Statistics

Taskmistress or Temptress?

Naomi Altman
Penn State University

Dec 2013
Recent Headlines

Lies, Damned Lies, and Medical Science

Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.
John Ioannidis, the man responsible for the fuss
Statistics as a set of methodologies in physical sciences probably begins in the 1800s.

- P-S. Laplace, C. F. Gauss and A-M. Legendre all worked on estimation errors in astronomy (around 1800).

- Sewall Wright, J. B. S. Haldane and R. A. Fisher are considered the founders of population genetics (around 1920).
### Advent of Academic Statistical Consulting Services

- Dept. of Biostatistics Johns Hopkins University 1918
- Rothamsted Experimental Station (now Research) 1919 - Fisher
- Harvard Medical Campus 1922
- Iowa State 1933
- Biometric Unit Cornell 1948
- ISI around 1950
- Many, many research programs now have Consulting Labs or equivalent.
Opinions differ on the role of statistics

- To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. *R.A. Fisher (father of statistics and quantitative genetics)*
Opinions differ on the role of statistics

- To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. *R.A. Fisher (father of statistics and quantitative genetics)*

- If your experiment needs statistics, you ought to have done a better experiment. *Ernest Rutherford (father of nuclear physics)*
Science and Statistics

Opinions differ on the role of statistics

- To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. *R.A. Fisher (father of statistics and quantitative genetics)*

- If your experiment needs statistics, you ought to have done a better experiment. *Ernest Rutherford (father of nuclear physics)*

- Do not put your faith in what statistics say until you have carefully considered what they do not say. *William W. Watt*
Who is studying statistics?

- At Penn State: 9500 took elementary statistics last year (out of 100,000).

- But: statistics is required only in the applied sciences, not bio, chemistry, physics, astronomy.

- Some disciplines have methodology courses but these vary in statistical thoughtfulness.
2005 - Study of Medical Journals

The study:

- *New England J. Medicine, JAMA, Lancet* and several high impact specialty journals.

- All papers with 1000 or more citations 1990-2003 investigated (49 papers).

- Comparison was with outcomes in studies or comparable or higher quality (sample size or design improvement).

- Note that medical research has a high level of input from statisticians compared to other disciplines.
## Ioannidis’ Result

### 2005 - Findings

- 4 papers reported negative results (contrary to previous studies of same intervention).
- 7 positive results were contradicted by subsequent studies.
- 7 positive results were stronger than found later.
- 20 positive results were replicated.
- 11 positive results were not followed up.
- Of the 6 nonrandomized trials, 5 were not replicated.
- Of the 39 randomized trials, 9 were not replicated.
- Less cited control studies had similar non-replicability results, but reported more negative results.
Ioannidis - The High Impact Paper

2005 - Why Most Published Research Findings are False

Ioannidis arguments:

- Bayes rule: When testing for rare events, false positives predominate the positives.
Ioannidis arguments:

- Bayes rule: When testing for rare events, false positives predominate the positives.
- Publication bias: Only interesting events are published: positive events are more interesting than negative.
Ioannidis arguments:

- Bayes rule: When testing for rare events, false positives predominate the positives.
- Publication bias: Only interesting events are published: positive events are more interesting than negative.
- Detection bias: Selective or distorted reporting, conflicts of interest, deliberate manipulation.
Ioannidis - The High Impact Paper

2005 - Why Most Published Research Findings are False

Ioannidis arguments:

- Bayes rule: When testing for rare events, false positives predominate the positives.

- Publication bias: Only interesting events are published: positive events are more interesting than negative.

- Detection bias: Selective or distorted reporting, conflicts of interest, deliberate manipulation

- Lack of independent replication
Ioannidis - The High Impact Paper

2005 - Why Most Published Research Findings are False

Ioannidis arguments:

- **Bayes rule**: When testing for rare events, false positives predominate the positives.
- **Publication bias**: Only interesting events are published: positive events are more interesting than negative.
- **Detection bias**: Selective or distorted reporting, conflicts of interest, deliberate manipulation
- **Lack of independent replication**
- **Selection of most significant events instead of proper evidence accumulation (meta-analysis) when there is replication**
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high.
- Small effect sizes.
- Large number of relationships tested without preliminary findings: lower prior probability of effect.
- High flexibility in designs, definitions, outcomes and analyzes: the search for significance.
- Rewards of research: winner takes all, so it pays to be first and to find something.
- Hot areas of research: research is rushed, lots of studies with low prior probability of effect.
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high
- Small effect sizes:
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high
- Small effect sizes:
- Large number of relationships tested without preliminary findings: lower prior probability of effect
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high
- Small effect sizes:
- Large number of relationships tested without preliminary findings: lower prior probability of effect
- High flexibility in designs, definitions, outcomes and analyzes: the search for significance
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- **Small studies**: if the false positive rate is controlled, the false negative rate is high
- **Small effect sizes**:
- **Large number of relationships tested without preliminary findings**: lower prior probability of effect
- **High flexibility in designs, definitions, outcomes and analyzes**: the search for significance
- **Rewards of research**: winner takes all, so it pays to be first and to find something
Ioannidis suggests several circumstances under which high false positive rates should be expected:

- Small studies: if the false positive rate is controlled, the false negative rate is high
- Small effect sizes:
- Large number of relationships tested without preliminary findings: lower prior probability of effect
- High flexibility in designs, definitions, outcomes and analyzes: the search for significance
- Rewards of research: winner takes all, so it pays to be first and to find something
- Hot areas of research: research is rushed, lots of studies with low prior probability of effect
Irreproducible results

Is the 0.05 level too lax?

Hayden (*Nature* November 2013) suggested that the problem is that \( p < 0.05 \) is not sufficiently stringent.

Not true: Consider the effects of multiple testing adjustments in bioinformatics (where we know the number of “events”).
Irreproducible results

Why don’t match our results match yours?

Let's assume there are 10 thousand features (genes, proteins).

- 2 careful labs replicate the same experiment.
- Both labs pick a sample size to achieve 80% power testing at level $\alpha = 0.05$.
- Both labs use the best possible methods and use FDR=0.05 to reject.

Shouldn’t they get the same results?
### Table: Results of testing

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.05$</th>
<th>FDR=0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance level</td>
<td>0.05</td>
<td>0.0015</td>
</tr>
<tr>
<td>Power</td>
<td>0.80</td>
<td>0.27</td>
</tr>
<tr>
<td>FDR</td>
<td>0.56</td>
<td>0.05</td>
</tr>
<tr>
<td>Total Discoveries</td>
<td>1250</td>
<td>283</td>
</tr>
<tr>
<td>Reproducible False Positives</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Reproducible True Positives</td>
<td>640</td>
<td>73</td>
</tr>
<tr>
<td>% Reproducible</td>
<td>52%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Suppose $\pi_0 = 0.90$. 

Matching results?
Of course, things are really more complex:

- Some of the features are correlated.
- Most studies are underpowered.
- Each feature has a different variance, so the power differs among the tests.
- There are systematic errors in the measurements so that some false positives and false negatives are more likely to reoccur.
Most of us are aware of the FDR (and also the False Negative Rate). But that is not the only problem. To see the problem, let’s consider some simulated data with $\pi_0 = 0.9$. I set 5% of the data to be 2.35-fold up and 5% 2.35-fold down (90% power with sample size 5).
Irreproducibility

But it is not so clear when the histogram is plotted proportional to the number of features.
The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Significant $p &lt; 0.05$</td>
<td>1303</td>
</tr>
<tr>
<td>False Discoveries</td>
<td>409</td>
</tr>
<tr>
<td>False Nondiscoveries</td>
<td>106</td>
</tr>
<tr>
<td>FDP</td>
<td>31%</td>
</tr>
</tbody>
</table>

Using the Benjamini and Hochberg method to obtain FDR = 0.05:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Significant FDR</td>
<td>120</td>
</tr>
<tr>
<td>False Discoveries</td>
<td>9</td>
</tr>
<tr>
<td>False Nondiscoveries</td>
<td>889</td>
</tr>
<tr>
<td>FDP</td>
<td>0.075%</td>
</tr>
</tbody>
</table>

The estimated mean "effect" is 2.91 although the actual effect size was 2.35.

Using effect size 2.02 (power = 0.8) there were no discoveries after BH adjustment.

Lessons

Without multiple testing adjustments, the FDR can be extremely high.

Multiple testing adjustments dramatically reduce power.

The effect size estimated from tests that survive after selection is dramatically biased up.
The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

<table>
<thead>
<tr>
<th>Number Significant $p &lt; 0.05$</th>
<th>1303</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Discoveries</td>
<td>409</td>
</tr>
<tr>
<td>False Nondiscoveries</td>
<td>106</td>
</tr>
<tr>
<td>FDP</td>
<td>31%</td>
</tr>
</tbody>
</table>

Screening with FDR

Using the Benjamini and Hochberg method to obtain FDR=0.05:

<table>
<thead>
<tr>
<th>Number Significant $FDR &lt; 0.05$</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Discoveries</td>
<td>9</td>
</tr>
<tr>
<td>False Nondiscoveries</td>
<td>889</td>
</tr>
<tr>
<td>FDP</td>
<td>0.075%</td>
</tr>
</tbody>
</table>

The estimated mean "effect" is 2.91 although the actual effect size was 2.35.
The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Significant $p &lt; 0.05$</td>
<td>1303</td>
</tr>
<tr>
<td>False discoveries</td>
<td>409</td>
</tr>
<tr>
<td>False nondiscoveries</td>
<td>106</td>
</tr>
<tr>
<td>FDP</td>
<td>31%</td>
</tr>
</tbody>
</table>

Screening with FDR

Using the Benjamini and Hochberg method to obtain FDR=0.05:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Significant $FDR &lt; 0.05$</td>
<td>120</td>
</tr>
<tr>
<td>False discoveries</td>
<td>9</td>
</tr>
<tr>
<td>False nondiscoveries</td>
<td>889</td>
</tr>
<tr>
<td>FDP</td>
<td>0.075%</td>
</tr>
</tbody>
</table>

The estimated mean "effect" is 2.91 although the actual effect size was 2.35.
Nature Publishing targets irreproducibility

- 2012 - A special series of articles on research reproducibility
  http://www.nature.com/nature/focus/reproducibility/index.html
- April 2013 - reporting checklist including study protocols and statistical methods
Nature Publishing targets irreproducibility

- 2012 - A special series of articles on research reproducibility
  http://www.nature.com/nature/focus/reproducibility/index.html

- April 2013 - reporting checklist including study protocols and statistical methods

- September 2013 - Points of Significance column launched by Nature Methods (Altman, Krzywinski (Evanko))

- September 2013 - statistical review panel assembled for all Nature Publishing journals
A tutorial column to encourage *statistical thinking*. 
Martin Krzywinski, Naomi Altman, Dan Evanko

**Topics to date:**

- Uncertainty, Populations and Sampling Distributions
- Error bars for the sample mean
- t-tests and p-values
- Power and Sample Size
- Boxplots (along with release of web tool to produce boxplots)

Topics selected by editorial team. Both frequentist and Bayesian methods will be featured.
**Points of Significance column**

**Error bars**

**T-tests**

**Power, sample size, multiple testing**

---

**Sample mean**

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.d.</td>
<td>0.0003</td>
<td>0.05</td>
</tr>
<tr>
<td>s.e.m.</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.005</td>
<td>0.05</td>
</tr>
</tbody>
</table>

**Repeated observations of expression**

- $\mu$, $\bar{x}$
- s.d., s.e.m.

**Distribution of expression values**

- $H_0$

**Distribution of average expression values**

- $\mu$, $\bar{x}$
Other Initiatives

Statistical Assistance for Underserved Groups

- Statisticians without Borders (ASA)
- LISA 2020 - Virginia Tech
Statistics Without Borders (SWB) is an Outreach Group of the American Statistical Association.

Comprised entirely of volunteers, we provide free statistical consulting to organizations and government agencies, particularly from developing nations, that do not have the resources for statistical services.

In support of non-partisan and secular activities, SWB promotes the use of statistics to improve the health and well-being of all people.

Our vision is to achieve and implement the best statistical practice in the service of others.
Laboratory of Interdisciplinary Statistical Analysis

Director: Eric Vance

- Initiative to create 20 regional statistical collaboration centers in developing countries.

- Train statistical collaborators in the US to go home and head up a center.

- Looking for high level buy-in (university president, head of government agency).
Doing a Better Job

- Train statistics students to handle real data.
- Change science curriculum to include statistics for everyone.
- Teach statistical *thinking* rather than methods and equations.
Doing a Better Job

- Train statistics students to handle real data.
- Change science curriculum to include statistics for everyone.
- Teach statistical *thinking* rather than methods and equations.

- Get involved in scientific journals:
  - tutorial level articles on methodology
  - *editorial process*
Doing a Better Job

- Train statistics students to handle real data.
- Change science curriculum to include statistics for everyone.
- Teach statistical *thinking* rather than methods and equations.

- Get involved in scientific journals:
  - tutorial level articles on methodology
  - *editorial process*

- Take consulting role seriously.
- Collaborate on and off campus.