Privacy and Reproducibility: Addressing Empirical Research Concerns

Daniel L. Goroff
Opinions here are his own.
Alfred P. Sloan

- Organized and ran GM
- Sloan School, Sloan Kettering, too
- Foundation (public goods business)
- Emphasized role of data in decision making
Trust Empirical Results?

- The Economist magazine says “no.” (2013)
- Argues that a third of all research is wrong!

So need:

1. Reproducibility?
2. Privacy protection?
3. False discovery protection?
I. Using Evidence

- Tom is either a Salesman or a Librarian.
- You find out he has a Quiet personality.
- Which is more likely, S or L?

(See Kahneman & Thaler)
Conditioning on Data

- Tom is either a Salesman or a Librarian.
- You find out he has a Quiet personality.
- Which is more likely, S or L? I.e., which is bigger: Pr (S | Q) or Pr (L | Q)?
Salesman Problem

• Large “conditional probability” that Tom is Quiet given that he is a Librarian, of course.
• But say Fred is either a Salesman or a Librarian. You know nothing else.
• Now which is more likely for Fred, S or L?
Conditional Probability

- Imagine two events: $A$ and $B$
- Prob of $A$ given $B$ is: $Pr(A \mid B) = \frac{Pr(A \& B)}{Pr(B)}$

$Pr(\text{Red} \mid \text{Blue}) = ?$

$Pr(\text{Blue} \mid \text{Red}) = ?$
Bayesian Updating

- Prob of $A$ given $D$ is: $\Pr(A \mid D) = \frac{\Pr(A \& D)}{\Pr(D)}$
- Bayes Law:
  $$\frac{\Pr(A \mid D)}{\Pr(B \mid D)} = \frac{\Pr(A \& D) / \Pr(D)}{\Pr(B \& D) / \Pr(D)} = \frac{\Pr(D \mid A)}{\Pr(D \mid B)} \times \frac{\Pr(A)}{\Pr(B)}$$
- Odds = Likelihood Ratio $\times$ Base Rate
Base Rate Fallacy

- Fred is either a Salesman or a Librarian?
- There are $\sim 100x$ as many S as L in the US.
- So which is more likely for Tom, S or L?

\[
\frac{\Pr(S | Q)}{\Pr(L | Q)} = \frac{\Pr(Q | S)}{\Pr(Q | L)} \times \frac{\Pr(S)}{\Pr(L)}
\]
Legal Decisions

A taxi was involved in a hit and run accident at night. There are two companies in town, a **Blue** Cab Company that has 15% of the taxis and a **Green** Cab Company that has 85%.

A witness at the scene identified the taxi involved in the accident as **Blue**. The witness was tested under similar visibility conditions, and made correct color identifications in 80% of the trial instances.

(Kahneman & Tversky)
Base Rate Fallacy

• Let \( W = \text{“witness says blue.”} \) Bayes:

\[
\frac{\Pr(G \mid W)}{\Pr(B \mid W)} = \frac{\Pr(W \mid G)}{\Pr(W \mid B)} \times \frac{\Pr(G)}{\Pr(B)}
\]

\[
\frac{\Pr(G \mid W)}{\Pr(B \mid W)} = \frac{.2 \times .85}{.8 \times .15} = \frac{17}{12}
\]

• See also medical inquiries, Monty Hall, etc.
Business Decisions

- Management Focus Magazine (1984):
- 85% of CEOs had a pet in high school!
- \( \text{Pr(DOG | CEO)} \) vs. \( \text{Pr(CEO | DOG)} \)
- Pet theories??
Hypothesis Testing

• Talk as if studying $\text{Pr}(\text{New Hypothesis} \mid \text{Data})$
• Classically, study $\text{Pr}(\text{Data} \mid \text{Null Hypothesis})$
• Reject Null if this $p$ is small
• Call it “statistically significant”?
• Ioannidis says most published research is wrong.
• *The Economist*’s cover story...
Many Findings Are Wrong?

Let $H_1 = \text{finding is true}$, $H_0 = \text{it is false}$.

Let $D = \text{data says the finding is true}$.

The Economist takes the Base Rate as $\frac{1}{9}$.

What do scholars usually accept as evidence?

$$\frac{\Pr(H_1 \mid D)}{\Pr(H_0 \mid D)} = \frac{\Pr(D \mid H_1)}{\Pr(D \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}$$
Hypotheses, Data & Bayes

\[
\frac{\Pr(H_1 \mid D)}{\Pr(H_0 \mid D)} = \frac{\Pr(D \mid H_1)}{\Pr(D \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}
\]

Here likelihood ratio also called the Bayes Factor.

Want the numerator (power) to be \( \alpha \geq .80 \) (e.a.)

Want the denominator to be \( p \leq .05 \) (e.p.)

So odds increase by a factor of at least \( 16 = \frac{.80}{.05} \)
Example (October 19, 2013 issue):

1000 hypotheses to test empirically
100 of these are actually true
.80 acceptable “power” < Pr(D | T)
.05 acceptable p > Pr(D | F)

Expected Outcome:

80 confirmed true = 80% of 100 that are true
+45 false positives = 5% of 900 that are false
125 publishable = 80 + 45

.64 fraction true = 80/125 = 16/25 (16:9 odds)
Unlikely results
How a small proportion of false positives can prove very misleading

1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.

2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.

3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.

$$\Pr(F \mid D) = \frac{45}{125} = \frac{9}{25}$$
Prior and Posterior Odds

Prior odds get multiplied by Bayes Factor >16

What to do if this is not good enough?

If take $p = .01$ and alpha = .90, can get BF >90

\[
\frac{\Pr(T \mid D)}{\Pr(F \mid D)} = \frac{\Pr(D \mid T)}{\Pr(D \mid F)} \times \frac{\Pr(T)}{\Pr(F)}
\]

\[
\frac{16}{9} = \frac{.80}{.05} \times \frac{1}{9}
\]
Reproduction Helps

- Prior odds get multiplied by >16 first time.
- Original odds multiplied by > 256 next time.
- In terms of $x = \text{fraction true findings/look true}$,
  \[ f(x) = \frac{16x}{1+15x} = \text{fraction after one test}. \]
- And $f(f(x)) = \frac{256x}{1+255x}$ after two tests.
- Second tests of 125 that initially look true yield
  ~67 apparently true of which ~64 really are.
Fraction of apparently true findings that really are after one test or two

Computed by Wolfram|Alpha
Fraction of apparently true findings that really are after one test or two

<table>
<thead>
<tr>
<th>x</th>
<th>f(x)</th>
<th>f(f(x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6400</td>
<td>0.9660</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8000</td>
<td>0.9846</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8727</td>
<td>0.9910</td>
</tr>
<tr>
<td>0.4</td>
<td>0.9143</td>
<td>0.9942</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9412</td>
<td>0.9961</td>
</tr>
<tr>
<td>0.6</td>
<td>0.9600</td>
<td>0.9974</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9739</td>
<td>0.9983</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9846</td>
<td>0.9990</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9931</td>
<td>0.9996</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

$0.64 = \frac{80}{125}$

$0.96 \approx \frac{64}{67}$
Empirical Research Enablers: Reproducibility

- Stan: Open Source Bayesian Software
- Center for Open Science
- Berkeley Initiative for Transparent Social Science
- Institute for Quantitative Social Science
- DataCite, DataVerse, ICPSR, CNRI
- RunMyCode, ResearchCompendia
Administrative Data Projects

• Council of Professional Associations on Federal Statistics
• LinkedIn, EBay, Mint, etc.
• Software Carpentry, Jupyter Notebooks
• Open Corporates, Legal Entity Identifiers
2. Privacy Protecting Research

- Protocols can impose obfuscation at 3 stages: input, computation, or output.

<table>
<thead>
<tr>
<th></th>
<th>Input</th>
<th>Computation</th>
<th>Output</th>
<th>Protocol Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Open Data</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>X</td>
<td>Data Enclave</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Nondisclosure Agreement</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>X</td>
<td></td>
<td>Anonymization</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td></td>
<td></td>
<td>Randomized Response</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Multiparty Computation</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Fully Homomorphic Encryption</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>X</td>
<td>X</td>
<td>Differential Privacy</td>
</tr>
</tbody>
</table>
Differential Privacy

• A concept and procedure for allowing aggregate statistical queries while provably protecting individuals’ privacy (see Dwork).

• Require that the addition or removal of a single individual from the dataset should have nearly zero effect on any information released.

• I.e., you can’t learn anything new about individuals. So eliminates harm from participation, (not findings).

• Randomness gives privacy but cost accuracy.
On the Map
3. False Discovery

- Publish if $p = \Pr(\text{Data} \mid \text{Null Hypothesis}) < .05$
- Often test multiple hypotheses but pretend not.
- Sometimes called p-hacking (see Simonsohn) or hypothesis fishing. Test depends on data.
- Reuse of data is methodologically problematic.
- Overfitting on idiosyncratic observations.
Methodological Improvements

• Stan: Open Source Bayesian Software
• Expert Prior Probability Elicitation
• AEA Registry: Study Design & Analysis Plans
• Peer Review of Registered Reports
• Differentially Privacy for Data Exploration
Thresholdout

- See Dwork et. al. in Science (2015)
- Use training set to explore hypotheses
- Use DP protocols to test on holdout set
- Can then reuse the holdout set
- ML algorithms can overfit otherwise
- Get more robust and reproducible findings
• Basic Research
  Deep Carbon Observatory
  Microbiology of the Built Environment

• Economic Performance and the Quality of Life
  Economic Institutions, Behavior, and Performance
  Working Longer

• STEM Higher Education
  The Science of Learning
  Advancement for Underrepresented Groups

• Public Understanding of Science, Technology, & Economics
  Radio, Film, Television, Books, Theater, New Media

• Digital Information Technology:
  Data and Computational Research
  Scholarly Communication
  Universal Access to Knowledge

• Sloan Research Fellowships
• Civic Initiatives