	.59520	.587320
*.9608	.10000	.112802	.9875	.67824	.671285
.9625	.11415	.128314	*.9881	.70000	.691709
.9650	.14505	.154860	.9900	.77007	.757932
.9675	.17526	.186224	.9925	.85072	.842845
.9700	.20826	.223054	*.9942	.90000	.894978
.9725	.25857	.265998	.9950	.92500	.919145
*.9747	.30000	.308413	*.9962	.95000	.948909
.9750	.30661	.315661	.9975	.97897	.976533
.9775	.36534	.372537			

* Interpolated values

Method B has been designed specifically to be a fair conversion strategy, suitable to both producer and consumer. This intention is particularly reflected in

Step 1f: Skewness in the simulated data convinced us that the B-plan should approximate the set of median, not mean, PA_I values. As such, the B-plan rejects more often than the INDIVIDUAL plan, for half of the profiles $\tilde{Q}$ considered in the simulation, and accepts more often for the other half. In this sense the producers and consumers risks associated with the conversion are equalized.

Steps 1a thru 1c: A fair GROUP plan should provide close approximation throughout the low, middle and high range of

median PA_I values. Our choice of q-interval and fine discretization of it into K=19 equi-spaced points ensures that the simulated median PA_I's will cover the entire spectrum of interest.

Step 2: A fair GROUP plan should also give equal consideration to the low, middle and high range of median PA_I values. Unfortunately, it is not possible to choose the u_i's in advance so as to get a balanced set of median PA_I values. For example, the raw (un-interpolated) data of Table 1, with almost half (9/19) of its simulated medians below 0.25, is considerably biased toward low PA_I values. Were a GROUP plan fitted to the raw data, a (95,1) B-plan would result, producing a fit that is slightly better than (94,1) in the low PA_I range, but worse elsewhere. For this reason, the B-plan has been based on $\bar{u}_i$'s corresponding to the more balanced set {.1,.3,.5,.7,.9} of interpolated median PA_I values. The insensitivity as to which data base is used, raw or interpolated, is typical and reassuring.

Step 3d: Del, the sum to be minimized, attaches equal weight to the approximation's lack of fit at each interpolated data point.

The set of candidate plans, including the (94,1) B-plan, along with their scores Del (step 3d), are presented in Table 2.

Table 2. Candidate B-plans for P_1

C	N_C	Del	C	N_C	Del
0	39	.35901	5	318	.41117
1	94	.03495	6	374	.46023
2	150	.14799	7	430	.50000
3	206	.26565	8	486	.53313
4	262	.34854	9	543	.56060

The N_C and C of Table 2 are highly correlated, with r=.999996 and regression N_C = 56C+38.2. In general, candidate B-plan OC curves are naturally forced to pivot about ($\bar{u}_3$,.5), the fixed "indifference point" (IP) determined in step 2b. How closely the OC curves approximate the IP depends on the other $\bar{u}_i$'s, especially when C is small, but their effect diminishes rapidly as C increases. Based on the IP only, Hald shows in [3], pp 195, that $N_C \sim aC+b$, where $a=1/(1-\bar{u}_3)$ and $b=(1+\bar{u}_3)/(3-3\bar{u}_3)$. By taking $\bar{u}_3$=.9822, a=56.18, b=37.12 and rounding to the nearest integer, the N_C's of Table 2 are duplicated, except when C=0,1 and 2, where you get 37, 93 and 149 respectively. This correlation can be exploited to economize the search for candidate plans, but, depending only on $\bar{u}_3$, does not constitute per se a reliable shortcut approach.

Before other examples are presented, an alternative conversion method forming the basis of [2] is described. It consists of taking N=n and letting C be the smallest X satisfying

$OC(AQL^m;N,X) \geq OC(AQL;n,c)$, where n,c and m are specified and AQL is defined by $OC(AQL;n,c) = .95$ (or .90). The intent here is to accept, with high probability, incoming product whose quality characteristics are all at AQL. From a consumer point of view, such an approach is intuitively unacceptable. The version proposed in [2], designated here as method A (for "alternative"), limits (n,c) and (N,C) to be MIL-STD-105D plans, and utilizes tabulated AQL values instead of exact ones. (N,C) determined in accordance with method A will be called an A-plan.

Table 3 presents sample conversions obtained by the two methods, for seven INDIVIDUAL plans having nominal AQL's of .996 (.4% AQL in [1]), at two values m=3 and m=14.

Table 3. Comparison of A,B-plans

	m=3		m=14	
	A-plan	B-plan	A-plan	B-plan
n -c	N -C	N -C	N -C	N -C
32-0	32-2	32-0	32-5	32-0
125-1	125-5	104-1	*125-14	* 94-1
200-2	200-5	160-2	200-21	140-2
315-3	315-7	247-3	↓	213-3
500-5	500-10	389-5		329-5
800-7	800-21	624-7		519-7
1250-10	1250-21	973-10	V	804-10

* A,B-plans for P_1, also depicted in Figure 1

Method B results imply that producer and consumer risks are more naturally balanced by taking C=c and N<n, than taking N=n and C>c, as suggested in [2]. Also, method B conversions produce average sample reductions (c>0) of 20% and 30% in the m=3 and m=14 cases respectively, as compared to no anticipated method A reductions, except those incidental to the tabular limitations of

[1], viz., the repeating (200,21) A-plan. Note also that method A misses the only sensible conversion when c=0, namely (N,C)=(n,c).

Figure 1 compares the A and B-plan OC curves of Table 3 (*), showing them in relation to the envelope [L(q),U(q)] defined by (2.1).

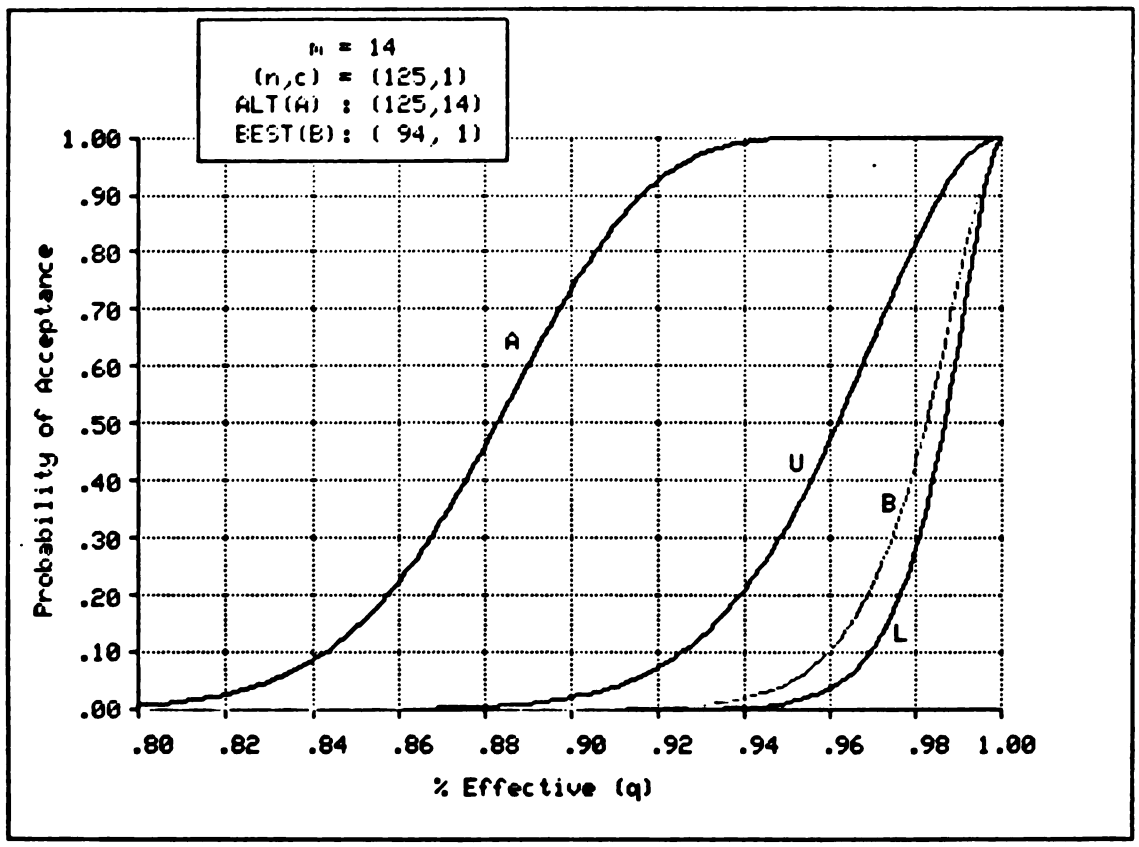


Figure 1. A,B-plans for P_1, with envelope [L,U]

The A-plan for P_1 is highly inconsistent with the original INDIVIDUAL plan: as such, method A is not a viable alternative.

Finally, Table 4 shows how B-plans change with m, for two different (n,c) INDIVIDUAL plans.

Table 4. B-plans, varying m

	m=2	m=3	m=5	m=7	m=10	m=15	m=20
n -c	N -C	N -C	N -C	N -C	N -C	N -C	N -C
50-1	44-1	42-1	40-1	39-1	38-1	38-1	37-1
315-3	271-3	246-3	227-3	222-3	218-3	213-3	210-3

As m increases, INDIVIDUAL sampling loosens and becomes less discriminating: with C=c, GROUP sampling can mimic this behavior by decreasing N from n to c as m goes from 1 to infinity.

8.0 CONCLUDING REMARKS

A general model has been described that allows one to simulate important performance measures of INDIVIDUAL sampling for a fixed q, e.g., PA_I. We have utilized simulated median PA_I's as target values for PA_G, thereby determining the B-plan conversion.

Not to be overlooked are the important considerations of how one responds to rejected lots and its impact on average outgoing quality (AOQ). The direct application of the B-plan without regard to alternative screening rules may substantially effect AOQ, and consequently average fraction inspected (AFI), as compared to the INDIVIDUAL plan. Future work will include an analysis of the conversion problem from the standpoint of AOQ, aimed at picking the "best" GROUP screening rule, given $(n,c)^m$ and its (N,C) B-plan.

APPENDIX 1

Proof of (2.1): to handle the constraints $0<q_j\leq 1$, set $q_j = e^{-x_j^2}$. By taking logs of objective function (1.2) and constraint (1.3), the problem takes the form

$$\text{optimize} \quad L(\tilde{X}) = \sum_{j=1}^{m} \ln\{OC(e^{-x_j^2};n,c)\} ,$$

$$\text{subject to} \quad G(\tilde{X}) = \sum_{j=1}^{m} x_j^2 = -\ln(q).$$

According to the Lagrange multiplier theory, extrema occur when all partials of $F(\tilde{X},\lambda) = L(\tilde{X})+\lambda\{G(\tilde{X})+\ln(q)\}$ are zero. The computations rely on

$$d/dq[OC(q;n,c)] = q^{n-c-1}(1-q)^c/B(n-c,c+1),$$

which is just a restatement of $OC(q;n,c) = BET(q;n-c,c+1)$, and

$$OC(q;n,c) = q^{n-c}(1-q)^c H(q), \text{ where } H(q) = \sum_{j=0}^{c} \binom{n}{c-j} \{q/(1-q)\}^j .$$

Therefore, for $1\leq i\leq m$,

$$d/dx_i[F(\tilde{X},\lambda)] = 2x_i\{\lambda - 1/[B(n-c,c+1)H(q_i)]\},$$

implying that extrema occur at vectors $\tilde{Q}$ whose components either (a) equal 1, or (b) satisfy, for q_i and q_j not 1, $H(q_i) = H(q_j)$. Since $H(q)$ is monotone increasing, $H(q_i) = H(q_j)$ implies $q_i = q_j$,

so to satisfy the constraint $d/d\lambda[F(\tilde{X},\lambda)] = G(\tilde{X})+\ln(q) = 0$, it follows that optimal $\tilde{Q}$ have $(m-k)$ components equal to 1, and k components equal to $q^{1/k}$, for $k=1,2,\ldots,m$. To determine which values of k correspond to the max and min, consider $g(t) = \{OC(q^{1/t};n,c)\}^t$. For integer $t \geq 1$, $g(t)$ is the objective function (1.2) evaluated at optimal $\tilde{Q}$ vectors. Since $g(t)$ is decidedly monotone increasing (see [7]), the min and max occur at $k=1$ and $k=m$ respectively, producing (2.1). Q.E.D.

APPENDIX 2

Proof of (4.2): make the change of variable

$$q_m = q / \prod_{i=1}^{m-1} q_i = P_m.$$

The joint density

$$j(q,q_1,q_2,\ldots,q_{m-1}) = \prod_{i=1}^{m-1} f_i(q_i)$$

$$= nlg(P_m;a,b_m) \prod_{i=1}^{m-1} \{nlg(q_i;a,b_i)/q_i\} .$$

The conditional joint density is

$$h(q_1, q_2, \ldots q_{m-1} \mid q) = j(q, q_1, q_2, \ldots, q_{m-1}) / nlg(q; a, S_1)$$

$$= F_1 F_2 / F_3 \quad .$$

where $$F_1 = \prod_{i=1}^{m-1} \{[\ln(1/q_i)]^{b_i - 1} / q_i\} \quad ,$$

$$F_2 = [\ln(1/P_m)]^{S_m - 1} / [\ln(1/P_1)]^{S_1 - 1} \quad ,$$

$$= \prod_{i=1}^{m-1} \{[\ln(1/P_{i+1})]^{S_{i+1} - 1} / [\ln(1/P_i)]^{S_i - 1}\} \quad ,$$

and $$F_3 = \{\prod_{i=1}^{m-1} \Gamma(b_i)\} / \Gamma(S_1) = \prod_{i=1}^{m-1} B(S_{i+1}, b_i) \quad .$$

Therefore $$h = \prod_{i=1}^{m-1} bet(T_i(q_i); S_{i+1}, b_i) dT_i(q_i)/dq_i \quad .$$

Integration of h with respect to q_i over $[P_i, x]$ if $i=k$, and $[P_i, 1]$ if $i \neq k$, for $i=1,2,\ldots,m-1$ and a specified k, then setting x back to q_i, produces (4.2). Q.E.D.

REFERENCES

1. MIL-STD-105D, Sampling Procedures and Tables for Inspection by Attributes, United States Department of Defense, Washington, D.C., 1964.
2. Bronken, C.J., "Proposed Matrices between Class and Individual Acceptable Quality Levels when Utilizing MIL-STD-105D", Proceedings, ARRADCOM Product Assurance Forum, May, 1979.
3. Hald, A., Statistical Theory of Sampling Inspection by Attributes, Academic Press, London, 1981.
4. Lorber, S.J., Letter, DRCQA-E, HQ DARCOM, "Reliability, Availability, and Maintainability (RAM) Specification Requirements", Feb 6, 1984.
5. Martz, H.F., Waller, R.A., Bayesian Reliability Analysis, John Wiley and Sons, 1982.
6. Mastran, D.V., Bayesian Assessments of Coherent Systems, Technical Report, Defense Program Analysis and Evaluation, Pentagon, Washington, D.C., 1973.
7. Roediger, P.A., Mardo, J.G., "An OC Curve Inequality", Problem 86-20*, Siam Review, Vol. 28, No. 4, Dec 1986.
8. Roediger, P.A., Mardo, J.G., and Loniewski, E.E., "Isim/Gsim, Fortran Code to Convert an Individual Plan to a Comparable Group Plan", Unpublished Technical Report.

DESIGN OF EXPERIMENTS CONFERENCE ATTENDEES
29-31 OCTOBER 1986

NAME	AGENCY
ADAMS, Lonnie MAJ	HQDA, DACS-DPD, Washington, D.C. 20310
ADKINS, Bud Mr.	USAMRSA,AMXMD-EL, Lexington, KY 40511-5101
BABA, Anthony Mr.	Harry Diamond Lab, 2800 Powder Mill Road, Adelphi, MD 20783
BAKER, William E. Mr.	USA Ballistic Res Lab, SLCBR-SE-D, APG, MD 21005
BALADI, George Y. Dr.	US Army Eng Waterways Exp St, P.O.Box 631, Vicksburg,MS 39180-0631
BASFORD, Kaye Dr.	MSI, Cornell University, Ithaca, NY 14853-0401
BATES, Carl B. Mr.	Math Stat Team, USACAA, 8120 Woodmont AV, Bethesda,MD 20814-2797
BISSINGER, Barney Dr.	Dept of Math, Pennsylvania State Univ, Middletown, PA 17057
BODT, Barry A. Mr.	USA Ballistic Res Lab, SLCBR-SE-D, APG, MD 21005-5056
BONKOWSKI, Ralph R. Mr.	Martin Marietta, 9532 Bay Vista Estates Blvd,Orlando,Fl 32819
BOX, George Prof.	Dir of Research, Univ of Wisconsin-Madison, Madison, WI 53705
BRANDON, Dennis Mr.	USAE Waterways Exp Station, P.O. BOX 631, Vicksburg,MS 39180
BRYSON, Marion R. Dr.	Dir, USACDEC, Fort Ord, CA 93941-7000
BURGE, Jay R. Mr.	Walter Reed Army Institute of Research, Washington, D.C. 20012
BURNS, Marian L. Ms	USA Matl Readiness Spt Act, AMXMD-ER, Lexington, KY 40511-5101
CELMINS, Alvars Dr.	USA Ballistic Res Lab, AMXBR-VLD-G, APG, MD 21005-5066
CORREIA, Charles A. Dr.	USA Logistics Center, Fort Lee, Virginia 23801
CRONIN, Terry Mr.	Ctr for Signal's Warfare,AMSEL-SW-PC,Warrenton,VA 22186-5100
DAVIS, Lynn Mr.	Chem Res Dev & Eng Ctr, SMCCR-ST, Aberdeen Proving Gd,MD 21010
DIACONIS, Persi Prof.	Dept of Mathematics, Harvard University, Cambridge, MA 02138
DRESSEL, Francis G. Dr.	USARO, SLCRO-MA, Research Triangle Park, NC 27709-2211
DUDEL, Helmut P. Mr.	USA Missile Com, Res,Dev & Eng Ctr, Redstone Arsenal,AL 35898-5248
DUTOIT, Eugene Dr.	USA Infantry School, Fort Benning, GA 31905
EDGE, SUSAN Mrs.	USA Communications Electronics Bd, ATSD, Ft Gordon,GA 30905
ESSENWANGER, Oskar Dr.	USA Missile Com, AMSMI-RD-RE-AP, Redstone Arsenal, AL 35898
FEDERER, Walter Prof	Cornell University, 337 Warren Hall, Ithaca, NY 14853-0401
GAVER, Donald Dr.	Naval Post Graduate School, Monterey, CA 93943
GEMAN, Stuart Prof	Div of Applied Mathematics, Brown University, Providence,RI 02912
GERSCH, Will Prof.	Dept of Opns Res, Naval Post Graduate School, Monterey, CA 93943
GOODWIN, John W. LCDR	Naval Post Graduate School, Monterey, CA 93940
GOTWALS, Edwin Mr.	AMSAA-LIRO, US Custom House,2d&Chestnut,Philadelphia,PA 19106
GRIMES, Fred Dr.	TCATA, CATD, Ft. Hood, Tx 76548-5065
HARRIS, Bernard Dr.	Dept of Statistics,Univ of Wisc,610 Walnut St, Madison, WI 53706
HARRIS, David Dr.	NSA, 9856 Softwater Way, Columbia, MD 21046
HUTCHINS, Charles CDR	Naval Post Graduate School, Monterey, CA 93943
JOHNSON, Dallas Dr.	Kansas State University
JOHNSON, Ronald L. Mr.	USA Res Dev & Eng Ctr,STRBE-JDS, Fort Belvoir, Va 22060
KASS, Richard A. Dr.	USAOTEA, CSTE-AV, 5600 Columbia Pike, Falls Church, VA 22041
KAUFFMAN, Tom Mr.	Rice University, 2929 Rolido #214, Houston, TX 77063
LAUNER, Robert L. Dr.	USARO, Math Sciences Div, SLCRO-MA,Res Triangle Park, NC 27709
LEHNIGK, S.H. Dr.	US Army Missile Com, AMSMI-RD-RE-OP, Redstone Ars, AL 35898
LOPES, Lewis Mr.	Sys Anal & Ctrl, 6665 Golfcrest Dr, San Diego, CA 92119
LUFKIN, Bradley M. Mr.	OTEA, 5600 Columbia Pike, Falls Ch, VA 22041
MILLER, Allen Mr.	USAOTEA, CSTE-AV, 5600 Columbia Pike, Falls Church, VA 22041-5115
NEUBERT, Chris Mr.	13416 Elevation Lane, Herndor, VA 22071-4007
O'MARA, Peter LTC	CDEC Test Board, Fort Lewis, WA 98433

PARZEN, Emanuel Dr.	Inst of Statistics,Texas A&M Univ., College Station,TX 77843
PASINI, Harold C. Mr.	USA Matl Sys Anal Act, AMXSY-RI, APG, MD 21005-5071
PURDUE, Peter Dr.	Naval Post Graduate School, Monterey, Ca 93943
ROEDIGER, Paul A. Mr.	USA Armament, Munitions & Chemical Com, Dover, NJ 07801-5001
ROHANI, Behzad Mr.	USAE Waterways Exp Station, P.O. Box 631, Vicksburg, MS 39180-0631
ROHRER, Todd Mr.	CDEC Test Board, Fort Lewis, WA 98433
RUSSELL, Carl T. Dr.	USA OTEA, CSTE-SP-M, 5600 Columbia Pike, Falls Church,VA 22041
SETHURAMAN, J. Prof.	Dept of Statistics, Florida State Univ, Tallahassee,FL 32306
SHALLHORN, Brian R. Mr.	Harry Diamond Lab, 2800 Powder Mill Road, Adelphi, MD 20783
SMITH, Laurel Dr.	Texas A&M Univ., Dept Of Stat, College Station, TX 77843
SU, Li Pi Ms.	USA Mat Readiness Spt Act, AMXMD-EL, Lexington, Ky 40511-5101
TANG, Douglas B. Dr.	Dept of Biostatitistics Applied Math, WRAIR, Washington,DC 20012
TAYLOR, Malcolm Dr.	Ballistic Res Lab, SLCBR-SECPO,APG, MD 21005-5006
TESSMER, Joseph M. Mr.	USAF/SASF, The Pentagon, Wash, D.C. 20330-5026
THOMPSON, James Prof.	Rice Univ,Dept of Math Science,P.O.Box 1892, Houston,TX 77251-1892
THRELKELD, David LCDR	Naval Post Graduate School, Monterey, CA 93943
TINGEY, Henry B. Prof.	USA Ballistic Res Lab, SLCBR-SE-D, APG, MD 21005-5056
TUKEY, John Prof.	Sr.Res.Stat, 408 Fine Hall, Washington Road, Princeton, NJ 08544
TYTULA, Thomas P. Dr.	USAMICOM, AMSMI-RD-SE-EA, Redstone Arsenal, AL 35898-5274
VANGEL, Mark Mr.	US Army Material Tech Lab, Watertown, MA 02171
WEBB, David W. Mr.	USA Ballistic Research Lab, SLCBR-SE-D, APG, MD 21005-5056
WEIGLE, Robert E. Dr.	Dir,US ARO, P.O.Box 12211,Res Triangle Park, NC 27709-2211
WEINBERGER, Marcus Dr.	Oprn Res & Anal Establ. Dept of Nat Def,Ottawa Ontario CAN KIA OK2
WEST, Larry Mr.	USA Test & Eval. Com, AMSTE-EV-O, APG, MD 21005-5066
WEST, William D. Mr.	USACDEC, Dir for Science & Technology, Fort Ord, CA 93941-7000
WILLNER, Lisa Ms.	Oprn Res & Anal Establ.,Dept of Nat Def,Ottawa Ontario CANADA KIA OK2
WINNER, Wendy A. Ms.	USA Ballistic Res Lab,SLCBR-SE-D,Aberdeen Proving Gd,MD 11005-5066
WOLD, Neil P. Mr.	USA Cold Regions Test Ctr, MT-A, APO Seattle, WA 98733-7850
WOMACK, Frank Mr.	USACAA, 8120 Woodmont, Avenue, Bethesda, MD 20814-2797
WOODRUFF, Brian W. Dr.	Air Force Ofc of Scientific Res/NM Bolling AFB, Wash,D.C. 20332

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 87-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCEEDINGS OF THE THIRTY-SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on Behalf of the Chief of Research, Development and Acquisition		12. REPORT DATE June 1987 13. NUMBER OF PAGES 408
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) US Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709-2211		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. The findings in this report are not to be construed as Official Department of the Army position, unless so designated by other authorized documents.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is a technical report from the Thirty-Second Conference on the Design of Experiments in Army Research, Development and Testing. It contains most of the papers presented at that meeting. These articles treat various Army statistical and design problems.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

intercropping
messy data
maximum likelihood estimation
factor anlaysis
one-sided tolerance limits
random allocation schemes
evaluation of expert opinion
incomplete block designs
significant interactions
desert camouflage
message filtering
rejection criteria
short time series
predictive analysis
quantile analysis
hypothesis testing
fuzzy random variables
tolerance limits
missing data
stress-strain properties
Baysesian image analysis
group sampling
sensitivity testing
system failure

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED