Multiple testing adjustments

mostad@chalmers.se

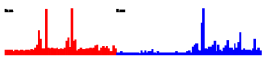
The multiple testing problem

- ”Rejecting a hypothesis at 5% significance level”: There is a 5% chance of rejecting a true hypothesis.
- Rejecting 10 hypotheses at 5%: There may be up to 50% chance of an incorrect rejection

High-throughput experiments: The problem becomes acute

Microarray data: 10.000s of hypotheses →

Example: Using co-expressed gene clusters to hunt for cis-regulatory elements (Nelander 2005)



Bonferroni P-value	Cluster coverage	Dataset coverage	GO term
1.00E+00	43% (3/7)	(56/9561)	motor activity
1.00E+00	29% (2/7)	(10/9561)	actin filament
1.00E+00	29% (2/7)	(23/9561)	muscle development
1.00E+00	29% (2/7)	(74/9561)	structural constituent of cytoskeleton

#	PFM	Cluster coverage	Dataset coverage	FDR
1	MA0083:SRF	86% (6/7)	1% (89/9561)	<2.5%
2	M00196:SRF|M00215:SRF	86% (6/7)	1% (114/9561)	<2.5%
3	M00152:SRF	71% (5/7)	1% (135/9561)	<2.5%
4	M00216:TATA	43% (3/7)	2% (220/9561)	<20%
5	M00059:YY1	43% (3/7)	3% (284/9561)	<20%
6	MA0090:TEF-1	43% (3/7)	3% (296/9561)	<20%

1	2	3	4	5	6	gene (MM)	gene (MS)	Annotation (mouse)
						Actg2	ACTG2	ACTIN, GAMMA 2, SMOOTH MUSCLE, ENTERIC
						Q9C330	MYH11	MYOSIN HEAVY CHAIN 11
						Lpp	LPP	LIPOMA PREFERRED PARTNER
						Tagln	TAGLN	TRANSGELIN (SMOOTH MUSCLE PROTEIN 22-ALPHA)
						Acta2	ACTA2	ACTIN, AORTIC SMOOTH MUSCLE (ALPHA-ACTIN 2)
						Myl9	MYL9	MYL9 PROTEIN (FRAGMENT)
						Lmod1	LMOD1	LEIOMODIN 1 (SMOOTH MUSCLE)

Many other examples!

Setup and notation:

Let S be the sample space of possible realities, and let $\theta \in S$.

Let $H = (H_1,...,H_N) : S \to \{0,1\}^N$ be a function specifying N "Hypotheses", where $H_i(\theta) = 0$ or 1 means that the hypothesis is false or true, respectively.

For every $\theta \in S$, let $T(\theta) = \left(T_1(\theta),...,T_N(\theta)\right)$ be a stochastic variable on $[0,1]^N$
$T(\theta)$ represents the collection of "test statistics", or more accurately the collection of resulting p-values, as we assume:

For any θ and any i such that $H_i(\theta) = 1$: $\quad T_i(\theta) \sim UNIFORM[0,1]$

Goal of analysis

Based on the test statistics (or p - values) $T(\theta)$, we want to predict the values of $H(\theta)$. In other words :

For a function $f : [0,1]^N \to \{0,1\}^N$ predicting values for $H(\theta)$ from $T(\theta)$, we study the error, i.e., we study the stochastic variable defined, for given θ and f, by

$Err(\theta, f) = (V, Z)$ where $V = \sum_{i=1}^{N} v_i$ and $Z = \sum_{i=1}^{N} z_i$ where

$v_i = H_i(\theta)(1 - f_i(T(\theta)))$

$z_i = (1 - H_i(\theta))\, f_i(T(\theta))$

Example

The Type I and Type II error rates are, for given values of θ and f, the expectations of V and Z, respectively :

$$E(V) = \sum_{i=1}^{N} E(v_i) = \sum_{i=1}^{N} H_i(\theta)\left(1 - E\left(f_i(T(\theta))\right)\right)$$

$$E(Z) = \sum_{i=1}^{N} E(z_i) = \sum_{i=1}^{N} (1 - H_i(\theta)) E\left(f_i(T(\theta))\right)$$

Note that, as usual, we cannot make computations for Type II errors whithout making more assumptions about the distribution of $T(\theta)$

Example

For a given $\alpha > 0$, define $f_\alpha : [0,1]^N \to \{0,1\}^N$ by

$$f_i(u) = \begin{cases} 0 & u_i < \alpha \\ 1 & u_i \geq \alpha \end{cases}$$

Then

$$E(V) = \sum_{i=1}^{N} H_i(\theta)\alpha \leq N\alpha$$

The family-wise error rate (FWER)

The FWER is defined, for given values of θ and f, as the probability $\Pr(V > 0)$

It measures, for the whole "family" of hypotheses, the probability of one or more Type I errors.

EXAMPLE: For f_α defined as above, we get

$$FWER = \Pr(V > 0) \leq \sum_{i=1}^{N} \Pr(v_i = 1)$$

$$= \sum_{i=1}^{N} H_i(\theta) \Pr\big(f_i(T(\theta)) = 0\big) = \sum_{i=1}^{N} H_i(\theta)\alpha \leq N\alpha$$

The Bonferroni correction

For a given $\alpha > 0$, define $f_{B,\alpha} : [0,1]^N \to \{0,1\}^N$ by

$$f_i(u) = \begin{cases} 0 & u_i < \alpha / N \\ 1 & u_i \geq \alpha / N \end{cases}$$

Then

$$E(V) = \sum_{i=1}^{N} H_i(\theta)\alpha / N \leq \alpha \quad \text{and}$$

$$FWER = \Pr(V > 0) \leq \sum_{i=1}^{N} H_i(\theta) \Pr\left(f_i(T(\theta)) = 0\right) = \sum_{i=1}^{N} H_i(\theta)\alpha / N \leq \alpha$$

The Bonferroni correction is thus said to control for FWER at level α

The Holm method

For a given $\alpha > 0$, define the Holm method $f_{H,\alpha} : [0,1]^N \to \{0,1\}^N$ by

- Sort the indices so that $u_1 \leq u_2 \leq \ldots \leq u_N$
- For i = 1,2,..., set $f_i(u) = 0$ as long as $u_i < \dfrac{\alpha}{N - i + 1}$ then set $f_i(u) = 1$ for the rest.

We get (by conditioning on $T(\theta)$ and reordering indices) :

$$FWER = \Pr(V > 0) \leq \sum_{i=1}^{N} H_i(\theta) \Pr\left(f_i(T(\theta)) = 0\right)$$

$$= \sum_{i=j}^{N} H_i(\theta) \frac{\alpha}{N - j + 1} \leq (N - (j-1)) \frac{\alpha}{N - (j-1)} = \alpha$$

Thus the Holm method controls FWER at level α

Adjusted p-values

Assume a function $f_{M,\alpha} : [0,1]^N \to \{0,1\}^N$ can be written as $f_{M,\alpha}(u) = f_\alpha(F(u))$, where f_α is the function defined before and $F : [0,1]^N \to [0,1]^N$ is some function.
Then F is called a p - value adjustment.

With adjusted p - values, one can "reject" and "accept" Hypotheses just as usual based on the adjusted p - values, while still getting for example control over FWER.

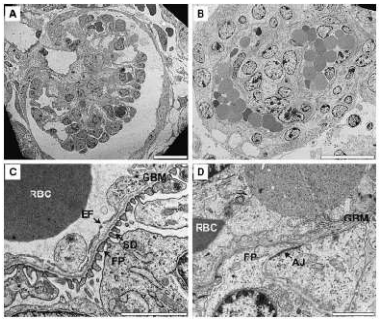
Examples

The function
$F(u_1,...,u_N) = (\min(1, Nu_1),..., \min(1, Nu_N))$
computes adjusted p - values for the Bonferroni method.

The function defined by first ordering p - values in increasing order and then computing
$F_j(u) = \max_{k=1,...,j}\left(\min\left(1, (N-k+1)u_k\right)\right)$
computes adjusted p - values for the Holm method.

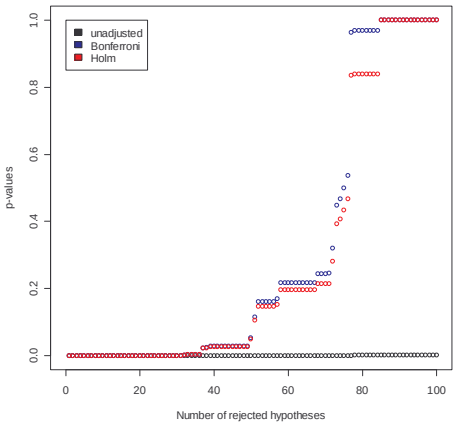
Example: EST mining

- Gene expression in the glomerulus in the kidney
- Libraries of ESTs were made from both newborn and adult mouse glomerulus
- Comparison with libraries from whole kidney to find glomerulus enrichment



Takemoto et.al.:Large-scale identification of genes implicated in kidney glomerulus development and function
He et.al.:Analysis of 15,000 mouse glomerular EST and identification of novel glomerular enriched genes

- For 573 genes with more than one EST in the glomerulus library
  - Hypotheses $H_1,\ldots,H_{573}$: there is no diff. exp.
  - Comparison between libraries for each gene: Test statistics $T_1,..,T_{573}$ from Fisher test.
  - We get unadjusted p-values $p_1,\ldots,p_{573}$
  - Adjusted p-values



Sidák adjusted p-values

The function defined by

$F_i(u) = 1-(1-u_i)^N$

computes adjusted p - values for the Sidák method.

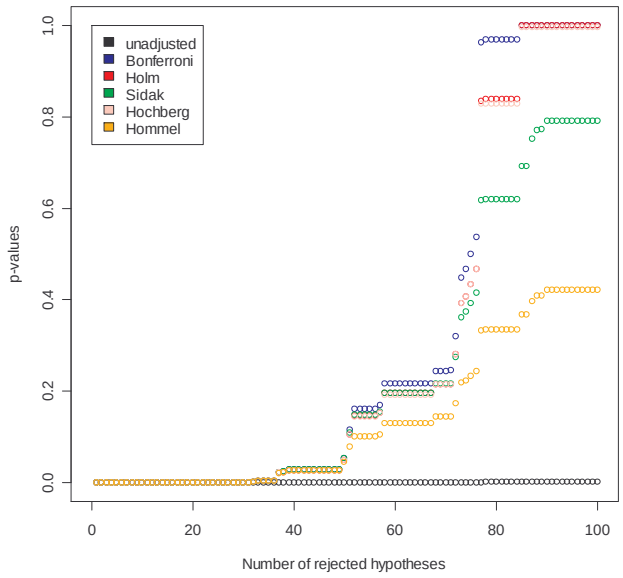
If we assume that the components of $T(\theta) = (T_1(\theta),...,T_N(\theta))$ are independent, one can easily show that this method controls FWER at level α. This can also be proven even under somewhat more general circumstances.

Other procedures controlling FWER

- Hochberg adjusted p-values (on sorted u_k):

$$F_i(u) = \min_{k=i,...,N}\left[\min((N-k+1)u_k, 1)\right]$$

- There is also a method by Hommel, and various other methods.
- They all require some assumption about the dependency in T(θ) to control FWER



The False Discovery Rate (FDR)

In addition to the stochastic variables V and Z defined above,

define a stochastic variable $R = \sum_{i=1}^{N} (1 - f_i(T(\theta)))$, and then

define Q as follows:

$$Q = \begin{cases} \dfrac{V}{R} = \dfrac{\sum_{i=1}^{N} H_i(\theta)(1 - f_i(T(\theta)))}{\sum_{i=1}^{N} (1 - f_i(T(\theta)))} & \text{when } R > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then the FDR is defined as the expectation of Q

(for fixed θ and f).

As for FWER, we can define adjusted p - values controlling for FDR

Examples

The Benjamini and Hochberg adjustment :

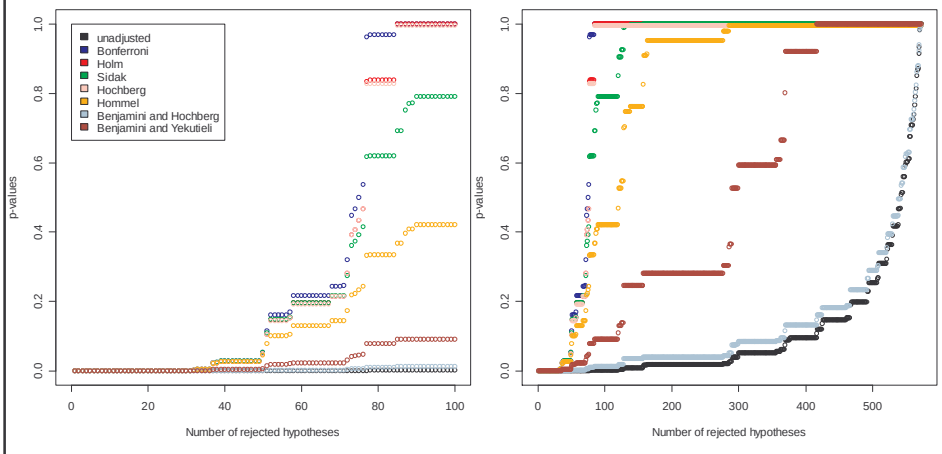
- Sort the indices so that $u_1 \leq u_2 \leq \ldots \leq u_N$
- Define $F_i(u) = \min_{k=i,\ldots,N} \left(\min(\frac{N}{k} u_k, 1) \right)$

This controls for FDR under some assumptions

The Benjamini and Yekutieli adjustment :

- Sort the indices so that $u_1 \leq u_2 \leq \ldots \leq u_N$
- Define $F_i(u) = \min_{k=i,\ldots,N} \left(\min(\frac{N}{k} (1 + \frac{1}{2} + \ldots + \frac{1}{N}) u_k, 1) \right)$

This always controls for FDR



Implementations

- Methods producing adjusted p-values from unadjusted p-values are easy to implement.
- In R, look at the function **p.adjust(...)**

Dependencies between test statistics

- The methods above focus on controlling various types of Type I error rates.
- To improve error bounds further, one needs to estimate the dependency structure in $T(\theta)$.
- This can sometimes be done using permutations of the data, when the test statistics are invariant under such permutations, assuming the null hypotheses.

Step-down max T adjusted p-values (Westfall and Young)

- Order hypotheses so that |T| is decreasing
- Do permutations of columns of data matrix:
  - Compute test statistic for each hypothesis
  - Adjust these, starting at the last, so that they are decreasing
- Estimate adjusted p-values as quantiles of observed |T| in simulated |T|'s for each hypothesis
- Enforce that adjusted p-values are increasing

The bioconductor multtest package

- This package implements a number of methods based on permutation and simulation:
  - Simple p-value adjustments
  - Step-down max T
  - Step-down min p
  - …

Different error rates:

- Family-wise error rate: $FWER = \Pr(V > 0)$
- False discovery rate: $FDR = E(V / R \mid R > 0)\Pr(R > 0)$
- Positive false discovery rate: $pFDR = E(V / R \mid R > 0)$
- Per comparison error rate: $PCER = E(V)/N$
- Per family error rate: $PFER = E(V)$

”Strong” control versus ”weak” control

Example: SAM: Finding differentially expressed genes

- Order hypotheses so that |T| is decreasing
- Use permutations to estimate the *expected* decreasing sequence of test statistics, under complete null hypothesis
- Form a qq-plot (SAM-plot) and select genes that are further than Δ away from the diagonal
- Estimate PFER by averaging over permutations.

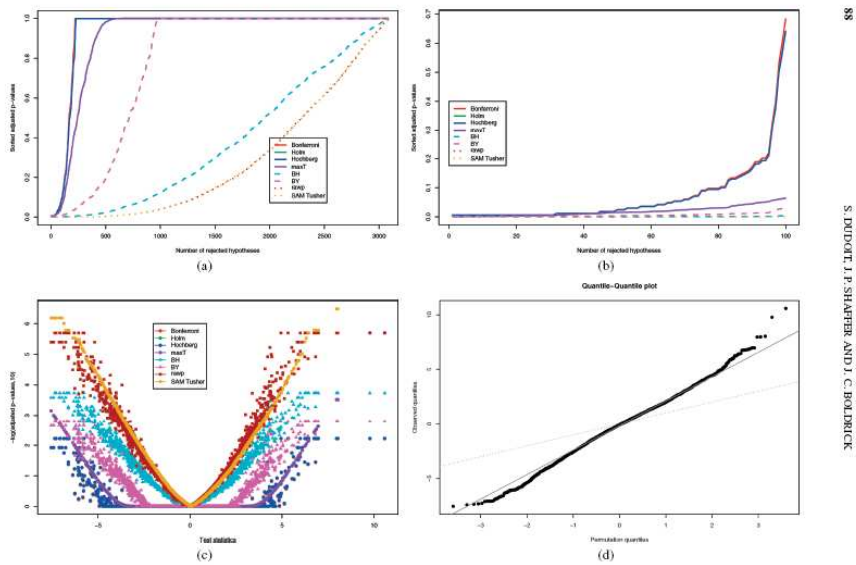


FIG. 5. *Leukemia study.* (a) *and* (b) *Plot of sorted permutation adjusted p-values,* $\tilde{p}^*_{(j)}$ *versus* j. *Panel* (b) *is an enlargement of panel* (a) *for the* 100 *genes with the largest absolute t-statistics* $|t_j|$. *Adjusted p-values for procedures controlling the* FWER, FDR *and* PCER *are plotted using solid, dashed and dotted lines, respectively.* (c) *Plot of adjusted p-values,* $-\log_{10}\tilde{p}^*_j$ *versus t-statistics* t_j. (d) *Quantile–quantile plot of t-statistics; the dotted line is the identity line and the dashed line passes through the first and third quartiles. Adjusted p-values were estimated based on* $B_{\text{perm}} = 500{,}000$ *random permutations of the ALL/AML labels, except for the SAM procedure for which* $B_{\text{sam}} = 1000$ *random permutations were used. Note that the results for the Bonferroni, Holm and Hochberg procedures are virtually identical, similarly for the unadjusted p-value (rawp) and SAM Tusher procedures.*

Comparisons of methods

- Classical statistical approach: To *prove inequalities* for *type I* error rates for given procedures
- Practical approach: Find actual error rates for real data, or under reasonable hypotheses (simulation studies)

References

- Dudoit, Shaffer, Boldrick: ”Multiple Hypothesis Testing in Microarray Experiments” (Stat. Sci. 2003)
- www.r-project.org, www.bioconductor.org
- Scott, Berger: ”An exploration of aspects of Bayesian multiple testing” (2003)

”Cheating” with FDR

How:

- You have a number of hypotheses you want to reject, but p-values are not quite good enough.
- Add to your hypotheses a number of untrue hypotheses, with low p-values.
- The number of rejections will rise, but not the number of false rejections, so your FDR improves, and you ”prove” the hypotheses you care about.