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*

Prolog

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

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

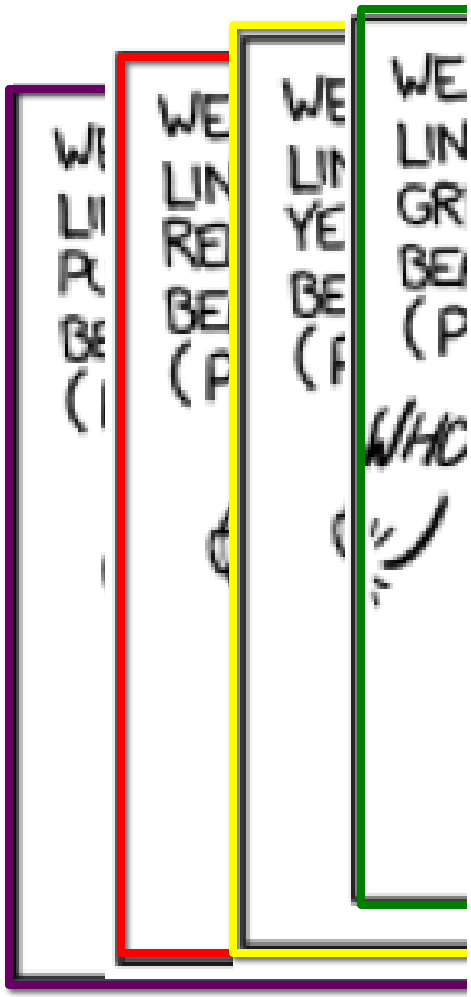
$$\text{Prob}(Z > |Y_{(k+)} - Y_{(1+)}| / \sigma_{\text{diff}}) < \alpha$$

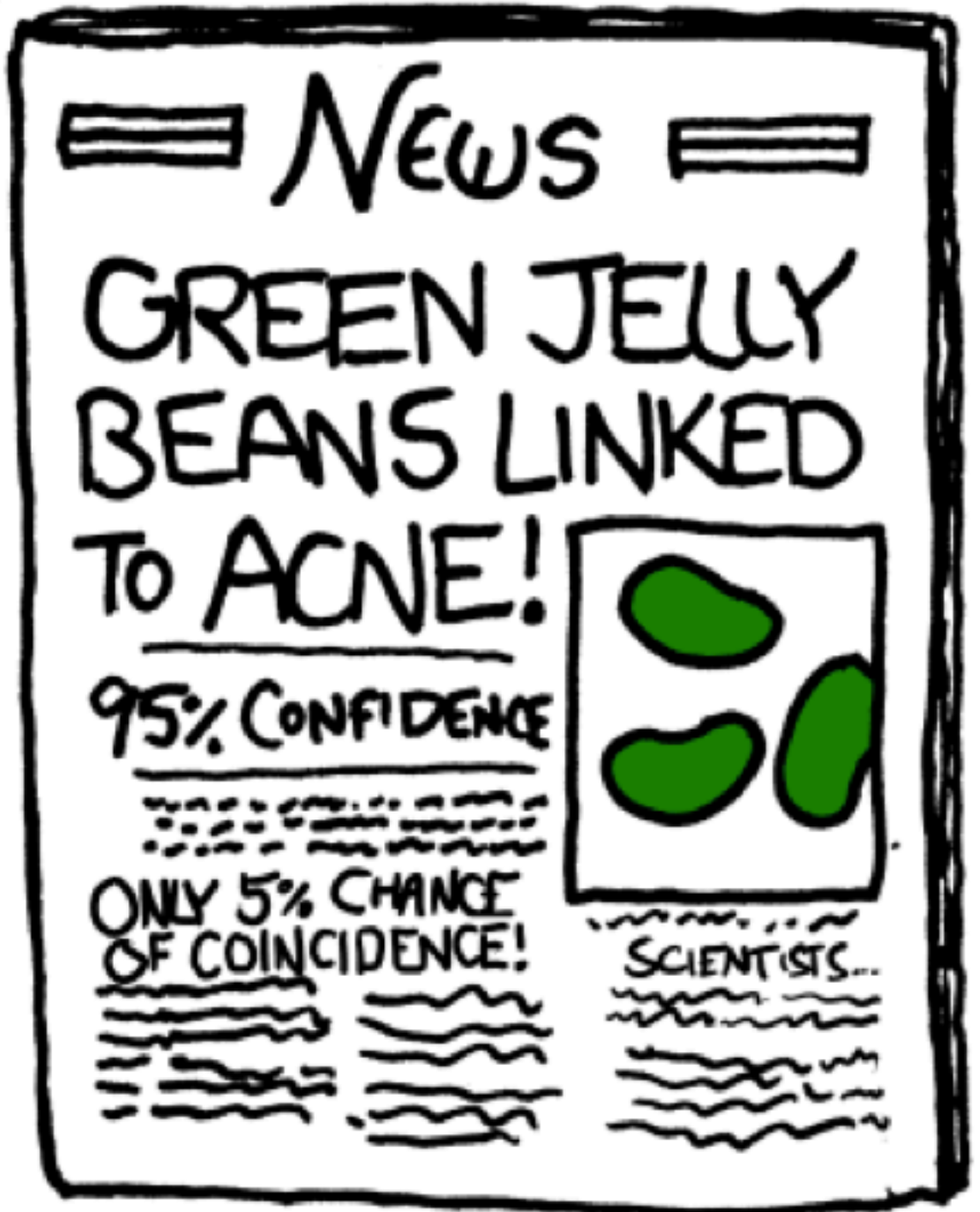
even if there is no difference. The larger k the worse it gets!

- 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.

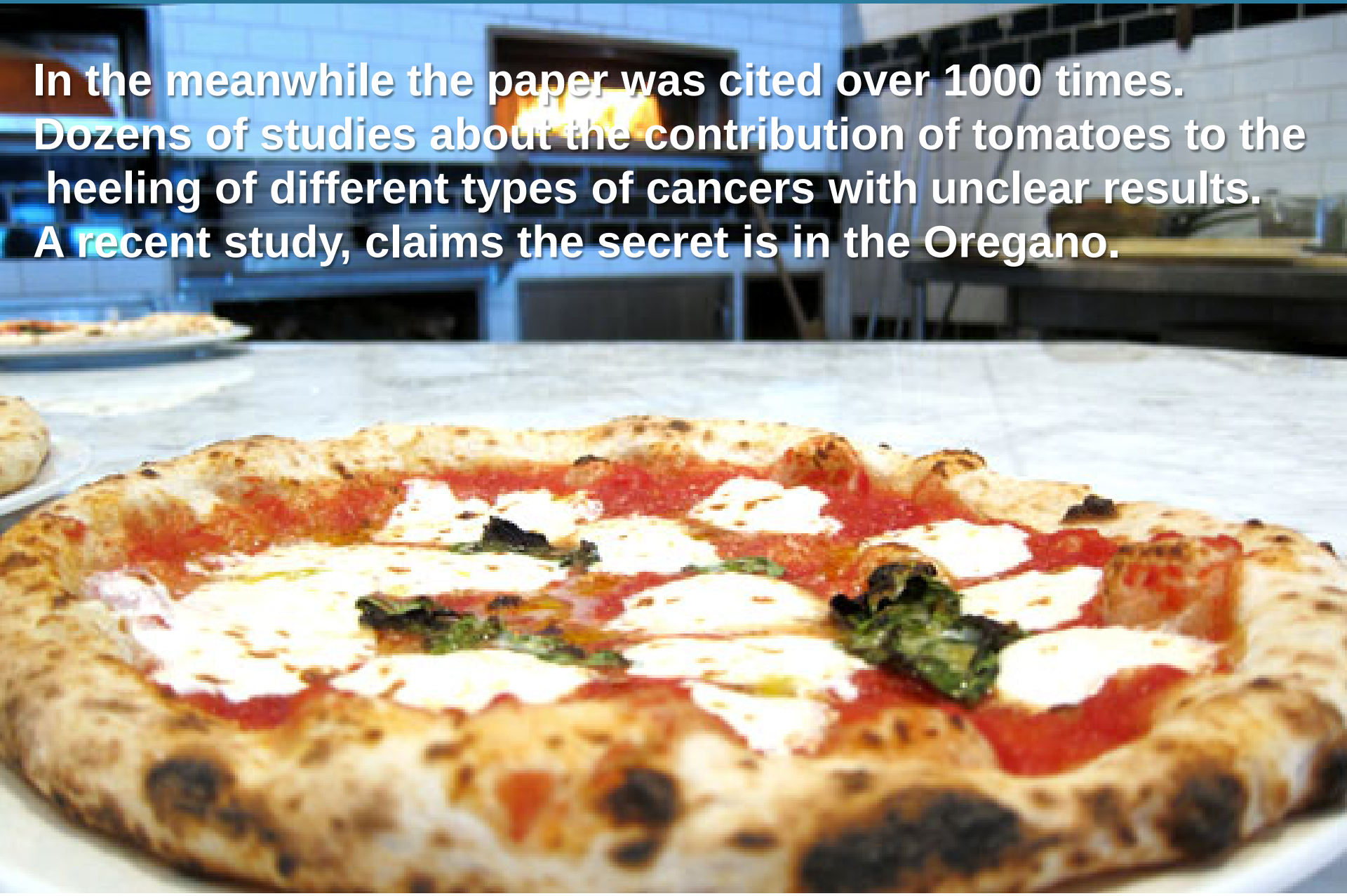
3 kinds of foods cleared the statistical significance bar –

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

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.

$$\textit{weak}\ \mathrm{FWER} = \Pr_{H_0}(V \geq 1).$$

FWER Protection

- 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 $E(V) \le \alpha$

(2) Tukey's procedure for pairwise comparisons:

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

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 $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$
  - 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.

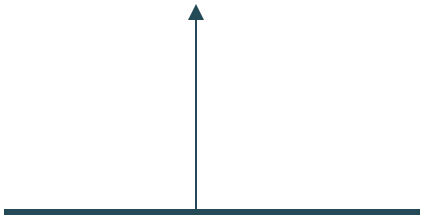
Always: $FWER \le \alpha$

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.

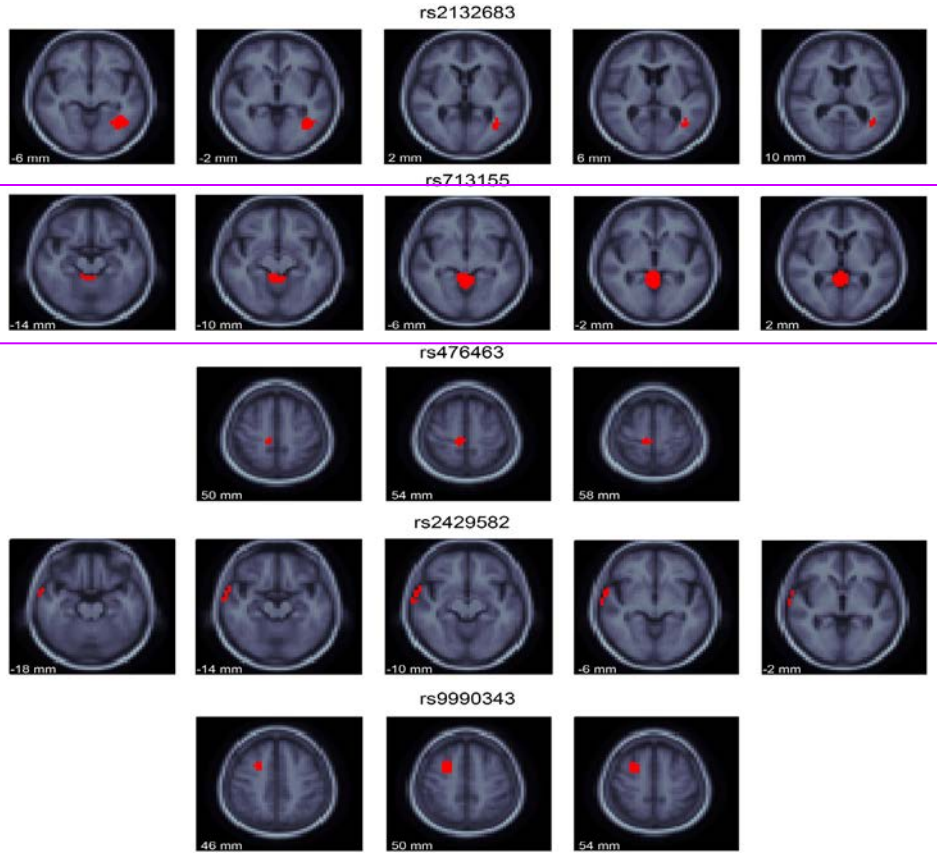
The increasing scale:

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, 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,

$$E(V/\text{number of tests performed}) = E(V/m) \leq \alpha$$

- FDR control assures

$$E(V/\text{number of tests selected}) = E(V/R) \leq \alpha$$

- 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 $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$
- Let

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

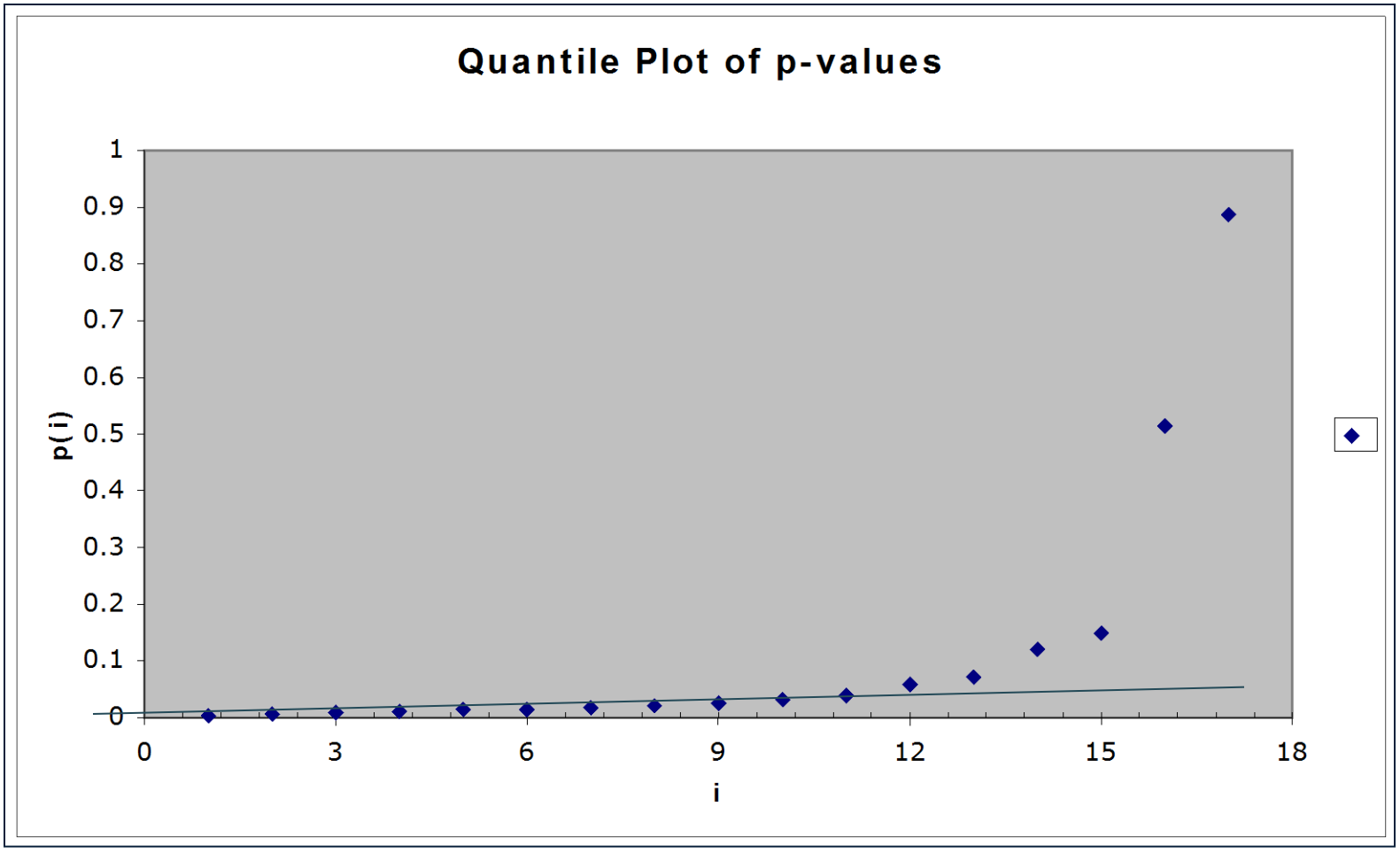
- Reject

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

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

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

- 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

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

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

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

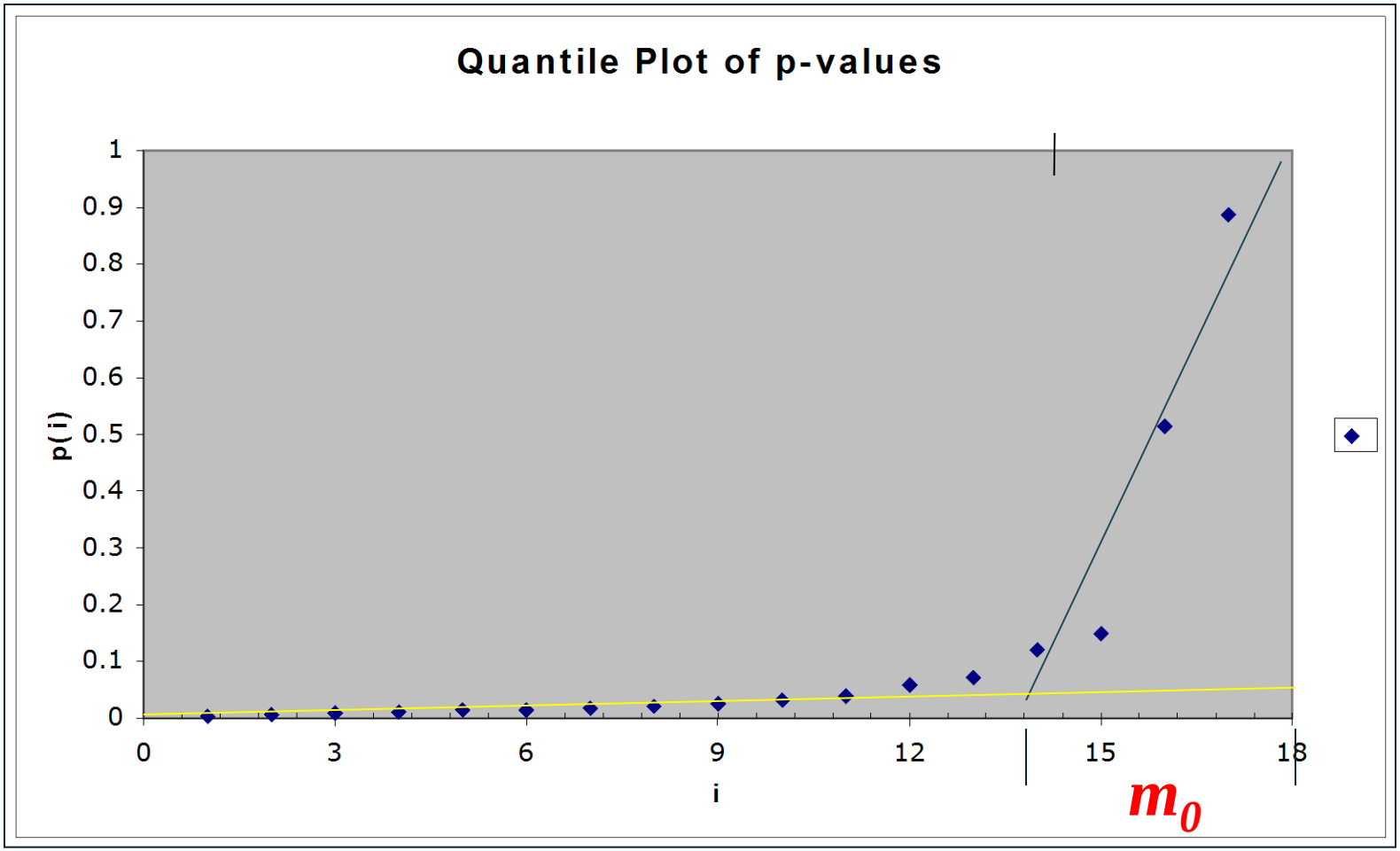
  (i.e. simulations, partial theoretical results)

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



Option 3: The step-down multi-stage procedure

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

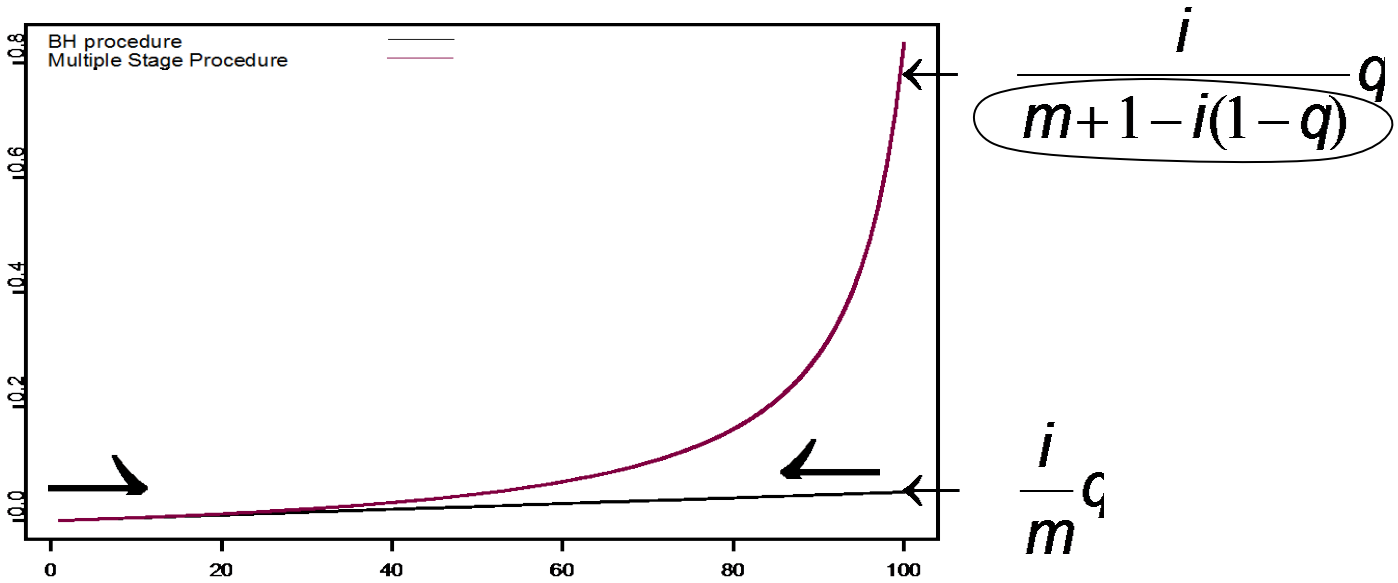
step to higher p-value reducing the size of the family by 1.
Stop with first non-rejection.

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



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

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(the marginal interval covers its parameter) = 0.95

or equivalently

Pr(the marginal interval fails to cover its parameter) = 0.05

- *With many such intervals*

Pr(some intervals fail to cover) > 0.05,

using Simultaneous CIs, (e.g. Bonferroni), assures ≤0.05

- Why bother? On the average over all parameters,

the expected proportion of intervals failing to cover ≤ *0.05* .

20 parameters to be estimated with 90% CIs

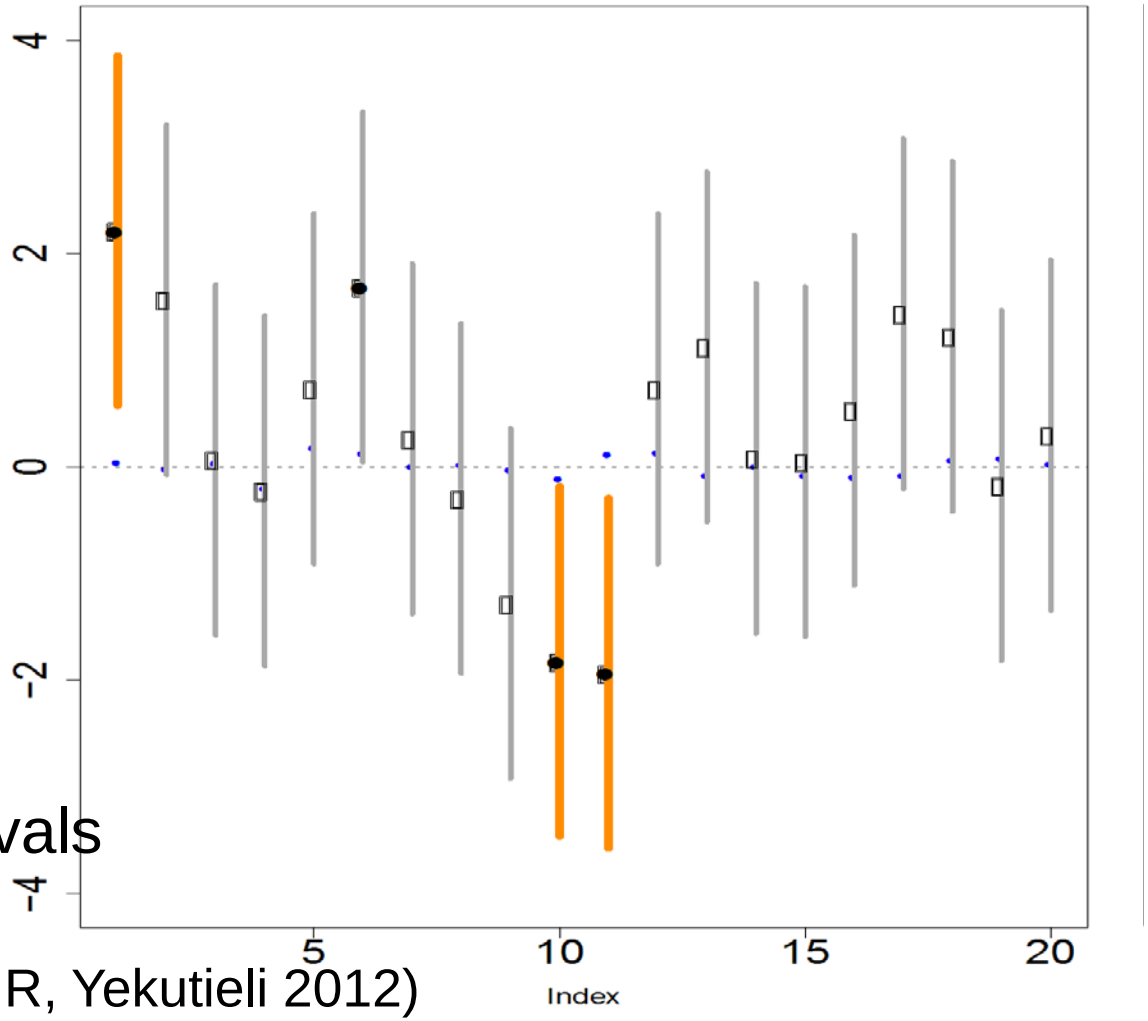
3/20 do not cover

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)

YB & Yekutieli ('05)

A selective CIs procedure uses the data T

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

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

$$FCR = E\left(\frac{\sum_{i \in S} \chi_{\{\theta_i \notin CI_i)\}}}{|S|}\right)$$

(|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 $i \in S(\boldsymbol{T})$,

construct a marginal *1- q* ***\|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 ≤ *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* <-> 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

Eleftheria Zeggini,[1,2]* Michael N. Weedon,[3,4]* Cecilia M. Lindgren,[1,2]* Timothy M. Frayling,[3,4]* Katherine S. Elliott,[2] Hana Lango,[3,4] Nicholas J. Timpson,[2,5] John R. B. Perry,[3,4] Nigel W. Rayner,[1,2] Rachel M. Freathy,[3,4] Jeffrey C. Barrett,[2] Beverley Shields,[4] Andrew P. Morris,[2] Sian Ellard,[4,6] Christopher J. Groves,[1] Lorna W. Harries,[4] Jonathan L. Marchini,[7] Katharine R. Owen,[1] Beatrice Knight,[4] Lon R. Cardon,[2] Mark Walker,[8] Graham A. Hitman,[9] Andrew D. Morris,[10] Alex S. F. Doney,[10] The Wellcome Trust Case Control Consortium (WTCCC),† Mark I. McCarthy,[1,2]‡§ Andrew T. Hattersley[3,4]‡

SCIENCE, 1 JUNE 2007

Main Table

rs	chr	position	A1	A2	Region	WTCCC 1924 cases 2938 controls OR (95% CI)	P_{add}	Replication meta-analysis 3757 cases 5346 controls OR (95% CI)	P_{add}	All UK sample meta-analysis 5681 cases 8284 controls OR (95% CI)	P_{add}	DGI 6529 cases 7252 controls OR (95% CI)	P_{add}	FUSION 2376 cases 2432 controls OR (95% CI)	P_{add}	All combined 14,586 cases 17,968 controls OR (95% CI)	P_{add}
rs8050136	16	52373776	A	C	*FTO*	1.27 (1.16–1.37)	2.0×10^{-8}	1.22 (1.12–1.32)	5.4×10^{-7}	1.23 (1.18–1.32)	7.3×10^{-14}	1.03 (0.91–1.17)	0.25	1.11 (1.02–1.20)	0.017	1.17 (1.12–1.22)	1.3×10^{-12}
rs10946398	6	20769013	A	C	*CDKAL1*	1.20 (1.10–1.31)	2.5×10^{-5}	1.14 (1.07–1.22)	8.3×10^{-5}	1.16 (1.10–1.22)	1.3×10^{-8}	1.08 (1.03–1.14)	2.4×10^{-3}	1.12 (1.03–1.22)	9.5×10^{-3}	1.12 (1.08–1.16)	4.1×10^{-11}
rs5015480	10	94455539	C	T	*HHEX*	1.22 (1.12–1.33)	5.4×10^{-6}	–	–	1.13 (1.07–1.19)	4.6×10^{-6}	1.14 (1.06–1.22)	1.7×10^{-4}	1.10 (1.01–1.19)	0.025	1.13 (1.08–1.17)	5.7×10^{-10}
rs1111875	10	94452862	C	T	*HHEX*	–	–	1.08 (1.01–1.15)	0.020								
rs10811661	9	22124094	C	T	*CDKN2B*	1.22 (1.09–1.37)	7.6×10^{-4}	1.18 (1.08–1.28)	1.7×10^{-4}	1.19 (1.11–1.28)	4.9×10^{-7}	1.20 (1.12–1.28)	5.4×10^{-8}	1.20 (1.07–1.36)	2.2×10^{-3}	1.20 (1.14–1.25)	7.8×10^{-15}
rs564398	9	22019547	C	T	*CDKN2B*	1.16 (1.07–1.27)	3.2×10^{-4}	1.12 (1.05–1.19)	8.6×10^{-4}	1.13 (1.08–1.19)	1.3×10^{-6}	1.05 (0.94–1.17)	0.5	1.13 (1.01–1.27)	0.039	1.12 (1.07–1.17)	1.2×10^{-7}
rs4402960	3	186994389	G	T	*IGF2BP2*	1.15 (1.05–1.25)	1.7×10^{-3}	1.09 (1.01–1.16)	0.018	1.11 (1.05–1.16)	1.6×10^{-4}	1.17 (1.11–1.23)	1.7×10^{-9}	1.18 (1.08–1.28)	2.4×10^{-4}	1.14 (1.11–1.18)	8.6×10^{-16}
rs13266634	8	118253964	C	T	*SLC30A8*	1.12 (1.02–1.23)	0.020	1.12 (1.04–1.19)	1.2×10^{-3}	1.12 (1.05–1.18)	7.0×10^{-5}	1.07 (1.00–1.16)	0.047	1.18 (1.09–1.29)	7.0×10^{-5}	1.12 (1.07–1.16)	5.3×10^{-8}
rs7901695	10	114744078	C	T	*TCF7L2*	1.37 (1.25–1.49)	6.7×10^{-13}	–	–	–	–	1.38 (1.31–1.46)	2.3×10^{-31}	1.34 (1.21–1.49)	1.4×10^{-8}	1.37 (1.31–1.43)	1.0×10^{-48}
rs5215	11	17365206	C	T	*KCNJ11*	1.15 (1.05–1.25)	1.3×10^{-3}	–	–	–	–	1.15 (1.09–1.21)	1.0×10^{-7}	1.11 (1.02–1.20)	0.014	1.14 (1.10–1.19)	5.0×10^{-11}
rs1801282	3	12368125	C	G	*PPARG*	1.23 (1.09–1.41)	1.3×10^{-3}	–	–	–	–	1.09 (1.01–1.16)	0.019	1.20 (1.07–1.33)	1.4×10^{-3}	1.14 (1.08–1.20)	1.7×10^{-6}

Odds ratio point and CI estimates for confirmed T2D susceptibility variants

Region	Odds ratio	0.95 CIs
• FTO	1.17	[1.12, 1.22]
• CDKAL1	1.12	[1.08, 1.16]
• HHEX	1.13	[1.08, 1.17]
• CDKN2B	1.20	[1.14, 1.25]
• CDKN2B	1.12	[1.07, 1.17]
• IGF2BP2	1.14	[1.11, 1.18]
• SLC30A8	1.12	[1.07, 1.16]
• TCF7L2	1.37	[1.31, 1.43]
• KCNJ11	1.14	[1.10, 1.19]
• PPARG	1.14	[1.08, 1.20]

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

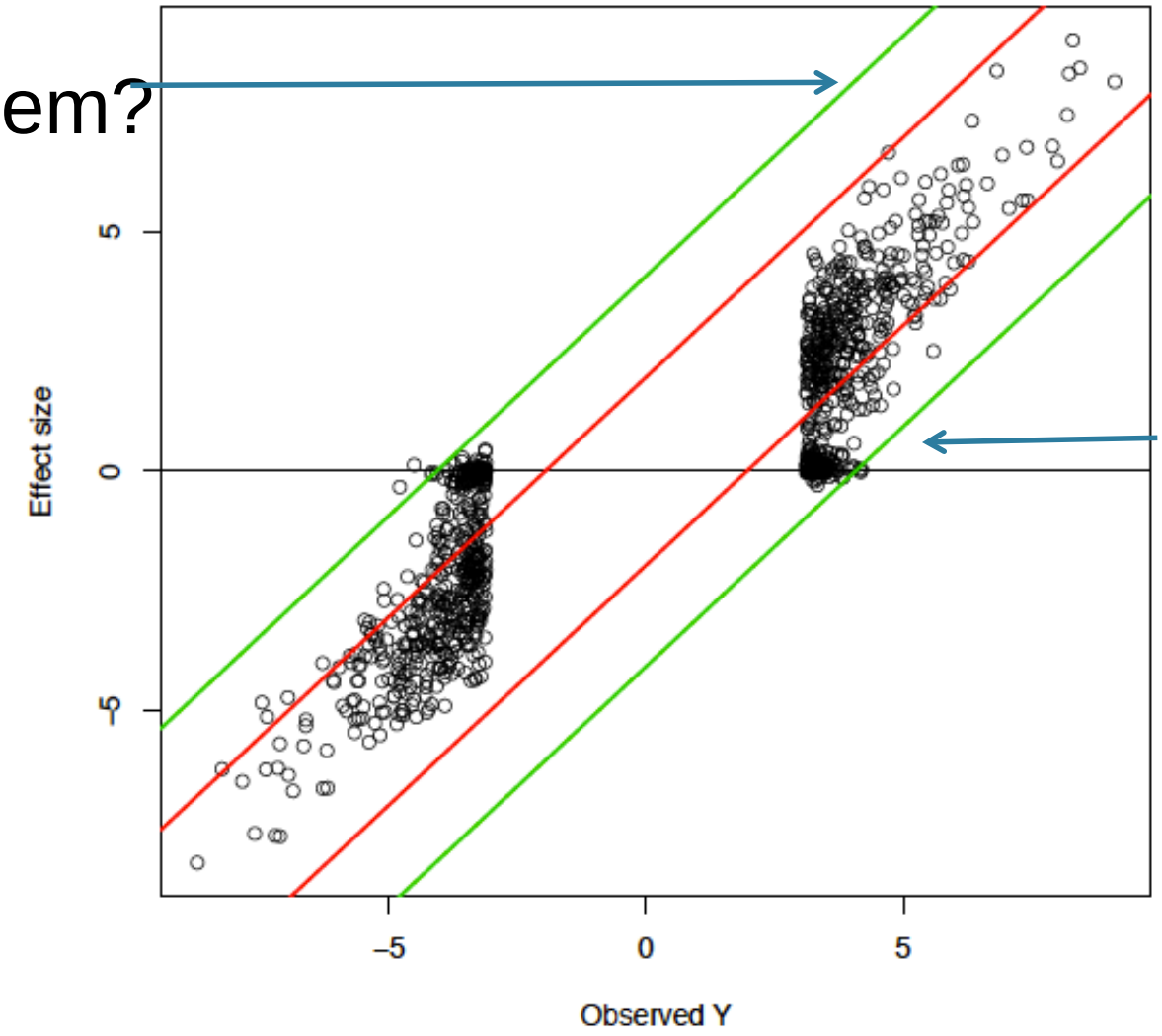
1-.05*10/400000

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

and FCR adjusted intervals

How well do we do?

Problem?



Success

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 θ_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

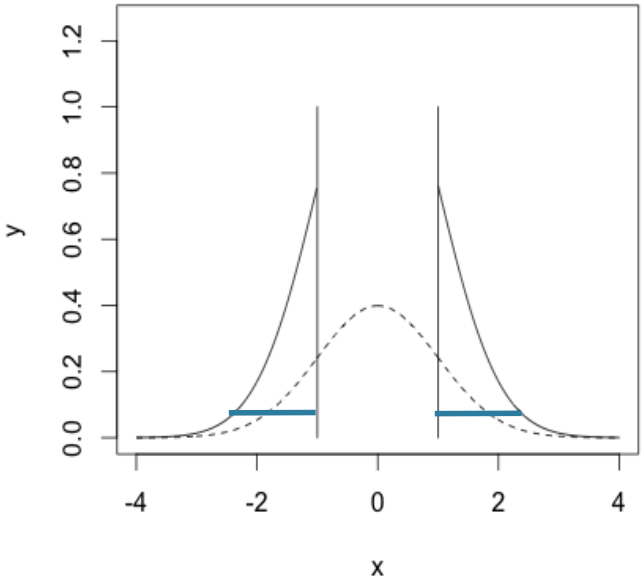
Utilize the nature of the selection process being used in order to improve selective inference CIs and tests

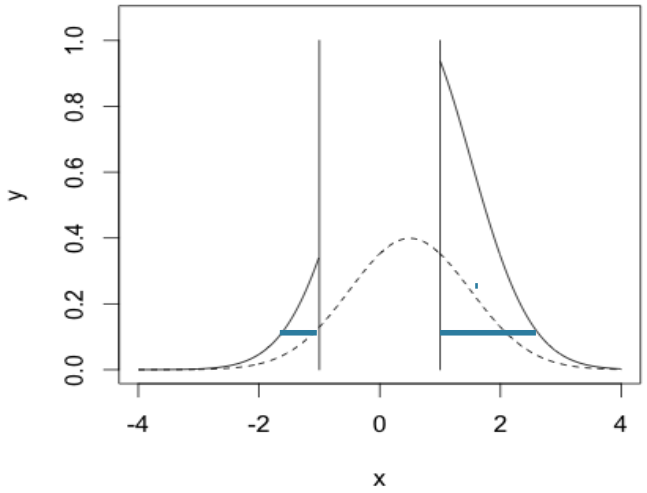
In particular selection of θ if its estimator is big enough

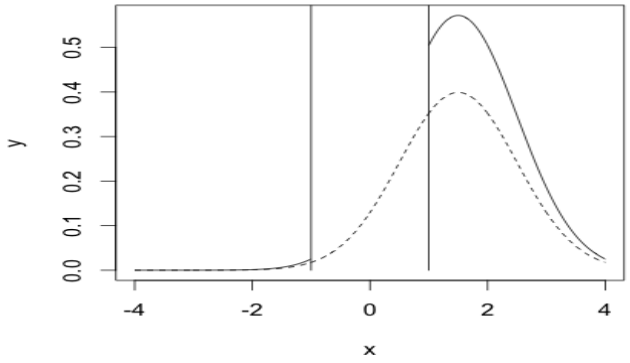
$$X=(Y \mid |Y| \geq c),$$

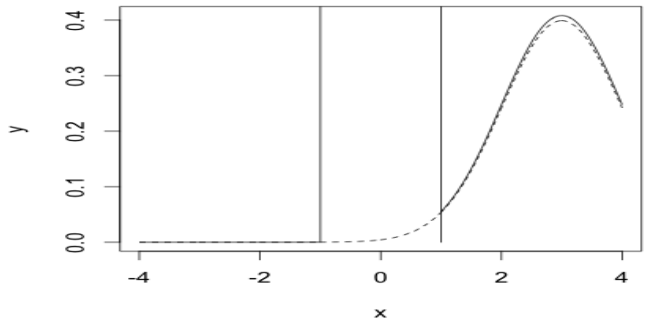
where c is either fixed or (simple) data dependent.

Weinstein, Fithian, YB (‘13)









The complication: θ 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 θ_0

Invert them to get conditional CIs.

*following YB,Hochberg & Stark ('98)

The intervals will also control the False Coverage Rate

Example

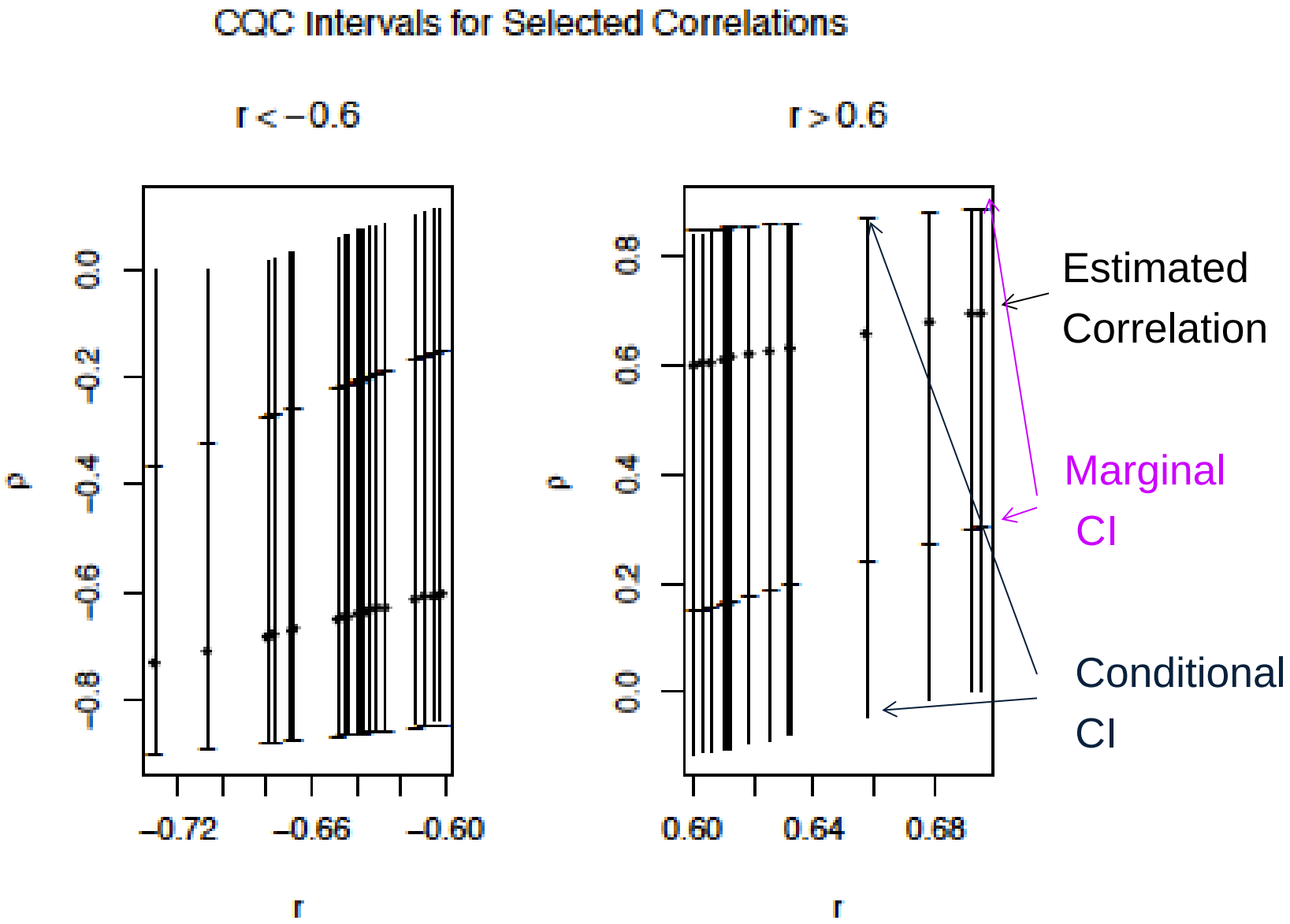
16 Subjects view 2 movie segments of different stress level. Recordings was made of:

- 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 $r \leq -0.6$.



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

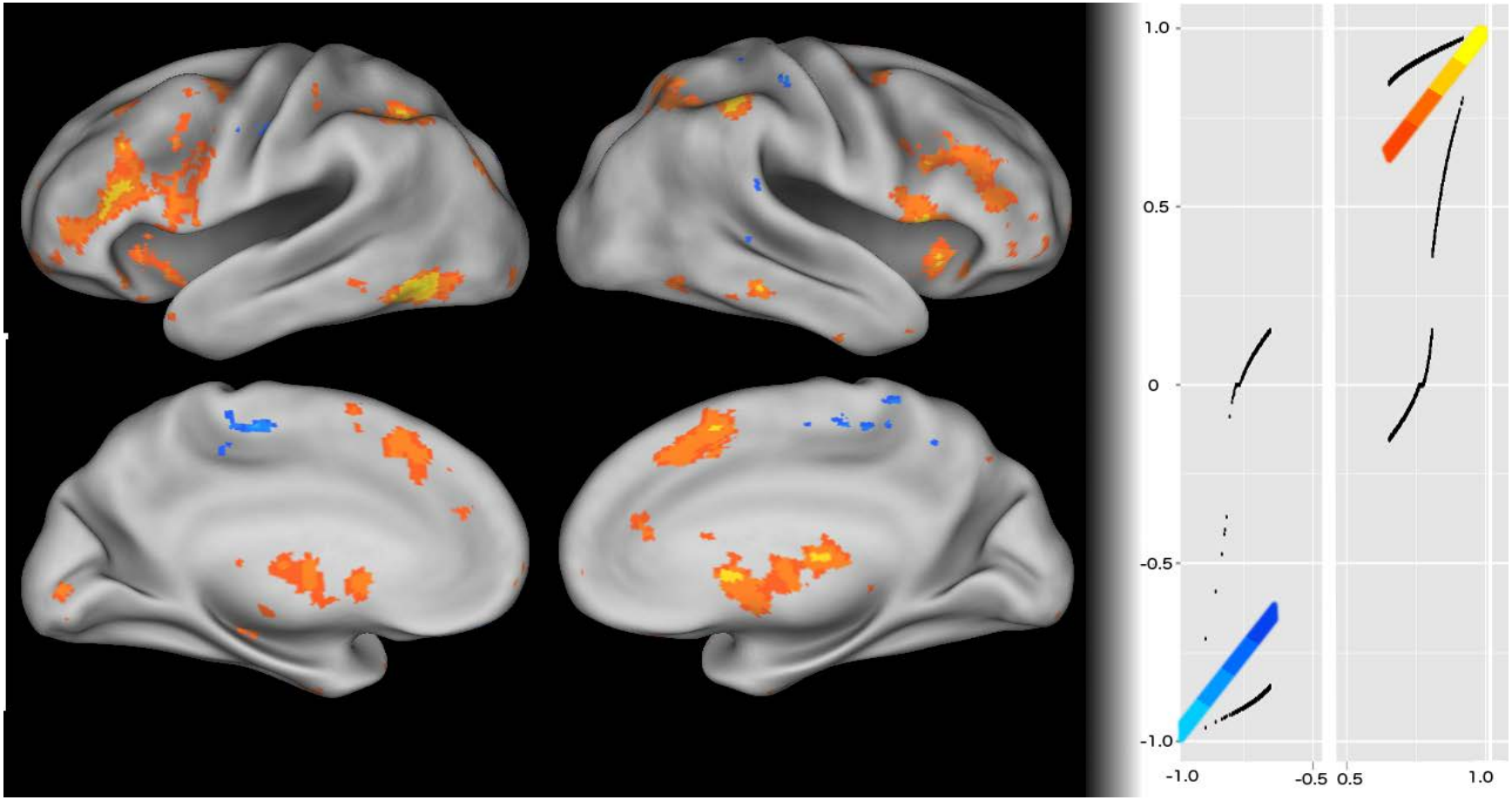
(named also ‘Circular reasoning’, ‘Double Dipping’)

and that:

Voodoo correlations are everywhere…

Their proposed solution: data splitting

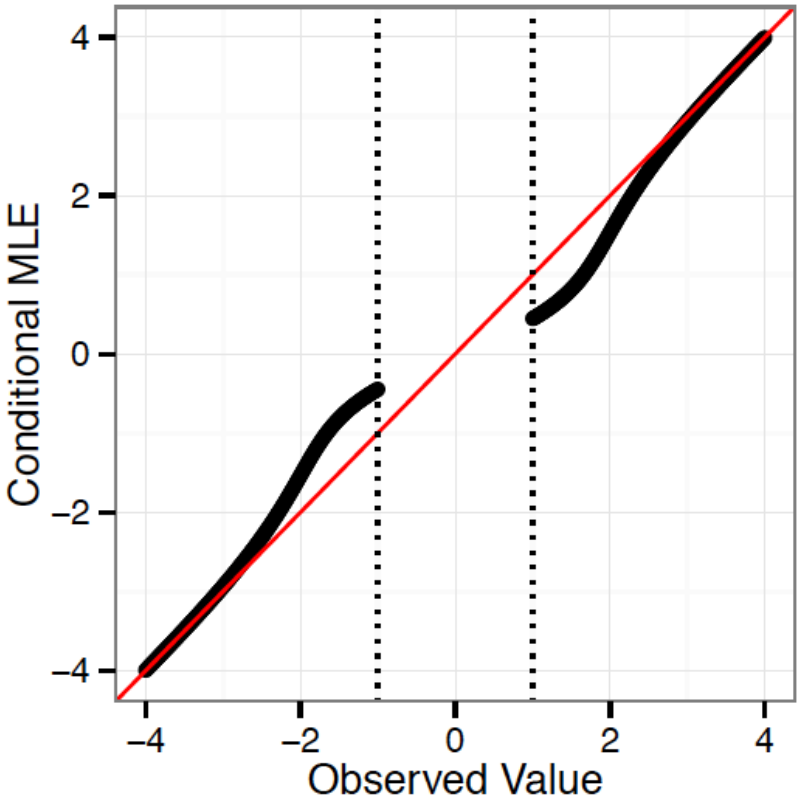
Addressing in-study ‘voodoo correlations’

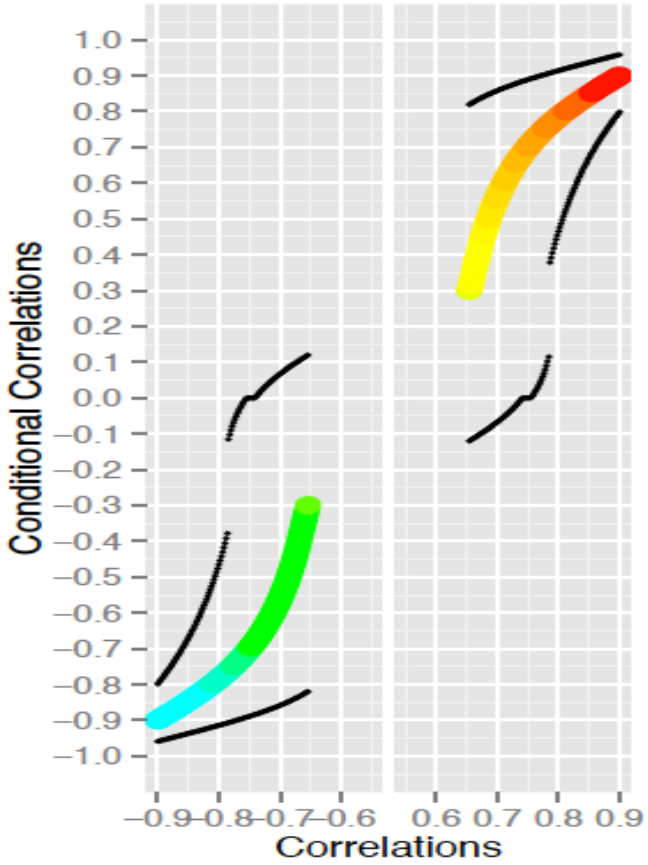


Confidence Calibration Plot: Observed correlations in significant voxels (B-H;FDR 0.1) encoding conditional confidence intervals as well.

Rosenblatt &YB ‘14+

Maximum conditional likelihood estimator





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 : $y_i=\mu_i+e_i$, $e_i \sim N(0,\sigma^2)$ $i=1,2,\ldots m$ *ind.*

The idea: For the prediction of linear function of μ_i

Screen: Threshold small coefficients

If $\mu_i^2 \le \sigma^2$ *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

$$p_{(k)} \le qk/m \quad <=> \quad |Y_i| \ge \sigma Z_{qk/2m}$$

- Estimate using

$$Y_i^{FDR} = 0 \text{ if } |Y_i| < \sigma Z_{qk/2m} \quad \text{(ignore)}$$

$$= Y_i \text{ if } |Y_i| \ge \sigma Z_{qk/2m} \quad \text{(report)}$$

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 $0 < r \leq 2$:

$$\Sigma(error)^2; \quad \Sigma|error|, \quad \#(errors)$$

relative to best "oracle" performance

Let #(parameters) -> infinity

Consider bodies of sparse signals such as:

- prop(non-zero coefficients) -> 0 (i.e. $p_0(m) \to 1$),
- 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

$$|\mu|_{(i)} \le C\, i^{-1/p} \qquad \textit{for all } i,\ p < 2$$

(if small μ_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:

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

m #number of parameters searched

k #number of parameters in current model

$$RSS(k) + \lambda_{k,m} k\sigma^2 =$$

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

An FDR testing based penalty:

$$RSS(k) + \lambda_{k,m} k\sigma^2 = RSS(k) + (\frac{1}{k} \sum_{i=1}^{k} Z^2_{\frac{iq}{2m}}) k\sigma^2$$

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)

Method	N.	Variables in the model	R^2
MS (at .05) FSR	*7*	BMI,S5,BP,AGE*SEX,BMI*BP, S3, SEX	*.53*
LARS	*16*	BMI, S5, BP, S3,BMI*BP, AGE*SEX, $S6^2$, BMI^2, AGE*BP, AGE*S6, SEX, S6, AGE*S5,AGE^2, SEX*BP, BP*S3	*.55*

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,

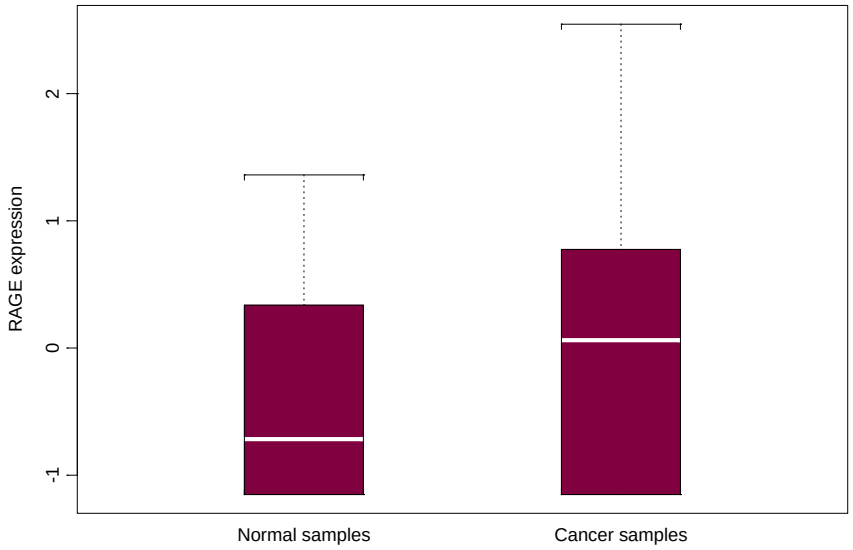
the top of which should be a “good” set of classifiers.

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

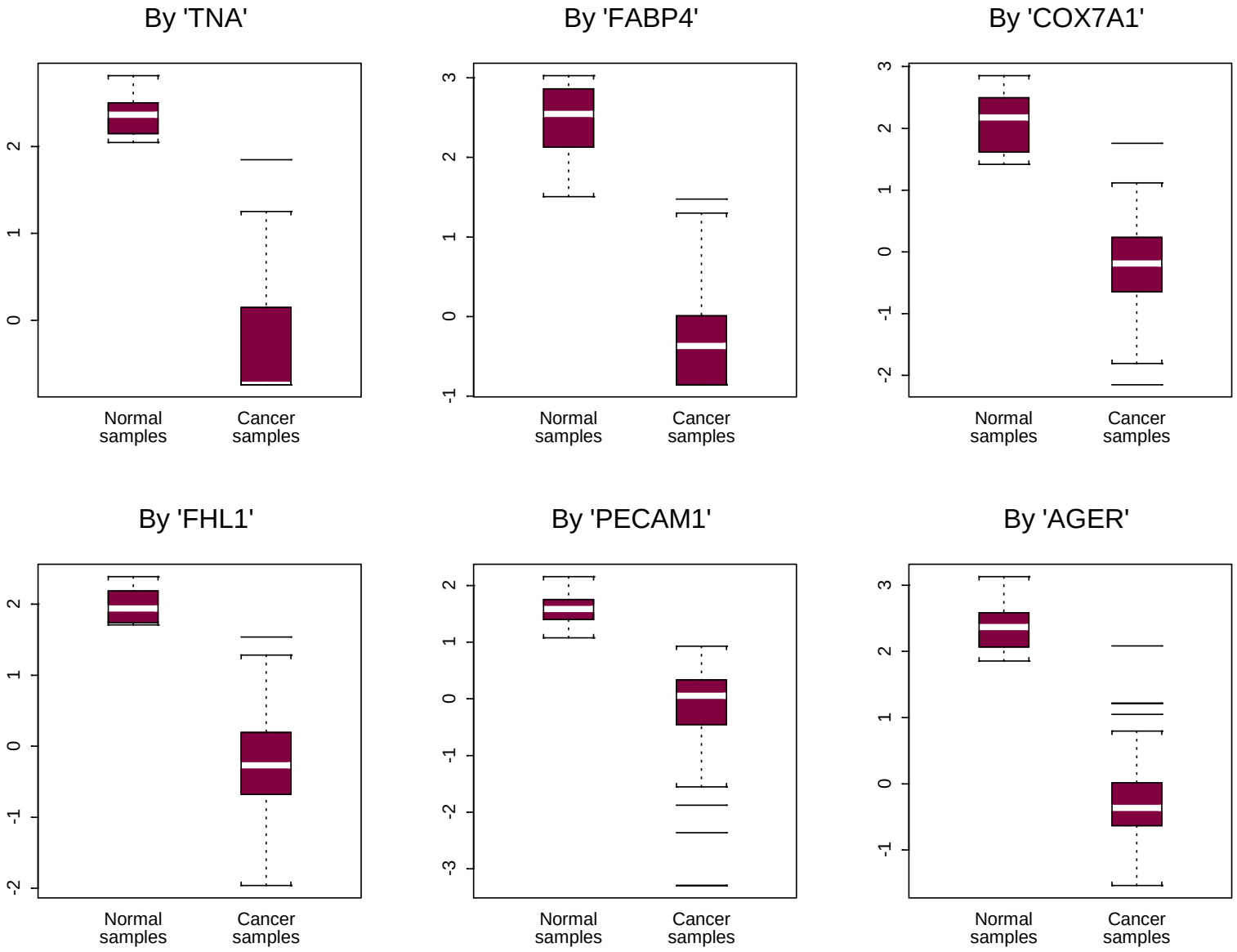
•Choose by forward (greedy) selection the features to enter the logistic model in order to minimizes the deviance plus FDR penalty.

•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

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,

Asymptotic p-values — Lockhart, Taylor, Tibshirani, J. Tibshirani,

Sequential Testing — G'Sell, Wager, Chouldechova,Tibshirani

- The fourth introduces "sorted l_1" version of FDR penalty

  Bogdan, van den Berg, Su, Candes

$$\|Y-\hat{Y}\|^2+\sigma^2\sum_{i=1}^{k} Z^2_{\frac{iq}{2m}} \qquad \|Y-\hat{Y}\|^2+\sigma^2\sum_{i=1}^{k} |\beta_{(i)}|\cdot|Z_{\frac{iq}{2m}}|$$

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.

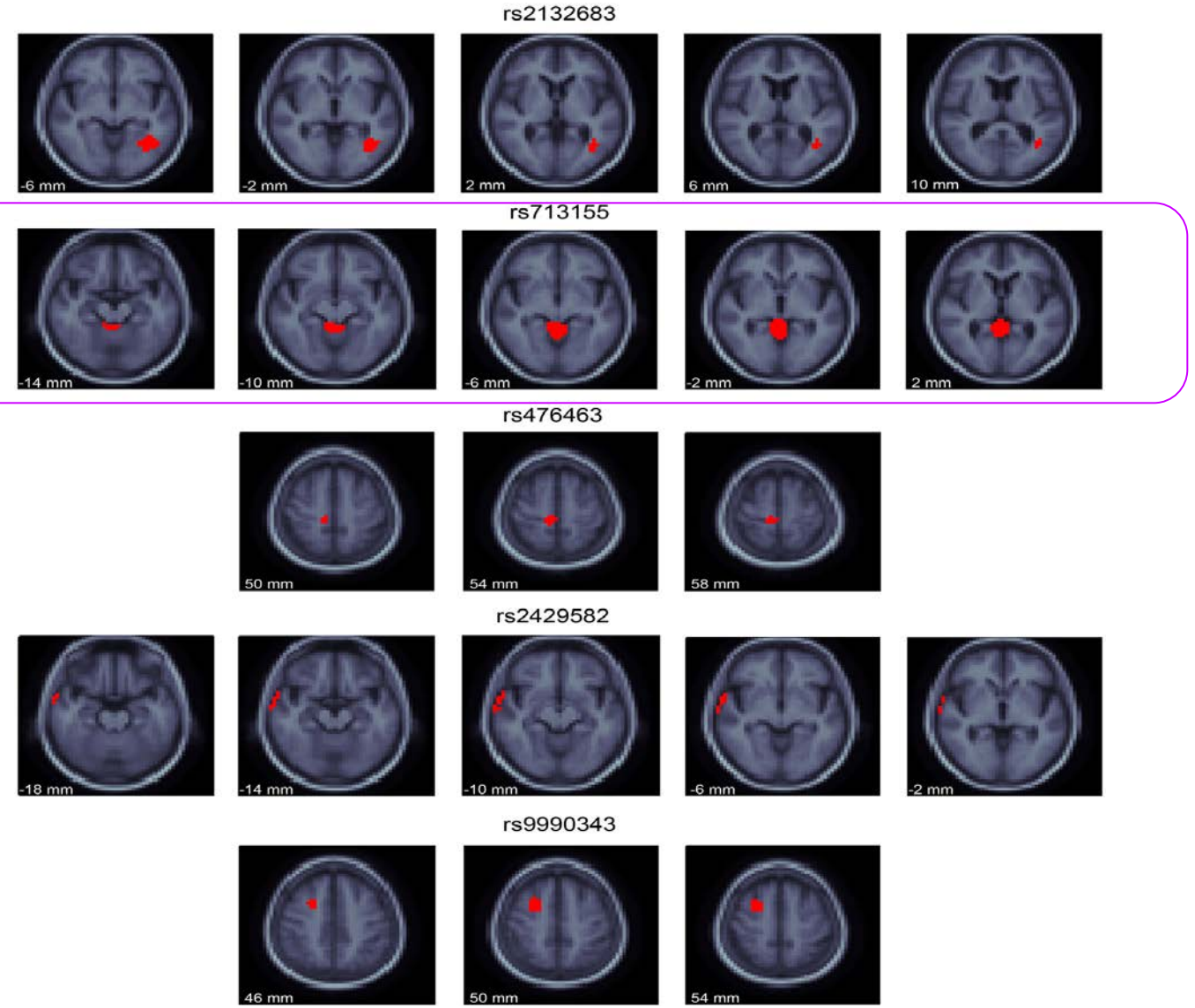
Testing selected families

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



S e p a r a t e vs joint FDR testing of families

Homogeneous case 50 families 10 hypotheses in each

$m_0/m \sim$ constant (< 1)

S e p a r a t e ~ Joint (scalability of the separate)

Heterogeneous case 50 families 10 hypotheses in each

$m_0/m=1$ for 49 families $m_0/m=0$ for 1 family

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

S e p a r a t e analysis: too liberal for the 49.

Overall FDR may reach .9

Efron's comment (2008)

Efron's comment (2008)

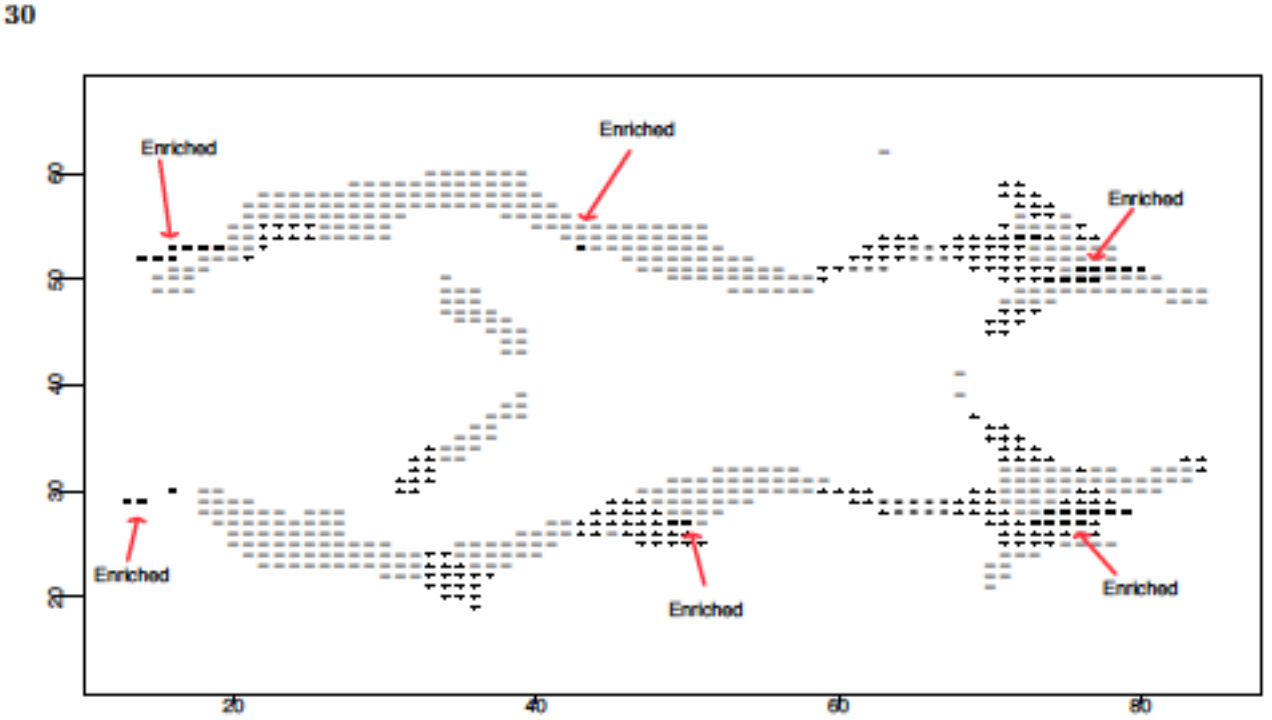


Fig 10: *Enrichment analysis of Imaging data, Panel D of Figure 1; z-value for original 15445 voxels have been averaged over "gene-sets" of neighboring voxels with city-block distance* ≤ 2. *Coded as "-" for* $\bar{z}_i < 0$, *"+" for* $\bar{z}_i \geq 0$; *solid rectangles, labeled as "Enriched", show voxels with* $\widehat{\text{fdr}}(\bar{z}_i) \leq 0.2$, *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!

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

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 $E(\mathcal{C})$ (FDR, but also FWER, and others)

on the average over the selected families means

$$E\left[\frac{\sum_{i\in S(P)} C_i}{|S(P)|}\right] \le q$$

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,

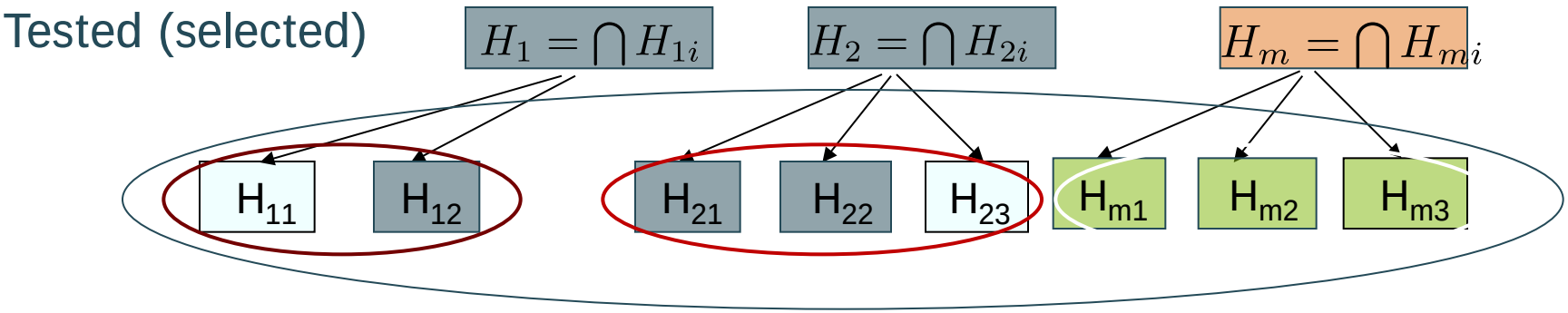
assures control on the average over the selected at level q

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

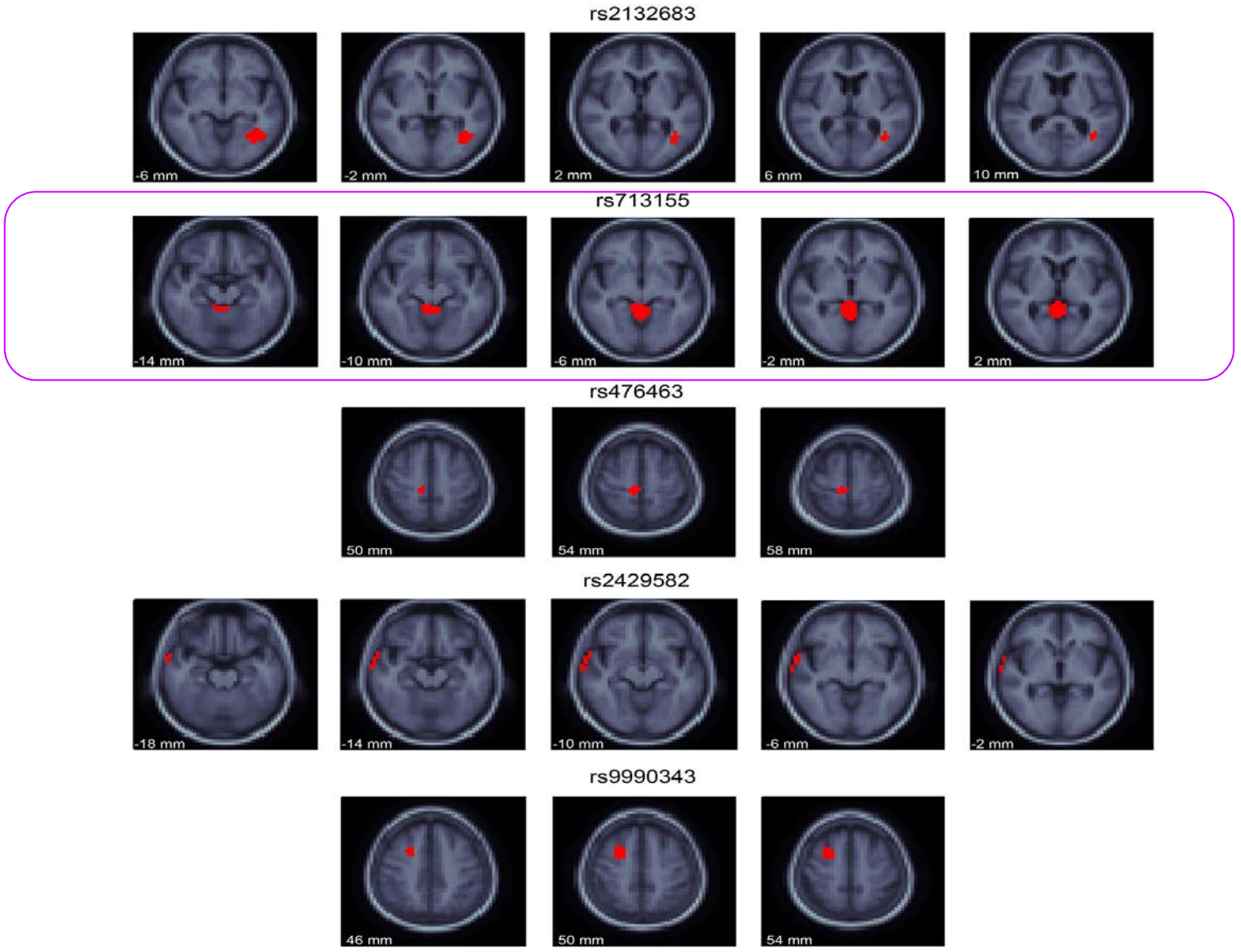
If we use the BH adjusted p-value to test the intersection of each family

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 Alzhiemer



Re-analysis of SNP-voxel data for Alzhiemer

- 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)

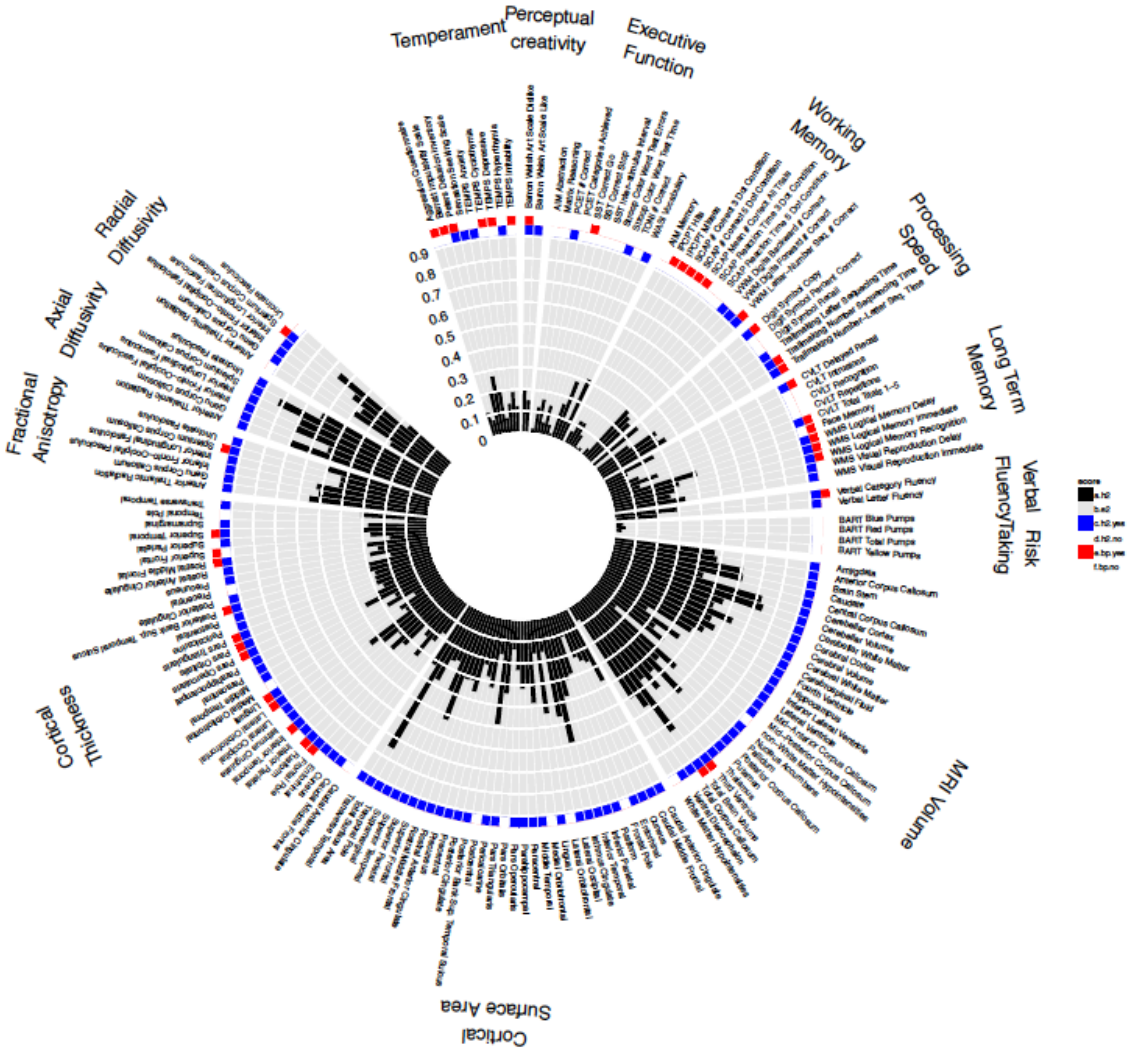
- Current work in Brain research: generalize the methodology to 3 or more levels

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): α-investment to control at level q the

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

Hypotheses arrive sequentially; in study i, test H_i with α_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
- Benjamini, Heller & Yekutieli (2009) ``Selective inference in complex research'' *Philosoph. Trans. of the Roy. Soc. A*.
- Donoho & Jin (2004) Higher criticism for detecting sparse heterogeneous mixtures , Annals of Statistics.
- Giovannucci et al. (1995) INTAKE OF CAROTENOIDS AND RETINOL IN RELATION TO RISK OF PROSTATE-CANCER. Journ. National Cancer Inst.
- Williams, Johns & Tukey (1999) Journal of Educational and Behavioral Statistics

References

- Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. M. (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann. Statist., 34, 584–653.
- Benjamini, Y. and Gavrilov, Y. (2009) A simple forward selection procedure based on false discovery rate control. Ann. Appl. Statist., 3, 179–198.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. Ann. Statist. 32 407–499. MR2060166
- FINNER, H., DICKHAUS, T. and ROTERS, M. (2009). On the false discovery rate and asymptotically optimal rejection curve. Ann. Statist.
- Benjamini,Y. andYekutieli,Y. (2005) False discovery rate controlling confidence intervals for selected parameters. J. Am. Statist. Ass., 100, 71–80.
- Benjamini, Y., Hochberg, Y., and Stark, P. B. (1998), “Confidence Intervals With More Power to Determine the Sign: Two Ends Constrain the Means,” Journal of the American Statistical Association, 93, 309–317.Donoho, D. and Jin, J. S. (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist., 32,962–994.

References

- Gavrilov, Y., Benjamini, Y. and Sarkar, S. (2009) An adaptive step-down procedure with proven FDR control under independence. Ann. Statist., 37, 619–629
- SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. Ann. Statist. 30 239–257. MR1892663)
- Weinstein, Fithian & Benjamini (2013) Selection Adjusted Confidence Intervals With More Power to Determine the Sign JASA http://dx.doi.org/10.1080/01621459.2012.737740
- Zeggini et al (2007) Replication of Genomwide association signals.. risk for type II diabetics. Science