Selective Inference and the False Discovery Rate

Yoav Benjamini
Tel Aviv University

Summer School – Ohrid, Macedonia

Supported by European Research Council grant: PSARPS
www.replicability.tau.ac.il
and by the European Human Project
Outline

1. Simultaneous and Selective inference
2. Testing with FDR control
3. False Coverage Rate
4. Estimation and Model Selection
5. More complex families
Saso’s first slide at the opening talk was
Data Mining: Prediction
The Statistical point of view: Prediction and Inference

Inference:
How close the model is the true always-unknown one.
Is it real? tests
How big? Estimate size
How far from the true value? Confidence intervals
Steel and Torrie (1960) bring from Erdman (1946):
6 groups of red clover plants, each inoculated with a different strain of Rhizobium bacteria.
5 measurements of Nitrogen content on each group (the standard textbook/manuals example)

\[ Y_{i+} \sim N(\mu, \sigma^2/5) \quad i=1,2,\ldots,6; \]

Interest in comparing strain effects
The first data-mining problem in Statistics

• Estimates
  \[ Y_{i+} - Y_{j+} \]

• Test the significance of the difference, with \( H_0: \mu_i = \mu_j \)
  via two-sample normal tests or t-tests

• Can do it by p-values
  \[ P-value = \text{Prob}_{H_0} \left( |Z| > \frac{|Y_{(i+)} - Y_{(j+)}}{|\sigma_{\text{diff}}}| \right) \]
  under \( H_0 \) \( P-value \sim U(0,1) \).

• To reject \( H_0 \) with the probability of type I error \( \leq \alpha \)
  (make a discovery with prob. to make a false discovery \( \leq \alpha \))

  Reject if \( P-value \leq \alpha \).
The first data-mining problem in Statistics

- Suppose we select the most promising groups’ difference
  \[ Y_{(k+)} - Y_{(1+)} \]
- With the $6 \times 5/2 = 15$ such tests, each at level $\alpha$
  \[ \text{Prob} ( Z > \frac{|Y_{(k+)} - Y_{(1+)}|}{\sigma_{\text{diff}}} ) < \alpha \]
even if there is no difference. The larger $k$ the worse it gets!
- In fact going back to the original paper we found 13 such groups resulting in $13 \times 12/2 = 78$ pairwise comparisons. With the limiting computing power of the 40s a large scale inference problem was encountered.

The multiple comparisons problem (procedures) MCP
Scientists study rumours: Eating coloured jelly beans causes acne.
"Unusual secrets are hidden in numbers. For instance, an orange car is less likely to have serious damages that are discovered only after the purchase...."
Giovanni and others (95) examined the possible effect of excess eating of 130 different kinds of foods on prostate cancer. 3 kinds of foods cleared the statistical significance bar – these are the only ones reported in the article’s abstract.

Eat ketchup and pizza to prevent prostate cancer

In the article itself all 130 results are reported but the abstract is usually the only information that passes on to the public – even to the professionals.

Selection by the abstract phenomenon
In the meanwhile the paper was cited over 1000 times. Dozens of studies about the contribution of tomatoes to the healing of different types of cancers with unclear results. A recent study, claims the secret is in the Oregano.
Some notations before we continue

1. The null hypotheses tested: $H_1, H_2, \ldots, H_m$.

$m_0$ of the $m$ hypotheses tested are true, we do not know which ones are true or even their number.

2. The result of any testing procedure is $R_i, i=1,2,\ldots,m$:
   
   - $R_i = 1$ if $H_i$ is rejected;
   - $R_i = 0$ if not.

   Let $V_i = 1$ if $R_i = 1$ but $H_i$ is true (a type I error was made);
   - $V_i = 0$ otherwise.

3. $R = \sum R_i$ # hypotheses rejected;
   - $V = \sum V_i$ # hypotheses rejected in error.

So, e.g.

$$\text{weak FWER} \leq \Pr_{H_0}(V \leq 1).$$
FWER Protection

• FamilyWise Error-Rate

For any configuration of true and null hypotheses

\[ FWER = \text{Prob}(V \geq 1) \]

Thus by assuring \( FWER \leq \alpha \), the probability of making even one type I error in the family, is controlled at level \( \alpha \):

Simultaneous Inference: all inference made are jointly correct up to the pre-specified error
Same for Confidence Intervals

Estimate \( m \) parameters by a confidence interval for each.

Define

\( V = \# \text{ of intervals failing to cover their respective parameter} \).

If for any configuration of parameters

\[
FWER = \text{Prob}(V \geq 1) \leq \alpha
\]

the set of such intervals is said to offer

Simultaneous Coverage at level \( 1-\alpha \)
Old and trusted solutions

If we test each hypothesis separately at level $\alpha_{\text{BON}}$

$$E(V) = E(\sum V_i) = \sum E(V_i) \leq m_0 \alpha_{\text{BON}} \leq m \alpha_{\text{BON}}$$

So to assure $E(V) \leq \alpha$ we may use $\alpha_{\text{BON}} = \alpha / m$

(Is any condition needed? )
This is

(1) The Bonferroni simultaneous inference procedure

that controls any configuration of hypotheses

$$\text{Expected number of errors } E(V) \leq \alpha$$
(2) **Tukey’s procedure** for pairwise comparisons:
Utilizes dependency by calculating the distribution of the studentized range statistics \( \frac{Y_{(k+)} - Y_{(1+)}}{s/n^{1/2}} \),

Known as **post-hoc analysis**
(3) Holm’s step-down procedure:

• Let $P_i$ be the observed p-value of the test for $H_i$

• Order the p-values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$

• If $P_{(1)} \leq \alpha/m$ Reject $H_{(1)}$

• If $P_{(2)} \leq \alpha/(m-1)$ Reject $H_{(2)}$

  ... 

• Until for the first time $P_{(k)} > \alpha/(m+1-k)$

• Then stop and reject no more.

Always: $FWER \leq \alpha$
## Significance of 8 Strain differences

<table>
<thead>
<tr>
<th>Behavioral Endpoint</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Lingering Time</td>
<td>0.0029</td>
</tr>
<tr>
<td># Progression segments</td>
<td>0.0068</td>
</tr>
<tr>
<td>Median Turn Radius (scaled)</td>
<td>0.0092</td>
</tr>
<tr>
<td>Time away from wall</td>
<td>0.0108</td>
</tr>
<tr>
<td>Distance traveled</td>
<td>0.0144</td>
</tr>
<tr>
<td>Acceleration</td>
<td>0.0146</td>
</tr>
<tr>
<td># Excursions</td>
<td>0.0178</td>
</tr>
<tr>
<td>Time to half max speed</td>
<td>0.0204</td>
</tr>
<tr>
<td>Max speed wall segments</td>
<td>0.0257</td>
</tr>
<tr>
<td>Median Turn rate</td>
<td>0.0320</td>
</tr>
<tr>
<td>Spatial spread</td>
<td>0.0388</td>
</tr>
<tr>
<td>Lingering mean speed</td>
<td>0.0588</td>
</tr>
<tr>
<td>Homebase occupancy</td>
<td>0.0712</td>
</tr>
<tr>
<td># stops per excursion</td>
<td>0.1202</td>
</tr>
<tr>
<td>Stop diversity</td>
<td>0.1489</td>
</tr>
<tr>
<td>Length of progression segments</td>
<td>0.5150</td>
</tr>
<tr>
<td>Activity decrease</td>
<td>0.8875</td>
</tr>
</tbody>
</table>

Bonferroni: \( \frac{.05}{17} = .0029 \)

Unadjusted
Unadjusted vs Simultaneous

In the search for food affecting Prostate Cancer,

3 food intakes were reducing with unadjusted significance with Bonferroni.
The increasing scale: Voxelwise Genome-Wise Association study

(Stein et al.’10)

- Goal: determine biological markers of Alzheimer’s disease by testing for associations between volume changes at voxels with genotype
Voxels searched: 32,000

SNPs:
- rs2132683
- rs713155
- rs476463
- rs2429582
- rs24990343

Number of tests: ~13,000,000,000

1 Voxels searched
1 SNPs
1 number of tests
A common feature of the larger applications

In these large problems:

• The selected are presented, highlighted, discussed. Their strength is displayed (p-values) The effect estimated

• Those inferences that are not selected are simply ignored: There are so many of them that even their identities are not reported, needless to say further details about the results of the inference for each
The increasing scale changes the goal

Tukey (1978): one should always control the FWER


\[ \text{# of comparisons} = \frac{35 \times (35-1)}{2} = 595 \]

There was a debate how to report results: with pairwise adjustment or without.

Their solution

Use the False Discovery Rate (FDR) approach
Outline

1. Simultaneous and Selective inference
2. Testing with FDR control
3. False Coverage Rate
4. Estimation and Model Selection
5. More complex families
The False Discovery Rate (FDR) criterion

Benjamini and Hochberg (89, 95)

\[ R = \# \text{ rejected hypotheses} = \# \text{ discoveries} \]

\[ V \text{ of these may be in error} = \# \text{ false discoveries} \]

The error (type I) in the entire study is measured by

\[ Q = \frac{V}{R} \quad R > 0 \]

\[ = 0 \quad R = 0 \]

i.e. the proportion of false discoveries among the discoveries (0 if none found)

\[ \text{FDR} = E(Q) \]

Does it make sense?
Does it make sense?

- Inspecting 100 features:
  
  2 false ones among 50 discovered - *bearable*
  
  2 false ones among 4 discovered - *unbearable*

  So this error rate is adaptive

- The same argument holds when inspecting 10,000

  So this error rate is scalable

- If nothing is “real” controlling the FDR at level $q$

  guarantees

  \[ \text{Prob}( V \geq 1 ) = E( V/R ) = FDR \leq q \]

- But otherwise

  \[ \text{Prob}( V \geq 1 ) \geq FDR \]

  So there is room for improving detection power
Reflections on goals

- Simultaneous inference: inference should hold jointly for all parameters in the family, and therefore jointly for any sub-family

- Selective inference: Inference should hold for the selected parameters the same way it holds for each parameter separately

“on the average over the selected”
• Instead of ignoring multiplicity, which still offers ‘control’ on the average,

\[ E(\frac{V}{\text{number of tests performed}}) = E(\frac{V}{m}) \leq \alpha \]

• FDR control assures

\[ E(\frac{V}{\text{number of tests selected}}) = E(\frac{V}{R}) \leq \alpha \]

• The above is hindsight. Our original motivation was a paper by Soric (’89) arguing that “most research discoveries might be false” when using 0.05 level testing.

• (See Ioannidis ’05 famous paper)
**FDR controlling procedures**

The BH (Linear Step-up ) procedure:

Let $P_i$ be the observed p-value of the test for $H_i$

- Order the p-values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$
- Let

  $$k = \max \{i : p(i) \leq (i/m)q\}$$

- Reject

  $$H_{(1)}, H_{(2)}, \ldots, H_{(k)}$$
# Significance of 8 Strain

<table>
<thead>
<tr>
<th>Behavioral Endpoint</th>
<th>Mixed</th>
<th>Linear StepUp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Lingering Time</td>
<td>0.0029</td>
<td>0.0029 = .05(1/17)</td>
</tr>
<tr>
<td># Progression segments</td>
<td>0.0068</td>
<td></td>
</tr>
<tr>
<td>Median Turn Radius (scaled)</td>
<td>0.0092</td>
<td></td>
</tr>
<tr>
<td>Time away from wall</td>
<td>0.0108</td>
<td></td>
</tr>
<tr>
<td>Distance traveled</td>
<td>0.0144</td>
<td></td>
</tr>
<tr>
<td>Acceleration</td>
<td>0.0146</td>
<td></td>
</tr>
<tr>
<td># Excursions</td>
<td>0.0178</td>
<td></td>
</tr>
<tr>
<td>Time to half max speed</td>
<td>0.0204</td>
<td></td>
</tr>
<tr>
<td>Max speed wall segments</td>
<td>0.0257</td>
<td></td>
</tr>
<tr>
<td>Median Turn rate</td>
<td>0.0320</td>
<td></td>
</tr>
<tr>
<td>Spatial spread</td>
<td>0.0388</td>
<td></td>
</tr>
<tr>
<td>Lingering mean speed</td>
<td>0.0588</td>
<td></td>
</tr>
<tr>
<td>Homebase occupancy</td>
<td>0.0712</td>
<td></td>
</tr>
<tr>
<td># stops per excursion</td>
<td>0.1202</td>
<td></td>
</tr>
<tr>
<td>Stop diversity</td>
<td>0.1489</td>
<td></td>
</tr>
<tr>
<td>Length of progression segments</td>
<td>0.5150</td>
<td></td>
</tr>
<tr>
<td>Activity decrease</td>
<td>0.8875</td>
<td>0.05 = .05(17/17)</td>
</tr>
</tbody>
</table>
The graphical way to look at it
FDR controlling procedures - adjusted p-values.

Westfall and Young (‘98), Storey (‘03)

- Order the p-values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$

- Let

$$k = \max \{ i : p_{(i)} \leq (i / m) q \}$$

or

$$k = \max \{ i : mp_{(i)} / i \leq q \}$$

- Define BH adjusted p-values, called q-values

$$p_{(i)}^{BH} = \max \{ j \leq i : mp_{(j)} / j \}$$

- Reject $H_{(i)}$ if

$$p_{(i)}^{BH} \leq q$$
FDR control of the BH procedure

If the test statistics are:

- Independent

\[ FDR = \frac{m_0}{m} q \]

- Independent and continuous

- Positive dependent

\[ FDR = \frac{m_0}{m} q \]

- General

\[ FDR = \frac{m_0}{m} q \left( \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m} \right) \approx \frac{m_0}{m} q \log(m) \]

YB&Hochberg (‘95). YB&Yekutieli (‘01)
Positive dependency

- Important cases covered by PRDS
  - Multivariate Normal with positive correlation
  - Absolute Studentized independent normal
  - (Studentized PRDS distribution, for q<.5)
  - Monotone latent variable $X | U=u$ ind. and co-monotone in $u$
- Important cases not covered by theory
  - Absolute (studentized) correlated normals
  - Pairwise comparisons
- But by practice
  (i.e. simulations, partial theoretical results)
Adaptive procedures that control FDR

- Recall the $m_0/m (= p_0)$ factor of conservativeness
- Hence: if $m_0$ is known, the BH procedure with $q i/ m(m/m_0) = q i/ m_0$ controls the FDR at $q$ exactly i.e. an "FDR Oracle"
- The adaptive procedure
  Estimate $m_0$ (or $p_0$) from the p-values

Schweder&Spjotvol (‘86), Hochberg&BY (‘90), BY&Hochberg (‘00) Storey (‘03)…
The graphical approach of Schweder & Spjotvoll

Quantile Plot of p-values

$m_0$
Option 3: The step-down multi-stage procedure

Holm: Starting with $p_{(1)}$, Compare $p_{(i)} \leq \alpha/(m+1-i)$;
step to higher p-value reducing the size of the family by 1.
  Stop with first non-rejection.

Multi-stage: Starting with $p_{(1)}$, compare $p_{(i)}$ to $q i/(m+1-i(1-q))$;
step to higher p-value reducing the size of the family by $1-q$.
  Stop with first non-rejection.
The step-down Multiple Stage procedure:

Let \( k = \max\{i : \forall j \leq i \ p(j) \leq \frac{q^j}{m+1-j(m+1-q)}\} \).

If such a \( k \) exists, reject the \( k \) associated hypotheses; otherwise reject no hypothesis.
Bayesian and Empirical Bayes approaches

- Started with Tusher et al (2001) in the context of gene expression analysis. Thresholding significance at a
- Storey (2012) \[ pFDR(a) = \frac{E(V(a)/R(a))}{R(a) > 0} \]
  \[ = \frac{FDR(a)}{Pr(R(a)>0)} \sim FDR \]
- Efron (’01), … until ‘Large Scale Inference’ Book (’10)
  \[ Fdr(a) = \frac{E(V(a))}{E(R(a))} \sim FDR \sim pFDR \]
  and the local FDR \[ fdr(x) = \frac{p_0 f_0(x)}{f(x)} \]
  \[ = \frac{p_0 f_0(x)}{p_0 f_0(x) + p_1 f_1(x)} \]
  and estimating \( p_0 \), \( f(x) \) and even \( f_0(x) \) makes it ‘empirical.
A well developed methodology addressing same goals.
Weighted FDR

- The approaches we have described take all hypotheses on equal footing
- Weighted procedures make distinctions, hypothesis $H_i$ receives weight $\omega_i$, $\sum \omega_i = m$, reflecting
- (a) Its importance $\text{YB & Hochberg (‘98)}$
  \[
  wFDR = E \left( \sum \omega_i V_i \right) / \left( \sum \omega_i R_i \right)
  \]
  it allows to assign monetary to decisions. Or,
- (b) The advantage it gets $\text{Genovese & Wasserman (‘06)}$
  \[
  p_i^* = p_i / \omega_i
  \]
  FDR defined, and tested, as before
- Both are underutilized
FDR a thing of the past?
Selective Inference, the False Discovery Rate, and analysis of neuro data

Part B

Yoav Benjamini
Tel Aviv University

Summer School – Ohrid, Macedonia

Research Supported by European Research Council grant: PSARPS
http://replicability.tau.ac.il
Outline

1. Simultaneous and Selective inference
2. Testing with FDR control
3. False Coverage Rate
4. Estimation and Model Selection
5. More complex families
One concern - different directions

- Marginal (standard) 95% Confidence Interval (CI) offers:
  \[ \Pr(\text{the marginal interval covers its parameter}) = 0.95 \]
  or equivalently
  \[ \Pr(\text{the marginal interval fails to cover its parameter}) = 0.05 \]
- With many such intervals
  \[ \Pr(\text{some intervals fail to cover}) > 0.05, \]
  using Simultaneous CIs, (e.g. Bonferroni), assures \( \leq 0.05 \)

- Why bother? On the average over all parameters,
  the expected proportion of intervals failing to cover \( \leq 0.05 \).
20 parameters to be estimated with 90% CIs

3/20 do not cover

3/4 CI do not cover when selected

These so selected 4 will tend to fail, or shrink back, when replicated

Selection of this form harms Bayesian Intervals as well (Wang & Lagakos ‘07 EMR, Yekutieli 2012)
The False Coverage-statement Rate (FCR)

A selective CIs procedure uses the data $T$

- to select $S(T) \subseteq \{1, 2, \ldots, m\}$
- to state confidence intervals for the selected

The False Coverage-statement Rate (FCR) of a selective CIs procedure is

$$FCR = E\left(\frac{\sum_{i \in S} \chi_{\{\theta_i \not\in CI_i\}}}{|S|}\right)$$

($|S|$ may be 0 in which case the ratio is 0)

FCR is the expected proportion of coverage-statements made that fail to cover their respective parameters
FCR adjusted selective CIs

(1) Apply a selection criterion $S(T)$
(2) For each $i \in S(T)$,
    construct a marginal $1 - q \frac{|S(T)|}{m}$ Conf. Int.

Thm: For any (simple) selection procedure $S()$, if the components of $T$ are independent or Positive Regression Dependent, the above Conf. Ints enjoy $FCR \leq q$.

Simple need not be that simple:
unadjusted testing, Bonferroni testing, BH, largest k…
If Test $\mu_i = 0$ & Select controlling FDR (with BH)
Select $i \leftrightarrow$ the FCR-adjusted CI doesn’t cover 0
Massive Selection - by a Table

Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes

SCIENCE, 1 JUNE 2007
## Main Table

<table>
<thead>
<tr>
<th>rs</th>
<th>chr</th>
<th>position</th>
<th>A1 A2 Region</th>
<th>WTCCC 1924 cases OR (95% CI)</th>
<th>Replication meta-analysis 3757 cases OR (95% CI)</th>
<th>All UK sample meta-analysis 5681 cases OR (95% CI)</th>
<th>DGI 7252 cases OR (95% CI)</th>
<th>FUSION 2432 cases OR (95% CI)</th>
<th>All combined 14,586 cases OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs8050136</td>
<td>16</td>
<td>52373776</td>
<td>A C FTO</td>
<td>1.27 (1.16–1.37)</td>
<td>1.22 (1.12–1.32)</td>
<td>1.23 (1.18–1.32)</td>
<td>1.03 (0.91–1.17)</td>
<td>0.25</td>
<td>1.11 (1.02–1.20)</td>
</tr>
<tr>
<td>rs10946398</td>
<td>6</td>
<td>20769013</td>
<td>A C CDKAL1</td>
<td>1.20 (1.10–1.31)</td>
<td>1.14 (1.07–1.22)</td>
<td>1.16 (1.10–1.22)</td>
<td>1.08 (1.03–1.14)</td>
<td>2.4x10^-3</td>
<td>1.12 (1.03–1.22)</td>
</tr>
<tr>
<td>rs5015480</td>
<td>10</td>
<td>94455539</td>
<td>C T HHEX</td>
<td>1.22 (1.12–1.33)</td>
<td>–</td>
<td>1.13 (1.07–1.19)</td>
<td>1.14 (1.06–1.22)</td>
<td>1.10</td>
<td>1.10 (1.01–1.19)</td>
</tr>
<tr>
<td>rs1111875</td>
<td>10</td>
<td>94452862</td>
<td>C T HHEX</td>
<td>–</td>
<td>1.08 (1.01–1.15)</td>
<td>4.6x10^-6</td>
<td>1.7x10^-4</td>
<td>0.020</td>
<td>1.00 (1.01–1.15)</td>
</tr>
<tr>
<td>rs10811661</td>
<td>9</td>
<td>22124094</td>
<td>C T CDKN2B</td>
<td>1.22 (1.09–1.37)</td>
<td>1.18 (1.08–1.28)</td>
<td>1.19 (1.11–1.28)</td>
<td>1.20 (1.12–1.28)</td>
<td>5.4x10^-3</td>
<td>1.20 (1.07–1.36)</td>
</tr>
<tr>
<td>rs564398</td>
<td>9</td>
<td>22019547</td>
<td>C T CDKN2B</td>
<td>1.16 (1.07–1.27)</td>
<td>1.12 (1.05–1.19)</td>
<td>1.13 (1.08–1.19)</td>
<td>1.05 (0.94–1.17)</td>
<td>0.05</td>
<td>1.13 (1.01–1.27)</td>
</tr>
<tr>
<td>rs4402960</td>
<td>3</td>
<td>186994389</td>
<td>G T IGFBP2</td>
<td>1.15 (1.05–1.25)</td>
<td>1.09 (1.01–1.16)</td>
<td>1.11 (1.05–1.16)</td>
<td>1.17 (1.11–1.23)</td>
<td>1.18</td>
<td>1.18 (1.08–1.28)</td>
</tr>
<tr>
<td>rs13266634</td>
<td>8</td>
<td>118253964</td>
<td>C T SLC30A8</td>
<td>1.12 (1.02–1.23)</td>
<td>1.12 (1.04–1.19)</td>
<td>1.12 (1.05–1.18)</td>
<td>1.07 (1.00–1.16)</td>
<td>0.047</td>
<td>1.18 (1.09–1.29)</td>
</tr>
<tr>
<td>rs7901695</td>
<td>10</td>
<td>114744078</td>
<td>C T TCF7L2</td>
<td>1.37 (1.25–1.49)</td>
<td>–</td>
<td>–</td>
<td>1.38 (1.31–1.46)</td>
<td>2.3x10^-3</td>
<td>1.34 (1.21–1.49)</td>
</tr>
<tr>
<td>rs5215</td>
<td>11</td>
<td>17365206</td>
<td>C T KCNJ11</td>
<td>1.15 (1.05–1.25)</td>
<td>–</td>
<td>–</td>
<td>1.15 (1.09–1.21)</td>
<td>1.0x10^-7</td>
<td>1.11 (1.02–1.20)</td>
</tr>
<tr>
<td>rs1801282</td>
<td>3</td>
<td>12368125</td>
<td>C G PPARG</td>
<td>1.23 (1.09–1.41)</td>
<td>–</td>
<td>–</td>
<td>1.09 (1.01–1.16)</td>
<td>0.019</td>
<td>1.20 (1.07–1.33)</td>
</tr>
</tbody>
</table>

**P_adj**
Odds ratio point and CI estimates for confirmed T2D susceptibility variants

<table>
<thead>
<tr>
<th>Region</th>
<th>Odds ratio</th>
<th>0.95 CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>• FTO</td>
<td>1.17</td>
<td>[1.12, 1.22]</td>
</tr>
<tr>
<td>• CDKAL1</td>
<td>1.12</td>
<td>[1.08, 1.16]</td>
</tr>
<tr>
<td>• HHEX</td>
<td>1.13</td>
<td>[1.08, 1.17]</td>
</tr>
<tr>
<td>• CDKN2B</td>
<td>1.20</td>
<td>[1.14, 1.25]</td>
</tr>
<tr>
<td>• CDKN2B</td>
<td>1.12</td>
<td>[1.07, 1.17]</td>
</tr>
<tr>
<td>• IGF2BP2</td>
<td>1.14</td>
<td>[1.11, 1.18]</td>
</tr>
<tr>
<td>• SLC30A8</td>
<td>1.12</td>
<td>[1.07, 1.16]</td>
</tr>
<tr>
<td>• TCF7L2</td>
<td>1.37</td>
<td>[1.31, 1.43]</td>
</tr>
<tr>
<td>• KCNJ11</td>
<td>1.14</td>
<td>[1.10, 1.19]</td>
</tr>
<tr>
<td>• PPARG</td>
<td>1.14</td>
<td>[1.08, 1.20]</td>
</tr>
</tbody>
</table>

Using marginal CI is more common than marginal tests. Alas, protecting from the effect of selection in testing does not solve the problem in estimation.
## Odds ratio point and CI estimates for confirmed T2D susceptibility variants

<table>
<thead>
<tr>
<th>Region</th>
<th>Odds ratio</th>
<th>0.95 CIs</th>
<th>FCR-adjusted CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>• FTO</td>
<td>1.17</td>
<td>[1.12, 1.22]</td>
<td>[1.05, 1.30]</td>
</tr>
<tr>
<td>• CDKAL1</td>
<td>1.12</td>
<td>[1.08, 1.16]</td>
<td>[1.03, 1.22]</td>
</tr>
<tr>
<td>• HHEX</td>
<td>1.13</td>
<td>[1.08, 1.17]</td>
<td>[1.02, 1.25]</td>
</tr>
<tr>
<td>• CDKN2B</td>
<td>1.20</td>
<td>[1.14, 1.25]</td>
<td>[1.07, 1.34]</td>
</tr>
<tr>
<td>• CDKN2B</td>
<td>1.12</td>
<td>[1.07, 1.17]</td>
<td>[1.00, 1.25]</td>
</tr>
<tr>
<td>• IGF2BP2</td>
<td>1.14</td>
<td>[1.11, 1.18]</td>
<td>[1.06, 1.23]</td>
</tr>
<tr>
<td>• SLC30A8</td>
<td>1.12</td>
<td>[1.07, 1.16]</td>
<td>[1.01, 1.24]</td>
</tr>
<tr>
<td>• TCF7L2</td>
<td>1.37</td>
<td>[1.31, 1.43]</td>
<td>[1.23, 1.53]</td>
</tr>
<tr>
<td>• KCNJ11</td>
<td>1.14</td>
<td>[1.10, 1.19]</td>
<td>[1.03, 1.26]</td>
</tr>
<tr>
<td>• PPARG</td>
<td>1.14</td>
<td>[1.08, 1.20]</td>
<td>[1.00, 1.30]</td>
</tr>
</tbody>
</table>

and FCR adjusted intervals
Success
Problem?

How well do we do?

Figure 1: Simulated example – scatter plot of $|Y_i| > 3.111$ components. $Y_i$ values are drawn on the abscissa of the plot, the ordinates are $\theta_i$ values. The red lines are marginal 0.95 CIs. The green lines are 0.05 FCR-adjusted CIs.
Adjusting to the selection procedure used

Utilize the nature of the selection process being used in order to improve selective inference CIs and tests

In particular selection of \( \theta \) if its estimator is big enough

\[ X = (Y \mid |Y| \geq c), \]

where \( c \) is either fixed or (simple) data dependent.

Weinstein, Fithian, YB (‘13)
The complication: $\theta$ is no longer only a shift parameter
Conditional Quasi-Conventional CI

Design acceptance region for testing $\theta = \theta_0$ that:

• Have correct level under the conditional distribution
• Are as short as possible
• Avoid including observations of opposite sign to $\theta_0$

Invert them to get conditional CIs.

*following YB, Hochberg & Stark (‘98)

The intervals will also control the False Coverage Rate
Example

16 Subjects view 2 movie segments of different stress level. Recordings was made of:

• Activity at voxels in the brain and
• The level of Cortisol in their blood

Goal: Estimate the correlation between these difference in activity and the difference in Cortisol levels across subjects, in the promising voxels.

• 14756 correlations - one for each voxel.
• Interest lies only with voxels for which the correlation is high: $|r| \geq 0.6$ (here: pre-determined).

• 15 voxels $r \geq 0.6$; 21 voxels with $r \leq -0.6$. 
CQC Intervals for Selected Correlations

Better than splitting to learning/testing; Software in JASA paper
Addressing ‘voodoo correlations’

Estimating quantities of interest correlated with brain activity from the same data used to locate the most promising ones. (Behavioral Neuroimaging).

Vul et al 2009 ‘blew the whistle’ on the practice.

It took a few years, heated debate, and a joint paper by 8 experts to realize the problem is of selective inference (named also ‘Circular reasoning’, ‘Double Dipping’) and that:

Voodoo correlations are everywhere…

Their proposed solution: data splitting
Addressing in-study ‘voodoo correlations’

Confidence Calibration Plot: Observed correlations in significant voxels (B-H;FDR 0.1) encoding conditional confidence intervals as well.  

Rosenblatt &YB ‘14+
Hedges ‘84, Zhong and Prentice ‘08
In Fithian, Sun Taylor terminology: 100% used for selection

Amit Meir and YB (‘15+)
Outline

1. Simultaneous and Selective inference
2. Testing with FDR control
3. False Coverage Rate
4. Estimation and Model Selection
5. More complex families
Motivation: Wavelets

Noisy signal: \( y_i = \mu_i + e_i, \quad e_i \sim N(0, \sigma^2) \quad i=1,2,\ldots m \quad \text{ind.} \)

The idea: For the prediction of linear function of \( \mu_i \)

Screen: Threshold small coefficients

If \( \mu_i^2 \leq \sigma^2 \) zeroing is better than estimating (screening)

- Testing whether \( \mu_i = 0 \) <=> Hard Thresholding
- Bonferroni <=> Universal threshold \( \sigma (2 \log(m))^{1/2} \)
  Donoho & Johnstone (‘94)
- FDR testing
Testimation

- $Y_i \sim N(\mu_i, \sigma^2)$ $i=1,2,\ldots,m$ independent
- Test using BH
  \[ p_{(k)} \leq \frac{qk}{m} \iff |Y_i| \geq \sigma Z_{\frac{qk}{2m}} \]
- Estimate using
  \[ Y_i^{FDR} = 0 \text{ if } |Y_i| < \sigma Z_{\frac{qk}{2m}} \text{ (ignore)} \]
  \[ = Y_i \text{ if } |Y_i| \geq \sigma Z_{\frac{qk}{2m}} \text{ (report)} \]
  amounting to hard thresholding
- Use $Y^{FDR}$ instead of $Y$
- Used to screen hundred of thousands of variables before complicated modeling (in genomics)
Testimation - some theory

Measure performance of estimating vector by expected \( l_r \)-loss \( 0 < r \leq 2 \):

\[
\sum (error)^2; \quad \sum |error|, \quad #(errors)
\]

relative to best “oracle” performance

Let \( #( parameters) \rightarrow \infty \)

Consider bodies of sparse signals such as:

- \( \text{prop( non-zero coefficients)} \rightarrow 0 \) \quad (i.e. \( p_0(m) \rightarrow 1 \)),
- size of sorted coefficients decays fast

Hard thresholding by FDR testing of the coefficients (with \( q<1/2 \)) is adaptively minimax simultaneously over bodies of sparse signals

Abramovich, YB, Donoho, & Johnstone (‘06)
What have we further learned from theory

1. Use $q < 1/2$ for minimax performance

1. FDR testing is relevant, and “works well”, even when no hypothesis is true

$$|\mu|_{(i)} \leq C i^{-1/p} \quad \text{for all } i, p < 2$$

(if small $\mu_i$ are moved to their null value 0 the estimation problem relative to oracle is harder)
Wider implications for model selection

Traditional model selection with penalized Residuals Sum of Squares (AIC, $C_p$), minimize:

$$RSS(k) + 2k\sigma^2$$

$k$ #number of parameters in current model

$m$ #number of parameters searched

Penalty per parameter $\lambda_{k,m}$ increases in $m$ decreases in $k$

An FDR testing based penalty:

$$RSS(k) + \lambda_{k,m}k\sigma^2 = RSS(k) + \left(\frac{1}{k} \sum_{i=1}^{k} \frac{Z_{iq}^2}{2m}\right)k\sigma^2$$
What is gained by introducing FDR penalty

Ex. 1: Diabetics data  (Efron et al ‘04)
Data: 442 Diabetics patients;
10 baseline measurements for each,
to which 45 interactions and 9 quadratic terms were
added (SEX^2=SEX…)

Dependent variable: A quantitative measure of
disease progress after one year.

Goal: predictive modeling from baseline data
Multiple-Stage FDR with $q=.05$

FSR: Introducing random explanatory variables and continuing until their proportion in the model reaches .05

Wu, Boos, & Stefanski (2007).

Least Angle Regression Selection (LARS)

<table>
<thead>
<tr>
<th>Method</th>
<th>N.</th>
<th>Variables in the model</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS (at .05)</td>
<td>7</td>
<td>BMI, S5, BP, AGE<em>SEX, BMI</em>BP, S3, SEX</td>
<td>.53</td>
</tr>
<tr>
<td>FSR</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LARS</td>
<td>16</td>
<td>BMI, S5, BP, S3, BMI<em>BP, AGE</em>SEX, S6^2, BMI^2, AGE<em>BP, AGE</em>S6, SEX, S6, AGE<em>S5, AGE^2, SEX</em>BP, BP*S3</td>
<td>.55</td>
</tr>
</tbody>
</table>
Ex 2: High dimensionality

Affecting classification and ranking algorithms

Example: Microarray dataset of 10 normal and 86 cancerous lung tissues (Beer, et al., ‘02), 7127 features, analyzed in Rupin’s Lab (Bionformatics, ’05)

The goal: Produce a stable ranked gene list, the top of which should be a “good” set of classifiers.
Rupin’s Lab Method:
(i) Producing 1000 different gene sets according to the SVM models of sizes 5 up to 100, on bootstrapped samples
(ii) ranking the genes according to their repeatability frequency in the ensemble of predictive gene sets.
Result: The gene with the highest score was “Rage”, its boxplot by two classes is presented below
Tool: selection adjusted regression

- Choose by forward (greedy) selection the features to enter the logistic model in order to minimize the deviance plus FDR penalty.
- Unlike the penalties in AIC, BIC or Cp where it linear in model size $k$; and is unaffected by the size of the pool of features $m$ from which selection takes place, the FDR penalty increases in $m$ and decreases in $k$.

  YB & Gavrilov ('13)

- Replicating 120 times by bootstrapping,

  In all replications only one gene is selected.
By 'TNA'

By 'FABP4'

By 'COX7A1'

By 'FHL1'

By 'PECAM1'

By 'AGER'

Normal samples  Cancer samples

Normal samples  Cancer samples

Normal samples  Cancer samples

Normal samples  Cancer samples

Normal samples  Cancer samples

Normal samples  Cancer samples
Post Model Selection Inference

• So far interest in model selection for prediction
• In last example tried to infer about the selected variables
• What interpretation can we give to the parameters in the model selected with FDR penalty?
  • Control of the “false dimensions rate” in the selected model?
• Not clear: Recall that as we move forward the parameters estimates (and the parameter estimated) change. (My hunch – controlled)

• Is the Forward Selection path essential?

How about $l_1$ LASSO (LARS) path?
Current work

- Three current papers out of Stanford teams deal with testing along the Lasso path, while controlling the size of the model using the FDR idea
  False Discovery/Selection/Variables Rate

Data splitting  
G'Sell, Hastie, Tibshirani, Lockhart, Taylor, Tibshirani, J. Tibshirani,
Asymptotic p-values  
G'Sell, Wager, Chouldechova, Tibshirani
Sequential Testing  
G'Sell, Hastie, Tibshirani, Lockhart, Taylor, Tibshirani, J. Tibshirani,

- The fourth introduces “sorted $l_1$” version of FDR penalty
  Bogdan, van den Berg, Su, Candes

\[
\| Y - \hat{Y} \|^2 + \sigma^2 \sum_{i=1}^{k} \frac{Z_{iq}^2}{2m} \quad \| Y - \hat{Y} \|^2 + \sigma^2 \sum_{i=1}^{k} |\beta(i)| \cdot \frac{Z_{iq}}{2m}
\]

More have come from Taylor (Stanford and his students)
Outline

1. Simultaneous and Selective inference
2. Testing with FDR control
3. False Coverage Rate
4. Estimation and Model Selection

5. More complex families
Recognizing a family

A family should best be defined by the danger of selective or simultaneous inference that is being faced:

A family is the richest set of inferences in an analysis, all achieving the same goal, from which one selected inference could be replaced by another selected inference for presentation, highlighting or action.

Different researchers can have different goals and thus define differently the families – still decisions can be defendable and with no arbitrariness.
Testing selected families

We select interesting/significant/promising families

We wish to test hypotheses within the selected families

and there too select the significant ones
The locations of associated voxels per SNP, for the 5 most associated SNPs
Separate vs joint FDR testing of families

Homogeneous case 50 families 10 hypotheses in each

\[ m_0/m \sim \text{constant} \ ( < 1) \]

Separate ~ Joint (scalability of the separate)

Heterogeneous case 50 families 10 hypotheses in each

\[ m_0/m = 1 \text{ for 49 families} \quad m_0/m = 0 \text{ for 1 family} \]

When Joint analysis: too liberal for 49, too conservative for the 1

Separate analysis: too liberal for the 49.

Overall FDR may reach .9

Efron’s comment (2008)
Efron’s comment (2008)

Fig 10: Enrichment analysis of Imaging data, Panel D of Figure 1; z-value for original 15445 voxels have been averaged over “gene-sets” of neighboring voxels with city-block distance $\leq 2$. Coded as “-” for $\bar{z}_i < 0$, “+” for $\bar{z}_i \geq 0$; solid rectangles, labeled as “Enriched”, show voxels with $\hat{fDR}(\bar{z}_i) \leq 0.2$, using empirical null.
Justifications for separate FDR testing

- If $Q_i$ is the false discovery proportion in family $i$, control $E(Q_i)$ separately for each family $i$, and get for free control of the average over all families!

\[ E \left( \frac{i}{m} Q_i \right) = \frac{i}{m} E(Q_i) = \frac{mq}{m} = q \]

Again, the “Don’t worry be happy” approach seems to work.

- But if only some of the families are selected based on the same data, control on the average over the selected ones is not guaranteed.
Selection adjusted separate testing of families

Let $P_i$ be the p-values for the hypotheses in family $i$, $S(P)$ data based selection procedure of families, $|S(P)|$ the (random) number of families selected.

The control of error $E(\mathcal{J})$ (FDR, but also FWER, and others) on the average over the selected families means

$$E\left[\frac{\sum_{i \in S(P)} C_i}{|S(P)|}\right] = q$$
Selection adjusted separate testing

For any ‘simple’ selection procedure $S(P)$, and for any error-rate of the form $E(C_i)$, if the $P_i$ across families are independent,

controlling $E(C_i) \leq q|S(P)|/m$ for all $i$,

assures control on the average over the selected at level $q$.

Note 1: if only one selected - amounts to $q/m$;

if all selected no adjustment needed

Note 2: If not ‘simple’ selection rule only the definition of $|S(P)|$ is more complicated, that’s all.
• **There was no restriction on the selection rule**

• In particular for each family calculate a p-value for the intersection hypothesis and test across families

Get:
* Within family FDR
* Average FDR over selected,
* Across families FDR (or any other error-rate).

Heller & YB (‘08), Sun & Wei (’10+) False Sets Rate YB, Bogomolov
Hierarchical BH testing

If we use the BH adjusted p-value to test the intersection of each family and use BH (or Bonferroni) to test within selected families, selection adjusted $FDR \leq q$ even under positive regression dependency.

A recent result of Guo & Sarkar et al (+12) for families of equal size shows that the over-all FDR is controlled when the second stage uses adaptive Bonferroni method.
Re-analysis of SNP-voxel data for Alzheimer
Re-analysis of SNP-voxel data for Alzheimer

- Family = the set of all association hypotheses for a specific SNP and all voxels (~34K)
  (So selection of families = selection of SNPs)
  Calculate p-value per SNP-family using Simes’ test.
- Test SNPs while controlling FDR over SNPs: 35 SNPs
- Test voxels within families of selected SNPs using BH at level .05*35/34,000
- For most SNPs ≤ 50 voxels; the max 400 voxels.
Other examples

- SNPs and gene expression (eQTL analysis)
  Peterson, Bogomolov, YB, Sabatti (Bioinformatics ‘15)
Family – all SNPs associations with a gene
Choose genes then SNPs in gene

- SNPs and multiple phenotypes (features)
  (Peterson et al. Gen. Epid. 2014)
Family – all phenotypes associations with a gene
Choose SNPs then phenotypes associated with this SNP
Other examples

• What predicts quantitative aspects of patients, each one separately? (Tal Kozlovski’s poster)
Family - the individual predictors within for each clinical variable
Select the clinical variables for which there is evidence for the entire model to predict (F-test)
Then select predictors within each selected model (Bonferroni)

• Current work in Brain research: generalize the methodology to 3 or more levels
Purpose: associate genes’ expression with
hierarchically organized measures of bipolar disease according to their clinical structure.
Multiple phenotypes
NIH Bipolar Disorder research with Chiara Sabatti
Selective inference challenges in open access data

Foster & Stein (‘08): $\alpha$-investment to control at level $q$ the

$$mFDR = \frac{E(V)}{E(R) + \nu}$$

Hypotheses arrive sequentially; in study $i$, test $H_i$ with $\alpha_i$;
if $H_i$ rejected $\alpha_{i+1} > \alpha_i$ (as only denominator increases) They
gave a simple and effective rule.

An optimal set of online rules for FDR:

Aharoni & Rosset (14); later by Javanmard & Montanari (15),

Note: order still need to be maintained
A potential outcome of this successful summer school

• A potential challenge
  Combine
  Hierarchical Testing Schemes
  with
  Hierarchical Prediction Schemes

Feasible? Useful? Worth a try
• Worry about the effect of selection
• It might be enough to assure properties ‘on the average over the selected’
• There are simple and useful methods for testing and confidence intervals
• The ideas seem important in other situations for the analysis of Big Data, or Large Scale Inference problems

• Many challenges still exist, more are coming.

Thanks
References

• Benjamini & Yekutieli (2005) "False Discovery Rate controlling confidence intervals for selected parameters". JASA


• Williams, Johns & Tukey (1999) Journal of Educational and Behavioral Statistics
References

References


