The Fate of Empirical Economics
When All Data are Private

John M. Abowd
Cornell University and U.S. Census Bureau*
Society of Government Economists
May 13, 2016

*This work was performed as a Cornell faculty member before I joined the U.S. Census Bureau in an executive capacity. No confidential data from any source were used to develop this talk.
First, a Live Demo!

• My assistants are now distributing sealed envelopes:

  CHOOSE YOUR OWN ENVELOPE
  DO NOT LET THE ASSISTANT CHOOSE

  PLEASE DO NOT OPEN
  THE ENVELOPE!!!!!!!!!!!
Step 1

OPEN THE ENVELOPE CAREFULLY (IT IS RESEALABLE)

REMOVE THE SHEET OF PAPER THAT SAYS “CIRCLE ONE”

LEAVE THE OTHER SHEET OF PAPER IN THE ENVELOPE
Step 2

WITHOUT TAKING THE OTHER SHEET OF PAPER OUT OF THE ENVELOPE:
READ THE QUESTION INSIDE YOUR ENVELOPE
MEMORIZE YOUR ANSWER
RESEAL THE ENVELOPE WITH THE QUESTION INSIDE
Step 3

ANSWER THE QUESTION ON THE SHEET OF PAPER THAT SAYS “CIRCLE ONE”
HAND YOUR ANSWER TO ONE OF THE ASSISTANTS
Step 4

WHILE THE ASSISTANTS ARE TABULATING THE DATA, SHRED YOUR QUESTION ENVELOPE IN ONE OF SHREDDERS IN THE ROOM
### Now, Let’s Analyze the Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>67</td>
<td>Sample size</td>
</tr>
<tr>
<td>Yes</td>
<td>28</td>
<td>Response to the survey question</td>
</tr>
<tr>
<td>No</td>
<td>39</td>
<td>Response to the survey question</td>
</tr>
<tr>
<td>beta_hat</td>
<td>41.8%</td>
<td>Raw percentage &quot;Yes&quot; estimate</td>
</tr>
<tr>
<td>Var[beta_hat]</td>
<td>0.0036</td>
<td>Sampling variance of raw proportion</td>
</tr>
<tr>
<td>StE[beta_hat]</td>
<td>6.0%</td>
<td>Standard error of raw percentage</td>
</tr>
<tr>
<td>Prec[beta_hat]</td>
<td>275</td>
<td>Sampling precision of raw proportion (inverse of sampling variance)</td>
</tr>
<tr>
<td>rho</td>
<td>0.5000</td>
<td>Probability that the sensitive question was asked</td>
</tr>
<tr>
<td>mu</td>
<td>0.5000</td>
<td>Probability of &quot;Yes&quot; on nonsensitive question</td>
</tr>
<tr>
<td>pi_hat</td>
<td>33.6%</td>
<td>Estimated percentage &quot;Yes&quot; to sensitive question</td>
</tr>
<tr>
<td>Var[pi_hat]</td>
<td>0.0145</td>
<td>Sampling variance of proportion &quot;Yes&quot; to sensitive question</td>
</tr>
<tr>
<td>StE[pi_hat]</td>
<td>12.1%</td>
<td>Standard error of percentage &quot;Yes&quot; to sensitive question</td>
</tr>
<tr>
<td>Prec[pi_hat]</td>
<td>69</td>
<td>Sampling precision of proportion &quot;Yes&quot; to sensitive question</td>
</tr>
<tr>
<td>Relative Precision</td>
<td>0.2500</td>
<td>Ratio of sampling precision of &quot;Yes&quot;: sensitive question/raw question</td>
</tr>
<tr>
<td>Bayes Factor</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>ln Bayes Factor</td>
<td>1.0986</td>
<td></td>
</tr>
</tbody>
</table>
Randomized Response

As a survey technique:


As a privacy-preserving data analysis system


The Basic Economics

• Scientific data quality is a pure public good (non-rival, non-excludable)

• Quantifiable privacy protection is also a pure public good (or “bad,” when measured as “privacy loss”) when supplied using the methods I will discuss shortly

• Computer scientists have succeeded in providing feasible technology sets relating the public goods: data quality and privacy protection

• These technology sets generate a quantifiable production possibilities frontier between data quality and privacy protection
The Basic Economics II

• We can now estimate the marginal social cost of data quality as a function of privacy protection—a big step forward

• The CS models are silent (or, occasionally, just wrong) about how to choose a socially optimal location on the PPF because they ignore social preferences

• To solve the social choice problem, we need to understand how to quantify preferences for data quality v. privacy protection

• For this we use the Marginal Social Cost of data quality and the Marginal Social Benefit of data quality, both measured in terms of the required privacy loss
Production Possibility Frontier

Data Quality (Relative Precision)

Privacy Protection (6 - ln Maximum Bayes Factor)
Production Possibility Frontier

Where computer scientists act like MSC = MSB

Where social scientists act like MSC = MSB
How Should We Measure MSB?

• Medical diagnosis example
• Consumer price index example
• Legislative apportionment example
• Generically: sum of all the marginal social benefits from every potential use
• Not: marginal social benefit of the highest-valued user (market solution)
Ideal Data Publication - Privacy Protection Systems

• To the maximum extent possible, *scientific analysis* should be performed on the *original confidential data*

• Publication of *statistical results* should respect a quantifiable *privacy-loss budget* constraint

• Data *publication algorithms* should provably *compose*

• Data *publication algorithms* should be provably *robust to arbitrary ancillary information*
Doing Data Analysis in This World

• Census Bureau already does this in some applications
  • OnTheMap
  • Survey of Income and Program Participation Synthetic Data
  • Synthetic Longitudinal Business Database

• Google does this
  • Randomized Aggregatable Privacy Preserving Ordinal Responses (tool for Cloud service providers to harvest browser data)

• Prototype systems allow medical record databases to do this
  • Privacy-preserving deep learning
  • Computational healthcare
Doing Data Analysis in This World

See:


Thank you!

Contacts:

john.abowd@cornell.edu