A Data Distortion by Probability Distribution

CHONG K. LIEW, UINAM J. CHOI, AND CHUNG J. LIEW
University of Oklahoma

This paper introduces data distortion by probability distribution, a probability distortion that involves three steps. The first step is to identify the underlying density function of the original series and to estimate the parameters of this density function. The second step is to generate a series of data from the estimated density function. And the final step is to map and replace the generated series for the original one. Because it is replaced by the distorted data set, probability distortion guards the privacy of an individual belonging to the original data set. At the same time, the probability distorted series provides asymptotically the same statistical properties as those of the original series, since both are under the same distribution. Unlike conventional point distortion, probability distortion is difficult to compromise by repeated queries, and provides a maximum exposure for statistical analysis.

Categories and Subject Descriptors: K.6.m [**Management of Computing and Information Systems**]: Miscellaneous—*security*; H.2.7 [**Database Management**]: Database Administration; H.2.m [**Database Management**]: Miscellaneous—*statistical databases*

General Terms: Management, Security

Additional Key Words and Phrases: Compromisability, point distortion, probability distortion, individual privacy, statistical database, microdata file

1. INTRODUCTION

The U.S. legal code[1] requires that sensitive information associated with a particular individual be protected from unauthorized release. At the same time, those data should be available to the public for a statistical analysis.

Inference control in statistical databases becomes an important issue, since many current types of research are heavily dependent upon statistical data that must preserve the confidentiality of individual information. Many researchers in this area have considered methods for controlling inference from statistical databases. Research effort has focused on finding a data set that provides statistical results similar to those of the original data set, while making it difficult for the curious user to identify information relating to a particular individual.

[1] For example, the U.S. Department of Health, Education and Welfare's regulation (45 CFR 46) requires the 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 subjects in research, development, or related activities."

Administrative support for this research was provided by the Division of Economics, The University of Oklahoma.
Authors' address: University of Oklahoma, Norman, OK 73019.

There has been a continuous search for a data distortion method that achieves these dual goals (i.e., provision of accurate statistical estimates and protection of individual privacy).

One method of inference control, against isolating a record by overlapping queries, is *partitioning the database* [2, 22]. 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. It is therefore impossible to isolate a record. Another method for such inference control is *microaggregation*: individuals are grouped and aggregated and statistics are computed for the aggregated value rather than individual data [11]. This technique increases uncertainty about the original information in the records as the size of the aggregated group of records grows.

Minimum overlap control is a third solution, suggesting that previously responded to queries be kept in file and that other queries that could lead to individual identification be refused [10]. The difficulty with this approach is tracking all the previous queries when many queries are requested. This control may also not be safe against queries that overlap by small amounts [6, 20].

Perturbing output is a technique that consists of rounding information up or down by some 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 [17].

The U.S. Census Bureau has used a technique that responds to queries that involve only a random subfile of the database, not the complete database. Even if some element of the subfile is identified, it may not be possible to learn which individual in the database was selected to be this element [13]. This technique is applicable to large databases only. Because a small random sample would be useless, other methods are needed to prevent compromise of small databases.

Recently, Denning [9] introduced a new inference control, called *random sample queries*, which deals directly with the basic principle of compromise by making it impossible for a user to precisely control the formation of query sets. Queries for relative frequencies and averages are computed using random samples drawn from query sets.

Conway and Strip [3] suggested *value distortion*, in which the value would be modified by some random quantity such that

$$Vd = Va + Vr,$$

where Vd is the distorted value, Va the actual value, and Vr a random deviate with a given distribution. The distribution is chosen to have an expected value of zero, so that Vd is an unbiased estimator of the true value Va. It is not always obvious what would constitute an appropriate distribution. If Va in the statistical database is symmetric, then the random deviate distribution should probably be symmetric. But if the Va is highly skewed, which is a common occurrence, then the choice of the distribution is much more difficult.

Hansen [13] has shown that the actual value is distorted by a random distortion rate, Cr, such that

$$Vd = Cr \cdot Va,$$

where Cr could be a randomly chosen interval.

If we compute a sum or an average on a large group, then the errors will tend to cancel each other out, 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 values of either input or output. This family of distortion is called *point distortion.*

Alternatively, the original series is considered as a random variable, associated with a probability distribution. If the underlying density function of the original series is determined, another set of data could be generated from the density function, and have asymptotically the same statistical properties as the original data set, since both originated from the same density function. This approach is appealing because the distorted data set not only preserves the basic statistical properties of the original series but is also sufficiently different from the original so as to protect the privacy individuals. This family of distortion is called *probability distortion.*

Data swapping is an early advocate of probability distortion, suggesting that the original data be replaced with a distorted version that maintains the same frequency count statistics as the original data. The difficulty with this approach is finding the general data swaps that preserve all frequency counts [5]. Since exact data swapping is not practically feasible, Reiss [21] suggested a feedback algorithm to find an approximate data swap 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 noncategorical data such as salary figures.

We present an alternative method of data distortion, which is workable for both categorical and noncategorical data sets. The basic hypothesis is that the original data is a sample from a population with a density function. Another sample from the same population can then be used as a distorted series to replace the original one. Since these two series share the same density function, the statistical properties of the original series such as frequency, subtotal, median, percentile, variance, and mean should be asymptotically the same as those of the distorted series. However, compromisability becomes difficult, since the distorted series is selected independently of the density function.

Once the original series is replaced by the distorted one, the replaced data set can be either placed on line to answer queries or released as microdata. Alternatively, it may be used as basic data from which statistical tabulations can be made. Suppose that the original series is of a dynamic variety whose value changes frequently over time, such as a salary file; the parameters of the density function should be updated and the corresponding new, distorted series should periodically replace the original data. When the original data are associated with other attributes such as sex, age, department, and place of birth, the distorted data should be mapped onto the original data to maintain consistency with other attributes.

2. DATA DISTORTION BY PROBABILITY DISTRIBUTION

A data set may be divided into confidential variables and nonconfidential variables. Consider a personal file that has current salary, salary raise this year, department, rank, sex, place of birth, and the highest degree earned. The current

salary and the salary raise may be considered confidential variables, and the remaining variables as nonconfidential.

Data distortion by probability distribution (probability distortion) necessitates three steps to compute the distorted series for confidential variables.

(a) Identification of the underlying density function of each of the confidential variables and estimation of the parameters associated with the density function.
(b) Generation of a distorted series for each confidential variable from the estimated density function.
(c) Mapping and replacement of the distorted series in place of the confidential series.[2]

We discuss each step.

2.1 Identification and Estimation

The original series, salary, for example, is screened to determine which of the predetermined density functions is best fitted for the data. The goodness of fit can be tested by the Kolmogorov–Simirnov maximum deviation test [15]. Currently available density functions for such identification are Poisson, exponential, normal, gamma, Weibull, log-normal, uniform, triangular, chi-square, and Erland-K distributions [18]. The Phillips computer package [18] is easy to use for the identification of the underlying density functions, and also provides goodness of fit test statistics—both Kolmogorov–Simirnov and Cramer–Von Mises statistics—and the parameter estimations of the density functions. In some cases, more than one density function can be acceptable at a given significance level. In such a case it is recommended that the density function showing the smallest Kolmogorov–Simirnov statistics be selected, for the obvious reason; namely, that this density has the most probable chance of accepting the null hypothesis. If none of the density functions fit the discrete data, we recommend using the frequency imposed distortion method.

The frequency imposed distortion method does not require the identification of any density functions. Instead, the original series is divided into several intervals (somewhere between 8 and 20). The frequencies within the interval are counted for the original series, and become a guideline to generate the distorted series. By using the uniform random number generating subroutine, the user generates a distorted series until its frequencies become the same as the frequencies of the original series. If the frequencies in some intervals overflow, they are simply discarded. The advantage of this method is that it can be used to distort most types of data. Small sampling properties of the frequency imposed distortion method are currently under investigation.

2.2 Data Generation

Once the best-fit density function is selected, its estimated parameters are supplied to its random value generating routine to produce the distorted series.

[2] Step (c) is needed when the distorted series is used to analyze statistics jointly with other attributes.

IMSL [14] has random number generating subroutines for each of all density functions identifiable in the Phillips package.

2.3 Mapping and Replacement

When distorted data are used for statistical analysis independently of other variables, ordered mapping is not necessary (i.e., sorting the distorted series and the original series in the same order and replacing each element of the original series with the corresponding distorted element). However, in many cases, the distorted series is used jointly with other variables for a statistical analysis. For example, a query response to an average salary by a specific age interval is the case in which both the distorted salary series and nonconfidential age data are used jointly for statistical analysis. Unless mapping is done, the average salary by age group becomes a meaningless value. In general, when the data set is a matrix in which the distorted data are used together with other variables for a statistical analysis, mapping is necessary to maintain consistency with other variables.

2.4 Asymptotic Properties of Probability Distortion

Probability distortion asymptotically preserves the statistical properties of the original series because the distorted values share asymptotically the same density function as the original series. Intuitively, the proof is as follows. Since the distorted series and the original series share the same density functions, both series will become the same series if the size of the sample becomes sufficiently large. As the sample size (n) approaches infinity (i.e., $n \rightarrow \infty$), both sample series converge to the population of the series. Therefore, the distorted series and the original series share asymptotically the same statistical properties [16, 19].

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 probability distortion.

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

Definition A. The sequence of random variables X_n is said to converge in distribution to a random variable X if $F_n \rightarrow F$ as $n \rightarrow \infty$.

This definition is provided by Rao [19, p. 96].

HELLY LEMMA. *Every sequence of a distribution function is weakly compact.*

The proof of this lemma is in [19, pp. 96–97].

HELLY–BRAY THEOREM. *$F_n \rightarrow F$ as $n \rightarrow \infty$ implies $\int g\,dF_n \rightarrow \int g\,dF$ for every bounded continuous function g.*

The proof of this theorem is given in [19, pp. 97–98].

LEMMA 1. *The moment-generating function* $(M_n(t))$ *of* X_n *asymptotically converges to the moment-generating function* $(M(t))$ *of* X *as* $n \to \infty$.

PROOF. By definition, $M_n(t) = \int \mathrm{EXP}(tX_n)\,dF_n$ and $M(t) = \int \mathrm{EXP}(tX)\,dF$. By definition A, above, $\mathrm{EXP}(tX_n) \to \mathrm{EXP}(tX)$ as $n \to \infty$. By setting $\mathrm{EXP}(tX) = g$, and by using the result of the Helly–Bray theorem, we complete the proof. □

COROLLARY 1. *The mean and standard deviation of* X_n *asymptotically converges to the mean and standard deviation of* X *as* $n \to \infty$.

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

LEMMA 2. *The cumulative density function* $(CF_n(a))$ *of* X_n *asymptotically converges to the cumulative density function* $(CF(a))$ *of* X *as* $n \to \infty$.

PROOF. By definition,

$$CF_n = \int_0^a F_n\,dF_n \quad \text{and} \quad CF(a) = \int_0^a F\,dF.$$

By statement A, above, $F_n \to F$ as $n \to \infty$. By setting $F = g$ and by using the results of the Helly–Bray theorem and the 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* X_n *asymptotically converge to those of* X.

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

The basic assumption of probability distortion is that the original series is a sample of size n drawn from a population with a density function. The distorted series becomes another sample of size n drawn from the same population. Since these two samples are independent, compromisability by repetition is not possible. However, as shown in the lemmas and corollaries, above, the mean, standard deviation, percentiles, and the median computed from both samples converge asymptotically to those of the population.[3]

3. A MONTE CARLO STUDY OF PROBABILITY DISTORTION

3.1 An Example of Data Distortion by Probability Distribution

We consider a hypothetical example of the faculty salaries of a business school that has four divisions (Finance, Economics, Management, and Accounting). The original series of the salaries by each division appear in Table I.

[3] Wilks [23, pp. 254–276] provides further properties of the asymptotic sampling theory.

Table I. Faculty Salary by Point Distortion (Unit: Thousand Dollars)

	Average Value of Each Observation when Number of Repetitions is N						
Group	Original	$N = 10$	$N = 30$	$N = 70$	$N = 100$	$N = 500$	$N = 1000$
Finance	19.600	24.453	20.063	20.261	20.605	19.509	19.625
	23.700	26.651	24.465	23.705	23.828	23.934	23.735
	27.300	29.652	28.054	28.628	28.169	27.812	27.680
	28.800	28.404	29.708	28.725	28.670	28.798	28.814
	29.900	30.644	29.486	29.210	29.450	29.888	29.756
	35.600	34.422	34.339	35.346	35.424	35.863	35.806
Economics	19.700	17.303	18.522	18.436	19.331	19.478	19.324
	25.600	26.948	26.457	26.005	25.775	25.715	25.751
	27.900	28.645	26.546	27.835	28.153	28.003	27.897
	29.200	27.518	28.143	27.973	28.535	29.133	29.139
	30.200	34.433	32.352	31.199	30.823	30.500	30.343
	33.300	32.365	32.863	32.225	31.892	32.949	32.878
	33.900	34.008	33.660	33.937	33.520	33.731	33.608
	45.300	46.320	46.350	46.416	45.803	45.157	45.245
Management	20.600	20.859	21.641	20.421	20.577	20.504	20.689
	26.900	25.229	24.904	26.490	25.876	26.438	26.583
	28.500	31.357	29.178	28.653	28.682	28.993	28.853
	29.200	30.502	28.983	28.180	27.985	28.772	28.839
	30.300	28.291	28.195	29.538	29.775	29.827	29.841
	32.600	34.010	31.068	31.767	31.298	32.019	32.292
	33.400	35.619	34.178	32.578	31.863	33.120	33.140
	34.800	36.299	35.076	35.165	35.352	34.707	34.903
	38.900	36.726	36.822	36.347	37.700	38.976	39.003
	41.300	43.709	42.780	41.134	41.601	41.209	41.371
	42.800	43.506	42.717	42.082	42.575	42.652	42.865
Accounting	22.800	24.437	22.876	23.113	22.894	22.891	22.824
	27.300	32.409	27.239	26.793	26.564	27.297	27.433
	28.700	30.781	29.710	28.587	29.000	28.547	28.856
	29.800	28.664	30.266	29.775	29.533	29.362	29.825
	32.600	33.625	32.569	32.622	32.594	32.533	32.492
	33.700	31.451	32.309	33.708	33.673	33.369	33.542
	35.600	38.657	35.852	36.297	35.486	35.675	35.522
	37.500	35.570	37.997	37.763	38.131	37.793	37.522
	42.800	41.128	40.751	41.809	42.406	43.108	43.114

A release of the original series will easily identify the salary of each faculty member. For example, a faculty member in Finance who receives $27,300 can easily estimate the salary of his or her colleagues by reckoning that the $19,600 figure must be the salary of a recently arrived assistant professor and that $35,600 is the salary of the division director. In this way the salary of every individual can be estimated. And the salary of a faculty member in any division can be calculated if the salary data is associated with other attributes such as age, rank, and the school where the final degree was earned.

To protect the privacy of the individual, we propose to distort the data by a probability distribution. This method requires three steps:

Step 1. Identification of the underlying density function. By using the original series, we compute the Kolmogorov–Smirnov (KS) statistics for each of the following density functions.

Density function	KS statistics (D)[4]	Remarks
Poisson	0.16601	
Exponential	0.43726	
Normal	0.11295	
Gamma	0.08597	
Weibull	0.14004	
Log-normal	0.07263	The best-fit
Uniform	0.20096	density
Triangular	0.17410	

The decision rule is as follows. If the computed D is smaller than the KS table value,[5] the null hypothesis (i.e., the sample drawn from the density function) is accepted. For example, the KS table value is 0.17909 when degrees of freedom are 34 and the significant level is 10 percent (one tail). We may conclude that at a 10 percent significance level the original series has the density function if its D value is smaller than 0.17909. In fact, the test includes all density functions, except that Exponential and Uniform are accepted as the original series' functions at a 10 percent significance level. However, we set the decision rule to choose the density function that yields the smallest D value, since such a choice will increase the maximum acceptance level. Under this decision rule, the log-normal 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. IMSL [14] has a subroutine to generate random numbers for the log-normal distribution. From this routine, an equal number of distorted series is generated to replace the original series. If this distorted series is used for a statistical analysis independently of other variables, it does not need to be mapped with the original series. However, if it is used with other variables in the same data set, mapping is necessary to maintain consistency with the other variables. When the estimated standard deviation is fairly large, the distorted set may have the smallest value and the largest value, which are quite different from the corresponding values of the original data. In such a case, we suggest a smoothing by averaging several replications of the distorted data instead of using the data generated first. Table I provides the smoothing series of the distorted data. An average of 30 replications provides a reasonably good distorted series.

3.2 The Small Sampling Efficiency of Probability Distortion

To investigate the small sampling efficiency of probability distortion, we generate a data set by point distortion by adding a random variable with zero mean and a constant variance to each of the original observations, that is

$$Z_i = X_i + \epsilon_i,$$

where Z_i = the ith observation of the distorted series by point distortion:

X_i = the ith observation of the original series;
ϵ_i = a random variable with zero mean and a standard deviation, the same as the original series (6.461 in this case).

[4] The KS statistic D is computed by $D = \text{Max}\,|F_a(X) - F_e(X)|$ where $F_a(X)$ is the observed cumulative relative frequency and $F_e(X)$ is the expected counterpart.
[5] See Lapin [15, pp. A-38 and A-39].

Table II. Number of Cases in which the Probabilty Distortion is Better than the Point Distortion (Total: 100 Cases)

Group	Mean	St. dev.	Min.	25th P.	Median	75th P.	Max.
Finance	74	84	70	83	74	79	88
Economics	71	78	73	73	72	84	71
Management	68	89	83	75	65	74	73
Accounting	58	89	80	78	64	69	82
Pooled	47	98	65	72	51	74	70

We give two criteria to compare the performance of probability distortion with that of point distortion: (1) accuracy in the estimation of statistics, and (2) the degree of compromisability if replications are permitted to the curious user.

3.2.1 *Accuracy of the Statistical Estimation.* As a measure of comparing the accuracy of the statistical estimation, we select the following commonly used statistics: mean, standard deviation, minimum value, 25th percentile, median, 75th percentile, and maximum value. These statistics are estimated by (1) the original series, (2) the point distorted series, and (3) the probability distorted series. The number of observations in both distorted series is the same as that of the original series (i.e., Finance 6, Economics 8, Management 11, Accounting 9, totaling 34 observations).

The seven statistics are computed once by the point distorted series and once by the probability distorted series, and results are compared with those computed by the original series.

First, we made 100 replications each of both distorted series. Out of the 100 replications, the number of cases in which the probability distorted series computes more accurately than the point distorted series is shown in Table II.

Except for the mean computation in the pooled groups, the probability distorted series computes all seven statistics in the five categories much better than the point distorted series. For example, in the pooled case, the probability distorted series computes the standard deviation 98 cases more accurately than the point distorted series does: out of 100 replications there are only two cases in which the point distorted series performs better than the probability distorted series. Similar results are obtained when the data are grouped and the group standard deviations are computed. The number of cases in which the probability distorted series performs better in the computation of the group standard deviation than the point distorted series does is 84 in Finance, 78 in Economics, 89 in Management, and 89 in Accounting. The probability distorted series computes the extreme value much better than the point distorted series. The number of cases in which the probability distorted series computes the minimum value more accurately than the point distorted series is 70 in Finance, 73 in Economics, 83 in Management, 80 in Accounting, and 65 in the pooled case. The results in the maximum value statistic are almost the same as those of the minimum case. Out of 100 replications, the number of cases in which the probability distorted series performs better than the point distorted series in the maximum calculation is 88 in Finance, 71 in Economics, 73 in Management, 82 in Accounting, and 70 in the pooled case. In the computation of percentiles (i.e., 25th, median, and 75th), the

probability distorted series is on average superior in a ratio of 3 to 1 over the point distorted series. For example, the probability distorted series computes the 25th percentile in the Finance group 83 cases out of 100 more accurately than the point distorted series. In the same category, the probability distorted series surpasses the point distorted series, 73 cases in Economics, 75 in Management, 78 in Accounting, and 72 in the pooled cases. In the 75th percentile computation, the number of cases in which the probability distorted series excel the point distorted series is 79 in Finance, 84 in Economics, 74 in Management, 69 in Accounting, and 74 in the pooled cases. A similar superiority of the probability distorted series over the point distorted series can be observed in the computation of the group mean or group median. When the seed for a random deviate generator value is changed, the basic result—the superiority of the probability distorted series over the point distorted series—remains the same, although there is a variation in number counting.

In the second experiment, 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 series. The AAME is computed by averaging the absolute deviation of the statistics computed by the distorted series and those computed by the original series, that is,

$$\text{AAME}_j = \frac{1}{G} \sum_{i=1}^{G} | 0_{ij} - \bar{D}_{ij} |.$$

The chi-square statistic (CSS) is computed by

$$\text{CSS}_j = \sum_{i=1}^{G} (0_{ij} - \bar{D}_{ij})^2 / 0_{ij}.$$

Where

0_{ij} = the original value of the jth statistic for the ith group.
$\bar{D}_{ij}$ = the average distorted value of the jth statistic in n replications for the ith group.
G = the number of the group (i.e., $G = 5$; Finance, Economics, Management, Accounting, and pooled).

The average absolute mean errors are calculated for each of the seven statistics by using both distorted series. When the number of replications is ten, the statistics computed by the probability distorted series have shown much smaller average absolute mean errors in all seven cases. For example, in the ten replications, the average absolute mean errors by the probability distorted series are 0.164 for the mean, 0.491 for 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 1.040 for the maximum value. In the same ten replications, the AAMEs by the point distorted series are 0.771 for the mean, 2.138 for standard deviation, and 2.016 for the minimum value, 1.354 for the 25th percentile, 1.062 for the median, 2.655 for the 75th percentile, and 1.199 for the maximum value. These AAMEs are much higher than those of the probability distorted series. It is interesting to observe that the probability distorted series performs better than the point distorted series even when the number of replications is relatively small (typically less than 70 replications). Out of the seven statistics computed, the

Table III. Grand Mean AAME for the Seven Statistics

	Number of Replications					
	10	30	70	100	500	1000
Probability distorted series	0.821	0.678	0.610	0.618	0.606	0.599
Point distorted series	1.599	1.386	1.319	1.267	1.232	1.231

number of cases in which the probability distorted series computes smaller average absolute mean errors than the point distorted series is 7 in 10 replications, 6 in 30 replications, and 6 in 70 replications. The minimum value is the only case in which the probability distorted series shows a larger average absolute mean error than the point distorted series in 30 and 70 replications. The average absolute mean error in the computation of the minimum value is 1.531 by the probability distorted series and 1.047 by the point distorted series for 30 replications. When the replications are increased to 70, the AAME becomes 1.200 by the probability distorted series and 1.014 by the point distorted series. The difference gets smaller as the number of replications increases. (See Tables IIIa and IIIb.)

When the number of replications increases sufficiently, to 500 or 1000, the result is mixed. Some statistics are computed better by the probability distorted series and others by the point distorted series. To compare the overall performance, we compute the grand mean of the average absolute mean errors of the seven statistics (i.e., the average AAME of the mean, standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum) for each level of replication. The results are shown in Table III.

The average AAME for the seven statistics by the probability distorted series is 0.821 for 10 replications, with a steady decrease to 0.599 for 1000 replications, whereas the averages for the point distorted series are 1.599 and 1.231, respectively. The results show that the probability distorted series performs better than the point distorted series in any number of replications. As the number of replications increases, the average AAME for the seven statistics decreases in both distorted series with a very slow converging rate.

We also compute the chi-square statistic to test the goodness of fit. Since the chi-square table value at a 0.05 significance level is 9.488,[6] we accept the hypothesis that the statistics computed by both distorted series are the same as those computed by the original series at 95 percent reliability.

3.2.2 *Compromisability of Distorted Data.* The primary purpose of data distortion is to guard individual privacy in the original data. If one set of distorted data replaces the original data, and no other sets of distorted data are released, there will be no problem of compromisability. However, if replications are permitted, some distorted series are easily compromisable whereas other distorted series are not. We say that distorted data are easily compromisable if any manipulation of the distorted series easily reveals the original data.

[6] Four degrees of freedom.

Table III(a). Parameter Estimation by Original and Distorted Data. Probability Distortion (N is Number of Repetitions)

Group	Parameters	Original	$N = 10$	$N = 30$	$N = 70$	$N = 100$	$N = 500$	$N = 1000$
Finance	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	27.867	27.804	27.830	27.838
	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
Economics	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
Management	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
Accounting	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	32.376
	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	42.800	41.828	42.049	42.410	42.222	42.276	42.371
Average ABS mean error	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	0.518
	75th P.		0.974	0.757	0.870	0.857	0.848	0.855
	Maximum		1.040	0.969	0.753	0.809	0.814	0.753
Chi-squares	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		0.170	0.162	0.115	0.128	0.129	0.118

Table III(b). Parameter Estimation by Original and Distorted Data. Point Distortion (N is Number of Repetitions)

Group	Parameters	Original	$N = 10$	$N = 30$	$N = 70$	$N = 100$	$N = 500$	$N = 1000$
Finance	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	24.453	20.063	20.261	20.605	19.510	19.626
	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	34.422	34.339	35.346	35.424	35.864	35.808
Economics	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	17.303	18.522	18.436	19.331	19.479	19.326
	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	46.320	46.350	46.417	45.803	45.159	45.247
Management	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	20.859	21.641	20.421	20.577	20.505	20.690
	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	43.506	42.717	42.082	42.575	42.654	42.867
Accounting	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	24.437	22.876	23.113	22.894	22.892	22.826
	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	41.128	40.751	41.809	42.406	43.109	43.116
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	18.664	17.123	16.946	17.231	16.716	16.642
	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	42.800	44.219	43.834	44.962	45.575	45.813	45.825
Average ABS mean error	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		2.016	1.047	1.014	0.772	0.676	0.695
	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		1.199	1.095	1.048	0.814	0.775	0.734
Chi-squares	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		1.659	0.447	0.469	0.345	0.428	0.454
	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		0.186	0.192	0.174	0.191	0.217	0.218

Table IV. Faculty Salary by Probability Distortion (Unit: Thousand Dollars)

	Average Value of Each Observation when Number of Repetitions is N						
Group	Original	$N = 10$	$N = 30$	$N = 70$	$N = 100$	$N = 500$	$N = 1000$
Finance	19.600	15.675	16.398	17.244	17.434	17.476	17.407
	23.700	24.398	23.939	23.844	23.793	23.750	23.753
	27.300	26.901	26.751	26.800	26.751	26.683	26.717
	28.800	29.299	29.004	28.934	28.858	28.974	28.954
	29.900	31.256	30.885	31.050	31.014	30.982	30.925
	35.600	35.955	36.189	36.215	36.223	36.260	36.296
Economics	19.700	19.271	19.541	19.885	20.139	20.115	20.091
	25.600	24.963	24.808	24.667	24.736	24.619	24.626
	27.900	27.533	27.240	27.334	27.292	27.309	27.330
	29.200	29.882	29.476	29.470	29.470	29.508	29.497
	30.200	31.600	31.284	31.492	31.480	31.448	31.395
	33.300	33.773	33.335	33.477	33.462	33.403	33.414
	33.900	35.210	34.969	35.075	35.055	35.058	35.065
	45.300	44.572	44.822	44.933	45.112	45.026	45.105
Management	20.600	22.042	21.671	21.682	21.799	21.696	21.688
	26.900	25.645	25.378	25.367	25.418	25.344	25.378
	28.500	27.953	27.765	27.831	27.767	27.860	27.881
	29.200	30.357	29.892	29.968	30.007	30.021	29.965
	30.300	32.053	31.728	31.926	31.934	31.916	31.867
	32.600	33.293	32.864	32.974	32.966	32.913	32.895
	33.400	34.314	33.894	34.029	34.005	33.938	33.975
	34.800	35.649	35.551	35.646	35.638	35.654	35.662
	38.900	38.966	38.534	38.605	38.560	38.622	38.662
	41.300	39.793	39.338	39.500	39.449	39.533	39.592
	42.800	40.626	40.523	40.796	40.723	40.710	40.780
Accounting	22.800	23.105	22.820	22.823	22.860	22.836	22.822
	27.300	26.182	26.063	26.117	26.105	26.019	26.077
	28.700	28.427	28.247	28.306	28.254	28.409	28.424
	29.800	30.785	30.488	30.596	30.573	30.492	30.458
	32.600	32.641	32.315	32.411	32.364	32.409	32.376
	33.700	34.660	34.423	34.538	34.505	34.478	34.516
	35.600	36.919	36.906	36.999	37.024	36.980	37.026
	37.500	37.658	37.561	37.669	37.673	37.723	37.782
	42.800	41.828	42.049	42.410	42.221	42.274	42.369

We compare the compromisability of probability distortion and point distortion. One popular way of identifying the original data is to average the replications of the distorted series if such replications are permitted. Tables I and IV show the average value of each datum when the number of replications is raised from 10, 30, 70, 100, 500, and 1000.

When data are distorted by point distortion, as few as 100 replications could easily identify the original data. However, when data are distorted by probability distortion, an increase in the number of replications will not improve compromisability.

We measure the degree of compromisability in terms of the average absolute percentage deviate from the original series and the distorted series; that is,

$$C(N) = \frac{1}{N} \Sigma \left(| X_i - Z_i | / X_i \right),$$

Table V. Degree of Compromisability (N is Number of Repetitions)

Probability 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 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	0.022	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

where $C(N)$ = the compromisability index when the number of replications is N:

X_i = the ith original observation,
Z_i = the ith distorted observation.

The point distorted series is easily compromisable when the replications reach more than 70. For example, the compromisability index of the point distorted series 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 70 replications are made. The compromisability index rapidly decreases as the number of replications reaches 1000. The index for the point distorted series when replication reaches 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 V).

In contrast to the point distorted series, the probability distorted series does not improve the compromisability by increasing the number of replications. For example, when 70 replications are made, the compromisability index of the probability distorted series 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 reach 1000, there is virtually no decrease in the compromisability index. The index for 1000 replications by the probability distorted series 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 case in which there are 10 replications, the compromisability indices of the probability distorted series are generally higher than those of the point distorted series, implying that the probability distorted series is much more difficult to compromise than the point distorted series. Tables I and IV provide average values for both of the distorted series and original values under varying replications. The reader can quickly find out that an increase in replication can be easily compromised by the point distorted series, but the same technique does not work in the case of the probability distorted series.

4. SUMMARY

We present a data distortion by probability distribution that requires three steps: (1) identification of the underlying density function, (2) generation of a distorted series from the density function, and (3) mapping of the distorted series onto the original series. (Step 3 is needed when the data are used jointly with other variables for statistical analysis.)

This data distortion provides a maximum exposure of the original data for statistical analysis while maintaining the confidentiality of an individual's record in the original data. The maximum exposure of the original data for statistical analysis was proved once by the asymptotic properties of the distorted series and another time by the performance of the small sampling experiments. In a Monte-Carlo study, we demonstrated the difficulty in compromising the probability distorted data even when replications are permitted.

The probability distorted series can be used to answer queries or be released as microstatistics. Any aggregation can be done from the probability distorted series and be released. In such a case, only one set of distorted series is assumed to be released. If the original data are of a dynamic variety, there should be a periodic update of the parameters of the density function and a new generation of the distorted series data from the up-dated density function.

REFERENCES

1. BECK, L. L. A security mechanism for a statistical database. *Database Syst. 5*, 3 (Sept. 1980), 316–338.
2. CHIN, F. Y., AND OZSOYOGULU, G. Securing in partitioned dynamic statistical databases. In *Proceedings COMPSAC '79* (Piscataway, N. J., 1979), IEEE, New York, 594–600.
3. CONWAY, R., AND STRIP, D. Selective partial access to a database. In *Proceedings of the ACM Annual Conference* (Houston, Tex., Oct. 20–22, 1976), ACM, New York, 85–89.
4. DALENIUS, T. Privacy transformations for statistical information systems. *J. Stat. Planning Inference 1.*
5. DALENIUS, T., AND REISS, S. P. Data-swapping: A technique for disclosure control. Computer Science Tech. Rep. 39, Brown Univ., July 1978.
6. DAVIDA, G. I., ET AL. Database security. *IEEE Trans. Softw. Eng. SE-4*, 6 (Nov. 1978), 531–533.
7. DENNING, D. E., AND DENNING, P. J. Data security. *ACM Comput. Surv. 11*, 3 (Sept. 1979), 227–249.
8. DENNING, D. E. A review of research on statistical database security. In *Foundations of Secure Computation*, R. A. DeMillo, et al., Eds., Academic Press, New York, 1978, 15–25.
9. DENNING, D. E. Secure statistical databases with random sample queries. *ACM Trans. Database Syst. 5*, 3 (Sept. 1980), 291–315.
10. DOBKING, D., JONES, A. K., AND LIPTON, R. J. Secure databases: protection against user inference. *ACM Trans. Database Syst. 4*, 1 (Mar. 1979), 97–106.
11. FEIGE, E. L., AND WATTS, H. W. Protection of privacy through microaggregation. In *Databases, Computers, and the Social Sciences*, R. L. Bisco, Ed., Wiley-Interscience, New York, 1970.
12. FELLEGI, I. P., AND PHILLIPS, J. L. Statistical confidentiality: Some theory and applications to data dissemination. *Ann. Econ. Soc. Measure 3*, 2 (1974), 399–409.
13. HANSEN, M. H. Insuring confidentiality of individual records in data storage and retrieval for statistical purposes. In *Proceedings 1971 AFIPS Fall Joint Computer Conference*, Vol. 39, AFIPS Press, Reston, Va., 579–585.
14. IMSL. *IMSL Library, Edition 8.* IMSL, Inc., Houston Tex., 1980.
15. LAPIN, L. *Statistics, Meaning and Method.* 2nd ed., Harcourt Brace, Jovanovich, New York, 1980.

16. LIEW, C. K. Inequality constrained least-squares estimation. *J. Am. Stat. Assoc. 71*, 365 (1976).
17. NARGUNDKAR, M. S., AND SAVELAND, W. Random rounding to prevent statistical disclosure. In *Proceedings of the American Statistical Association: Social Statistics Section* (1972), 382–385.
18. PHILLIPS, D. T. Applied goodness of fit testing. In *O. R. Monograph Series 1*, AIIE-0R-72-1, American Institute of Industrial Engineers, Atlanta, Ga., 1972.
19. RAO, C. R. *Linear Statistical Inference and Its Applications.* John Wiley, New York, 1965.
20. REISS, S. P. Medians and database security. In *Foundations of Secure Computation*, R. A. DeMillo, et al., Eds., Academic Press, New York, 1978, 57–91.
21. REISS, S. P. Practical data-swapping: The first steps. In *Proceedings 1980 Sympsoium on Security and Privacy* (Apr. 1980), IEEE, New York, 38–45.
22. YU, C. T., AND CHIN, F. Y. A study of the protection of statistical databases. In *Proceedings ACM SIGMOD International Conference on Management of Data* (1977), ACM, New York, 169–181.
23. WILKS, S. S. *Mathematical Statistics.* John Wiley, New York, 1962.

Received August 1981; revised June 1982; accepted March 1985