A review of challenges in high dimensional multiple inferences

Yoav Benjamini
Tel Aviv University, Israel

Mathematical Methods in Modern Statistics
Luminy, 2017

www.replicability.tau.ac.il

Collaborative research with many

Ruth Heller, Dani Yekutieli, Tzviel Frostig, Tel Aviv U
Marina Bogomolov, *Technion*
Jonathan Rosenblatt, *Ben Gurion U*

Philip Stark, Will Fithian, *UC Berkeley*
Chiara Sabatti, Assaf Weinstein, *Stanford*
Yotam Hechtlinger, Carnegie Mellon
Christine Peterson, AMC, Texas

ERC – Advanced Research Grant

Practical Statistical Approaches to Replicability Problems in Life Sciences (PSARPS)

The Replicability Problems in Science

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

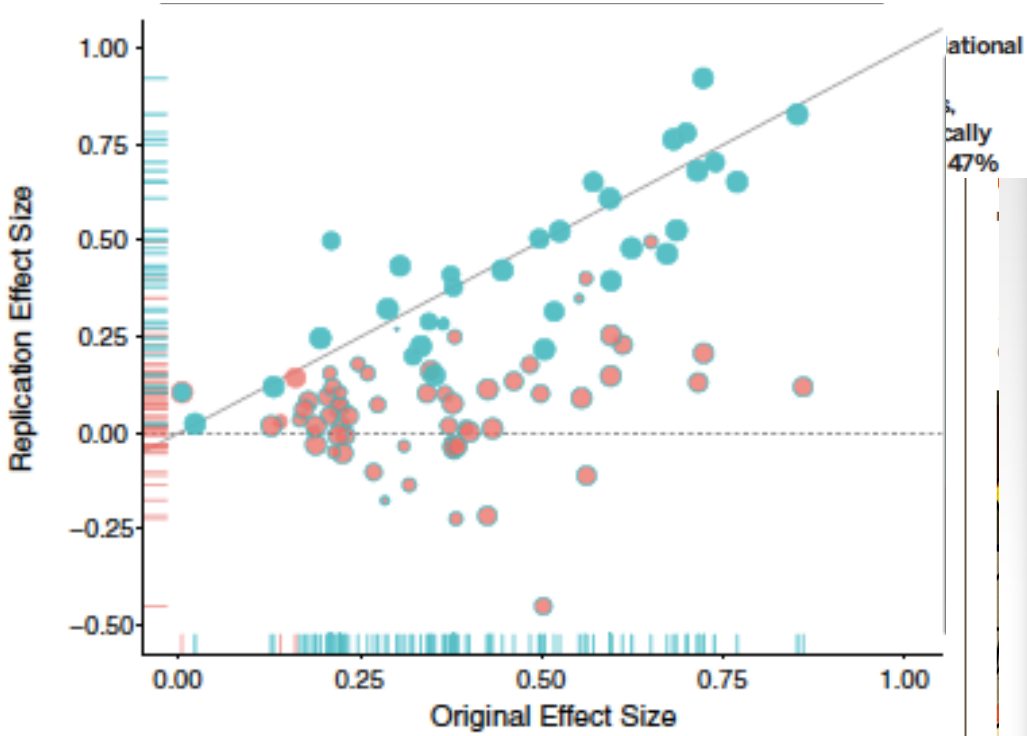


Fig. 3. Original study effect size versus replication effect size (corr
Diagonal line represents replication effect size equal to original effect size.
replication effect size of 0. Points below the dotted line were effects in the o
original. Density plots are separated by significant (blue) and nonsignificant

Open access, freely available online

ch Findings

ational

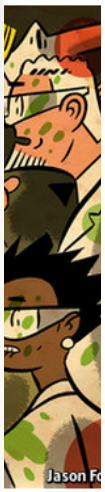
cally

47%

ology Cul

not

weet 1,852



logist, in a

Reproducibility/Replicability

- Reproduce the study: from the original data, through analysis, to get same figures and conclusions
- Replicability of results: replicate the entire study, from enlisting subjects through collecting data, and analyzing the results, in a similar but not necessarily identical way, yet get essentially the same results.

(Biostatistics, Editorial 2010, Nature Editorial 2013, NSF 2015)

- A confusion about terminology:

" reproducibilty is the ability to replicate the results…"

in a paper on "reproducibility is not replicability"

We can therefore assure reproducibility of a single study

but only enhance its replicability

Enhancing Replicability

At the level of the single study?

All agree

1. Well and transparently designed experiment
2. Reproducible data analysis and computation

    (Nature ’13, NIH in Nature ’14, Science ’14)

Also

3. Statistical methodology that enhances replicability

    But what is it?

    What problems should it address?

It's the p-values' fault

- ***Psychological Science*** "… seeks to aid researchers in shifting from reliance on NHST … we have published a tutorial by Cumming ('14), a leader in the new-statistics movement…"
- 9. Do not trust any p value.
- 10. Whenever possible, avoid using statistical significance or p-values; simply omit any mention of null hypothesis significance testing (NHST).
- 14. Prefer 95% CIs to SE bars. Routinely report 95% CIs…

Ban them!

Basic and Applied Social Psychology

Editorial by Trafimow & Marks Feb 24, 2015

- *"From now on, BASP is banning the NHSTP…prior to publication, authors will have to remove all vestiges of the NHSTP (p -values, t -values, F –values, statements about "significant" differences or lack thereof, and so on).*

Is it the p-values' fault?

Given the the attack on the p-value, a year long process started by American Statistical Association (ASA).

ASA Board's statement about p-values (Am. Stat. 2016):

- Opens: The p-value "can be useful"
- Then comes: a list of "do not" "is not" and "should not" "leads to distortion" – all warnings phrased about the p-value.
- It concludes*:* "In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. "

It's the p-values' fault!

"We're finally starting to get rid of the p-value tyranny"

Replicability with significance

"*We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results.*"

Fisher (1935) "The Design of Experiments".

What other approaches were mentioned?

Confidence intervals

Prediction intervals

Estimation

Likelihood ratios

Bayesian methods

Bayes factor

Credibility intervals,

What other approaches were mentioned?

- Confidence intervals
- Prediction intervals
- Estimation
- Likelihood ratios
- Bayesian methods
- Bayes factor
- Credibility intervals,

Epidemiology: a p-values free zone

Epidemiology: a p-values free zone

- Giovannucci et al. (1995) look for relationships between more than a hundred types of food intakes and the risk of prostate cancer
- The abstract reports only three (marginal) 95% confidence intervals (CIs), apparently only for those relative risks whose CIs do not cover 1.

Using false coverage rate controlling CIs,

which addresses selection by constructing

(1-α[#selected / # in pool])100% CIs – all three cover 0.

This is not unusual across science:

Our analysis of the 100 in the Psychology reproducibility project:

# of inferences per study (4-700, average 72);

Only 11 (very very partially) addressed selection

Two main statistical obstacles to replicability are relevant to all statistical methods

Addressing selective Inference

Addressing the relevant variability

Selective inference

Inference on a selected subset of the parameters that turned out to be of interest

after viewing the data!

Worry about the effect of selection on properties of inference

How is selection manifested?

In-study selection - evident in the published work:

- Selection by the Abstract, a Table, a Figure
- Selection by highlighting those passing a threshold
- Selection by modeling: AIC, C_p, BIC, FDR, LASSO,…

Selective inference

Out-of-study selection - not evident in the published work

File drawer problem / publication bias

The garden of forking path

p-hacking

Taylor's "Inferactive Data Analysis" – confession

My goal: to review the area only as needed to present the challenges (so biased & not complete):

Well formulated (Math) challenges

Conceptual challenges

What goes on in research?

In Medicine (not for registering drugs)

We conducted an in deep analysis of 100 papers from the NEJM 2002-2010. All had multiple endpoints

- # of endpoints in a paper 4-167 ; mean=27
- In 80% the issue of multiplicity was entirely ignored (in none fully addressed)

Psychology

Our analysis of the 100 in the reproducibility project:

- # of inferences per study (4-700, average 72);

  Only 11 (partially) addressed selection

Defending the p-value

- It's the first defense line against being fooled by randomness – needs minimal modeling assumptions
- Significant difference gives sign determination (the null need not be precisely true)
- Threshold for decision (selection) –

  essential in modern science

  (likelihood ratio, posterior odds,…, are all subject to selection)
- In some emerging branches of science it's the only way to compare across conditions: GWAS, fMRI, Microbiom, Brain Networks.

Framework

Observe $\mathbf{Y}=(Y_1,Y_2,\ldots,Y_m)$

where $Y_i \sim F_\mu$ with $E(Y_i)=\mu_i$; $H_{0i}: \mu_i = \tau_i$; $i=1,2,...,m$

or $i=1,2,\ldots$ or $i \in A$

Regular (marginal) test

$$Pr(\text{ reject } H_{0i} \text{ when it is true }) \leq \alpha$$

Regular (marginal) confidence interval $CI_i(\mathbf{Y})$:

$$Pr(\mu_i \notin CIi(Y)) \leq \alpha$$

Data dependent selection rule

$$S(\mathbf{Y}) \subset \{1,2,\ldots m)$$

Four approaches to address the effect of selection

Error-rates

A. *Simultaneous over all possible selection rules* *(SoP)*

B. *Simultaneous over the selected* *(SoS)*

C. *On the average over the selected* *(FDR/FCR)*

D. *Conditional over the selected* *(CoS)*

A. Simultaneous over all selection rules

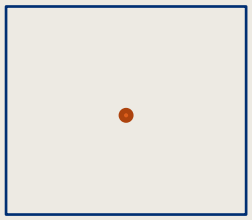
The *FamilyWise error-rate (FWER) :*

For testing: $\mathcal{R}=\Sigma\mathcal{R}_i$ is number rejected $\mathcal{V}=\Sigma\mathcal{V}_i$ rejected in error

$$Pr(\mathcal{V} \geq 1) \leq \alpha$$

For CIs $$Pr(\exists\ i,\ \mu\downarrow i \notin CIi(Y)) \leq \alpha$$

Now, for any $S(\mathbf{Y}) \subset \{1,2,\ldots m\}$ the same properties hold



Methods: Bonferroni: work at α/m

Sidak: Under independence work at $1-(1-\alpha)^{1/m}$

Royen: N(μ,Σ) work also at $1-(1-\alpha)^{1/m}$ *(convex & symmetric region)*

POst Selection Inference on any linear model selected estimated

Berk et al '15 (U of Penn team)

Challenge 1

Y~N(0, Σ) $T=\max(|Y\downarrow 1|, (|Y\downarrow 2|,\ldots, (|Y\downarrow m|)$

$$\Sigma_{i,j} = \begin{cases} 1, & \text{if } i = j \\ \rho_{i,j}, & \text{otherwise} \end{cases}$$

$$\overline{\Sigma} = \begin{cases} 1, & \text{if } i = j \\ \overline{\rho}, & \text{otherwise} \end{cases}$$

Use average correlation and be conservative (less than under ind.)

$$P_{H_0,\Sigma}(T \geq t) \leq P_{H_0,\overline{\Sigma}}(T \geq t)$$

Proof for m=3 $\rho_{i,j} \geq 0$ done ; for m>3 (done?) (Cohen & Krieger)

For any $\rho_{i,j}$ as long as average ≥ 0
For $r_{i,j}$ (estimated) ? Better than exact for m > n(?)
For $Y_i \sim t_\nu$?

Currently calculation for large m infeasible! “Bonferroni” in GWAS

SoP often too harsh: Natalizumab study

Natalizumab, was examined by Ghosh et al (NEJM, 2003) for the treatment of Crohn's disease.

Comparing 3 regimes with placebo; 4 measures of success;

at 5 time points; Total 51 endpoints

1 primary endpoint: Treatment by 2 infusions of 6mg/kg dose remission measured at week 6

Other 50 described as secondary endpoints

The result for the primary endpoint was not significant ($p = 0.533$);

27 secondary endpoints at $p \leq 0.05$ were considered as discoveries

Study reported as a success

Would not have been reported as such using FWER control

B. Simultaneous over selected (SoS)

The selection rule *S* is determined before observing the data.
e.g.: the largest; the five largest; forward selection with stopping rule

For tests $\Pr(\sum i \varepsilon S(Y) \uparrow \blacksquare V \downarrow i \geq 1) \leq a$

For CIs $Pr(\exists i \in \boldsymbol{S(Y)}, \ \mu \downarrow i \downarrow \notin CIi(Y)) \leq \alpha$

Obviously SoP => SoS

Inference on being largest

Example : Studying the effect of a pre-specified risk factor β:

By itself? Adjusted for Age and Gender? Best of them?

The abstract carries merely the largest of the two β's and its 95% confidence interval

Goal: Design a (single) Conf. Int. for the largest

Can we do better than $\beta_{max} \pm se(\beta_{max}) z_{1-\alpha/4}$?

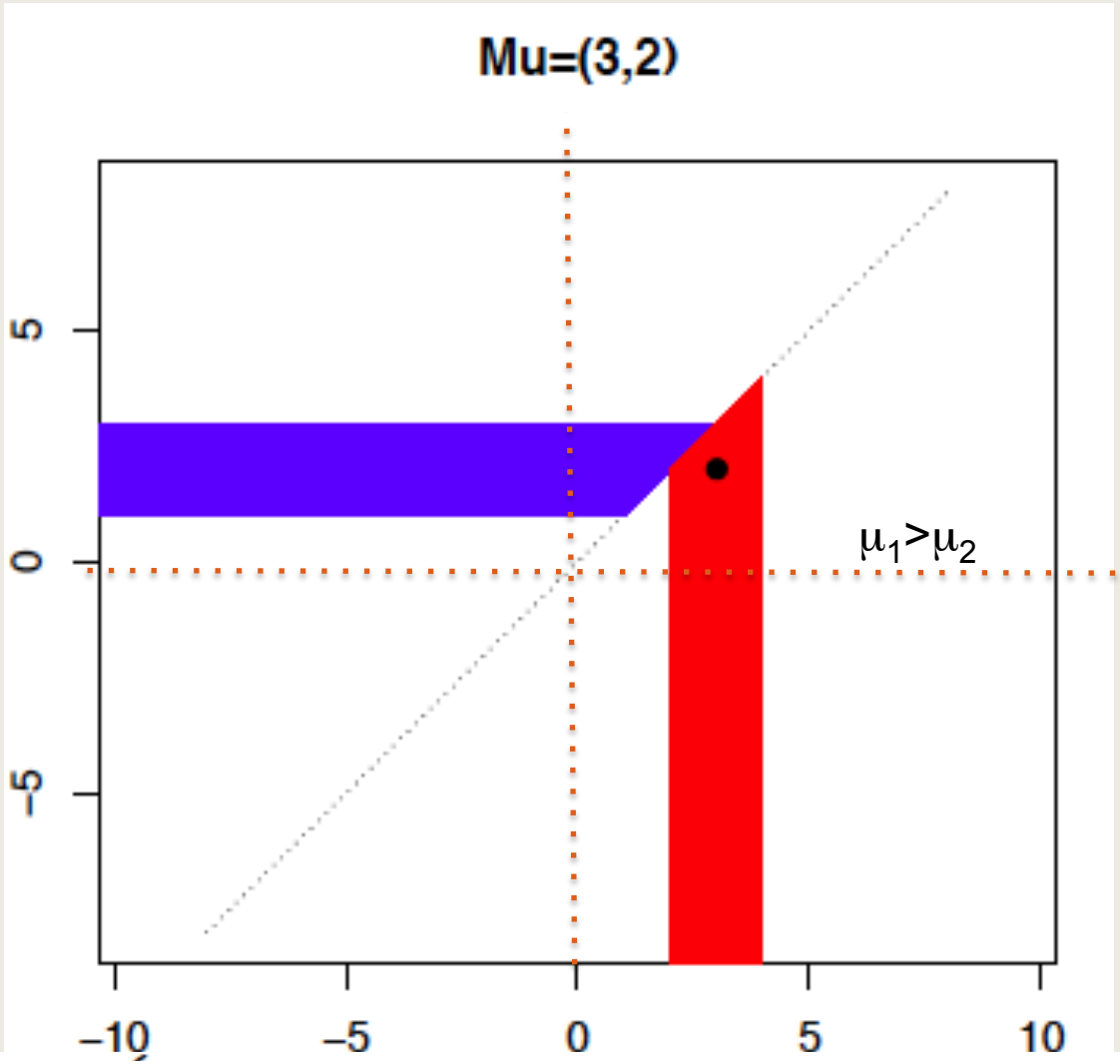
Surprising result: $c=z_{1-\alpha/2}$

Hence CI for max of two is like CI for one parameter!

Holds for correlated bivariate normal as well

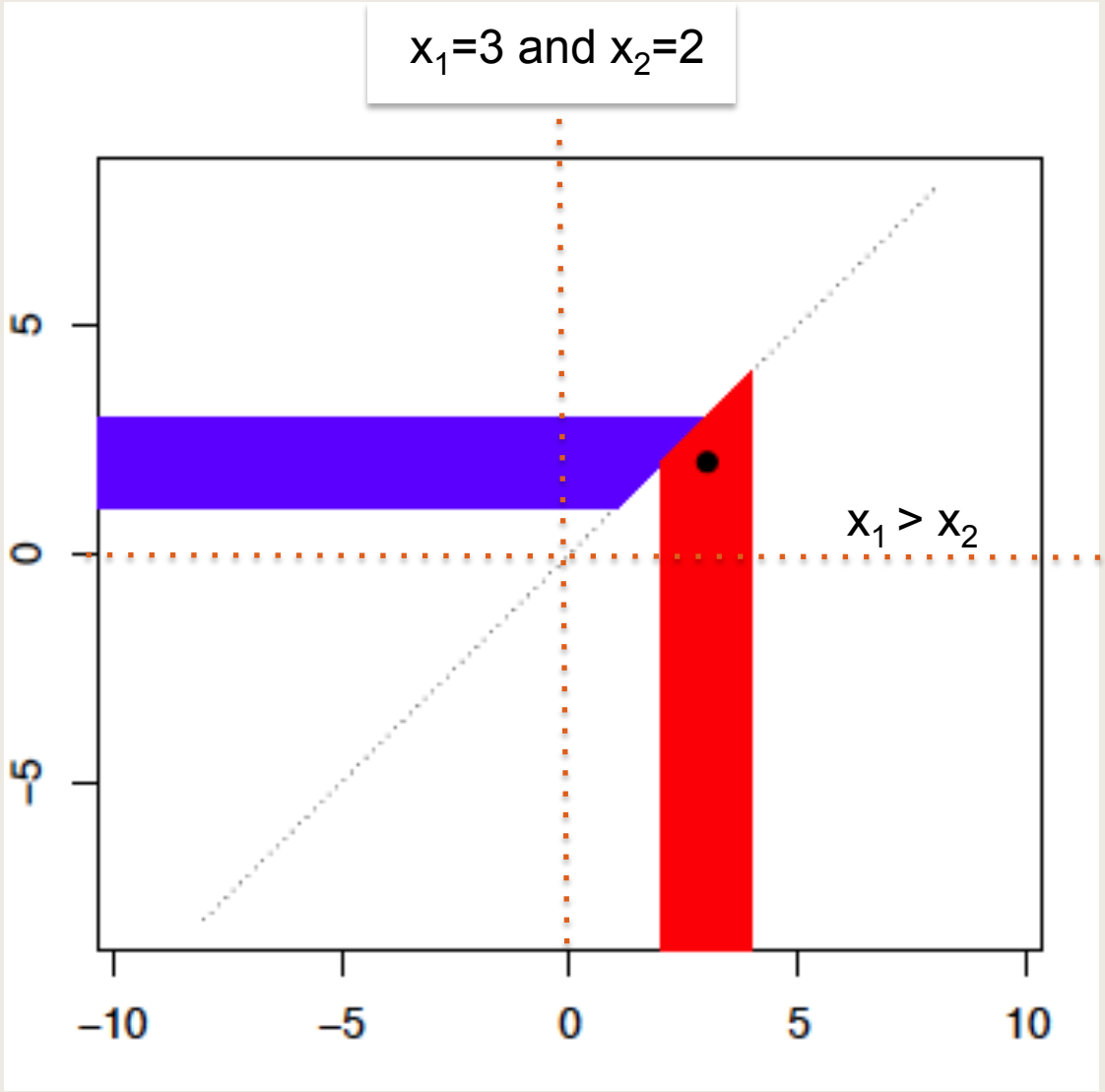
Hechtlinger Stark YB ('15+)

Acceptance regions for μ (non-equivariant)



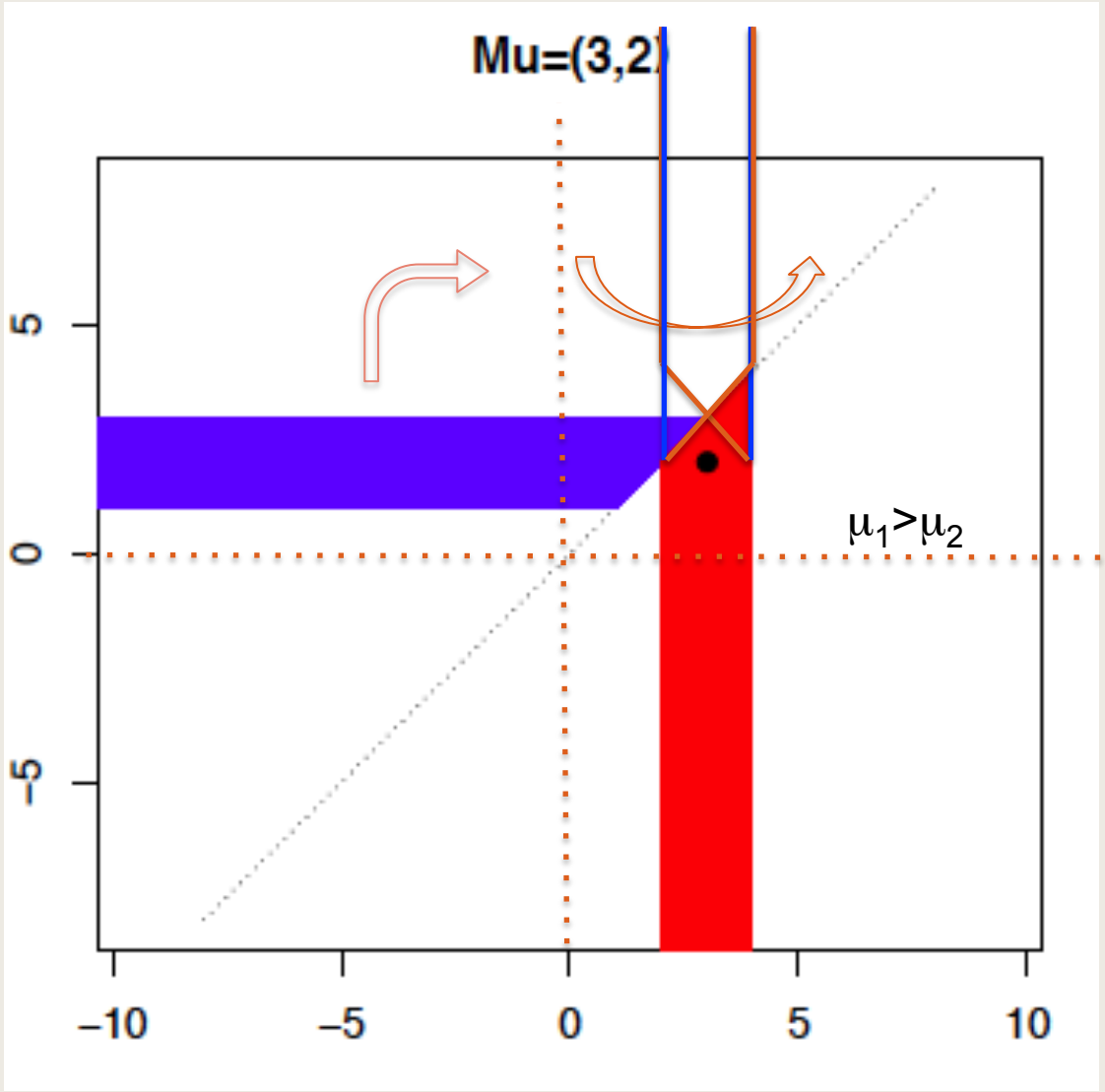
$$P(Z \in (-c,c)) = F(c)^2 - F(-c)^2 = [F(c)+F(-c)]\,[F(c)-F(-c)] = [F(c)-F(-c)] = P(X_1 \in (-c,c)).$$

Confidence intervals for μ_i (x_i the largest)

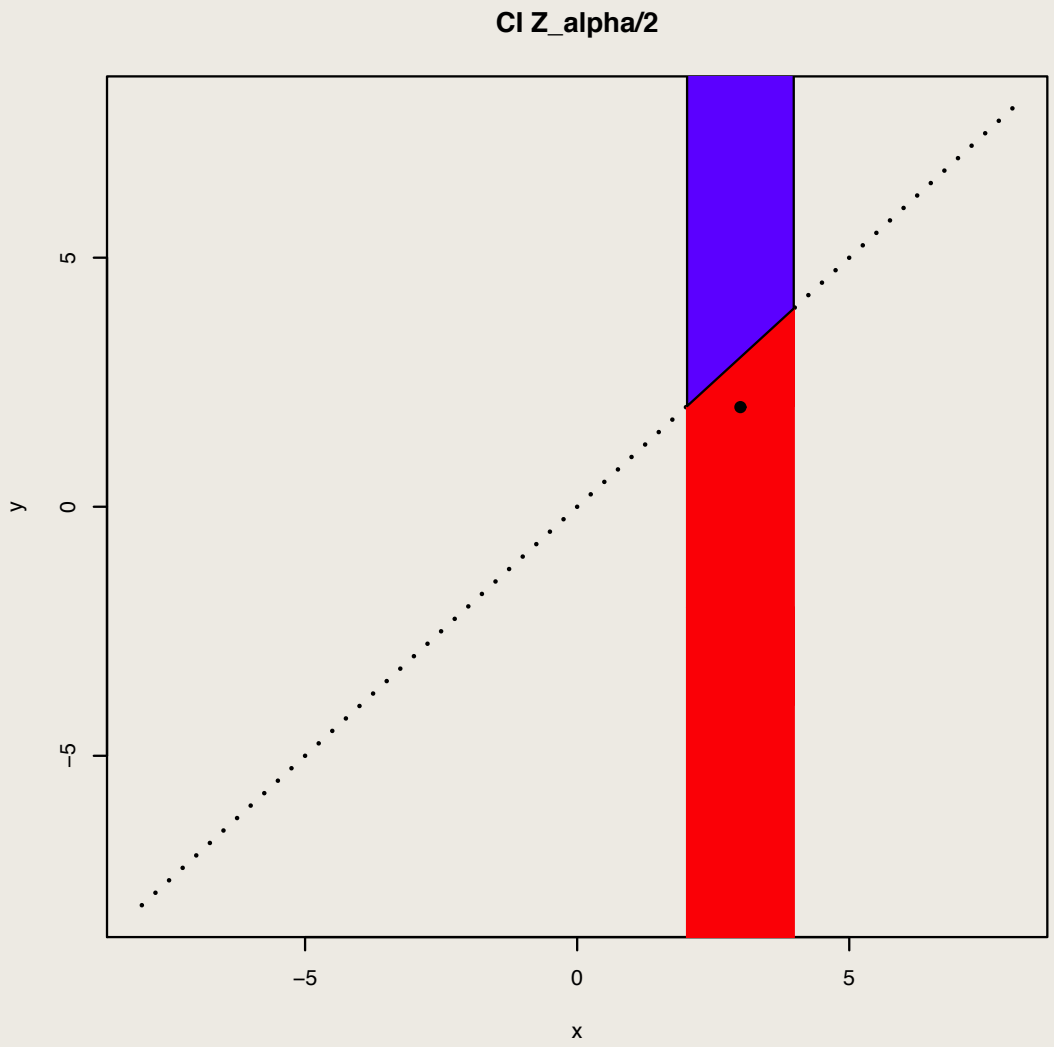


How large should c be?

Acceptance regions for μ : How big should c be?



So $c=z_{1-\alpha/2}$: CI for max of two is like CI for one parameter!

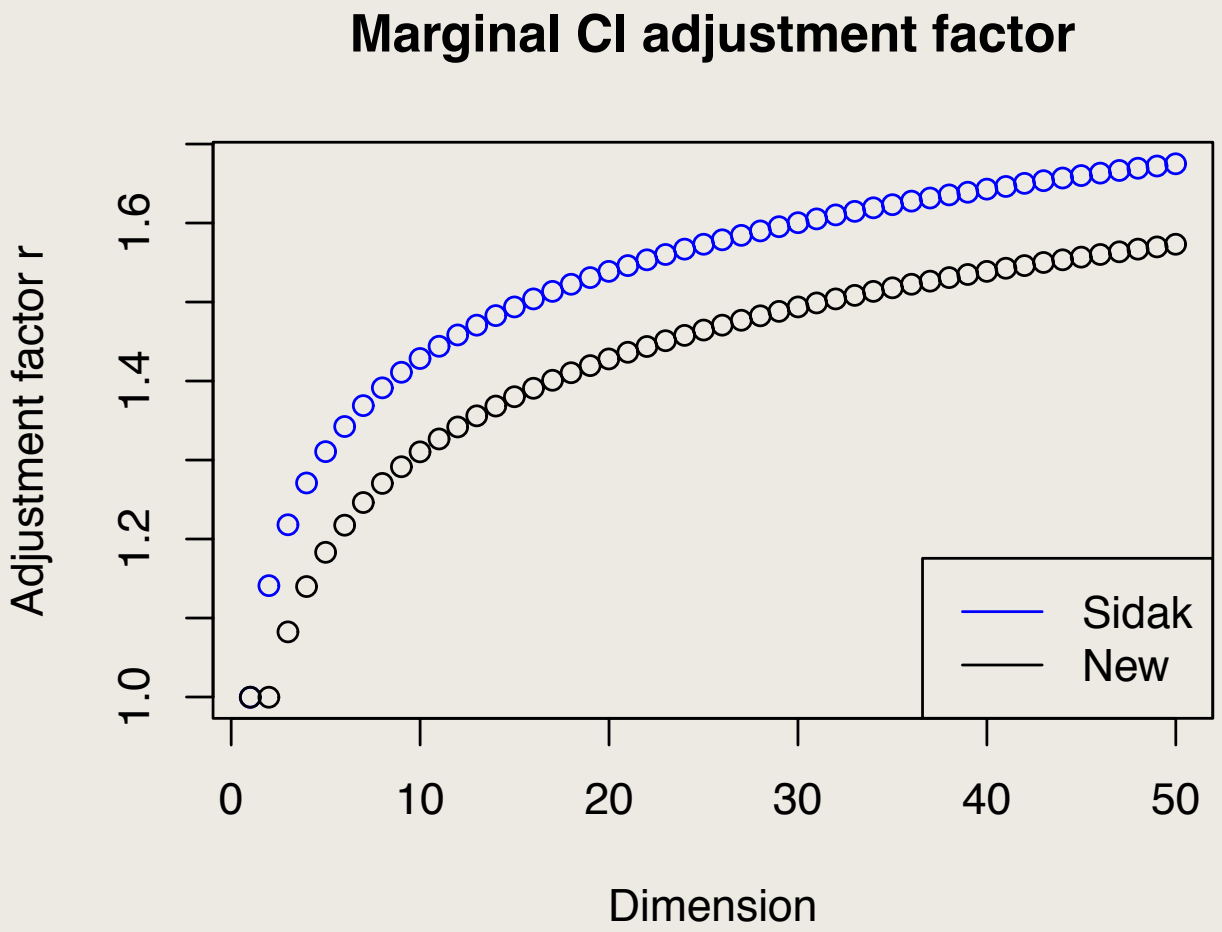


Holds for correlated bivariate normal as well

- Result: $c=z_{1-\alpha/2}$
- Hence CI for max of two is like CI for one parameter!
- Holds for correlated bivariate normal as well

Hechtlinger, Stark & YB ‘17

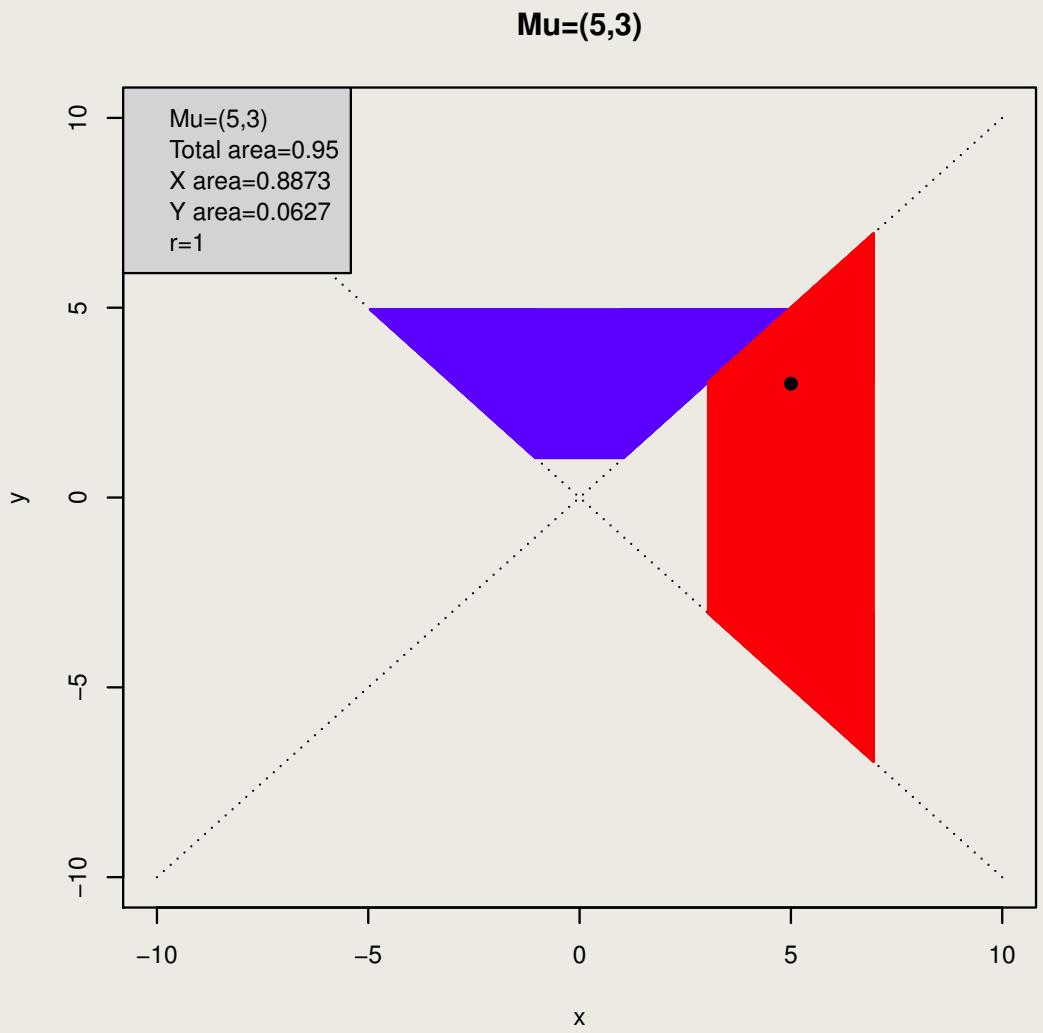
For maximum of m>2

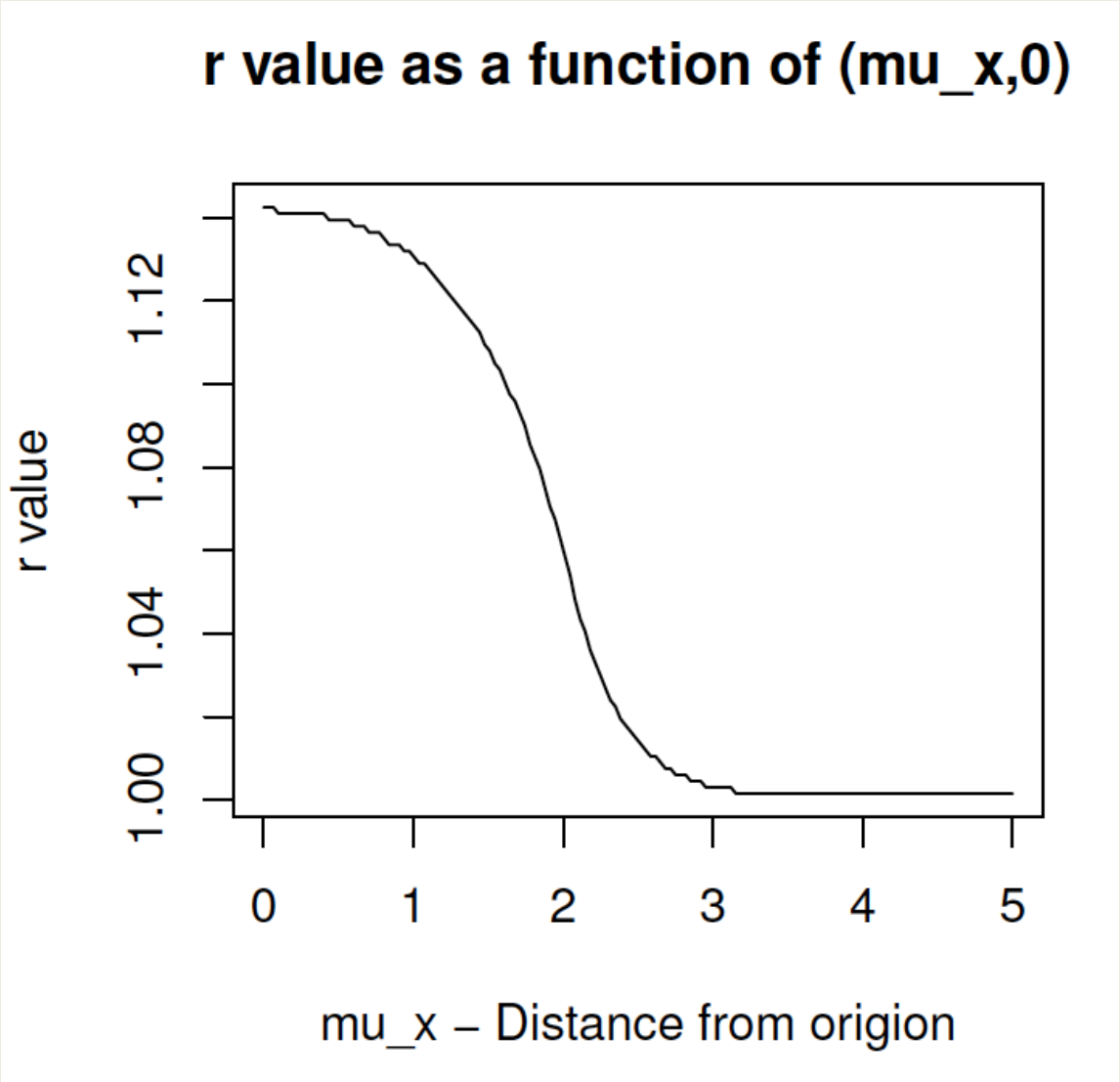


In general under independence:

one-sided maximum modulus interval

$\max |x_i|$


Mu=(5,3)
Mu=(5,3)
Total area=0.95
X area=0.8873
Y area=0.0627
r=1
10
5
0
−5
−10
y
−10
−5
0
5
10
x



When away from the origin, again no inflation is needed

Conjectures:

The least favorable situation for max k out of m is

$$(0,0,..,0,\underbrace{\mu, \mu,.., \mu}_{k-1})$$

Challenges 2:

(i) The CI remains conservative under dependency

(ii) How far can we go with this approach?

Simultaneous over the Selected when selecting by

max(abs), Forward Selection, Lasso, etc

See Goeman & Solari (‘13); Next talk by Roquain (?)

(iii) Point estimator

Importance: Registering analysis plan for reproducibility

C. On the average over the selected

Rephrase the False Discovery Rate (FDR) for testing:
$S(\boldsymbol{Y})$ selects the rejected hypotheses; $R = |S(\boldsymbol{Y})|$
V is the number in S(Y) of type I errors

So $FDP = V/R = (\sum i \varepsilon S(Y) \uparrow V\downarrow i \ / |S(Y)|$ *if* $R>0$

$= 0$ *if* $R=0$

And

$FDR = E(FDP)$

FDR is the expected average # errors over the selected

For Conf. Int. define False Coverage-statement Rate (FCR)

as above by setting $V\downarrow i$ =1 for a non-covering interval

A reminder of the BH

Sort $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$; $k = max(j \mid p_{(j)} \le q\, j/m)$;

Reject H_{0j} , for all $j \le k$ (none if no j exists)

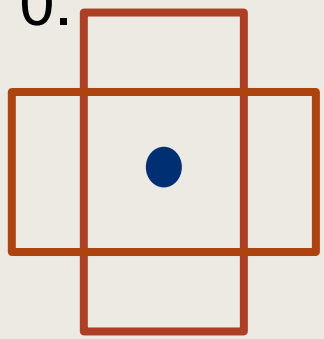
$p^{BH}_{(k)} = min(p_{(j)} m/j \mid j \ge k)$ FDR-adjusted p-values (q-values)

BH controls FDR at level q for (i) ind. test statistics,

(ii) Positive Regression Dependent on a Subset (PRDS => MTP_2), (iii) One-sided tests Gaussian with $\Sigma \ge 0$.

(iv)Two-sided studentized ind. Gaussian.

Royen's result does not yield more because



Challenge 3:

Prove that the BH is conservative for two-sided tests for Gaussian dist'd test statistics under any correlation structure

Reiner showed by simulations that the worst case is $\rho_{ij}=1$

Cohen showed that if $\rho_{ij}=1$ only for the $m_0<m$ true H_{oi} , then

$$(1.5) \quad FDR \le \alpha \frac{m_0}{m} \left\{ 1 + \sum_{j=m_0+1}^{m} 1/j \right\}$$

Hence conservative.

Further Support: Candes (?) Heller (?) Roquain (?)

Natalizumab study endpoints:

Ingnoring selection 27; FWER None FDR 18

The use of weights

Weighted False Discovery proportion

$$wFDP\ (= V/R) = \left(\sum i \varepsilon S(Y)\uparrow\ w_i V_i\downarrow\right) \Big/ \left(\sum i \varepsilon S(Y)\uparrow\ v_i R_i\downarrow\right) \quad \text{if } R=0$$

And

$$wFDR = E(wFDP)$$

YB & Hochberg ('89) Genovese & Wasserman ('03) Ramdas et al (17+)

The current US practice for Primary & Secondary endpoints:
well described by $w_i=v_i=0$ for secondary endpoints
$w_i=v_i=1$ for the single primary endpoint.

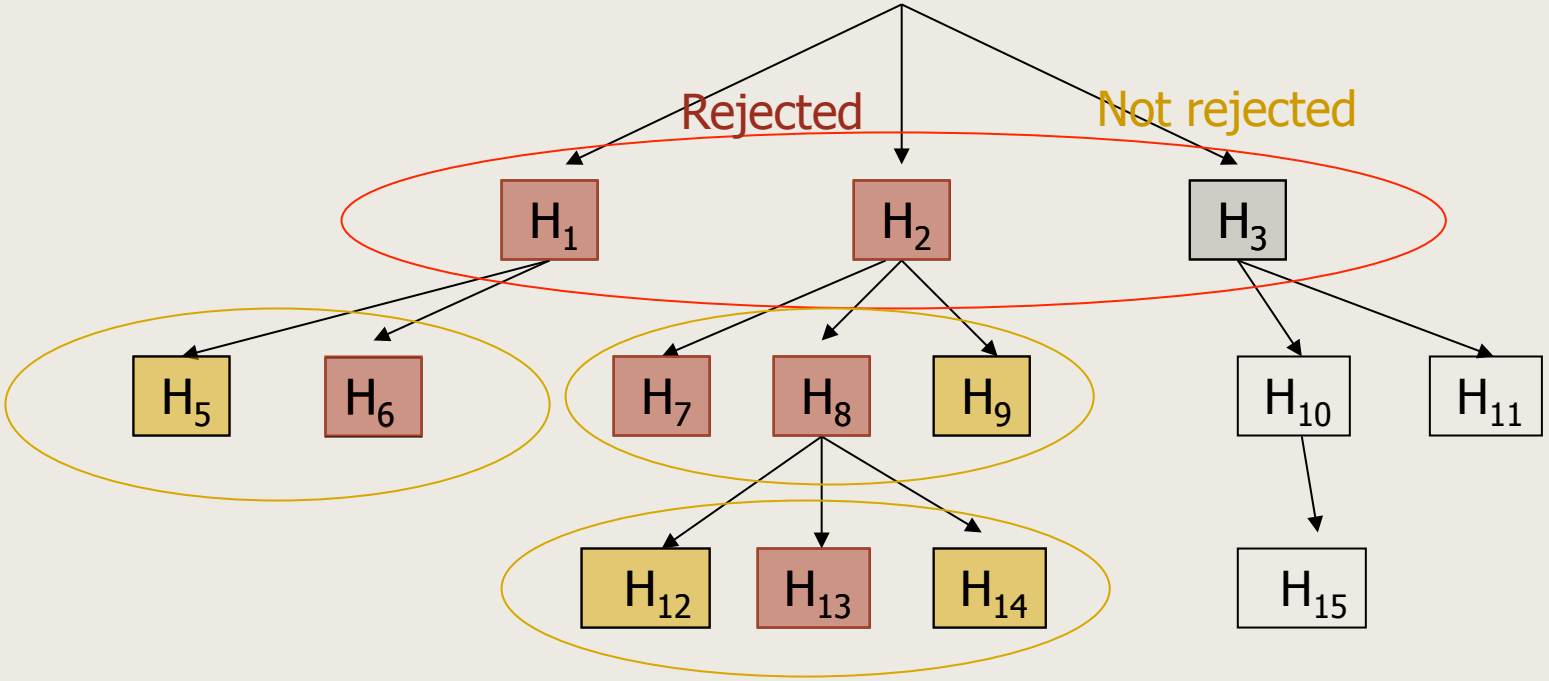
YB & Cohen ('16)

Yet secondary advantages are allowed to be on the package

Challenge 4: The use of weights in particular settings

The hierarchical framework

(Yekutieli et al `06, Yekutieli ‘08)



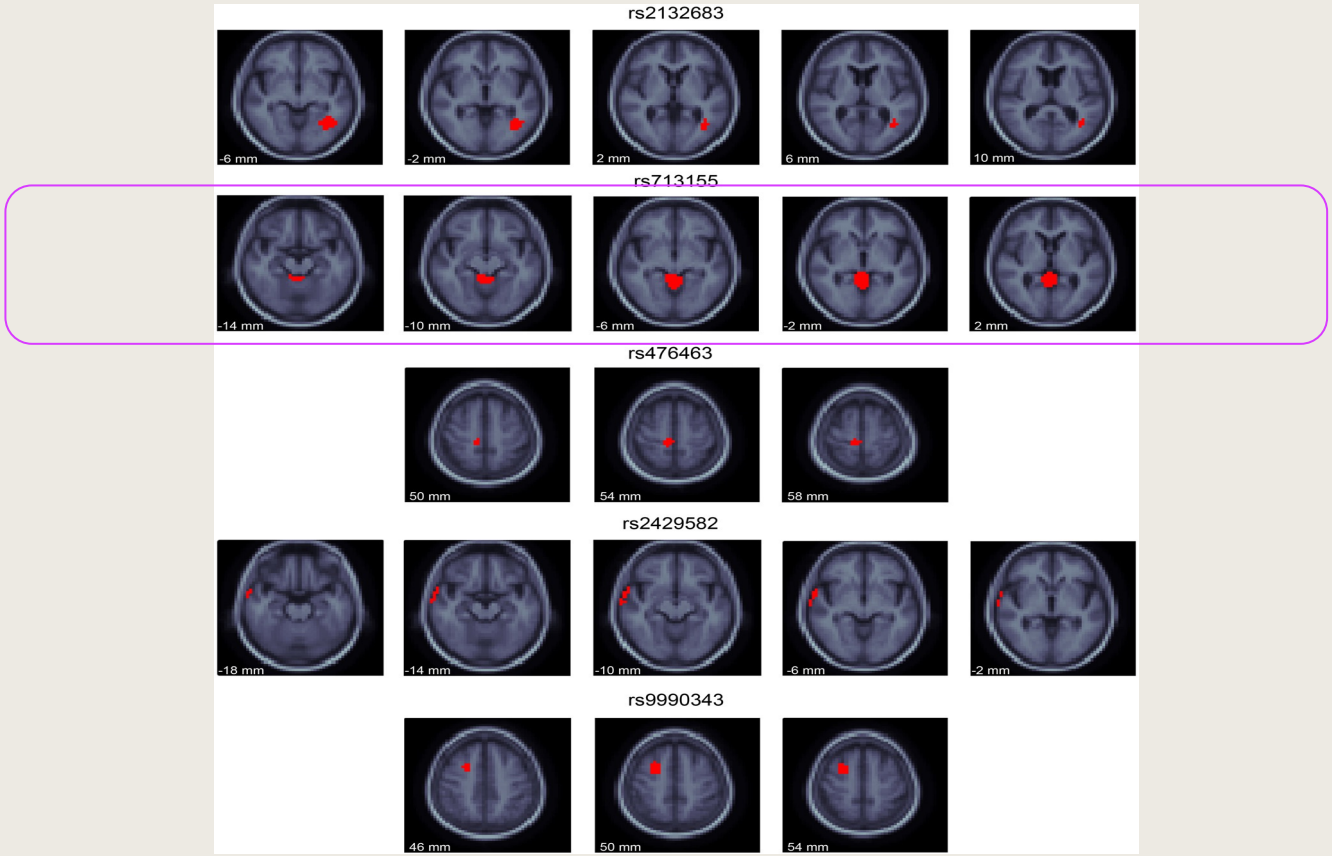
1. Arrange hypotheses in sub-families corresponding to a single parent hypotheses
2. Test sub-family of a rejected parent hypothesis by the procedure in BH at q

Meaningful concepts. Beautiful results, but they require independence between test statistics of a parent and its sub-family

Inference on selected families

Goal: Association between volume changes at voxels with genotype (Stein et al.'10)

1 ← Voxels searched → 32,000



1

SNPs

448,000

number of tests ~ 13,000,000,000

Selection adjusted testing of families

Let H_{ij} be the the hypotheses in family F_i, $j=1,..m_i$; $i=1,\ldots,m$
with $\boldsymbol{Y}=\{ Y_{ij}\}$ or with p-values $\boldsymbol{P}=\{ p_{ij}\}$)

$S(\boldsymbol{P})$ is a selection procedure of families.

$|S(\boldsymbol{P})|$ the (random) number of families selected.

The control of error E(C) (FDR, FWER, False Exceedance rate and others) on the average over the selected families means

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

BH over all hypotheses may be too liberal on the family level!

(BH-q, BH-**R**q/m) - hierarchical testing

Get p-value for the intersection hypothesis $\cap j\uparrow\blacksquare H\downarrow ij$ in F_i ,

$$p_i = min(p_{i(j)} m_i / j)$$

Test the families using BH-q with p_i; select the rejected **R**.

Within each selected family use BH at level q*(**R** / m)

(1)
$$E(\sum i\in S(P)\uparrow\blacksquare Q\downarrow i \ /|S(P)|\ ) = q/m$$
$$\sum i\uparrow\blacksquare m\downarrow 0 i \ / m\downarrow i \leq q$$

YB & Bogomolov '14

(2) FDR ≤ q across families;
(3) FDR ≤ q overall.

More general selection rules; Multi-level structures

Challenge 5:

Develop methods that address such complex structures

Tree structured families

Heller's talk, Sabatti's talk, Foygel-Barber & Ramadas ('16)

But much more to be done:

Addressing dependency

Topological features (peaks in level sets, cusps)

Siegmund's talk?

Graphs

Adaptive methods

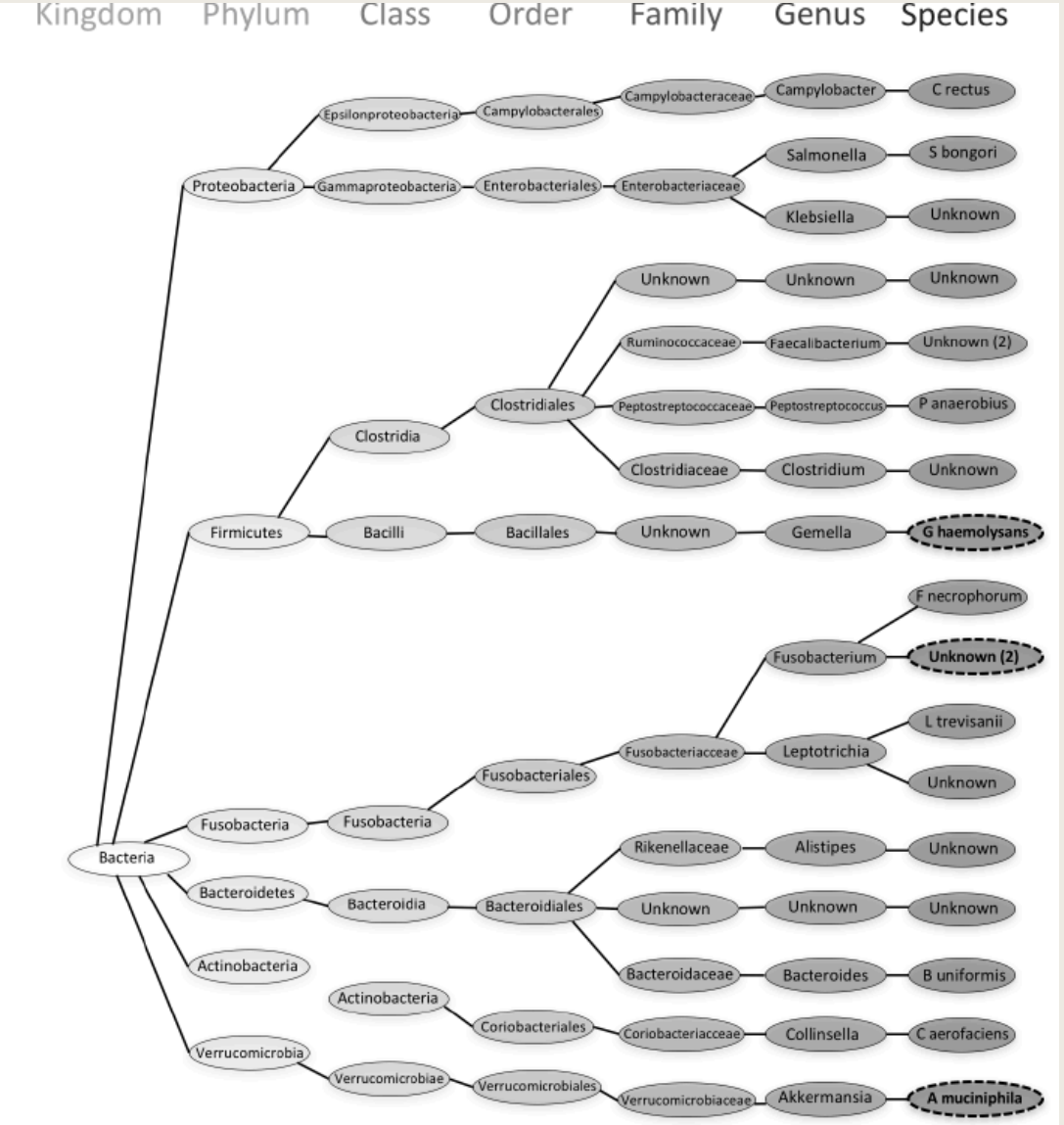
One more example and a challenge:

Association of gut microbiome with Colon Cancer

- # microorganisms 496
- N= 177 (86 tumors)
- Abundance determined by rDNA & compared between cancer and normal

- 4 were discovered by TreeFDR and not by BH

- 14 were discovered by BH and not by TreeFDR, 13 of which were of unknown type

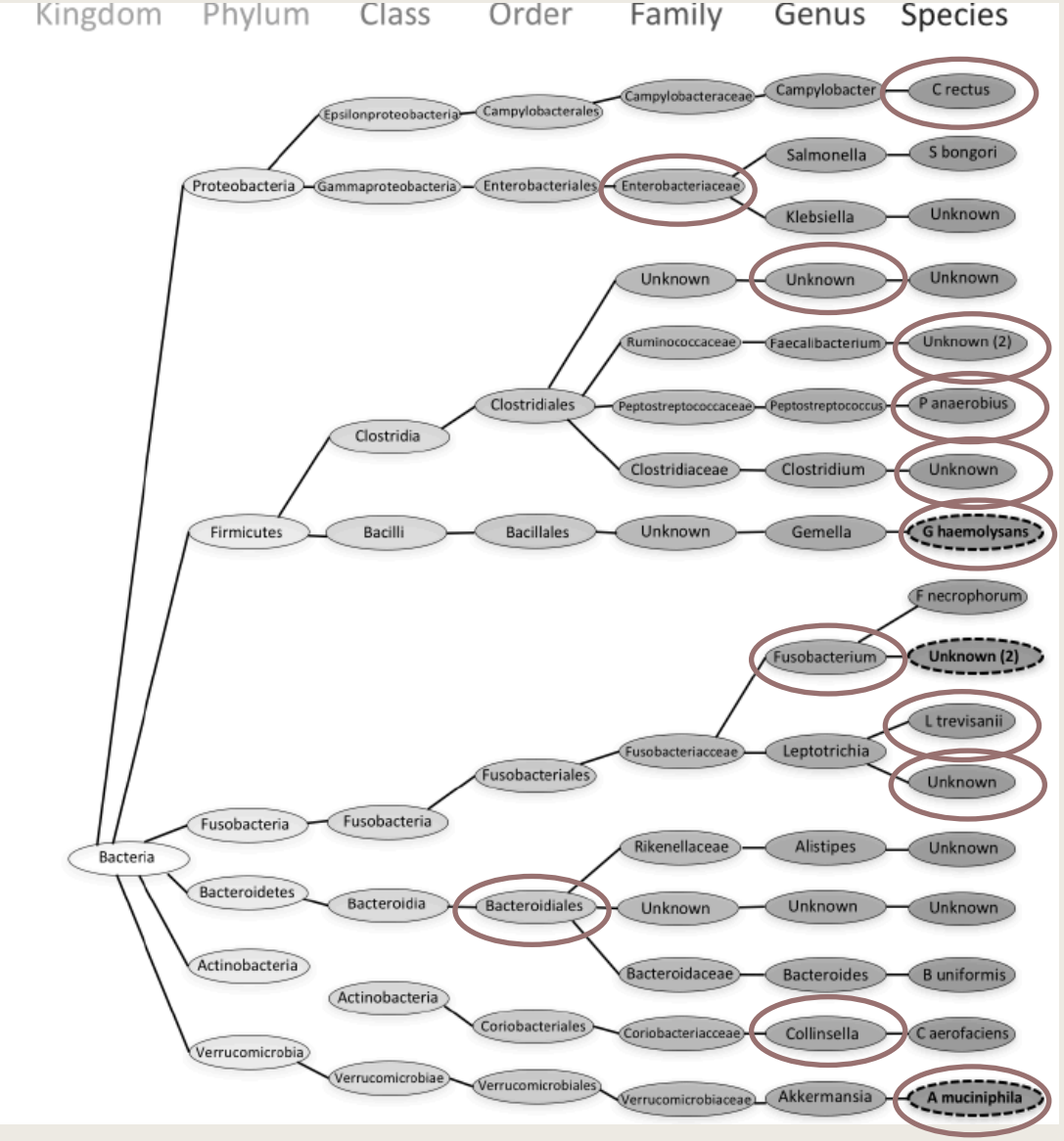


Association of gut microbiome with Colon Cancer

Beyond offering selective inference at the levels of Phylum, Class, Order, etc. which always result in the identification of at least one Species,

There is interest in inference that stops at a family, genus etc. with no particular species identified, namely the end-nodes family

Challenge 6: Theory for inference on end-nodes (under dependency between parent and its children) is still lacking



Association of gut microbiome with Colon Cancer

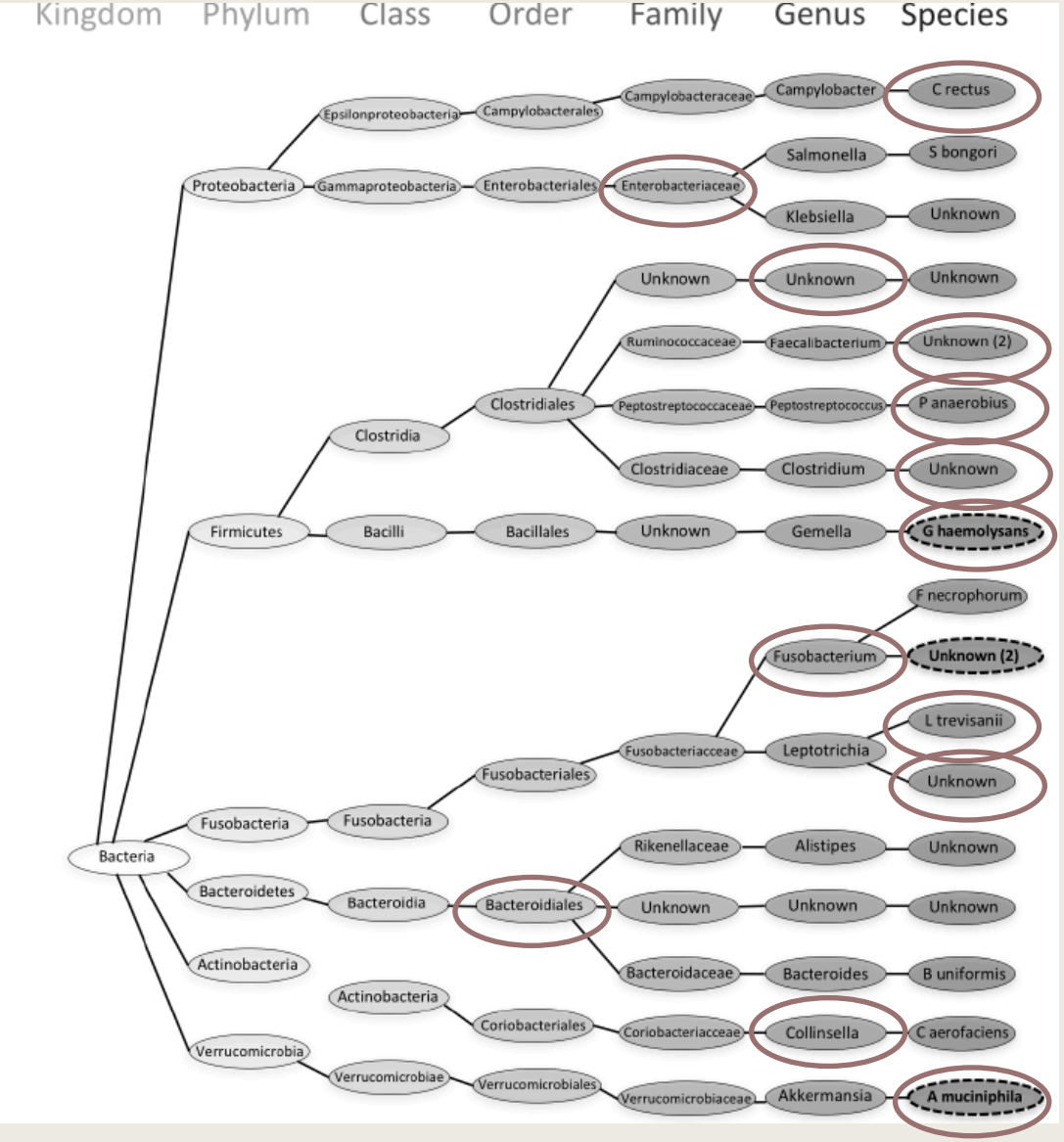
Hence tests such as

Higher Criticism
Simes-Hotelling

Can be very useful

Offering more power at a node at the expense of no rejection of children hypotheses

Theory for inference on end-nodes (under dependency between parent and its children) is still lacking



Challenge 7:

Avoiding the use of p-values for testing

Use of negative controls (fMRI, Genomics)

Knockoff for testing (Candes Foygel-Barber)

A synthetic way of generating matched negative control

Challenge 7: can it be used for selective CIs? Selective estimators?

Testimation

FDR thresholding has exact asymptotic minimaxity (including the constant) adaptively over bodies of sparse signals l_r , $0 \le r < 2$,

(Abramovich, YB, Donoho Johnstone '06)

(i) The usual average squared error over entire vector including those thresholden to 0 was considered

(ii) Independence assumed

Challenge 8: What happens when we consider average squared error over the selected?

Challenge 9: This testimator is a Penalized Model Selection for orthogonal X. How about its performance for non-orthogonal X? (hints in Gavrilov et al '13' Bogdan et al '15 SLOPE) Abramovich's talk?

D. Conditional over selected

Selecting from *m* features by a selection rule $S(\boldsymbol{Y})$

For *each of the selected ones,*
construct a marginal conditional confidence interval

$$Pr(\mu_{\downarrow i} \notin CIi(Y) \mid i \in S(\boldsymbol{Y})) \leq \alpha$$

E.g. Select the largest one; Bigger than 2; p-value ≤.01 ;
Coefficients in the Lasso

eQTL: The TreeQTL application

Peterson, C., M. Bogomolov, Y. Benjamini and C. Sabatti (2015, 2015),

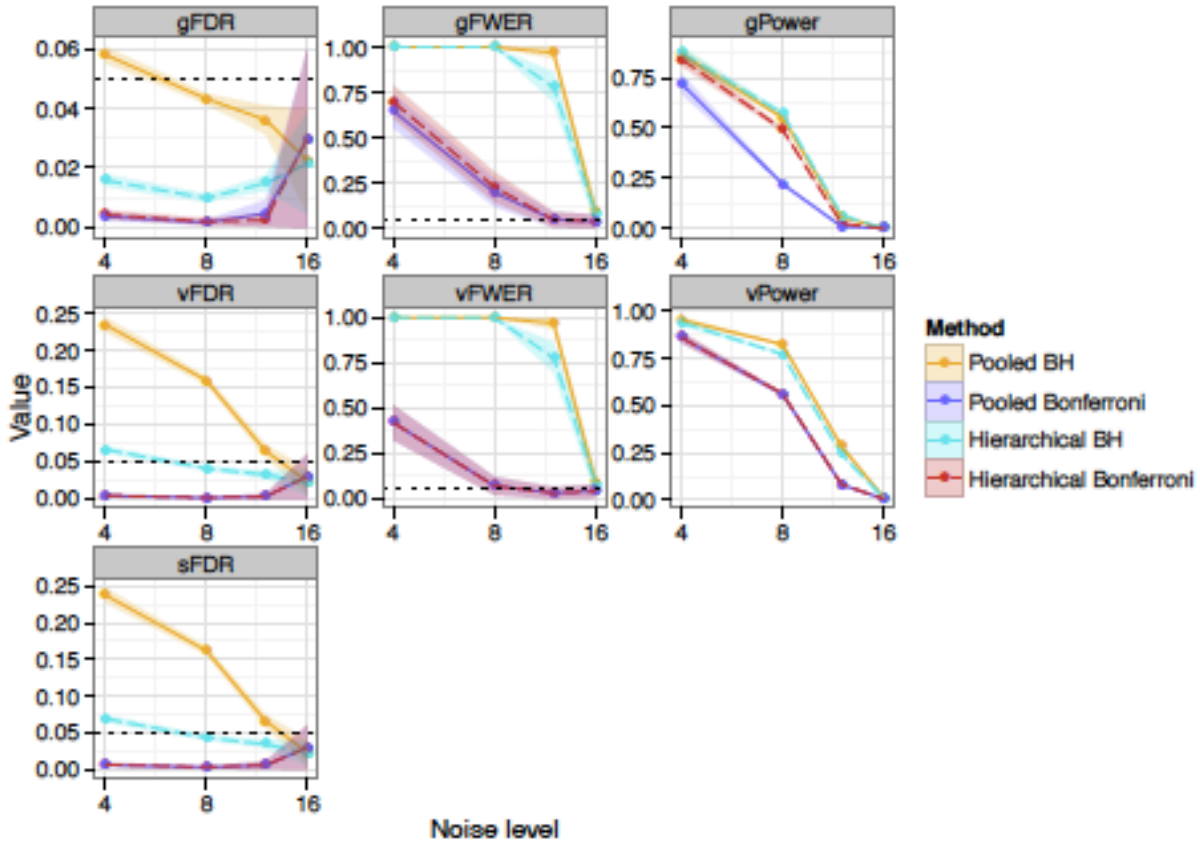


Figure 5: Error rates and power for four multiple-testing strategies applied to simulated data starting from real genotypes. The lines show the average, the shaded areas report the standard error over 100 iterations, and the dotted horizontal lines mark the 0.05 level.

Utilizing the selection procedure used

Select μ_i if its estimator is big enough

$$X_i = (Y_i \mid |Y_i| \geq c),$$

where c is fixed

or (simple) data dependent $c(\boldsymbol{Y})$.

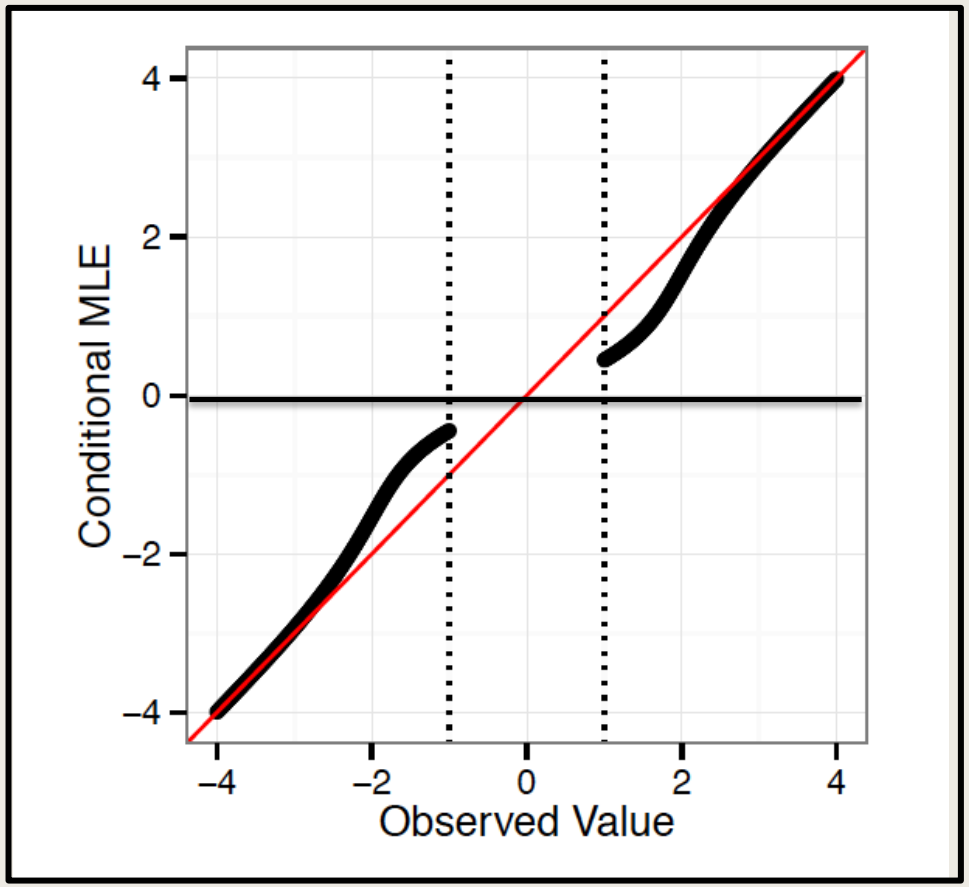
Conditional density -> Acceptance region for each parameter

(non-equivariant) with short 0-crossing -> inverting to get

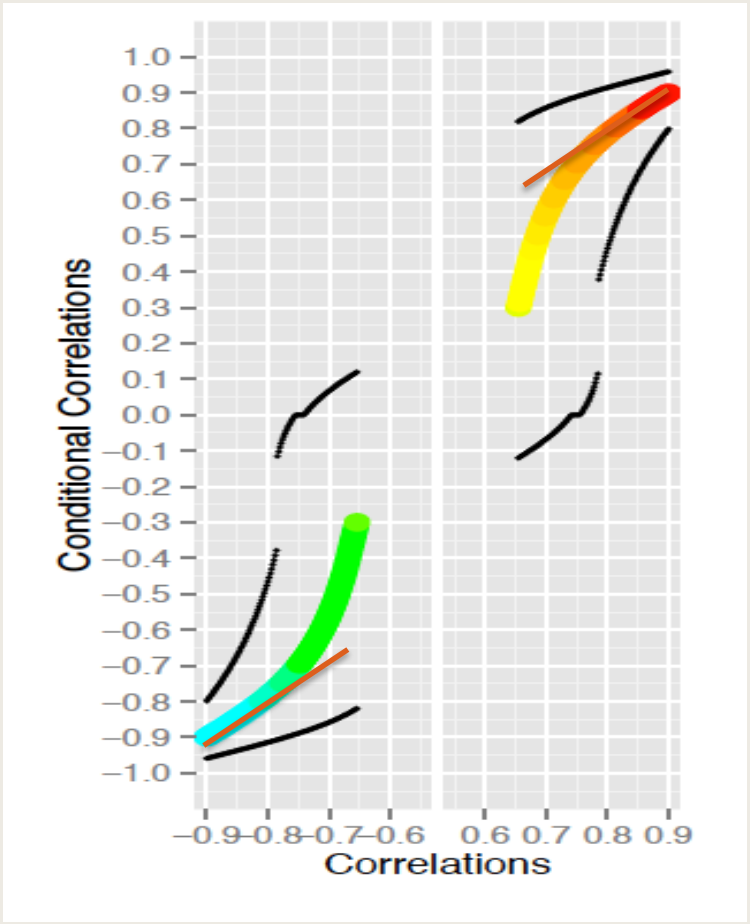
Conditional CIs -> offer FCR

Hedges ('84) for meta-analysis, Zhong &Prentice ('08) asymptotic dist'n in GWAS, Weinstein Fithian YB ('13)

Conditional MLE

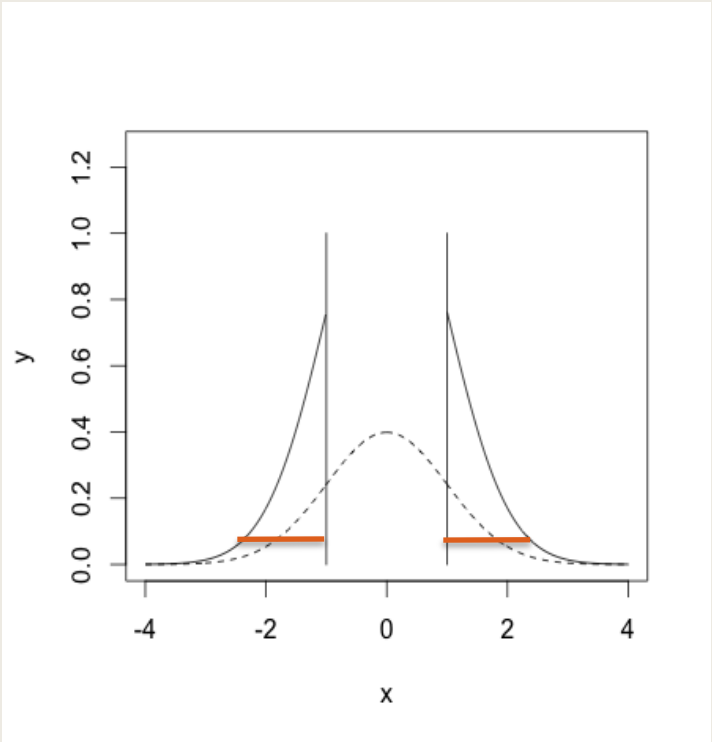


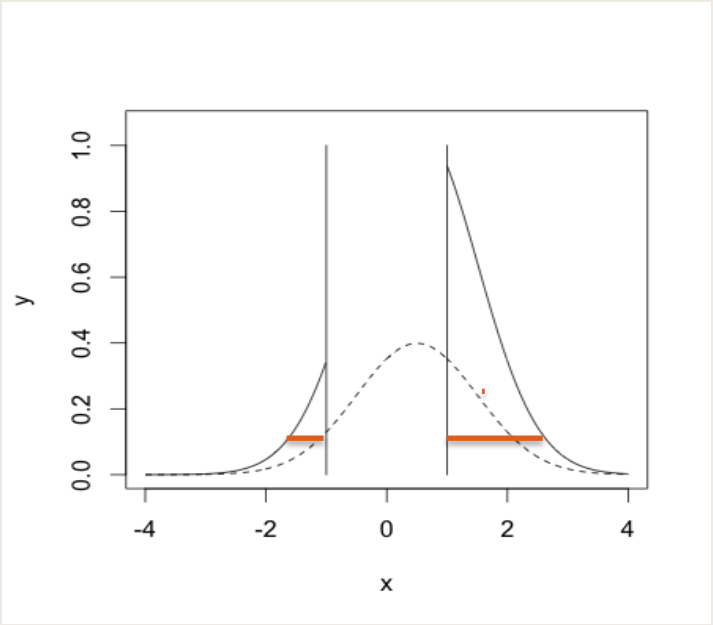
Cond. MLE and CI for correlation

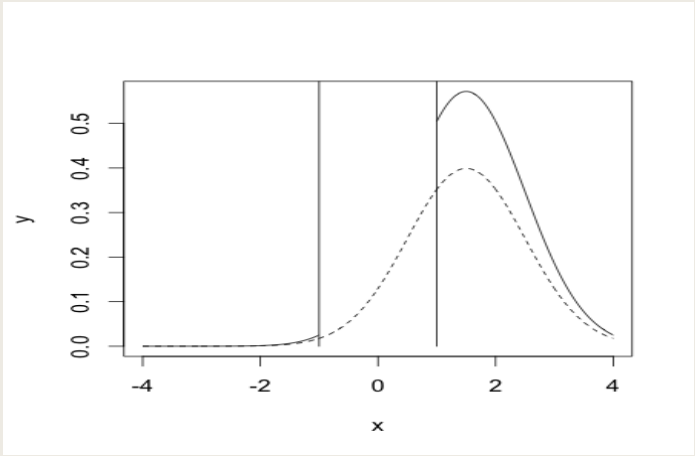


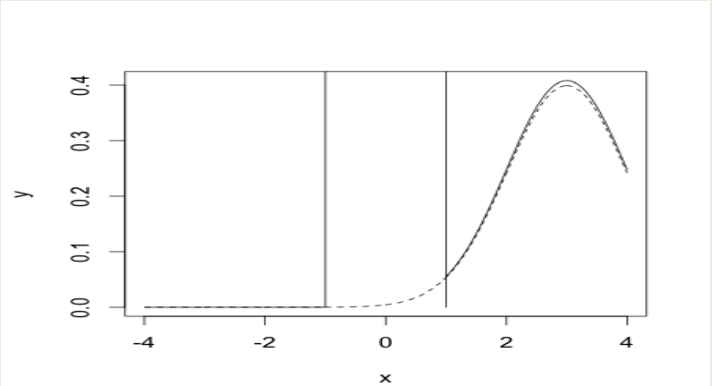
Hedges ‘84, Zhong and Prentice ’08, Fithian, Sun, Taylor (16) YB and Meir (16+)

Can be used to address ‘publication bias’









The complication: θ is no longer only a shift parameter

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

on

With estimators and CIs conditional on p-value ≤ 0.05

77% fall in Cond. CI Instead of 47%

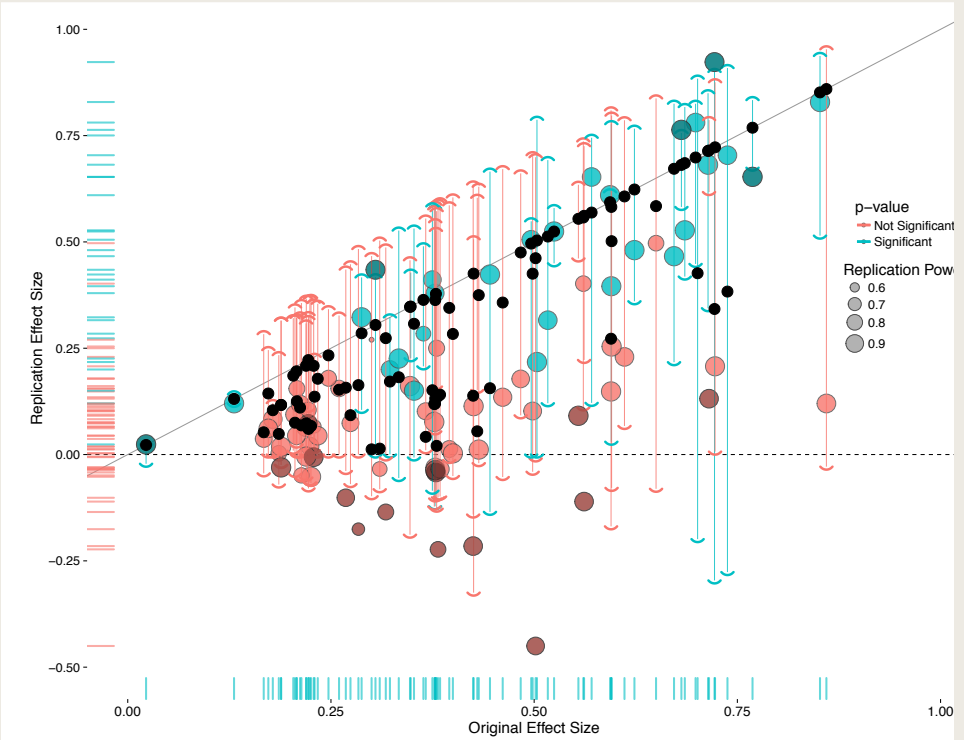
- Principle 6*:* "…a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis"

  Based on Val Johnson's '14 PNAS paper which offers Bayesian argument for $p \leq 0.005$ threshold

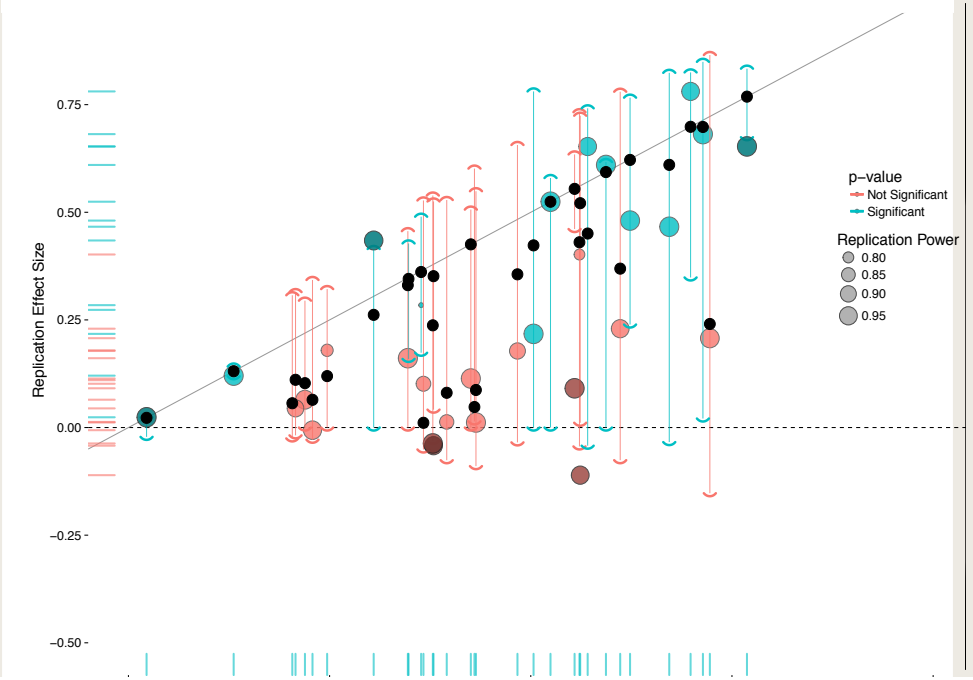
The problem though is being close to the threshold,

not to where the threshold is set

Thresholding at
p-value ≤ 0.05



Thresholding at
p-value ≤ 0.005



More recent work on conditional inference

J. Taylor with coauthors and students:

- P-values and CI post selection of a model using a Lasso
  Asymptotic p-values Lockhart, Taylor, Tibshirani, J.Tibshirani (14)
  Exact post-selection inference Lee, Dennis Sun, Yuekai Sun,Taylor (13+)
- P-values & CIs post screening variables with marginal significance then fitting using some model selection (AIC, Lasso, …) Tian Lotfus, Reid, Choi (14+,14+,14+,15+) Fithian et al (’16+)
- Post selecting a parameter with |estimator| > threshold Fithian, Sun, Taylor (14+)
- Inference on the average over a level set – YB Irizarri & Taylor (‘16+)
- Inferatory Data Analysis Talk

Heller’s Talk

Challenge 10 (mine)

- I feel more comfortable with conditional analysis when replication can be done while controlling the conditioning event (p≤.05; expression ≥ two-fold; effect size ≥1/3)
- Should it bother me? Are there situations where the penance (power loss) of conditional analysis can be tolerated but it is less natural than other approaches?

Challenge 11

Addressing Exploratory Data Analysis

(the garden of forking path, p-value hacking,…)

- Let the researcher do exploratory data analysis in a free and uncommitted way.

But

- Document in the analysis software the analysis path taken: do automatically reproducible computing (no confession needed).
- Offer conditional inference given the taken path

Summing up

1.The importance of selective inference

- The dangers of selective inference in testing are recognized even by the researchers, though usually when m > 1K (4K)
- But it is a quite killer of replicability even when m>10.
- There is well developed practice to address testing
- Adjusting for selection in estimation and confidence intervals is rarely practiced, leading to dwindling results upon replication.

2. There is more than one approach to selective inference

SoP => SoS => FDR/FCR <= CoS

3. This research area is active

- Some well formulated challenges
- Many conceptual challenges awaiting development

11 were presented; challenge 12 is left for you to formulate

Thanks!

www.replicability.tau.ac.il

1888 1999

The industrialization of the scientific process

1950 2010

Thanks

www.replicability.tau.ac.il

Source of the problems

But notice: Replicability problems became more severe only recently,

In my opinion: because of

The industrialization of the scientific process

Industrialization of the scientific process

- Compare to changes in car manufacturing process
- Internal Combustion production started in 1888 by Benz
- 5 cars per year separately and manually manufactured.

66

What about Confidence Intervals?

The use of Marginal (standard) 95% Confidence Interval

on the selected few may be deceivingly optimistic.

Indeed,

on the average over all parameters,

the expected proportion of intervals failing to cover $\leq \alpha$:

V_i making a non-covering confidence interval

$$E(\Sigma V_i / m) = \Sigma E(V_i)/m = m\,\alpha/m$$

But

20 parameters to be estimated with 90% CIs

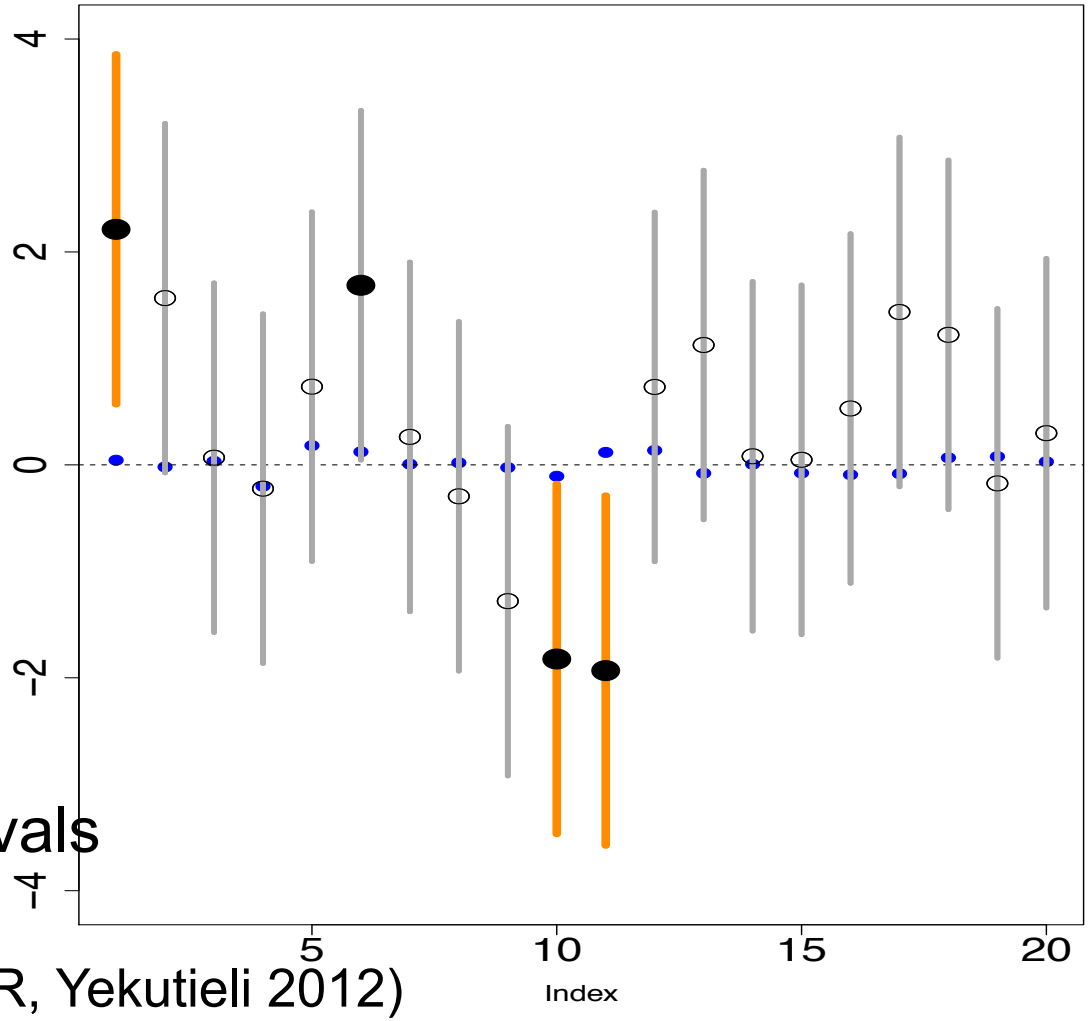
3/20 do not cover

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)



Inference on the average over the selected – selective inference

we wish at least to assure that the property of the individual inference will still hold on the average over the selected

The False Coverage-statement Rate (FCR) of a selective CIs procedure is the expected proportion of coverage statements made that fail to cover their respective parameters

There are general FCR controlling CIs

Selecting from *m* features the 'interesting ones' by S(Y)

#(selected) = |S(Y)|

For *each of the selected ones,*

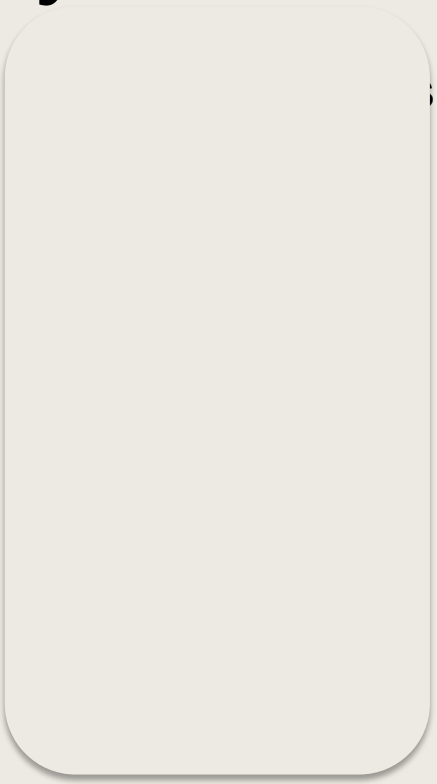
construct a marginal *1- q** |S(Y)| */ m* Conf. Int.

Thm: holds for "simple" selection rules under positive regression dependency

YB & Yekutieli '05

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 CIs is common even in large problems

Alas protecting from the effect of selection in testing does not solve the problem in estimation

In examples

- Pizza and Prostate 3 discoveries

  FWER screening X FCR screening X

- Natalizumab study

  FWER screening X FDR screening V

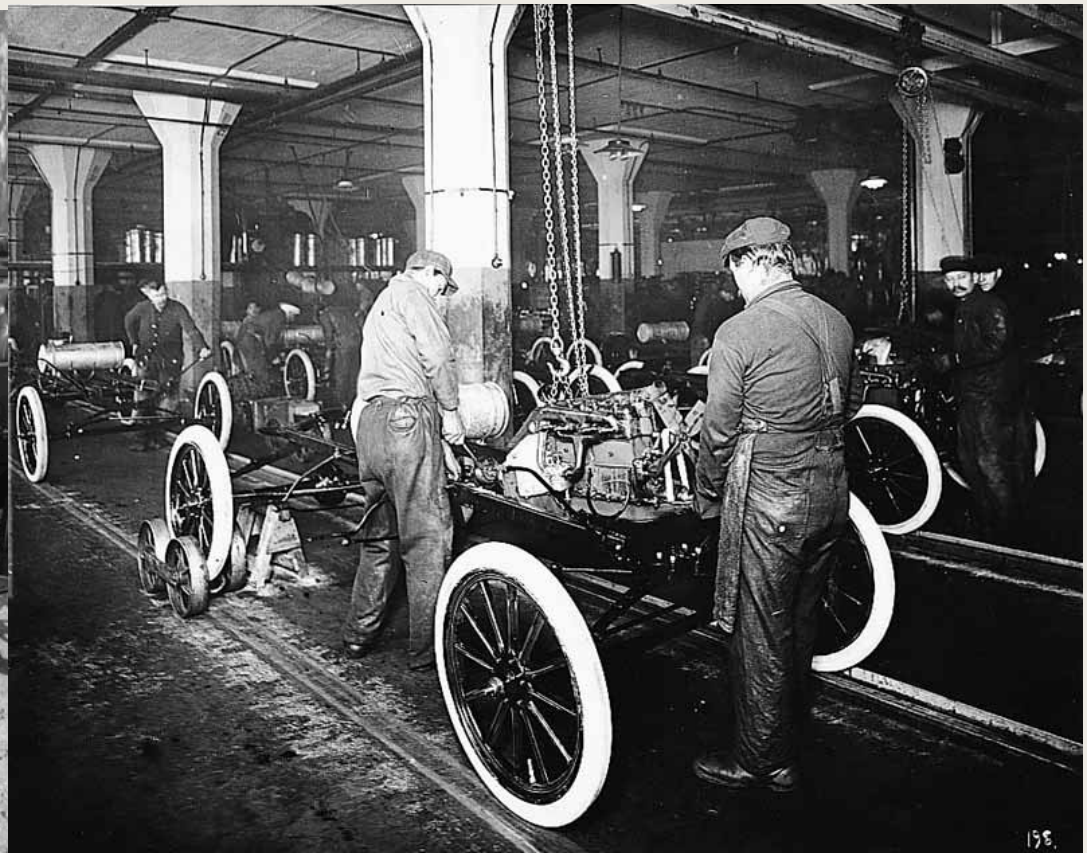
~~27~~ 18

Car manufacturing

In 1902 Olds implemented the first production line.

It manufactured a car every two hours, ~1500 per year:

In 1914 Ford's T-model, 4 cars per hour, ~12,000 per year.

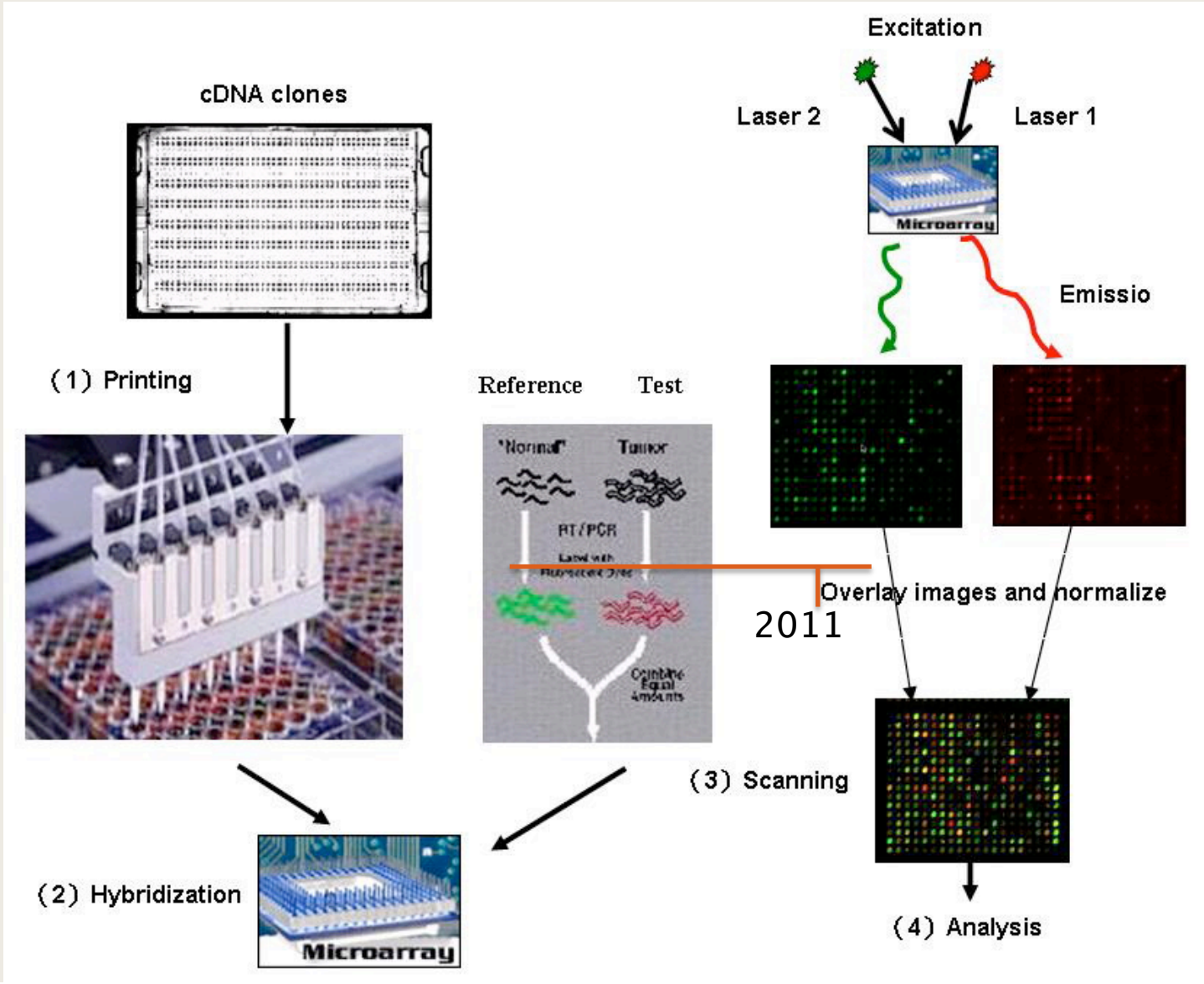


Car manufacturing

The robotic production line started in Japan in the 1980's

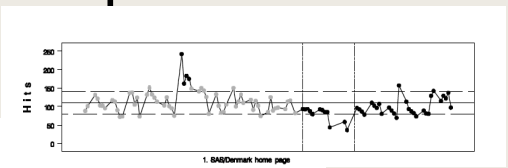
This is the way 70 millions new cars are manufactured each year. People design and supervise.

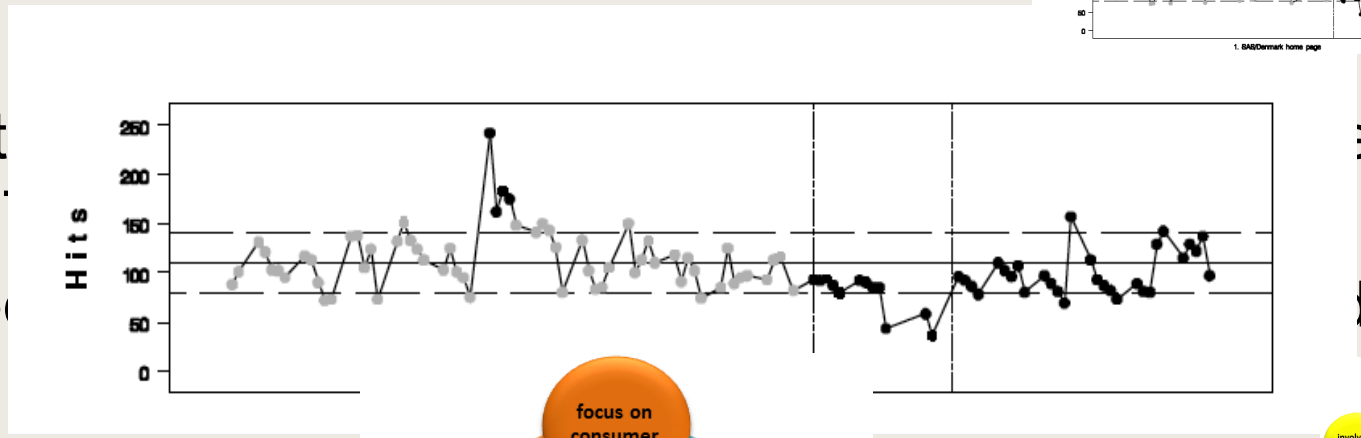
Industrialization of the scientific process: gene expression analysis



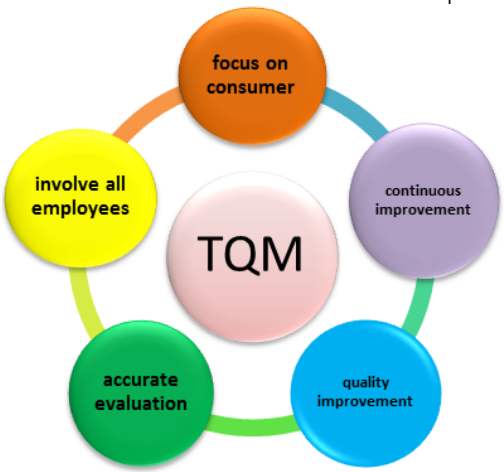
Assuring quality of cars

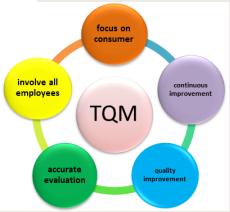
- During the manual manufacturing period – the mechanic
- In the production lines of the 50's – statistical process control





- With t… e 80's TQM … starte… d-wide.



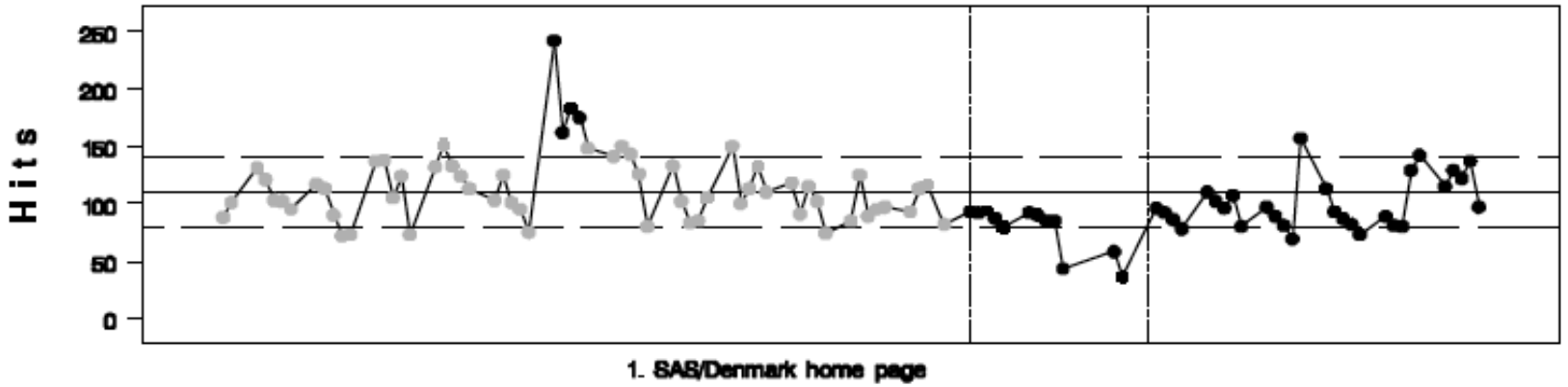


The lesson: methods … riate for manual manufacturing are … riate for mass production.

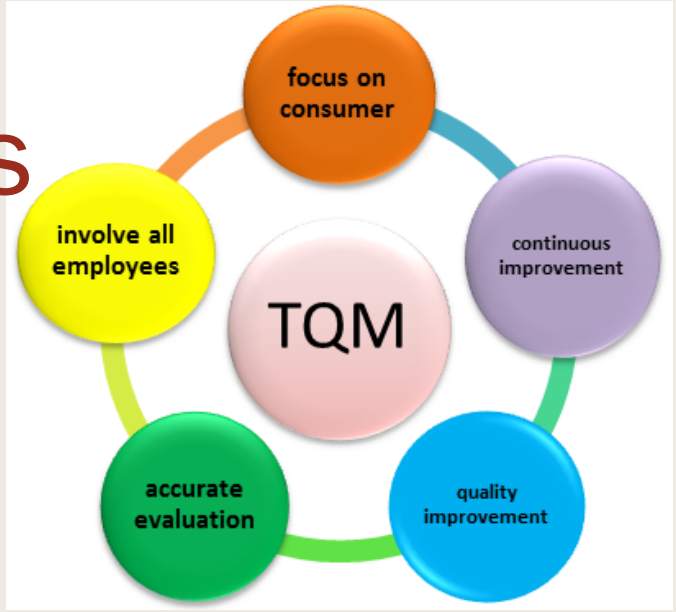
Assuring quality of cars

During the manual manufacturing period – the mechanic

In the production lines of the 50's – statistical process control



Assuring quality of cars



With the beginning of automated production line in the 80's TQM Total Quality Management

started in Japan and the methods were adopted world-wide.

The lesson: methods that were appropriate for manual manufacturing are no longer appropriate for mass production.

Assuring quality of scientific research

Methods developed in the 20's for testing a single hypothesis, like Fisher's 1/20 rule,

or during the 50's for testing a few hypotheses,

are no longer appropriate when used after selection from the huge pool of potential discoveries made available to researchers in modern industrialized research.

New methods are needed, and that's what MCP2013 is about.

Conclusion

Lack of appropriateness of design and transparency in reporting is only a part of the replicability problem

The importance of statistical issues is recognized but sometimes the solutions are ill conceived

Addressing selective inference is the major statistical challenge in assuring replicability

Taking too narrow a view about variability is a second major challenge (not addressed in this talk)

Estimating the science-wise FDR

- Ioanidis wrote about what may happen

- Jager & Leek (‘14) tried to estimate it:
  Mined the Abstracts of 5 top medical journal over 10 years
  Collected all p-values < 0.05; Estimated FDR at ~15%

- Analyzing a sample of 25 papers
  The problem seems more severe (and different).
  # p-value ≤ 0.05 in the paper >> in the abstract, yet in 19 of the 25 papers the smallest p-value in the paper appeared in the abstract. Again, evidence of selection.

  Even more selection with CIs

YB& Hechtlinger ‘14