The Data Privacy Problem: Computer Science, Statistics and Future Directions

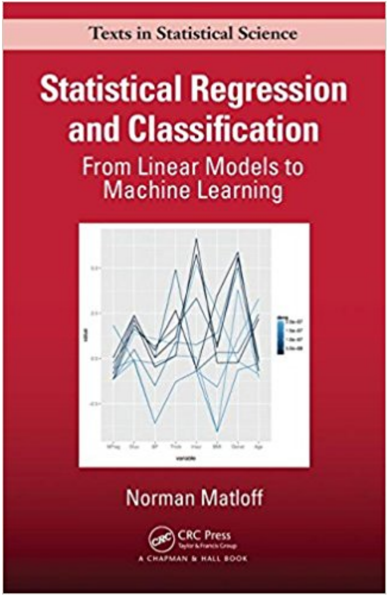
Norm Matloff
University of California at Davis

SAE2017

These will be slides available at
`http://heather.cs.ucdavis.edu/sae2017.pdf`

Shameless Promotion

Shameless Promotion



Out July 28!

(A longheld plan — decades — now finally got around to it.)

Where I Am Coming From

Where I Am Coming From

- Born and raised in LA.
- PhD in Pure Math, UCLA (theoretical probability)
- Was one of the founders of UC Davis Stat Dept. Did applied stat methodology.
- Later switched to CS Dept. — but still, much of my research is statistical.
- **New to SAE field**.

Plan of the Talk

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.
- CS vs. Stat — "Never the twain shall meet."

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.
- CS vs. Stat — "Never the twain shall meet."
- My old *Biometrika* paper.

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.
- CS vs. Stat — "Never the twain shall meet."
- My old *Biometrika* paper.
    - Regression averaging.

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.
- CS vs. Stat — "Never the twain shall meet."
- My old *Biometrika* paper.
    - Regression averaging.
    - Application to SAE.

Plan of the Talk

- Overview of Statistical Disclosure (SDC) Control methods, Then and Now.
- CS vs. Stat — "Never the twain shall meet."
- My old *Biometrika* paper.
    - Regression averaging.
    - Application to SAE.
    - Application to SDC.

Statistical Data Security: Overview

Statistical Data Security: Overview

Commonly-used example:

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.
- Need statistical access, e.g. regression analysis, to investigate discrimination claim.

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.
- Need statistical access, e.g. regression analysis, to investigate discrimination claim.
- But want to protect privacy of individuals.

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.
- Need statistical access, e.g. regression analysis, to investigate discrimination claim.
- But want to protect privacy of individuals.
- Say snooper knows there is just one female electrical engineer, Ms. X.

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.
- Need statistical access, e.g. regression analysis, to investigate discrimination claim.
- But want to protect privacy of individuals.
- Say snooper knows there is just one female electrical engineer, Ms. X.
- He submits a "statistical" query: Mean salary of all female EEs.

Statistical Data Security: Overview

Commonly-used example:

- Gender discrimination lawsuit.
- Need statistical access, e.g. regression analysis, to investigate discrimination claim.
- But want to protect privacy of individuals.
- Say snooper knows there is just one female electrical engineer, Ms. X.
- He submits a "statistical" query: Mean salary of all female EEs. Thus snooper learns Ms. X's salary.

Methodology: History

Methodology: History

- Long history, going back to 1980s or even earlier.

Norm Matloff
University of California at Davis

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).
- Fancy name now: Statistical Disclosure Control.

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).
- Fancy name now: Statistical Disclosure Control.
- Computer Science picks up the issue: *Differential privacy* (Dwork, 2006); major research issue now in CS.

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).
- Fancy name now: Statistical Disclosure Control.
- Computer Science picks up the issue: *Differential privacy* (Dwork, 2006); major research issue now in CS.
- Warnings:

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).
- Fancy name now: Statistical Disclosure Control.
- Computer Science picks up the issue: *Differential privacy* (Dwork, 2006); major research issue now in CS.
- Warnings:
    - There is no fully-satisfactory method.

Methodology: History

- Long history, going back to 1980s or even earlier.
- Current state of the art: See e.g. books by (G. Duncan *et al*, 2011); (Hundepool *et al*, 2012).
- Fancy name now: Statistical Disclosure Control.
- Computer Science picks up the issue: *Differential privacy* (Dwork, 2006); major research issue now in CS.
- Warnings:
  - There is no fully-satisfactory method.
  - Significant divergence between CS and Stat views.

Methodology: General categories

Methodology: General categories

- Data suppression.

Methodology: General categories

- Data suppression.
    - Suppression (replacement by NA) of small cells in contingency tables.

Methodology: General categories

- Data suppression.
  - Suppression (replacement by NA) of small cells in contingency tables.
- Data perturbation.

Methodology: General categories

- Data suppression.
    - Suppression (replacement by NA) of small cells in contingency tables.
- Data perturbation.
    - Rounding.

Methodology: General categories

- Data suppression.
  - Suppression (replacement by NA) of small cells in contingency tables.
- Data perturbation.
  - Rounding.
  - Data swapping.

Methodology: General categories

- Data suppression.
  - Suppression (replacement by NA) of small cells in contingency tables.
- Data perturbation.
  - Rounding.
  - Data swapping.
  - Noise addition.

Methodology: General categories

- Data suppression.
  - Suppression (replacement by NA) of small cells in contingency tables.
- Data perturbation.
  - Rounding.
  - Data swapping.
  - Noise addition.
- Most/all methods are in these categories.

NM's "Pillow" Theorem

NM's "Pillow" Theorem

Pound down on one part of a fluffy pillow, and another part will pop up. :-)

NM's "Pillow" Theorem

Pound down on one part of a fluffy pillow, and another part will pop up. :-)

Any SDC method suffers from some combination of

NM's "Pillow" Theorem

Pound down on one part of a fluffy pillow, and another part will pop up. :-)

Any SDC method suffers from some combination of

- increased bias

NM's "Pillow" Theorem

Pound down on one part of a fluffy pillow, and another part will pop up. :-)

Any SDC method suffers from some combination of

- increased bias
- increased variance

NM's "Pillow" Theorem

Pound down on one part of a fluffy pillow, and another part will pop up. :-)

Any SDC method suffers from some combination of

- increased bias
- increased variance
- insufficient protection of privacy

Example: Cell suppression

Example: Cell suppression

- Again think of the company with just 1 female EE.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.
- Creates a bias, potentially substantial.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.
- Creates a bias, potentially substantial. E.g. say $X^{(i)}$ has lots of rare values, but is correlated (in whatever sense) with $X^{(j)}$.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.
- Creates a bias, potentially substantial. E.g. say $X^{(i)}$ has lots of rare values, but is correlated (in whatever sense) with $X^{(j)}$. Attenuates correlation.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.
- Creates a bias, potentially substantial. E.g. say $X^{(i)}$ has lots of rare values, but is correlated (in whatever sense) with $X^{(j)}$. Attenuates correlation.
- "Protection" may be illusory. E.g. snooper queries total salary of all EEs, then for male EEs, and subtracts to get female EE wage.

Example: Cell suppression

- Again think of the company with just 1 female EE. (A "small area." More on this later.)
- Say policy is to render as NAs all cells of size ≤ 1.
- Creates a bias, potentially substantial. E.g. say $X^{(i)}$ has lots of rare values, but is correlated (in whatever sense) with $X^{(j)}$. Attenuates correlation.
- "Protection" may be illusory. E.g. snooper queries total salary of all EEs, then for male EEs, and subtracts to get female EE wage.

  Various schemes to cope with this, but all complex and of unclear value.

Norm Matloff
University of California at Davis

Example: Noise Addition

Example: Noise Addition

- Add random noise to each variable.

Example: Noise Addition

- Add random noise to each variable.
- A favorite of the CS crowd in the early 80s, now replaced in CS by differential privacy.

Example: Noise Addition

- Add random noise to each variable.
- A favorite of the CS crowd in the early 80s, now replaced in CS by differential privacy.
- Variance/protection tradeoff.

Example: Noise Addition

- Add random noise to each variable.
- A favorite of the CS crowd in the early 80s, now replaced in CS by differential privacy.
- Variance/protection tradeoff.
- But attenuates relations among variables.

Example: Noise Addition

- Add random noise to each variable.
- A favorite of the CS crowd in the early 80s, now replaced in CS by differential privacy.
- Variance/protection tradeoff.
- But attenuates relations among variables.
- Can add noise with same covariance matrix as the data to try to remedy (Matloff, 1986); (Kim, 1986); (Tendick and Matloff, 1994).

Example: Noise Addition

- Add random noise to each variable.
- A favorite of the CS crowd in the early 80s, now replaced in CS by differential privacy.
- Variance/protection tradeoff.
- But attenuates relations among variables.
- Can add noise with same covariance matrix as the data to try to remedy (Matloff, 1986); (Kim, 1986); (Tendick and Matloff, 1994).
- Presents a problem with discrete/categorical variables.

Divergence between CS and Stat

Divergence between CS and Stat

- Famously noted in (Breiman, 2001), but divergence is arguably even worse today.

Divergence between CS and Stat

- Famously noted in (Breiman, 2001), but divergence is arguably even worse today.
- Somewhat simplified summary (my view, not Breiman's):

Divergence between CS and Stat

- Famously noted in (Breiman, 2001), but divergence is arguably even worse today.
- Somewhat simplified summary (my view, not Breiman's):

	stat	CS
data source	sample from pop.	"it just exists"
math tools	asymptotics	famous prob. ineqs.
research funding	poor	generous
extern. perception	relic of the past	exciting panacea

Divergence between CS and Stat

- Famously noted in (Breiman, 2001), but divergence is arguably even worse today.
- Somewhat simplified summary (my view, not Breiman's):

	stat	CS
data source	sample from pop.	"it just exists"
math tools	asymptotics	famous prob. ineqs.
research funding	poor	generous
extern. perception	relic of the past	exciting panacea

- Examples: Deep learning; differential privacy.

Random Perturbation in the Discrete/Categrical Case

Random Perturbation in the Discrete/Categrical Case

- Data swapping: "Trade some of X's variables for Y's." Again, the attenuation issue is a problem.

Random Perturbation in the Discrete/Categrical Case

- Data swapping: "Trade some of X's variables for Y's." Again, the attenuation issue is a problem.
- Log-linear models, e.g. (Manrique-Vallier and Reter, 2012).

Random Perturbation in the Discrete/Categrical Case

- Data swapping: "Trade some of X's variables for Y's." Again, the attenuation issue is a problem.
- Log-linear models, e.g. (Manrique-Vallier and Reter, 2012). Users view only log-lin fit, so original data hidden.

Random Perturbation in the Discrete/Categrical Case

- Data swapping: "Trade some of X's variables for Y's." Again, the attenuation issue is a problem.
- Log-linear models, e.g. (Manrique-Vallier and Reter, 2012). Users view only log-lin fit, so original data hidden.
- (Matloff and Tendick, 2015) — next slide

Work in Progress

Work in Progress

- Works with any data, continuous, discrete etc.

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.
- For unit $r_i = (W_{i1}, ..., W_{ip})$ in original, w.p. q replace r_i by $r_i' = (W_{i1}', ..., W_{ip}')$ as follows:

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.
- For unit $r_i = (W_{i1}, ..., W_{ip})$ in original, w.p. q replace r_i by $r_i' = (W_{i1}', ..., W_{ip}')$ as follows:
  (a) Find ϵ-neighborhood S of r_i.

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.
- For unit $r_i = (W_{i1}, ..., W_{ip})$ in original, w.p. q replace r_i by $r_i' = (W_{i1}', ..., W_{ip}')$ as follows:
  (a) Find ϵ-neighborhood S of r_i.
  (b) For $j = 1, ..., p$, *independently* set W_{ij}' to be a random value chosen from the values of variable j in S

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.
- For unit $r_i = (W_{i1}, ..., W_{ip})$ in original, w.p. q replace r_i by $r_i' = (W_{i1}', ..., W_{ip}')$ as follows:
  (a) Find ϵ-neighborhood S of r_i.
  (b) For $j = 1, ..., p$, *independently* set W_{ij}' to be a random value chosen from the values of variable j in S
- **Key point:** We are not estimating the joint distribution of the p variables at all!

Work in Progress

- Works with any data, continuous, discrete etc.
- Say we have p variables.
- For unit $r_i = (W_{i1}, ..., W_{ip})$ in original, w.p. q replace r_i by $r_i' = (W_{i1}', ..., W_{ip}')$ as follows:
  (a) Find ϵ-neighborhood S of r_i.
  (b) For $j = 1, ..., p$, *independently* set W_{ij}' to be a random value chosen from the values of variable j in S
- **Key point:** We are not estimating the joint distribution of the p variables at all!
- Just sampling from the marginal distributions suffices. Can prove this works for small ϵ.

Differential Privacy

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition.

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.
- But can be applied much more generally, e.g. with randomized response surveys.

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.
- But can be applied much more generally, e.g. with randomized response surveys.
- Not just a method, but a "philosophy." Very formal math definition of privacy.

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.
- But can be applied much more generally, e.g. with randomized response surveys.
- Not just a method, but a "philosophy." Very formal math definition of privacy. Can't fit on slide, but basically asks, How much will a function of the data change if one row changes?

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.
- But can be applied much more generally, e.g. with randomized response surveys.
- Not just a method, but a "philosophy." Very formal math definition of privacy. Can't fit on slide, but basically asks, How much will a function of the data change if one row changes?
- Lots of impressive uses of inequalities, e.g. Chernoff.

Differential Privacy

(Disclaimer: I have only limited exposure to DP, as I do not consider it to answer the questions of interest to statisticians.)

- In its most common form, just (Laplace-distributd) noise addition. Most implementations do NOT deal with the attenuation problem.
- But can be applied much more generally, e.g. with randomized response surveys.
- Not just a method, but a "philosophy." Very formal math definition of privacy. Can't fit on slide, but basically asks, How much will a function of the data change if one row changes?
- Lots of impressive uses of inequalities, e.g. Chernoff. But not focused on estimation, standard errors etc.

Norm Matloff
University of California at Davis

Connecting to SAE (I)

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

- Regression average (RA) for improved estimation of means.

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

- Regression average (RA) for improved estimation of means.
- Estimate the regression function $m(t) = E(Y|X = t)$, say with a parametric model, $m(t) = g(t, \theta)$.

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

- Regression average (RA) for improved estimation of means.
- Estimate the regression function $m(t) = E(Y|X = t)$, say with a parametric model, $m(t) = g(t, \theta)$.
- $\mu = EY$, $\widehat{\mu} = \overline{Y}$.

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

- Regression average (RA) for improved estimation of means.
- Estimate the regression function $m(t) = E(Y|X = t)$, say with a parametric model, $m(t) = g(t, \theta)$.
- $\mu = EY$, $\widehat{\mu} = \overline{Y}$.
- Set $\breve{\mu}$ to average value of $\widehat{m}$ over data,

$$\breve{\mu} = \frac{1}{n} \sum_i g(X_i, \widehat{\theta})$$

Connecting to SAE (I)

My old (ancient) *Biometrika* paper:

- Regression average (RA) for improved estimation of means.
- Estimate the regression function $m(t) = E(Y|X = t)$, say with a parametric model, $m(t) = g(t, \theta)$.
- $\mu = EY$, $\widehat{\mu} = \overline{Y}$.
- Set $\check{\mu}$ to average value of $\widehat{m}$ over data,

$$\check{\mu} = \frac{1}{n} \sum_i g(X_i, \widehat{\theta})$$

Connecting to SAE (I), cont'd.

Norm Matloff
University of California at Davis

Connecting to SAE (I), cont'd.

- No assumption on $F_{Y|X}$

Connecting to SAE (I), cont'd.

- No assumption on $F_{Y|X}$
- Asympt. distribution derived.

Connecting to SAE (I), cont'd.

- No assumption on $F_{Y|X}$
- Asympt. distribution derived.
- Can prove for parametric regression models

$$AVar(\check{\mu}) < AVar(\widehat{\mu})$$

Connecting to SAE (I), cont'd.

- No assumption on $F_{Y|X}$
- Asympt. distribution derived.
- Can prove for parametric regression models

$$AVar(\check{\mu}) < AVar(\widehat{\mu})$$

except if g is linear regression with a constant term.

Connecting to SAE (II)

Connecting to SAE (II)

("Standard" statistical setting, not explicitly fine-pop. model.)

Connecting to SAE (II)

("Standard" statistical setting, not explicitly fine-pop. model.)

Say have unit-level data, but in one area A have X data but little or no Y data.

Connecting to SAE (II)

("Standard" statistical setting, not explicitly fine-pop. model.)

Say have unit-level data, but in one area A have X data but little or no Y data.

Use RA to estimate area mean:

Connecting to SAE (II)

("Standard" statistical setting, not explicitly fine-pop. model.)

Say have unit-level data, but in one area A have X data but little or no Y data.

Use RA to estimate area mean:

$$\check{\mu} = \frac{1}{n(A)} \sum_{X_i \text{ in } A} g(X_i, \widehat{\theta})$$

Connecting to SAE (II)

("Standard" statistical setting, not explicitly fine-pop. model.)

Say have unit-level data, but in one area A have X data but little or no Y data.

Use RA to estimate area mean:

$$\check{\mu} = \frac{1}{n(A)} \sum_{X_i \text{ in } A} g(X_i, \widehat{\theta})$$

(Assumes same θ in all areas.)

Back to SDC

Back to SDC

- Back to the example of gender discrimination lawsuit.

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.
- To investigate discrimination claim, may wish to estimate μ, population mean salary EY for female EEs.

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.
- To investigate discrimination claim, may wish to estimate μ, population mean salary EY for female EEs. (Simple case here, to keep exposition simple.)

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.
- To investigate discrimination claim, may wish to estimate μ, population mean salary EY for female EEs. (Simple case here, to keep exposition simple.)
- Say have data on age, education and so on, in vector X for each worker,

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.
- To investigate discrimination claim, may wish to estimate μ, population mean salary EY for female EEs. (Simple case here, to keep exposition simple.)
- Say have data on age, education and so on, in vector X for each worker, and have Y values but want to keep them hidden.

Back to SDC

- Back to the example of gender discrimination lawsuit.
- Say have k female EEs, k small.
- To investigate discrimination claim, may wish to estimate μ, population mean salary EY for female EEs. (Simple case here, to keep exposition simple.)
- Say have data on age, education and so on, in vector X for each worker, and have Y values but want to keep them hidden.
- Solution: Use RA over those X values.

Conclusions

Conclusions

- No really satisfactory solution to SDC problem, IMO.

Conclusions

- No really satisfactory solution to SDC problem, IMO.
- But here I introduced two new ones anyway, both works in progress.

Conclusions

- No really satisfactory solution to SDC problem, IMO.
- But here I introduced two new ones anyway, both works in progress.
- The second solution also is new methodology for SAE.