CONTROLLING THE FALSE DISCOVERY RATE:

A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING

By Yoav Benjamini and Yosef Hochberg (1995)

Presented by: Zhiyu Chen

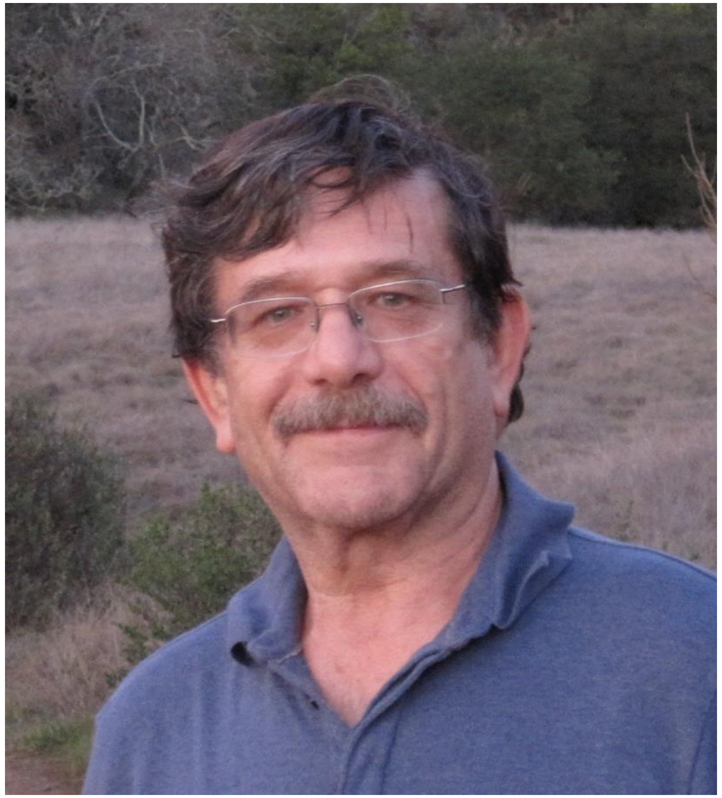
Outline

- Background
- False Discovery Rate
  - Definition
  - Properties
  - Examples
- FDR Controlling Procedure
- Power Comparisons

BACKGROUND

Authors

Authors -- Yoav Benjamini



- Yoav Benjamini (Hebrew: יואב בנימיני; born 5 January 1949) is an Israeli statistician best known for the development of:
  - the false discovery rate (FDR) criterion
  - the Benjamini-Hochberg (BH)
  - The Benjamini-Yekutieli (BY) procedures
- He is currently The Nathan and Lily Silver Professor of Applied Statistics at Tel Aviv University.

Authors – Yosef Hochberg

- Yosef Hochberg (Hebrew: יוסף הוכברג; 1945 – December 3, 2013) was an Israeli statistician and professor of statistics at Tel Aviv University. He is best known for the development (with Yoav Benjamini) of:
  - the false discovery rate (FDR) criterion
  - the Benjamini–Hochberg (BH) procedure for controlling the FDR rate
  - Hochberg's step-up procedure for controlling the family-wise error rate.

BACKGROUND

Multiple Testing and Classical MCPs

Multiple Testing

- Multiple Testing/ Multiplicity/ Multiple Comparison:
  - Consider a set of statistical Inference simultaneously
  - Or estimates a subset of parameters selected
- Multiple comparisons problem:
  - Occurs when conducting multiple comparisons
  - The number of erroneous inference is related to the number of inferences

Classical MCPS

- Classical Multiple Comparison Procedures(MCPs):
  - **Goal:** Control the risk of making false discoveries
  - **Common methods:**
    - Bonferroni Correction -> controls the FWER
    - Holm's Step-Down Procedure (1979)
    - Hochberg's Step-Up Procedure (1988)

Familywise Error Rate (FWER)

- m null hypothesis: $H_1, H_2, \ldots, H_m$

	$\boldsymbol{H_0}$ is True	$\boldsymbol{H_A}$ is True	Total
Test significant	V	S	R
Test non-significant	U	T	m-R
Total	$\boldsymbol{m_0}$	$\boldsymbol{m - m_0}$	m

- FWER is the probability of making at least one Type I error.
- $\mathbf{FWER} = \boldsymbol{Pr(V \geq 1) = 1 - PR(V = 0)}$
- Control: $\boldsymbol{FWER \leq \alpha}$
  - In the weak sense: $\boldsymbol{m_0 = m}$
  - In the strong sense: guarantee for any configuration of true and non-true $\boldsymbol{H_0}$.

Issues with Classical MCPs

- Mismatch with real data
    - Assumption of multivariate normality
- Less powerful than per-comparison approaches
- Not all applications need control of the FWER

FALSE DISCOVERY RATE

False Discovery Rate(FDR) –A New Error Metric

- Inspiration:
  - Spjøtvoll (1972)
  - Sorić (1989)

- Proposal: control FDR – the expected proportion of false discoveries among the rejected hypothesis.

Definition of False Discovery Rate

- The proportion of falsely rejected null hypothesis:
  - $Q = V/(V+S)$
- FDR = expectation of Q = Q_e
  - $Q_e = E(Q) = E\left\{\frac{V}{V+S}\right\} = E(\frac{V}{R})$

	H_0 is True	H_A is True	Total
Test significant	V	S	R
Test non-significant	U	T	m-R
Total	m_0	$m - m_0$	m

Properties of FDR

1. If all null hypotheses are true $(m_0 = m)$, then V=R, and the FDR is equal to FWER.

2. If some null hypotheses are false $m_0 < m$, then $V \leq R, so\ FDR \leq FWER.$

Implication:

- FDR control is less stringent than FWER control.
- Allows for more rejections -> greater statistical power.
- FDR controls balances: the need for scientific discovery & the need to limit false positives.

Examples

1. Clinical Trials with Multiple Endpoints

- Scenario: A new treatment is compared to a standard one.
- Multiple null hypotheses tested on various outcome measures.
- Goal:
  - Make as many discoveries as possible
  - Subject to control of the FDR
- FDR vs. FWER
  - FWER control is too strict – even a few false discoveries may not invalidate the overall conclusion.
  - FDR control allows more flexibility by tolerating a small proportion of false discoveries.

Examples

2. Multiple-Subgroup Problem

- **Scenario:** Compare treatments across different subgroups.
- **Goal:** Make separate treatment decisions for each subgroup.
- **FDR control** is ideal when you expect some misses but still need actionable insights.

3. Screening Problem

- **Scenario:** Screening of various chemicals for potential drug development.
- **Goal:** Weed out uninteresting effects and highlight promising ones.
- **FDR control** ensures that not too many false leads are passed to the costly second phase.

Evaluating Alternatives to FDR

Option 1. Control Q = V/R in each realization

- Problem: When all hypotheses are true, even one rejection means Q = 1.

Option 2. Control (V/R|R>0)

- Problem: Still equals 1 when all hypotheses are true.

Option 3. Soric's Metric: Q = E[V]/r

- Problem: Not a true expectation or conditional expectation -> not controllable in practice.

Option 4. E[V]/E[R]

- Problem: Equals 1 when all nulls are true -> not controllable.

THE FDR CONTROLLING PROCEDURE

The FDR Controlling Procedure

Given m hypotheses $H_1, \dots, H_m$ with corresponding p-values: $P_1, \dots, P_m$

- Step 1: Order the p-values: $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$
- Step 2: Choose a desired FDR level q^*
- Step 3: Let k be the largest i for which: $P_{(i)} \leq \frac{i}{m} \cdot q^*$, then reject all $H_{(i)}, i = 1, 2, \dots, k$. (1)

Theoretical Guarantee of the Procedure

- **Lemma:** For any $\mathbf{0} \leq \boldsymbol{m_1} \leq \boldsymbol{m}$ independent p-values corresponding to true null hypotheses, and for any values that the $\boldsymbol{m_1} = \boldsymbol{m} - \boldsymbol{m_0}$ p-values corresponding to the false null hypotheses can take, the multiple-testing procedure defined by procedure (1) above satisfies the inequality:

$$E\left(Q \middle| P_{m_0+1} = p_1, \dots, P_m = p_{m_1}\right) \leq \frac{m_0}{m} q^*$$

- **Theorem 1.** For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at $\boldsymbol{q}^*$.

Example of FDR Controlling Procedure

- Case Study: Myocardial Infarction Trial (Neuhaus et al., 1992)
  - **Study goal:** Compare rt-PA vs. APSAC for acute myocardial infarction
  - 421 patients, 4 families of hypothesis (focus on d) family)
    - d) Cardiac & other events post-treatments (15 hypotheses)

Example of FDR Controlling Procedure

- **15 p-values:** 0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.0000
- FWER Control
  - Bonferroni's procedure at alpha = 0.05: only 3 smallest p-values are significant.
  - Hochberg's procedure: same result.
- FDR Control ($\boldsymbol{q^* = 0.05}$)
  - Compare each $\boldsymbol{p_{(i)}}$ with 0.05i/15, starting with $\boldsymbol{p_{(15)}}$
  - Found: $p_{(4)} = \mathbf{0.0095} \leq \frac{\mathbf{4}}{\mathbf{15}}\mathbf{0.05} = \mathbf{0.013}$
  - Reject 4 hypotheses with p-values less than or equal to 0.013

Connection to Other Procedures

- Sime's Procedure (1986):
  - Purpose: Test the global null hypothesis that all m null hypotheses are true
  - Steps:
    - Order the p-values: $P_{(1)}, \dots, P_{(m)}$
    - Reject the global null if: $\exists i \; such \; that \; P_{(i)} \leq \frac{i\alpha}{m}$

- Hochberg's Procedure (1988):
  - Purpose: Control the FWER in the strong sense
  - Steps:
    - Order the p-values: $P_{(1)}, \dots, P_{(m)}$
    - Let k be the largest i for which $P_{(i)} \leq \frac{i\alpha}{m+1-i}$, then reject all $H_{(i)}, i = 1, 2, \dots, k$.

Another Look at FDR controlling procedure

-- View FDR Controlling procedure as a maximization problem:

- Theorem 2: The FDR controlling procedure given by expression (1) is the solution of the following constrained maximization problem:
  - Choose α that maximizes the number of rejections at this level, $r(\alpha)$, subject to the constraint $\frac{\alpha m}{r(\alpha)} \leq q^*$.
  - Where:
    - $r(\alpha)$: number of hypotheses rejected at level α.
    - m: total number of hypotheses
    - q^*: desired FDR Level

Power Comparisons

- Compare the FDR Controlling procedure with
  - The Bonferroni's Method (FWER control)
  - The Hochberg's Method (FWER control)

- $\boldsymbol{q^* = \alpha = 0.05}$

- Goal: Evaluate how many true effects each method can detect, under various conditions

Power Comparisons

- Simulation Design:
  - Testing of m = {4, 8, 16, 32, 64} hypotheses.
  - Proportions of true nulls: 3m/4, m/2, m/4 and 0.
  - Non-zero means: grouped at L/4, L/2, 3L/4 and L, with L = 5 and L = 10.
  - Configurations of effect size distribution:
    - D: Linearly decreasing number of hypotheses away from 0 in each group.
    - E: Equal number of hypotheses in each group.
    - I: Linearly increasing number of hypotheses away from 0 in each group.

Findings:

1. The power of all the methods decreases when the number of hypotheses tested increased.

2. The FDR Controlling procedures consistently shows higher average power.

3. The advantage increase with:

    a. The number of non-null hypothesis.

    b. m.

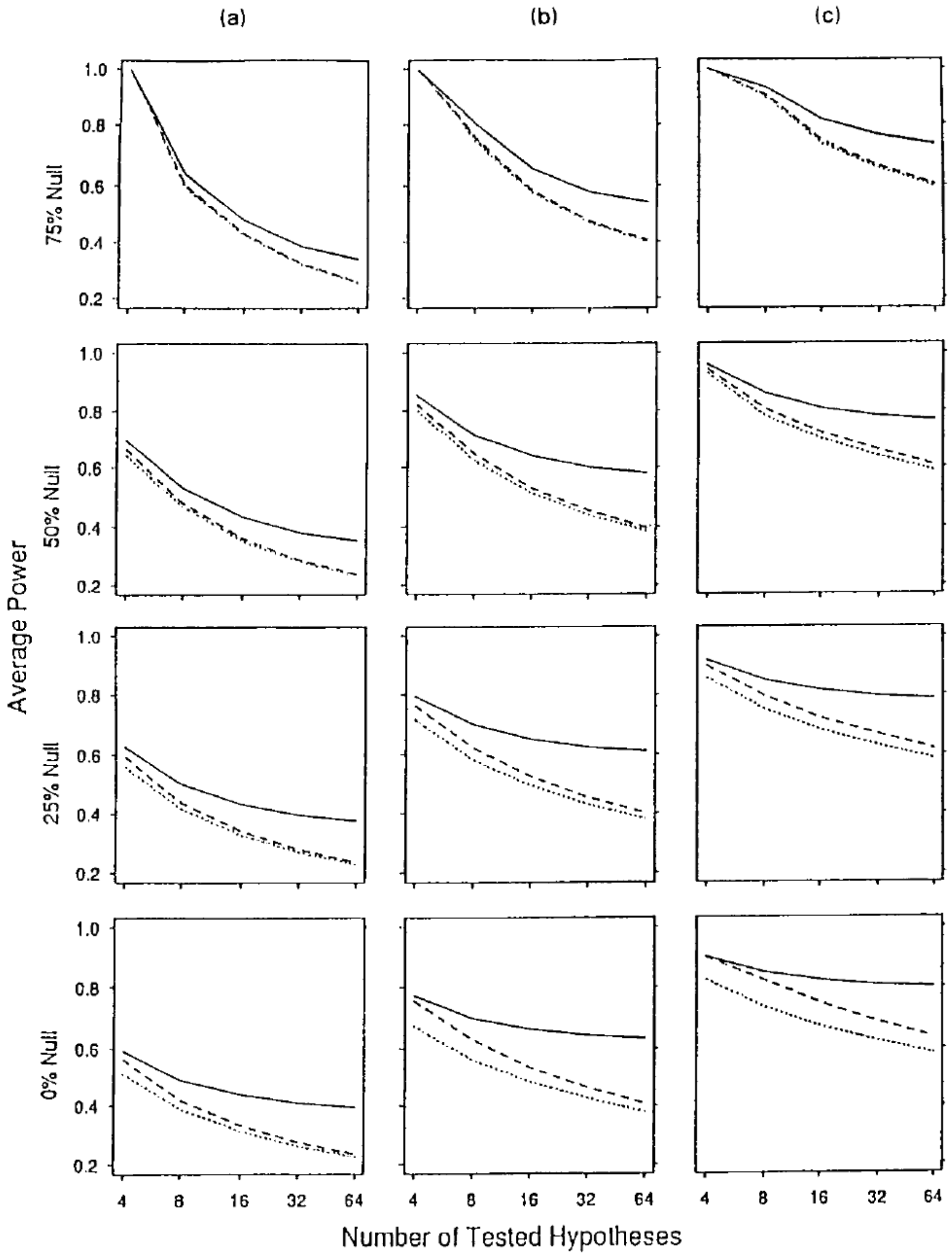


Fig. 1. Simulation-based estimates of the average power (the proportion of the false null hypotheses which are correctly rejected) for two FWER controlling methods, the Bonferroni (······) and Hochberg's (1988) (-----) methods, and the FDR controlling procedure (———): (a) decreasing; (b) equally spread; (c) increasing

Summary – Controlling the FDR

- Problem:
  - Traditional multiple testing methods control Familywise Erro Rate(FWER).
  - FWER is too conservative in large-scale testing -> low power.
- Contribution:
  - Introduce a new criterion: False Discovery Rate(FDR) = $\boldsymbol{E}\left[\frac{V}{R}\right]$ = Expected proportion of false discoveries among rejections.
  - FDR-controlling has higher power than FWER-controlling methods.
  - FDR-controlling is especially effective when many nulls are false.

THANK YOU