What Can Research on Data Confidentiality Teach Us about Data Quality?

Alan F. Karr

National Institute of Statistical Sciences

DIMACS Data Quality Workshop: February 3–4, 2011

NISS
The Statistics Community Serving the Nation

My World View

From Karr et al. (2006):

> *Data quality is the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions.*

Put differently, DQ measures the capability of data to support *sound decisions based on statistical inferences drawn from the data*.

Therefore, DQ is a decision problem: quality comes only at a cost, which may be economic or not.

Surveys are a case in which tradeoffs are explicit

The Question Underlying the Research

Can knowledge about controllable DQ effects in the context of data confidentiality (DC) inform knowledge about uncontrollable DQ effects in other contexts? (And *vice versa*?)

The Central Tension

Context Official statistics agencies, which must both

- Protect confidentiality of data and privacy of data subjects
- Make data, or at least information derived from data, available for research, policy and other purposes

Problem DQ (data utility) conflicts directly with disclosure risk

But What do we mean by risk and utility?

Compounding Factor One person's risk is another person's utility. Put differently, it is hard to distinguish legitimate users from intruders.

Statistical Disclosure Limitation (SDL) Strategies

Restricted Access to "Real" Data At centers or via licensing

Restricted Analyses User submits analysis (e.g., SAS code), agency reviews it, performs it if it is deemed safe and reports subset of results following disclosure review

Altered Analyses User submits analysis, agency performs it and alters results before reporting them to user

** **Public Microdata Releases** Agency alters data and makes them available publicly

What is Disclosure?

Identity Disclosure Record-level identification of subjects (individuals or establishments), essentially always by linkage to a dataset containing identifiers

Attribute Disclosure Of sensitive attributes, such as income or health status

Inferential Disclosure On the basis of a statistical model

Note No concept of harm or loss

SDL for Microdata

Agency Goal Alter the data before release, converting *original* database $\mathcal{D}_{\text{original}}$ to *masked* database $\mathcal{D}_{\text{masked}}$, ideally in way that decreases risk a lot and decreases utility only a little.

Agency Must Decide

- How to measure risk
- How to measure utility
- How to make the tradeoff

Examples 1: The Truth But Not the Whole Truth

- Drop explicit identifiers (name, address, SSN, . . .)
- Suppress cells in tables (usually, of small counts)
- Coarsen values (rounding, category aggregation, top-coding, . . .)

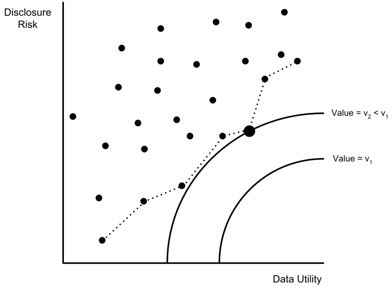
Examples 2: Not the Truth

- Microaggregation
- Noise addition
- Data swapping
- Imputation/Synthetic data
- Combinations (example: microaggregation followed by addition of noise with same covariance structure as original data)

Risk–Utility Paradigm

Steps

1. Create multiple candidates for $\mathcal{D}_{\mathrm{masked}}$
2. Assign quantified risk and utility to each
3. Agency can then make principled decision, exploiting risk-utility frontier



Some Things We Don't Understand

- Query interaction: answering one query makes others more risky
- Transparency
- How to deal with survey weights (more later)

Core Idea

Basis Utility is the opposite of distortion

In symbols,

$$DU(\mathcal{D}_{\text{masked}}) = -d(\mathcal{D}_{\text{masked}}, \mathcal{D}_{\text{original}}),$$

where d is a metric between datasets

Broad But Blunt Measures

- For categorical data, Hellinger distance between associated contingency tables
- For numerical data, Kullback-Liebler distance between estimated multi-dimensional densities
- For any data, output of classifier or propensity score model applied to (“stacked”) union of $\mathcal{D}_{\text{original}}$ and $\mathcal{D}_{\text{masked}}$

Inference-Based Measures, Typically Analysis-Specific

- For categorical data, log-likelihood of $\mathcal{D}_{\text{masked}}$ under log-linear model fit to $\mathcal{D}_{\text{original}}$
- For numerical data, overlap of confidence regions for regression models fit to $\mathcal{D}_{\text{masked}}$ and $\mathcal{D}_{\text{original}}$

A Problem: What's Between Blunt and Narrow?

Verification Servers

Idea User receives information from agency about fidelity of analysis performed on $\mathcal{D}_{\text{masked}}$ to same analysis performed on $\mathcal{D}_{\text{original}}$

Issues Precision of fidelity measures, analyses that subset the data too finely

Components for DQ

Recapitulating notation,

- True database $\mathcal{D}_{\text{true}}$ (exists only conceptually, if that)
- Actual database $\mathcal{D}_{\text{actual}}$
- $\mathcal{K}$ = available knowledge, especially about how $\mathcal{D}_{\text{true}}$ became $\mathcal{D}_{\text{actual}}$

The Main Idea in this Talk

Analogy

DC	DQ
Original Data $\mathcal{D}_{\text{original}}$	True Data $\mathcal{D}_{\text{true}}$
Masked Data $\mathcal{D}_{\text{masked}}$	Actual Data $\mathcal{D}_{\text{actual}}$

DC Compute $d(\mathcal{D}_{\text{masked}}, \mathcal{D}_{\text{original}})$ to measure utility

DQ Compute $d(\mathcal{D}_{\text{actual}}, \mathcal{D}_{\text{true}})$ to measure quality, but can't, so what about $d(\mathcal{D}_{\text{actual}}, \widehat{\mathcal{D}_{\text{true}}})$?

Bayesian View: Compare Decisions and Analyses

- Statistical analyses are vector-valued functions $\mathbf{f}(\mathcal{D})$ of a database $\mathcal{D}$
- So use $d(\mathbf{f}(\mathcal{D}_{\text{actual}}), \mathbf{f}(\mathcal{D}_{\text{true}}))$, where d is a numerical measure of the fidelity of inferences
- Therefore, have to construct estimate

$$\widehat{\mathbf{f}(\mathcal{D}_{\text{true}})} = \int_D \mathbf{f}(d)\, dP\{\mathcal{D}_{\text{true}} = d | \mathcal{D}_{\text{actual}}, \mathcal{K}\}$$

and use $d(\mathbf{f}(\mathcal{D}_{\text{actual}}), \widehat{\mathbf{f}(\mathcal{D}_{\text{true}})})$

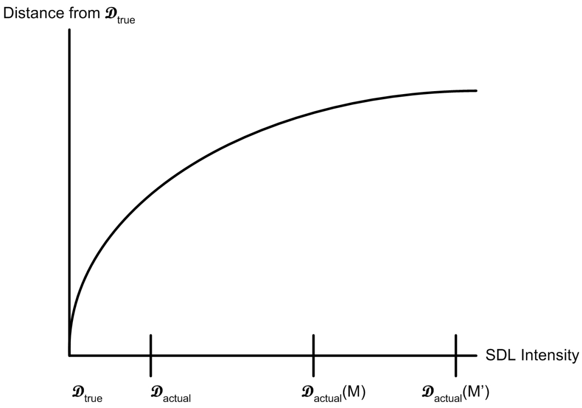
How to Estimate $\mathbf{f}(\mathcal{D}_{\text{true}})$?

Goal Understand and reason about $d(\mathbf{f}(\mathcal{D}_{\text{actual}}), \widehat{\mathbf{f}(\mathcal{D}_{\text{true}})})$, and ultimately about $d(\mathbf{f}(\mathcal{D}_{\text{actual}}), \mathbf{f}(\mathcal{D}_{\text{true}}))$

Strategy Apply statistical disclosure limitation (SDL) procedures M with varying intensities to $\mathcal{D}_{\text{actual}}$, yielding altered databases $\mathcal{D}_{\text{actual}}(M)$, and use differences $d(\mathbf{f}(\mathcal{D}_{\text{actual}}, \mathbf{f}(\mathcal{D}_{\text{actual}}(M)))$ to estimate $d(\mathbf{f}(\mathcal{D}_{\text{actual}}), \mathbf{f}(\mathcal{D}_{\text{true}}))$

The Hope Since DQ problems *attenuate structure in data*, intentionally lowering DQ (as done in DC) might be insightful about the extent to which DQ has already been lowered

Pictorial Depiction of “The Hope”

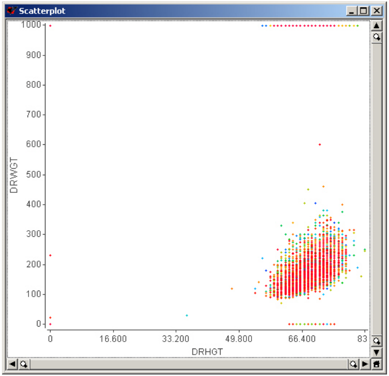


This is a Challenge!

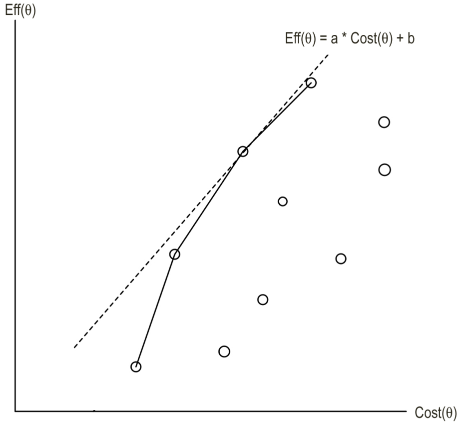
The Need Models for DQ degradation that reflect the underlying processes, and the fundamental role of people in these processes

The Issue Do extant SDL methods *M* at all resemble these processes?

Complexity of the Challenge



The Goal: Produce This Tool for Informing Decisions



Feasible Path

- Clean-up strategy S produces an (ostensibly) improved database $\mathcal{D}_{\text{cleaned}}(S)$. The extent to which inferences drawn from $\mathcal{D}_{\text{cleaned}}(S)$ are closer to those from $\mathcal{D}_{\text{true}}$ than those drawn from $\mathcal{D}_{\text{actual}}$ measures the effectiveness of S.
- Would like to—but can't—employ the improvement

$$\begin{aligned}\mathrm{Eff}(S, \mathbf{f}, \mathcal{D}_{\text{actual}}) &= d\Big(\mathbf{f}(\mathcal{D}_{\text{actual}}), \mathbf{f}(\mathcal{D}_{\text{true}})\Big) \\ &\quad - d\Big(\mathbf{f}(\mathcal{D}_{\text{cleaned}}(S)), \mathbf{f}(\mathcal{D}_{\text{true}})\Big)\end{aligned}$$

- But *can examine*

$$\mathrm{Eff}^*(S, \mathbf{f}, \mathcal{D}_{\text{actual}}) = -d\Big(\mathbf{f}(\mathcal{D}_{\text{cleaned}}(S)), \mathbf{f}(\mathcal{D}_{\text{actual}})\Big)$$

What Can We Conclude?

Can Say If $\mathrm{Eff}^*(S, \mathbf{f}, \mathcal{D}_{\mathrm{actual}})$ 0, inferences have not changed, so S was ineffective

Cannot Say If $\mathrm{Eff}^*(S_1, \mathbf{f}, \mathcal{D}_{\mathrm{actual}}) > \mathrm{Eff}^*(S_2, \mathbf{f}, \mathcal{D}_{\mathrm{actual}})$, then S_1 is more effective than S_2

Can Say If $\mathrm{Eff}^*(S_1, \mathbf{f}, \mathcal{D}_{\mathrm{actual}}) > \mathrm{Eff}^*(S_2, \mathbf{f}, \mathcal{D}_{\mathrm{actual}})$, then S_1 has changed $\mathcal{D}_{\mathrm{actual}}$ more than S_2 has

Prediction

The Need Predictive models for effectiveness, of the form

$$\widehat{\mathrm{Eff}}(\theta) = f(\theta) + \text{uncertainty}$$

No clue about how to construct such models

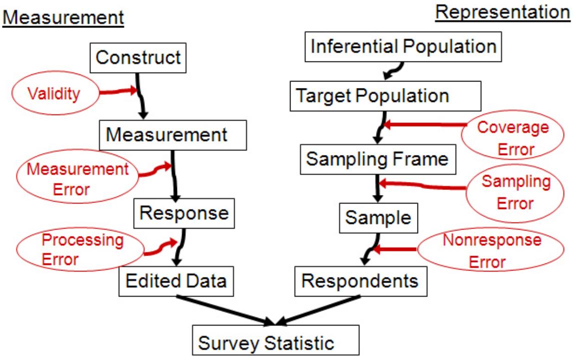
Cost

Cost models

$$\text{Cost}(\theta) = g(\theta) \qquad \text{[+ uncertainty]}$$

are even further away, especially if both process and opportunity costs must be included

Total Survey Error



Idle Speculation: Weights

Surveys Each record has a weight interpretable as the number of elements in the population that it represents (and reflecting, sample design, nonresponse, . . .), and weighted analyses are performed

DQ For inference purposes, could records be assigned weights reflecting confidence that they are "correct"? (In some settings, this is done already, when "bad" records are discarded.)