High-throughput Testing

Noah Simon and Richard Simon

July 2016

Testing vs Prediction

On each of n patients measure

y_i - single binary outcome

(eg. progression after a year, PCR)

$\mathbf{x}_i$ - p-vector of features

(eg. SNPs, gene expression values)

Want to test for x_j with different means in the two classes; for

- Variable selection in predictive modeling
- Learning underlying biology

Testing for a single feature

For a single j calculate two-sample t-statistic:

$$T_j = \frac{\bar{x}_j^{(c)} - \bar{x}_j^{(d)}}{s_j},$$

s_j is your favorite estimate of standard error

Compare to the cutoff of corresponding t-distribution

Reject if T_j is sufficiently large

Testing many features

With many tests we need to think more carefully about error

Do we want to limit

- probability of even a single false rejection?

  familywise error rate

- expected proportion of false rejections?

  false discovery rate

Controlling familywise error rate

Find t so that

$$P_{H_0}(\text{any} \quad T_j > t) \leq \alpha$$

Note.

$$P_{H_0}(\text{any} \quad T_j > t) = P_{H_0}(max\, T_j > t)$$

For independent statistics, this gives us “Sidak’s procedure”:

Reject H_j if $p_j \leq 1 - (1-\alpha)^{1/(\#\text{tests})}$

What about under dependence?

eg. What if the expression values are dependent (with unknown structure)?

Conservative Estimate (Bonferroni)

$$P\left(max\ T_j > t\right) \leq \left(\#tests\right) * P\left(T > t\right)$$

Gives us the test:

Reject H_j if $p_j \leq \frac{\alpha}{(\#\text{tests})}$

Improvements

This can be improved using the "Holm" procedure:

1. Order the p-values (lowest to highest) $p_{(1)}, p_{(2)}, \ldots$
2. Find the first k with

$$p_{(k)} > \frac{\alpha}{(\#\text{tests}) + 1 - k} \qquad \left[\text{vs} \quad \frac{\alpha}{(\#\text{tests})}\right]$$

3. reject hypotheses corresponding to $p_{(1)}, \ldots, p_{(k)}$

Less conservative; not much less though

False Discovery Rates

Family-wise Error Rate vs False Discovery Rate

If we call 50 features significant, may not care about 1 or 2 false positives.

Care more about

$$FDP = \frac{\#\text{ False Rejections}}{\#\text{ Total Rejections}}$$

and

$$FDR = E[FDP].$$

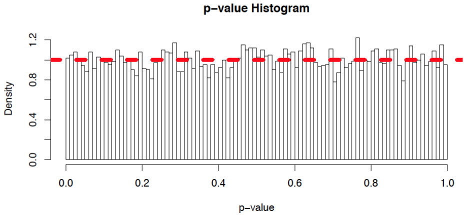
Estimating FDR

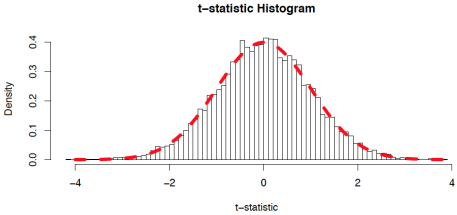
How many rejections do I expect if I:	
Run 100 null tests at 0.05 level?	(5)
How about for 1000 tests?	(50)
How about p tests, at level α?	$(\alpha \times p)$

Estimating FDR

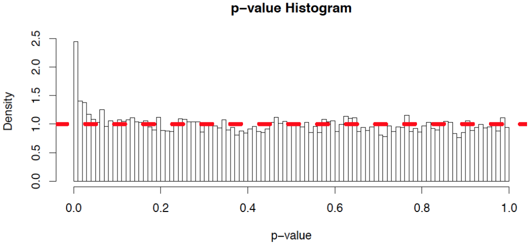
What's a reasonable FDR estimate if I:	
Expect 5 significant results under a global null, and see 20	$(1/4)$
Run 10000 tests, at level 0.001 and find 20 significant	$(1/2)$
Run p tests, at level α and find k significant	$(p\alpha/k)$

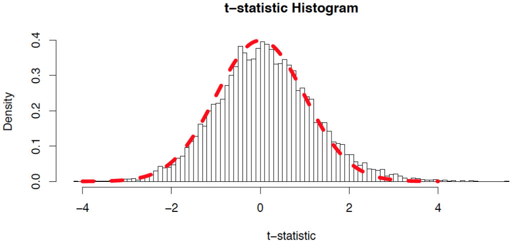
Under Global Null



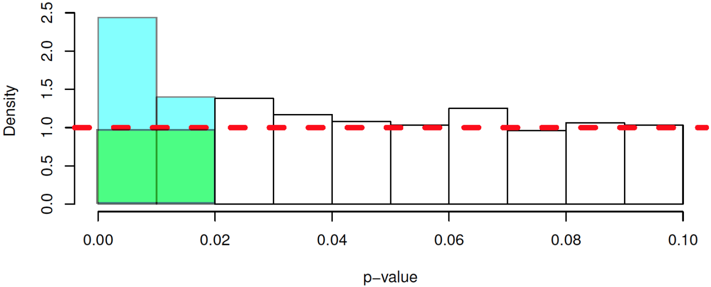


With 1000 non-zero δ_j of varying size





FDR estimate



Formally

Benjamini and Hochberg (under independence/positive dependence):

Find the maximum order statistic (k) such that

$$\frac{p_{(k)} * (\#\text{tests})}{k} \leq \alpha$$

Reject all j with $p_j < p_{(k)}$.

This controls FDR at α.

Comparison to Bonferroni

Benjamini and Hochberg:

Find the maximum order statistic (k) such that

$$p_{(k)} \leq \frac{\alpha k}{(\#\text{tests})}$$

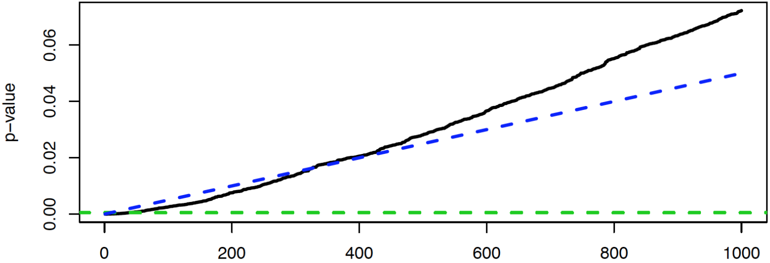
Reject all j with $p_j < p_{(k)}$.

As opposed to Bonferroni:

Reject p_j if

$$p_j \leq \frac{\alpha}{(\#\text{tests})}$$

FDR estimate



Formally

Benjamini and Yekutieli (under arbitrary dependence):

Find the maximum order statistic (k) such that

$$\frac{p_{(k)} * (\#\text{tests}) \left[\sum_{i=1}^{(\#\text{tests})} 1/i\right]}{k} \leq \alpha$$

Reject all j with $p_j < p_{(k)}$.

This controls *FDR* at α under arbitrary dependence.

---

note. $\sum_{i=1}^{m} 1/i \approx \log(m)$

Significance Analysis of Microarrays (SAM)

For BH, use $\alpha * (\#\text{tests})$ to estimate number of false positives.

SAM cleverly uses permutations:

For a cutoff t, want to estimate $\mathrm{E}\left[\#\left\{T_j > t\right\}\right]$:

1. Permute class labels
2. With the new labels calculate a null set of statistics $T_1^{null}, \ldots, T_{(\#\text{tests})}^{null}$
3. calculate the number of these null statistics that exceed t.

Run the above many times, and average the number of exceedences.

Estimation

For ease of exposition, assume we have a pooled se, and equal class sizes.

Can think of

$$T_j/\sqrt{n} = \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}}{\sqrt{n}s_j} \dot{\sim} N\left(\delta_j, 1/n\right),$$

where

$$\delta_j = \frac{\mu_j^{(1)} - \mu_j^{(2)}}{\sigma_j}$$

δ_j quantifies the separation between the two classes for feature j.

A reasonable measure of practical significance

A bad way to estimate δ_j

Suppose we

1. Calculate our many t-statistics
2. use Benjamini-Yekutieli procedure (with FDR of 0.01) and find 10 significant features

How do we estimate their corresponding δs?

How about with $\hat{\delta}_j = T_j/\sqrt{n}$?

NO. This induces a systematic bias.

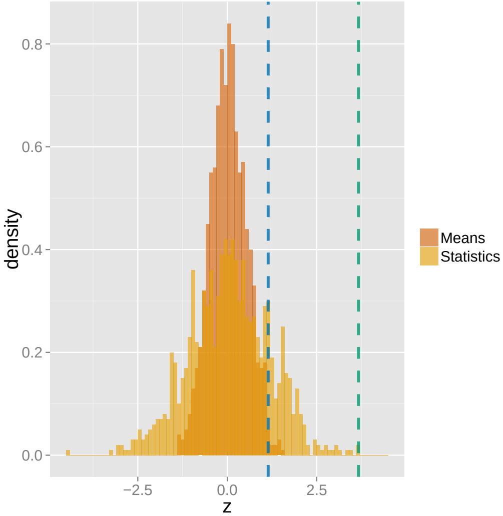
Selection Bias / Multiplicity

We have selected the most extreme statistics

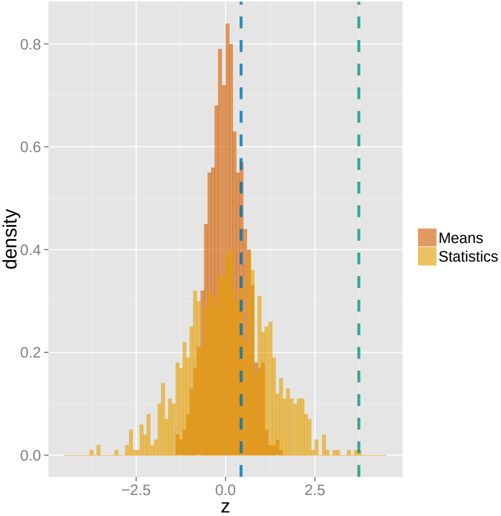
While we have adjusted for this in testing if $\delta_j = 0$...

We must also use an adjustment in estimating nonzero δ_j.

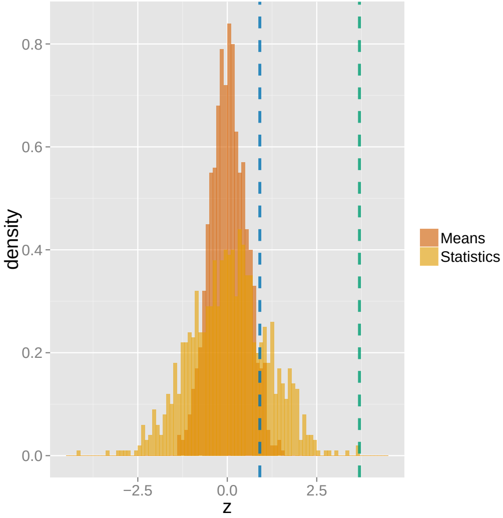
Winner’s Curse



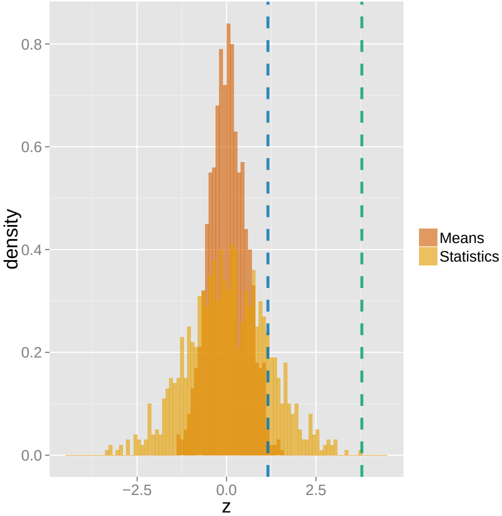
Winner’s Curse



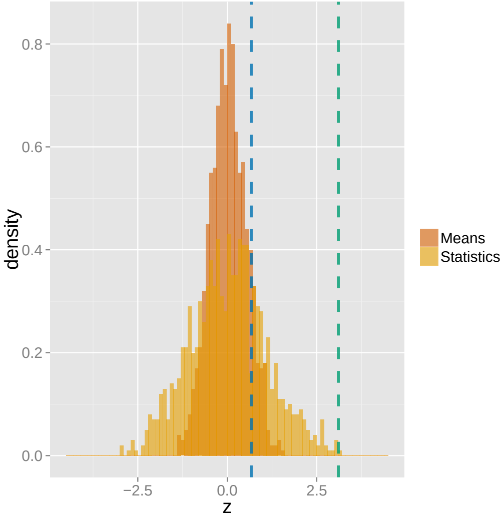
Winner’s Curse



Winner's Curse



Winner's Curse



Correcting Selection Bias

One popular correction approach uses Empirical Bayes (Efron)

Assume that $\delta_j \sim g(\cdot)$ for some prior g.

We observe $T_j = \delta_j + N(0, 1/n)$

This implies $T_j \sim f(\cdot)$ with $f = \phi * g$

Use a smoother to estimate f by $\hat{f}$ from data

Deconvolve $\hat{f}$ and ϕ to get $\hat{g}$.

Calculate bayesian posterior with prior $\hat{g}$

Empirical Bayes Correction

Actually correct from a frequentist viewpoint (compound decision theory)

Assumes independence (small - moderate departures ok in practice)

Decent R support.

Takeaways

Multiplicity Correction is important in testing:

- Family-wise Error Rate control (often too conservative)
- False Discovery Rate control (more appropriate)

Also need to adjust in effect-size estimation!

- Empirical Bayes