Selective Inference and the False Discovery Rate

Yoav Benjamini
Tel Aviv University

Simon's Institute for the Theory of Computing
UC Berkeley

Preparation Supported by European Research Council grant: PSARPS
http://replicability2.wordpress.com/

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 first data-mining problem in Statistics (in Science?)

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)

$$Y_{i+} \sim N(\mu_i, \sigma^2/5) \qquad i=1,2,\ldots,6;$$

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 $= \text{Prob}_{H_0}(|Z| > |Y_{(i+)} - Y_{(j+)}| / \sigma_{diff})$

  under H_0 *P-value* ~*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*5/2=15** such tests, each at level α

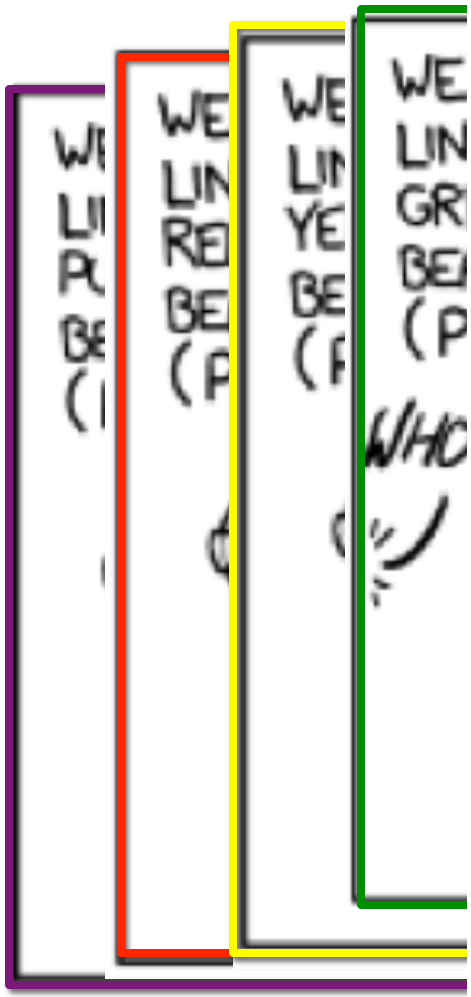
$$\text{Prob}(Z > |Y_{(k+)} - Y_{(1+)}| / \sigma_{\text{diff}}) < \alpha$$

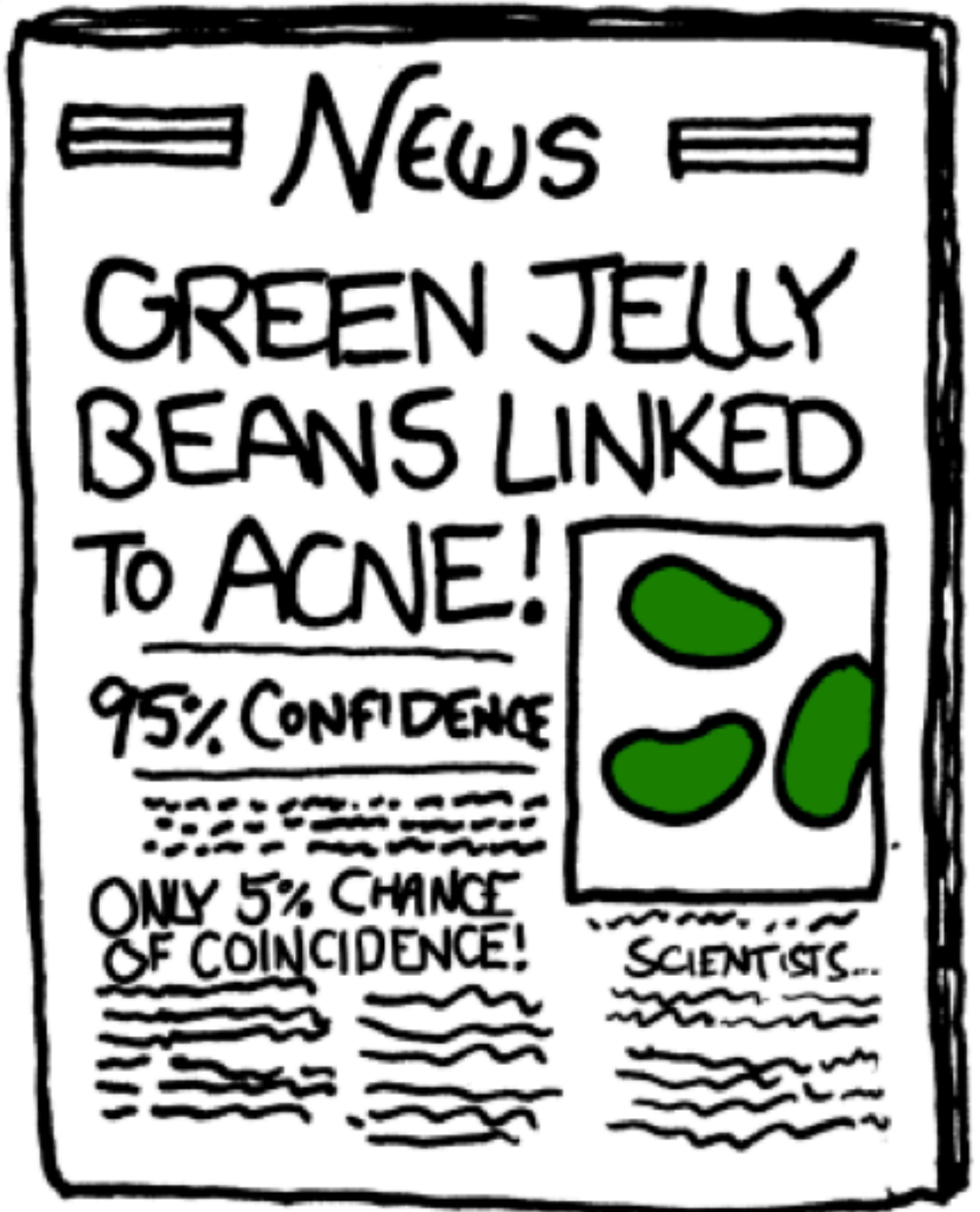
even if there is no difference. The larger k the worse it gets!

- In fact going back to the original paper we found 13 such groups resulting in **13*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

The lethal combination o





Not only Jelly Beans

"Unussual secrets are hidden in numbers. for instance, an orange car is less likely to have serious damages that are discovered only after the purchase…."

Data mining from KAGGLE website

THE MARKER IT 2.5.2012

Not only colors

Giovanni and others (95) examined the possible effect of excess eating of 130 different kinds of foods on prostate cancer.

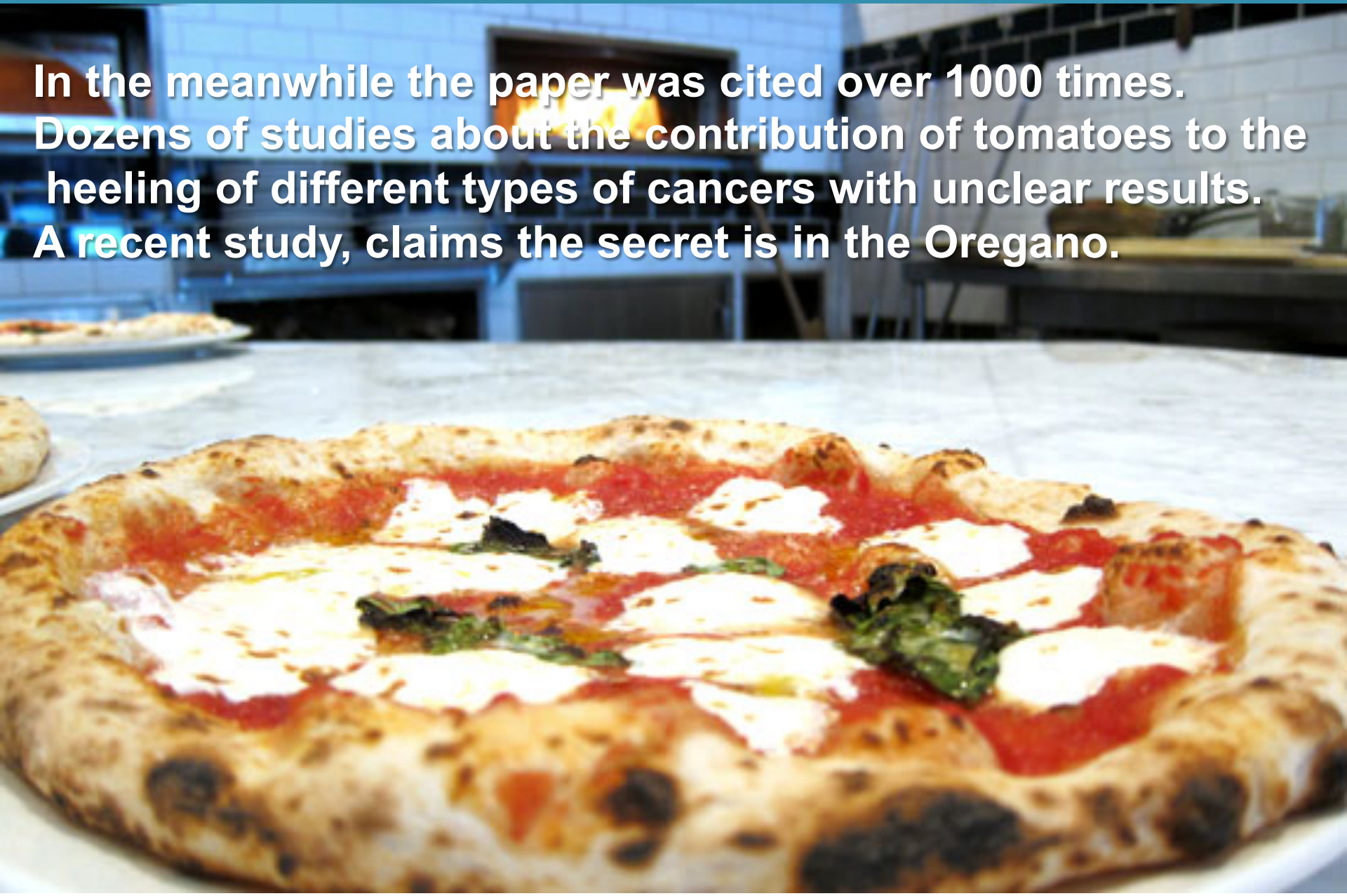
3 kinds of foods cleared the statistical significance bar –

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



Selective inference

1st line of protection when comparing multiple groups

Since the danger seems largest when $\mu_1=\mu_2=\ldots=\mu_k$

- Test first this (single) hypothesis via F-test at level α.
  - If not significant STOP
  - If significant continue with t-tests at level α as before

Fisher's protected LSD (Least Significant Difference)

But: protection is offered only when $\mu_1=\mu_2=\ldots=\mu_k$

Define such protection as the control of the

FamilyWise Error-Rate in the weak sense.

$$\Pr_{H_0}\left(making\ even\ one\ type\ I\ error\right) \le \alpha.$$

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;
$= 0$ if not

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

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

So, e.g.

$$weak\ \mathrm{FWER} \equiv \Pr_{H_0}\left(V \geq 1\right).$$

2nd line of protection

- The FamilyWise Error-Rate

For any configuration of true and null hypotheses

$$FWER = Prob(V \geq 1)$$

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

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 = # of intervals failing to cover their respective parameter.

If for any configuration of parameters

$$FWER = Prob(V \geq 1) \leq \alpha$$

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 α_{BON}

$$E(V)=E(\Sigma V_i) = \Sigma\, E(V_i) \le m_0\, \alpha_{BON} \le m\ \alpha_{BON}$$

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

(Is any condition needed?)

This is

(1) The Bonferroni simultaneous inference procedure

that controls any configuration of hypotheses

Expected number of errors $E(V) \le \alpha$

Old and trusted solutions

As the Bonferroni procedure assures $E(V) \leq \alpha$

This also assures $Pr(V \geq 1) \leq \alpha$ Because:

$$E(V) = 0Pr(V=0)+1Pr(V=1)+2Pr(V=2)+\ldots+mPr(V=m)$$
$$\geq 0 + 1Pr(V=1)+1Pr(V=2)+\ldots+1Pr(V=m)$$
$$= 0 + Pr(V \geq 1)$$

So, when using $\alpha_{BON} = \alpha/m$ for individual tests, or for CIs

$$FWER = Prob(V \geq 1) \leq E(V) \leq \alpha$$

(again no condition is needed)

More…

If the test statistics are independent,
and we test each hypothesis separately at level α_{SID}

$$Prob(V \geq 1) = 1 - Prob(V = 0) = 1 - (1 - \alpha_{SID})^{m_0} \leq 1 - (1 - \alpha_{SID})^{m} \leq \alpha$$

So to assure $Prob(V \geq 1) \leq \alpha$ we may use

$$\alpha_{SID} = 1 - (1 - \alpha)^{1/m}$$

(2) This is Sidak's multiple testing procedure

Note: If $m_0 = m$ equalities

Idea: we used dependency structure to get a better test.

How much better?

$$\alpha_{SID} = 1-(1-\alpha)^{1/m} \sim 1-(1-\alpha/m - \alpha^2/2m) = \alpha_{BON}+(\alpha^2/2m)(m-1)/m$$

Even for small m (=10) very little gain: .00511 instead of .005

(3) Tukey's procedure for pairwise comparisons:

Utilizes dependency by calculating the distribution of the studentized range statistics $(Y_{(k+)} - Y_{(1+)})/(s/n^{1/2})$,

same idea but larger gain.

Known as post-hoc analysis

Newer solutions

Stepwise procedures that make use of observed p-values:

(4) Holm's procedure:

- Let P_i be the observed p-value of the test for H_i
- Order the p-values $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$
  - If $P_{(1)} \le \alpha/m$ Reject $H_{(1)}$
  - If $P_{(2)} \le \alpha/(m-1)$ Reject $H_{(2)}$
  - …
  - Until for the first time $P_{(k)} > \alpha/(m+1-k)$
- Then stop and reject no more.

Always: $FWER \le \alpha$

Example; exploratory behavior of mice

NIH: Phenotyping Mouse Behavior High throughput screening of mutant mice

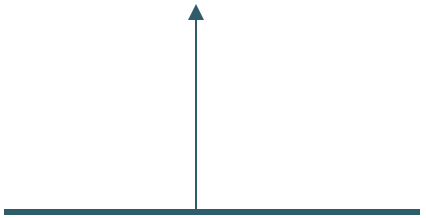
Comparing between 8 inbred strains of mice
At 3 labs: Golani at TAU, Elmer MPRC, Kafkafi NIDA

Significance of 8 Strain differences

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

Bonferroni

.05/17=.0029



Unadjusted

Unadjusted vs Simultaneous

In the search for food affecting Prostate Cancer,

3 food intakes were reducing with unadjusted significance

0 with Bonferroni.

But why care about simultaneous inference?

If a single null is true and tested at level α
on the average,
the proportion of times a type I errors is made is α.
So why should we not worry only about the proportion
of times a test/CI results in error ?
This property has is called The Per Comparison Error-Rate
where for any configuration of hypotheses

$$PCER = E(V/m) = E(V)/m$$

Testing at (nominal) level α assures per comparison level is α; *amounts to "don't worry – be happy" approach.*

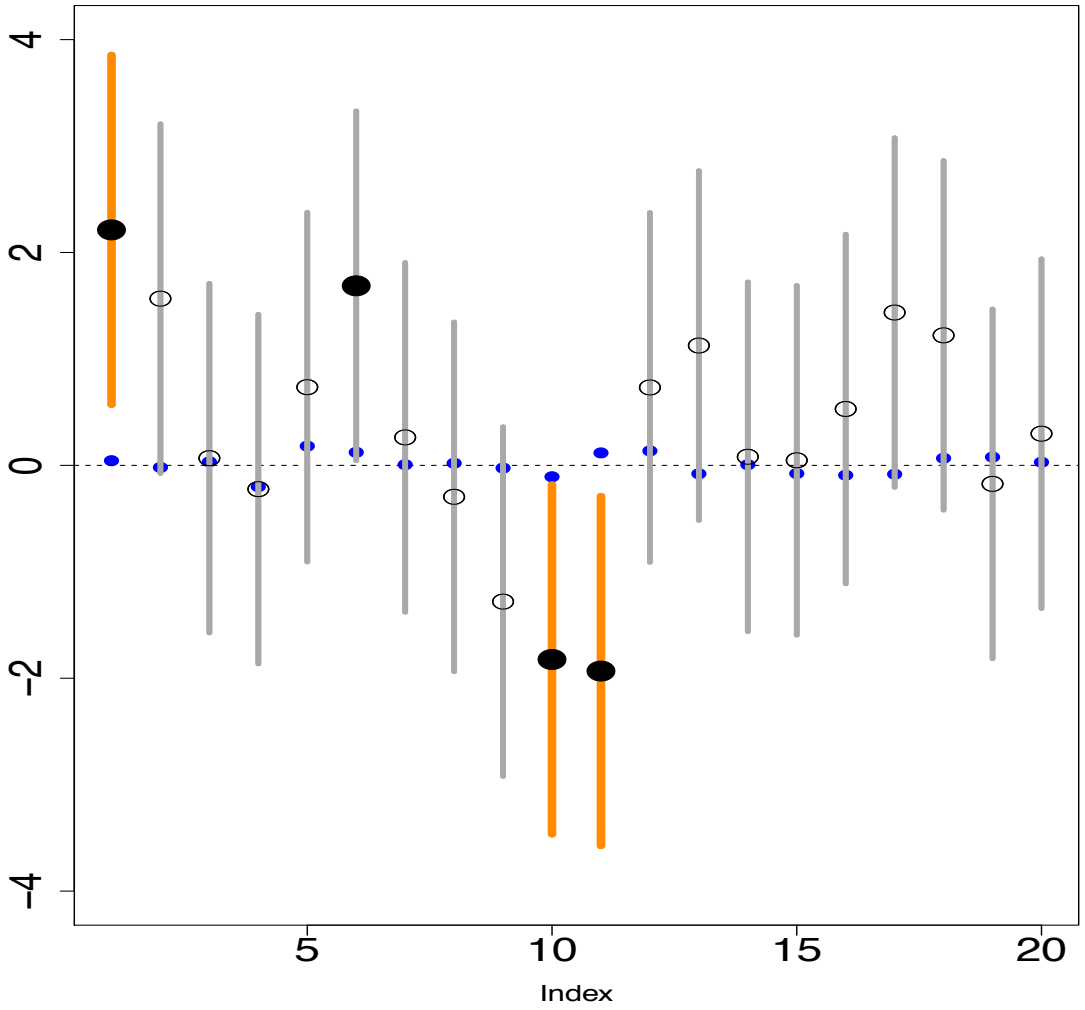
But when we select…

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



Selective inference

- The goal for selective inference is modest:

keep the original property of the individual inference to hold

on the average over the selected .

Selective inference for multiple confidence intervals:

average lack of coverage over the selected to be ≤ α.

- The goal for simultaneous inference is more ambitious:

The property for individual inference should hold simultaneously for all parameters, and therefore simultaneously for any selected subset

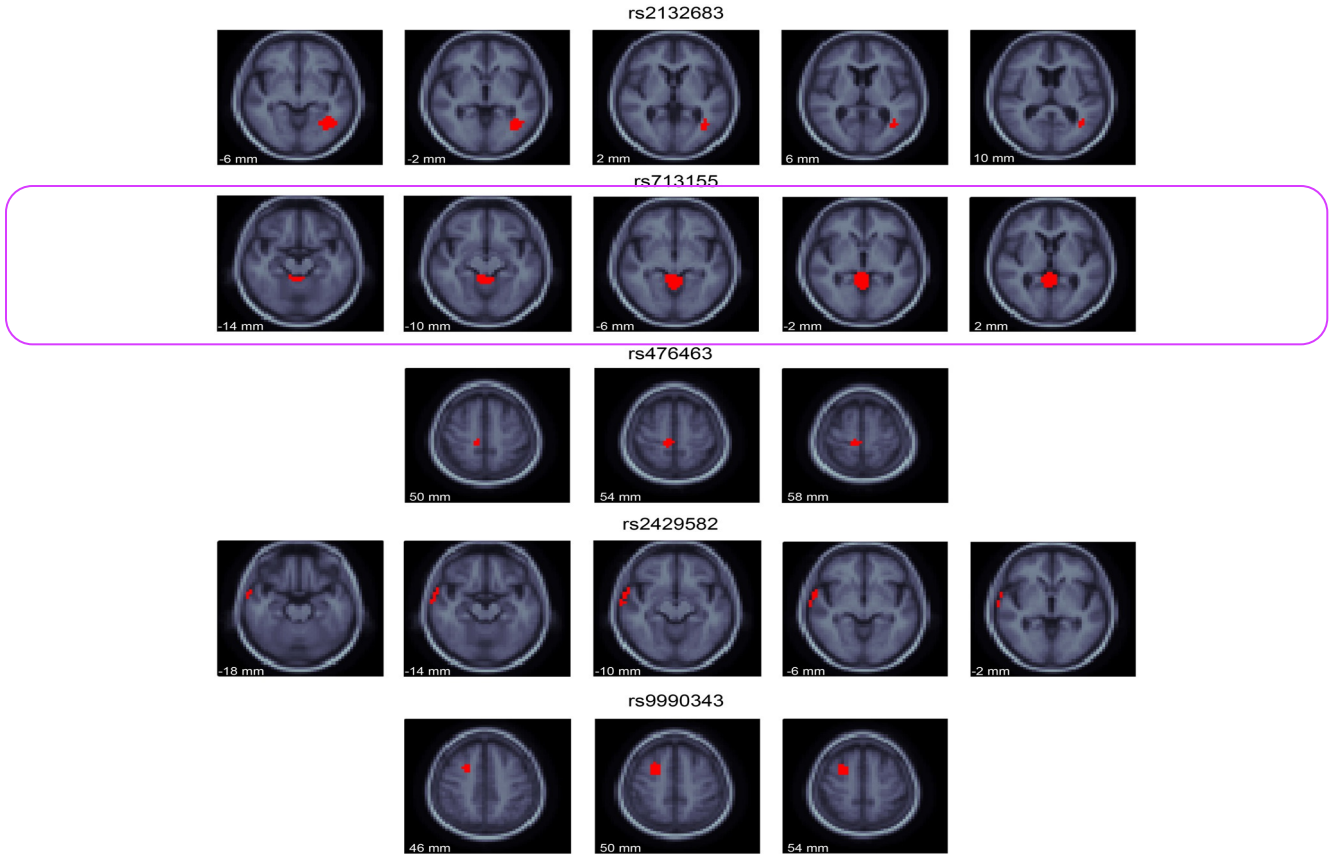
simultaneous inference => selective inference

The increasing scale:
Voxelwise Genome-Wise Association study

(Stein et al.'10)

- Alzheimer's Disease Neuroimaging Initiative (ADNI) study: 2003-2008
- Goal: determine biological markers of Alzheimer's disease by testing for associations between volume changes at voxels with genotype

1 ← Voxels searched → 32,000



1 ↓ SNPs ↓ 448,000

number of tests ~ 13,000,000,000

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

Tukey et al ('94,2000): National assessment of Educational Progress , comparing 35 States in US

# of comparisons 35*(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 = # rejected hypotheses = # discoveries

V of these may be in error = # false discoveries

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

$$Q = \frac{V}{R} \qquad R > 0$$
$$= 0 \qquad R = 0$$

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

$$FDR = E(Q)$$

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

$$Prob(V \geq 1) = E(V/R) = FDR \leq q$$

- *But otherwise*

$$Prob(V \geq 1) \geq FDR$$

So there is room for improving detection power

Reflections on goals

- Simultaneous inference: inference should hold jointly for all parameters in the family

- 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,

$$E(V/number\ of\ \ tests\ \ performed) = E(V/m) \le \alpha$$

- FDR control assures

$$E(V/number\ of\ \ tests\ selected) = E(V/R) \le \alpha$$

- 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)

Historical Perspective (II)

- *Shaffer (1997) brought back forgotten references*
  - *Eklund (unpublished work in Swedish)*
  - *Seeger(1968) about Eklund's: FWER controlled in weak sense, but not in the strong sense*
- *Simes (1986) suggested to extend his global test of the single intersection hypothesis to multiple inferences*
- *Hommel (1988) about Simes:FWER not controlled in the strong sense*
- *Hochberg (1988) and Hommel (1988)*

  the series is constants are q/(m+1-i)
- *Sen (1998a) points out to the classical Ballot Theorem*
- *Names: FDR procedure (SAS); Benjamini-Hochberg (BH); Linear Step-up*

FDR controlling proceures

The BH (Linear Step-up)procedure:

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

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

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

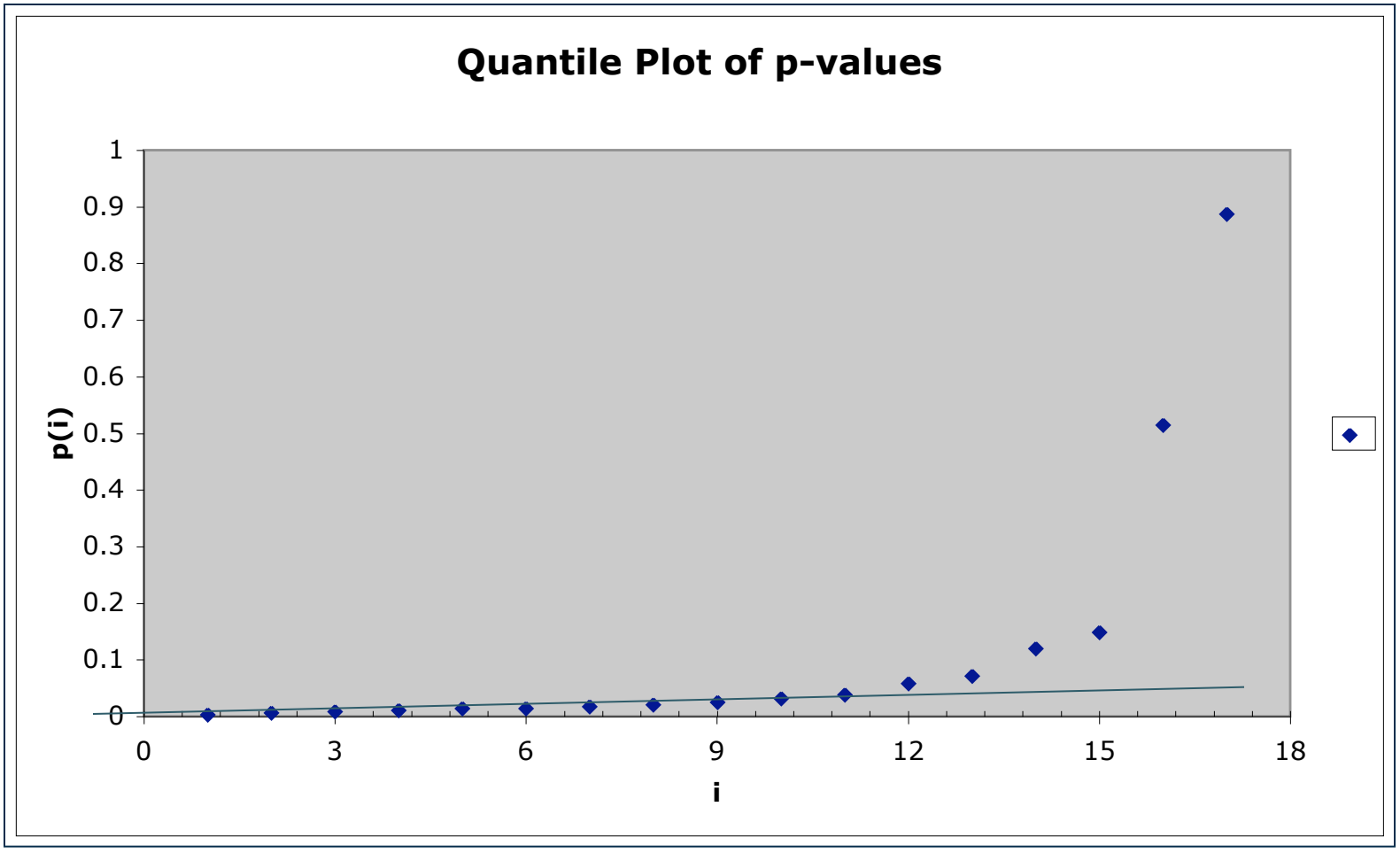
- Reject

$$H_{(1)}, H_{(2)}, \ldots, H_{(k)}$$

Significance of 8 Strain

Behavioral Endpoint	Mixed	Linear StepUp
Prop. Lingering Time	0.0029	0.0029 =.05(**1**/17)
# Progression segments	0.0068	
Median Turn Radius (scaled)	0.0092	
Time away from wall	0.0108	
Distance traveled	0.0144	
Acceleration	0.0146	
# Excursions	0.0178	
Time to half max speed	0.0204	
Max speed wall segments	0.0257	
Median Turn rate	0.0320	
Spatial spread	0.0388	
Lingering mean speed	0.0588	
Homebase occupancy	0.0712	
# stops per excursion	0.1202	
Stop diversity	0.1489	
Length of progression segments	0.5150	
Activity decrease	0.8875	0.05 =.05(**17**/17)

The graphical way to look at it



FDR controlling proceures - adjusted p-values.

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

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

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

or

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

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

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

- Reject $H_{(i)}$ $$p_{(i)}^{BH} \le q$$

FDR control of the BH procedure

If the test statistics are :

- Independent

$$FDR \le \frac{m_0}{m} q$$

- independent and continuous

$$FDR = \frac{m_0}{m} q$$

- Positive dependent

$$FDR \le \frac{m_0}{m} q$$

- General

$$FDR \le \frac{m_0}{m} q\,(1 + 1/2 + 1/3 + \ldots + 1/m)$$

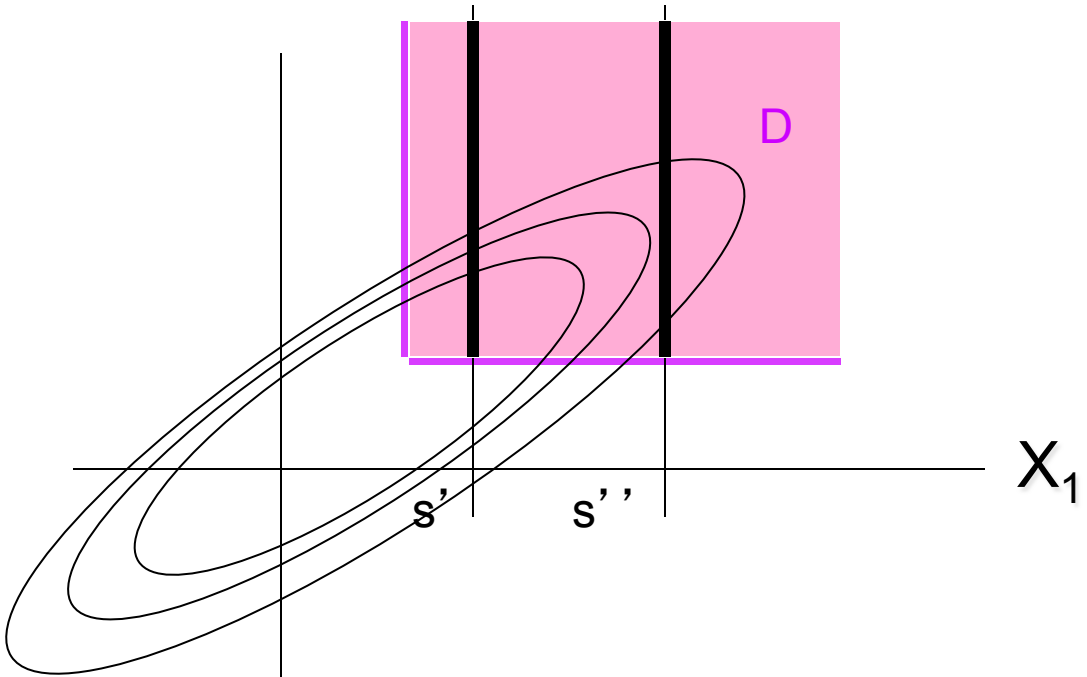
$$\approx \frac{m_0}{m} q \log(m)$$

YB&Hochberg (‘95). YB&Yekutieli (‘01)

Positive dependency

- Positive Regression Dependency on the subset of true null hypotheses :
- If the test statistics are $\mathbf{X}=(X_1,X_2,\ldots,X_m)$:
  - For any increasing set D, and H_{0i} true

    Prob($\mathbf{X}$ in D | $X_i=s$) is increasing in s



Some common properties

- If X is positive dependent in any of the above senses,

taking co-monotone (all increasing / all decreasing) transformations in all coordinates will leave the dependency unchanged.

So: If the test statistics X are PRDS on I_0,

So are the left-tailed p-values, via $F_0(X_i)$ for all i

Or right-tailed p-values, via $1-F_0(X_i)$

(but not the two-sided)

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)

Open problem for two-sided tests

Nevertheless, for the linear step-up procedure, Simulation results show that the worst case is when all test statistics have correlation 1. Can it be proved?

Under this assumption, the FDR can be analytically written, and it is shown that

$$FDR \le q m_0/m(1+\Sigma_{j=m0+1} (1/j)/2) \le q$$

Reiner '07, Rami '11

So still conservative.

More open problems about dependency

- If the test statistics are :
  - All Pairwise Comparisons: $\boldsymbol{x_i - x_j}$ $i,j=1,2,\ldots k$

$$FDR \le \frac{m_0}{m} q$$

even though correlations between pairs of comparisons are both + and -

Based on many simulation studies:

including Williams, Jones, & Tukey (‘94,’ 99);

And on theoretical analysis by Yekutieli (‘08)

Adaptive procedures

FDR control of the BH.

- Independent

$$FDR \le \frac{m_0}{m} q$$

- independent and continuous

$$FDR = \frac{m_0}{m} q$$

- Positive dependent

$$FDR \le \frac{m_0}{m} q$$

- General

$$FDR \le \frac{m_0}{m} q\,(1 + 1/2 + 1/3 + \ldots + 1/m)$$

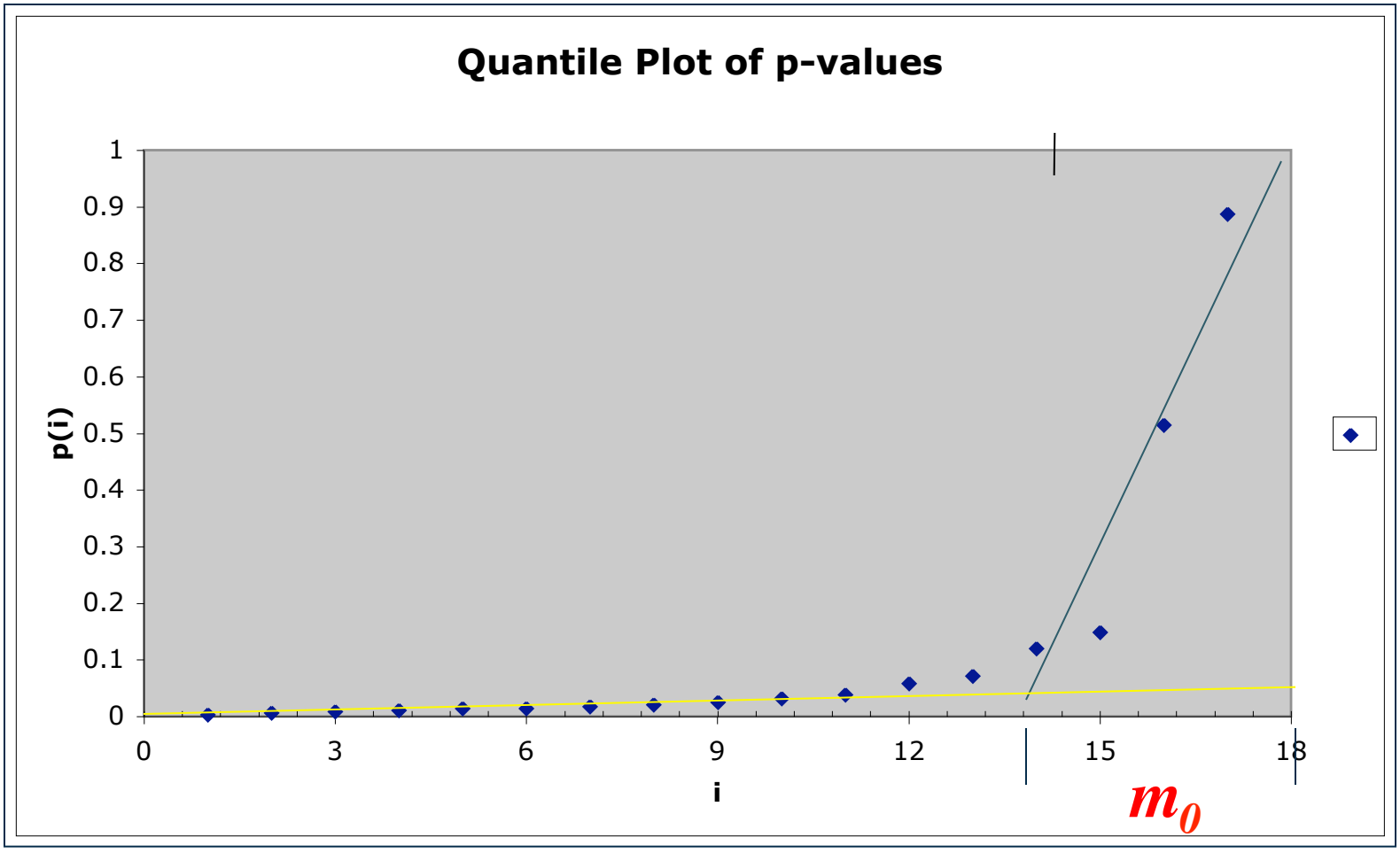
$$\approx \frac{m_0}{m} q \log(m)$$

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)

The graphical approach of Schweder & Spjotvoll



The adaptive FDR controlling procedure

YB and Hochberg (1989,2000)

- Use BH with q

  if nothing rejected stop
- Estimate m_0 =(m+1-k) / (1-$p_{(k)}$)
- Then use the Linear Stepup with **q*** = q **m/** m_0

FDR is controlled under independence (simulations)

Power is greatly increased

Option 1:Two stage procedure

YB, Yekutieli, Krieger(2006)

Stage I: Use the BH with $q/(1+q)$, rejecting r_1;

if $r_1=0$ stop

Stage II: Estimate $m_0 = (m- r_1)(1+q)$,

Then use it again with $q^*= q\, m/ m_0$

Proven FDR control under independence

Significance of 8 Strain differences

Behavioral Endpoint	.05/(1.05)(i/17)	$p_{(i)}$	.05i/[(1.05)*(17-8)]
Prop. Lingering Time	0.0028	0.0029	
# Progression segments	0. 0056	0.0068	
Median Turn Radius (scaled)	0.0084	0.0092	
Time away from wall	0.0112	0.0108	
Distance traveled	0.0140	0.0144	
Acceleration	0.0168	0.0146	
# Excursions	0.0196	0.0178	
Time to half max speed	0.0224	0.0204	
Max speed wall segments	0.0252	0.0257	
Median Turn rate	0.0280	0.0320	
Spatial spread	0.0308	0.0388	
Lingering mean speed	0.0336	0.0588	
Homebase occupancy	0.0364	0.0712	
# stops per excursion	0.0392	0.1202	
Stop diversity	0.0420	0.1489	
Length of progression segments	0.0448	0.5150	
Activity decrease	0.0476	0.8875	

Option 2:

Storey, Taylor, Siegmund ('04) modified Storey ('03):
Pre-determined λ ***(say λ=1/2)***:

into $m_0=(m+\mathbf{1}-\#(p_i \geq \lambda) \ / \ (1-\lambda)$
and added the condition for rejection $\mathbf{p_i \leq \lambda}$

- Proven FDR control under independence
  asymptotic control for weak dependency
- Most powerful under independence
- Fails to control FDR for all PRDS
  (for equally correlated FDR may double)

Significance of 8 Strain differences

Behavioral Endpoint	$p_{(i)}$	$p_{(i)}$	0.05 (i/6)
Prop. Lingering Time	0.0029	0.0029	
# Progression segments	0.0068	0.0068	
Median Turn Radius (scaled)	0.0092	0.0092	
Time away from wall	0.0108	0.0108	
Distance traveled	0.0144	0.0144	
Acceleration	0.0146	0.0146	
# Excursions	0.0178	0.0178	
Time to half max speed	0.0204	0.0204	
Max speed wall segments	0.0257	0.0257	
Median Turn rate	0.0320	0.0320	
Spatial spread	0.0388	0.0388	
Lingering mean speed	0.0588	0.0588	
Homebase occupancy	0.0712	0.0712	0.108
# stops per excursion	0.1202	0.1202	0.116
Stop diversity	0.1489	0.1489	0.125
Length of progression segments	0.5150	0.5150	0.133
Activity decrease	0.8875	0.8875	0.141

$2 = \#\{P_i \geq .5\}$

$m_0=(2+1)/(1-1/2)=6$

Why does dependency matter?

How do the bias and variance of $\hat{m}_0$ affect $E(m_0/\hat{m}_0)$?

Taylor expansion:
$E(m_0/\hat{m}_0) = 1 - bias/m_0 + bias^2/m_0^2 + variance/m_0^2.$

A surprising result
If one restricts Storey's estimator to $\lambda=q$ *(= say .05)*
performance under dependency improves dramatically

(Blanchard &Roquain '08,'09)

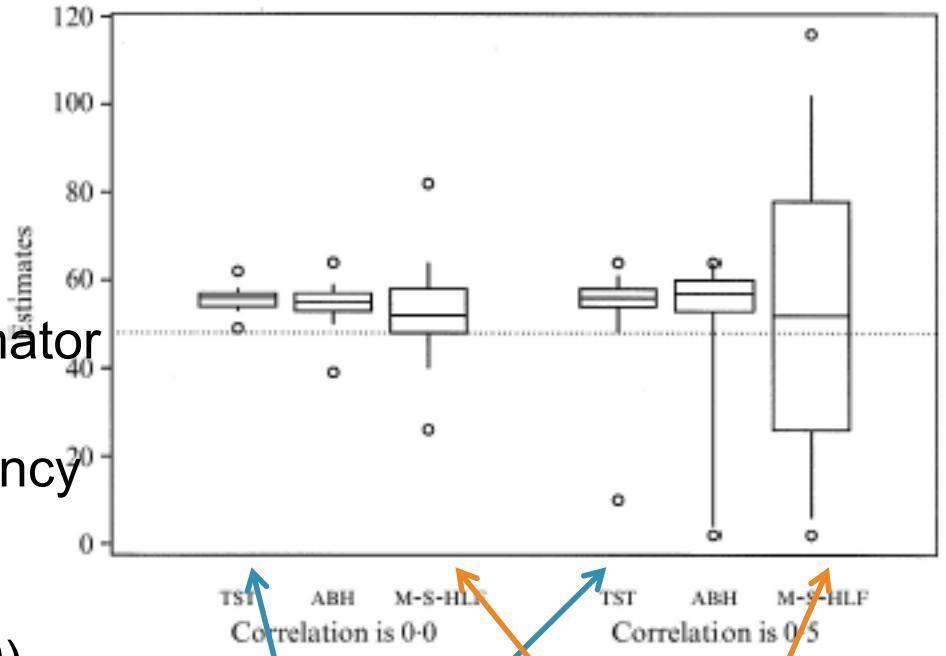


Fig. 3: The simulated distribution of the estimators $\hat{m}_0$ used in the TST, ABH and M-S-HLF adaptive procedures for estimating the number of true hypotheses with independent and positively correlated statistics for the case of $m_0 = 48$ and $m = 64$. Each box displays the median and quartiles as usual. The whiskers extend to the 5% and the 95% quantiles. The circles are located at the extremes, i.e. the 0·01% and 99·99% percentiles.

Two-Stage

M-Storey λ=1/2

Option 3: The step-down multi-stage procedure

Starting with m_0=m+1-1(1-q) -q Use m_{0i} = (m+1-i(1-q))
Namely

Let $k = \max\{i : \forall j \leq i \; p_{(j)} \leq \frac{qj}{m+1-j(1-q)}\}$.
If such a k exists, reject the k associated hypotheses; otherwise reject no hypothesis.

Gavrilov et al (2010) under independence
Finner et al (2010) showed its asymptotic optimality in *m*

Option 3: The step-down multi-stage procedure

Holm: Starting with $p_{(1)}$, Compare $p_{(i)} \leq \alpha/(m+1-i)$;

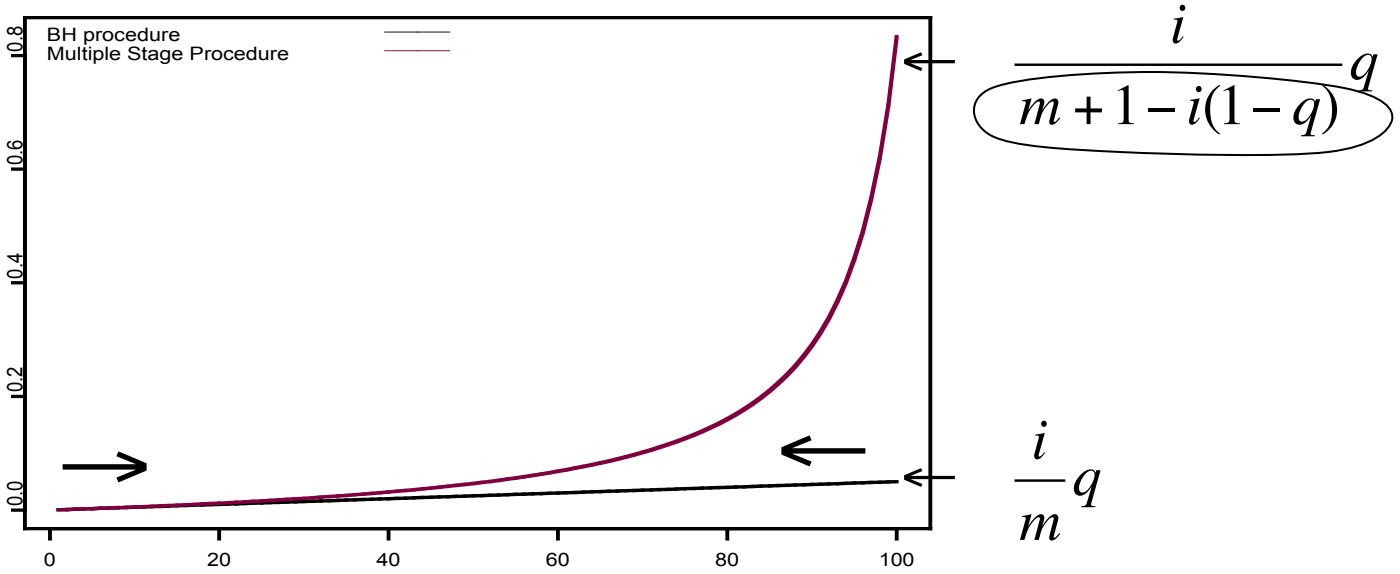
step to higher p-value reducing the size of the family by 1.
Stop with first non-rejection.

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.

Let $k = \max\{i : \forall j \leq i \; p_{(j)} \leq \frac{qj}{m+1-j(1-q)}\}$.
If such a k exists, reject the k associated hypotheses;
otherwise reject no hypothesis.

The step-down Multiple Stage procedure:



FDR controlling properties by Gavrilov et al (‘10)
Asymptotic Optimality Shown by Finner et al (‘10)

Bayesian and Empirical Bayes approaches

- Started with Tusher et al (2001) in the context of gene expression analysis . Thresholding significance at a
- Storey (2012) $pFDR(a)=E(V(a)/R(a) \mid R(a)>0)$

$$=FDR(a)/Pr(R(a)>0) \sim FDR$$

- Efron ('01),… until 'Large Scale Inference' Book ('10)

$$Fdr(a) =E(V(a))/E(R(a)) \sim FDR \sim pFDR$$

and the local FDR $fdr(x) = p_0 f_0(x)/f(x)$

$$= p_0 f_0(x)/ (p_0 f_0(x)+p_1 f_1(x))$$

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 ω_i , $\Sigma \omega_i = m$, reflecting
- (a) Its importance YB & Hochberg ('98)

$$wFDR = E(\Sigma \omega_i V_i) / (\Sigma \omega_i R_i))$$

it allows to assign monetary to decisions. Or,

- (b) The advantage it gets Genovese & Wasserman ('06)

$$p_i^* = p_i / \omega_i$$

FDR defined, and tested, as before

- Both are underutilized

FDR a thing of the past?