

## INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.
2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.
3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.
4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.
5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

**University  
Microfilms  
International**

300 N. Zeeb Road  
Ann Arbor, MI 48106



8314760

Choi, Uinam Jung

INFERENCE CONTROL IN STATISTICAL DATABASES: A DATA  
DISTORTION BY PROBABILITY DISTRIBUTION

*The University of Oklahoma*

PH.D. 1983

University  
Microfilms  
International 300 N. Zeeb Road, Ann Arbor, MI 48106



THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

INFERENCE CONTROL IN STATISTICAL DATABASES: A  
DATA DISTORTION BY PROBABILITY DISTRIBUTION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

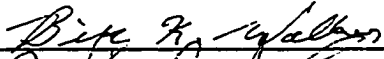


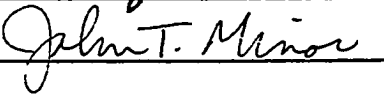
UINAM JUNG CHOI

Norman, Oklahoma

1983

INFERENCE CONTROL IN STATISTICAL DATABASES: A  
DATA DISTORTION BY PROBABILITY DISTRIBUTION

APPROVED BY

DISSERTATION COMMITTEE

## ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to Dr. Bill K. Walker, chairman of the dissertation committee, for his guidance and supervision in this study.

The author is indebted to Dr. Chong K. Liew for his inspiration and criticism, and appreciation is expressed to Dr. John T. Minor and Dr. John C. Thompson for their beneficial suggestions.

Finally, the author would like to express his gratitude to his wife for her patience and understanding throughout the graduate study. He also wishes to thank his parents for their encouragement.

INFERENCE CONTROL IN STATISTICAL DATABASES: A  
DATA DISTORTION BY PROBABILITY DISTRIBUTION

By

Uinam Jung Choi

Major Professor: Bill K. Walker

The legal code on "The Protection of Human Subjects in Research Activities" requires that sensitive information about an individual should be protected from unauthorized release and at the same time, those data should be available for statistical analysis. To meet these conflicting goals, recent research efforts focus on creating distorted data which is not easily compromisable and yet preserves the statistical properties of the original data.

An efficient and effective data distortion technique is introduced. This "Probability Data Distortion," which is not easily compromisable and has asymptotically the same statistical properties as the original data, is significantly different from the conventional Point Data Distortion technique which adds random errors to the original values. This mechanism, the data distortion by probability distribution, is resistant to compromise and provides better exposure for statistical analysis than do the existing data distortion techniques.

## TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	vii
LIST OF ILLUSTRATIONS . . . . .	viii
 Chapter	
I. INTRODUCTION . . . . .	1
1.0 Statement of the Problem . . . . .	1
1.1 Examples of the Existing Problem . . . . .	3
1.2 Outline of the Research . . . . .	5
II. REVIEW OF RESEARCH ON INFERENCE CONTROL . . . . .	9
2.0 Introduction . . . . .	9
2.1 Existing Inference Control Mechanisms . . . . .	10
2.2 Inference Control by Data Distortion . . . . .	11
III. PROBABILITY DATA DISTORTION . . . . .	16
3.0 Introduction . . . . .	16
3.1 Probability Data Distortion . . . . .	19
3.1.1 Identification and Estimation . . . . .	21
3.1.2 Data Generation . . . . .	23
3.1.3 Mapping and Replacement . . . . .	23
3.2 The Asymptotic Properties of the Probability Data Distortion . . . . .	24
IV. A MONTE CARLO STUDY OF THE PROBABILITY DATA DISTORTION . .	31
4.0 Introduction . . . . .	31
4.1 An Example of Probability Data Distortion . . . . .	31
4.2 Small Sampling Efficiency of Probability Data Distortion . . . . .	34
4.3 Compromisability of the Probability Data Distortion . . . . .	40

V. FREQUENCY IMPOSED DATA DISTORTION . . . . .	53
5.0 Introduction . . . . .	53
5.1 A Monte Carlo Study . . . . .	54
5.1.1 Point Data Distortion . . . . .	57
5.1.2 Probability Data Distortion . . . . .	57
5.1.3 Frequency Imposed Data Distortion . . . . .	58
5.1.4 Frequency Imposed Probability Data Distortion . . . . .	58
5.2 Empirical Results . . . . .	59
5.2.1 Accuracy of Parameter Estimation . . . . .	59
5.2.2 Accuracy in Computation of Conditional Statistics . . . . .	63
5.2.3 The Compromisability of the Distorted Data . . . . .	66
5.3 Summary . . . . .	68
VI. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS . . . . .	83
6.1 Summary . . . . .	83
6.2 Conclusions . . . . .	85
6.3 Suggestions for Further Research . . . . .	86
REFERENCES . . . . .	88
APPENDIX A . . . . .	93
APPENDIX B . . . . .	97
APPENDIX C . . . . .	100
APPENDIX D . . . . .	101

## LIST OF TABLES

TABLE	Page
4.1 Grand Mean of AAME of the Seven Statistics . . . . .	39
4.2 Parameter Estimation by Original and Distorted Data (Probability Data Distortion) . . . . .	45
4.3 Parameter Estimation by Original and Distorted Data (Point Data Distortion) . . . . .	46
4.4 Average Absolute Mean Error . . . . .	47
4.5 CHI-SQUARES . . . . .	48
4.6 Number of Cases in which the Probability Data Distortion is Better than the Point Data Distortion . . . . .	49
4.7 Degree of Compromisability . . . . .	50
4.8 Faculty Salary by Probability Data Distortion . . . . .	51
4.9 Faculty Salary by Point Data Distortion . . . . .	52
5.1 Frequency Table by Salary Range . . . . .	55
5.2 Frequency Table by Rank and Department . . . . .	55
5.3 Accuracy in Computation in 7 Statistics . . . . .	70
5.4 Degree of Compromisability . . . . .	81

## LIST OF ILLUSTRATIONS

FIGURE	Page
4.1 Compromisability Index (Pooled Case) . . . . .	44
5.1 Compromisability Index (Total Case) . . . . .	69

INFERENCE CONTROL IN STATISTICAL DATABASES: A  
DATA DISTORTION BY PROBABILITY DISTRIBUTION

CHAPTER I

INTRODUCTION

1.0 Statement of the Problem

Computers have now become such powerful tools in the hands of both users and producers of data that the dimensions of the confidentiality problem have been transformed. Automated information systems have the ability to process more records more rapidly than ever before.

Some data bases are used primarily for statistical purposes; such databases are typically released or made public either through the publication of a set of statistical tabulations corresponding to the data base with identifying information removed, or through responses to user queries. It has been shown in both cases that confidential individual information can often be compromised by such a statistical presentation.

Controlling inference from statistical databases is a problem of recent interest. Statistical databases contain data on many sensitive characteristics, associated with particular individuals, which must

be carefully protected. At the same time, however, users should be allowed access to such databases for statistical purposes in a manner which does not violate the privacy of a particular individual. For example, the user should be allowed access to information on the average salary of professors at the University of Oklahoma, but denied access to information on the salary of a particular professor. Privacy is a legal and social concept, which refers to the right of an individual to control the collection, storing and dissemination of data about himself. The right to privacy can conflict with society's need for free flow of information.

Threats to data privacy and to security in statistical databases may arise from an intruder. While statistical databases are a necessity for many purposes, they can be a danger if confidential information can readily be obtained from them. Thus there is a need to provide statistical information while preserving the confidentiality of the underlying data. This is the problem of statistical database inference control.

The problem of inference control has grown with the spread of computer installations. Confidentiality is now a concern of every organization which uses a database. The statistical databases must be protected from disclosure and other illegal compromise. The problem is now much more complex and much more important than ever before.

Inference control in statistical databases will become an even more important issue since many current types of research, such as business forecasting, medical research, and political decision making, are heavily dependent upon statistical data. This control has never been

easy or straight forward, and is made particularly difficult by the possibility that a user may ask a series of permitted queries and correlate enough statistical summaries to compromise confidential information.

### 1.1 Examples of the Existing Problem

The problems existing in statistical databases will be illustrated through the use of a few examples.

Definition 1:

The complement of a set  $A$  with respect to the space  $\Omega$ , denoted by  $\bar{A}$ , is the set of all points that are in  $\Omega$  but not in  $A$ .

Theorem 1:

$$A \cdot \bar{A} = \phi; A + \bar{A} = \Omega; (\bar{\bar{A}}) = A.$$

Definition 2:

A query is a statistical question which can be asked about the database. A query is called a permitted query if it is permitted by the system; otherwise it is called a restricted query (Denning (1978)).

The query set,  $X_C$ , is the set of records satisfying a characteristic  $C$ , which is an arbitrary logical formula using categorical values as terms connected by operators AND ( $\cdot$ ), OR ( $+$ ), and NOT ( $\bar{\phantom{x}}$ ).

$$\text{COUNT } (C) = \{|X_C| \mid |X_C| \text{ is the size of } X_C\}$$

$$\text{TOTAL } (C;j) = \{ \sum_{i \in X_C} V_{ij} \mid V_{ij} \text{ is the value in the } i\text{th record of } j\text{th field} \}$$

$$\text{SELECT } (C;j) = \{ \text{SELECT } V_{ij} \mid \text{SELECT IS MEDIAN, MAX, MIN, MODE, etc . . . .} \}$$

Definition 3:

A Tracker is a set of auxiliary characteristics which is added to the original characteristics in the formation of a query. When the auxiliary characteristics form permitted queries, the user subtracts out the effects of the auxiliary characteristics to determine the answer to the query for the original characteristics (Denning, Denning, and Schwartz (1979)).

EXAMPLE 1:

If a user knows Prof. M in the computer science department is married, then the second query reveals that he got a pay raise.

$$Q_1 = \text{COUNT (Professor} \cdot \text{Married} \cdot \text{CS)} = 5$$

$$Q_2 = \text{COUNT (Professor} \cdot \text{Married} \cdot \text{CS} \cdot \text{Pay raise)} = 5$$

If we know  $k$  of his characteristics, then we can find  $C_{k+1}$  such that

$$\text{COUNT } (C_1 \cdot C_2 \cdot \dots \cdot C_k) = \text{COUNT } (C_1 \cdot C_2 \cdot \dots \cdot C_k \cdot C_{k+1})$$

then he also possesses characteristic  $C_{k+1}$ .

If  $\text{COUNT } (C_1 \cdot C_2 \cdot \dots \cdot C_k) \neq \text{COUNT } (C_1 \cdot C_2 \cdot \dots \cdot C_k \cdot C_{k+1})$ , then we can not determine whether he has characteristic  $C_{k+1}$  unless  $\text{COUNT } (C_1 \cdot C_2 \cdot \dots \cdot C_k) = 1$ .

EXAMPLE 2:

Let's consider these five queries.

$$Q_1 = \text{TOTAL (A,B,C; Salary)} = \$67,000$$

$$Q_2 = \text{TOTAL (D,E,F; Salary)} = \$71,000$$

$$Q_3 = \text{TOTAL (A,D,G; Salary)} = \$60,000$$

$$Q_4 = \text{TOTAL (B,E,G; Salary)} = \$65,000$$

$$Q_5 = \text{TOTAL (C,F,G; Salary)} = \$73,000$$

The G's salary can be determined by

$$(Q_3 + Q_4 + Q_5 - Q_1 - Q_2)/3 = \$20,000.$$

EXAMPLE 3:

Suppose a user wishes to learn Prof. Y's salary. This can be calculated using the "Tracker" technique, if the user knows Prof. Y is the only female professor in the Computer Science department.

$$Q_1 = \text{TOTAL} (\text{Female} \cdot \text{CS}; \text{Salary}) + \text{TOTAL} (\overline{\text{Female}} \cdot \overline{\text{CS}}; \text{Salary}) = \$1,600,000$$

$$Q_2 = \text{TOTAL} (\text{Female} \cdot \text{CS} + \text{CS}; \text{Salary}) + \text{TOTAL} (\text{Female} \cdot \text{CS} + \overline{\text{CS}}; \text{Salary}) = \$1,623,000$$

Then Prof. Y's salary can be determined by

$$Q_2 - Q_1 = \text{TOTAL} (\text{Female} \cdot \text{CS}; \text{Salary}) = \$23,000.$$

EXAMPLE 4:

$$Q_1 = \text{MEDIAN} (A, B, C; \text{Salary}) = \$28,000$$

$$Q_2 = \text{MEDIAN} (D, E, C; \text{Salary}) = \$28,000$$

If no two individual's salaries are same, C's salary can be determined. Since C is the only individual common to both queries, the returned median value \$28,000 must be C's salary.

## 1.2 Outline of the Research

Many researchers in this area have considered methods for controlling such inference from statistical databases. The security problem of a statistical database is to limit the use of the statistical database so that only statistical information is available and no sequence of queries is sufficient to deduce private information about any individual.

Several studies have reported conditions which, if imposed on the contents of a database, guarantee its security. In most cases, these conditions depend on the users having little previous knowledge about the information in the database. Since users have prior information, these methods are not practical in reality. All of these policies are either not sufficient to enforce total security or require very large amounts of computation to detect compromising queries. It has been shown that all of these policies require further study to insure total security (Schwartz (1977), Chin (1978)). Dobkin, Jones and Lipton (1979) show that if the responses to queries are exact values, the data is easily compromised.

In general, inference control schemes impose restrictions on the system. A good protection scheme should provide security to a reasonable extent, possibly by restricting the information to be released to users, and also should maintain the statistical properties of the database.

This research is an attempt to develop an efficient and effective mechanism for protecting the privacy of an individual from user inference and for making the cost of compromising statistical databases unacceptably high. An efficient inference control mechanism, Probability Data Distortion, is introduced. The three major strong points of this technique are:

- (1) The resultant microdata remain useful for statistical purposes.
- (2) There is a high degree of confidentiality in the microdata.

- (3) This technique can be applied to a small database as well as a large data base.

The research presented in this dissertation may be divided into three general categories:

- (1) Theoretical development and mathematical justification of the Probability Data Distortion mechanism.
- (2) Applications and compromisability test of this new approach and comparisons to previously existing methods.
- (3) Monte Carlo studies of the usefulness of this technique.

First a basic method for identifying the probability distribution of the data set and generating the distorted data by a probability distribution function is considered. The theoretical developments and asymptotic properties of data distortion by probability distribution have been presented in the form of statistical theorems and proved mathematically to show these properties are satisfied for data from large samples.

The small sampling properties of Probability Data Distortion are shown with a hypothetical example, and the effectiveness of the mechanism is illustrated using simulation. Monte Carlo studies are used as a basis for evaluation of the Probability Data Distortion method in application to statistical databases. The efficiency and compromisability of Probability Data Distortion are tested with small-sample data and compared to Point Data Distortion.

Finally the usefulness of this mechanism is demonstrated, using a real salary data set. The statistics of the probability distorted data set and the original data set are compared to determine the

efficacy of the method. It is also demonstrated that this method could be applied to databases to achieve both confidentiality of individual information and accuracy of the statistical properties by introducing Frequency Imposed Data Distortion which doesn't require identification of the underlying density function. Frequency Imposed Probability Data Distortion which is a hybrid of Probability Data Distortion and Frequency Imposed Data Distortion is also introduced.

## CHAPTER II

### REVIEW OF RESEARCH ON INFERENCE CONTROL

#### 2.0 Introduction

Statistical databases sometimes contain sensitive information which is associated with particular individuals and which needs to be protected. The legal code requires that sensitive information associated with a particular individual be protected from unauthorized release. For example, the U.S. Department of Health, Education and Welfare's regulation (45 CFR 46) requires protection of "The rights and welfare of individuals who may be exposed to the possibility of physical, psychological or social injury while they are participating as a subject in research, development, or related activities." At the same time, users should be allowed access to this data base for statistical analysis as long as they do not infringe upon the privacy of a particular individual. This inference control depends upon the queries allowed and the amount of initial information in the possession of a user. Hoffman and Miller (1970) have shown that a user can combine the answers to some specific statistical queries and some previous knowledge of an individual's personal information to find out more about him.

## 2.1 Existing Inference Control Mechanisms

One method of inference control against isolating a record by overlapping queries is Partitioning database (Chin-Ozsoyoglu (1978) Yu-Chin (1977)). Records are stored in groups, each containing at least some predetermined number of records. Queries may apply to any set of groups, but never to subsets of records within any group. Therefore it is impossible to isolate a record.

A variant of the Partitioning database is called Microaggregation: individuals are grouped and aggregated and statistics are computed for the aggregated set rather than for individuals (Feige-Watts (1970)). This technique increases uncertainty about the original information in the records as the size of the aggregated group of records is increased.

"Minimum overlap control" inhibits responses to queries that have more than a predetermined number of records in common with any prior query since this could lead to identification of an individual (Dobkin, et al (1979)). The difficulty with this approach is keeping track of all previous queries when the number of requests is large, and also that two users might cooperate to fool the system if the requests of each taken separately are not suspicious enough to be detected. This control also may not be safe against queries that overlap by small amounts (Davids, et al (1978), Reiss (1978)).

"Perturbing output" is a technique which consists of rounding the output up or down by a small amount before the answer to a query is released. Rounding by adding random values from a set with zero-mean is insecure since the correct answer can be deduced by averaging a

sufficient number of responses to the same query (Nargundkar-Saveland (1972)).

The U.S. Census Bureau has used the technique of responding only to queries which involve a random subfile of the data base, but not the complete data base. Even if some element of the subfile is identified, it may not be possible to learn which individual in the data base was selected to be this element (Hansen (1971)). This technique is applicable only to large data bases. A small random sample would not be statistically significant and would not represent the statistical properties of the data. For this reason, random sampling has been ignored as a possible inference control.

Surveys of statistical data base security can be found in Denning and Denning (1979), Denning (1978) and Hoffman (1977). It has been shown that all of these schemes either do not provide sufficient security or are impractical to implement.

Denning (1980) introduced a new inference control, called random sample queries. The random sample queries control deals directly with the basic principle of compromise by making it impossible for a user to control precisely the formation of query sets. Queries for relative frequencies and averages are computed using random samples drawn from the query sets. This technique is also effective only for reasonably large data bases.

## 2.2 Inference Control by Data Distortion

The research effort has focused on finding a data set which provides statistical results similar to those of the original data set while preventing the curious user from identifying information relating

to a particular individual. There has been a continuous search for a data distortion method which achieves these dual goals; i.e., accurate statistical estimation and protection of the privacy of an individual.

In some applications it is desirable to provide microdata in response to statistical query. If this microdata is derived directly from the actual data base, then compromise is certain. The data distortion technique is explored as a means of inference control in statistical database while enabling research scholars to utilize the information for analytic purposes. This technique provides protection of the data through addition of a random error of relatively small variance and zero-mean to the original data values. If the random variables are produced by a process whose statistical characteristics are properly chosen, the statistical properties of the distorted data are not altered.

Conway and Strip (1976) suggested that the value would be modified by some random quantity, such that

$$V_d = V_a + V_r$$

where  $V_d$ : Distorted value

$V_a$ : Actual value

$V_r$ : A random variable with zero mean and a constant standard deviation.

The distribution is chosen to have an expected value of zero, so the  $V_d$  is an unbiased estimator of the true value  $V_a$ . What would constitute an appropriate distribution, however, is not always obvious. If the population of values  $V_a$  in the statistical database is symmetric, then the random deviate distribution should probably also be symmetric. But

if the population of  $V_a$  is highly skewed, which is a common occurrence, then the choice of the distribution is much more difficult.

This random modification of data to avoid disclosure has been considered extensively for various applications in the U.S. Bureau of Census over the past decade, but its actual application has been quite limited.

Recently Beck (1980) has shown a clever approach to protection from compromise by repeated queries, and introduces a formula for his scheme such that

$$V_d = V_a + C_r \cdot (V_a - \bar{V}_a) + V_r$$

where  $\bar{V}_a$  is the mean value of the actual data over the query set. In this equation  $C_r$  and  $V_r$  are independent random variables generated for each record with expected values of  $C_r$  and  $V_r$  equal to zero. Normal distributions were used to generate  $C_r$  and  $V_r$ . This scheme does not guarantee that the accuracy of statistical properties of the data set will be maintained.

In most cases, protection of privacy is assured if an individual's record is only slightly distorted. For satisfactory privacy, the level of distortion of the data should be sufficiently high to control user inference from the statistical data base, yet low enough for the distorted values of the data to be used in statistical analysis. For this purpose we have to minimize the loss of useful information in the actual data after the data has been distorted.

Value Dissociation is another technique which was suggested by Conway and Strip (1976). With this approach, each value in the data is exchanged with a value from the same field in some record different

from the actual record it represents, in such a way as to preserve certain statistical properties. This technique has the following advantages.

- (1) The distribution from which the random deviates are obtained is automatically appropriate in relation to the distribution of the values of the actual data.
- (2) Since the actual values are unchanged, certain statistical properties of those values are preserved.

There is, however, no known algorithm by which the actual values might be dissociated.

Hansen (1971) shows that the actual value might be distorted by a random distortion rate,  $C_r$ , such that

$$V_d = C_r \cdot V_a$$

where  $C_r$ : Randomly chosen interval.

If we compute a sum or an average on a large group, then the errors will tend to cancel out each other, so that the relative error variance in a sum from a large group is much smaller than the relative error in each single item. These data distortions are made on the value of either input or output. This family of data distortion is called "Point Data Distortion."

Alternatively, the original data is considered as a random variable which is associated with a probability distribution. If the underlying density function of the original data is determined, another set of data could be generated from the density function which will have, asymptotically, the same statistical properties as the original data set since they originate from the same density function. This approach is

appealing since the distorted data set not only preserves the basic statistical properties of the original data but also is sufficiently different from the original to protect the privacy of individuals. This family of data distortion is called "Probability Data Distortion."

"Data Swapping" is an early application of this Probability Data Distortion. It suggests that the original data be replaced with the data of other records which have the same frequency count statistics as those of the original data. The difficulty with this approach is finding the general data swaps which preserve all frequency counts (Dalenius and Reiss (1978)). Since exact Data Swapping is practically not feasible, Reiss (1980) suggests a feedback algorithm to find an Approximate Data Swapping on a categorical data set. Approximate Data Swapping is still in an experimental stage and its computational efficiency has yet to be proved. Furthermore, Approximate Data Swapping is not feasible for non-categorical data such as salary figures.

In implementing a data distortion mechanism, we must guarantee that the data is distorted in such a manner as to preserve statistical properties. "How do we modify the actual data?" This is the question which many researchers in this area have been trying to answer.

This new technique, Probability Data Distortion, will guarantee all statistical properties within certain confidence limits. This technique may be used both to produce microdata and to release statistical tabulations so that confidentiality is not violated. After carefully studying all of the existing methods just discussed, it is believed "Data Distortion" will be the most valuable technique in the near future.

## CHAPTER III

### PROBABILITY DATA DISTORTION

#### 3.0 Introduction

A data set may be divided into confidential variables and non-confidential variables. Consider a personnel file which contains current salary, salary raise, department, rank, sex, age, and the highest degree earned. The current salary and raise may be considered as confidential variables and the remaining variables as non-confidential variables.

The basic idea of Data Distortion is to construct a new data set that is equivalent to the original in terms of statistical properties, where the new data is sufficiently different from the original so that sensitive information cannot be compromised. As a consequence, a statistical analysis based on distorted data will be less accurate than one based on original data. This "Loss of Information" may be looked upon as the price paid for the protection of confidential information from the compromise: the better the protection that is wanted, the higher the price that must be paid. We must clearly try to strike a rational balance between two conflicting objectives: (1) to provide

good protection; and (2) accurate statistical estimates. The difference between the actual data and the distorted data might not affect the statistical uses to which the data are put as long as the amount of difference is reasonable, and the probability distributions are the same.

Most previous data distortions were made on the values of observations. If the values of data are distorted on point, the underlying probability distribution of the distorted data is not guaranteed to be the same as the probability distribution of the original data. The distorted data may have another probability distribution which is quite different from that of the original data. This deviation becomes severe when the original data has a non-symmetric type of distribution such as a skewed distribution. Because of this, many important statistical properties of the original data are not preserved by the point distorted data. Another difficulty with the Point Data Distortion is the relatively high compromisability; i.e., repeated queries could identify the original value by averaging since the expected value of the rounding errors is usually specified to be zero.

We present an alternative method of data distortion which will generate distorted data by probability distribution function. The basic hypothesis is that the original data is a sample from a population with a probability density function. Then another sample from the same population can be used as a distorted data set to replace the original one. Since these two data sets (original set and distorted set) share the same density function, the statistical properties of the original data

set such as frequencies, subtotals, median, percentiles, and mean should be asymptotically same as those of the distorted data set.

The distorted data sets can be generated if we know the population density function of the original data set. The population density function tells us how likely or probable each value of the original data is. The distortion is done by pulling a sample from the population density function. Compromisability becomes difficult since there is no guarantee that a sample (i.e., original data) should be the average of all other samples (i.e., distorted data) if these samples were drawn independently from the same population.

The statistical data which we observe have discrete values. The population density is a continuous function, so this function could generate as many discrete sets of data as we want. The original data set is replaced by one of the generated data sets from the same probability distribution. This family of data distortion is called "Probability Data Distortion" in contrast to "Point Data Distortion."

What are the advantages of Probability Data Distortion as compared to the traditional Point Data Distortion? The probability distorted data more accurately produce the statistical properties of the original data than do conventional point distorted data which consist of random errors added to the original values. This is true because Probability Data Distortion asymptotically preserves the statistical properties of the original data, and the probability distorted data asymptotically share the same density function with the original data set. Even in the case of small samples, it is difficult to compromise

the probability distorted data and so it protects the privacy of the individual from snoopers.

A prerequisite for effective use of this mechanism is that the actual data set should be totally replaced by the distorted one in the database. Once the original data set is replaced by the distorted one, the replaced data set can either be placed on line to answer queries or released as microdata. Also it may be used as basic data to form a statistical tabulation. This technique is highly acceptable for statistical tabulation because all the data are available and appropriate calculations can be performed. Moreover, Probability Data Distortion doesn't share the basic weakness of Point Data Distortion since it guarantees that statistics will be preserved. Suppose that the original data set is a dynamic one whose values are changed frequently over time, such as a salary data set; the parameters of the density function should be updated and the corresponding new distorted data set should be put in place of the original periodically. When the original data are associated with other variables the distorted data should be mapped onto the original data to maintain consistency with other variables.

We will discuss the Probability Data Distortion mechanism in section 3.1 and the asymptotic properties of this mechanism in section 3.2.

### 3.1 Probability Data Distortion

The data distortion by probability distribution, Probability Data Distortion, requires three steps to compute the distorted data set for confidential variables.

- Step 1: Identification of the probability density function for each confidential variable and estimation of the parameters associated with the density function.
- Step 2: Generation of a distorted data set for each confidential variable from the estimated density function.
- Step 3: Mapping and replacement of the confidential data by distorted data.

Definition 1:

Any function  $f(\cdot)$  with domain the real line and counterdomain  $[0, \infty)$  is defined to be a probability density function if and only if

$$(i) f(x) \geq 0 \text{ for all } x$$

$$(ii) \int_{-\infty}^{\infty} f(x) dx = 1.$$

The moments of a distribution are the expectations of the powers of the random variable which has the given distribution.

The "kth moment about the origin" is defined as

$$\mu_k' = E(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx$$

The first moment  $\mu_1'$ , is called the mean of  $X$ . The moments about any arbitrary point  $a$  are defined as

$$E[(X-a)^k] = \int_{-\infty}^{\infty} (x-a)^k f(x) dx$$

and when  $a$  is put equal to the mean, we have the "kth moment about the mean":

$$\mu_k = E[(X-\mu_1')^k] = \int_{-\infty}^{\infty} (x-\mu_1')^k f(x) dx$$

we have

$$\begin{aligned} \mu_1 &= \int_{-\infty}^{\infty} x f(x) dx - \mu_1' \int_{-\infty}^{\infty} f(x) dx \\ &= \mu_1' - \mu_1' = 0 \end{aligned}$$

and

$$\begin{aligned}
 \mu_2 &= \int_{-\infty}^{\infty} (x - \mu_1)^2 f(x) dx \\
 &= \int_{-\infty}^{\infty} [x^2 - 2x\mu_1 + (\mu_1)^2] f(x) dx \\
 &= \mu_2' - 2\mu_1\mu_1' + (\mu_1')^2 \\
 &= \mu_2' - (\mu_1')^2.
 \end{aligned}$$

This second moment about the mean is called the variance of  $x$ . The mean and variance play an important role in statistical data analysis. Purely as descriptive measures of the distribution, the mean represents a central value of the random variable and the variance represents the scatter around the central value. The third moment about the mean is used to describe the symmetry or skewness of a distribution and the fourth moment about the mean is similarly used to describe its peakedness, or kurtosis.

On the basis of the original data, we find the underlying probability density function and estimate its parameters. The distorted data are generated from the probability density function. These distorted and original data are sorted in the same order and the original data are replaced by the distorted data.

If the probability density functions of the actual data and the distorted data are the same, then we can claim that all the moments of these two sets of data are the same.

### 3.1.1 Identification and Estimation

This technique requires an identification of the probability density function which represents the original data set. The original

data is screened to determine which of the predetermined functions is best fitted to the data. The test of agreement between a theoretical probability distribution and the distribution of a set of original data constitutes a "Goodness of fit" test. "Goodness of fit" testing is accomplished by comparing the distribution of the actual data with the theoretical probability distribution. The goodness of fits can be tested by the Kolmogorov-Smirnov goodness of fit test. The Kolmogorov-Smirnov goodness of fit test gives us some indication of how well the original data points fit the probability density function which will be used to generate the distorted data.

Currently available density functions for such identification are Poisson, Exponential, Normal, Gamma, Weibull, Lognormal, Uniform, and Triangular distributions (Phillips (1972)). The Phillips computer package is easy to use for the identification of the underlying density function, and allows the user to test a set of  $n$  observations against the theoretical probability density functions using Chi-square, Kolmogorov-Smirnov, Cramer-Von Mises, and Moments "Goodness of fit" tests. This computer package also computes the best estimate of the parameters of the density function. Estimators of the chosen function's parameters must be found for use in generating distorted data. In some cases more than one density function could be acceptable at a given significance level. In this case selection of the density function which shows the smallest Kolmogorov-Smirnov statistics is recommended for the obvious reason that acceptance of the null hypothesis is most probable when using that function. If none of the density functions fits to the discrete data, we recommend use of a Frequency Imposed Data

Distortion method. The Frequency Imposed Data Distortion method doesn't require identification of any density function. This topic will be discussed in Chapter V.

### 3.1.2 Data Generation

Once the best fitted density function is selected, the estimated parameters of the density function are supplied to the density function's random number generating routine to produce the distorted data. The IMSL (1980) has random number generating subroutines for each of the density functions identifiable through the Phillips package. The number of distorted values generated from the density function matches the number of values in the original data. Suppose the original data has  $N$  observations then  $N$  distorted observations are generated from the density function.

### 3.1.3 Mapping and Replacement

When the distorted data are used for a statistical analysis independent of other variables, ordered mapping, i.e., sorting the distorted data and the original data in the same order and replacing each element of the original data with the corresponding distorted one is not necessary. However, in most cases, the distorted data are used in conjunction with other variables for a statistical analysis. For example, a response to a query for average salary by specific age interval is a case in which both distorted data and non-confidential age data are used jointly with other attributes for the statistical analysis. Unless the mapping is done, the average salary by age group becomes a meaningless value. In general, if the data set is a matrix in

which the distorted data set is a joint variable with other variables then the ordered mapping is necessary to maintain consistency with other attributes.

### 3.2 The Asymptotic Properties of the Probability Data Distortion

Definition 1. Random Sample.

Let the random variables  $X_1, X_2, \dots, X_n$  have joint density

$$g(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \dots F(x_n)$$

where the density of each  $X_i$  is  $F(x_i)$ . Then  $X_1, X_2, \dots, X_n$  is said to be a random sample of size  $n$  from the population with density  $F(x)$ .

Definition 2. Sampled Population.

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with density  $F(\cdot)$ ; then this population is called the sampled population.

Definition 3. Statistic.

A statistic is a function of observable random variables, which is itself an observable random variable, which does not contain any unknown parameters.

The basic assumption of the Probability Data Distortion is that the original data set is a sample of size  $n$  drawn from a population with a certain distribution, and the distorted data set is another sample of size  $n$  drawn from the same population.

If the sample is not large, we may inquire how we could find the probability density function and statistics when the distribution depends upon unknown parameters which may have any values within a range. The answer is that the parameters are not really unknown; they can be estimated, and the estimators approach to the population

parameters as the sample size increases. In the limit as  $n$  becomes infinite the parameters are known exactly, and the distribution becomes unique.

The statistics such as mean, standard deviation, percentiles, minimum, and maximum values computed from both samples (i.e., distorted and original) converge asymptotically to those of the population.

**Definition 4. Moment Generating Function.**

Let  $X$  be a random variable with density  $F(\cdot)$ . The expected value of  $e^{tX}$  is defined to be the moment generating function of  $X$  if the expected value exists for every value of  $t$  in some interval  $-h < t < h$ ;  $h > 0$ . The moment generating function, denoted by  $M(t)$ , is

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} F(x) dx.$$

**Statement A:**

If the moment generating function of a random variable exists, then this moment generating function uniquely determines the corresponding distribution function.

A particular moment or a few of the moments may give little information about the distribution, but the entire set of moments will uniquely determine the actual distribution, and also the density function determines a set of moments  $\mu'_1, \mu'_2, \dots$ , when they exist.

We evaluate the asymptotic properties of the distorted data.

**We define:**

$X$ : a population

$F$ : a density function associated with the random variable  $X$ .

$x_n$ : a random variable of size  $n$  drawn from the population  $X$ .

$F_n$ : a density function associated with  $x_n$ ;

We make the following statement, definition, Helly Lemma, and Helly-Bray Theorem to prove the asymptotic properties of the Probability Data Distortion.

Statement B:

When a sample size  $n$  increases sufficiently large to become population size,  $F_n$  converges to  $F$  (i.e.,  $F_n \rightarrow F$  as  $n \rightarrow \infty$ ).

Theorem 1:<sup>1</sup>

Let  $X$  be a random variable and  $g(\cdot)$  a nonnegative function with domain the real line; then

$$P[g(X) \geq k] \leq \frac{E[g(X)]}{k} \text{ for every } k > 0.$$

Proof:

Assume that  $X$  is a continuous random variable with probability density function  $F(X)$ ; then

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x)F(x)dx = \int_{\{x:g(x) \geq k\}} g(x)F(x)dx \\ &\quad + \int_{\{x:g(x) < k\}} g(x)F(x)dx \geq \int_{\{x:g(x) \geq k\}} g(x)F(x)dx \\ &\geq \int_{\{x:g(x) \geq k\}} kF(x)dx = kP[g(X) \geq k] \end{aligned}$$

Divide by  $k$ , we get the result.

Theorem 2: Weak law of large numbers.<sup>2</sup>

Let  $F(\cdot)$  be a density with mean  $\mu$  and finite variance  $\sigma^2$ , and let  $\bar{x}_n$  be the sample mean of a random sample of size  $n$  from  $F(\cdot)$ . Let  $\epsilon$  and  $\delta$  be any two specified numbers satisfying  $\epsilon > 0$  and  $0 < \delta < 1$ .

<sup>1</sup>Proof is shown in Mood, Graybill, and Boes (1974), page 71.

<sup>2</sup>Proof is shown in Mood, Graybill, and Boes (1974), pages 232-233.

If  $n$  is any integer greater than  $\sigma^2/\epsilon^2\delta$ , then

$$P[-\epsilon < \bar{X}_n - \mu < \epsilon] \geq 1 - \delta.$$

Proof:

Theorem 1 stated that  $P[g(X) \geq k] \leq E[g(X)]/k$  for every  $k > 0$ , random variable  $X$ , and nonnegative function  $g(\cdot)$ .

Equivalently,  $P[g(X) < k] \geq 1 - E[g(X)]/k$ .

Let  $g(X) = (\bar{X}_n - \mu)^2$  and  $k = \epsilon^2$ ; then

$$\begin{aligned} P[-\epsilon < \bar{X}_n - \mu < \epsilon] &= P[|\bar{X}_n - \mu| < \epsilon] \\ &= P[|\bar{X}_n - \mu|^2 < \epsilon^2] \geq 1 - \frac{E[(\bar{X}_n - \mu)^2]}{\epsilon^2} \\ &= 1 - \frac{(1/n)\sigma^2}{\epsilon^2} \geq 1 - \delta \end{aligned}$$

for  $\delta > \sigma^2/n\epsilon^2$  or  $n > \sigma^2/\epsilon^2\delta$ .

Definition 5.

The sequence of the random variables  $x_n$  is said to converge in distribution to a random variable  $X$  if  $F_n(x_n) \rightarrow F(x)$  as  $n \rightarrow \infty$ .

This definition is provided by Rao (1965), page 96.

Helly Lemma:<sup>3</sup>

Every sequence of distribution functions is weakly compact.

Helly-Bray Theorem:<sup>4</sup>

$F_n(x_n) \rightarrow F(x)$  as  $n \rightarrow \infty$  implies  $\int g dF_n \rightarrow \int g dF$  for every bounded continuous function  $g$ .

<sup>3</sup>Proof is shown in Appendix B.

<sup>4</sup>Proof is shown in Appendix B.

Theorem 3:<sup>5</sup>

If a variable which depends upon the sample size  $n$  has a moment generating function that approaches the moment generating function of a second variable, then the distribution function of the first variable must approach the distribution function of the second variable as  $n \rightarrow \infty$ .

Proof:

If the moment generating functions associated with two density functions  $F(x)$  and  $u(x)$  are the same and if the difference  $F(x) - u(x)$  has a power series expansion about the origin, then  $F(x) = u(x)$ , identically.

We have

$$F(x) - u(x) = a + bx + cx^2 + \dots, \quad (3-1)$$

and we form

$$\int_{-\infty}^{\infty} [F(x) - u(x)]^2 dx = \int_{-\infty}^{\infty} (a + bx + cx^2 + \dots) [F(x) - u(x)] dx,$$

replacing one  $[F(x) - u(x)]$  by (3-1).

On integration, the right hand-side reduces to 0, for the corresponding moments of the two random variables are given to be the same. As the left-hand integrand is nonnegative, it is evident that for the left-hand integral to be 0, we must have  $F(x) = u(x)$ , identically.

Lemma 1.

The moment generating function ( $M_n(t)$ ) of  $x_n$  asymptotically converges to the moment generating function ( $M(t)$ ) of  $X$  as  $n \rightarrow \infty$ .

Proof:

$$\text{By definition 4, } M_n(t) = \int_{-\infty}^{\infty} e^{tx_n} dF_n.$$

---

<sup>5</sup>Proof is shown in Freeman (1963), pages 32-33.

By definition 5,  $e^{tx_n} \rightarrow e^{tX}$  as  $n \rightarrow \infty$ .

By setting  $e^{tX} = g$ , and by using the result of Helly-Bray Theorem, the proof is completed.

Corollary 1:

The mean and standard deviation of  $x_n$  asymptotically converges to the mean and standard deviation of  $X$  as  $n \rightarrow \infty$ .

Proof:

This is direct result of Lemma 1 since the first and the second derivatives of the moment generating function evaluated at  $t = 0$  become mean and variance of the random variable respectively.

Lemma 2:

The cumulative density function ( $CF_n(a)$ ) of sample  $x_n$ , asymptotically converges to the cumulative density function ( $CF(a)$ ) of population  $X$  as  $n \rightarrow \infty$ .

Proof:

By definition,

$$CF_n(a) = \int_0^a F_n(x_n) dF_n \text{ and } CF(a) = \int_0^a F(x) dF$$

By the statement B,  $F_n \rightarrow F$  as  $n \rightarrow \infty$ . By setting  $F = g$  and by using the results of Helly-Bray Theorem and Helly Lemma we complete the proof. The Helly Lemma proves the existence of the upper bound of the cumulative density function.

Corollary 2:

The percentiles and median of sample  $x_n$  asymptotically converge to those of population  $X$ .

Proof:

The Corollary 2 is a direct result of Lemma 2 since percentiles and median are same if the cumulative density functions are same.

---

Wilks (1972) provides further properties of the asymptotic sampling theory on pages 254-274.

## CHAPTER IV

### A MONTE CARLO STUDY OF THE PROBABILITY DATA DISTORTION

#### 4.0 Introduction

The results of a Monte Carlo study concerning small samples are shown in the following three sections. The first section shows how data is generated using Probability Data Distortion. In the next section, the accuracy of the statistical estimation is tested by comparing average absolute mean error and Chi-square statistics for each set of distorted data (i.e., point distorted data and probability distorted data). In the last section, the compromisability index is computed and the compromisability of Point Data Distortion and Probability Data Distortion are compared.

#### 4.1 An Example of Probability Data Distortion

Consider a hypothetical example of the faculty salary of a business school which has four divisions (Finance, Economics, Management and Accounting). The original data for salary by division are in Table 4.8.

Release of the original data will lead to easy identification of the salary of each professor. For example, a faculty member in

finance who receives \$27,300 can easily deduce the salary of his colleagues. He can figure out that the \$19,600 must be the salary of a recently arrived assistant professor and \$35,600 would be the salary of the division director. In this way, he can easily guess the salary of a faculty in any division if the salary data is associated with other attributes such as age, rank, sex, and the school where the final degree was earned.

To protect confidential information about each individual, we propose to distort the data by a probability distribution. This method requires three steps:

Step 1. Identification of the underlying density function.

Using the original data, we compute Kolmogorov-Smirnov (K-S) statistic for each of the following density functions.

Density Function	K-S Statistics (D)	Remarks
Poisson	0.16601	
Exponential	0.43726	
Normal	0.11295	
Gamma	0.08597	
Weibull	0.14004	
Lognormal	0.07263	← the best fit
Uniform	0.20096	
Triangular	0.17410	

The K-S statistic D is computed by:

$$D = \max_{\text{all } i} |F_a(X_i) - F_e(X_i)|$$

Where  $F_a(X_i)$ : The  $i$ th observed cumulative relative frequency

$F_e(X_i)$ : The  $i$ th expected cumulative relative frequency.

The choice of the probability density function is based on the following criteria. If the computed  $D$  is smaller than the K-S table value, then the null hypothesis (i.e., the hypothesis that the sample is drawn from the density function being tested) is accepted. For example, the K-S table value is .179 (see Appendix C) when degrees of freedom are 34 and the significance level is 10% (one tail). We may conclude that, at the 10% significance level, the original data has this density function if the obtained  $D$  value is smaller than .179. In fact, the test statistics of all density functions except the Exponential and Uniform distributions are accepted as the density function of the original data at the 10% significance level. However, we set the decision rule to choose the density function which yields the smallest  $D$  value since such choice will maximize the probability of acceptance. Under this decision rule, the Lognormal distribution, with an estimated mean of 31.212 and an estimated standard deviation of 6.674, is selected as the underlying density function of the faculty salary data.

Definition 1:

The Lognormal distribution is the model for a random variable whose Logarithm follows a normal distribution. The Lognormal density function is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} x^{-1} e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}}$$

where  $x > 0$

$$-\infty < \mu < \infty$$

$$\sigma > 0.$$

### Step 2. Data generation.

IMSL (1980) has a subroutine to generate random numbers for the Lognormal distribution. A matching number of random numbers are generated and used as distorted data to replace the original data.

### Step 3. Mapping and replacement.

If this distorted data is used for a statistical analysis independently of other variables, it doesn't need to be mapped onto the original data. However, if it is used with other variables in the same data set, the mapping is necessary to maintain consistency with other data. When the estimated standard deviation is fairly large, the distorted data may have a smallest value and a largest value which are quite different from the corresponding values in the original data. In this case, it is suggested that the average of several replications of the distorted data be used rather than the values generated first. Table 4.8 shows the mean values of the repeatedly distorted data. Monte Carlo study shows that an average of 30 replications provides reasonably good distorted data.

#### 4.2 Small Sampling Efficiency of Probability Data Distortion

To investigate the small sampling efficiency of Probability Data Distortion, we generate a data set by Point Data Distortion by adding a random variable with zero mean and constant variance to each of the original observations, i.e.,

$$Z_i = X_i + \varepsilon_i$$

where  $Z_i$  = the  $i$ th observation of the distorted data by a point data distortion

$X_i$  = the  $i$ th observation of the original data

$\varepsilon_i$  = a random variable with mean of zero and standard deviation the same as the original data (6.461 in this case).

We select two criteria to compare the performance of the Probability Data Distortion with that of the Point Data Distortion.

- (1) Accuracy of parameter estimation.
- (2) Degree of Compromisability if replications are accessible by the curious user.

As measures for comparing the accuracy of statistical estimations we select the following commonly used statistics: mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum. These statistics are estimated by following:

- (1) Original data set.
- (2) The data set which is generated by Point Data Distortion.
- (3) The data set which is generated by Probability Data Distortion.

The number of observations in both distorted data sets is the same as that of the original data set (i.e., Finance 6, Economics 8, Management 11, Accounting 9 thus totaling 34 observations).

The seven statistics are computed once using the point distorted data and again using the probability distorted data; these results are then compared to the same statistics computed using the original data.

First, we made 100 replications each of both distorted data sets. Out of the 100 replications the number of cases which the probability distorted data results in more accurate estimators than does the

point distorted data is tabulated in Table 4.6. This table shows comparisons between the original data and the distorted data.

Except for computation of the mean in the pooled groups, the probability distorted data estimates all seven statistics in the five categories much better than the point distorted data. For example, in the pooled case, the probability distorted data estimates the standard deviation more accurately than does the point distorted data in 98 cases. This means that out of 100 replications, there are only 2 cases in which the point distorted data performs better than the probability distorted data. Similar results are obtained when the data is grouped and the group standard deviations are computed. The number of cases in which the probability distorted data performs better in estimation of the group standard deviation than does the point distorted data are 84 in Finance, 78 in Economics, 89 in Management, and 89 in Accounting. The probability distorted data estimates the extreme value much better than does the point distorted data. The number of cases in which the probability distorted data estimates the minimum value more accurately than does the point distorted data is 65 in Finance, 71 in Economics, 71 in Management, 80 in Accounting, and 85 in the pooled case. The results for estimation of the maximum value are almost the same as those for the minimum. Out of 100 replications, the number of cases in which the probability distorted data performs better than the point distorted data in estimating the maximum value is 82 in Finance, 68 in Economics, 80 in Management, 70 in Accounting, and 76 in the pooled case. In estimating the percentiles (i.e., 25th, median, and 75th), the probability distorted data performs with an average superiority of

3 to 1 over the point distorted data. For example, the probability distorted data estimates the 25th percentile in the Finance group more accurately than does the point distorted data in 83 cases out of 100. In the same category, the probability distorted data surpasses the point distorted data in 73 cases in Economics, 75 in Management, 78 in Accounting and 72 in pooled cases. In estimation of the 75th percentile the number of cases in which the probability distorted data excels the point distorted data is 79 in Finance, 84 in Economics, 74 in Management, 69 in Accounting, and 74 in pooled cases. Similar superiority of the probability distorted data over the point distorted data can be observed in the estimation of group mean or group median. When the seed for the random deviate generator is changed, the basic result (i.e., the superiority of the probability distorted data over the point distorted data) remains the same even though there is some variation in the number of superior replications.

In the second portion of the simulation, the average absolute mean error (AAME) and the Chi-square statistic (CSS) are computed to compare the accuracy of the statistical estimation of each distorted data set. The average absolute mean error is computed by averaging the absolute deviation between the statistics computed by the distorted data and those computed by the original data, i.e.,

$$AAME_j = \frac{1}{G} \sum_{i=1}^G |O_{ij} - \bar{D}_{ij}|$$

The Chi-square statistic (CSS) is computed by:

$$CSS_j = \sum_{i=1}^G (O_{ij} - \bar{D}_{ij})^2 / O_{ij}$$

Where  $O_{ij}$ : the original value of the  $j$ th statistics for  $i$ th group  
 $\bar{D}_{ij}$ : the average distorted value of the  $j$ th statistics in  $n$  replications for the  $i$ th group  
 $G$  : Number of groups (i.e.,  $G=5$ ; Finance, Economics, Management, Accounting and Pooled)

The average absolute mean errors are calculated for each of the seven statistics using both distorted data sets. When the number of replications is 10, the statistics resulting from the probability distorted data have shown much smaller average absolute mean errors in all seven cases. The average absolute mean errors and the Chi-square statistics for both data distortions are shown in Table 4.4 and Table 4.5. Out of a total 42 statistics (i.e., 7 statistics each for 10, 30, 70, 100, 500, and 1000 replications) computed, there are only 6 cases in which the point distorted data results in smaller average absolute mean errors than does the probability distorted data. For example, in 10 replications, the average absolute mean errors for the probability distorted data are 0.164 for the mean, 0.491 for the standard deviation, 2.005 for the minimum value, 0.435 for the 25th percentile, 0.641 for the median, 0.974 for the 75th percentile, and 0.991 for the maximum value. In the same 10 replications, the average absolute mean errors for the point distorted data are 0.771 for mean, 2.138 for standard deviation, and 2.502 for minimum value, 1.354 for 25th percentile, 1.062 for median, 2.655 for 75th percentile, and 3.427 for maximum value. Clearly, these average absolute mean errors are much higher than those for the probability distorted data.

It is interesting to observe that the probability distorted data performs better than the point distorted data even when the number of replications is relatively small (typically smaller than 100 replications). The mean and median are the only cases which the probability distorted data shows larger average absolute mean error than the point distorted data in 500 and 1000 replications. The average absolute mean errors in the computation of the mean are 0.101 for the probability distorted data and 0.094 for the point distorted data for 500 replications. When the replications are increased to 1000, the average absolute mean error of the mean becomes 0.109 for the probability distorted data and 0.070 for the point distorted data. The average absolute mean errors in the estimation of the median by the probability distorted data are 0.550 for 70 replications, 0.564 for 100 replications, 0.534 for 500 replications, and 0.518 for 1000 replications whereas those by the point distorted data are 0.450, 0.398, 0.444, and 0.424. To compare the overall performance, the grand mean of the average absolute mean errors of the seven statistics (i.e., average of AAME of mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) was computed for each level of replications. The results are in Table 4.1.

TABLE 4.1  
GRAND MEAN OF AAME OF THE SEVEN STATISTICS  
(n: Number of Replications)

METHOD	n=10	n=30	n=70	n=100	n=500	n=1000
Prob. Data Distortion	0.814	0.670	0.610	0.608	0.598	0.593
Point Data Distortion	1.987	2.014	1.983	2.008	2.031	2.049

The average AAME of the seven statistics for the probability distorted data is 0.814 for 10 replications and steadily decreases to 0.593 for 1000 replications whereas those for the point distorted data are 1.987 and 2.049 respectively. The results show that the probability distorted data performs better than the point distorted data in any number of replications. As the number of replications is increased, the average AAME of the seven statistics decreases in the probability distorted data with a very slow convergence rate.

Chi-square statistics were also computed to test the goodness of fit. Since the Chi-square table value at the 0.05 significance level is 9.488 with 4 degrees of freedom, we accept the hypothesis that the statistics computed from both distorted data sets are the same as those computed from the original data with 95% reliability.

#### 4.3 Compromisability of Probability Data Distortion

Compromise occurs when a user deduces confidential information of which he was previously unaware from the responses to one or more queries. We say that a database has been positively compromised if the value of a particular data item is known, and that it has been negatively compromised if it is known that a data item does not have a certain value. Partial compromise occurs when information about at least one individual is deduced, and complete compromise occurs when everything in the database is deduced. A database is strongly secure if it cannot be compromised either positively or negatively; it is weakly secure if only positive compromise is impossible.

The primary purpose of data distortion is to guard individual privacy in the original data. Ideally, data should be distorted so that

no user could deduce the original information of any individual. If one set of distorted data replaces the original data, and no other set of distorted data is released, there will be no problem of compromisability. However, if replications are permitted, some distorted data are easily compromisable whereas other distorted data are not. It is said that a distorted data is compromisable if any manipulation of the distorted data easily reveals the original observations.

Comparisons are now made between the compromisability of the Probability Data Distortion and that of the Point Data Distortion. One popular way of identifying the original observation is by averaging replications of the distorted data if repeated queries are permitted. Table 4.8 and Table 4.9 show the average value of each observation when the number of replications is raised to 10, 30, 70, 100, 500, and 1000.

When data is distorted by the Point Data Distortion, as few as 100 repetitions could easily identify the original observations. However, when data is distorted by the Probability Data Distortion, an increase in the number of repetitions won't increase the compromisability by any appreciable amount.

The degree of compromisability is measured in terms of the average absolute percentage deviation of the distorted data from the original data:

$$\text{i.e., } C(N) = \frac{1}{T} \sum_{i=1}^T (|O_i - \bar{D}_i|/O_i)$$

where  $C(N)$  = compromisability index when number of repetitions is  $N$

$O_i$  = the  $i$ th original observation

$\bar{D}_i$  = the  $i$ th mean value of the distorted observations in  $N$  replications

$T$  = number of observations.

The point distorted data is easily compromisable when replications reach more than 70. For example, the compromisability index of the point distorted data becomes 0.019 for Finance, 0.027 for Economics, 0.022 for Management, 0.01 for Accounting, and 0.019 for the Pooled case when the 70 replications are made. The compromisability index in the point distorted data rapidly decreases as the number of replications reaches to 1000. The compromisability index of the point distorted data when the replication becomes 1000 is 0.005 for Finance, 0.007 for Economics, 0.007 for Management, 0.003 for Accounting, and 0.006 for the Pooled case (See Table 4.7).

In contrast with the point distorted data, the probability distorted data does not show increased compromisability through increased replications as is shown in Figure 4.1. For example, when 70 replications are made, the compromisability index of the probability distorted data is 0.034 for Finance, 0.021 for Economics, 0.033 for Management, 0.019 for Accounting, and 0.027 for the Pooled case. Even when the replications are raised to 1000, there is virtually no decrease in the compromisability index. The index in 1000 replications by the probability distorted data is 0.032 for Finance, 0.021 for Economics, 0.032 for Management, 0.018 for Accounting, and 0.026 for the Pooled case. Except for the 10 replications case, the compromisability indices of the probability distorted data are generally higher than those of the point distorted data, implying that the probability distorted data

is much more difficult to compromise than is the point distorted data. Table 4.7 provides compromisability index values of both distorted data sets and the original data under different replications. The reader can easily see that the point distorted data can be compromised by an increase in replications but this technique does not work in the case of the probability distorted data.

FIGURE 4.1  
COMPRACMISABILITY INDEX  
(POOLED CASE)

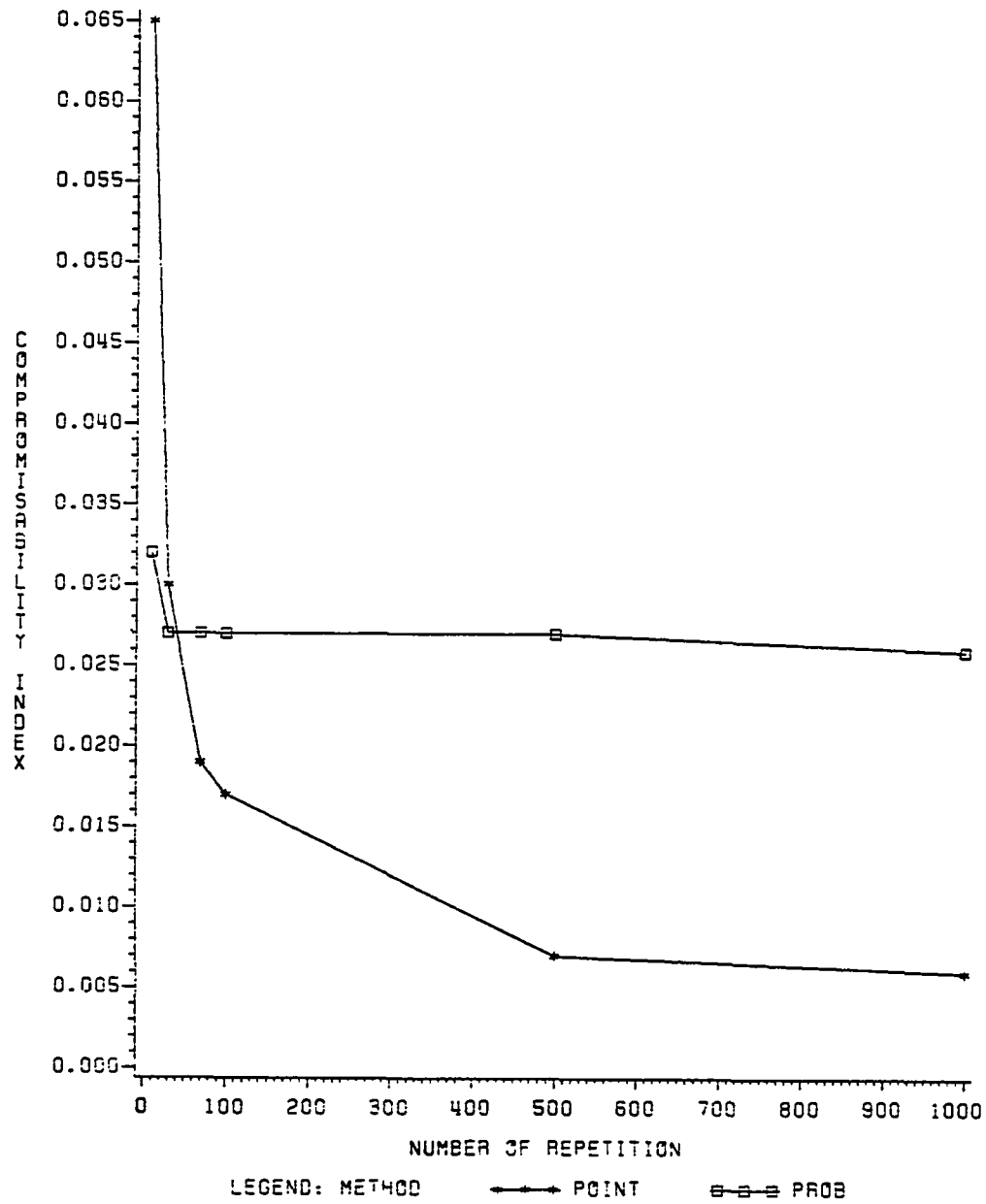


TABLE 4.2

PARAMETER ESTIMATION BY ORIGINAL AND DISTORTED DATA  
PROBABILITY DATA DISTORTION (N: NUMBER OF REPETITIONS)

GROUP	PARAMETERS	ORIGINAL	N=10	N=30	N=70	N=100	N=500	N=1000
FIN	MEAN	27.483	27.247	27.194	27.348	27.346	27.356	27.344
	ST. DEV.	5.476	6.944	6.787	6.551	6.500	6.507	6.536
	MINIMUM	19.600	15.675	16.398	17.244	17.434	17.477	17.409
	25TH P.	23.700	24.398	23.939	23.844	23.794	23.752	23.754
	MEDIAN	28.050	28.100	27.877	27.067	27.804	27.830	27.837
	75TH P.	29.900	31.256	30.885	31.050	31.014	30.984	30.927
	MAXIMUM	35.600	35.955	36.189	36.215	36.224	36.261	36.298
ECON	MEAN	30.638	30.850	30.685	30.792	30.844	30.812	30.817
	ST. DEV.	7.440	7.620	7.628	7.605	7.602	7.595	7.618
	MINIMUM	19.700	19.271	19.541	19.885	20.140	20.116	20.093
	25TH P.	26.750	26.248	26.024	26.001	26.014	25.965	25.980
	MEDIAN	29.700	30.741	30.380	30.481	30.475	30.479	30.448
	75TH P.	33.600	34.491	34.152	34.276	34.259	34.232	34.241
	MAXIMUM	45.300	44.572	44.823	44.933	45.113	45.027	45.107
MGT	MEAN	32.664	32.790	32.467	32.575	32.570	32.566	32.579
	ST. DEV.	6.596	6.032	6.055	6.103	6.060	6.095	6.113
	MINIMUM	20.600	22.042	21.672	21.682	21.799	21.697	21.690
	25TH P.	28.850	29.155	28.828	28.899	28.887	28.942	28.925
	MEDIAN	32.600	33.294	32.864	32.974	32.966	32.914	32.897
	75TH P.	36.850	37.307	37.043	37.126	37.099	37.139	37.164
	MAXIMUM	42.800	40.626	40.523	40.796	40.723	40.712	40.782
ACCT	MEAN	32.311	32.467	32.319	32.430	32.398	32.404	32.430
	ST. DEV.	5.965	6.038	6.162	6.242	6.210	6.232	6.258
	MINIMUM	22.800	23.105	22.820	22.823	22.860	22.838	22.823
	25TH P.	28.700	28.427	28.247	28.306	28.255	28.411	28.426
	MEDIAN	32.600	32.641	32.315	32.412	32.364	32.410	32.378
	75TH P.	35.600	36.919	36.906	36.999	37.024	36.981	37.028
	MAXIMUM	42.800	41.828	42.049	42.410	42.222	42.276	42.371
POOLED	MEAN	31.179	31.270	31.078	31.195	31.196	31.191	31.201
	ST. DEV.	6.460	6.631	6.614	6.597	6.569	6.584	6.611
	MINIMUM	19.600	15.675	16.398	17.244	17.434	17.477	17.409
	25TH P.	27.300	26.901	26.751	26.800	26.751	26.684	26.719
	MEDIAN	30.050	31.428	31.085	31.271	31.247	31.217	31.162
	75TH P.	34.800	35.649	35.551	35.646	35.638	35.656	35.664
	MAXIMUM	45.300	44.572	44.823	44.933	45.113	45.027	45.107

TABLE 4.3  
PARAMETER ESTIMATION BY ORIGINAL AND DISTORTED DATA  
POINT DATA DISTORTION (N: NUMBER OF REPETITIONS)

GROUP	PARAMETERS	ORIGINAL	N=10	N=30	N=70	N=100	N=500	N=1000
FIN	MEAN	27.483	29.038	27.686	27.646	27.691	27.635	27.571
	ST. DEV.	5.476	7.548	8.129	8.033	7.928	8.175	8.201
	MINIMUM	19.600	18.664	17.123	16.946	17.231	16.716	16.642
	25TH P.	23.700	24.202	21.545	22.149	22.244	22.147	21.994
	MEDIAN	28.050	29.040	27.671	27.787	27.790	27.644	27.615
	75TH P.	29.900	33.854	33.511	32.713	32.615	33.082	33.015
	MAXIMUM	35.600	39.426	38.593	38.494	38.477	38.579	38.544
ECON	MEAN	30.638	30.942	30.612	30.503	30.479	30.585	30.525
	ST. DEV.	7.440	9.803	9.763	9.924	9.689	9.509	9.594
	MINIMUM	19.700	16.362	16.712	16.512	16.806	17.068	16.936
	25TH P.	26.750	24.600	24.249	23.947	24.044	24.329	24.199
	MEDIAN	29.700	30.882	30.028	30.022	30.002	30.142	30.050
	75TH P.	33.600	36.599	36.398	36.299	36.324	36.298	36.322
	MAXIMUM	45.300	47.015	46.833	46.980	46.288	46.069	46.121
MGT	MEAN	32.664	33.282	32.322	32.032	32.117	32.476	32.582
	ST. DEV.	6.596	8.760	8.931	8.692	9.022	9.057	9.109
	MINIMUM	20.600	19.429	19.231	18.260	18.090	17.712	17.749
	25TH P.	28.850	27.661	26.328	26.639	26.345	26.930	27.023
	MEDIAN	32.600	32.557	31.693	31.866	31.981	32.487	32.577
	75TH P.	36.850	38.769	37.806	37.366	37.673	38.110	38.171
	MAXIMUM	42.800	47.797	47.209	46.229	46.961	47.035	47.328
ACCT	MEAN	32.311	32.969	32.174	32.274	32.254	32.288	32.350
	ST. DEV.	5.965	7.805	8.222	8.414	8.633	8.590	8.558
	MINIMUM	22.800	21.173	19.398	19.556	19.292	19.366	19.448
	25TH P.	28.700	27.308	26.860	26.812	26.830	27.104	27.291
	MEDIAN	32.600	34.167	33.261	32.790	32.441	32.199	32.190
	75TH P.	35.600	37.379	36.715	37.183	37.252	37.160	37.207
	MAXIMUM	42.800	44.219	43.834	44.962	45.575	45.813	45.825
POOLED	MEAN	31.179	31.900	31.062	30.963	30.987	31.127	31.152
	ST. DEV.	6.460	8.709	8.974	8.960	9.016	9.042	9.084
	MINIMUM	19.600	14.162	12.679	12.614	12.938	12.594	12.547
	25TH P.	27.300	25.761	24.957	25.010	24.916	25.055	25.020
	MEDIAN	30.050	31.581	30.964	30.790	30.699	30.910	30.953
	75TH P.	34.800	37.424	36.669	36.735	36.867	37.099	37.132
	MAXIMUM	45.300	50.475	50.853	50.324	50.451	50.683	50.761

TABLE 4.4  
AVERAGE ABSOLUTE MEAN ERROR  
(N: NUMBER OF REPETITIONS)

PROBABILITY DATA DISTORTION

PARAMETER	N=10	N=30	N=70	N=100	N=500	N=1000
MEAN	0.164	0.128	0.103	0.108	0.101	0.109
ST. DEV.	0.491	0.478	0.429	0.415	0.416	0.433
MINIMUM	2.005	1.531	1.200	1.206	1.159	1.178
25TH P.	0.435	0.398	0.367	0.372	0.367	0.351
MEDIAN	0.641	0.487	0.550	0.564	0.534	0.158
75TH P.	0.974	0.757	0.870	0.857	0.848	0.855
MAXIMUM	0.991	0.914	0.749	0.731	0.764	0.706

POINT DATA DISTORTION

PARAMETER	N=10	N=30	N=70	N=100	N=500	N=1000
MEAN	0.771	0.165	0.236	0.233	0.094	0.070
ST. DEV.	2.138	2.416	2.417	2.470	2.487	2.522
MINIMUM	2.502	3.431	3.683	3.588	3.769	3.796
25TH P.	1.354	2.272	2.149	2.184	1.947	1.955
MEDIAN	1.062	0.638	0.450	0.398	0.444	0.424
75TH P.	2.655	2.070	1.909	1.996	2.200	2.219
MAXIMUM	3.427	3.104	3.038	3.190	3.276	3.356

TABLE 4.5  
CHI-SQUARES  
(N: NUMBER OF REPETITIONS)

PROBABILITY DATA DISTORTION

PARAMETERS	N=10	N=30	N=70	N=100	N=500	N=1000
MEAN	0.005	0.005	0.002	0.003	0.002	0.002
ST. DEV.	0.451	0.373	0.267	0.250	0.250	0.263
MINIMUM	1.686	1.103	0.625	0.559	0.527	0.555
25TH P.	0.042	0.040	0.037	0.039	0.040	0.037
MEDIAN	0.115	0.057	0.077	0.076	0.072	0.066
75TH P.	0.160	0.107	0.135	0.133	0.128	0.129
MAXIMUM	0.159	0.154	0.114	0.121	0.124	0.115

POINT DATA DISTORTION

PARAMETERS	N=10	N=30	N=70	N=100	N=500	N=1000
MEAN	0.133	0.006	0.015	0.013	0.002	0.001
ST. DEV.	3.595	4.669	4.662	4.875	5.010	5.129
MINIMUM	2.302	3.809	4.093	3.821	4.202	4.260
25TH P.	0.387	0.969	0.881	0.911	0.722	0.741
MEDIAN	0.235	0.075	0.042	0.032	0.042	0.043
75TH P.	1.177	0.829	0.667	0.685	0.818	0.821
MAXIMUM	1.698	1.463	1.239	1.424	1.533	1.609

TABLE 4.6  
 NUMBER OF CASES IN WHICH THE PROBABILITY DATA DISTORTION IS  
 BETTER THAN THE POINT DATA DISTORTION (TOTAL 100 CASES)

GROUP	MEAN	ST DEV	MIN	25TH P	MEDIAN	75TH P	MAX
FINANCE	74	84	65	83	74	79	82
ECONOMICS	71	78	71	73	72	84	68
MANAGEMENT	68	89	71	75	65	74	80
ACCOUNTING	58	89	80	78	64	69	70
POOLED	47	98	85	72	51	74	76

TABLE 4.7  
 DEGREE OF COMPROMISABILITY  
 (N: NUMBER OF REPETITIONS)

PROBABILITY DATA DISTORTION

GROUP	N=10	N=30	N=70	N=100	N=500	N=1000
FINANCE	0.053	0.042	0.034	0.032	0.032	0.032
ECONOMICS	0.025	0.019	0.021	0.022	0.022	0.021
MANAGEMENT	0.036	0.033	0.033	0.034	0.033	0.032
ACCOUNTING	0.021	0.019	0.019	0.020	0.019	0.018
POOLED	0.032	0.027	0.027	0.027	0.027	0.026

POINT DATA DISTORTION

GROUP	N=10	N=30	N=70	N=100	N=500	N=1000
FINANCE	0.088	0.027	0.019	0.019	0.007	0.005
ECONOMICS	0.057	0.037	0.027	0.018	0.006	0.007
MANAGEMENT	0.052	0.036	0.021	0.023	0.010	0.007
ACCOUNTING	0.072	0.019	0.010	0.009	0.006	0.003
POOLED	0.065	0.030	0.019	0.017	0.007	0.006

TABLE 4.8  
FACULTY SALARY BY PROBABILITY DATA DISTORTION  
(UNIT: THOUSAND DOLLARS)

AVERAGE VALUE OF EACH OBSERVATION  
(N: NUMBER OF REPETITIONS)

GROUP	ORIGINAL	N=10	N=30	N=70	N=100	N=500	N=1000
FIN	19.600	15.675	16.398	17.244	17.434	17.477	17.409
	23.700	24.398	23.939	23.844	23.794	23.752	23.754
	27.300	26.901	26.751	26.800	26.751	26.684	26.719
	28.800	29.299	29.004	28.934	28.858	28.976	28.956
	29.900	31.256	30.885	31.050	31.014	30.984	30.927
	35.600	35.955	36.189	36.215	36.224	36.261	36.298
ECON	19.700	19.271	19.541	19.885	20.140	20.116	20.093
	25.600	24.963	24.808	24.668	24.736	24.620	24.628
	27.900	27.533	27.240	27.334	27.292	27.310	27.332
	29.200	29.882	29.476	29.470	29.470	29.509	29.498
	30.200	31.600	31.285	31.492	31.480	31.449	31.397
	33.300	33.773	33.335	33.478	33.463	33.404	33.415
	33.900	35.210	34.969	35.075	35.055	35.060	35.067
	45.300	44.572	44.823	44.933	45.113	45.027	45.107
MGT	20.600	22.042	21.672	21.682	21.799	21.697	21.690
	26.900	25.645	25.378	25.367	25.418	25.345	25.380
	28.500	27.953	27.765	27.831	27.767	27.861	27.883
	29.200	30.357	29.892	29.968	30.007	30.022	29.967
	30.300	32.053	31.728	31.926	31.935	31.917	31.869
	32.600	33.294	32.864	32.974	32.966	32.914	32.897
	33.400	34.314	33.894	34.029	34.005	33.940	33.977
	34.800	35.649	35.551	35.646	35.638	35.656	35.664
	38.900	38.966	38.535	38.605	38.560	38.623	38.664
	41.300	39.793	39.338	39.500	39.450	39.535	39.594
	42.800	40.626	40.523	40.796	40.723	40.712	40.782
ACCT	22.800	23.105	22.820	22.823	22.860	22.838	22.823
	27.300	26.182	26.063	26.117	26.105	26.021	26.079
	28.700	28.427	28.247	28.306	28.255	28.411	28.426
	29.800	30.785	30.488	30.596	30.573	30.494	30.460
	32.600	32.641	32.315	32.412	32.364	32.410	32.378
	33.700	34.660	34.423	34.538	34.506	34.479	34.518
	35.600	36.919	36.906	36.999	37.024	36.981	37.028
	37.500	37.658	37.561	37.669	37.673	37.724	37.784
	42.800	41.828	42.049	42.410	42.222	42.276	42.371

TABLE 4.9  
FACULTY SALARY BY POINT DATA DISTORTION  
(UNIT: THOUSAND DOLLARS)  
  
AVERAGE VALUE OF EACH OBSERVATION  
(N: NUMBER OF REPETITIONS)

GROUP	ORIGINAL	N=10	N=30	N=70	N=100	N=500	N=1000
FIN	19.600	24.453	20.063	20.261	20.605	19.510	19.626
	23.700	26.651	24.465	23.706	23.828	23.935	23.736
	27.300	29.652	28.054	28.628	28.169	27.813	27.682
	28.800	28.404	29.708	28.725	28.670	28.800	28.815
	29.900	30.644	29.486	29.210	29.450	29.890	29.758
	35.600	34.422	34.339	35.346	35.424	35.864	35.808
ECON	19.700	17.303	18.522	18.436	19.331	19.479	19.326
	25.600	26.948	26.457	26.005	25.775	25.716	25.753
	27.900	28.645	26.546	27.835	28.153	28.004	27.899
	29.200	27.518	28.143	27.973	28.535	29.134	29.141
	30.200	34.433	32.352	31.200	30.824	30.501	30.344
	33.300	32.365	32.863	32.225	31.893	32.950	32.880
	33.900	34.008	33.660	33.937	33.521	33.732	33.609
	45.300	46.320	46.350	46.417	45.803	45.159	45.247
MGT	20.600	20.859	21.641	20.421	20.577	20.505	20.690
	26.900	25.229	24.904	26.490	25.876	26.439	26.585
	28.500	31.357	29.178	28.654	28.682	28.995	28.855
	29.200	30.502	28.983	28.180	27.985	28.773	28.841
	30.300	28.291	28.195	29.538	29.775	29.828	29.843
	32.600	34.010	31.069	31.767	31.298	32.021	32.293
	33.400	35.619	34.178	32.579	31.864	33.121	33.141
	34.800	36.299	35.076	35.165	35.352	34.709	34.905
	38.900	36.726	36.822	36.347	37.700	38.978	39.005
	41.300	43.709	42.780	41.134	41.602	41.210	41.373
	42.800	43.506	42.717	42.082	42.575	42.654	42.867
ACCT	22.800	24.437	22.876	23.113	22.894	22.892	22.826
	27.300	32.409	27.239	26.793	26.564	27.298	27.435
	28.700	30.781	29.710	28.588	29.000	28.549	28.858
	29.800	28.664	30.266	29.775	29.534	29.363	29.827
	32.600	33.625	32.569	32.622	32.594	32.535	32.494
	33.700	31.451	32.310	33.708	33.673	33.371	33.544
	35.600	38.657	35.852	36.297	35.486	35.676	35.524
	37.500	35.570	37.997	37.763	38.131	37.795	37.524
	42.800	41.128	40.751	41.809	42.406	43.109	43.116

## CHAPTER V

### FREQUENCY IMPOSED DATA DISTORTION

#### 5.0 Introduction

We introduced data distortion by probability distribution, Probability Data Distortion, which preserves asymptotically the statistical properties of the original data. One drawback of the Probability Data Distortion at the present time is the limited choice offered by available density functions.

To make this mechanism more flexible, we introduce Frequency Imposed Data Distortion which doesn't require identification of an underlying density function. Instead, the original data is divided into several intervals and the frequency in each interval is recorded. These frequencies are used as guidelines to generate the distorted data. By using a uniform random number generating routine, distorted data points are generated so that the frequency in the distorted data in each interval coincides with that of the original data.

Another method which we introduce in this chapter is the Frequency Imposed Probability Data Distortion which is a hybrid of the Probability Data Distortion and the Frequency Imposed Data Distortion.

Like Probability Data Distortion, this method requires identification of a density function which is best fitted to the original data set. Just as in Frequency Imposed Data Distortion, the original data set is divided into several intervals and the proper number of distorted data points are generated for each interval. The best fitted density function is used to generate random numbers to be used as the distorted data set, and the frequency in each interval of the distorted data set is forced to be equal to the frequency in the matching interval of the original data set.

Through the use of actual faculty salary data, the performance of the two data distortion methods in terms of accuracy in parameter estimation and in compromisability can be compared. Accuracy is determined by the capability to maintain the original statistical parameters such as mean, standard deviation, maximum value, minimum value, and percentiles. Compromisability is measured by the capability to protect individual information from compromise.

Section 5.1 briefly discusses the procedure by which each method generates the distorted data, Section 5.2 describes the empirical results, and Section 5.3 provides a brief concluding remark.

### 5.1 A Monte Carlo Study

The 1982-83 faculty salary data file of the College of Business Administration, The University of Oklahoma, was chosen as the original data. The salary file contains the salary of each faculty member by rank, department, sex, and contract period. We changed all twelve-month salaries to nine-month equivalents for this study. The frequency distribution of the salary data is shown in Table 5.1.

TABLE 5.1  
FREQUENCY TABLE BY SALARY RANGE

Salary Range	Frequencies
\$20K - \$25K	8
\$25K - \$30K	18
\$30K - \$35K	29
\$35K - \$40K	13
\$40K - \$45K	9
\$45K - \$50K	5
\$50K - \$55K	2
Total	84

A joint frequency table by rank and department is shown in Table 5.2.

TABLE 5.2  
FREQUENCY TABLE BY RANK AND DEPARTMENT

Rank	Department						Total
	ECON	ACCT	FIN	EAP	MKT	MGT	
Assistant Prof	6	6	4	4	1	7	28
Associate Prof	3	1	5	3	6	5	23
Full Prof	9	9	3	1	3	8	33
Total	18	16	12	8	10	20	84

Release of the original salary data information as it is can easily lead to the identification of the salary of each faculty member during the 1982-83 academic year since there are small frequencies in the various rank and department classifications. For example, there is only one associate professor in the Accounting department. Some of these professors are appointed for twelve months instead of nine months. This indicates that he is in an administrative position as a department director. Furthermore, the sex indication will easily sort out the female from the male professors.

Suppose we wish to guard the privacy of an individual from the identification of his salary since faculty salary like a student's grade may reflect the individual's ability. We consider the salary data as confidential information and the other data such as rank, department, sex, and contract period as non-confidential information.

Which data distortion method will best reproduce the actual statistics of the faculty salary data while protecting the privacy of an individual at minimum cost?

Seven popular statistical parameters: mean, standard deviation, 25th percentile, median, 75th percentile, maximum, and minimum are selected for measurement of statistical accuracy.

We consider four different data distortion methods:

- (1) Point Data Distortion
- (2) Probability Data Distortion
- (3) Frequency Imposed Data Distortion
- (4) Frequency Imposed Probability Data Distortion.

Distorted data sets are generated through the following procedures.

#### 5.1.1 Point Data Distortion

The distorted data is computed by adding a random variable to each original observation. The random variable is generated with zero mean and the standard deviation (6.818) of the original data as shown in Section 4.2.

#### 5.1.2 Probability Data Distortion

The faculty salary data was screened to find the best fitted density function. The Kolmogorov-Smirnov statistic was used to identify such density function. These are computed by using the Phillip's computer package (1972), and are shown below:

Density Function	K-S Statistics	Remarks
Poisson	0.17616	
Exponential	0.46387	
Normal	0.11822	
Gamma	0.09133	
Weibull	0.14780	
Lognormal	0.07885	← The best fit
Uniform	0.26255	
Triangular	0.23225	

The K-S table value is 0.117 (see Appendix C) when the degrees of freedom are 84 and the significant level is 10% (one tail). The test statistics of both the Gamma and Lognormal distributions are accepted as the density function for this actual salary data at 10% significant level. However, since the K-S statistic for Lognormal is smaller than

the one for Gamma, the lognormal distribution is selected as the density function best fitted to the original data. By using the lognormal random number generating routine (IMSL (1980)) with an estimated mean of 33.738 and an estimated standard deviation of 6.760, the distorted data were generated and then mapped onto the original salary data.

#### 5.1.3 Frequency Imposed Data Distortion

No attempt is made to identify the underlying density function from the original data. The salary data is divided into seven intervals and the frequency within each interval is counted. Through the use of a uniform random number generating routine, the distorted data are generated so that the frequency counts of the distorted data for each interval will be the same as those for that interval in the original data. For example, there are 8 original data points in the salary range between 20K and 25K. Eight numbers between 20K and 25K are generated using the uniform random number generating routine. A similar method is used to generate the distorted values for the remaining intervals.

#### 5.1.4 Frequency Imposed Probability Data Distortion

Frequency Imposed Probability Data Distortion is a hybrid of Probability Data Distortion and Frequency Imposed Data Distortion. This method requires essentially the same procedure as Probability Data Distortion except that it continues to generate distorted data until the frequency of data points in each interval becomes the same as that in the original data. The original data are grouped into seven intervals as shown on Table 5.2. Previously the lognormal density function was

chosen as the best fitted function for the salary data. The distorted data are generated by using this function and at the same time the frequency in each interval is forced to become the same as that in the original data. If the lognormal random number generator placed more than 8 data points in the first interval, any excess points are simply discarded. The major advantage of this approach is that it forces the distorted data to be the same as the original data not only in overall density function but also in frequency by interval. This double restriction should improve the accuracy of the statistical parameter estimations.

## 5.2 Empirical Results

### 5.2.1 Accuracy of Parameter Estimation

One hundred replications of the 84 salary observations were made by employing each of the four data distortion methods. For each replication, seven parameters (i.e., mean, standard deviation, 25th percentile, median, 75th percentile, maximum, and minimum) were computed and the results were compared.

All four methods estimate mean and median fairly close to those of the original data. The mean and median of the original salary data are 33.74 and 32.80 respectively. The average value of means of 100 replications is 33.62 by Point Data Distortion, 33.79 by Probability Data Distortion, 33.66 by Frequency Imposed Data Distortion, and 33.63 by Frequency Imposed Probability Data Distortion. The standard deviation by means of 100 replications is 0.73 by Point Data Distortion, 0.71 by Probability Data Distortion, 0.12 by Frequency Imposed Data Distortion, and 0.13 by Frequency Imposed Probability Data Distortion.

Point Data Distortion shows very poor performance in the estimation of the 25th and 75th percentiles. It consistently overestimates the 75th percentile, and underestimates the 25th. The 75th percentile of the original salary data is 37.47, and Point Data Distortion yields an average of 39.81 for the 75th percentile in 100 replications which is higher than the 75th percentile of the original salary data.

In general, Point Data Distortion gives poor estimates of the maximum and minimum values. The maximum and minimum values in the original salary data are 54.00 and 21.00 respectively. The Point Data Distortion consistently overestimates the maximum value and underestimates the minimum. In 100 replications, the average of maximum values computed through Point Data Distortion is 59.50 which is far higher than the original maximum of 54.00. In the same experiment, Point Data Distortion yields an average minimum of 11.96 which is much smaller than the original minimum salary of 21.00.

Point Data Distortion also overestimates the standard deviation of the salary data. The average standard deviation computed using Point Data Distortion in 100 replications is 9.65 which is much higher than that of the original standard deviation of 6.82.

The statistics computed using Probability Data Distortion, Frequency Imposed Data Distortion, and Frequency Imposed Probability Data Distortion yield values very close to those of the original. For example, the average standard deviation in 100 replications is 6.77 for Probability Data Distortion, 7.34 for Frequency Imposed Data Distortion, and 7.19 for Frequency Imposed Probability Data Distortion. As reported earlier, the standard deviation of the original salary data is 6.82. In

the same 100 replications, the average of the maximum value is 50.33 for Probability Data Distortion, 53.12 for Frequency Imposed Data Distortion, and 53.01 for the Frequency Imposed Probability Data Distortion. The original maximum value of the salary data is 54.00. The average minimum value in 100 replications is 17.22 for Probability Data Distortion, 20.43 for Frequency Imposed Data Distortion, and 18.29 for Frequency Imposed Probability Data Distortion. The minimum value of the original data is 21.00.

To make the performance of four distortion methods comparable, weights are assigned; 4 for the closest to the original, 3 for the second closest, 2 for the third and 1 for the last. For each replication, 84 distorted observations are generated by using each of the four distortion methods. Using the 84 observations, the mean, standard deviation, minimum value, 25th percentile, median, 75th percentile, and maximum value are computed. These statistics are compared with those computed from the original salary data, and by using the ranking system, we score the numerical rating of each distortion method.

For the mean computation, the average ranking score in 100 replications is 1.80 by Point Data Distortion, 1.73 by Probability Data Distortion, 3.42 by Frequency Imposed Data Distortion, and 3.05 by Frequency Imposed Probability Data Distortion. This implies that Frequency Imposed Data Distortion leads to better estimation of the original mean than does any other method. Frequency Imposed Probability Data Distortion places second.

Frequency Imposed Data Distortion performs best in the estimation of the minimum value, maximum value, median, and mean in the

experiment. Frequency Imposed Data Distortion scores an average of 3.81 for minimum value, 3.01 for median, 3.48 for maximum value, and 3.42 for mean. However, Frequency Imposed Probability Data Distortion closely follows in second place. The average ranking of 100 replications by Frequency Imposed Probability Data Distortion is 2.67 for the minimum value, 3.00 for the median, 2.91 for the maximum value, and 3.05 for the mean.

In the same experiment, Frequency Imposed Probability Data Distortion performs best among these four distortion methods in estimating the 25th percentile and the 75th percentile with average ranking scores of 3.20 and 3.34 respectively. Probability Data Distortion edges Frequency Imposed Probability Data Distortion with an average ranking score of 3.26 to 3.24 in estimating the standard deviation.

The last experiment involves counting the number of times each technique resulted in the best estimation. In estimating the standard deviation, Probability Data Distortion places first 53 times out of 100 replications and is followed by Frequency Imposed Probability Data Distortion with 35, and by Frequency Imposed Data Distortion with 12. Point Data Distortion fails to make a single first place in the standard deviation estimation. Generally, the extreme values are estimated very well through Frequency Imposed Data Distortion. For example, the minimum value of original salary data was 82 times out of 100 most closely estimated through Frequency Imposed Data Distortion. Similarly, Frequency Imposed Data Distortion shows 61 first places in 100 replications on the maximum value. Frequency Imposed Data Distortion also performs well in estimating the mean and median. Out of 100 replications,

Frequency Imposed Data Distortion scores 57 first places in mean estimation, and 40 first places in median estimation. Frequency Imposed Probability Data Distortion scores 31 first places in the mean and 30 first places in the median estimations.

Composite scores are computed to compare overall performance of the four distortion methods. As a composite score, the grand mean of average ranking scores of seven statistics (i.e., average of the average ranking scores of mean, standard deviation, minimum, the 25th percentile, median, the 75th percentile, and maximum in 100 replications) was chosen. The first place in composite score goes to Frequency Imposed Data Distortion with 3.16 closely followed by Frequency Imposed Probability Data Distortion with 3.06, Probability Data Distortion places third with a score of 2.30 and Point Data Distortion is last with a score of 1.49.

Similar conclusions can be obtained when the first place scores are summed over 7 statistics. Total first places by Frequency Imposed Data Distortion amounts to 312 out of 700 (or 45%). It is followed by Frequency Imposed Probability Data Distortion 227 (or 32%). Probability Data Distortion is the third by showing 124 first places (or 18%), and Point Data Distortion comes to last place with 37 first places (or 5%) in all seven statistics.

#### 5.2.2 Accuracy in Computation of Conditional Statistics

Performance of the distorted data when used jointly with other non-confidential variables is now evaluated. The original salary data are divided into six departmental groups. For each departmental

classification, group mean, group standard deviation, maximum value within the group, minimum value within the group, the 75th percentile, the 25th percentile, and median within each group are calculated once using original and again for each of the distorted data sets.

For each statistic, the result computed from the distorted data is compared with that from the original salary data. Since seven statistics (i.e., mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) are used, the rank scores within each group are added to produce subtotal rank score for the department. For example, the subtotal rank score of the Accounting subgroup is 22.29 by Frequency Imposed Probability Data Distortion. The 22.29 implies that the average rank score of seven statistics is around 3.18 by Frequency Imposed Probability Data Distortion.

Then, this subgroup total is summed to total rank. For example, total rank score by Frequency Imposed Probability Data Distortion is 133.50 or an average of 3.18. This average 3.18 is an indicator of overall performance of the distorted data in estimating the group statistics.

The overall winner in group parameter estimation is Frequency Imposed Probability Data Distortion with an average rank score of 3.18, closely followed by Frequency Imposed Data Distortion which has an average rank score of 3.12. A distant third place goes to Probability Data Distortion with the average rank score of 2.29 and the last place finisher is Point Data Distortion with the average rank score of 1.42.

In this experiment, 42 statistics (i.e., 7 statistics by each of 6 departments) were computed through each of four distortion methods

in 100 replications. Out of possible 4200, Frequency Imposed Probability Data Distortion scores 1767 first places (or 42.07%), followed by Frequency Imposed Data Distortion with 1603 first places (or 38.17%). Probability Data Distortion finishes third with 621 first places (or 14.79%) and Point Data Distortion finishes last with 209 first places (or 4.98%).

Another experiment involves the classification of the salary data by professional rank (i.e., Assistant Professor, Associate Professor, and Full Professor). There are 21 statistics (i.e., 7 statistics by each of those 3 groups) to be computed through each of four distortion methods in 100 replications. The result is compared by using the same weight system (i.e., 4 for the best, . . . , 1 for the worst). The overall winner of this experiment is still Frequency Imposed Probability Data Distortion which scores 3.17. However, it is closely followed by the Frequency Imposed Data Distortion which scores 3.13. The third place goes to Probability Data Distortion with 2.28 and Point Data Distortion takes last place. In the number one rank counting, Frequency Imposed Probability Data Distortion takes 853 first places out of a possible 2100. Frequency Imposed Data Distortion is a close second with 828 first places (or 39.43%). The third and fourth places go to Probability Data Distortion and Point Data Distortion which make 297 and 122 first places respectively.

Next we evaluate the accuracy of estimation on seven statistics within each department and within each professional rank. Using the same rank score, we sum the total rank score in departmental classification. In the computation of the seven statistics within the Accounting

subgroup, Frequency Imposed Probability Data Distortion scores 22.29 (or an average of 3.18) and it is followed by Frequency Imposed Data Distortion, which has a total rank score of 20.72 (or an average of 2.96). Probability Data Distortion finishes third with a total rank score of 17.37 (or average of 2.48). Point Data Distortion makes a distant fourth place with total score of 9.62 (or an average of 1.37). Frequency Imposed Probability Data Distortion continuously excels in estimation of subgroup statistics in the Economics, Management, Marketing, Associate Professor, and Full Professor groups. However, in three occasions out of nine subgroups, Frequency Imposed Data Distortion narrowly beats Frequency Imposed Probability Data Distortion. In all cases, Point Data Distortion takes a distant last place and Probability Data Distortion takes a firm third place. Table 5.3 shows the details of these statistics.

### 5.2.3 The Compromisability of the Distorted Data

Some of the distorted data will converge to the original and will be easily compromisable if we average repeated observations. The compromisability of each of these four distortion methods is now compared.

The degree of compromisability is measured by a compromisability index which is the average absolute percentage deviation from the original observation as shown in Section 4.3, and average of the distorted observations in  $N$  replications.

In this experiment, the number of replications is raised from 10 to 1000 to observe whether there is any trend toward decreasing

compromisability index as the number of replications increases. At 10 replications, the compromisability index becomes 0.08266 for Point Data Distortion, 0.06980 for Probability Data Distortion, 0.05302 for Frequency Imposed Data Distortion, and 0.05325 for Frequency Imposed Probability Data Distortion. At 70 replications, the compromisability index decreased to 0.02117 for Point Data Distortion, 0.02896 for Probability Data Distortion, 0.01209 for Frequency Imposed Data Distortion, and 0.01357 for Frequency Imposed Probability Data Distortion. As the number of replications is raised to 100, 500, and 1000, there is a sharp decrease in the compromisability index for Point Data Distortion. At 100 replications, the compromisability index for Point Data Distortion reaches 0.01677 and the index decreases to 0.00552 at 1000 replications. As the number of replications goes up, such a rapid decrease in the compromisability index is a clear indication that Point Data Distortion fails to guard the privacy of the individual if such replications are accessible by the user.

In contrast, the compromisability index remains stable in spite of a drastic increase in the number of replications of the other three distortion methods. As shown on Figure 5.1, the compromisability index of Probability Data Distortion is decreased from 0.02896 in 70 replications to 0.02651 in 1000 replications. Similarly, the compromisability index of Frequency Imposed Probability Data Distortion has gone down from 0.01357 in 70 replications to 0.01272 in 1000 replications. In the case of Frequency Imposed Data Distortion the compromisability index has actually increased from 0.01209 to 0.01505 as the number of replications is raised from 70 to 1000. It is evident that the data distorted by

Probability Data Distortion, by Frequency Imposed Data Distortion, or by Frequency Imposed Probability Data Distortion is very difficult to compromise by averaging the observations even with a very large number of replications. However, Point Data Distortion can be easily compromised if sufficient replications are permitted. Table 5.4 shows the details of the compromisability index.

When the frequency is imposed, the distorted data performs better for obvious reasons. However, such frequency imposition did not yield increased compromisability. Whether the frequency is imposed or not, compromisability in the case of Probability Data Distortion remains quite low.

### 5.3 Summary

Frequency Imposed Probability Data Distortion takes more computer time than any other distortion mechanism because it requires several steps to generate distorted data. Second in consumption of computer time is Probability Data Distortion. Frequency Imposed Data Distortion consumes the least computer time and doesn't require any special package to identify the underlying density function of the original data.

In summary, Frequency Imposed Data Distortion is easy to use, takes less computer time, produces very accurate statistics, and is very difficult to compromise. It appears to be the best data distortion method if the underlying density function of the original data is not easily identifiable.

FIGURE 5.1  
COMPROMISABILITY INDEX  
(TOTAL CASE)

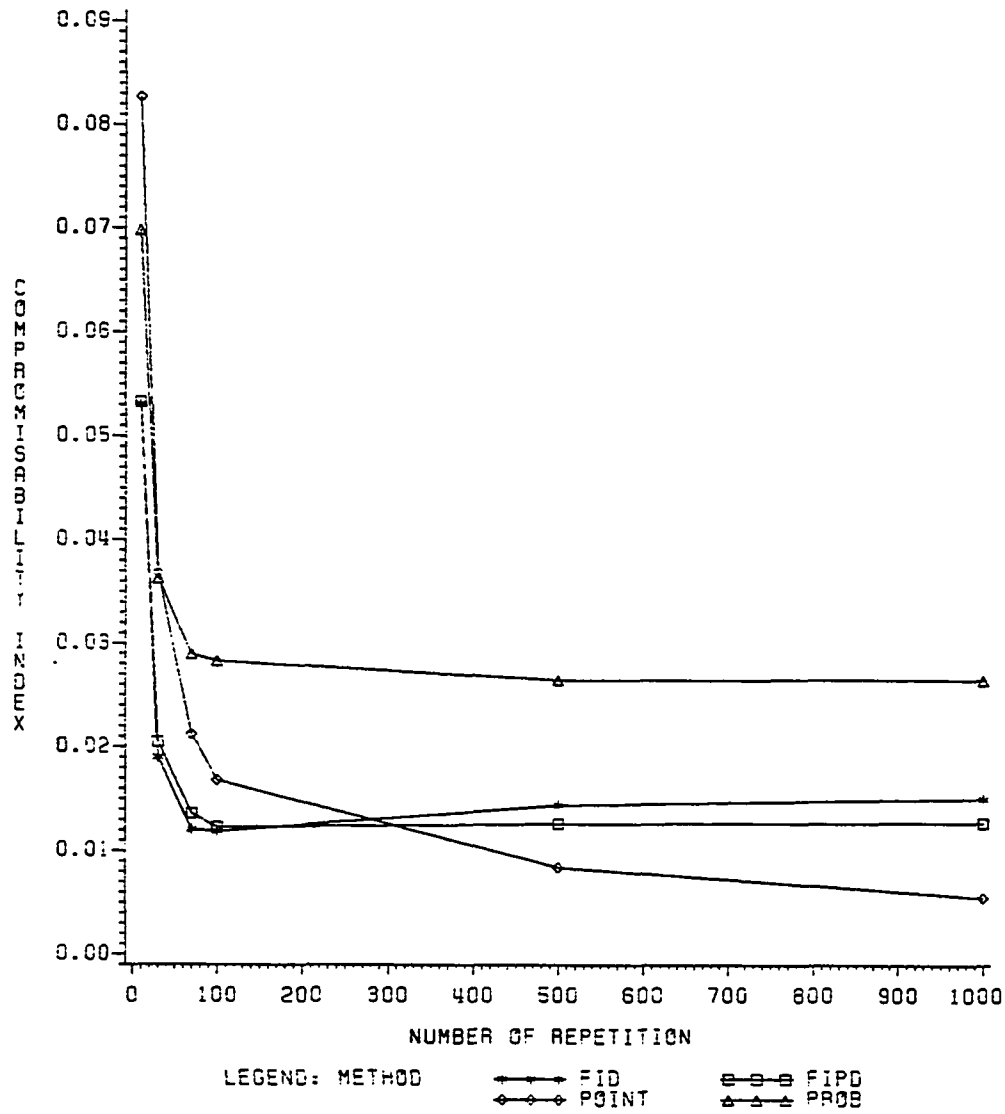


TABLE 5.3  
ACCURACY IN COMPUTATION IN 7 STATISTICS  
(SUMMARY)

DISTORTION METHODS

	POINT		PROBABILITY		FID		FIPD	
	AVERAGE SCORE	TOTAL 1ST PLACES	AVERAGE SCORE	TOTAL 1ST PLACES	AVERAGE SCORE	TOTAL 1ST PLACES	AVERAGE SCORE	TOTAL 1ST PLACES
(A) CLASSIFICATION BY DEPARTMENT								
ACCT (16)*	1.37	31	2.48	150	2.96	226	3.18	293
EAP ( 8)	1.57	48	2.21	82	3.36	373	2.85	197
ECON (18)	1.38	36	2.42	148	3.06	240	3.15	276
FIN (12)	1.37	30	2.15	70	3.33	337	3.15	263
MGT (20)	1.41	32	2.23	92	3.03	230	3.32	346
MKT (10)	1.39	32	2.22	79	2.97	197	3.42	392
-----								
TOTAL (84)		209		621		1603		1767
PERCENTAGE	1.42	4.98%	2.29	14.79%	3.12	38.17%	3.18	42.07%
(B) CLASSIFICATION BY RANK								
A PROF (28)	1.45	47	2.40	117	3.14	303	3.02	233
AS PROF (23)	1.26	18	2.21	76	3.19	280	3.34	326
F PROF (33)	1.56	57	2.22	104	3.05	245	3.16	294
-----								
TOTAL (84)		122		297		828		853
PERCENTAGE	1.42	5.81%	2.28	14.14%	3.13	39.43%	3.17	40.62%
(C) NO CLASSIFICATION								
TOTAL		37		124		312		227
PERCENTAGE	1.49	5.29%	2.30	17.71%	3.16	44.57%	3.06	32.43%

\*THE FIGURE INSIDE OF THE PARENTHESIS IS THE NUMBER OF OBSERVATIONS.

# ACCOUNTING

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	35.85	9.53	19.08	29.27	35.85	42.26	53.12		
PROB	35.90	6.27	21.81	32.73	35.15	39.85	45.58		
FID	35.90	7.09	22.22	31.87	33.86	40.05	48.35		
FIPD	35.78	6.78	22.19	32.03	33.94	39.96	47.27		
ORG	35.83	6.41	23.70	32.50	33.80	39.95	47.00		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.66	1.68	4.48	2.76	2.13	2.64	4.69		
PROB	0.77	0.54	1.61	0.92	0.83	1.11	1.49		
FID	0.16	0.25	0.87	0.36	0.28	0.24	0.88		
FIPD	0.23	0.23	1.23	0.42	0.30	0.26	0.90		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	5.00	2.00	9.00	8.00	2.00	1.00	4.00	31.00	4.43
PROB	16.00	49.00	20.00	33.00	4.00	7.00	21.00	150.00	21.43
FID	49.00	9.00	32.00	20.00	46.00	52.00	18.00	226.00	32.29
FIPD	30.00	40.00	39.00	39.00	48.00	40.00	57.00	293.00	41.86
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.45	1.15	1.66	1.37	1.44	1.32	1.23	9.62	1.37
PROB	2.30	3.24	2.48	2.82	1.78	2.06	2.69	17.37	2.48
FID	3.26	2.37	2.93	2.71	3.38	3.39	2.68	20.72	2.96
FIPD	2.99	3.24	2.93	3.10	3.40	3.23	3.40	22.29	3.18

# ENVIRONMENTAL ANALYSIS AND POLICY

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	26.90	8.34	14.71	21.05	26.93	32.73	39.05		
PROB	26.08	6.65	16.75	20.17	26.60	31.59	35.19		
FID	26.53	5.17	20.53	21.41	26.40	31.08	33.98		
FIPD	26.15	5.87	18.30	20.54	26.76	31.12	34.06		
ORG	26.94	4.84	21.00	22.25	26.86	31.30	33.70		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	2.24	1.89	4.63	2.68	2.97	3.18	4.24		
PROB	0.93	0.92	3.01	1.49	1.00	0.94	0.99		
FID	0.25	0.29	0.51	0.65	0.45	0.37	0.35		
FIPD	0.54	0.75	1.87	1.51	0.50	0.31	0.31		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	9.00	2.00	4.00	14.00	10.00	8.00	1.00	48.00	6.86
PROB	20.00	2.00	3.00	17.00	18.00	12.00	10.00	82.00	11.71
FID	55.00	81.00	82.00	51.00	30.00	31.00	43.00	373.00	53.29
FIPD	16.00	15.00	11.00	18.00	42.00	49.00	46.00	197.00	28.14
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.74	1.32	1.65	2.12	1.52	1.43	1.22	11.00	1.57
PROB	2.42	2.05	2.00	2.23	2.44	2.27	2.08	15.49	2.21
FID	3.38	3.80	3.82	3.29	2.88	3.01	3.36	23.54	3.36
FIPD	2.46	2.83	2.53	2.36	3.16	3.29	3.34	19.97	2.85

ECONOMICS

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	32.81	9.68	16.47	25.53	32.38	39.73	50.68		
PROB	32.35	7.13	23.18	25.30	30.33	38.87	42.76		
FID	32.33	7.72	22.80	25.25	29.81	38.96	44.50		
FIPD	32.26	7.52	22.76	25.36	29.88	38.79	44.23		
ORG	32.59	7.08	25.00	25.75	30.07	38.85	44.55		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.35	1.58	4.39	2.26	1.95	2.61	3.94		
PROB	0.83	0.65	1.59	1.31	1.07	1.07	1.29		
FID	0.19	0.23	0.82	0.24	0.18	0.50	0.44		
FIPD	0.21	0.29	1.06	0.32	0.17	0.54	0.65		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	10.00	4.00	3.00	9.00	0.0	9.00	1.00	36.00	5.14
PROB	10.00	46.00	46.00	18.00	8.00	17.00	3.00	148.00	21.14
FID	48.00	13.00	23.00	27.00	39.00	35.00	55.00	240.00	34.29
FIPD	32.00	37.00	28.00	46.00	53.00	39.00	41.00	276.00	39.43
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.56	1.23	1.20	1.73	1.23	1.51	1.20	9.66	1.38
PROB	2.12	3.17	3.08	2.06	2.15	2.33	2.00	16.91	2.42
FID	3.31	2.48	2.80	2.97	3.24	3.09	3.50	21.39	3.06
FIPD	3.01	3.12	2.92	3.24	3.38	3.07	3.30	22.04	3.15

# FINANCE

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	33.63	9.80	20.07	27.22	32.41	38.44	55.00		
PROB	33.84	6.36	26.54	30.60	32.64	36.27	50.40		
FID	33.45	7.11	26.36	30.04	31.87	35.04	53.38		
FIPD	33.52	7.08	26.70	30.07	31.89	34.99	53.51		
ORG	33.60	7.14	26.55	30.40	32.00	34.86	54.00		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	2.18	1.62	4.39	3.06	2.55	2.93	5.99		
PROB	0.79	0.83	1.12	0.98	0.88	1.03	3.00		
FID	0.19	0.32	0.50	0.15	0.34	0.14	1.18		
FIPD	0.24	0.60	0.58	0.11	0.29	0.16	2.28		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	6.00	3.00	1.00	4.00	5.00	1.00	10.00	30.00	4.29
PROB	8.00	9.00	18.00	14.00	9.00	2.00	10.00	70.00	10.00
FID	45.00	65.00	41.00	36.00	37.00	50.00	63.00	337.00	48.14
FIPD	41.00	23.00	40.00	46.00	49.00	47.00	17.00	263.00	37.57
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.41	1.27	1.15	1.28	1.28	1.32	1.90	9.61	1.37
PROB	2.18	2.26	2.47	2.25	2.15	1.83	1.94	15.08	2.15
FID	3.26	3.57	3.19	3.11	3.21	3.42	3.52	23.28	3.33
FIPD	3.15	2.90	3.19	3.36	3.36	3.43	2.64	22.03	3.15

MANAGEMENT

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	34.85	9.54	18.52	28.22	34.25	40.76	55.03		
PROB	34.91	6.03	26.00	29.49	34.17	38.17	47.65		
FID	34.82	7.15	25.80	28.79	33.11	37.87	51.80		
FIPD	34.80	6.82	26.11	29.12	33.25	37.57	51.20		
ORG	34.71	6.62	26.00	29.30	33.18	37.27	51.45		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.36	1.64	3.85	2.30	1.93	2.27	5.26		
PROB	0.78	0.59	1.14	0.93	0.97	1.04	2.25		
FID	0.21	0.24	0.43	0.47	0.39	0.59	1.15		
FIPD	0.18	0.20	0.56	0.36	0.42	0.57	0.90		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	7.00	0.0	0.0	8.00	7.00	3.00	7.00	32.00	4.57
PROB	7.00	20.00	19.00	16.00	6.00	16.00	8.00	92.00	13.14
FID	43.00	8.00	41.00	29.00	40.00	31.00	38.00	230.00	32.86
FIPD	43.00	72.00	40.00	47.00	47.00	50.00	47.00	346.00	49.43
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.57	1.09	1.01	1.61	1.68	1.26	1.66	9.88	1.41
PROB	2.03	2.62	2.58	2.30	1.84	2.46	1.79	15.62	2.23
FID	3.22	2.62	3.16	2.87	3.22	2.97	3.18	21.24	3.03
FIPD	3.18	3.67	3.25	3.22	3.26	3.31	3.37	23.26	3.32

MARKETING

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	36.10	8.53	22.80	30.53	35.95	41.76	49.86		
PROB	36.49	4.81	28.62	33.70	35.47	40.02	43.97		
FID	36.37	6.21	27.86	32.74	34.13	41.03	46.75		
FIPD	36.26	5.94	28.21	32.77	34.11	40.68	46.07		
ORG	36.10	5.74	28.50	32.75	33.68	40.70	45.40		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	2.16	1.80	4.57	3.47	2.53	3.13	4.28		
PROB	0.82	0.52	1.11	1.00	0.99	1.08	1.36		
FID	0.20	0.27	0.52	0.44	0.29	0.63	0.83		
FIPD	0.16	0.22	0.52	0.42	0.28	0.39	0.70		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	3.00	1.00	2.00	3.00	11.00	9.00	3.00	32.00	4.57
PROB	17.00	7.00	16.00	6.00	7.00	13.00	13.00	79.00	11.29
FID	24.00	21.00	25.00	40.00	39.00	30.00	18.00	197.00	28.14
FIPD	56.00	71.00	57.00	51.00	43.00	48.00	66.00	392.00	56.00
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.37	1.18	1.17	1.30	1.73	1.55	1.45	9.75	1.39
PROB	2.28	2.18	2.60	2.10	1.75	2.25	2.38	15.54	2.22
FID	2.96	2.96	2.86	3.24	3.21	2.92	2.65	20.80	2.97
FIPD	3.39	3.68	3.37	3.36	3.31	3.28	3.52	23.91	3.42

# ASSISTANT PROFESSOR

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	27.78	7.44	12.01	22.81	27.73	32.90	42.71		
PROB	27.49	4.57	17.64	24.43	27.67	31.31	34.53		
FID	27.31	3.92	20.58	24.24	27.19	30.85	33.50		
FIPD	27.31	4.19	18.21	24.37	27.53	30.88	33.52		
ORG	27.88	3.64	21.00	25.00	27.50	31.20	33.50		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.32	0.91	3.79	1.70	1.86	1.78	3.47		
PROB	0.92	0.66	2.71	1.33	1.03	0.82	0.86		
FID	0.21	0.22	0.57	0.52	0.55	0.33	0.42		
FIPD	0.25	0.34	1.66	0.42	0.51	0.30	0.41		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	25.00	0.0	0.0	6.00	8.00	8.00	0.0	47.00	6.71
PROB	30.00	13.00	6.00	25.00	12.00	20.00	11.00	117.00	16.71
FID	26.00	64.00	83.00	28.00	30.00	32.00	40.00	303.00	43.29
FIPD	19.00	23.00	11.00	41.00	50.00	40.00	49.00	233.00	33.29
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	2.21	1.02	1.18	1.53	1.71	1.52	1.00	10.17	1.45
PROB	2.48	2.48	2.35	2.43	2.19	2.47	2.37	16.77	2.40
FID	2.67	3.55	3.80	2.91	2.88	2.91	3.23	21.95	3.14
FIPD	2.64	2.95	2.67	3.13	3.22	3.10	3.40	21.11	3.02

ASSOCIATE PROFESSOR

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	32.71	7.50	18.30	27.79	32.65	37.71	47.49		
PROB	33.15	3.71	26.64	30.07	33.80	35.45	40.01		
FID	32.47	3.73	26.30	29.79	32.77	34.13	41.01		
FIPD	32.63	3.52	26.82	29.90	33.01	34.21	40.70		
ORG	32.54	3.44	26.55	30.07	32.85	33.72	40.70		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.55	1.29	3.79	2.01	1.89	2.12	4.51		
PROB	0.74	0.45	1.12	0.96	0.90	0.92	1.07		
FID	0.18	0.15	0.46	0.17	0.42	0.35	0.60		
FIPD	0.16	0.18	0.59	0.17	0.39	0.29	0.48		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	5.00	0.0	0.0	0.0	9.00	4.00	0.0	18.00	2.57
PROB	6.00	18.00	15.00	11.00	8.00	3.00	15.00	76.00	10.86
FID	51.00	23.00	40.00	32.00	39.00	53.00	42.00	280.00	40.00
FIPD	38.00	59.00	45.00	57.00	44.00	40.00	43.00	326.00	46.57
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.51	1.01	1.00	1.24	1.71	1.27	1.06	8.80	1.26
PROB	1.84	2.59	2.56	2.20	1.87	1.94	2.44	15.44	2.21
FID	3.40	2.89	3.19	3.07	3.17	3.49	3.15	22.36	3.19
FIPD	3.25	3.51	3.25	3.49	3.25	3.30	3.35	23.40	3.34

FULL PROFESSOR

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	39.45	8.95	20.91	33.73	39.44	45.14	58.63		
PROB	39.46	5.02	29.74	36.83	39.00	42.54	50.51		
FID	39.93	6.27	29.22	35.76	38.93	43.95	53.29		
FIPD	39.70	6.14	29.39	35.65	38.66	43.59	53.65		
ORG	39.53	5.97	30.00	35.55	38.85	43.45	54.00		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	1.23	1.02	3.48	1.93	1.64	1.87	4.29		
PROB	0.93	0.56	1.15	1.02	1.11	1.24	2.84		
FID	0.18	0.20	0.43	0.47	0.47	0.60	1.15		
FIPD	0.20	0.28	0.32	0.42	0.63	0.67	2.20		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	11.00	0.0	0.0	4.00	20.00	9.00	13.00	57.00	8.14
PROB	17.00	6.00	31.00	8.00	17.00	16.00	9.00	104.00	14.86
FID	19.00	33.00	25.00	42.00	39.00	31.00	56.00	245.00	35.00
FIPD	53.00	61.00	44.00	46.00	24.00	44.00	22.00	294.00	42.00
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.88	1.03	1.00	1.46	2.00	1.73	1.85	10.95	1.56
PROB	2.23	2.19	2.87	2.02	2.22	2.14	1.90	15.57	2.22
FID	2.56	3.24	2.93	3.18	3.02	2.94	3.46	21.33	3.05
FIPD	3.33	3.54	3.20	3.34	2.76	3.19	2.79	22.15	3.16

## TOTAL

METHOD	MEAN	STD	MIN	25TH	MEDIAN	75TH	MAX	TOTAL	AVERAGE
AVERAGE VALUE OF 100 REPEATED VALUES									
POINT	33.62	9.65	11.96	27.00	33.14	39.81	59.50		
PROB	33.79	6.77	17.22	29.16	33.85	38.41	50.33		
FID	33.66	7.34	20.43	28.56	32.67	38.10	53.12		
FIPD	33.63	7.19	18.29	28.83	32.82	37.83	53.01		
ORG	33.74	6.82	21.00	29.00	32.80	37.47	54.00		
STANDARD DEVIATION OF 100 REPEATED VALUES									
POINT	0.73	0.64	3.75	1.10	0.94	1.31	3.80		
PROB	0.71	0.52	2.48	1.00	0.85	0.94	3.18		
FID	0.12	0.15	0.38	0.40	0.42	0.63	1.22		
FIPD	0.13	0.18	1.83	0.43	0.44	0.56	1.97		
NUMBER OF BEST CASES OUT OF 100 REPETITIONS									
POINT	7.00	0.0	0.0	4.00	18.00	5.00	3.00	37.00	5.29
PROB	5.00	53.00	8.00	22.00	12.00	14.00	10.00	124.00	17.71
FID	57.00	12.00	82.00	31.00	40.00	29.00	61.00	312.00	44.57
FIPD	31.00	35.00	10.00	43.00	30.00	52.00	26.00	227.00	32.43
AVERAGE RANK OF 100 REPETITIONS (MAX = 4)									
POINT	1.80	1.00	1.18	1.31	2.16	1.40	1.57	10.42	1.49
PROB	1.73	3.26	2.34	2.51	1.83	2.36	2.04	16.07	2.30
FID	3.42	2.50	3.81	2.98	3.01	2.90	3.48	22.10	3.16
FIPD	3.05	3.24	2.67	3.20	3.00	3.34	2.91	21.41	3.06

TABLE 5.4  
 DEGREE OF COMPROMISABILITY  
 (N: NUMBER OF REPETITIONS)

METHOD	N=10	N=30	N=70	N=100	N=500	N=1000
ACCT						
POINT	0.05175	0.03414	0.01481	0.01398	0.00638	0.00589
PROB	0.02968	0.02597	0.02304	0.02198	0.02256	0.02255
FID	0.01679	0.01689	0.01671	0.01638	0.01594	0.01608
FIPD	0.00855	0.01003	0.01172	0.01171	0.01242	0.01214
EAP						
POINT	0.13820	0.05247	0.02774	0.02813	0.00908	0.00695
PROB	0.11672	0.06467	0.05291	0.05318	0.05164	0.05211
FID	0.09984	0.02232	0.01223	0.01313	0.01664	0.01770
FIPD	0.08667	0.03921	0.03361	0.03306	0.03732	0.03773
ECON						
POINT	0.08282	0.04218	0.02238	0.01924	0.00716	0.00367
PROB	0.09315	0.03851	0.02779	0.02457	0.02444	0.02397
FID	0.08561	0.03110	0.02142	0.02090	0.02205	0.02221
FIPD	0.08278	0.03184	0.01939	0.01824	0.01696	0.01695
FIN						
POINT	0.12932	0.05388	0.02930	0.02333	0.00674	0.00764
PROB	0.11776	0.05743	0.03994	0.03559	0.02420	0.02303
FID	0.07678	0.02286	0.00880	0.00848	0.00960	0.00995
FIPD	0.08006	0.02386	0.00986	0.00807	0.00851	0.00874
MGT						
POINT	0.11125	0.05092	0.02590	0.02466	0.00603	0.00387
PROB	0.11849	0.05231	0.03373	0.03204	0.02308	0.02201
FID	0.10064	0.03370	0.01628	0.01388	0.01238	0.01221
FIPD	0.10130	0.03429	0.01630	0.01248	0.00749	0.00718

TABLE 5.4 (Continued)

METHOD	N=10	N=30	N=70	N=100	N=500	N=1000
MKT						
POINT	0.09352	0.03819	0.01702	0.01765	0.00596	0.00485
PROB	0.09965	0.05020	0.03832	0.03649	0.03141	0.03089
FID	0.08874	0.03142	0.01939	0.01636	0.01219	0.01167
FIPD	0.08779	0.03216	0.01621	0.01258	0.00787	0.00751
ASSISTANT PROFESSOR						
POINT	0.13156	0.05904	0.02991	0.02702	0.00905	0.00623
PROB	0.08850	0.03624	0.02878	0.02742	0.02904	0.02928
FID	0.08950	0.02645	0.01521	0.01485	0.01995	0.02130
FIPD	0.08406	0.03024	0.02337	0.02087	0.02386	0.02454
ASSOCIATE PROFESSOR						
POINT	0.06708	0.02530	0.01736	0.01515	0.00811	0.00432
PROB	0.10540	0.05176	0.03859	0.03480	0.02613	0.02368
FID	0.07856	0.02414	0.01196	0.01003	0.00918	0.00922
FIPD	0.07879	0.02603	0.01222	0.00945	0.00643	0.00608
FULL PROFESSOR						
POINT	0.08502	0.03808	0.01969	0.01656	0.00766	0.00483
PROB	0.08833	0.03966	0.03043	0.02831	0.02730	0.02692
FID	0.08477	0.03479	0.02128	0.01923	0.01470	0.01416
FIPD	0.07840	0.02748	0.01575	0.01342	0.00952	0.00922
TOTAL						
POINT	0.08266	0.03682	0.02117	0.01677	0.00838	0.00552
PROB	0.06980	0.03627	0.02896	0.02830	0.02647	0.02651
FID	0.05302	0.01898	0.01209	0.01185	0.01442	0.01505
FIPD	0.05325	0.02049	0.01357	0.01230	0.01259	0.01272

## CHAPTER VI

### SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

#### 6.1 Summary

Recent interest has focused on creating a statistical database which insures the privacy of the individual and yet provides maximum exposure of the original data for statistical analysis. The research effort in this area has produced three groups of data distortion methods: Data Suppression, Data Aggregation, and Data Transformation.

The Data Suppression group suggests that some of the attributes in the data set should be either partitioned or suppressed so that the user can't sort one individual out of the file (Chin-Ozsoyoglu (1978) and Yu-Chin (1977)). Limited exposure of the database is a disadvantage of this approach.

The Data Aggregation group advocates that the data should be released in aggregation (Feige-Watts (1970)). The major drawback of this approach is significant loss of information from the original data set.

The Data Transformation group proposes that the confidential data should be replaced by distorted data while keeping non-confidential

data in its original position. The data transformation could be done by distorting the original data on point or by frequency of the data distribution. Data transformation by point distortion is usually done by adding or multiplying a random number to the original data. The random number is usually chosen so that its mean is zero and standard deviation is the same as that of the original data (Beck (1980), Conway-Strip (1976), and Hansen (1971)). The data transformation of a frequency distribution can be performed either by data swapping or by the identification of an underlying density function.

Data Swapping replaces the value of the original data with another value from other record so that such swapping can maintain the same statistical properties of the original data (Dalenius-Reiss (1978)). The major drawback of Data Swapping is that no efficient way of data swapping is known. Reiss suggests Approximate Data Swapping which is still in the infant stage (Reiss (1979), Reiss (1980)).

Data transformation by a probability distribution has been introduced here. The basic assumption of this approach is that the original data is a sample from a population with a certain density function, and the distorted data is another sample from the same population. Since these two data sets are random samples from the same density function, they share asymptotically similar statistical properties. Unlike Data Swapping, the distorted data is easily obtainable by Probability Data Distortion. This technique requires three steps:

- Step 1: Identification of the underlying density function
- Step 2: Generation of distorted data from the probability distribution function
- Step 3: Mapping of the distorted data onto the original data

The basic idea behind Probability Data Distortion is that the distorted data can be generated if we know the density function with the true parameters of the original data. From this density function and the estimated parameters, we can generate distorted data which are comparable to the original data. Probability Data Distortion offers significantly greater protection of privacy and usefulness of the distorted data for statistical analysis than any other previously known data distortion mechanism.

Probability Data Distortion provides maximum exposure of the original data for statistical analysis while protecting confidential information on an individual from compromise. This has been proven both through the asymptotic properties of the distorted data and through the performance of small sampling experiments. In the Monte Carlo study, the low compromisability of the probability distorted data was demonstrated.

This probability distorted data can be used to answer queries or can be released as microdata. Any aggregation can be done from the probability distorted data, and only one set of distorted data is assumed to be released. If the original data is of a dynamic nature, there should be periodic updates of the parameters and the probability density function. The new distorted data should be generated from the updated density function.

## 6.2 Conclusions

Most of known value distortion techniques have some degree of difficulty in recreating all of the statistical properties of the original data. In implementing a distortion mechanism, we must guarantee

that the original data is distorted in such a manner as to preserve those statistical properties.

Four data distortion methods: namely, Point Data Distortion, Probability Data Distortion, Frequency Imposed Data Distortion, and Frequency Imposed Probability Data Distortion have been compared in terms of accuracy of statistical estimation, compromisability, and computational burden.

Our preliminary finding is that unless the underlying density function of the original data closely matches one of the density functions in Phillips' computer package (1972), Frequency Imposed Data Distortion and Frequency Imposed Probability Data Distortion perform better in re-creation of statistical parameters and in protection from compromise than Probability Data Distortion. Point Data Distortion performs very poorly in both parameter estimation and compromisability. The Probability Data Distortion consistently performs better than Point Data Distortion in all statistical estimations and in the compromisability test. Data generated through Point Data Distortion becomes easily compromisable as the number of replications is increased but data obtained by Probability Data Distortion resists compromise even when replications are permitted.

### 6.3 Suggestions for Further Research

This study of data distortion by probability distribution, Probability Data Distortion, is just a first step. It is acknowledged that Probability Data Distortion is more costly than the data distortion methods in current use when applied to dynamic databases.

Probability Data Distortion and Frequency Imposed Probability Data Distortion require identification of the underlying density function of the original data. One drawback of these methods is the limited number of density functions for which identification routines are currently available.

It is suggested that future study be directed toward developing an efficient algorithm which will find a general density function and the estimated parameters no matter what the probability distribution of the original data. In conjunction with this, a general random number generating routine is needed which (using the density function and estimated parameters found by that algorithm) will generate random numbers to be used as distorted data.

This discussion has been limited to the application of Probability Data Distortion to non-categorical data. This research should be continued to examine the possible application of this technique to categorical data. The concepts involved would be the same as those employed here but the techniques for implementation might be different. Of course, the transformed categorical data should preserve the distribution and the estimated parameters of the original data.

## REFERENCES

- Achugbue, J. O., and Chin, F. Y. "The Effectiveness of Output Modification by Rounding for Protection of Statistical Data Bases," Infor 17:3 (August 1979), pp. 209-218.
- Alagar, V. S., Blanchard, B., and Glaser, D. "Effective Inference Control Mechanisms for Securing Statistical Databases," AFIPS Conf. Proc. 1981 NCC, pp. 443-452.
- Beck, L. L. "A Security Mechanism for Statistical Databases," ACM Trans. Database Syst., 5:3 (September 1980), pp. 316-338.
- Boruch, R. F. "Maintaining Confidentiality in Educational Research: A Systematic Analysis," Am. Psycho. 26 (1971), pp. 413-430.
- Campbell, D. T., et al. "Confidentiality-preserving Modes Access to Files and to Interfile Exchange for Useful Statistical Analysis," Eval. Q. 1:2 (May 1977), pp. 269-299.
- Chin, F. Y. "Security in Statistical Databases for Queries with Small Count," ACM Trans. Database Syst., 3:1 (March 1978), pp. 92-104.
- Chin, F. Y., and Ozsoyoglu, G. "Security in Partitioned Dynamic Statistical Databases," in "PROC. COMPSAC '79" IEEE Service CTR., Piscataway, N.J., 1979, pp. 594-600.
- Chin, F. Y., and Ozsoyoglu, G. "Security in Statistical Databases for Sum and Count Queries," Information Privacy 1:4 (March 1979), pp. 148-153.
- Chin, F. Y., and Ozsoyoglu, G. "Statistical Database Design," ACM Trans. Database Syst. 6:1 (March 1981), pp. 113-139.
- Conway, R., and Strip, D. "Selective Partial Access to a Data Base," Proc. of ACM Annual Conf., Oct. 20-22, 1976, Houston, Texas, pp. 85-89.

- Dalénus, T. "Privacy Transformations for Statistical Information Systems," J. of Statistical Planning and Inference, 1 (1977), pp. 73-86.
- Dalénus, T. "Towards a Methodology for Statistical Disclosure Control," Statistisk Tidskrift 15 (1977), pp. 429-444.
- Dalénus, T., and Reiss, S. P. "Data-Swapping: A Technique for Disclosure Control," Computer Science Tech. Rep. 39, Brown Univ., Providence, R.I., July 1978.
- Davida, G. I., et al. "Database Security," IEEE Trans. Software Engr. SE-4:6 (November 1978), pp. 531-533.
- DeMillo, R. A., Dobkin, D., and Lipton, R. J. "Even Databases that Lie Can Be Compromised," IEEE Trans. Software Engr. SE-4:1 (January 1977), pp. 73-75.
- Denning, D. E., and Denning, P. J. "Data Security," Comp. Surveys 11:3 (September 1979), pp. 227-249.
- Denning, D. E. "Are Statistical Data Bases Secure?" AFIPS Conf. Proc. 1978 NCC, 47, 1978, pp. 525-530.
- Denning, D. E. "A Review of Research on Statistical Data Base Security," in Foundation of Secure Computation, R. A. DeMillo, et al (Eds.) New York: Academic Press, 1978, pp. 15-25.
- Denning, D. E. "Secure Statistical Databases with Random Sample Queries," ACM Trans. Database Syst. 5:3 (September 1980), pp. 291-315.
- Denning, D. E., Denning, P. J., and Schwartz, M. D. "The Tracker: A Threat to Statistical Database Security," ACM Trans. Database Syst. 4:1 (March 1979), pp. 76-96.
- Denning, D. E. "Complexity Results Relating to Statistical Confidentiality," Computer Science and Statistics: 12th Ann. Symp. Interface, Waterloo, Canada, May 1979, pp. 252-256.
- Denning, D. E., and Schlörner, J. "A Fast Procedure for Finding a Tracker in a Statistical Database," ACM Trans. Database Syst. 5:1 (March 1980), pp. 88-102.
- Dobkin, D. P., Jones, A. K., and Lipton, R. J. "Secure Databases: Protection Against User Inference," ACM Trans. Databases Syst. 4:1 (March 1979), pp. 97-106.
- Dobkin, D. P., Lipton, R. J., and Reiss, S. P. "Aspects of the Database Security Problem," Proc. of Conf. on Theoretical Computer Sci., Waterloo, Canada, 1977, pp. 262-274.

- Feige, E. L., and Watts, H. W. "Protection of Privacy through Micro-aggregation," in Databases, Computer, and the Social Sciences, R. L. Bisco (Eds.), New York: Wiley-Interscience, 1970.
- Fellegi, I. P. "On the Question of Statistical Confidentiality," J. Amer. Stat. Assoc., 67:337 (March 1972), pp. 7-18.
- Fellegi, I. P., and Phillips, J. L. "Statistical Confidentiality: Some Theory and Applications to Data Dissemination," Annals Econ. Soc. Measure, 3:2 (1974), pp. 399-409.
- Fernandez, E. B., Summers, R. C., and Wood, C. Database Security and Integrity, Addison-Wesley Pub. Co., 1979.
- Freeman, H. Introduction to Statistical Inference, Addison-Wesley Pub. Co., 1963.
- Hansen, M. H. "Insuring Confidentiality of Individual Records in Data Storage and Retrieval for Statistical Purposes," in Proc. 1971 AFIPS Fall Jt. Comptr. Conf., 39, Montvale, N.J.: AFIPS Press, pp. 579-585.
- Haq, M. I. "Security in a Statistical Data Base," Proc. Amer. Soc. Info. Sci., 11 (1974), pp. 33-39.
- Haq, M. I. "Insuring Individuals' Privacy from Statistical Data Base Users," AFIPS Conf. Proc. 43 (1975), pp. 941-946.
- Hoffman, L. J. Modern Methods for Computer Security and Privacy, Englewood Cliffs, N.J.: Prentice-Hall, 1977.
- Hoffman, L. J., and Miller, W. F. "Getting a Personal Dossier from a Statistical Data Bank," Datamation 16:5 (May 1970), pp. 74-75.
- IMSL. IMSL Library, Edition 8, Houston, IMSL Inc., 1980.
- Kam, J. B., and Ullman, J. D. "A Model of Statistical Databases and Their Security," ACM Trans. Database Syst. 2:1 (March 1977), pp. 1-10.
- Mood, A. M., Graybill, F. A., and Boes, D. C. Introduction to the Theory of Statistics, 3rd Edition, McGraw-Hill Book Co., 1974.
- Nargundkar, M. S., and Saveland, W. "Random Rounding to Prevent Statistical Disclosure," Am. Statist. Assoc. Proc., Social Statistics Section, (1972), pp. 382-385.
- Palme, J. "Software Security," Datamation 20:1 (January 1974), pp. 51-55.

- Phillips, Don T. "Applied Goodness of Fit Testing," O.R. Monograph Series, no. 1, AIIE-OR-72-1, American Institute of Industrial Engineers, Atlanta, Ga., 1972.
- Rao, C. R. Linear Statistical Inference and Its Applications, John Wiley & Sons, New York, 1965.
- Reed, I. S. "Information Theory and Privacy in Data Banks," Proc. AFIPS, 42 (1973), pp. 581-587.
- Reiss, S. P. "Security in Databases: A Combinatorial Study," J. ACM 26:1, pp. 45-57.
- Reiss, S. P. "The Practicality of Data Swapping," Computer Science Tech. Rep. 48, Brown Univ., Providence, R.I., July 1979.
- Reiss, S. P. "Practical Data-Swapping: The First Steps," Proc. 1980 Symp. on Security to Privacy, IEEE (April 1980), pp. 38-45.
- Reiss, S. P. "Medians and Database Security," in Foundations of Secure Computation, R. A. DeMillo, et al. (Eds.), New York: Academic Press, 1978, pp. 57-91.
- Schlörer, J. "Identification and Retrieval of Personal Records from a Statistical Data Bank," Methods Inf. Med. 14:i (Jan. 1975), pp. 7-13.
- Schlörer, J. "Disclosure from Statistical Databases: Quantitative Aspects of Trackers," ACM Trans. Database Syst. 5:4 (Dec. 1980), pp. 467-492.
- Schlörer, J. "Security of Statistical Databases: Multidimensional Transformation," ACM Trans. Database Syst. 6:1 (March 1981), pp. 95-112.
- Schwartz, M. D. "Inference from Statistical Databases," Ph.D. Th., Comptr. Sci. Dept., Purdue Univ., August 1977.
- Schwartz, M. D., Denning, D. E., and Denning, P. J. "Linear Queries in Statistical Databases," ACM Trans. Database Syst. 4:2 (June 1979), pp. 156-167.
- Turn, R., and Ware, W. H. "Privacy and Security Issues in Information Systems," IEEE Trans. Computer C-25:12 (Dec. 1976), pp. 1353-1361.
- Ullman, J. D. Principles of Database Systems, Potomac, Maryland: Computer Science Press, 1980.
- Van Leeuwen, J. "On Compromising Statistical Data Bases with a Few Known Elements," Infor. Processing Letters, 8:3, pp. 149-153.

- Wilks, S. S. Mathematical Statistics, John Wiley & Sons, Inc., New York - London, 1962.
- Wong, E. "A Statistical Approach to Incomplete Information in Database Systems," ACM Trans. Database Syst. 7:3 (Sep. 1982), pp. 470-488.
- Yu, C. T., and Chin, F. Y. "A Study on the Protection of Statistical Databases," Proc. ACM SIGMOD Int. Conf. Management of Data, 1977, pp. 169-181.
- Zehna, Peter W. Probability Distributions and Statistics, Allyn and Bacon, Inc., Boston, 1970.

## APPENDIX A

The "Goodness of Fit" tests:

(ref: Phillips, Don T., "Applied Goodness of Fit Testing,"  
O.R. Monograph Series, No. 1, AIIE, Atlanta, Ga., 1972)

### The Chi-Square Test

$$\chi^2 = \sum \frac{k (f_o - f_e)^2}{f_e}$$

where

$f_o$  = observed frequency for each class or interval

$f_e$  = expected frequency for each class or interval predicted  
by the theoretical distribution

$\sum$  = sum over all  $k$  classes or intervals

If  $\chi^2 = 0$ , then the observed and theoretical frequencies agree exactly, whereas if  $\chi^2 > 0$  they do not. The larger the value of  $\chi^2$ , the greater is the discrepancy between the observed and expected. If  $\chi^2 > 0$ , we must compare our calculated value against the tabulated values of  $\chi^2$ .

### The Kolmogorov-Smirnov Test

The test as developed by Kolmogorov and Smirnov consists of comparing the sample cumulative distribution function with the theoretical cumulative distribution function at each sample observation. The test statistic is the maximum deviation between the two functions at any point in the sample. From a sample of size  $n$  containing data points such that  $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$ , let  $F(x_i)$  = the theoretical cumulative distribution function specified in the null hypothesis and  $S(x_i)$  = the sample cumulative distribution function for any given observation  $x_i$ . The sample test statistic is:

$$D = \max_{\text{all } i} |F(x_i) - S(x_i)|$$

The resulting maximum sample test statistic is compared with a critical value, referenced by the size of the sample  $n$ , and a chosen level of significance.

### The Cramer-Von Mises Test

The Cramer-Von Mises test is similar to the Kolmogorov-Smirnov test in that it consists of a comparison of the cumulative theoretical distribution function with the cumulative frequency distribution function of the sample. To apply the test it is necessary first to arrange the sample data in increasing order. The data points are then treated separately without the necessity of grouping as in the Chi-square test. The test statistic is given by:

$$\omega^2 = \int_{-\infty}^{\infty} [F(x) - S(x)]^2 dF(x)$$

where:  $F(x)$  and  $S(x)$  are the cumulative sample distribution function

and cumulative theoretical distribution function respectively. Using an observed sample of size  $n$ , the integral can be approximated by

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^n \left[ \frac{2i-1}{2n} - F(x_i) \right]^2$$

or multiplying by  $n$

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^n \left[ \frac{2i-1}{2n} - F(x_i) \right]^2$$

The sample test statistic,  $n\omega^2$ , can now be compared to a critical value.

#### The Moments Test

$$v_1 = \frac{\delta_3^2}{\delta_2^3}$$

as a measure of skewness and

$$v_2 = \frac{\delta_4}{\delta_2^2}$$

as a measure of kurtosis

where:

$$\delta_2 = \int_x (x-\delta)^2 f(x) dx$$

$$\delta_3 = \int_x (x-\delta)^3 f(x) dx$$

$$\delta_4 = \int_x (x-\delta)^4 f(x) dx$$

$$\delta = \int_x xf(x) dx$$

For convenience, we define

$$\tilde{v}_1 = \sqrt{v_1} = \delta_3 / \delta_2^{3/2}$$

$$\tilde{v}_2 = (v_2 - 3) = \frac{\delta_4 - 3\delta_2^2}{\delta_2^2}$$

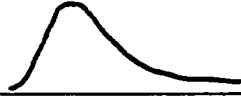
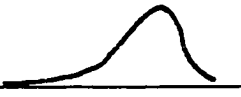




Hence, for a normal distribution

$$\tilde{v}_1 = 0$$

and

$$\tilde{v}_2 = 0.$$

These two quantities also reflect the nature of the skewness and kurtosis by their sign.

$\tilde{v}_1 > 0$		Skewed right
$\tilde{v}_1 < 0$		Skewed left
$\tilde{v}_2 > 0$		Platykurtic
$\tilde{v}_2 < 0$		Leptokurtic
$\tilde{v}_2 = 0$		Mesokurtic
$\tilde{v}_1 = 0$		Symmetrical

## APPENDIX B

### HELLY LEMMA AND HELLY-BRAY THEOREM

(ref: Rao, C. R. Linear Statistical Inference and Its Applications, John Wiley & Sons, New York, 1965.)

(i) HELLY LEMMA. Every sequence of distribution functions is weakly compact, that is, there is a subsequence which tends to a function at all continuity points of the latter.

Let  $D = \{r_k\}$  be the set of all rationals. Since  $F_n(r_1)$  is bounded, there exists a convergent subsequence. Consider the sequence  $\{F_{n_1}(x)\}$  which converges for the particular value  $x = r_1$ . From the sequence  $\{F_{n_1}(x)\}$  we can extract another subsequence  $\{F_{n_2}(x)\}$ , in a similar way, which converges at  $x = r_2$ , and of course such a sequence will converge at  $x = r_1$  also, and so on. Let us consider a sequence formed by the first member of  $\{F_{n_1}\}$ , the second member of  $\{F_{n_2}\}$ , . . . . Such a sequence of functions  $\{F_s\}$  necessarily converges for all  $x \in D$ , and the limiting function  $F_D$  defined for all  $x \in D$  is bounded and non-decreasing. Let, for any  $x$

$$F(x) = \sup_{r_i < x} F_D(r_i). \quad (B-1)$$

By definition  $F$  is continuous from the left, bounded, and nondecreasing.

We shall show that the subsequence determined actually converges to  $F(x)$  as defined in (B-1) at all continuity points of  $F$ . Let  $x$  be such a point. Then we can find a sequence of rational values  $(x_i^<, x_i^>)$  such that  $x_i^< < x < x_i^>$  and  $F(x_i^>) - F(x_i^<) \rightarrow 0$  as  $i \rightarrow \infty$ . For any pair  $(x_i^<, x_i^>)$  we have the obvious relationship

$$F_s(x_i^<) \leq F_s(x) \leq F_s(x_i^>) \quad (B-2)$$

for each  $s$ , where  $\{F_s(x)\}$  is the sequence tending toward  $F_D(x)$ . Taking limits of functions in (B-2), we have

$$F_D(x_i^<) \leq \underline{\lim} F_s(x) \leq \overline{\lim} F_s(x) \leq F_D(x_i^>)$$

for each  $i$ . Since the difference  $F_D(x_i^>) - F_D(x_i^<)$  can be made arbitrarily small,  $\lim F_s(x)$  exists and is equal to  $F(x)$  defined in (B-1). Observe that  $F(x)$  is also in the interval  $[F_D(x_i^<), F_D(x_i^>)]$ .

(ii) HELLY-BRAY THEOREM.  $F_n \rightarrow F \Rightarrow \int g dF_n \rightarrow \int g dF$  for every bounded continuous function  $g$ .

Choose two continuity points  $a, b (a < b)$  of  $F$  and write

$$\begin{aligned} \int_{-\infty}^{\infty} g dF_n - \int_{-\infty}^{\infty} g dF \\ = \int_{-\infty}^a g(dF_n - dF) + \int_a^b g(dF_n - dF) + \int_b^{\infty} g(dF_n - dF). \end{aligned} \quad (B-3)$$

Let  $|g| < c$ . Then the modulus of the first integral in (B-3) gives

$$|\int_{-\infty}^a g dF_n - \int_{-\infty}^a g dF| < c \int_{-\infty}^a dF_n + c \int_{-\infty}^a dF = c[F_n(a) + F(a)].$$

If  $a$  is sufficiently small,  $F(a)$  is small and so also is  $F_n(a)$  for all  $n > n_0$ . Hence  $c[F_n(a) + F(a)] < \varepsilon/5$  for a suitable choice of  $a$  and  $n_0$ . Similarly the third integral in (B-3) is  $< \varepsilon/5$  for a suitable choice of  $b$  and  $n_0$ .

In the finite interval  $(a, b)$ ,  $g$  is uniformly continuous. Let us divide  $(a, b)$  into  $m$  intervals

$$x_0 = a < x_1 < \dots < x_{m-1} < b = x_m$$

where  $x_1, \dots, x_{m-1}$  are continuity points of  $F$  and such that

$$[g(x) - g(x_i)] < \varepsilon/5$$

for  $(x_i < x < x_{i+1})$  uniformly for all  $i$ . Define the function

$$g_m(x) = g(x_i), \quad x_i \leq x < x_{i+1}.$$

Then

$$\begin{aligned} \int_a^b g_m(x) dF_n &= \sum g(x_i) [F_n(x_{i+1}) - F_n(x_i)] \\ &\rightarrow \sum g(x_i) [F(x_{i+1}) - F(x_i)] = \int_a^b g_m dF \end{aligned}$$

as  $n \rightarrow \infty$ , so that for any given  $m$

$$\int_a^b g_m (dF_n - dF) < \frac{\varepsilon}{5}$$

for sufficiently large  $n$ . But

$$\begin{aligned} \int_a^b g dF_n - \int_a^b g dF &= \int_a^b (g - g_m) dF_n + \int_a^b g_m (dF_n - dF) + \int_a^b (g - g_m) dF \\ &< \int_a^b \frac{\varepsilon}{5} dF_n + \frac{\varepsilon}{5} + \int_a^b \frac{\varepsilon}{5} dF < \frac{3}{5} \varepsilon \end{aligned}$$

for sufficiently large  $n$ . Hence the difference (B-3) is  $< \varepsilon$  which proves the desired result.

APPENDIX C  
CRITICAL VALUES OF THE KOLMOGOROV-SMIRNOV STATISTIC\*

One-Sided Test p = .90		.95	.975	.99	.995	p = .90 .95 .975 .99 .995						
Two-Sided Test p = .80		.90	.95	.98	.99	p = .80 .90 .95 .98 .99						
100	n = 1	.900	.950	.975	.990	.995	n = 21	.226	.259	.287	.321	.344
	2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
	3	.565	.636	.708	.785	.829	23	.216	.247	.275	.307	.330
	4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
	5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
	6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
	7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
	8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
	9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
	10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
	11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
	12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
	13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
	14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
	15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
	16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
	17	.250	.286	.318	.355	.381	37	.172	.196	.218	.244	.262
	18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
	19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
	20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252
Approximation for n > 40							$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	

\*Adapted from Miller, L. H. "Table of Percentage Points of Kolmogorov Statistics," Journal of the American Statistical Association, 51, 1956, pp. 111-121.

## APPENDIX D

### PROBABILITY SPACE

(ref: Zehna, Peter W. Probability Distributions and Statistics, Allyn and Bacon, Inc., Boston, 1970.)

In the study of this dissertation we have developed three inference control mechanisms (Probability Data Distortion, Frequency Imposed Data Distortion, and Frequency Imposed Probability Data Distortion) to generate a distorted data which would be mapped into the original data.

Definition 1.

Let  $S$  be a sample description space,  $Y$  a class of subsets of  $S$  and  $F$  a functional defined on  $Y$ . The triple  $(S, Y, F)$  is called a Probability Space.

A random variable  $X$  is a real-valued function whose domain is  $S$ , that is,  $X: S \rightarrow S_1$ .

Definition 2.

Let  $(S_1, Y_1, F_x)$  be the probability space induced by the random variable  $X$ . For each  $x \in S_1$ , let

$$F_n(x) = F_x((-\infty, x]) = F([X \leq x]).$$

Then  $F_n: S_1 \rightarrow S_1$  is the probability distribution function for  $X$ .

A Probability Data Distortion technique is a mathematical model whose theoretical structure is a probability space and in which the actual data of  $S$  are used to identify the theoretical probability distribution function  $F$  which is required to generate the probability distorted data. The basic idea of Probability Data Distortion is that a subset of  $S$  could be used as a distorted data set.

If  $F: O \rightarrow D$  is a function from  $O$  to  $D$ , we will say:  $F$  maps  $O$  to  $D$ , which each element of  $O$  there corresponds a unique image in  $D$ . The distorted data set  $D$  will be referred as the image set of original data set  $O$ .

Probability Data Distortion requires to have a unique image, so  $F$  is said to be one-to-one. We show a typical point  $a \in O$  being matched with its image  $d$  as shown in the figure D-1.

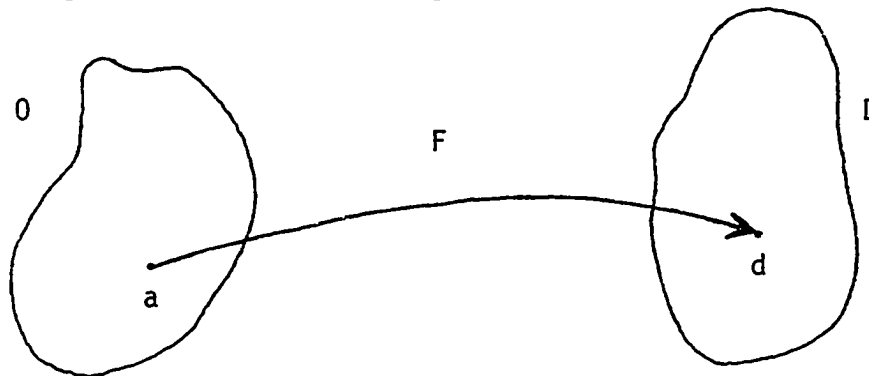


Figure D-1

Let  $F: O \rightarrow D$  be a function and let  $\hat{O}$  and  $\hat{D}$  be the respective power sets of  $O$  and  $D$ . Associated with  $F$  is a natural relation in  $\hat{O}$  and  $\hat{D}$ , which we will denote  $\tilde{F}$ . This relation is defined by the rule

$$\tilde{F}(C) = \{d; d = F(a), a \in C\} \text{ for each } C \in \hat{O}.$$

Definition 3.

Let  $P = (S, Y, F)$  be a probability space. Let  $X$  be a mapping of  $S$  into the real numbers such that

$$S_x = \{\omega | X(\omega) \leq x\} \in Y$$

for all real  $x$ . Then we say  $X$  is a random variable defined on  $P$ . Thus a random variable is a  $P$ -measurable function on  $P$ . We call

$$F(x) = P[S_x] = P[X \leq x]$$

the distribution function of  $X$ .

Definition 4.

Suppose  $\{D_n\}_{n=1}^{\infty}$  is a sequence of random variables all defined on the same sample space. Let  $F_n$  be the distribution function of  $D_n$ , and suppose  $F$  is a distribution function. Let  $X$  be a random variable with distribution function  $F$ . Then we say that  $D_n$  converges in distribution to  $X$  provided

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

at each  $x$  where  $F$  is continuous.

When we say  $X$  is a random variable with distribution function  $F_n$ , it is assumed that there is some underlying probability space  $(S, Y, F)$  and  $X: S \xrightarrow{\text{onto}} R_x$ , which are real numbers.

The random variable  $X$  induces another probability space  $(S_x, Y_x, F_x)$ . In this sense, the original probability space  $(S, Y, F)$  can be replaced by this new one, and the defining condition  $F_x(D) = F([X \in D])$  preserves the probability distribution function. Once we generate the distorted data set  $D$ , those are a member of  $R_x$ . If

$D: S \rightarrow S_1$  is a different random variable, this also induces a new probability space  $(S_d, Y_d, F_d)$  and the only difference in these two spaces lies in the functional forms of  $F_x$  and  $F_d$ .

When  $S$  is uncountable we have no means of characterizing a probability function on  $Y$  unless  $S = S_1$ . In general by defining a function from  $S$  to  $S_1$ , we can replace  $S$  by  $S_1$ . Even though it is easy to characterize the probability functions when  $S$  is countable for, real numbers possess an order properties.

When  $X$  is discrete, the original data can be ordered from smallest to largest. As shown previously in Frequency Imposed Data Distortion technique, the distribution function is considered as a step function with simple discontinuities at each point of  $R_x$ . For if  $x_i$  and  $x_j$  are two successive members of  $R_x$  with  $x_i < x < x_j$ . Then

$$F(x) = \sum_{t \leq x} F_n(t) = \sum_{t \leq x_i} F_n(t) + \sum_{x_i \leq t \leq x} F_n(t) = F_n(x_i) + \sum_{x_i < t \leq x} F_n(t)$$

But, for  $x_i < t < x_j$ , we know that  $F_n(t) = 0$  so that

$$F_n(x) = F_n(x_i) \text{ for all } x_i < x < x_j.$$

Hence  $F_n$  is constant over the interval  $(x_i, x_j)$ . Such a step function is another way of characterizing the probability distribution of a certain discrete random variable and is used as the distribution of such random variable in Frequency Imposed Data Distortion.