2020 SISG Module 8: Bayesian Statistics for Genetics
Lecture 9: Bayes and Frequentist Testing

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Review of p-Values and Bayes Factors

The Statistical Set-Up

We review frequentist and Bayesian test procedures.

- We begin with a very simple situation in which we have a single parameter of interest θ.
- Assume the null of interest is

$$H_0 : \theta = 0$$

  with θ, for example, a treatment difference, or a log odds ratio, or a log hazard ratio.
- We assume an analysis yields a statistic T for which large values indicate departures from the null – asymptotically χ^2_1.
- For example, the squared Wald statistic, $T = \widehat{\theta}^2/V$, with V the asymptotic variance of the MLE[1].
- An alternative is the likelihood ratio statistic.

[1] $T = Z^2$ where Z is the Z-score

Types of Testing

- The observed p-value is,

$$p = \Pr(T > t_{\text{obs}} | H_0)$$

  where t_{obs} is a number that is evaluated for the data at hand.
- The p-value is not saying anything about the probability of the null being true!!
- To report p only, gives a pure significance test.
- A small p-value can arise because:
  - H_0 is true but we were "unlucky".
  - H_0 is not true.

  – to decide which explanation is responsible depends crucially on the prior belief on whether H_0 is true or not.

Key question: How small is small?

Types of Testing

- A test of significance sets a cut-off value (e.g. $\alpha = 0.05$) and rejects H_0 if $p < \alpha$.

  Again: How to pick α?

- A type I error is to reject H_0 when it is true, and a test of significance controls the type I error (whereas a pure significance test does not).
- A type II error occurs when H_1 is true but H_0 is not rejected.
- A hypothesis test goes one step further and specifies an alternative hypothesis.
- A decision is then taken as to which of H_0 and H_1 is chosen.
- The celebrated Neyman-Pearson lemma shows that for fixed α-level the likelihood ratio statistic maximizes the power.
- Wouldn't it be more reasonable to balance type I and type II errors?

The Dangers of Fixed Significance Levels

- Example: Sample, $Y_1, \ldots, Y_n$ of size n from N$(\theta, 1)$,

$$H_0 : \theta = 0, \quad H_1 : \theta = 1.$$

  Obvious that we should reject H_0 for $\overline{Y}_n > k(n)$, a constant[2].

- The table below illustrates the problems of choosing a fixed α, regardless of sample size — imbalance in α and β as a function of n:

n	α	β	$k(n)$
1	0.01	0.91	2.33
25	0.01	0.0038	0.46
100	0.01	8×10^{-15}	0.23

- Also: Statistical versus practical significance.
- For both p-values and α levels we need thresholds that decrease as a function of the sample size n. Pearson (1953, p. 68), "...the quite legitimate device of reducing α as n increases".

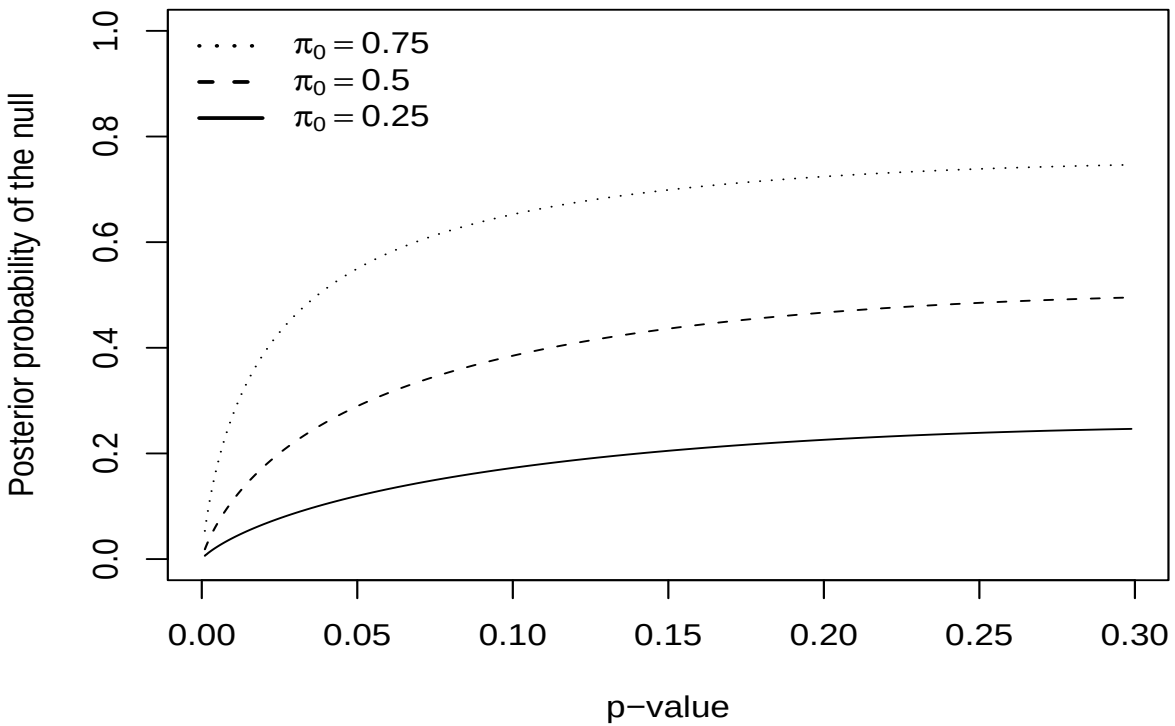
[2]Note that the threshold for $T = [\overline{Y}_n/(1/\sqrt{n})]^2$ is constant

A quite remarkable result!

- With $\pi_0 = \Pr(H_0)$, Sellke *et al.* (2001) show that:

$$\Pr(H_0 | \text{ data }) \geq \left\{ 1 - \frac{1}{2.72\, p \log p} \times \frac{1 - \pi_0}{\pi_0} \right\}^{-1} \qquad (1)$$

- A small p-value doesn't translate to a small probability that the null is not true.



Why does anyone use *p*-values?

- Historically, it was usual to carry out well-powered (single) experiments, and the prior on the alternative was not small.
- With respect to (1) and with $\pi_0 = 0.5$:
  - p-value $= 0.05$ gives $\Pr(H_0 | \text{ data }) > 0.29$.
  - p-value $= 0.01$ gives $\Pr(H_0 | \text{ data }) > 0.11$.
- Scientists well-calibrated in their own discipline?
- Perhaps, but if you're going to be subjective, why not be formal about it?
- Aside: Reason for lack of replication in observational epidemiology? Along with confounding, data dredging, measurement error,...

Calibrating α-Levels

- We want $\Pr(H_0|\text{ data })$, where "data" corresponds to the event $T > t_{\text{fix}}$, but to obtain this we must specify alternatives – consider a simple alternative, say $H_1 : \theta = \theta_1$.
- Then,

$$\begin{aligned}
\text{Posterior Odds of } H_0 &= \frac{\Pr(H_0|\text{ data })}{\Pr(H_1|\text{ data })} \\
&= \frac{\Pr(T > t_{\text{fix}}|H_0)}{\Pr(T > t_{\text{fix}}|H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)} \\
&= \frac{\alpha}{1-\beta} \times \text{ Prior Odds of } H_0
\end{aligned}$$

- For ranking associations (which does not involve the prior odds if constant across tests): must consider the power, $\Pr(\text{ data }|H_1)$.
- For calibration: must consider the prior odds of H_0.

A Sanity Check via a Simple Example

- The model:

$$Y_i|\theta \sim_{iid} \mathrm{N}(\theta, \sigma^2), \quad \sigma^2 \text{ known},$$

  $i = 1, \ldots, n$.

- The distribution of the MLE is:

$$\widehat{\theta} = \overline{Y} \sim \mathrm{N}(\theta, V)$$

  with $V = \sigma^2/n$,

$$T = \frac{n\overline{Y}^2}{\sigma^2}.$$

- Null and alternative hypotheses are

$$H_0 : \theta = 0, \quad H_1 : \theta \neq 0.$$

A Sanity Check via a Simple Example

- Under H_1 assume the prior $\theta \sim \mathrm{N}(0, W)$.
- Recall from previous lectures that the evidence in the data for a pair of hypotheses is summarized in the Bayes factor:

$$\mathrm{BF} = \frac{p(\boldsymbol{y}|H_0)}{p(\boldsymbol{y}|H_1)} = \frac{\prod_{i=1}^{n} \mathrm{N}(y_i|0, \sigma^2)}{\int_\theta \prod_{i=1}^{n} \mathrm{N}(y_i|\theta, \sigma^2) \times \mathrm{N}(\theta|0, W) d\theta}.$$

Testing: decision theory

A reminder of the ingredients for decision theory;

- Loss function $L(\theta, d)$: how bad it would be if the truth were θ but you took decision d. (Optimists: note we could equivalently define utility as $-L(\theta, d)$ — how good it would be – economists do this)
- Expected posterior loss $\mathrm{E}[\, L(\theta, d)\,]$ – loss for some decision d averaged over posterior uncertainty

The Bayes rule is the decision d that minimizes $\mathrm{E}[\, L(\theta, d)\,]$ – but for testing, d is 0 or 1, so this means checking whether

$$\mathrm{E}[\, L(\theta, d = 0)\,] \leq \mathrm{E}[\, L(\theta, d = 1)\,],$$

i.e., do we expect less loss deciding $d = 0$ or $d = 1$?

Testing

		Truth	
		$\theta = 0$	$\theta \neq 0$
Decision	$d = 0$	0	L_1
	$d = 1$	L_2	0

With respect to this table, the posterior expected cost associated with the decision d is

$$\mathrm{E}[L(\theta, d)] = L(\theta = 0, d)\,\Pr(\theta = 0|\boldsymbol{y}) + L(\theta \neq 0, d)\,\Pr(\theta \neq 0|\boldsymbol{y}).$$

The two possible decisions (report $\theta = 0$ or $\theta \neq 0$) the expected losses are:

$$\begin{aligned} \mathrm{E}[L(\theta, d = 0)] &= 0 \times \Pr(\theta = 0|\boldsymbol{y}) + L_2\,\Pr(\theta \neq 0|\boldsymbol{y}) \\ \mathrm{E}[L(\theta, d = 1)] &= L_1\,\Pr(\theta = 0|\boldsymbol{y}) + 0 \times \Pr(\theta \neq 0|\boldsymbol{y}) \end{aligned}$$

Testing

We now have to find the decision that minimizes the posterior expected loss, as a function of $\Pr(\theta \neq 0|\boldsymbol{y}) = \Pr(\theta|\boldsymbol{y})$.

A little rearrangement leads to reporting $\theta \neq 0$ if

$$\Pr(\theta \neq 0|\boldsymbol{y}) \geq \frac{L_1}{L_1 + L_2} = \frac{1}{1 + L_2/L_1} = \frac{1}{1 + R},$$

or equivalently

$$\Pr(\theta = 0|\boldsymbol{y}) < \frac{1}{1 + R}.$$

Examples:

If $L_1 = L_2$ ($R = 1$), report $\theta \neq 0$ if $\Pr(\theta \neq 0|\boldsymbol{y}) \geq \frac{1}{2}$.

If $L_1 = 3 \times L_2$ ($R = 1/3$), report $\theta \neq 0$ if $\Pr(\theta \neq 0|\boldsymbol{y}) \geq \frac{3}{4}$.

If $L_2 = 3 \times L_1$ ($R = 3$), report $\theta \neq 0$ if $\Pr(\theta \neq 0|\boldsymbol{y}) \geq \frac{1}{4}$.

A Sanity Check via a Simple Example

- We take $W = \sigma^2$, which corresponds to the "unit information prior" of Kass and Wasserman (1995) (this choice not so important).
- With a prior odds, PO, and ratio of costs of type II to type I errors, R, this gives the decision rule to reject H_0:

$$\begin{aligned} \text{Posterior Odds} &= \text{BF} \times \text{PO} \\ &= \sqrt{1+n} \times \exp\left(-\frac{T}{2}\frac{n}{1+n}\right) \times \text{PO} < R \end{aligned}$$

- Notice how this depends on T and n.

A Bayesian Test Statistics Threshold

- Rearrangement gives a threshold for rejection of:

$$T > \frac{2(1+n)}{n} \log\left(\frac{\text{PO}}{\text{R}}\sqrt{1+n}\right)$$

- For relatively large prior odds on the null PO: require T to be larger (more evidence).
- For relatively large cost of Type II errors R (so that we are averse to type II error, i.e. missing signals): require T to be smaller (less evidence).
- Not such a simply summarization for n but, beyond a certain point, as n gets larger, we require larger T (more evidence).
- The above should be contrasted with the usual frequentist approach of

$$T > \text{const}$$

with the constant usually chosen to control the type I error.

A Bayesian Test Statistic Threshold

- The table below evaluates the probability of rejection given H_0. We assume $R = 1$.
- For $\pi_0 = 0.5$ and $n = 20, 50, 100$ the thresholds give ≈ 0.05 — the situation in which this infamous threshold was first derived?

	$\pi_0 = 0.25$	$\pi_0 = 0.50$	$\pi_0 = 0.95$
$n = 10$	0.64	0.10	0.0025
$n = 20$	0.35	0.074	0.0022
$n = 50$	0.18	0.045	0.0016
$n = 100$	0.12	0.031	0.0011
$n = 1000$	0.030	0.0085	0.00034

Calibration with *p*-values

- The ABF can be inverted to give a rule for Z^2 that depends on PO, R and n (as with the simple example presented previously).
- For more details, see Wakefield (2009).
- Figure 1 shows the behavior of this rule as a function of the sample size n, and for different choices of the prior on the alternative π_1 and the ratio of costs of type II to type I errors.
- The curves have the expected ordering and, as n gets large, a greater and greater level of evidence is required.

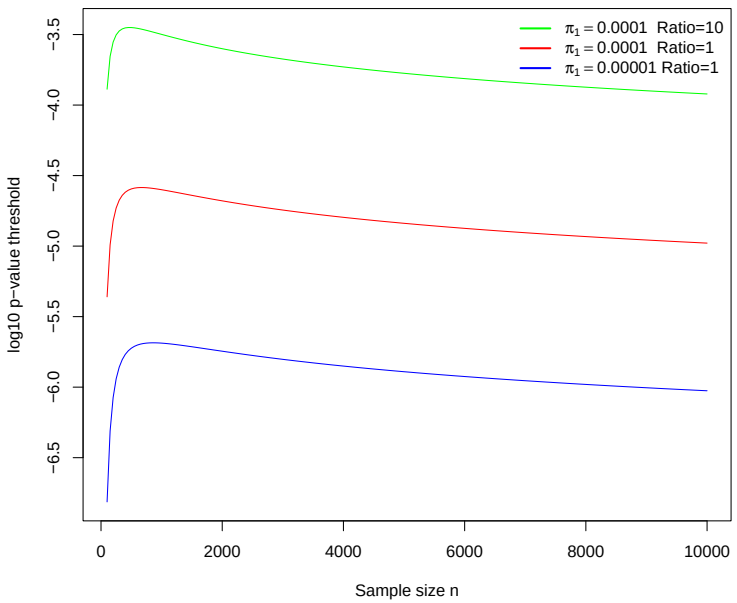


Figure 1: Regression threshold, on $\log_{10}(p)$-value scale, vs sample size.

This is as we would expect because as the sample size increases we want both Type I and Type II errors to go to zero.

Summary for Single Tests

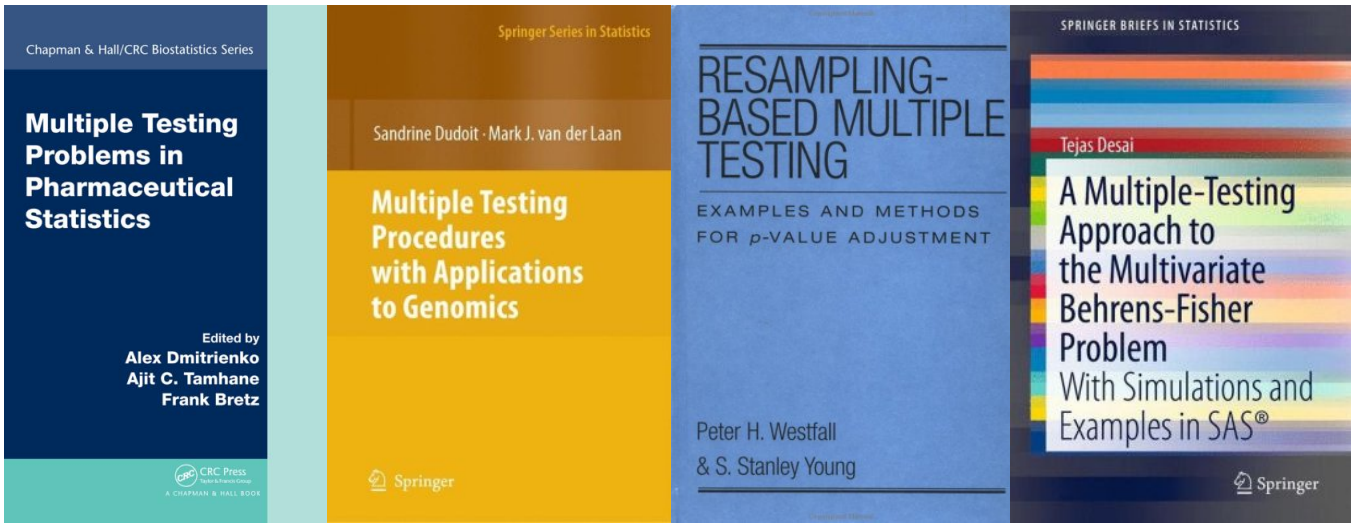
- p-values are widely misinterpreted but are not going away and so it's important to interpret correctly.
- p-values are hard to calibrate without knowing the sample size/power.
- Thresholds for significance should increase (i.e., p-values should be smaller) as n increases.
- Bayes factors provide an alternative (and they produce thresholds with desirable properties), but they are not without their issues (prior specification, calibration,...).
- To get at $\Pr(H_0|$ data $)$ you can't get away from specifying $\pi_0 = \Pr(H_0)$, and the posterior probability is horribly sensitive to the value chosen.
- Better to use estimation procedures if you can.

Multiple Testing

Motivation for Multiple Testing

We have covered testing procedures, both frequentist and Bayesian, in the context of single tests.

How to proceed, when multiple tests are envisaged, is a big topic:



A lot of interest lately, given the advent of technologies that allow huge numbers of experiments to be performed.

As with single tests, this topic is controversial.

Motivating Example

- We follow a running example with data from a microarray study of 102 men, 52 with prostate cancer and 50 normal controls (Efron and Hastie, 2016).
- Gene expression levels were measured for $m = 6033$ genes.
- A two-standard t-test was carried out.

Motivating Example

- A transformation was made so that the resultant statistic z_i, has distribution under the null:

$$H_{0i} : z_i \sim \mathrm{N}(0, 1),$$

for $i = 1, \ldots, m$ genes.

- Under the alternative:

$$H_{1i} : z_i \sim \mathrm{N}(\mu_i, 1),$$

for $i = 1, \ldots, m$ genes.

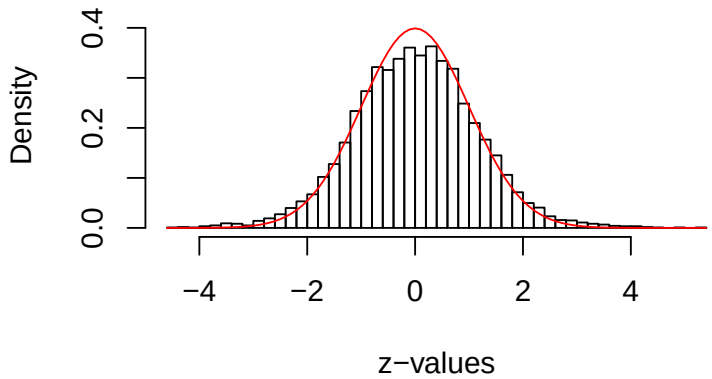


Figure 2: Histogram of z-values for prostate microarray study, with $\mathrm{N}(0, 1)$ distribution in red.

- The aim is to find genes with non-zero μ_i.

Framework for Multiple Testing

Possibilities with m tests and when K are flagged as requiring further attention:

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m - K$	K	m

- m_0 is the number of true nulls.
- B is the number of type I errors.
- C is the number of type II errors.

Problem: To select a rule that will determine K.

We discriminate between:

- A sensible criterion.
- How the criterion should depend on sample size.

The Family-Wise Error Rate

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m - K$	K	m

- The family-wise error rate (FWER) is the probability of making at least one Type I error, i.e.

$$\Pr(B \geq 1 | \text{ all } H_0 \text{ true }).$$

- Let B_i be the event that the i-th null is incorrectly rejected, so that $B = \cup_{i=1}^{m} B_i$ is the total number of incorrectly rejected nulls.

The Family-Wise Error Rate

- The FWER is given by:

$$\begin{aligned} \text{FWER} = \Pr(B \geq 1 | \text{ all } H_0 \text{ true }) &= \Pr\left(\cup_{i=1}^{m} B_i | \text{ all } H_0 \text{ true }\right) \\ &\leq \sum_{i=1}^{m} \Pr(B_i | \text{ all } H_0 \text{ true }) \\ &= m\alpha^\star \end{aligned}$$

  where $\alpha^\star$ is the level for each test.
- This is true regardless of whether the tests are independent or not.
- Bonferroni takes $\alpha^\star = \alpha/m$ to give FWER $\leq \alpha$.
- Example: For control at $\alpha = 0.05$ with $m = 6033$ tests take $\alpha^\star = 0.05/6033 = 8.3 \times 10^{-6}$.
- Such stringent rules lead to a loss of power, but not ridiculous if you think there is a reasonable chance that all nulls could be true (but α should depend on n, in particular should decrease as n gets larger and larger).

Sidak Correction

- If tests are independent:

$$\begin{aligned}
\Pr(B \geq 1) &= 1 - \Pr(B = 0) \\
&= 1 - \Pr(\cap_{i=1}^{m} B_i') \\
&= 1 - \prod_{i=1}^{m} \Pr(B_i') \\
&= 1 - (1 - \alpha^\star)^m = \text{ FWER}
\end{aligned}$$

- So to achieve FWER $= \alpha$ take p-value threshold as $\alpha^\star = 1 - (1 - \alpha)^{1/m}$ — the Sidak correction (Sidák, 1967).
- Example: with $m = 500K$ tests take

$$\alpha^\star = 1 - (1 - 0.05)^{1/500{,}000} = 1.03 \times 10^{-7}.$$

Holm's Procedure

Holm's procedure Holm (1979) offers a modest improvement over Bonferroni.

Let

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(i)} \leq \cdots \leq p_{(m)},$$

with corresponding null hypotheses $H_{0(i)}$.

Then, proceed as follows:

1. Let i_0 be the smallest index i such that

$$p_{(i)} > \frac{\alpha}{m - i + 1}.$$

2. Reject all null hypotheses $N_{0(i)}$ for $i < i_0$ and accept all with $i \geq i_0$.

It can shown that Holm's procedure controls FWER at level α and is slightly less conservative.

Expected Number of False Discoveries

We describe an alternative criterion.

For $i = 1, \ldots, m$ tests let B_i again be the 1/0 random variable representing whether the null was incorrectly rejected or not, so that $B = \cup_{i=1}^{m} B_i$.

The expected number of false discoveries (EFD), with significance level α for each test, is given by

$$\text{EFD} = \text{E}[B] = \sum_{i=1}^{m} \text{E}[B_i] = m\alpha$$

if all nulls are true.

Expected False Discoveries

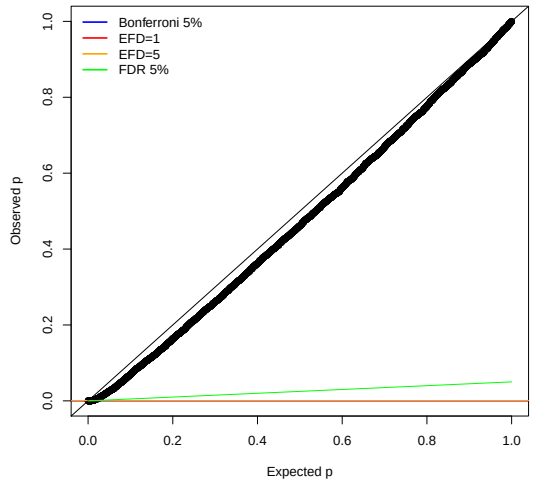
For m_0 true nulls: $\mathrm{E}[B] = m_0\alpha$, but m_0 is unknown, so all we can say is

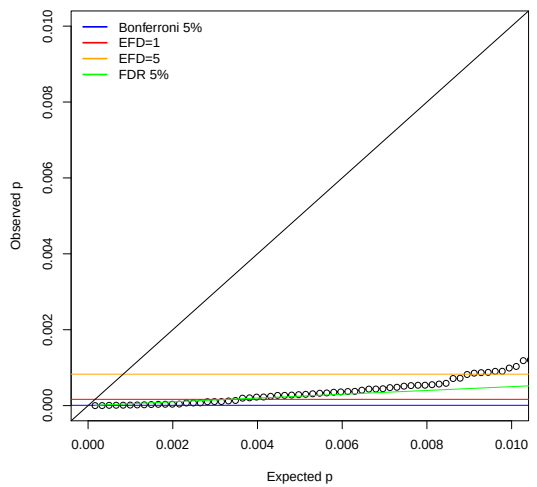
$$\mathrm{EFD} = \mathrm{E}[B] \leq m\alpha.$$

- In a GWAS context suppose $m = 500K$ and $\alpha = 0.05$; this gives $\mathrm{EFD} \leq 25,000$, so conventional levels will clearly not work!
- We can easily put an upper bound on the EFD.
- For example, if we set $\alpha = 1/m$ the expected number of false discoveries is bounded by 1.
- With $\alpha = 5/m$ the expected number of false discoveries is bounded by 5.
- Compare to Bonferroni which controls the FWER via α/m.

Prostate Cancer Example

- We begin by plotting,the observed p-values versus those expected under the null, i.e. $i/(m+1)$ for $i = 1, \ldots, m = 6033$.
- Hard to tell what is going on here... even when we focus in on bottom left, which is the small p-value area of region of interest.





Prostate Cancer Example

- We stretch the scale in Figure 5 by taking $-\log_{10}$, the area where the action is, is now top right.
- On this scale, a value of 2 corresponds to a p-value of 0.01, and a value of 3 corresponds to a p-value of 0.001.
- Bonferroni, $p = 0.05/m = 8.3 \times 10^{-6}$, or $-\log_{10}(p) = 5.1$, flags only 3 genes as worthy of attention – this is the consequence of such a conservative criteria of following a procedure in which the probability of making any type I errors is 0.05.
- Holm's procedure gives the same 3.

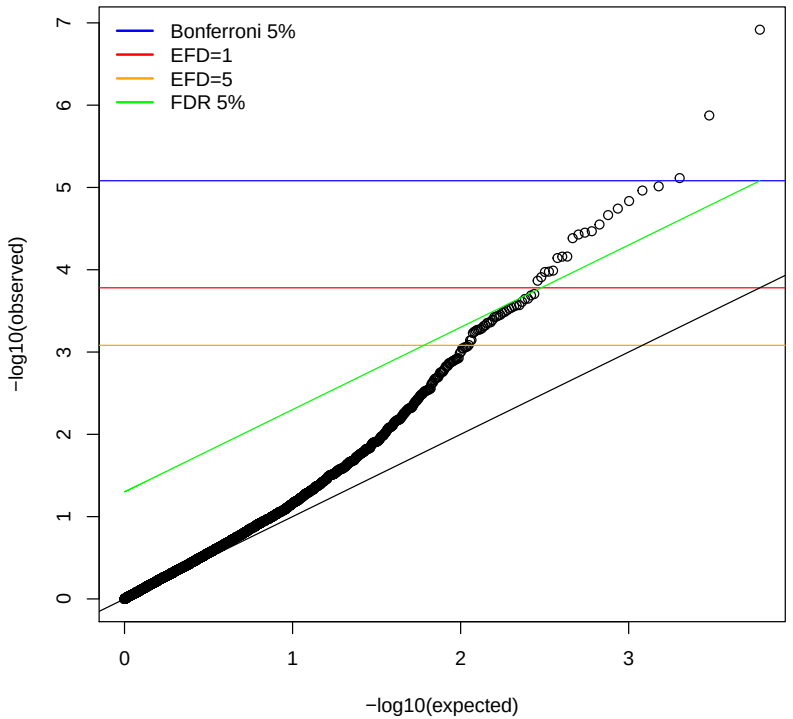


Figure 3: Observed versus expected p-values, on $-\log_{10}$ scale.

Prostate Cancer Example

- The EFD=1 gives a p-value threshold of $1/6033 = 0.00017$, or $-\log_{10}(p) = 3.78$ and gives 21 flagged genes – with an expected false discovery of 1.
- The EFD=5 gives a p-value threshold of $5/6033 = 0.00083$, or $-\log_{10}(p) = 3.08$ and gives 54 flagged genes.
- As always with frequentist procedures there is no way of knowing anything about the 21 (or 54) specifically, the EFD=1 or 5 is the average over repeated uses of this procedure.

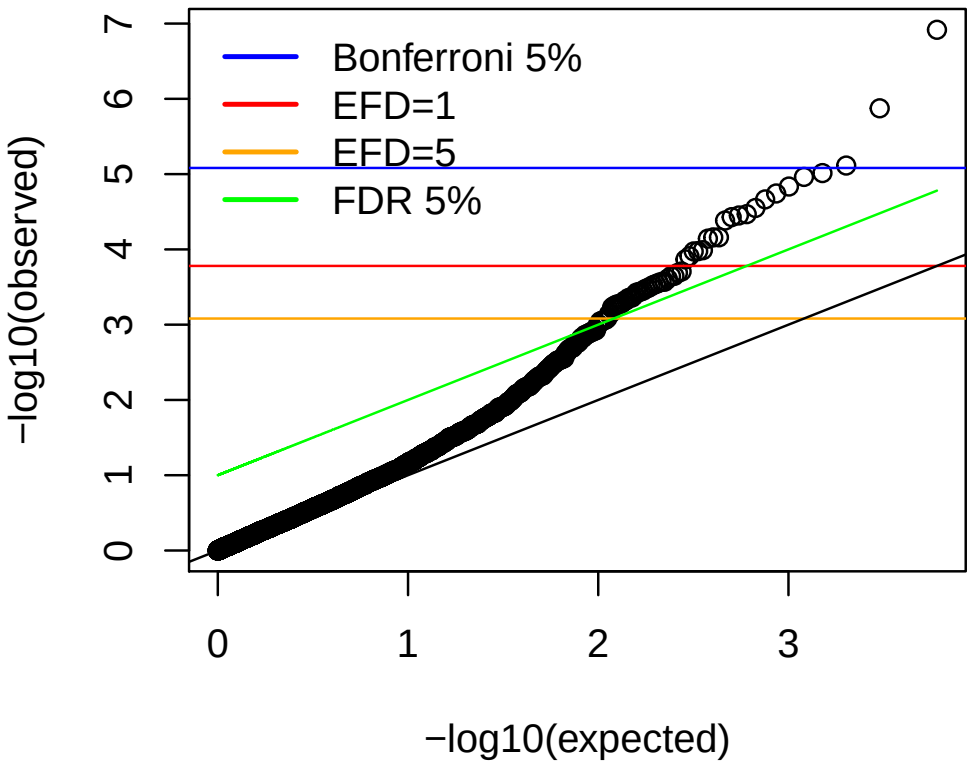


Figure 4: Observed versus expected p-values, on $-\log_{10}$ scale.

False Discovery Rate

A very popular criterion is the false discovery rate (FDR).

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m - K$	K	m

Define the false discovery proportion (FDP) as the proportion of incorrect rejections:

$$\text{FDP} = \begin{cases} \frac{B}{K} & \text{if } K > 0 \\ 0 & \text{if } K = 0 \end{cases}$$

Then the false discovery rate (FDR), the expected proportion of rejected nulls that are actually true nulls, is given by

$$\text{FDR} = \text{E}[\text{FDP}].$$

This is the usual frequentist thing – under hypothetical replication of the experiment and application of the procedure the proportion of flagged features which are actually null.

False Discovery Rate

We describe an algorithm for controlling the FDR.

Consider the following procedure for independent p-values:

1. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.
2. Define $l_i = i\alpha/m$ and $R = \max\{i : P_{(i)} < l_i\}$ where α is the value for which we would like FDR control.
3. Then define the p-value threshold as $P_T = P_{(R)}$.
4. Reject all H_{0i} for which $P_i \leq P_T$.

Benjamini and Hochberg (1995) show that if this procedure is applied, then regardless of how many nulls are true (m_0) and regardless of the distribution of the p-values when the null is false

$$\text{FDR} \leq \frac{m_0}{m}\alpha < \alpha.$$

This is incredible!

False Discovery Rate

If all the signals are null, then $B = K$ (all rejections are false) and

$$\text{FDR} = \text{E}\left[\frac{B}{K}\right] = 1 \times \Pr(B \geq 1) = \text{FWER}.$$

FDR in this form and with extensions, e.g. Storey and Tibshirani (2003) has been successfully used in the microarrays field, where the number of non-null associations is not small.

Unfortunately less successful in a GWAS, because the proportion of nulls is very close to 1.

Prostate Cancer Example

- With a 5% FDR, 21 signals are flagged (not shown on figure).
- With a 10% FDR, 59 signals are flagged.
- Again, we cannot say anything about specific signals but under repeated use of this procedure we are using 10% of the signals we flag as significant, will actually be null.
- We definitely can't say that for any of the signals we have flagged there is a 10% chance that the null is true.
- With a 20% FDR, 106 signals are flagged (not shown on figure).

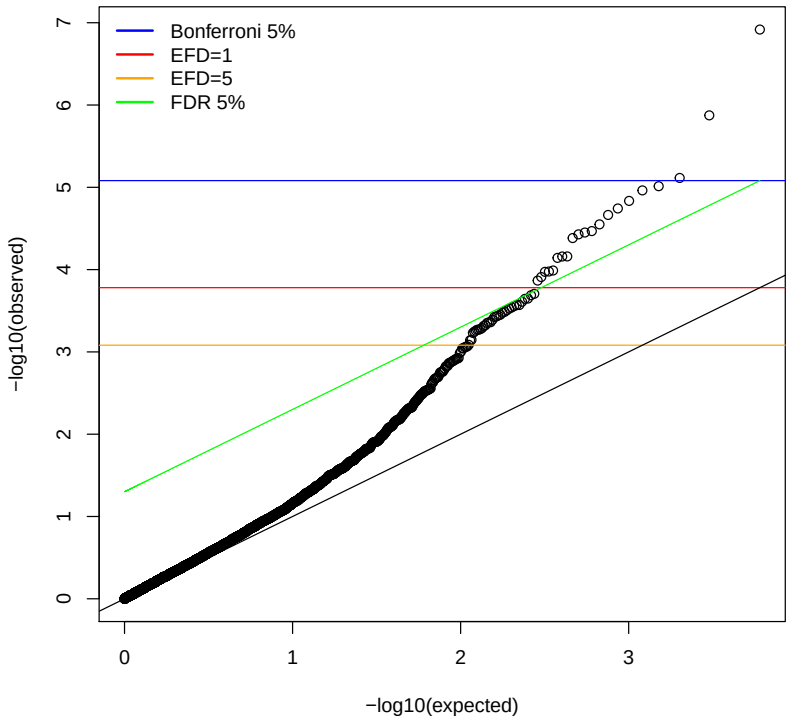


Figure 5: Observed versus expected p-values, on $-\log_{10}$ scale.

q-values

The algorithm of Benjamini and Hochberg (1995) begins with a desired FDR and then provides the p-value threshold.

Storey (2002) proposed an alternative method by which, for any fixed rejection region, a criteria closely related to FDR, the positive false discovery rate

$$\text{pFDR} = \text{E}[B/K \mid K > 0],$$

may be estimated[3].

We assume rejection regions of the form $T > t_{\text{fix}}$ and consider the pFDR associated with regions of this form, which we write as $\text{pFDR}(t_{\text{fix}})$.

[3] this handles the event $K = 0$ differently to the previously-defined FDR

q-values

We define, for $i = 1, \ldots, m$ tests, the random variables $H_i = 0/1$ corresponding to null/alternative hypotheses and test statistics T_i.

Then, with $\pi_0 = \Pr(H = 0)$ and $\pi_1 = 1 - \pi_0$ independently for all tests:

$$\text{pFDR}(t_{\text{fix}}) = \frac{\Pr(T > t_{\text{fix}} \mid H = 0) \times \pi_0}{\Pr(T > t_{\text{fix}} \mid H = 0) \times \pi_0 + \Pr(T > t_{\text{fix}} \mid H = 1) \times \pi_1}.$$

Consideration of the false discovery odds:

$$\frac{\text{pFDR}(t_{\text{fix}})}{1 - \text{pFDR}(t_{\text{fix}})} = \frac{\Pr(T > t_{\text{fix}} \mid H = 0)}{\Pr(T > t_{\text{fix}} \mid H = 1)} \times \frac{\pi_0}{\pi_1}$$

explicitly shows the weighted trade-off of type I and type II errors, with weights determined by the prior on the null/alternative.

q-values

Storey (2003) rigorously shows that

$$\text{pFDR}(t_{\text{fix}}) = \Pr(H = 0 \mid T > t_{\text{fix}}).$$

giving a Bayesian interpretation.

In terms of p-values, the rejection region corresponding to $T > t_{\text{fix}}$ is of the form $[0, \gamma]$.

Let P be the random p-value resulting from a test.

Under the null, $P \sim \text{U}(0, 1)$, and so

$$\begin{aligned} \text{pFDR}(\gamma) &= \frac{\Pr(P \leq \gamma \mid H = 0) \times \pi_0}{\Pr(P \leq \gamma)} \\ &= \frac{\gamma \times \pi_0}{\Pr(P \leq \gamma)}. \end{aligned} \tag{2}$$

From this expression, the crucial role of π_0 is evident.

q-values

- Storey (2002) estimates (2), using uniformity of p-values under the null, to produce the estimates

$$\widehat{\pi}_0 = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)} \quad (3)$$

$$\widehat{\Pr}(P \leq \gamma) = \frac{\#\{p_i \leq \gamma\}}{m} \quad (4)$$

  with λ chosen via the bootstrap to minimize the mean-squared error for prediction of the pFDR.
- The expression (3) calculates the empirical proportion of p-values to the right of λ, and then inflates this to account for the proportion of null p-values in $[0, \lambda]$.

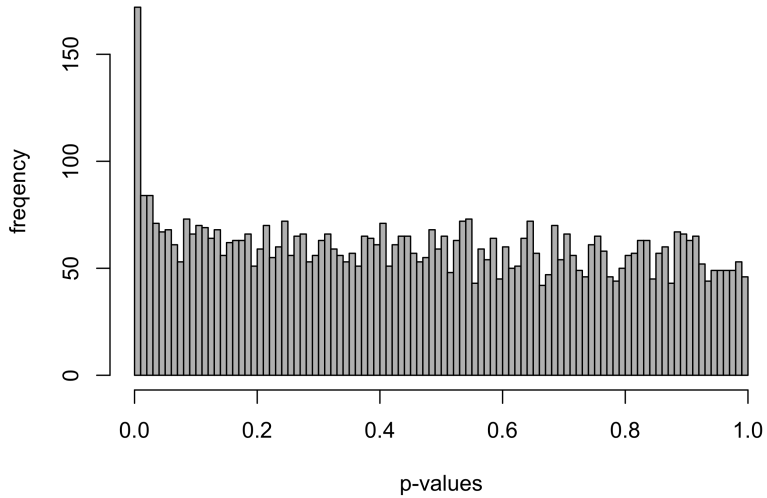


Figure 6: Histogram of p-values for prostate cancer example.

- π_0 is estimated as 0.854 for the prostate cancer data.
- 71 genes flagged at 10% FDR level.

q-values

This method highlights the benefits of allowing the totality of p-values to estimate fundamental quantities of interest such as π_0.

The q-value is the minimum FDR that can be attained when a particular test is called significant.

We give a derivation of the q-value and, following Storey (2002).

To make the argument simpler, suppose we have a test statistic T that is χ^2_1 under the null.

Then define a set of nested rejection regions $\{\Gamma\}$ where these sets could be of the form

$$\Gamma = [t, \infty)$$

where $-\infty \leq t \leq \infty$.

q-values

Then,

$$p\text{-value}(t) = \inf_{\{\Gamma : t \in \Gamma\}} \Pr(T \in \Gamma \mid H = 0)$$

is the p-value corresponding to an observed statistic t.

For example, p-values of 0.05 and 0.10 correspond to $\Gamma = [3.84, \infty)$ and $\Gamma = [2.71, \infty)$, respectively.

The q-value is defined as

$$q\text{-value}(t) = \inf_{\{\Gamma : t \in \Gamma\}} \Pr(H = 0 \mid T \in \Gamma) \tag{5}$$

Therefore, for each observed statistic t_i there is an associated q-value.

The q-value is the minimum pFDR that can be attained when calling that feature significant.

For example, if a particular feature has a q-value of 0.17, then if we call this feature significant, the expected proportion of false positives incurred is 17%.

q-values

- Recall,

$$\text{pFDR}(\gamma) \quad = \quad \frac{\gamma \times \pi_0}{\Pr(P \leq \gamma)}.$$

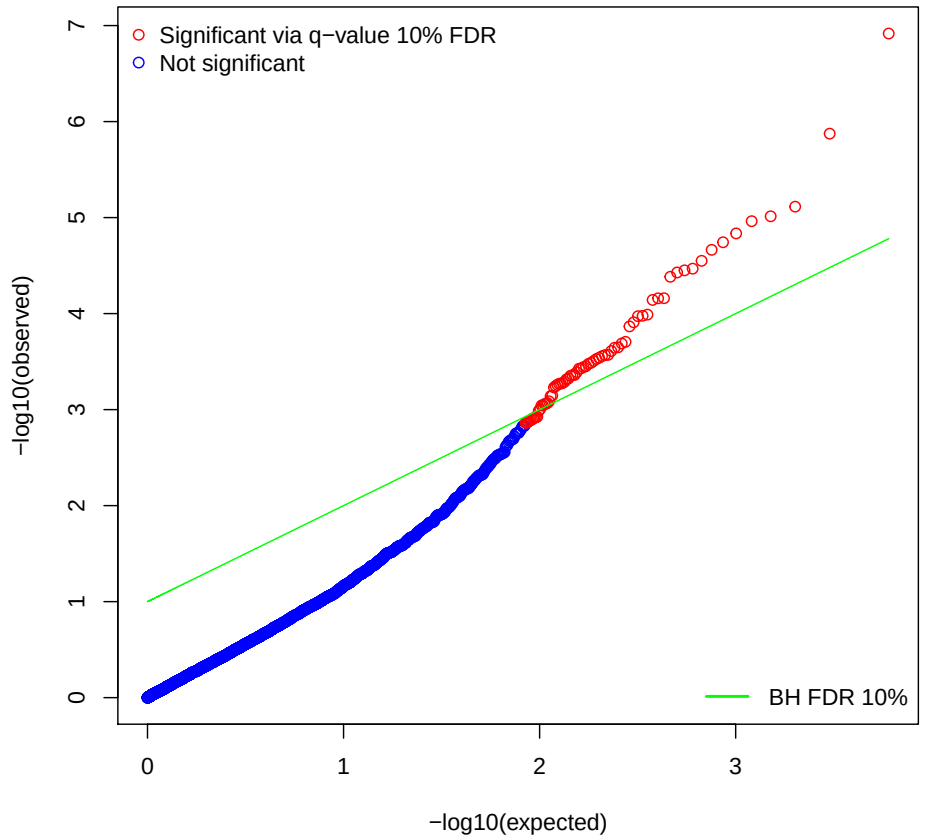
- As we have noted, a common mistake is to say that the p-value is the probability a feature is a false positive, i.e., to equate with $\Pr(H_0|$ data $)$.
- We stress that the q-value is also not the probability that the feature is a false positive.
- The q-values can be estimated from the p-values via,

$$\widehat{q}(p) = \inf_{\gamma \geq p} \text{pFDR}(\gamma).$$

- The mathematical definition of the q-value is the minimum FDR that can be attained when calling that feature significant.

q-values

- The order of the *q*-values is the same as the order of the *p*-values (as with Bonferroni and EFD).
- 71 genes are flagged with a 10% FDR using the *q*-value approach, recall that the Benjamini-Hochberg algorithm gave 59 genes at this level.



q-values and "Local" FDR

It can be shown that,

$$\Pr(H_0 \mid T > t_{\text{obs}}) < \Pr(H_0 \mid T = t_{\text{obs}}) \tag{6}$$

- So the evidence for H_0 given the exact ordinate is always greater than that corresponding to the tail area.
- This fits in with the Sellke *et al.* (2001) result we saw earlier, see Figure 7.

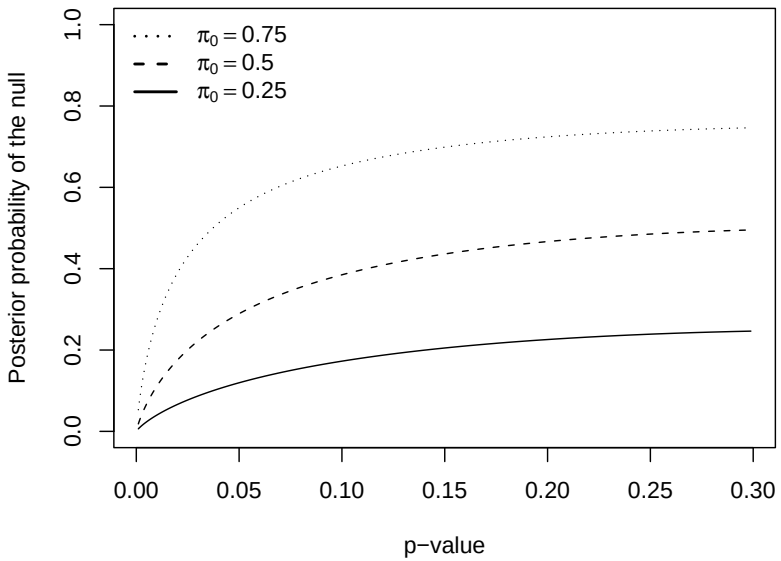


Figure 7: Sellke *et al.* (2001) relationship between posterior probability of the null and the p-value.

q-values

When one decides upon a value of FDR (or pFDR) to use in practice, the sample size should again be taken into account, since for large sample size one would not want to tolerate as large an FDR as with a small sample size.

Again, we would prefer a procedure that was consistent.

However, as in the single test situation, there is no prescription for deciding how FDR should decrease with increasing sample size.

Prostate cancer

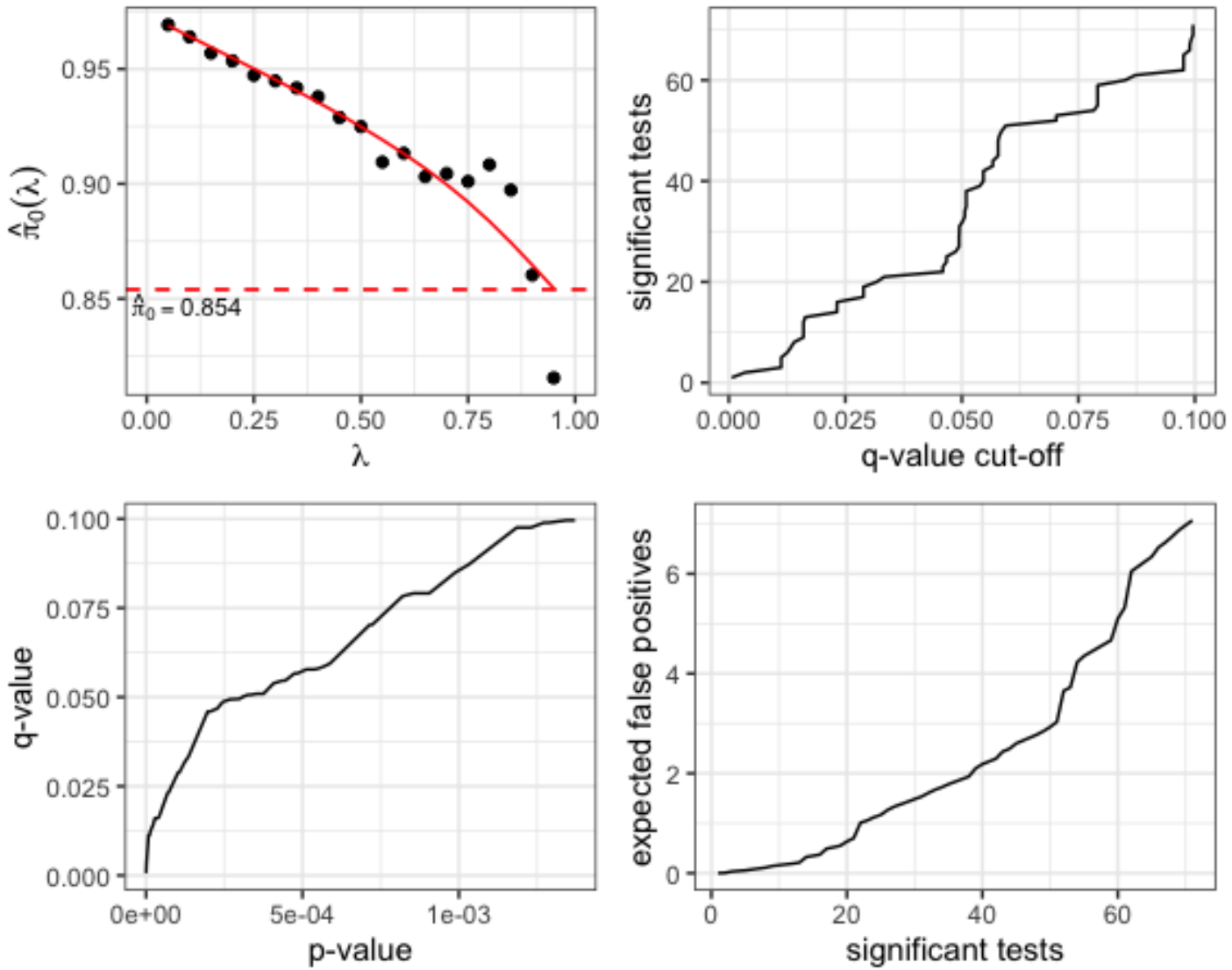


Figure 8: *q*-value plots for prostate cancer data.

Bayesian False Discoveries/Non-Discoveries

- In a Bayesian approach, based on Bayes factors we have a rule to flag a single association as noteworthy if:

$$\begin{aligned} \text{Posterior Odds} &= \text{Bayes Factor} \times \text{Prior Odds} \\ &< R \end{aligned}$$

  where R is the ratio of costs of type II to type I errors.
- In a multiple testing situation in which m associations are being examined nothing, in principle, changes.
- We simply apply the same rule m times, perhaps changing the priors if we have different priors for different associations.
- The choice of threshold, R, and hence the procedure, does not depend on: the number of tests being carried out[4].

[4] unless the prior on the null, or the ratio of costs of errors depends on the number of tests

Bayesian False Discoveries/Non-Discoveries

- As we have seen, the Bayes factor depends, crucially, on the sample size.
- In contrast, multiple testing based on *p*-values (e.g. Bonferroni/Sidak) does not depend on the sample size but, crucially, on the number of tests *m*.
- We have already noted that *p*-value calibration is very difficult, and we would like a procedure by which *p*-value thresholds decrease to zero with increasing sample size.
- The same would also be required of EFD or FDR based procedures.

Bayesian False Discoveries/Non-Discoveries

To summarize in the case of normal test statistics:

The Bayesian decision is based on the Z score and on the sample size, n, but not on the number of tests, m.

In contrast:

The Bonferroni decision is based on the Z score and on the number of tests, m, but not on the sample size, n.

Bayesian Multiple Testing

In a Bayesian context, for a single test:

- If we call a hypothesis noteworthy then $\Pr(H_0 | \text{ data })$ is the probability of a false discovery.
- If we call a hypothesis not rejected then $\Pr(H_1 | \text{ data })$ is the probability of a false non-discovery.

A Key Point: A Bayesian analysis of a single SNP alone, or the same SNP from multiple SNPs will produce the same decision (assuming the prior is the same).

Bayesian False Discoveries/Non-Discoveries

In a multiple-hypothesis testing situation (and assuming ordered so the first K are rejected), we have

$$\begin{aligned}
\text{Expected number of false discoveries} &= \sum_{i=1}^{K} \Pr(H_{0i}|\ \text{data}_i) \\
\text{Proportion of false discoveries} &= \frac{1}{K}\sum_{i=1}^{K} \Pr(H_{0i}|\ \text{data}_i) \\
\text{Expected number of false non-discoveries} &= \sum_{i=K+1}^{m} \Pr(H_{1i}|\ \text{data}_i) \\
\text{Proportion of false non-discoveries} &= \frac{1}{m-K}\sum_{i=K+1}^{m} \Pr(H_{1i}|\ \text{data}_i).
\end{aligned}$$

Bayesian Multiple Testing

- In the frequentist approaches, the expected FDR is with respect to infinite hypothetical identical situations; the above Bayesian approach we have posterior summaries so they are conditional on the data (and are also dependent on the model).
- Another important difference is that the Benjamini Hochberg FDR was defined with respect to a tail-area, whereas these Bayesian measures condition on $\boldsymbol{Y} = \boldsymbol{y}$.

Local FDR Estimation

The local FDR corresponding to a test statistic z_0 is defined as

$$\text{FDR}(z_0) = \Pr(\text{gene } i \text{ is null} | z_i = z_0).$$

Note: The local bit just refers to conditioning on a particular value, rather than a tail area.

We have

$$\text{FDR}(z) = \frac{\pi_0 f_0(z)}{f(z)}.$$

In practice $f(z)$ is replaced by $\widehat{f}(z)$.

Empirical Bayes method

- Efron's local FDR (Efron *et al.*, 2001) uses a two-groups model to estimates the proportion of null/signal genes as a function of Z_i.

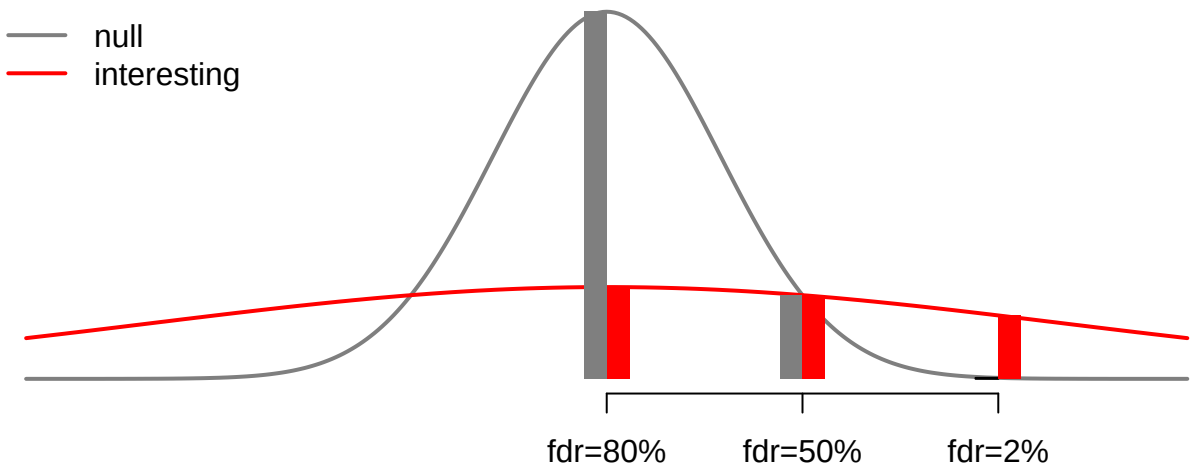


Figure 9: Local FDR.

- Estimating the 'null' component from the middle of the data, subtracting it from an overall density estimate, we can estimate local FDR, denoted FDR(Z).

Prostate cancer example

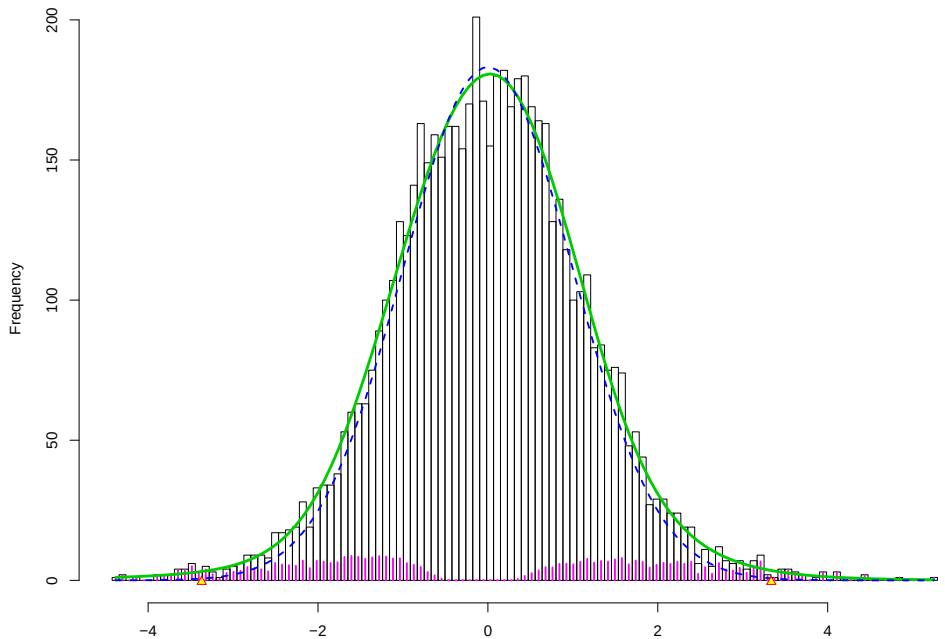


Figure 10: Local FDR for prostate cancer data. Blue dashed curve is distribution if all null. The green solid line is the spline-based estimate of the mixture density f. Pink are non-null signals. $\widehat{\pi}_0 = 0.932$.

We find 25 genes with $\widehat{\text{FDR}}(Z_i) < 0.1$.

Local false sign rates

Stephens (2017) has recently proposed an approach building on previous ideas and with a number of benefits:

- Assumes effect sizes are drawn from a unimodal distribution which allows more accurare inference (lower variance), provided the assumption holds, and convenient computation.
- The method requires two inputs for each feature, estimate and uncertainty, in contrast to *p*- and *q*-values, so accounts for the power of the test/experiment.
- It provides an estimate of the effect size, along with uncertainty (i.e., the posterior distribution).

Empirical Bayes (EB) is used for estimation, as its computationally simpler – this adaptive shrinkage method is referred to as ash.

Local false sign rates: Specific details

The approach takes as input an estimate $\widehat{\beta}_i$ and standard error s_i for the i-th signal and then (also) builds a hierarchical mixture model.

The posterior for β is

$$p(\beta_i|\widehat{\beta}_i, s_i) \propto p(\widehat{\beta}_i|\beta_i, s_i) \times p(\beta_i),$$

and the prior for β is assumed to be independent from a unimodal pdf with

$$p(\beta_i) = \pi_0\delta_0(\beta_i) + \sum_{k=1}^{K} \pi_k \mathrm{N}(\beta_i|0, \sigma_k^2),$$

where $\delta_0(\cdot)$ is a point mass at 0, and the mixing proportions π_k are to be estimated, as are the mixture variances σ_k^2.

Local false sign rates: Specific details

- Estimation would focus on the posterior distribution $p(\beta_i|\widehat{\boldsymbol{\beta}}, \boldsymbol{s}, \widehat{\boldsymbol{\pi}})$.
- To gauge significance of observation i, we can examine the local FDR (Efron, 2008):

$$\text{LFDR}_i \quad = \quad \Pr(\beta_i = 0|\widehat{\boldsymbol{\beta}}, \boldsymbol{s}, \widehat{\boldsymbol{\pi}})$$

  This is the probability of being incorrect if we were to declare significance when actually null.
- This measures reflects the classic focus on whether an effect is exactly zero, and Stephens (2017) prefers the local false sign rate (LFSR).
- The LFSR is the probability that we would make an error in the sign if we were forced to declare it either positive or negative (a Type S error).
- Formally,

$$\text{LFSR}_i = \min\left[\Pr(\beta_i \geq 0|\widehat{\boldsymbol{\beta}}, \boldsymbol{s}, \widehat{\boldsymbol{\pi}}), \Pr(\beta_i \leq 0|\ \widehat{\boldsymbol{\beta}}, \boldsymbol{s}, \widehat{\boldsymbol{\pi}})\right].$$

Local false sign rates

Example: Suppose that

$$\begin{aligned} \Pr(\beta_i < 0 | \widehat{\pi}, \widehat{\beta}, \boldsymbol{s}) &= 0.95 \\ \Pr(\beta_i = 0 | \widehat{\pi}, \widehat{\beta}, \boldsymbol{s}) &= 0.03 \\ \Pr(\beta_i > 0 | \widehat{\pi}, \widehat{\beta}, \boldsymbol{s}) &= 0.02 \end{aligned}$$

Then, $\text{LFSR}_i = \min(0.05, 0.98) = 0.05$.

This LFSR corresponds to the fact that, given these results, we would guess that β_i is negative, with probability 0.05 of being wrong.

The LFDR_i is 0.03 in this example.

Local false sign rates

Small values of LFDR_i indicate we can be confident that $\beta_i \neq 0$, while small values of LFSR_i indicate we can be confident in the sign of β_i.

Being confident in the sign of an effect implies we are confident it is non-zero, and $\text{LFSR}_i \geq \text{LFDR}_i$

In this sense, LFSR is a more conservative measure of significance than LFDR.

LFSR: Prostate data

For the prostate cancer data, $\widehat{\pi}_0 = 0.84$.

```
head(out$result[,1:5],5)
        betahat sebetahat NegativeProb PositiveProb       lfsr
1  0.394234285 0.2656347   0.029291126   0.15911902 0.8408810
2  0.703227359 0.1900733   0.001392831   0.90157430 0.0984257
3 -0.006046081 0.2173035   0.056465210   0.05461837 0.9435348
4 -0.239860003 0.2106722   0.122247036   0.02933164 0.8777530
5 -0.033913878 0.2412444   0.063317423   0.05393860 0.9366826
```

```
head(out$result[,6:10],5)
      svalue       lfdr     qvalue PosteriorMean PosteriorSD
1 0.67579181 0.81158985 0.65031517   0.0350449209  0.10830331
2 0.04354673 0.09703287 0.04299731   0.4186750962  0.20706512
3 0.86172602 0.88891642 0.80947053  -0.0003780271  0.05434714
4 0.74093749 0.84842132 0.71902507  -0.0211092735  0.08025763
5 0.85043901 0.88274398 0.79390938  -0.0020380021  0.05940802
```

LFSR: Illustration of shrinkage

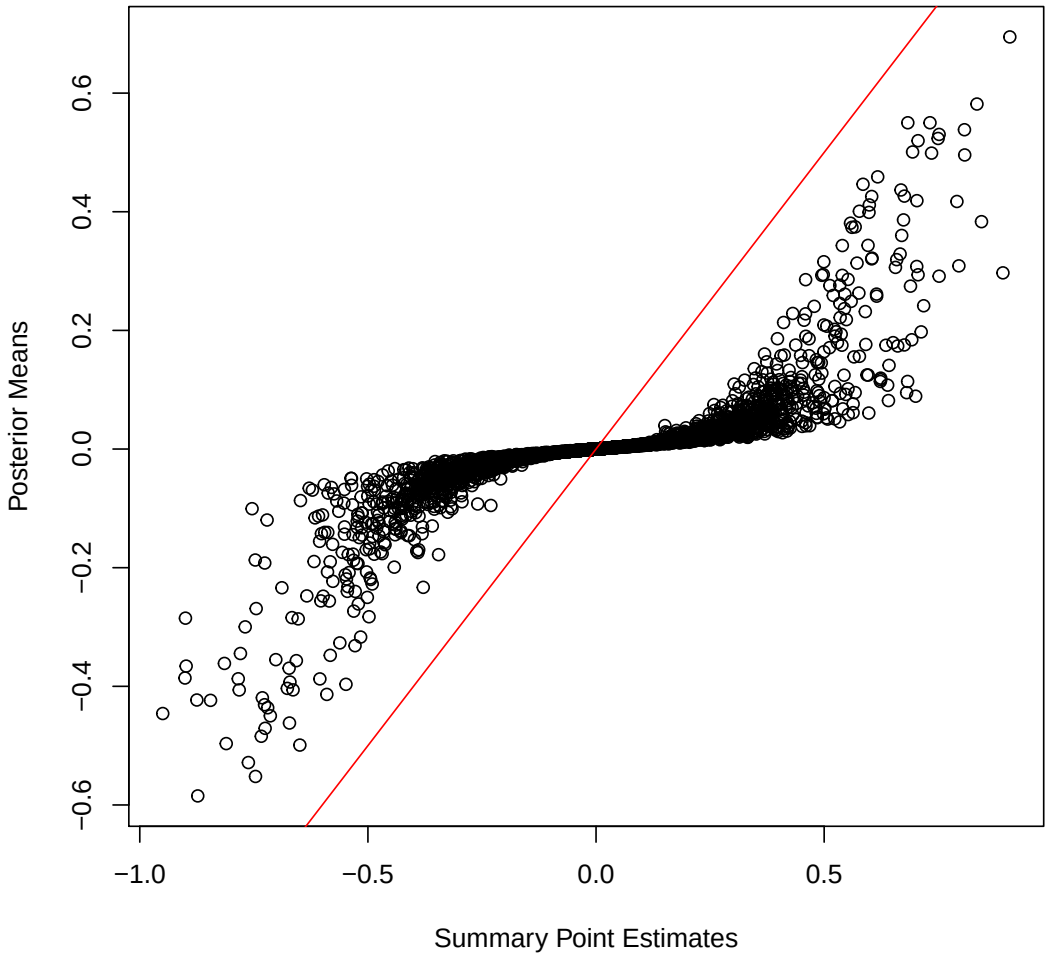


Figure 11: Posterior mean versus summary point estimates.

LFSR: comparison with LFDR

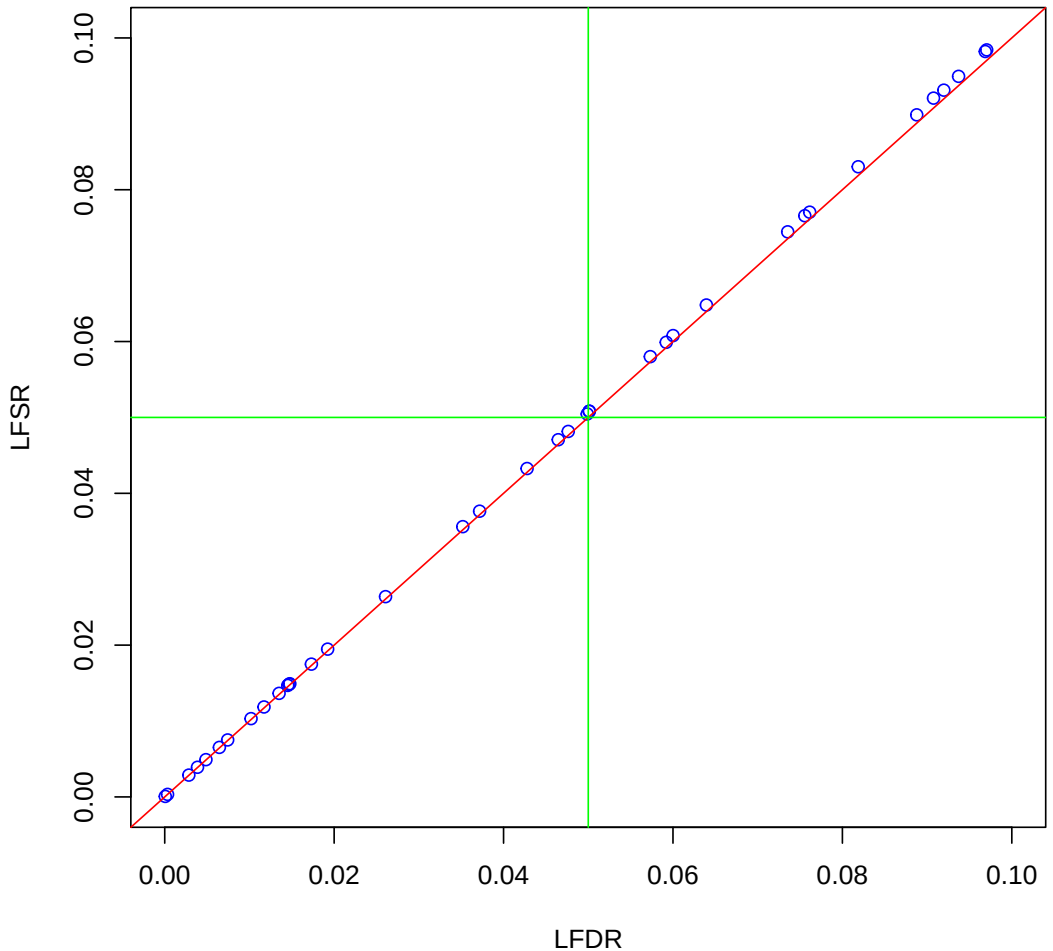


Figure 12: LFSR versus LFDR.

Conclusions

Conclusions

- Bayesian analysis is attractive in a multiple testing context, but the results are very sensitive to the prior on the proportion of nulls, π_0.
- Fast methods are required for large m (e.g. in a GWAS context) of tests, which is still a drawback for many Bayesian approaches.
- Such priors can have a major impact on rankings and posterior probabilities.

Summary

What to do with multiple comparisons is a difficult problem:

- Apart from doing nothing, the only truly 'default' method is Bonferroni, which may not answer a relevant question, and/or may not answer it very well.
- Bonferroni is poorly-understood, as are other methods.
- If we use estimation (for example, via a hierarchical model) we can avoid multiple comparison problems (though care in the model specification needed).
- There are many summaries of techniques, see for example Efron and Hastie (2016).
- Stephens (2017) is a very good discussion of modern techniques.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Efron, B. (2008). Microarrays, empirical Bayes and the two groups model (with discussion). *Statistical Science*, **23**, 1–47.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

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.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, **5**, 189–211.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Pearson, E. (1953). Discussion of “Statistical inference” by D.V. Lindley. *Journal of the Royal Statistical Society, Series B*, **15**, 68–69.

Sellke, T., Bayarri, M., and Berger, J. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, **55**, 62–71.

Sidák, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**, 626–633.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, **18**, 275–294.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the *q*-value. *The Annals of Statistics*, **31**, 2013–2035.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Science*, **100**, 9440–9445.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The*

Official Publication of the International Genetic Epidemiology Society, **33**, 79–86.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

Westfall, P., Johnson, W., and Utts, J. (1995). A Bayesian perspective on the bonferroni adjustment. *Biometrika*, **84**, 419–427.

Appendix

Bayes Bonferroni

There is a prior that results in a Bayesian Bonferroni-type correction[5].

If the prior probabilities of each of the nulls are independent with $\pi_{0i} = \pi_0$ for $i = 1, \ldots, m$.

Then the prior probability that all nulls are true is

$$\Pi_0 = \Pr(H_1 = 0, \ldots, H_m = 0) = \pi_0^m$$

which we refer to as prior P_1, and let $\alpha_{i,\mathrm{B}}$ be the posterior probability of the null under this prior for gene i.

Example if $\pi_0 = 0.5$ and $m = 10$, $\Pi_0 = 0.00098$, which may not reflect the required prior belief.

[5] The following describes a very idealized setting where the data model and prior are both normal

Bayes Bonferroni

Suppose instead that we wish to fix the prior probability that all of the nulls are true at Π_0.

A simple way of achieving this is to take $\pi_{0i} = \Pi_0^{1/m}$, a specification we call prior P_2.

Westfall *et al.* (1995) show that for independent tests

$$\begin{aligned} \alpha^{\star}_{i,\mathrm{B}} &= \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_2) \\ &\approx m \times \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_1) \\ &= m \times \alpha_{i,\mathrm{B}}. \end{aligned}$$

So a Bayesian version of a Bonferroni-like result is recovered.

As we have seen before, the posterior probability on the null, is strongly dependent on the prior on the null.

Multiple testing: Does Bayes help?

Efron's $\widehat{\text{FDR}}(z)$ is an 'empirical Bayes' method – it 'borrows strength' from the collection $z_i, i = 1, \ldots, m$, to say what happens at specific z.

Hierarchical models also do this, using prior assumptions of exchangeability to motivate borrowing strength across subgroups.

As shown by Gelman *et al.* (2012)[6], this is not the same as, for example, Bonferroni.

They also discuss Type S errors, which are sign errors, i.e., saying an association is positive when it is truly negative.

[6] In a paper entitled, 'Why We (Usually) Don't Have to Worry About Multiple Comparisons'

Multiple testing: Does Bayes help?

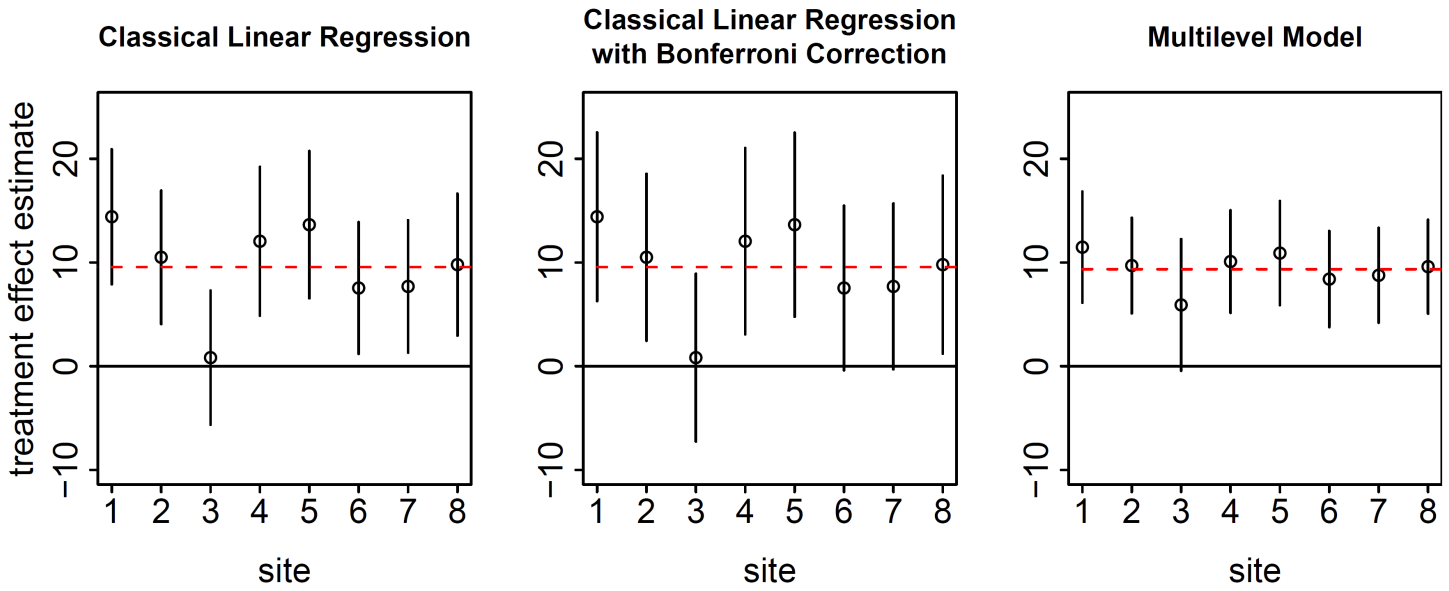


Figure 13: Point and 95% intervals, reproduction of Figure 1 from Gelman *et al.* (2012).

Compared to simpler methods, multilevel approaches do allow better inference on vectors of parameters – generally by trading some bias for reduced variance.

Bayes Mixture Model

We consider the mixture model described in Chapter 4 of Wakefield (2013).

The sampling model is $Y_i|\mu_i \sim \text{N}(\mu_i, \sigma_i^2)$, where the σ_i^2 are assumed known.

We specify a mixture model for the collection $[\mu_1, ..., \mu_m]$, with

$$\mu_i = \begin{cases} 0 & \text{with probability } \pi_0 \\ \sim \text{N}(\delta, \tau^2) & \text{with probability } \pi_1 = 1 - \pi_0 \end{cases}$$

We use mixture component indicators $H_i = 0/1$ to denote the zero/normal membership model for transcript i.

Bayes Mixture Model

Collapsing over μ_i gives the three stage model:

Stage One:

$$Y_i \mid H_i, \delta, \tau, \pi_0 \sim_{ind} \begin{cases} \mathsf{N}(0, \sigma_i^2) & \text{if } H_i = 0 \\ \mathsf{N}(\delta, \sigma_i^2 + \tau^2) & \text{if } H_i = 1. \end{cases}$$

Stage Two: $H_i \mid \pi_1 \sim_{iid} \text{Bernoulli}(\pi_1)$, $i = 1, ..., m$.

Stage Three: Independent priors on the common parameters:

$$p(\delta, \tau, \pi_0) = p(\delta)p(\tau)p(\pi_0).$$

Bayes Mixture Model

We illustrate the use of this model with

$$\begin{aligned} p(\delta) &\propto 1, \\ p(\tau) &\propto 1/\tau \\ p(\pi_0) &= 1, \end{aligned}$$

so that we have improper priors for δ and τ^2.

The latter choice still produces a proper posterior, because we have fixed variances at the first stage of the model.

Implementation is via a Markov chain Monte Carlo algorithm; Exercise 4.4 of Wakefield (2013) derives details of the algorithm.

Bayes Mixture Model

For transcript i, we may evaluate the posterior probabilities of the alternative

$$\begin{aligned}
\Pr(H_i = 1 \mid y_i) &= \mathrm{E}[H_i \mid \boldsymbol{y}] \\
&= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\mathrm{E}(H_i \mid \boldsymbol{y}, \delta, \tau^2, \pi_0)\right] \\
&= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\Pr(H_i = 1 \mid \boldsymbol{y}, \delta, \tau^2, \pi_0)\right] \\
&= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\frac{p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2) \times \pi_1}{p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2) \times \pi_1 + p(\boldsymbol{y} \mid H_i = 0) \times \pi_0}\right]
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2, \pi_0) &= [2\pi(\sigma_i^2 + \tau^2)]^{-1/2} \exp\left[-\frac{(y_i - \delta)^2}{2(\sigma_i^2 + \tau^2)}\right] \\
p(\boldsymbol{y} \mid H_i = 0, \delta, \tau^2, \pi_0) &= [2\pi\sigma_i^2]^{-1/2} \exp\left[-\frac{y_i^2}{2\sigma_i^2}\right].
\end{aligned}$$

Bayes Mixture Model

Expression (7) averages $\Pr(H_i = 1 \mid \boldsymbol{y}, \delta, \tau^2, \pi_0)$ with respect to the posterior $p(\delta, \tau^2, \pi_0 \mid \boldsymbol{y})$, and may be simply evaluated via

$$\frac{1}{T}\sum_{t=1}^{T} \frac{p(\boldsymbol{y} \mid H_i = 1, \delta^{(t)}, \tau^{2(t)})\pi_1^{(t)}}{p(\boldsymbol{y} \mid H_i = 1, \delta^{(t)}, \tau^{2(t)}, \pi_0^{(t)})\pi_1^{(t)} + p(\boldsymbol{y} \mid H_i = 0)\pi_0^{(t)}}$$

given samples $\delta^{(t)}, \tau^{2(t)}, \pi_0^{(t)}$, $t = 1, \ldots, T$, from the Markov chain.

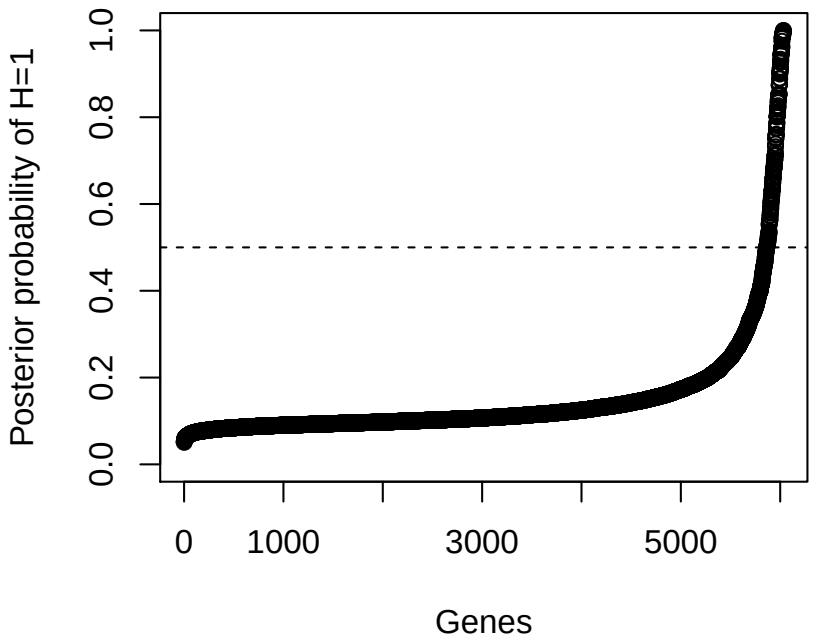


Figure 14: Posterior probability of alternative for prostate cancer.