Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

A General Methodology for Masking Output from Remote Analysis Systems

Prepared by Krish Muralidhar*, Christine M. O'Keefe**, and Rathindra Sarathy***

* Gatton College of Business & Economics, University of Kentucky, Lexington, KY 40506, USA
Email: krishm@uky.edu
** CSIRO Computational Informatics, GPO Box 664, Canberra ACT 2602, AUSTRALIA
Email: Christine.OKeefe@csiro.au
*** Spears School of Business, Oklahoma State University, Stillwater, OK 74078, USA
Email: rathin.sarathy@okstate.edu

A General Methodology for Masking Output from Remote Analysis Systems

Krish Muralidhar[*], Christine M. O'Keefe[**], and Rathindra Sarathy[***]

[*] Gatton College of Business & Economics, University of Kentucky, Lexington, KY 40506, USA
Email: krishm@uky.edu
[**] CSIRO Computational Informatics, GPO Box 664, Canberra ACT 2602, AUSTRALIA
Email: Christine.OKeefe@csiro.au
[***] Spears School of Business, Oklahoma State University, Stillwater, OK 74078, USA
Email: rathin.sarathy@okstate.edu

Abstract: Remote analysis systems are being considered as an effective approach for providing individuals the ability to perform analysis on data held by statistical agencies and to receive the results of such analysis. One of the major obstacles to the wider implementation of these systems is the lack of a masking mechanism that will ensure that the masked response will be both useful and prevent disclosure. It is also desirable that this mechanism is easy to implement and automate so as to minimize human intervention. In this study, we describe the bootstrap mechanism that satisfies these requirements.

1 Introduction

National Statistical Agencies and other data custodian agencies hold a wealth of data regarding individuals and organizations, collected from censuses, surveys and administrative sources. In many cases, these data are made available to external researchers, for the investigation of questions of social and economic importance. In all cases, researcher access must comply with privacy legislation and regulation, as well as confidentiality assurances provided to data subjects by the custodian agency. The challenge of making useful data available for research while protecting confidentiality is often characterized as a trade-off between disclosure risk and data utility.[1]

Data custodians have traditionally offered a number of data access modes, including releasing confidentialised or masked data files and establishing secure on-site data centers. However, growing data access demands from increasingly sophisticated researchers and improving global communications infrastructure are prompting data custodians to look for new ways of meeting researcher demand, including establishing secure virtual data centres where trained staff check all output to ensure acceptably low disclosure risk.

In this paper, we will be concerned with remote analysis systems (RAS), which accept a query from an analyst, run it on data held in a secure environment, and then

[1] Due to space restrictions, we have presented a very condensed version of the paper. The full version has been submitted to an academic journal. Interested readers are requested to contact one of the authors to obtain a copy of the full version of the paper.

return the results to the analyst. RAS are normally understood to implement automated measures to ensure acceptably low disclosure risk including preventing the researcher from accessing the underlying data. RAS may be useful in allowing broader access than virtual data centres which reveal the underlying data, and certainly address the scalability issues of manual output checking. It is generally believed that the future of data access will involve RAS as one of the range of available data access modes. We can describe the desirable characteristics of a masking mechanism for RAS to possess the following characteristics:

(1) Easy to implement and automate,
(2) Provide appropriate responses,
(3) Provide adequate security, and
(4) Must be robust.

Despite growing interest from confidentiality researchers and practitioners, an as-yet-unsolved challenge is still to design automated output confidentialization measures for RAS that achieve a provably good balance between disclosure risk and utility. It is the purpose of this paper to address this unsolved challenge for RAS that offer traditional statistical analyses (and specifically exclude tabular data release). The bootstrap mechanism we propose is very simple, yet the characteristics of the bootstrap assure us that it is also highly effective.

2 Brief summary of RAS confidentialization methods

A Stage-based conceptual model for RAS is shown in Figure 1, introduced in (O'Keefe & Chipperfield, in press).

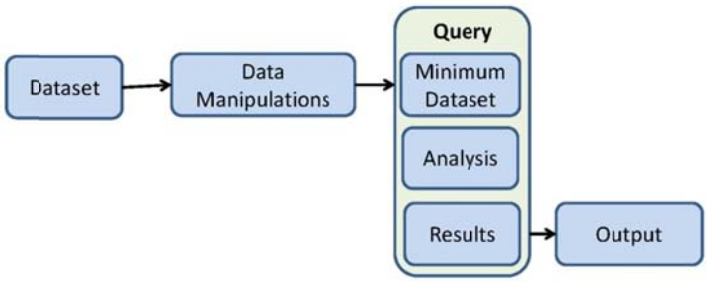

Figure 1 Stages of operation of a RAS

In order to generate masked output, a system administrator can introduce different types of confidentiality measures at different stages in the process. Normally, there is some masking of the underlying data, some restriction on the analyses as well as some output masking. This type of RAS is currently under development in Australia (Chipperfield & O'Keefe, submitted) and the United States (Lucero et al 2011).

A comprehensive review of confidentialization measures for RAS is provided in O'Keefe & Chipperfield (in press). Based on this review, in the remainder of the paper we assume that the data administrator of the RAS has implemented many of the data selection, manipulation, and analysis protection measures (including methods to prevent massively repeated queries) outlined in O'Keefe and Chipperfield (in press).

In addition to these confidentiality protection measures, it is necessary to modify or mask statistical output from RAS has to be modified to prevent disclosure by modifying the output resulting from a query. Output modification by using noise offers a simple alternative for this purpose.[2]

It would, in principle, be possible to use disclosure control methods designed for confidentializing microdata such as controlled suppression and noise-based perturbation (additive, multiplicative, and magnitude perturbation as well as controlled rounding) for confidentializing remote analysis outputs. In *additive perturbation*, the statistic is modified by the addition of random noise with mean zero and specified variance. In *multiplicative perturbation*, the statistic is multiplied by a random value selected from a distribution with mean one and specified variance. We remark that rounding is equivalent to an additive perturbation of the output by an amount which is up to (plus or minus) the rounded amount. In fact, additive, multiplicative and magnitude perturbation mechanisms are closely related and in most cases, each of them can be expressed as a noise-based perturbation (Muralidhar et al 1995, Kim and Winkler 1995).

One important advantage of noise-based perturbation (rounding, additive, or multiplicative) is its simplicity and generalizability. The approach could be used for any numerical response without the need for a great deal of prior analysis of the characteristics of the original data or the specific statistical technique underlying the query. Under this approach, the true response is modified using noise based on a pre-specified policy, then released. However, the result of such modification may not provide either the desired level of confidentiality or desired level of utility (or both), as we illustrate below.

3 Issues with perturbation approaches

In this section we illustrate the limitations of noise-based perturbation for confidentializing statistical analysis outputs, in terms of loss of data utility. Since the noise distribution is chosen without deep analysis of the underlying data structure, using noise-based perturbation of the statistical analysis result means that the modified response may not be consistent with the characteristics of the true response.

[2] We do not consider differential privacy based Laplace noise addition for several reasons. Please refer to the complete paper for the details of this discussion.

In this illustration, we consider a query which relates to the variance of a heavily skewed data set (Gamma distribution with scale parameter = 0.05, shape parameter = 10, and population variance = 5.00). The query data set (the records that satisfy the qualifying attributes of the query) consists of 100 records with variance = 3.80. Assume that the data administrator is considering two options: (1) Multiplicative perturbation using a $Uniform(0.9,1.1)$ distribution, and (2) Additive perturbation using $Normal(0, 0.25)$. The choice of the particular noise is for illustration purposes only. Figure 2 shows the true value, the range of the multiplicative and normal noise.

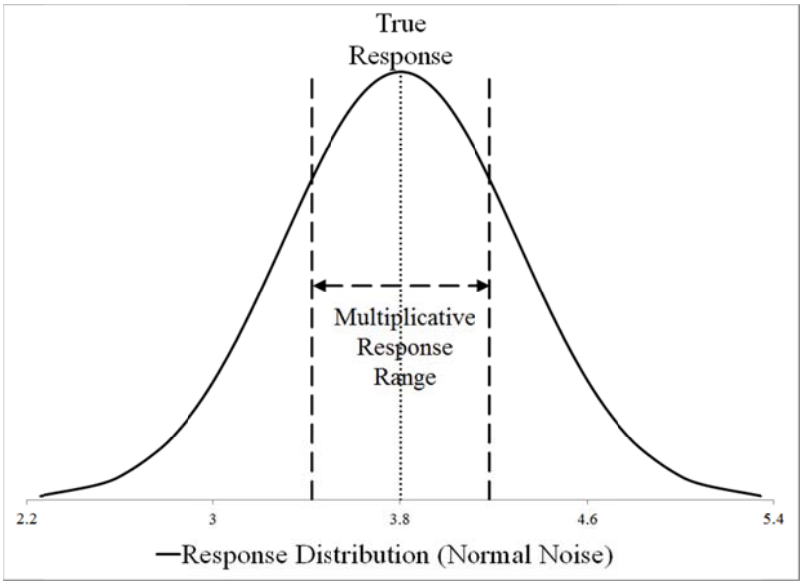

Figure 2. The distribution of the perturbed responses for sample variance

Multiplicative perturbation is purely a function of the true response, while additive noise is based only on the noise variance. The problem with the above approach is that the data administrator has no means to verify whether the perturbation is too low, just right, or too high. If the perturbation is too low, it presents a disclosure risk; if the perturbation is too high, it results in a useless response. But how can the data administrator verify that the perturbation level is appropriate?

One possibility is for the data administrator to consider the sampling distribution of the sample variance (that is, the distribution of the sample variance when all possible samples of size 100 are drawn from this population). The standard deviation of this distribution (the standard error of the sample variance) would present a measure of the dispersion of this statistic. The data administrator can then decide to use noise that is proportional to the standard error. This approach would be similar to the approach used in microdata perturbation where noise variance proportional to the variance of the underlying data set is often used (Hundepool et al 2012).

It is well known that, when the sample is drawn from a *normal population*, the sampling distribution of the sample variance follows a Chi-square distribution (Mood

et al. 1974). When responding to queries regarding the sample variance, the data administrator could use the Chi-square distribution to select or develop an appropriate masking mechanism.

Unfortunately, unlike the robust normality assumption based on the central limit theorem for the sampling distribution of the sample mean, the sampling distribution of the sample variance is very sensitive to the population distribution. When the sample is drawn from a population other than the normal, the Chi-square assumption can result in incorrect inferences. For the purposes of this example, let us assume that the data is actually from a heavily skewed distribution that occurs frequently in business data (Hundepool et al 2012). **Error! Reference source not found.**3 shows the sampling distribution of the sample variance under the assumption of normality and, for the same population, the simulated sampling distribution of the sample variance (constructed by sampling from the specified population).

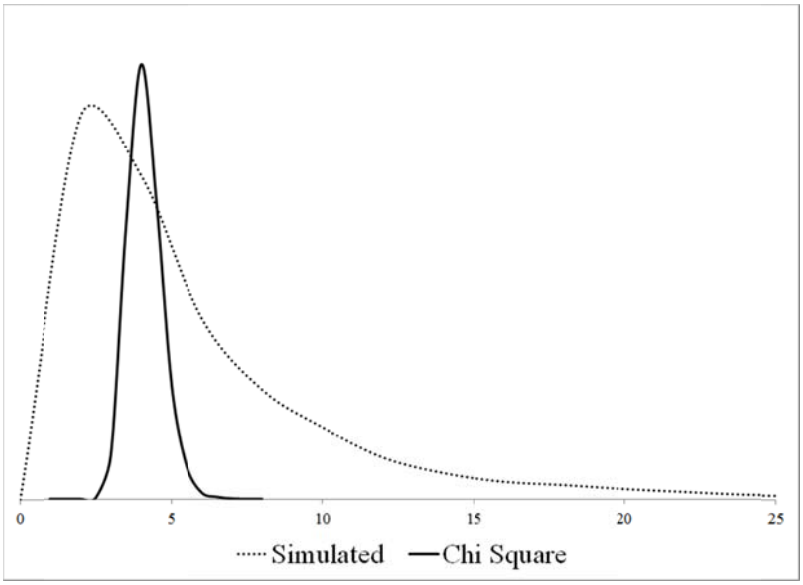

Figure 3. Distribution of sample variance using simulation and Chi-square

It is evident that the sampling distribution of the sample variance under the assumption of normality is quite different from the simulated sampling distribution of the sample variance. If the data administrator were to use the parametric Chi-square distribution to specify the noise level (for either multiplicative or additive perturbation), it is likely to be small relative to the standard error of the sample variance.

In general, in order to provide a response that is appropriate (useful) and adequate (prevents disclosure), it is necessary to have knowledge of the sampling distribution of the statistic. The sampling distribution of the statistic is essentially the distribution of the value of the statistic when random samples of a given size are chosen from the population. Releasing the masked value from this distribution is essentially the

equivalent of selecting another random sample and releasing the value of the statistic from this new sample in place of the original response and provides the same level of usefulness as the original response. The masked value selected in this manner also provides adequate confidentiality protection since it represents an independent sample from the population and is completely unrelated to the original data. The intruder has no way to identify the particular sample that was generated. Unfortunately, as we have seen, constructing the sampling distribution of the statistic can be extremely difficult except in a few simple scenarios such as the mean where the central limit theorem assures us that the sampling distribution of the sample mean will be approximately normal for relatively large sample sizes and is also robust. For other statistics, this is not the case.

One of the key objectives of RAS is to provide the users with the ability to perform many types of statistical analyses while ensuring that the responses are both useful and prevent disclosure of potentially sensitive data. As noted earlier, it is highly desirable that the confidentialization measures in a RAS be automated and require limited or no manual intervention. We have shown in this section that traditional masking mechanisms are unable to achieve these objectives. In the following section, we describe a new procedure that better fulfils these objectives.

4 A description of the bootstrap mechanism

The concept of the bootstrap was originally proposed by Efron (1979). [3] Let θ and $\hat{\theta}$ represent the parameter and statistic of interest. The bootstrap method is designed to estimate the standard error of a statistic or to construct an empirical sampling distribution of a statistic by resampling, with replacement, from the data that has been collected. According to Efron and Tibshirani (1993),

> … the bootstrap data points $x_1^*, x_2^*, \dots, x_n^*$ are a random sample of size n drawn *with* replacement rom the population of n objects $(x_1, x_2, \dots, x_n)$.

Let $\hat{\theta}_j^*$ represent the value of the statistic from bootstrap sample j. The entire process of collecting bootstrap samples is repeated B times resulting in a collection of $\widehat{\boldsymbol{\theta}}^* = \left(\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*\right)$. Efron (1979) showed that $\widehat{\boldsymbol{\theta}}^*$ allows us to compute the standard error of $\hat{\theta}$ with reasonable accuracy. As $B \to \infty$, the standard error of $\widehat{\boldsymbol{\theta}}^*$ approaches the standard error of $\hat{\theta}$ for data sets $\boldsymbol{x}_j^*$ of size n randomly sampled from $\hat{F}$ (the observed empirical distribution function) (Efron and Tibshirani 1993). The collection $\left(\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*\right)$**,** when represented as a histogram is often referred to as the bootstrap distribution of $\hat{\theta}$. Singh (1981) showed that the distribution of $\left(\hat{\theta}^* - \hat{\theta}\right)$

[3] Note that we are using the traditional statistical notation only for ease of understanding. We do not contend that the records in the query subset are independently, identically distributed.

provides a fairly accurate estimate of the distribution of $(\hat{\theta} - \theta)$. This is an important characteristic of the bootstrap distribution that allows for the construction of confidence intervals for θ based on the percentiles of the distribution of $\hat{\theta}^*$. For a comprehensive discussion of the bootstrap methods and its applications, see Efron and Tisbshirani (1986, 1993).[4]

The primary objective of any masking mechanism to provide a response that closely emulates the true characteristics of $\hat{\theta}$. Thus, if we consider the query set as $\boldsymbol{x}$ as the "population", then a random bootstrap sample $\boldsymbol{x}^*$ is the equivalent of drawing a random sample from a population. With $B \to \infty$, the statistic $\hat{\theta}^*$ computed from this random bootstrap sample has the following characteristics:

(1) The distribution of $\hat{\theta}^*$ closely approximates the sampling distribution of $\hat{\theta}$,
(2) If $\hat{\theta}$ is an unbiased estimator, then $E(\hat{\theta}^*) = \hat{\theta}$, and
(3) Variance of $\hat{\theta}^* = \sigma^2_{\hat{\theta}}$, the variance of $\hat{\theta}$.

There is an additional advantage of using the bootstrap in the RAS context. Since the bootstrap "substitutes considerable amounts of computation in place of theoretical analysis." (Efron and Tibshirani 1986), in the traditional application of the bootstrap, it would be necessary to repeat the process of selecting the bootstrap sample hundreds or thousands of times (and obviously impossible to construct the bootstrap with $B \to \infty$). This *may* present computational difficulties for some complicated statistics. In the RAS context, even this is not a problem; we are not interested in estimating the standard error of the statistic or the bootstrap distribution. We are simply interested in generating a masked response $\hat{\theta}^*$ which has the desirable properties listed above. Hence, *all that is required is to draw a single bootstrap sample* $\boldsymbol{x}^*$ *of size n drawn from* $\boldsymbol{x}$ *with replacement, and release* $\hat{\theta}^*$ *computed using* $\boldsymbol{x}^*$**.**

The *Bootstrap Mechanism* that we propose in this paper is simple and straightforward. The procedure is implemented as follows.

(1) Assume a query requesting some summary statistic from data set $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$. Let $\hat{\theta}$ be the *true value* of the statistic calculated from the data set. Note that both $\boldsymbol{x}$ and $\hat{\theta}$ could be multivariate.
(2) Select a random sample of size *n*, *with replacement*, from the original query set $\boldsymbol{x}$ resulting in a new data set $\boldsymbol{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$. Note that since $\boldsymbol{x}^*$ is

[4] The bootstrap and the related jackknife have been previously addressed in the disclosure limitation literature by the following studies: Jones and Adam (1989), Heer (1993), Fienberg (1994, 1996), Domingo-Ferrer & Mateo-Sanz (1999), Raghunathan et al (2003), Heitzig (2005, 2006), Ichim (2010), and Melville and McQuaid (2012), but not for the general purpose of masking output resulting from statistical analyses.

drawn *with replacement* from $\boldsymbol{x}$, some values in $\boldsymbol{x}$ may appear in $\boldsymbol{x}^*$ multiple times, while others may not appear at all.

(3) Compute the value of $\hat{\theta}^*$ (the value of the summary statistic from the sample $\boldsymbol{x}^*$).

(4) Release $\hat{\theta}^*$.

That the bootstrap satisfies the first requirement of an effective masking mechanism is evident. That it satisfies the last three requirements follow from the theory of the statistical bootstrap.

In summary, the bootstrap mechanism provides a response from the bootstrap distribution (the empirical sampling distribution) of the statistic given the observed data. From the data utility perspective, we are assured that the response from the bootstrap mechanism is appropriate since it is the closest equivalent of selecting a random sample from the given data (without actually drawing such a sample). From a disclosure risk perspective, in order to provide a useful response, the noise cannot be any more or any different from the sampling distribution of the statistic. Thus, the bootstrap mechanism induces just the adequate level of noise in order ensure an appropriate response.

4.1 Utility of the bootstrap mechanism

In this section, we illustrate the usefulness of the bootstrap mechanism by revisiting the motivating illustrations that we provided in Section 3. The example that was used earlier related to the sample variance from a heavily skewed data set. Figure 4 provides the simulated sampling distribution of the sample variance (using the population information) and the bootstrap distribution constructed using only the sample data (B = 10000). The bootstrap distribution is robust and provides a good approximation of the sampling distribution of the sample variance. By contrast, in Figure 3 we saw that the sampling distribution of sample variance under the normality assumption was very different from the simulated distribution. Using the bootstrap distribution would enable the data administrator to provide a masked response that is both appropriate and adequate as shown in Figure 4.

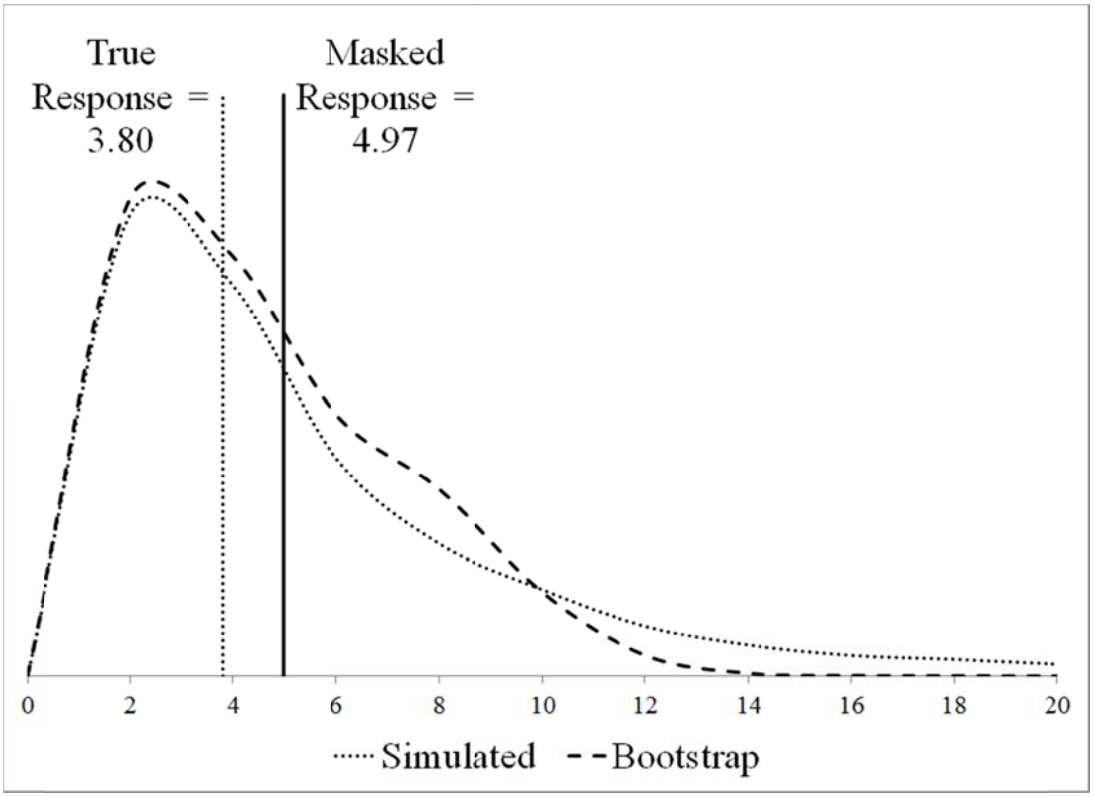

Figure 4. Simulated and bootstrap distribution of the sample variance

It is important to note that the data administrator *does not have to construct the bootstrap distribution*. It is only provided for information purposes. All that the data administrator has to do is to select a bootstrap sample (sample n items with replacement from the original sample) and release the value of the variance for this sample. The theory of the bootstrap assures us that the resulting masked response constitutes an appropriate response since it is the closest equivalent of drawing an independent sample and releasing the value of the statistic (sample variance) from this sample. It provides an adequate response since the noise variance induced by the bootstrap mechanism (variance of the masked value) equals the variance of the sampling distribution of the sample variance. One of the biggest strengths of the bootstrap is that it is non-parametric and is hence robust.[5] And the bootstrap mechanism can be easily extended to multiple statistics that result from a single analysis (such as regression analysis).

[5] It is precisely because the bootstrap is robust that most popular software packages (such as SPSS and SAS) offer the bootstrap as a non-parametric alternative for estimating the standard error.

4.2 Information loss and disclosure risk characteristics

One of the key features of the bootstrap mechanism is that the theory behind the statistical bootstrap allows us to derive the information loss and disclosure risk characteristics of the bootstrap mechanism for output masking. Due to space limitations, we are unable to provide the details here. A full version of the paper has been submitted to an academic journal. Interested readers are requested to contact one of the authors for a copy of the complete paper.

5 Future research and conclusions

There are several interesting research avenues that can be pursued in the further development of the bootstrap mechanism:

(1) The use of Bayesian bootstrap (Rubin 1981) for improving inferential effectiveness,
(2) Releasing multiple responses selected from the bootstrap distribution and deriving the data utility and disclosure risk characteristics of such a mechanism,
(3) Using the bootstrap mechanism for tabular data, and
(4) Relationship between differential privacy, particularly smooth sensitivity (Nissim et al 2007), and the bootstrap mechanism.

In RASs, the real challenge is to ensure confidentiality protection while also ensuring that responses in terms of analysis outputs are reasonable. Given a query set selected on some qualifying attributes, one way to address this challenge would be to select an independent random sample from the population, conduct the analysis on the new sample, and release the output computed from this new sample. However, doing this for every query, or even for one query in most situations, is infeasible. Without the ability to gather a new sample, we argue that the best alternative is to use the bootstrap. While the bootstrap distribution will not provide the same usefulness and disclosure risk characteristics as drawing an independent random sample, it provides a good estimate of the sampling distribution of the statistic given the observed data. Hence, the response computed from the bootstrap sample is reasonable – it is a value that comes from the sampling distribution of the statistic (one that can be observed if calculated on a different sample of the population with the same characteristics). In order to provide a reasonable response, we argue that the bootstrap mechanism also provides the best confidentiality protection – any more or different noise would result in an unreasonable response.

References

Please contact the authors for the list of references.