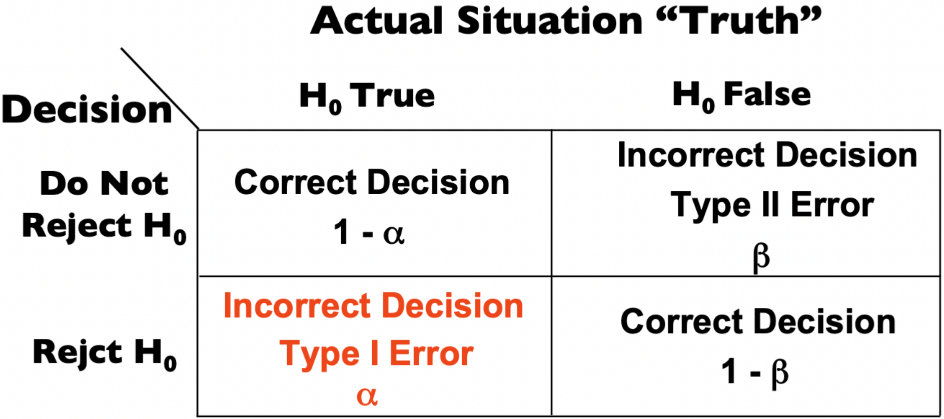
BIOS 835: Multiple Testing

November 28, 2023

Type I and II Errors



	Actual Situation "Truth"	
Decision	H_0 True	H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$\alpha = \Pr(\text{Type I Error})$ $\quad \beta = \Pr(\text{Type II Error})$

Why Multiple Testing Matters

Genomics/Neuroimaging/Big Data = Lots of Data = Lots of Hypothesis Tests

- A typical microarray experiment might result in performing 10000 separate hypothesis tests.
- If we use a standard p-value cut-off of 0.05, we'd expect 500 genes to be deemed "significant" by chance.

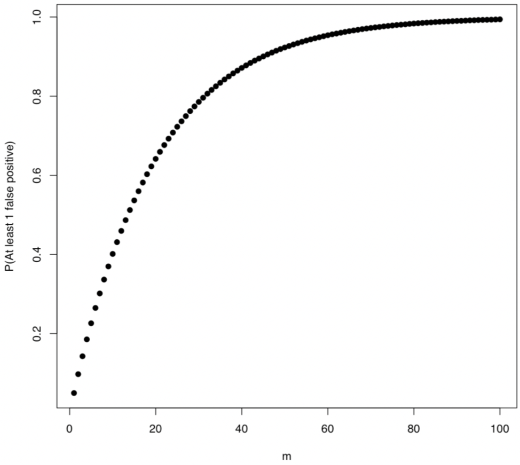
Why Multiple Testing Matters

Genomics/Neuroimaging/Big Data = Lots of Data = Lots of Hypothesis Tests

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

$$\begin{aligned}
\text{Pr(Making an error)} &= \alpha \\
\text{Pr(Not making an error)} &= 1-\alpha \\
\text{Pr(Not making an error in } m \text{ tests)} &= (1-\alpha)^m \\
\text{Pr(Making at least 1 error in } m \text{ tests)} &= 1-(1-\alpha)^m
\end{aligned}$$

Probability of At Least 1 False Positive



Counting Errors

Assume we are testing $H_1, H_2, \ldots, H_M$, where $H_m = 1$ if the mth test is non-null.

	Not Rejected	Rejected	Total
$H_m = 0$	T	$\boldsymbol{V}$	M_0
$H_m = 1$	U	S	M_1
	$M - R$	R	M

- $M_0 = \#$ of True null hypotheses
- $M_1 = \#$ of False null hypotheses (Signals)
- $\boldsymbol{R} = \#$ rejected hypotheses
- $\boldsymbol{V} = \#$ of Type I errors (false positives)

What Does Correcting for Multiple Testing Mean?

- When people say “adjusting p-values for the number of hypothesis tests performed” what they mean is controlling the Type I error rate
- Very active area of statistics - many different methods have been described
- Although these varied approaches have the same goal, they go about it in fundamentally different ways

Different Approaches To Control Type I Errors

- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses,

$$PCER = E(V)/M$$

- Per-family error rate (PFER): the expected number of Type I errors,

$$PFE = E(V).$$

- Family-wise error rate: the probability of at least one type I error

$$FEWR = P(V \geq 1)$$

Different Approaches To Control Type I Errors

- False discovery rate (FDR) is the expected proportion of Type I errors among the rejected hypotheses
$$FDR = E(V/R|R > 0)P(R > 0)$$
- Positive false discovery rate (pFDR): the rate that discoveries are false
$$pFDR = E(V/R|R > 0)$$
- Marginal false discovery rate (mFDR): the expected number of Type I errors over the expected number of rejections
$$mFDR = E(V)/E(R)$$
generally $mFDR > FDR$.

Digression: p-values

- Implicit in all multiple testing procedures is the assumption that the distribution of p-values is "correct"
- This assumption often is not valid for genomics/neuroimaging data where p-values are obtained by asymptotic theory (and are unadjusted)
- Thus, resampling methods are often used to calculate calculate p-values, though this doesn't account for the lack of adjustment.

Permutations

1. Analyze the problem: think carefully about the null and alternative hypotheses
2. Choose a test statistic
3. Calculate the test statistic for the original labeling of the observations
4. Permute the labels and recalculate the test statistic
    - Do all permutations: Exact Test
    - Randomly selected subset: Monte Carlo Test
5. Calculate p-value by comparing where the observed test statistic value lies in the permuted distributed of test statistics

Example: What to Permute?

1. Gene expression matrix of m genes measured in 4 cases and 4 controls

Gene	Case 1	Case 2	Case 3	Case 4	Control 1	Control 2	Control 3	Control 4
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M	X_{M1}	X_{M2}	X_{M3}	X_{M4}	X_{M5}	X_{M6}	X_{M7}	X_{M8}

FWER

- Many procedures have been developed to control the Family Wise Error Rate (the probability of at least one type I error):
- Two general types of FWER corrections:
  1. **Single step:** equivalent adjustments made to each p-value
  2. **Sequential:** adaptive adjustment made to each p- value

Single Step Approach: Bonferroni

- Very simple method for ensuring that the overall Type I error rate of α is maintained when performing m independent hypothesis tests
- Rejects any hypothesis with p-value $\leq \alpha/M$.
- The adjusted p-values are:
$$\tilde{p}_j = \min(Mp_j, 1)$$
- For example, if we want to have an experiment wide Type I error rate of 0.05 when we perform $10,000$ hypothesis tests, we'd need a p-value of $0.05/10000 = 5 \times 10^{-6}$ to declare significance

Objections to Bonferroni Corrections

"*Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference*" Perneger (1998)

- Counter-intuitive: interpretation of finding depends on the number of other tests performed
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. Bmj, 316(7139), 1236-1238.

FWER: Sequential Adjustments

- The simplest sequential method is Holm's Method
  - Order the unadjusted p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$
  - For control of the FWER at level α, the step-down Holm adjusted p-values are

$$\tilde{p}_{(j)} = \min\{(M - j + 1)p_{(j)}, 1\}$$

  - If $\tilde{p}_{(j)} < \alpha$ then proceed to $j + 1$, otherwise stop.
  - Each $\tilde{p}_{(j)}$ is multiplied by a different factor.
- For example, when $M = 10000$:

$$\tilde{p}_{(1)} = 10000p_{(1)}, \tilde{p}_{(2)} = 9999p_{(2)}, \ldots, \tilde{p}_{(M-1)} = 2p_{(M-1)}, \tilde{p}_{(M)} = p_{(M)}$$

Objections to FWER Corrections

- FWER is appropriate when you want to guard against ANY false positives
- However, in many cases (particularly in genomics/neuroimaging) we can live with a certain number of false positives
- In these cases, the more relevant quantity to control is the false discovery rate (FDR)

Benjamini and Hochberg FDR

- To control FDR at level α:
  - Order the unadjusted p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$
  - For a given α, find the largest k such that $p_{(k)} \leq \frac{k}{M}\alpha$. That is, let

$$k = \underset{i \in (1,\ldots,M)}{\arg\max} \left\{ i \cdot I\left(p_{(i)} \leq \frac{i\alpha}{M} \right) \right\} = \underset{i \in (1,\ldots,M)}{\arg\max} \left\{ i \cdot I\left(\frac{M}{i} p_{(i)} \leq \alpha \right) \right\}.$$

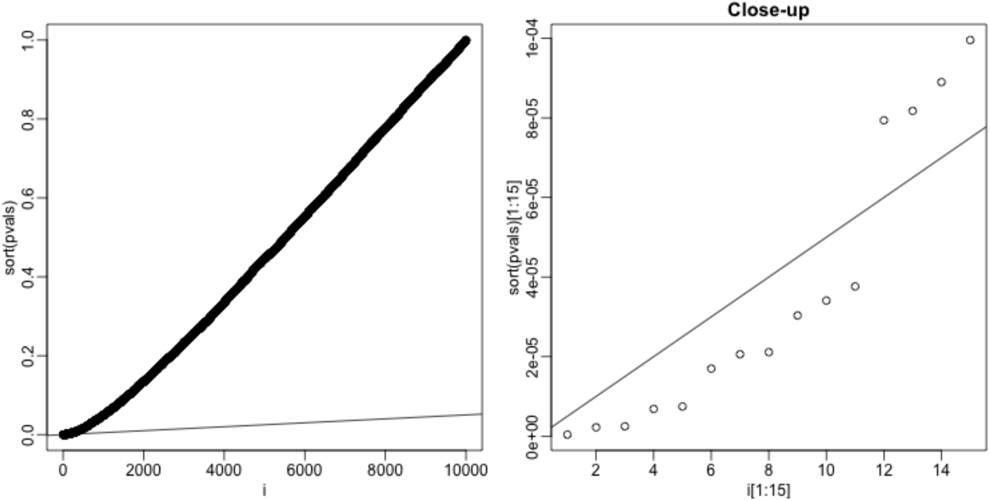
  - Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.
- The BH procedure is valid when the M tests are independent, and also in various scenarios of dependence (but not all) and has

$$E(FDR_{BH}) \leq \alpha M_0/M \leq \alpha$$

BH FDR Example

Rank (j)	P-value	$\alpha(j/M)$	Reject H_0?
1	0.0008	0.005	1
2	0.009	0.01	1
3	0.165	0.015	0
4	0.205	0.02	0
5	0.396	0.025	0
6	0.45	0.03	0
7	0.641	0.035	0
8	0.781	0.04	0
9	0.9	0.045	0
10	0.993	0.05	0

BH FDR Example



Different Approaches To Control Type I Errors

- The BH procedure controls the FDR

$$FDR = E(V/R|R > 0)P(R > 0)$$

- Storey's q-value approach controls the pFDR

$$pFDR = E(V/R|R > 0)$$

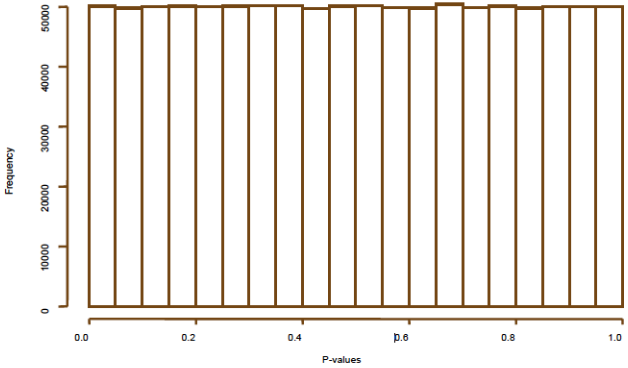
- Since $P(R > 0) \sim 1$ in most experiments FDR and pFDR are very similar
- Omitting $P(R > 0)$ facilitated development of a measure of significance in terms of the FDR for each hypothesis.

What's a q-value?

- q-value is defined as the minimum FDR that can be attained when calling that "feature" significant (i.e., expected proportion of false positives incurred when calling that feature significant).
- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)
- The q-values use an estimate of the proportional of null hypotheses (π_0).

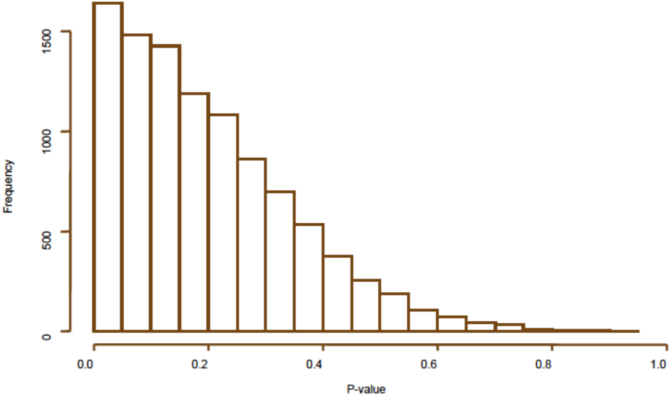
Estimating The Proportion of Truly Null Tests

- Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1



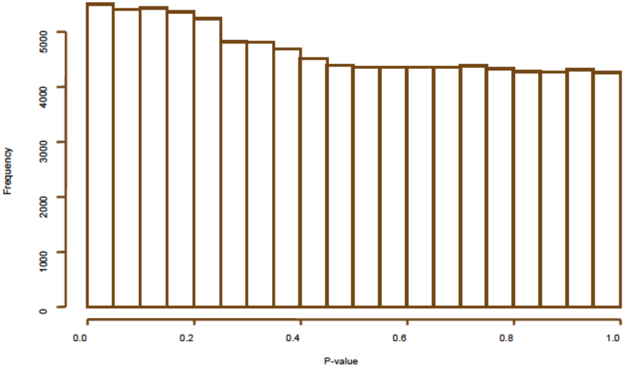
Estimating The Proportion of Truly Null Tests

- Under the alternative hypothesis p-values are skewed towards 0



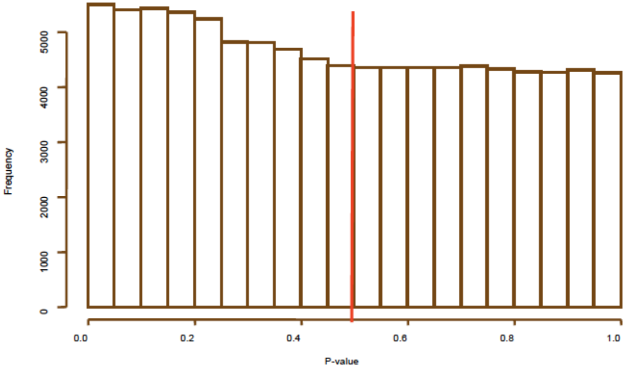
Estimating The Proportion of Truly Null Tests

- Combined distribution is a mixture of p-values from the null and alternative hypotheses



Estimating The Proportion of Truly Null Tests

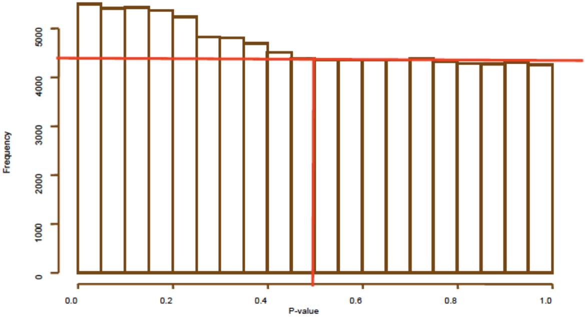
- For p-values greater than say 0.5, we can assume they mostly represent observations from the null hypothesis



Definition of Storey's π_0

- π_0 = Proportion of true null hypotheses, which Storey and Tibshiriani (2003) estimate with

$$\hat{\pi}_0(\lambda) = \frac{\sum_{i=1}^{M} I(p_i > \lambda)}{M(1-\lambda)}$$



Controling the pFDR with q-values

- Estimate $\hat{\pi}_0$.
- Order the unadjusted p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$
- The q-value of $p_{(M)}$ is equal to

$$\hat{q}(p_{(M)}) = \min_{t \geq p_{(M)}} \widehat{FDR}(t) = \min_{t \geq p_{(M)}} \left\{ \frac{\hat{\pi}_0 M t}{\sum_{i=1}^{M} I(p_i \leq t)} \right\} = \hat{\pi}_0 p_{(M)}$$

- For $i = M-1, M-2, \ldots, 2, 1$ calculate

$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \left\{ \frac{\hat{\pi}_0 M t}{\sum_{i=1}^{M} I(p_i \leq t)} \right\} = \min \left(\frac{M}{i} \hat{\pi}_0 p_{(i)}, \hat{q}(p_{(i+1)}) \right)$$

- Reject all hypotheses such that $\hat{q}(p_i) \leq \alpha$

Benjamini and Hochberg FDR

- To control FDR at level α:
  - Order the unadjusted p-values such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$
  - For a given α, find the largest k such that $p_{(k)} \leq \frac{k}{M}\alpha$. That is, let

$$k = \underset{i \in (1,\ldots,M)}{\arg\max} \left\{ i \cdot I\left(\frac{M}{i} p_{(i)} \leq \alpha \right) \right\}.$$

  - Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.
- The BH procedure is valid when the M tests are independent, and also in various scenarios of dependence (but not all) and has

$$E(FDR_{BH}) \leq \alpha \frac{M_0}{M} = \alpha\pi_0 \leq \alpha$$

Similarities between Storey and BH

- For the BH procedure, the BH adjusted q-values can be defined as

$$q_{BH(i)} = \min\left(\frac{Mp_{(i)}}{i}, 1\right),$$

  and we reject $H_{(i)}$ if $\min_{j=i,i+1,\dots,M}\{q_{BH(j)} \leq \alpha\}$.
- Similarly, with the Storey procedure the q-values are

$$\hat{q}(p_{(i)}) = \min\left\{\hat{\pi}_0 \frac{Mp_{(i)}}{i}, \hat{q}(p_{(i+1)})\right\}$$

  and we reject $H_{(i)}$ if $\hat{q}(p_{(i)}) \leq \alpha$ (the minimum isn't necessary since they are non-increasing by definition).

Similarities between Storey and BH

- So, ignoring the constraints

$$q_{BH(i)} = \frac{Mp_{(i)}}{i} \geq \hat{\pi}_0 \frac{Mp_{(i)}}{i} = \hat{q}(p_{(i)})$$

  since $\hat{\pi}_0 \leq 1$, and the Storey procedure will result in at least as many rejections then the BH procedure.
- We do not need to use the Storey estimate of π_0 and there are many other options (e.g., Efron et al., 2001; Sun and Cai, 2007; Storey, 2007).
- Recently, many procedure allow for covariate data to be used to estimate π_0 (Li and Barber, 2017, 2019; Lei and Fithian, 2018, Ignatiadis and Huber, 2021).

Two groups model

The multiple-testing problem can often be simplified as follows:

- We are given hypotheses $\boldsymbol{H} = (H_1, \ldots, H_m)$ and we observe test statistics $\boldsymbol{T} = (T_1, \ldots, T_m)$.
- Let $H_i = 1$ if the ith null hypothesis is false (i.e., predictor i is non-null).
- Further, $H_1, \ldots, H_m \sim^{ind} \text{Bernoulli}(1 - \pi_0)$.
- Then
$$T_i | H_i \sim (1 - H_i) F_0 + H_i F_1$$
- This is the popular "two-groups" model for multiple testing.

Two groups model

- The marginal cdf of T is the mixture distribution

$$F(t) = \pi_0 F_0(t) + (1 - \pi_0) F_1(t),$$

and the pdf is

$$f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t)$$

- What is the probability that test i is null given T_i, π_0, f_0 and f_1?

Local false discovery rate

- The local false discovery rate is defined as

$$lfdr(t) = \Pr(null | T = t) = \frac{\pi_0 f_0(t)}{f(t)}$$

- Then the posterior probability of a case being null given that its test statistics (T) is less than some value t is then

$$\begin{aligned} Fdr(t) = \Pr(null | T \le t) &= \frac{\int_{-\infty}^{t} lfdr(T) f(T) dT}{\int_{-\infty}^{t} f(T) dT} \\ &= \frac{\pi_0 F_0(t)}{F(t)} = E\{lfdr(T) | T \le t\} \end{aligned}$$

Local false discovery rate and the BH procedure

- We said T was a test-statistics. If T was a p-value what is f_0 and F_0?
- What if we use F equal to the empirical distribution function and set $\pi_0 = 1$?

- What does this look like?

Why Multiple Testing Matters

- So, using the local false discovery rate or the two-groups model opens up many possibilities for multiple testing.
- Any procedure that can estimate π_0, f_0 and f can be used.
- This lead people to notice that using p-values can be not the best way to do things.
- Efron et al. (2001) and Sun and Cai (2007) were some of the first to exploit possible asymmetry in the data to increase power.
- Sun and Cai (2007) provided the theory to show that data driven approaches under the two-groups model control the FDR.

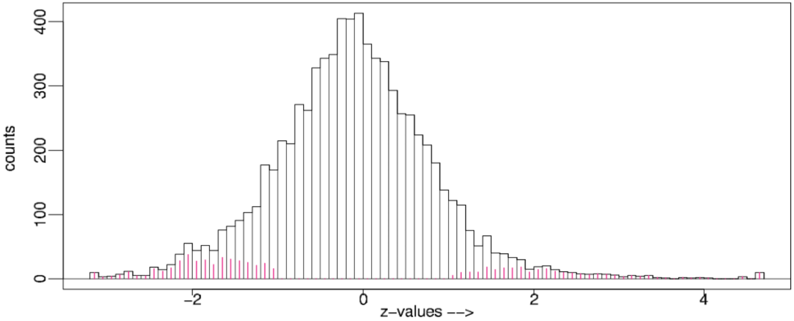
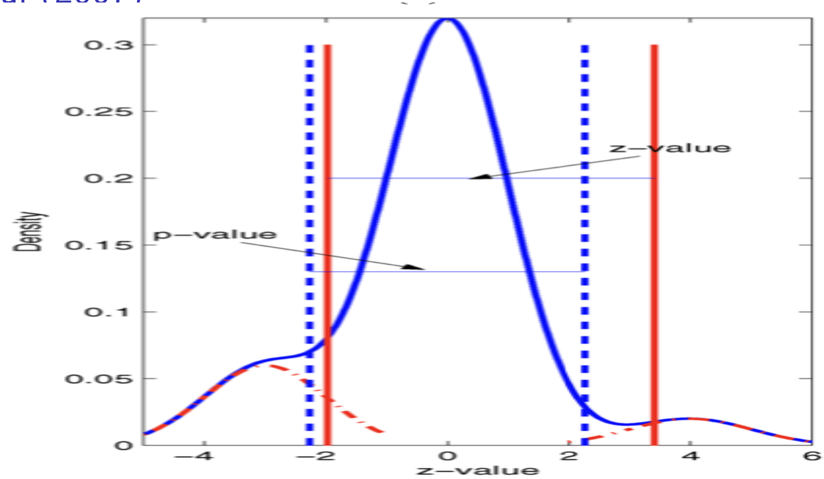


Figure 1: *Histogram of 7680 z-values from an HIV microarray experiment. Short vertical bars are estimated "thinned counts" of non-null genes, as explained in Section 5. (Extreme values have been truncated, giving small bars at each end.) Data from van't Wout et al. (2003), discussed in Gottardo et al. (2004).*

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment", Journal of the American Statistical Association 96, 1151-1160.

Sun and Cai (2007)



Sun, W. and Cai, T. T. (2007). Oracle and adaptive com- pound decision rules for false discovery rate control. J. Amer. Statist. Assoc. 102 901–912.

Estimation methods

- There have been many published methods to estimate π_0.
- Usually people will assume that f_0 is a standard normal or t-distribution (for test statistics).
- Estimating f is actually quite simple and can be done with standard methods of density estimation.

Controling the FDR with two-groups model

Using *Fdr*

- Estimate $\hat{\pi}_0$, $\hat{f}$ (assume f_0 is some distribution).
- For each i calculate
$$Fdr_i = \frac{\hat{\pi}_0 F_0(T_i)}{\hat{F}(T_i)}$$
- Order the unadjusted test statistics such that $Fdr_{(1)} \leq Fdr_{(2)} \leq \cdots \leq Fdr_{(m)}$
- For a given α, find the largest k such that $Fdr_{(k)} \leq \alpha$.
- Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.

Controling the FDR with two-groups model

Using *lfdr*

- Estimate $\hat{\pi}_0$, $\hat{f}$ (assume f_0 is some distribution).
- For each i calculate
$$lfdr_i = \frac{\hat{\pi}_0 f_0(T_i)}{\hat{f}(T_i)}$$
- Order the unadjusted test statistics such that $lfdr_{(1)} \leq lfdr_{(2)} \leq \cdots \leq lfdr_{(m)}$
- For a given α, find the largest k such that
$$\frac{1}{k}\sum_{j=1}^{k} lfdr_{(k)} \leq \alpha.$$
- Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.

Incorporation of side-information

Modern methods have considered incorporating “side-information” about the tests to improve the power of the procedure.

- Suppose, z_j is “side-information” about test j.
- Then, we may consider using

$$lfdr(t|z_j) = \Pr(null|T = t, Z = z_j) = \frac{\pi_0(z_j)f_0(t)}{f(t|z_j)}$$

where the prior probability the test is a null (π_0) and the distribution of the test-statistics (f) are a function of the side-information.

Side-information example

- Consider some neuroimaging data where $X_{ij} = 1$ if voxel j from patient i was damaged ($X_{ij} = 0$ otherwise).
- The X_{ij} are random, and for some j most $X_{ij} = 0$ and $\bar{X}_j \approx 0$.
- The $\bar{X}_j$ are likely related to f.
- We also have $d_j = (d_{j1}, d_{j2}, d_{j3})$ the coordinates of voxel j.
- The coordinates might be related to $\pi_0(z_j)$

Lei, L., & Fithian, W. (2018). *AdaPT: an interactive procedure for multiple testing with side information.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 80(4), 649-679.
Zhang, M. J., Xia, F., & Zou, J. (2019). *Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing.* Nature communications, 10(1), 3433.
Zhang, X., & Chen, J. (2022). *Covariate adaptive false discovery rate control with applications to omics-wide multiple testing.* Journal of the American Statistical Association, 117(537), 411-427.