Statistics for the Big Data Era

Emmanuel Candès

Goals for these lectures

1. Emphasize the importance of modern statistical inference problems

2. Discuss ways in which the statistical community responds to pressing contemporary problems
Agenda

Lecture 1: The big data era and the problem of selective inference

- Classical inference
- The reproducibility crisis
- Multiple testing and false discovery rate
- The Benjamini-Hochberg (BHq) procedure
- Interpretation of BHq
- Inference after selection
The Beginning of Statistical Inference
The problem of induction (Hume)
No matter how many instances of white swans we might have observed, this does not justify the conclusion that all swans are white.

Deductive methods of testing
An hypothesis can only be empirically tested, and only after it has been advanced; predictions are deduced from the theory... and compared with the results of practical applications and experiments... they can be falsified or corroborated.

K. Popper (1902-1994)
Operationalization

R. A. Fisher (1890-1962)
The lady tasting tea

“A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested.”

The null hypothesis

“Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.”

The p-value

“It is open to the experimenter to be more or less exacting in respect to the smallness of the probability he would require before he would be willing to admit that his observations have” disproved the null.
p-values

Probability of observing in the data a discrepancy from what is expected under the standing hypothesis as large as that we observe, when the standing (null) hypothesis is true

\[ f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \]
An example: the 1954 Salk vaccine trial

Randomized, controlled, double-blinded trial

750,000 children

Randomized

Control

Placebo

200,000 children
rate of polio
71 per 100,000

Refusals

350,000 children
rate of polio
46 per 100,000

Treatment

Refusals

200,000 children
rate of polio
28 per 100,000

Vaccine
The null hypothesis

<table>
<thead>
<tr>
<th>Group</th>
<th>Size</th>
<th>Rate per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>200,000</td>
<td>28</td>
</tr>
<tr>
<td>Control</td>
<td>200,000</td>
<td>71</td>
</tr>
</tbody>
</table>


null hypothesis: vaccine is ineffective and observed difference is due to chance

p-value: what is the chance (under the null) of observing a discrepancy as high as the one we have observed?
Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>56</td>
<td>199,964</td>
<td>200,000</td>
</tr>
<tr>
<td>Control</td>
<td>142</td>
<td>199,858</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>399,802</td>
<td>400,000</td>
</tr>
</tbody>
</table>

There are 400,000 balls: 198 are black and the others are white. Select 200,000 at random (make 2 groups). The p-value is the chance of having a group with at least 142 black balls. This chance is less than 1 in 1 billion!
Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>56</td>
<td>199,964</td>
<td>200,000</td>
</tr>
<tr>
<td>Control</td>
<td>142</td>
<td>199,858</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>399,802</td>
<td>400,000</td>
</tr>
</tbody>
</table>

There are 400,000 balls: 198 are black and the others are white.
Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>56</td>
<td>199,964</td>
<td>200,000</td>
</tr>
<tr>
<td>Control</td>
<td>142</td>
<td>199,858</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>399,802</td>
<td>400,000</td>
</tr>
</tbody>
</table>

- There are 400,000 balls: 198 are black and the others are white
- Select 200,000 at random (make 2 groups)
Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>56</td>
<td>199,964</td>
<td>200,000</td>
</tr>
<tr>
<td>Control</td>
<td>142</td>
<td>199,858</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>399,802</td>
<td>400,000</td>
</tr>
</tbody>
</table>

- There are 400,000 balls: 198 are black and the others are white
- Select 200,000 at random (make 2 groups)
- p-value is the chance of having a group with at least 142 black balls
Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>56</td>
<td>199,964</td>
<td>200,000</td>
</tr>
<tr>
<td>Control</td>
<td>142</td>
<td>199,858</td>
<td>200,000</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>399,802</td>
<td>400,000</td>
</tr>
</tbody>
</table>

- There are 400,000 balls: 198 are black and the others are white
- Select 200,000 at random (make 2 groups)
- p-value is the chance of having a group with at least 142 black balls

This chance is less than 1 in 1 billion!
Approximate p-value

- \( Y_0/Y_1 \) number of cases in control/treatment group
- \( Y_i \sim \text{Binomial}(n, p_i) \) \((n = 200,000)\)
- Under \( H_0 \), \( p_0 = p_1 = p \)
- p-value is chance of obs. a diff. \( Y_0 - Y_1 \) at least as large as what we've seen
Approximate p-value

- $Y_0/Y_1$ number of cases in control/treatment group
- $Y_i \sim \text{Binomial}(n, p_i)$ ($n = 200,000$)
- Under $H_0$, $p_0 = p_1 = p$
- p-value is chance of obs. a diff. $Y_0 - Y_1$ at least as large as what we've seen

Under $H_0$

\[
\begin{align*}
\mathbb{E}(Y_i) &= np \\
\text{Var}(Y_i) &= np(1 - p)
\end{align*}
\implies
\begin{align*}
\mathbb{E}(Y_0 - Y_1) &= 0 \\
\text{Var}(Y_0 - Y_1) &= 2np(1 - p)
\end{align*}
\]
Approximate p-value

- \( Y_0/Y_1 \) number of cases in control/treatment group
- \( Y_i \sim \text{Binomial}(n, p_i) \) \((n = 200,000)\)
- Under \( H_0, p_0 = p_1 = p \)
- p-value is chance of obs. a diff. \( Y_0 - Y_1 \) at least as large as what we’ve seen

Under \( H_0 \)

\[
\begin{align*}
\mathbb{E}(Y_i) &= np \\
\text{Var}(Y_i) &= np(1-p)
\end{align*}
\]

\( \implies \)

\[
\begin{align*}
\mathbb{E}(Y_0 - Y_1) &= 0 \\
\text{Var}(Y_0 - Y_1) &= 2np(1-p)
\end{align*}
\]

Central limit theorem

\[
\frac{Y_0 - Y_1}{\sqrt{2np(1-p)}} \approx \mathcal{N}(0, 1)
\]

\[
f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}
\]

A beautiful thing!
Approximate p-value

- $Y_0/Y_1$ number of cases in control/treatment group
- $Y_i \sim \text{Binomial}(n, p_i) \ (n = 200, 000)$
- Under $H_0$, $p_0 = p_1 = p$
- p-value is chance of obs. a diff. $Y_0 - Y_1$ at least as large as what we’ve seen

Under $H_0$

$$
\begin{align*}
\mathbb{E}(Y_i) &= np \\
\text{Var}(Y_i) &= np(1 - p) \\
\mathbb{E}(Y_0 - Y_1) &= 0 \\
\text{Var}(Y_0 - Y_1) &= 2np(1 - p)
\end{align*}
$$

Central limit theorem

$$
\frac{Y_0 - Y_1}{\sqrt{2np(1 - p)}} \ \sim \ \mathcal{N}(0, 1)
$$

Plug $\hat{p} = (Y_0 + Y_1)/2n$ above and use $t$-distribution instead
Distribution of p-value

Test statistic $T$ for one-sided test

$$P\text{-val} = \mathbb{P}_{H_0}(T \geq T_{obs} \mid T_{obs}) \quad \text{(this is a random variable)}$$

Under $H_0$

$$P\text{-val} \sim \text{Unif}(0, 1) \quad \text{(if } T \text{ is continuous)}$$
Has Something Gone Wrong?
The Economist

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

“I SEE a train wreck looming,” warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as “priming.” Priming studies suggest that decisions can be influenced by apparently irrelevant actions or events that took place just before the cusp of choice. They have been a boom area in psychology over the past decade, and some of their insights have already made it out of the lab and into the toolkits of policy wonks keen on “nudging” the populace.
Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Systematic attempts to replicate widely cited priming experiments have failed

- Amgen could only replicate 6 of 53 studies they considered landmarks in basic cancer science
- HealthCare could only replicate about 25% of 67 seminal studies

Early report (Kaplan, ’08): 50% of Phase III FDA studies ended in failure
“Significance chasing”

“Publication bias” (file drawer effect, Rosenthal ’79)

“Selective reporting” (“researcher’s degrees of freedom”)

“Why most published research findings are false” (Ioannidis, ’05)
Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

Modelling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, $\alpha$. Assuming that $c$ relationships are being probed in the field, the expected values of the $2 \times 2$ table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance.

It can be proven that most claimed research findings are false.
Since 1955, The Journal of Irreproducible Results has offered “spoofs, parodies, whimsies, burlesques, lampoons and satires” about life in the laboratory. Among its greatest hits: “Acoustic Oscillations in Jell-O, With and Without Fruit, Subjected to Varying Levels of Stress” and “Utilizing Infinite Loops to Compute an Approximate Value of Infinity.” The good-natured jibes are a backhanded celebration of science. What really goes on in the lab is, by implication, of a loftier, more serious nature.

It has been jarring to learn in recent years that a reproducible result may actually be the rarest of birds. Replication, the ability of another lab to reproduce a finding, is the gold standard of science, reassurance that you have discovered something true. But that is getting harder all the time. With the most accessible truths already discovered, what remains are often subtle effects, some so delicate that they can be conjured up only under ideal circumstances, using highly specialized techniques.
Personal and societal concern

Great danger in seeing erosion of public confidence in science
Great danger in seeing erosion of public confidence in science

Seems like scientific community is beginning to respond
Reproducibility Initiative

http://validation.scienceexchange.com/
Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the *Nature* research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters that can introduce bias or influence robustness, and provide precise characterization of key reagents that may be subject to biological variability, such as cell lines and antibodies. The checklist also consolidates existing policies about data deposition and presentation.

We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editor’s discretion and at the referees’ suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require the commitment of funders, institutions, researchers and publishers. It is encouraging that NIH institutes have led community discussions on this topic and are considering their own recommendations. We urge others to take note of these and of our initiatives, and do whatever they can to improve research reproducibility.
NAS President’s address: April 27, 2015
Big data and a new scientific paradigm

Collect data first $\implies$ Ask questions later

- Large data sets available prior to formulation of hypotheses
- Need to quantify “reliability” of hypotheses generated by data snooping

Very different from hypothesis-driven research
Big data and a new scientific paradigm

Collect data first  $\implies$  Ask questions later

- Large data sets available prior to formulation of hypotheses
- Need to quantify “reliability” of hypotheses generated by data snooping

Very different from hypothesis-driven research

What does statistics have to offer?

- Account for “look everywhere” effect
- Understand reliability in the context of all hypotheses that have been explored
Most discoveries may be false: Sorić ('89)

1000 hypotheses to test
Most discoveries may be false: Sorić ('89)

1000 hypotheses, 100 potential discoveries
Most discoveries may be false: Sorić ('89)

1000 hypotheses, 100 potential discoveries
Most discoveries may be false: Sorić ('89)

Power $\approx 80\% \implies$ true positives $\approx 80$
False positives (5\% level) $\approx 45$  \implies  \text{False discovery rate} $\approx 36\%$
Most discoveries may be false: Sorić (’89)

Power $\approx 80\% \implies$ true positives $\approx 80$
False positives (5% level) $\approx 45$ $\implies$ False discovery rate $\approx 36\%$
Most discoveries may be false: Sorić ('89)

Power $\approx 30\% \implies$ False discovery rate $\approx 60\%$

More false negatives than true positives!
Example: meta-analysis in neuroscience

Button et al. (2013) *Power failure: why small sample size undermines the reliability of neuroscience*
The False Discovery Rate and the Benjamini-Hochberg Procedure
False Discovery Rate (FDR): Benjamini & Hochberg ('95)

\( H_1, \ldots, H_n \) hypotheses subject to some testing procedure

\[
\text{FDR} = \mathbb{E} \left[ \frac{\text{#false discoveries}}{\text{#discoveries}} \right] \quad \text{‘0/0 = 0’}
\]

- Natural type I error
- Under independence (and PRDS) simple rules control FDR (BHq)
- Widely used → enormous influence on medical research
FDR control with BHq (under independence)

FDR: expected proportion of false discoveries

- Sorted \( p \)-values: \( p_1 \leq p_2 \leq \ldots \leq p_n \) (from most to least significant)
- Target FDR \( q \)

![Sorted p-values and line iq/n](image)

The cut-off is adaptive to number of non-nulls.
FDR control with BHq (under independence)

FDR: expected proportion of false discoveries

- Sorted $p$-values: $p_1 \leq p_2 \leq \ldots \leq p_n$ (from most to least significant)
- Target FDR $q$

The cut-off is adaptive to number of non-nulls
Earlier work on multiple comparisons

Henry Scheffe
John Tukey

Westfall and Young (’93)

Rupert Miller
Formal definition of FDR

Table: Outcomes of a multiple testing problem

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U$</td>
<td>$V$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T$</td>
<td>$S$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R$</td>
<td>$R$</td>
<td>$n$</td>
</tr>
</tbody>
</table>
Formal definition of FDR

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U$</td>
<td>$V$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T$</td>
<td>$S$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R$</td>
<td>$R$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

**Table:** Outcomes of a multiple testing problem

- **Family wise error rate:** $\text{FWER} = P(V \geq 1)$
- **False discovery proportion:**
  
  $$\text{FDP} = \frac{V}{\max(R, 1)} = \begin{cases} 
  \frac{V}{R} & \text{if } R \geq 1 \\
  0 & \text{otherwise}
  \end{cases}$$

- **False discovery rate:**
  
  $$\text{FDR} = E[\text{FDP}]$$
The BHq step-up procedure and FDR control

- Sort \( p \)-values \( p(1) \leq p(2) \leq \cdots \leq p(n) \)
- Fix \( q \in [0, 1] \) and find
  \[
  \hat{i} = \max\{i : p(i) \leq q_i/n\}
  \]
- Reject hypotheses with \( p_j \leq p(\hat{i}) \)
  (make no rejection if \( \{i : p(i) \leq q_i/n\} \) is empty)
The BHq step-up procedure and FDR control

- Sort $p$-values $p(1) \leq p(2) \leq \ldots \leq p(n)$
- Fix $q \in [0, 1]$ and find
  $$\hat{i} = \max\{i : p(i) \leq qi/n\}$$
- Reject hypotheses with $p_j \leq p(\hat{i})$
  (make no rejection if $\{i : p(i) \leq qi/n\}$ is empty)

**Theorem (Benjamini-Hochberg '95)**

For independent test statistics $[p$-values$]$, BHq controls the FDR at level $q$:

$$\text{FDR} = \frac{n_0}{n} q \leq q$$
Proof

\[
FDR = \mathbb{E} \left[ \frac{\sum_{i \in \mathcal{H}_0} 1\{H_i \text{ rejected}\}}{1 \lor R} \right]
\]
Proof

\[ \text{FDR} = \mathbb{E} \left[ \frac{\sum_{i \in \mathcal{H}_0} 1\{H_i \text{ rejected}\}}{1 \lor R} \right] \]

Setting \( V_i = 1\{H_i \text{ rejected}\} \), for all \( i \in \mathcal{H}_0 \)

\[ \frac{V_i}{1 \lor R} = \sum_{k=1}^{n} \frac{V_i 1\{R = k\}}{k} \]
Proof

\[ \text{FDR} = \mathbb{E} \left[ \frac{\sum_{i \in \mathcal{H}_0} 1\{H_i \text{ rejected}\}}{1 \lor R} \right] \]

Setting \( V_i = 1\{H_i \text{ rejected}\} \), for all \( i \in \mathcal{H}_0 \)

\[ \frac{V_i}{1 \lor R} = \sum_{k=1}^{n} \frac{V_i 1\{R = k\}}{k} \]

Let \( R(p_{i \rightarrow 0}) \) be \# rejections BHq commits by setting \( i \)th \( p \)-value to 0

\[ R(p_{i \rightarrow 0}) = R(p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots, p_n) \]
Proof

\[
\text{FDR} = \mathbb{E} \left[ \sum_{i \in \mathcal{H}_0} \frac{1 \{ H_i \text{ rejected} \}}{1 \lor R} \right]
\]

Setting \( V_i = 1 \{ H_i \text{ rejected} \} \), for all \( i \in \mathcal{H}_0 \)

\[
\frac{V_i}{1 \lor R} = \sum_{k=1}^{n} \frac{V_i 1 \{ R = k \}}{k}
\]

Let \( R(p_{i \rightarrow 0}) \) be \# rejections BHq commits by setting \( i \)th \( p \)-value to 0

\[
R(p_{i \rightarrow 0}) = R(p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots, p_n)
\]

Claim:

\[
V_i 1 \{ R(p) = k \}
\]
\[
= V_i 1 \{ R(p_{i \rightarrow 0}) = k \}
\]
Proof

\[ \text{FDR} = \mathbb{E} \left[ \sum_{i \in \mathcal{H}_0} \frac{1\{H_i \text{ rejected}\}}{1 \lor R} \right] \]

Setting \( V_i = 1\{H_i \text{ rejected}\} \), for all \( i \in \mathcal{H}_0 \)

\[ \frac{V_i}{1 \lor R} = \sum_{k=1}^{n} \frac{V_i \{R = k\}}{k} \]

Let \( R(p_i \rightarrow 0) \) be \# rejections BHq commits by setting \( i \)th \( p \)-value to 0

\[ R(p_i \rightarrow 0) = R(p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots, p_n) \]

Claim:

\[ V_i 1\{R(p) = k\} = V_i 1\{R(p_i \rightarrow 0) = k\} \]

Illustration of \( R(p_i \rightarrow 0) \)
\[
\frac{V_i}{1 \land R} = \sum_{k=1}^{n} \frac{V_i \{ R(p_i \to 0) = k \} \cdot \sum_{k=1}^{n} 1\{p_i \leq qk/n\} 1\{R(p_i \to 0) = k\}}{k}
\]
\[
\frac{V_i}{1 \lor R} = \sum_{k=1}^{n} \frac{V_i 1\{R(p_i \to 0) = k\}}{k} \\
= \sum_{k=1}^{n} \frac{1\{p_i \leq qk/n\} 1\{R(p_i \to 0) = k\}}{k}
\]

With \( F_i = \{p_1, \ldots, p_i-1, p_{i+1}, \ldots, p_n\} \)

\[
\mathbb{E}\left[ \frac{V_i}{1 \lor R} \mid F_i \right] = \sum_{k=1}^{n} \frac{q}{n} 1\{R(p_i \to 0) = k\} = \frac{q}{n}
\]
\[
\frac{V_i}{1 \vee R} = \sum_{k=1}^{n} \frac{V_i 1\{R(p_i \to 0) = k\}}{k}
\]

= \sum_{k=1}^{n} \frac{1\{p_i \leq qk/n\} 1\{R(p_i \to 0) = k\}}{k}

With \( \mathcal{F}_i = \{p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n\} \)

\[
\mathbb{E} \left[ \frac{V_i}{1 \vee R} \mid \mathcal{F}_i \right] = \sum_{k=1}^{n} \frac{q}{n} 1\{R(p_i \to 0) = k\} = \frac{q}{n}
\]

Hence

\[
\text{FDR} = \sum_{i \in \mathcal{H}_0} \mathbb{E} \left[ \frac{V_i}{1 \vee R} \right] = \frac{qn_0}{n}
\]
Empirical process viewpoint of BH(q)

Empirical CDF of p-values

\[ \hat{F}(t) = \# \{ i : p_i \leq t \} / n \]

\[ \alpha i / n \]

\[ t_0 / n \]

\[ t / \alpha \]
Empirical process viewpoint of BH(q)

Empirical CDF of p-values

\[ \hat{F}(t) = \frac{\#\{i : p_i \leq t\}}{n} \]
Empirical process viewpoint of BH(q)

Empirical CDF of p-values

\[ \hat{F}(t) = \frac{\# \{ i : p_i \leq t \}}{n} \]

BH threshold

\[ \tau = \max \left\{ p(i) : p(i) \leq q \frac{i}{n} \right\} = \max \left\{ p(i) : p(i) \leq q \hat{F}(p(i)) \right\} \]

\[ = \max \left\{ t \in \{ p_1, \ldots, p_n \} : t/\hat{F}(t) \leq q \right\} \]

with \( \tau = q/n \) if set above is empty
Empirical process viewpoint of BH(q)

Empirical CDF of p-values

\[ \hat{F}(t) = \frac{\# \{ i : p_i \leq t \}}{n} \]

BH rejects all hypotheses with \( p_i \leq \tau \)

\[ \tau = \max \left\{ t : \frac{t}{\hat{F}(t) \lor 1/n} \leq q \right\} \]
Estimating FDP

Table: Outcomes if rejecting for $p_i \leq t$

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U(t)$</td>
<td>$V(t)$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T(t)$</td>
<td>$S(t)$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R(t)$</td>
<td>$R(t)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

$FDP(t) = \max(R(t), 1)$

$FDR(t) = E[FDP(t)]$
### Estimating FDP

#### Table: Outcomes if rejecting for $p_i \leq t$

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U(t)$</td>
<td>$V(t)$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T(t)$</td>
<td>$S(t)$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R(t)$</td>
<td>$R(t)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

$$FDP(t) = \frac{V(t)}{\max(R(t), 1)}$$

$$FDR(t) = \mathbb{E}[FDP(t)]$$
Estimating FDP

Table: Outcomes if rejecting for \( p_i \leq t \)

<table>
<thead>
<tr>
<th>( H_0 ) true</th>
<th>( H_0 ) accepted</th>
<th>( H_0 ) rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) true</td>
<td>( U(t) )</td>
<td>( V(t) )</td>
<td>( n_0 )</td>
</tr>
<tr>
<td>( H_0 ) false</td>
<td>( T(t) )</td>
<td>( S(t) )</td>
<td>( n - n_0 )</td>
</tr>
<tr>
<td></td>
<td>( n - R(t) )</td>
<td>( R(t) )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

\[
FDP(t) = \frac{V(t)}{\max(R(t), 1)}
\]

\[
FDR(t) = \mathbb{E}[FDP(t)]
\]

Q? How do we estimate FDP(\( t \)?)
Estimating FDP

Table: Outcomes if rejecting for $p_i \leq t$

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U(t)$</td>
<td>$V(t)$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T(t)$</td>
<td>$S(t)$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R(t)$</td>
<td>$R(t)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

$$FDP(t) = \frac{V(t)}{\max(R(t), 1)}$$

$$FDR(t) = \mathbb{E}[FDP(t)]$$

Q? How do we estimate FDP($t$)?

$$\mathbb{E} V(t) = n_0 t \leq nt$$
### Estimating FDP

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ accepted</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$U(t)$</td>
<td>$V(t)$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$T(t)$</td>
<td>$S(t)$</td>
<td>$n - n_0$</td>
</tr>
<tr>
<td></td>
<td>$n - R(t)$</td>
<td>$R(t)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

**FDP**

\[
FDP(t) = \frac{V(t)}{\max(R(t), 1)}
\]

**FDR**

\[
FDR(t) = \mathbb{E}[FDP(t)]
\]

Q? How do we estimate $FDP(t)$?

\[
\mathbb{E}V(t) = n_0t \leq nt \implies \widehat{FDP}(t) = \frac{nt}{\max(R(t), 1)}
\]
This is BHq!

\[
\hat{\text{FDP}}(t) = \frac{nt}{\max(R(t), 1)} = \frac{t}{\max(\hat{F}(t), 1/n)}
\]
This is BHq!

\[
\hat{\text{FDP}}(t) = \frac{nt}{\max(R(t), 1)} = \frac{t}{\max(\hat{F}(t), 1/n)}
\]

Being as liberal as possible while maintaining estimate below target is BHq!

\[
\max \left\{ t : \hat{\text{FDP}}(t) \leq q \right\} = \tau_{BH}
\]
This is BHq!

\[
\hat{\text{FDP}}(t) = \frac{nt}{\max(R(t), 1)} = \frac{t}{\max(\hat{F}(t), 1/n)}
\]

Being as liberal as possible while maintaining estimate below target is BHq!

\[
\max \left\{ t : \hat{\text{FDP}}(t) \leq q \right\} = \tau_{\text{BH}}
\]

- BH theorem says we can invert estimate to control FDR
- Leads to new proof of FDR control via martingale theory
From FDR to Selective Inference
From FDR to selective inference

\[
FDR = \mathbb{E} \left[ \frac{\#\text{false discoveries}}{\#\text{discoveries}} \right] \quad \text{Type I error averaged over selected discoveries}
\]
From FDR to selective inference

\[
\text{FDR} = \mathbb{E} \left[ \frac{\#\text{false discoveries}}{\#\text{discoveries}} \right]
\]

Type I error averaged over selected discoveries

Selective inference
Wish that original property of inference holds on the average over the selected

E.g. for subset of parameters that happen to be of interest after viewing the data
Confidence intervals

Data $Y$ about parameter $\theta$

**Confidence interval**

A $95\%$ (resp. $100(1 - \alpha)\%$) CI is an interval $CI = [\theta_{lo}(Y), \theta_{hi}(Y)]$ s. t.

\[
P(CI \ni \theta) \geq 0.95 \quad \text{(resp. } 1 - \alpha)\]

Example: Salk trial

<table>
<thead>
<tr>
<th>Group</th>
<th>Size</th>
<th>Rate per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>200,000</td>
<td>28</td>
</tr>
<tr>
<td>Control</td>
<td>200,000</td>
<td>71</td>
</tr>
</tbody>
</table>

Incidence rate after vaccination $p_1$

Incidence rate without vaccination $p_0$

May wish confidence interval about $p_0 - p_1$, $p_0/p_1$, ...
Confidence intervals

Data $Y$ about parameter $\theta$

Confidence interval

A 95% (resp. 100(1 − $\alpha$)%) CI is an interval $\text{CI} = [\theta_{lo}(Y), \theta_{hi}(Y)]$ s. t.

$$\mathbb{P}(\text{CI} \ni \theta) \geq 0.95 \quad (\text{resp. } 1 - \alpha)$$

Example: Salk trial

<table>
<thead>
<tr>
<th>Group</th>
<th>Size</th>
<th>Rate per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>200,000</td>
<td>28</td>
</tr>
<tr>
<td>Control</td>
<td>200,000</td>
<td>71</td>
</tr>
</tbody>
</table>

- Incidence rate after vaccination $p_1$
- Incidence rate without vaccination $p_0$

May wish confidence interval about $p_0 - p_1$, $p_0/p_1$, ...
Multiple parameters $\theta_1, \ldots, \theta_n$

Regular 95% CI

\[ \mathbb{E} \left( 1(\text{Cl}_i \text{ covers } \theta_i) \right) \geq 0.95 \implies \mathbb{E} \left[ \frac{\sum_{i=1}^{n} 1(\text{Cl}_i \text{ covers } \theta_i)}{n} \right] \geq 0.95 \]

Expected proportion is all right
After selection…

“In a large number of 95% confidence intervals, 95% of them contain the population parameter […] but it would be wrong to imagine that the same rule also applies to a large number of 95% interesting confidence intervals”

Sorić (’89)
After selection...

“In a large number of 95% confidence intervals, 95% of them contain the population parameter [...] but it would be wrong to imagine that the same rule also applies to a large number of 95% interesting confidence intervals”

Sorić ('89)

90% CI
- 17/20 over all cover
- 1/4 over selected cover!

Will tend to fail when replicated

Errors of Type S & M (Gelman & Tuerlinckx, '00)
Ignoring multiplicity of intervals is more common than ignoring multiplicity of tests

- A CI is more informative than just hypothesis testing, so we are already doing better science...

- The CIs give the right coverage on average: for 95% CI’s

\[ \mathbb{E} \frac{\text{# covering CIs}}{\text{# estimated}} \geq 0.95 \]
Observations from Y. Benjamini

Ignoring multiplicity of intervals is more common than ignoring multiplicity of tests

- A CI is more informative than just hypothesis testing, so we are already doing better science...

- The CIs give the right coverage on average: for 95% CI’s

\[
\mathbb{E} \left\{ \frac{\# \text{ covering CIs}}{\# \text{ estimated}} \right\} \geq 0.95
\]

Anecdotes: \( p \)-values have a bad rep and many journals ban the use of \( p \)-values but not of CIs...
Ignoring selection when reporting confidence intervals

From Benjamini ('11)
For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white. The results were "plain as day," recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The P value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the P value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame.

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the P value, which is neither as reliable nor as objective as most scientists assume. "P values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

P values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny. One researcher suggested rechristening the methodology "statistical hypothesis inference testing," presumably for the acronym it would yield.

The irony is that when UK statistician Ronald Fisher introduced the P value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.
The Basic and Applied Social Psychology (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

**Question 1.** Will manuscripts with p-values be desk rejected automatically?

**Answer to Question 1.** No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about 'significant' differences or lack thereof, and so on).

**Question 2.** What about other types of inferential statistics such as confidence intervals or Bayesian methods?

**Answer to Question 2.** Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP. In the NHSTP, the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding. Regarding confidence intervals, the problem is that, for example, a 95% confidence interval does not indicate that the parameter of interest has a 95% probability of being within the interval. Rather, it means merely that if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter. Analogous to how the NHSTP fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been
False Coverage Rate (FCR): Benjamini & Yekutieli (’05)

- Collect data about \((\theta_1, \ldots, \theta_n)\)
- Then select subset \(S \subset \{1, \ldots, n\}\)

\[
FCR = \mathbb{E} \left[ \frac{\#\{i \in S : \text{Cl}_i \text{ does not cover } \theta_i\}}{|S|} \right] \quad '0/0 = 0'
\]

Coverage is averaged over the selected!
Collect data about \((\theta_1, \ldots, \theta_n)\)

Then select subset \(S \subset \{1, \ldots, n\}\)

\[
\text{FCR} = \mathbb{E} \left[ \frac{\# \{ i \in S : \text{Cl}_i \text{ does not cover } \theta_i \} }{|S|} \right] \\
\text{‘0/0 = 0’}
\]

Coverage is averaged over the selected!

Q? Can we control the FCR?
Collect data about \((\theta_1, \ldots, \theta_n)\)

Then select subset \(S \subset \{1, \ldots, n\}\)

\[
\text{FCR} = \mathbb{E} \left[ \frac{\#\{i \in S : \text{Cl}_i \text{ does not cover } \theta_i\}}{|S|} \right]
\]

‘0/0 = 0’

Coverage is averaged over the selected!

Q? Can we control the FCR?

A? Yes (Lecture 3)
The decline effect and the scientific method: The New Yorker

September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered

Selective inference: want inference to hold when selecting promising leads

- Addresses the reproducibility issue (at least partially)
- Statisticians extraordinarily engaged
- Lots of activity
- Needs even much more attention