# 2018 SISG MODULE 20: Bayesian Statistics for Genetics
# Lecture 9: Bayesian and Frequentist Testing

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

# Outline

# Multiple Testing: Motivating Data Description

- The Vitamin Intervention for Stroke Prevention (VISP) trial is an NIH-funded, multi-center, double-blind, randomized, controlled clinical trial.
- More detail in Wakefield *et al.* (2014).
- The aim is to determine whether a daily intake of high dose folic acid and vitamins B6 and B12 was associated with cardiovascular endpoints.
- We examine data on $n = 1670$ individuals, with 837 randomized to the high dose and 833 to the low dose.

# Motivating Data Description

- The outcome is the intermediary variable homocysteine level: high levels in blood are associated with cardiovascular disease.
- In the VISP trial, levels were measured longitudinally but for simplicity we take as outcome the difference between the baseline and the first post-baseline measurements: $Y$ will represent this difference.
- The change was -0.37 $\mu$mol/L in the low dose group versus -2.36 $\mu$mol/L in the high dose group, i.e., a difference of -1.99 $\mu$mol/L ($p < 2 \times 10^{-16}$).

# Marker-Specific Treatment Effects

- An increasingly important venture is examining treatment effects by marker (e.g. SNP): a particular type of gene-environment interaction.
- Historically, candidate gene studies were popular, but now genome-wide scans are also being performed, see Daly (2010) for a review.
- Pharmacogenomics-related traits: Drug response, susceptibility to adverse drug reactions,...
- **Key Statistical Point:** The estimated interactions are based on subgroups of varying sizes, so that the power varies substantially across tests.
- In the **VISP** trial, there are $J = 803,122$ SNPs and suppose we define subgroups as having at least one copy of the minor allele.

The Pharmacogenomics Journal (2012), 1–9

www.nature.com/tpj

ORIGINAL ARTICLE

# Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis

M Man[1], SL Close[2], AD Shaw[3], GR Bernard[4], IS Douglas[5], RJ Kaner[6], D Payen[7], J-L Vincent[8], S Fossceco[1], JM Janes[1], AG Leishman[1], L O'Brien[1], MD Williams[9] and JGN Garcia[10]

[1]*Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA;* [2]*Department of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN, USA;* [3]*Department of Anesthesiology, Duke University Medical Center, Durham, NC, USA;* [4]*Division of Allergy, Pulmonary and Critical Care, Vanderbilt University School of Medicine, Nashville, TN, USA;* [5]*Denver Health Medical Center, University of Colorado, School of Medicine, Denver, CO,*

Management of severe sepsis, an acute illness with high morbidity and mortality, suffers from the lack of effective biomarkers and largely empirical predictions of disease progression and therapeutic responses. We conducted a genome-wide association study using a large randomized clinical trial cohort to discover genetic biomarkers of response to therapy and prognosis utilizing novel approaches, including combination markers, to overcome limitations of single-marker analyses. Sepsis prognostic models were dominated by clinical variables with genetic markers less informative. In contrast, evidence for gene–gene interactions were identified for sepsis treatment responses with genetic biomarkers dominating models for predicting therapeutic responses, yielding candidates for replication in other cohorts.
*The Pharmacogenomics Journal* advance online publication, 7 February 2012; doi:10.1038/tpj.2012.1

**Keywords:** drotrecogin alfa (activated); epistasis; genetic markers; genome-wide association study; polymorphism; severe sepsis

Aim: To identify marker-defined populations with improved response to DAA (for treatment of severe sespis).

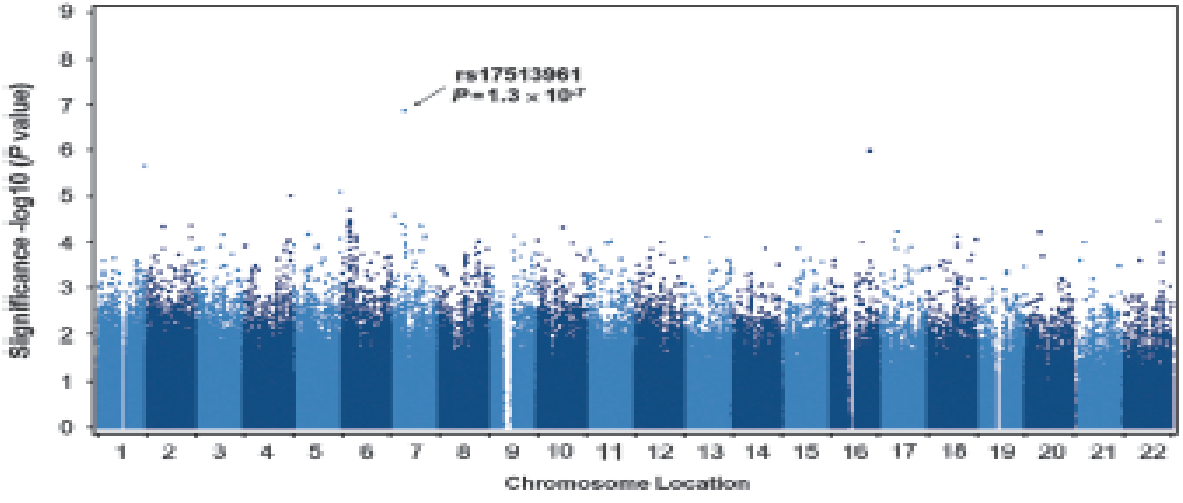



**Figure 1** Representative Manhattan plot from GWAS in the entire cohort for genotype BB vs not BB. All possible combinations of genotypes representing a dominant, heterozygous and recessive inheritance were evaluated. Three pairs of comparisons AA vs not AA, AB (heterozygous) vs not AB and BB vs not BB were completed. This figure represents the plot of chromosome position and *P*-values for homozygous genotype (BB) vs heterozygous (AB) or homozygous for the other allele (AA), or BB vs not BB. (Manhattan plots of GWAS results for genotype AA vs not AA and AB vs not AB are shown in Supplementary Figure 1). GWAS = genome-wide association study.

# The Statistical Set-Up

Before considering sample size and power calculations we review frequentist and Bayesian test procedures.

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

$$H_0 : \theta = 0$$

  with $\theta$, 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.
- 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 given by:

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

  where $t_{\text{obs}}$ is a number that is evaluated for the data at hand.
- 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 $\alpha$?

- 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 $\alpha$-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 normal$(\theta, 1)$,

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

  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 $\alpha$, regardless of sample size — imbalance in $\alpha$ and $\beta$ as a function of $n$:

| $n$ | $\alpha$ | $\beta$ | $k(n)$ |
|---|---|---|---|
| 1 | 0.01 | 0.91 | 2.33 |
| 25 | 0.01 | 0.0038 | 0.46 |
| 100 | 0.01 | $8 \times 10^{-15}$ | 0.23 |

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

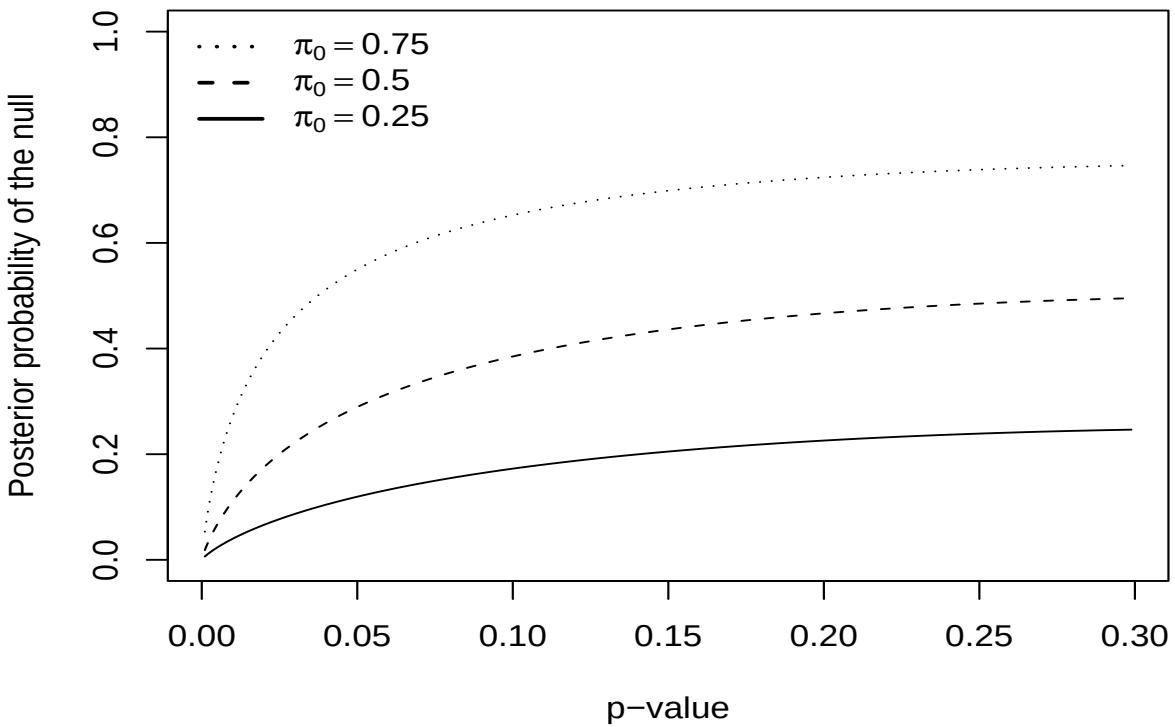
[2]Note that the threshold for $T = [\overline{Y}_n/(1/\sqrt{n})]^2$ is constant

# 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 $\alpha$-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:

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

where N is shorthand for the density of a normal random variable.

# 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 $\theta$ 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 $\approx 0.05$ — 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 with $W$ not depending on $n$ is consistent (you get the right answer with a lot of data) if one of $H_0$ or $H_1$ is true, whereas the "*p*-value" Bayes factor is not.
- The original 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 $\pi_1$ and the ratio of costs of type II to type I errors.
- Larger values on the $y$ axis correspond to less extreme test statistics.
- The curves have the expected ordering and, as $n$ gets large, a greater and greater level of evidence is required.
- 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.

# A Bayes Factor Threshold

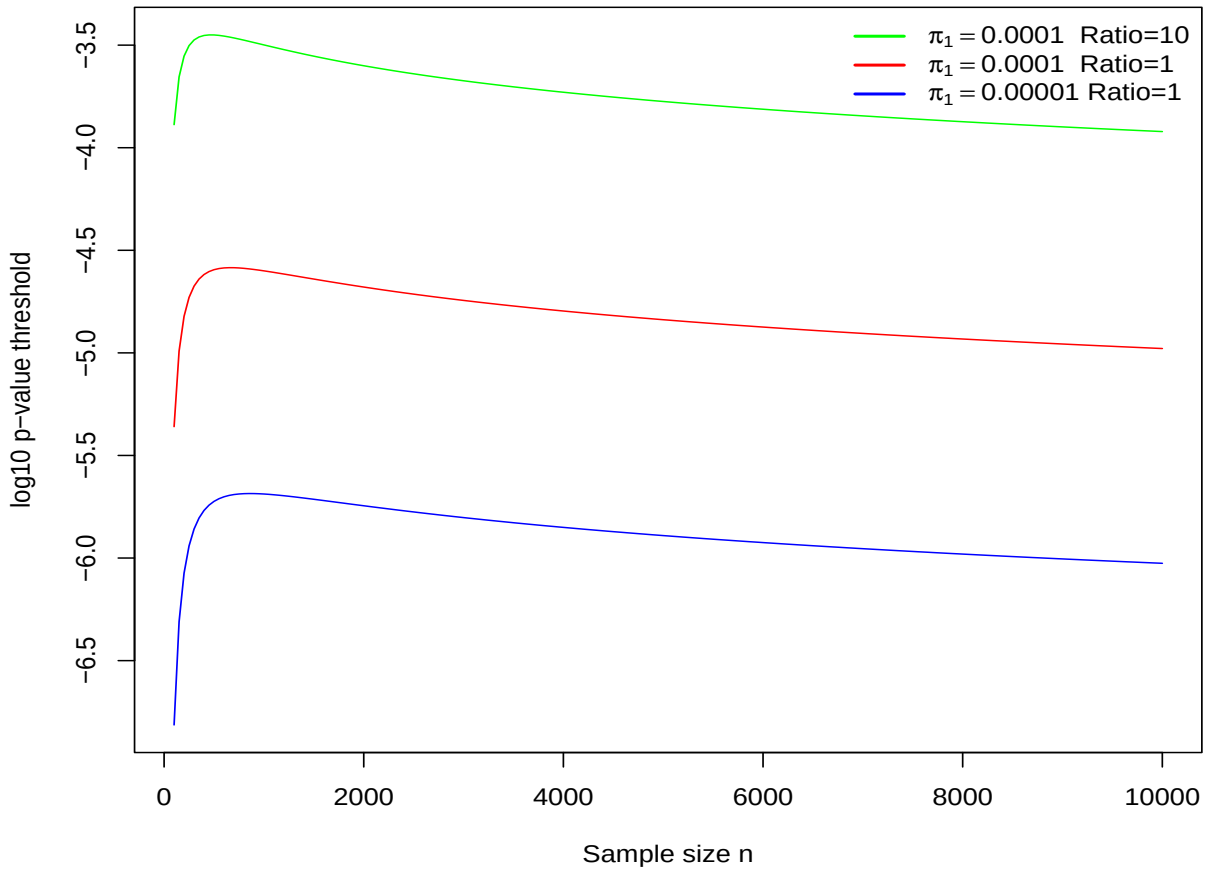



Figure 1: Threshold for rejection, on the $\log_{10}(p)$-value scale, versus sample size. Notice how the threshold is decreasing with increasing sample size.

# Loss functions: introduction

In Bayesian decision theory, the loss function is a required ingredient that must be specified.

"How bad can the decision be", is captured by the loss function – how bad the decision $d$ would be, if the truth were $\beta$.

The loss function clearly depends on the aim of the analysis (what is the decision problem):

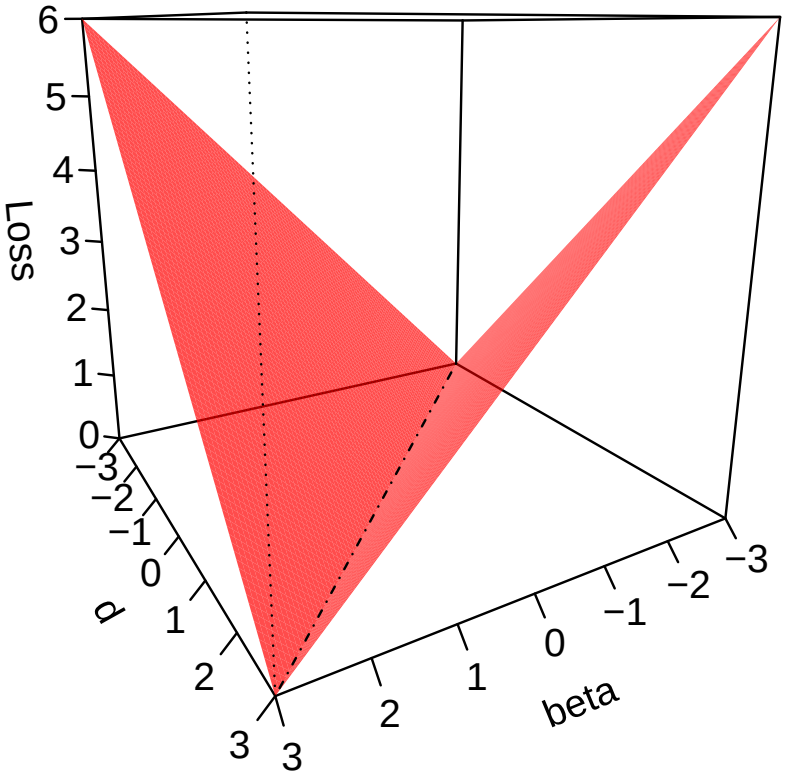
- Reporting a single summary of location for a parameter of interest: point estimation.
- Report a range of values for a parameter of interest: interval estimation.
- To decide between two hypotheses.

# Loss functions: point estimation

Two possible loss functions for point estimation are displayed below.

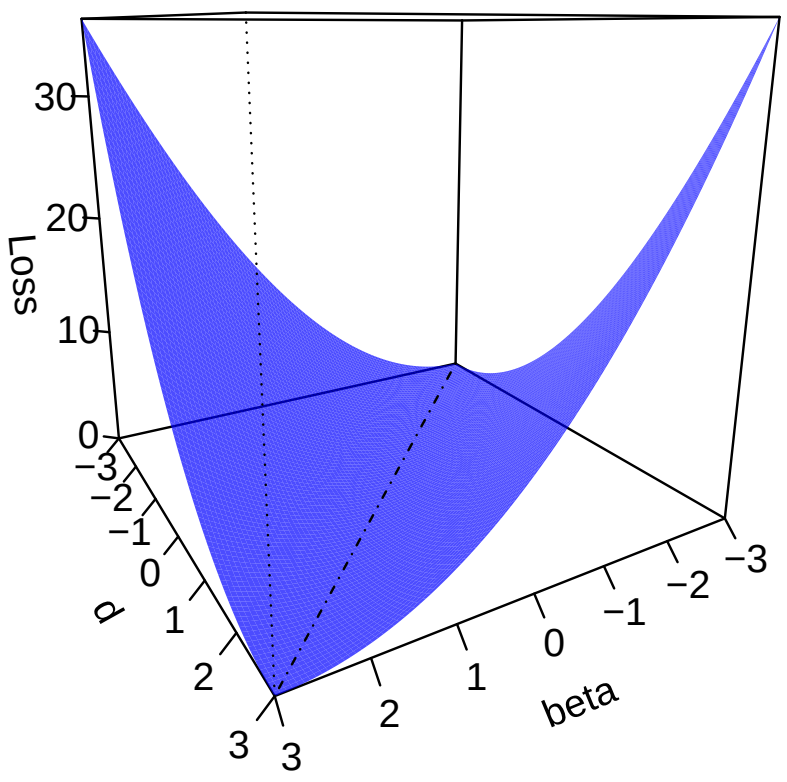
## Absolute loss

$$L(\beta, d) = |\beta - d|$$




## Quadratic loss

$$L(\beta, d) = (\beta - d)^2$$



# Point Estimation

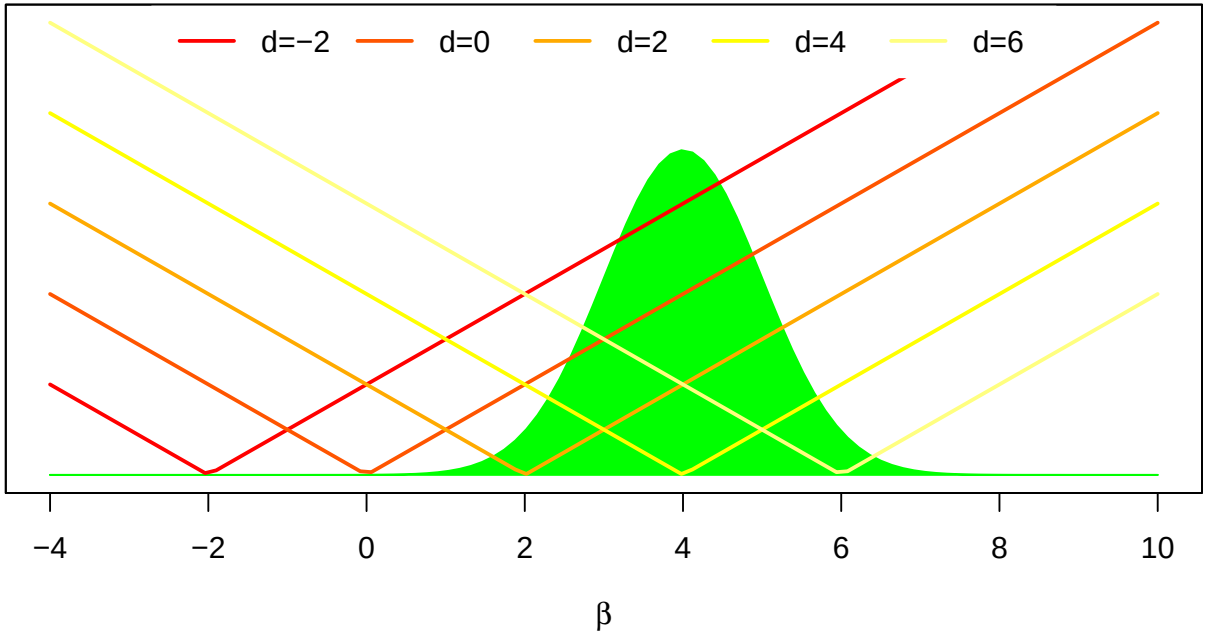
The game is to choose $d$, with $\beta$ begin uncertain.

Bayesian decision theory chooses $d$ by minimizing the expected posterior loss:

$$d^B = \text{argmin}_d \text{E}[L(\beta, d)|\boldsymbol{y}] = \text{argmin}_d \int L(\beta, d)\pi(\beta|\boldsymbol{y})\, d\beta.$$

# Point Estimation

For this green posterior and absolute loss $L(\beta, d) = |\beta - d|$, which choice of $d$ minimizes the expected loss?




This optimal decision $d^B$ (called the Bayes rule) is the posterior median.

# Minimizing posterior expected loss

Why minimize the expected posterior loss?

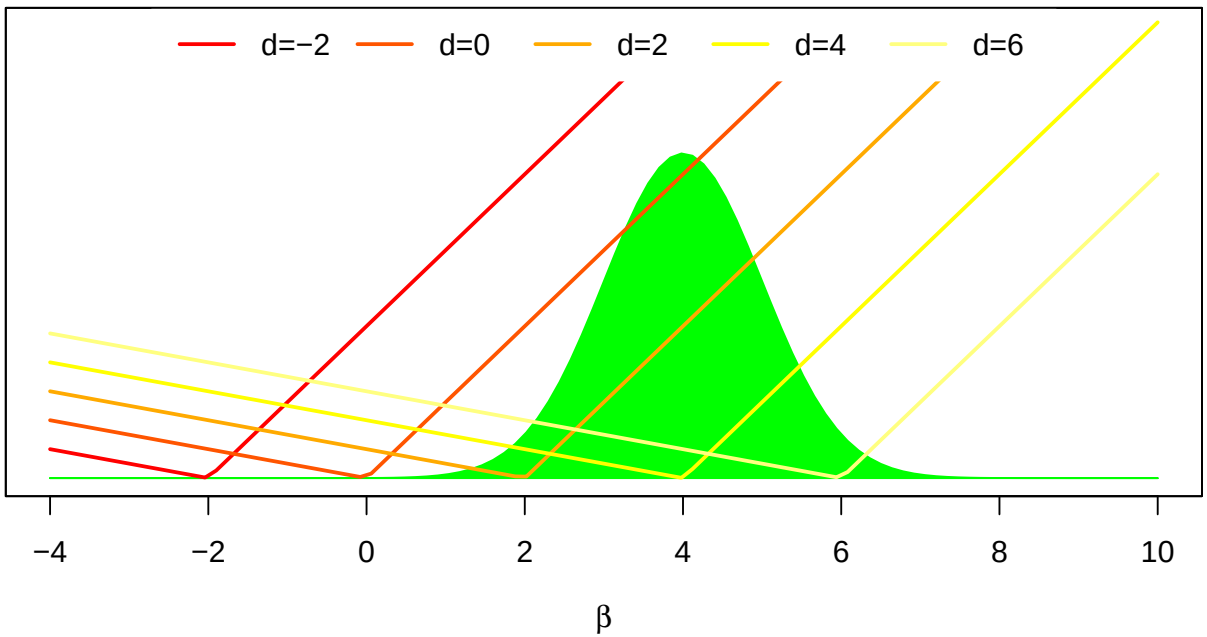
- Minimizing $L(\beta, d)$ averaged – sensibly – over (posterior) uncertainty.
- Good frequentist properties! Using the Bayes rule in repeated experiments, $d^B$ minimizes the loss You (i.e. person with Your prior) would expect to suffer.
- Can't be too awful! Complete class theorems show that, essentially, any rule that isn't a Bayes rule will have worse loss, at least sometimes.

If these seem unconvincing/esoteric, note that any method of choosing between decisions implicitly has a loss function $L(\beta, d)$, so it's useful to explicitly consider different choices.

# Point estimation

A more complex loss function; suppose under-estimates of $\beta$ are worse than over-estimates.

Example: Number of disease cases is $\beta$, aim is prediction for resource planning.




Compared to the earlier symmetric losses, how does the Bayes rule change?

# Loss functions: introduction

The math of this[3]

$$L(\beta, d) = \begin{cases} \alpha|\beta - d|, & d < \beta \quad \text{(i.e., under-estimate)} \\ (1-\alpha)|\beta - d|, & d > \beta \quad \text{(i.e., over-estimate)} \end{cases}$$

The Bayes rule here is to report the $\alpha$ quantile (i.e. $\alpha \times 100\%$ percentile) of the posterior.

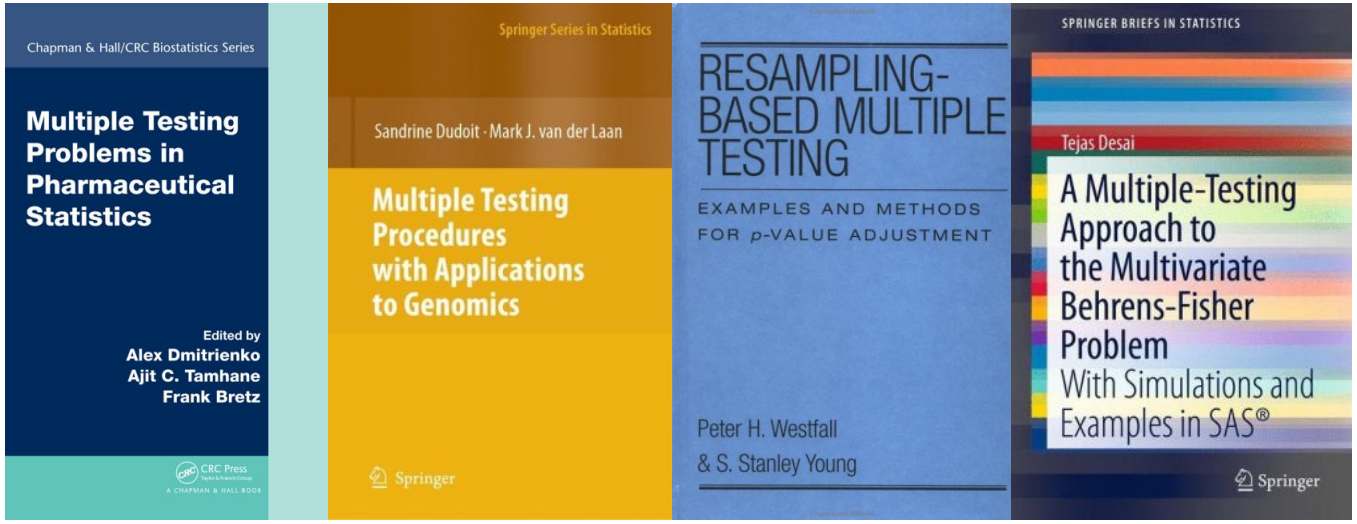
Example: if $\alpha = 3/4$ (under-estimation 3 times as bad as over-estimation) we report that value of $\beta$ that puts 75% of the mass of the posterior to the left and 25% to the right, i.e., we are overestimating.

[3] previous example would have $\alpha > 1/2$

# 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:




which has seen a lot of interest lately, given the advent of technologies that allow huge numbers of experiments to be performed.

As with testing, 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.

Gene expression levels were measured for 6033 genes.

A two-standard t-test was carried out, and then 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 $i = 1, \ldots, N = 6033$ genes.

Under the alternative:

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

The aim is to find genes with non-zero $\mu_i$.

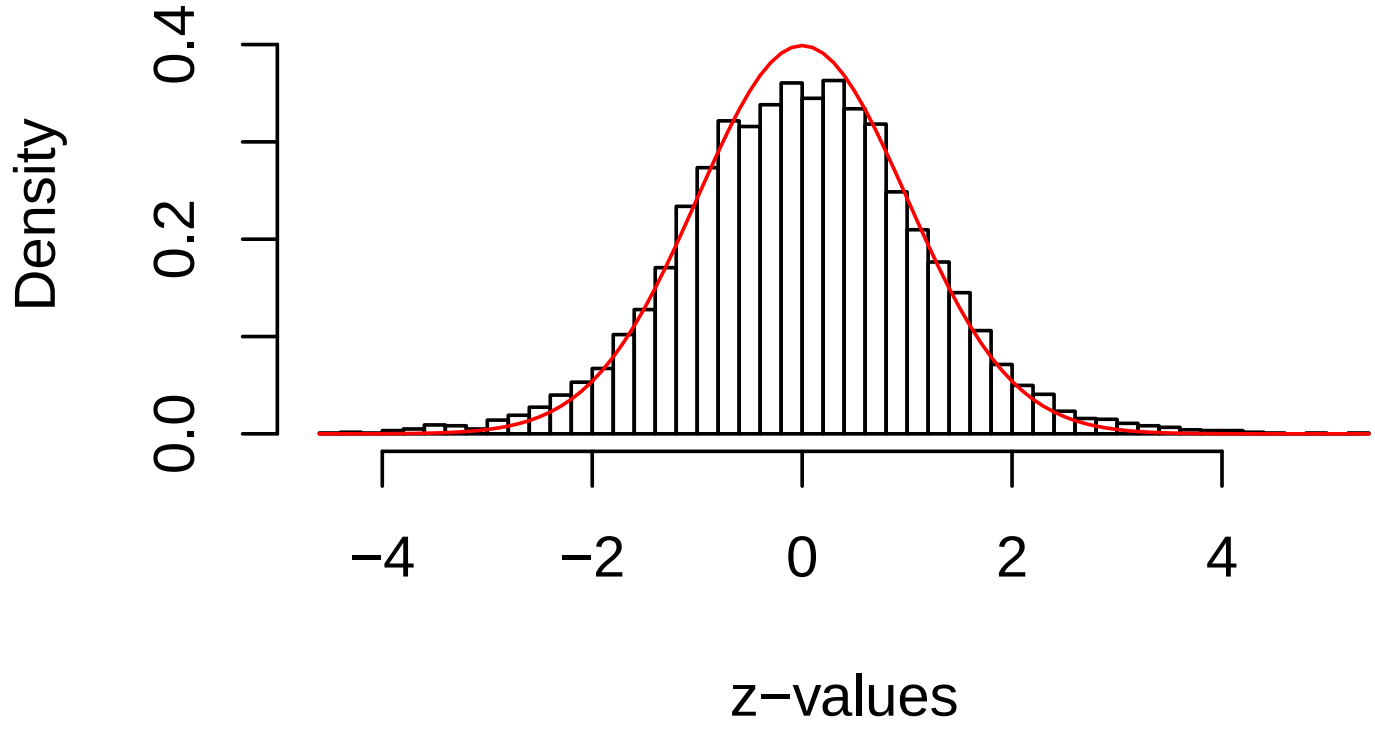



Figure 2: Histogram of *z*-values for prostate microarray study.

# 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.

# Motivation for Bonferroni

Suppose we have $m$ independent predictors (e.g., suppose we are in the GWAS situation).

Let $p^\star$ be the smallest $p$-value out of $m$ obtained.

The rationale for $p$-values is there relationship to the incorrect rejection of null hypotheses; when we have $m$ tests and report the smallest then we need to think about the properties of $p$-values with respect to the complete procedure.

For example, suppose all $m$ nulls are true, then

$$1 - \underbrace{(1 - p^\star)^m}_{\text{Don't reject any}}$$

is the probability of false rejection of at least one null; e.g., $m = 20$, $p^\star = 0.02$ gives 0.33, so a third of the time we will reject at this level.

# 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 $m = 500K$ tests take $\alpha^\star = 0.05/500,000 = 10^{-7}$.
- 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 $\alpha$ should depend on $n$, in partcular should decrease as $n$ gets larger and larger).

# Sidak Correction

- If all the tests are independent (same calculation as before):

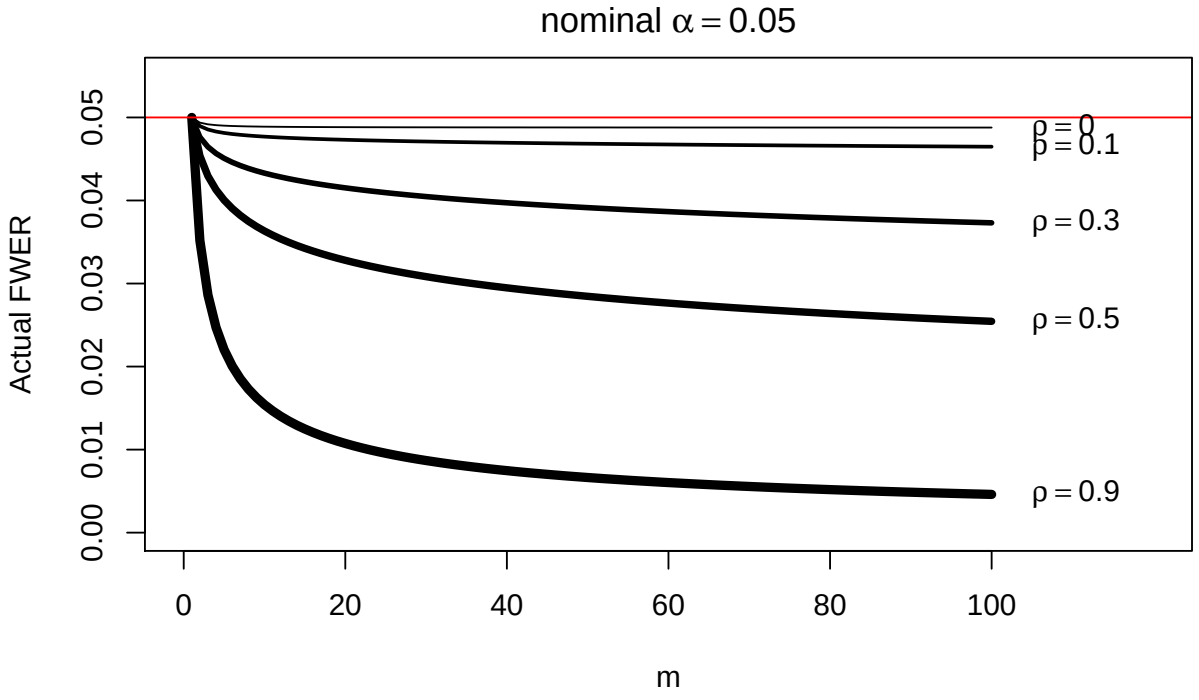
$$\begin{aligned}
\Pr(B \geq 1) &= 1 - \Pr(B = 0) \\
&= 1 - \Pr(\cap_{i=1}^{m} B_i') \\
&= 1 - \prod_{i=1}^{m} \Pr(B_i') \\
&= 1 - (1 - \alpha^\star)^m
\end{aligned}$$

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

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

# Bonferroni under dependence

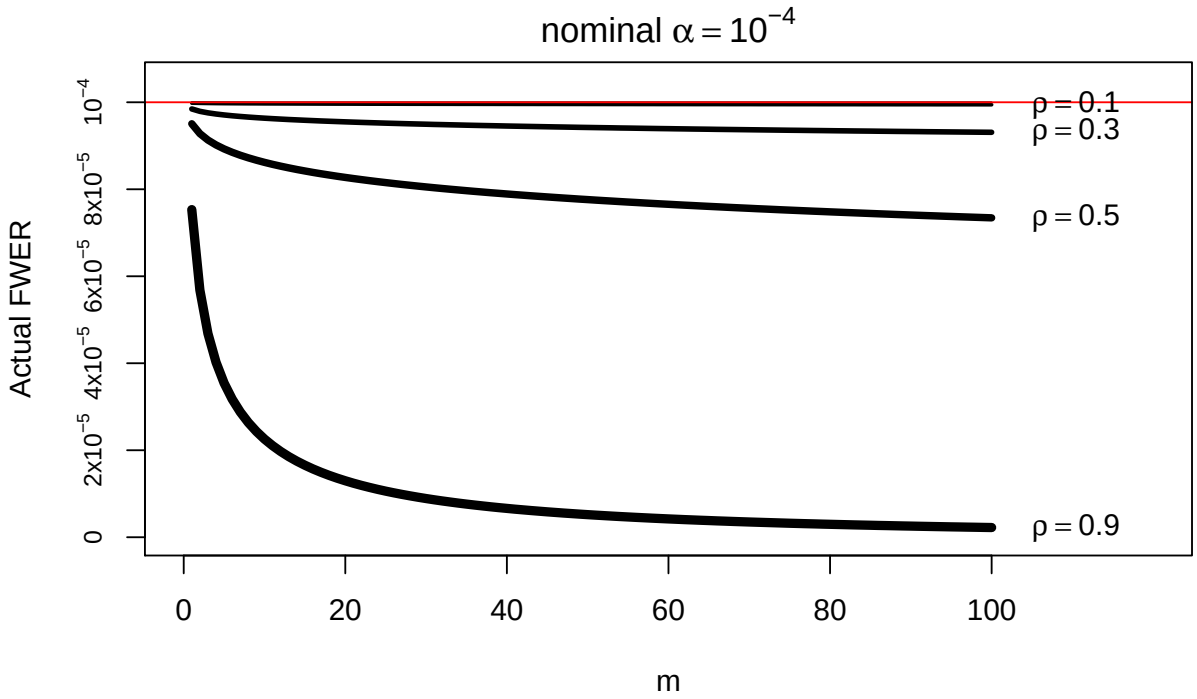
For dependent tests, the situation is messier. For $p_j$ based on $Z_j$, where all pairs of $Z_j, Z_{j'}$ have correlation $\rho$;




- For modest $\rho$, not hugely conservative – but the number of tests $m$ does matter

# Bonferroni under dependence

Repeating the calculation for a lower $\alpha$, Bonferroni's degree of conservatism changes;




- A family of $> 100$ highly-correlated tests is unlikely in practice; 'blocks' of highly-correlated tests are more plausible
- For tiny $\alpha$, e.g. $5 \times 10^{-8}$, conservatism often *very* minor

# Bayes Bonferroni

There is a prior that results in a Bayesian Bonferroni-type correction[4].

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 $\alpha_B$ be the posterior probability of the null under this prior.

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

[4] The following describes a very idealized setting where the data 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 $\Pi_0$.

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

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

$$\alpha_{\mathrm{B}}^{\star} = \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_2) \approx m \times \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_1) = m \times \alpha_{\mathrm{B}}.$$

So a Bayesian version of Bonferroni is recovered.

# 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 $\alpha$ for each test, is given by

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

if all nulls are true.

# Expected False Discoveries

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

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

- In a GWAS context suppose $m = 500K$ and $\alpha = 0.05$; this gives 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 $\alpha/m$.

# 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

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

# 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 $\alpha$ is the value at 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 algorithm was originally proposed by Simes (1986) to control the FWER.

# False Discovery Rate

If all the nulls are true then $B = K$ (all rejections are false) and

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

FDR in this form and with extensions, e.g. Storey and Tibshirani (2003) (description of the $q$-value methodology) have 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.

# $q$-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[5].

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 $\text{pFDR}(t_{\text{fix}})$.

[5] 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

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

# *q*-values

From this expression, the crucial role of $\pi_0$ is evident.

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

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

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

with $\lambda$ 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 $\lambda$, and then inflates this to account for the proportion of null *p*-values in $[0, \lambda]$.

# *q*-values

This method highlights the benefits of allowing the totality of *p*-values to estimate fundamental quantities of interest such as $\pi_0$.

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), first define a set of nested rejection regions

$$\{t_\alpha\}_{\alpha=0}^{1}$$

where $\alpha$ is such that

$$\Pr(T > t_\alpha \mid H = 0) = \alpha.$$

# $q$-values

Then,

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

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

The $q$-value is defined as

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

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

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

# $q$-values

The $q$-values are estimated as

$$\widehat{q}_i(p_i) = \min_{t \geq p_i} \widehat{\text{pFDR}}(t).$$

Note that in some papers (Storey and Tibshirani, 2003) the $q$-values is defined in terms of the FDR, since often $m$ is large and $\Pr(D > 0) \approx 1$ and FDR $\approx$ pFDR.

# $q$-values

It can be shown that,

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

so that the evidence for $H_0$ given the exact ordinate is always greater than that corresponding to the tail area.

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.

# Illustration from Storey and Tibshirani (2003)

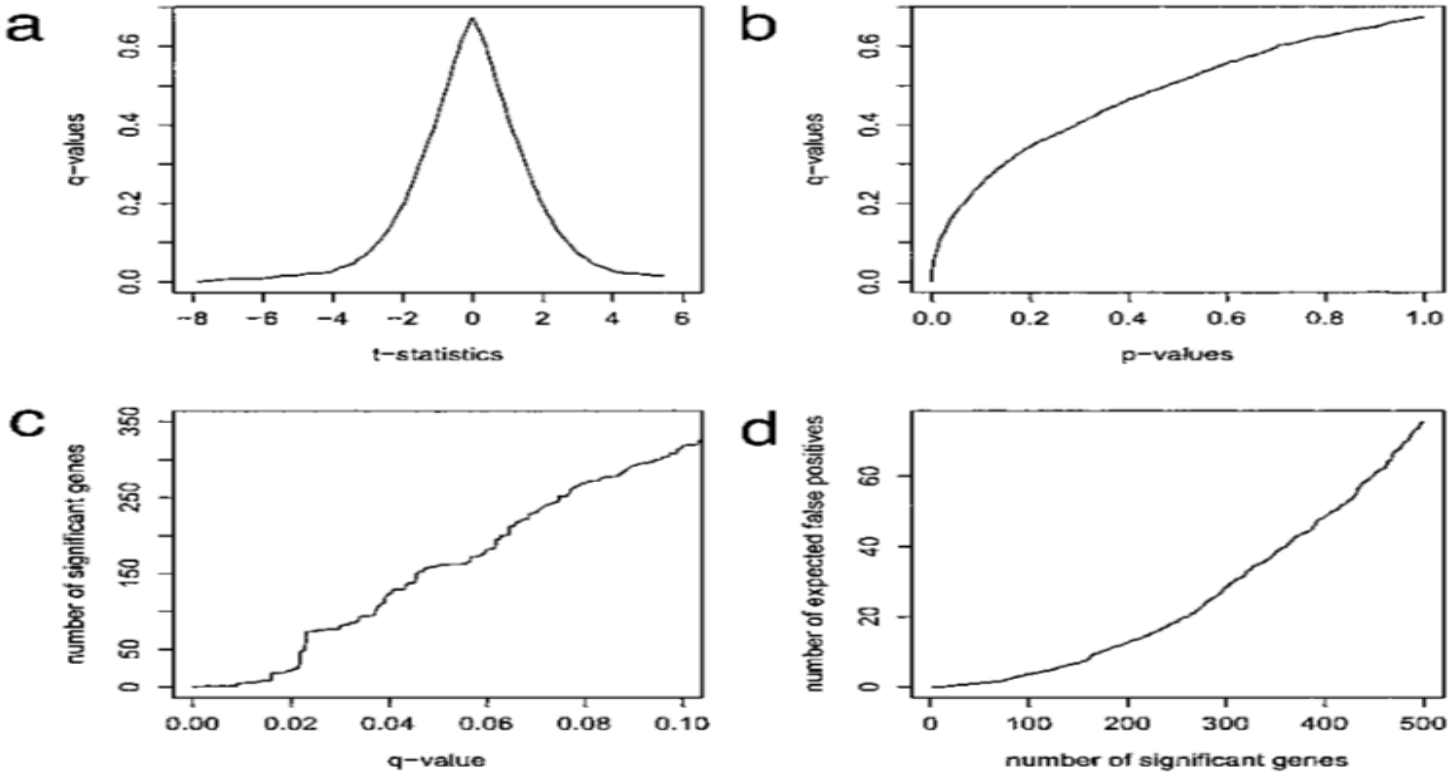



**Fig. 2.** Results from the Hedenfalk *et al.* (14) data. (*a*) The *q* values of the genes versus their respective *t* statistics. (*b*) The *q* values versus their respective *p* values. (*c*) The number of genes occurring on the list up through each *q* value versus the respective *q* value. (*d*) The expected number of false positive genes versus the total number of significant genes given by the *q* values.

# Simulated Example

- We illustrate control by the family-wide error rate (FWER), the expected number of false discoveries (EFD) and the false discovery rate (FDR).
- We simulate data for $m = 100$ tests, with $m_1 = 5$ being non-null.
- True table:

| | Non-Flagged | Flagged | |
|---|---|---|---|
| $H_0$ | $A$ | $B$ | 95 |
| $H_1$ | $C$ | $D$ | 5 |
| | $m - K$ | $K$ | 100 |

# Prostate Cancer Example

- We begin by plotting, in Figure 3 the observed $p$-values versus those expected under the null, i.e. $i/(m+1)$ for $i = 1, \ldots, m = 6033$.
- Hard to interpret, so we truncate the scales in Figure 4.
- Finally we stretch the scale in Figure 5 to show $-\log_{10}$ the observed $p$-values versus expected $p$-values.
- 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.
- We see that the FWER is very conservative ($p = 0.05/m = 8.3 \times 10^{-6}$, or $-\log_{10}(p) = 5.1$) and only flags 3 genes as being significant.
- The EFD=1 gives a $p$-value threshold of $1/6033 = 0.00017$, or $-\log_{10} p = 3.78$ and gives 21 flagged genes.
- The EFD=5 gives a $p$-value threshold of $5/6033 = 0.00083$, or $-\log_{10} p = 3.08$ and gives 54 flagged genes.
- The FDR control at 5% gives the green diagonal line and rejects 21 tests.

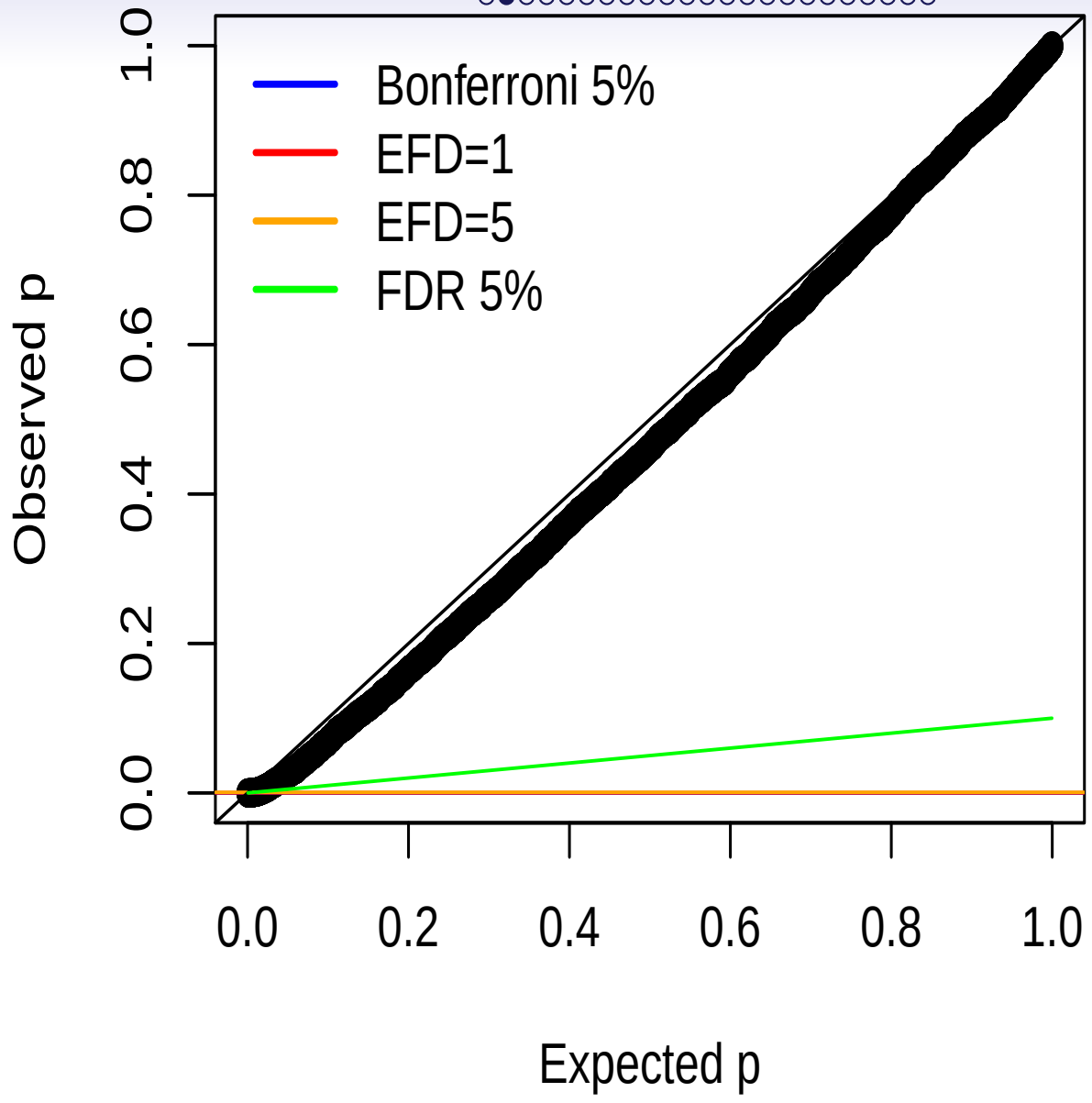



Figure 3: Observed versus expected *p*-values.

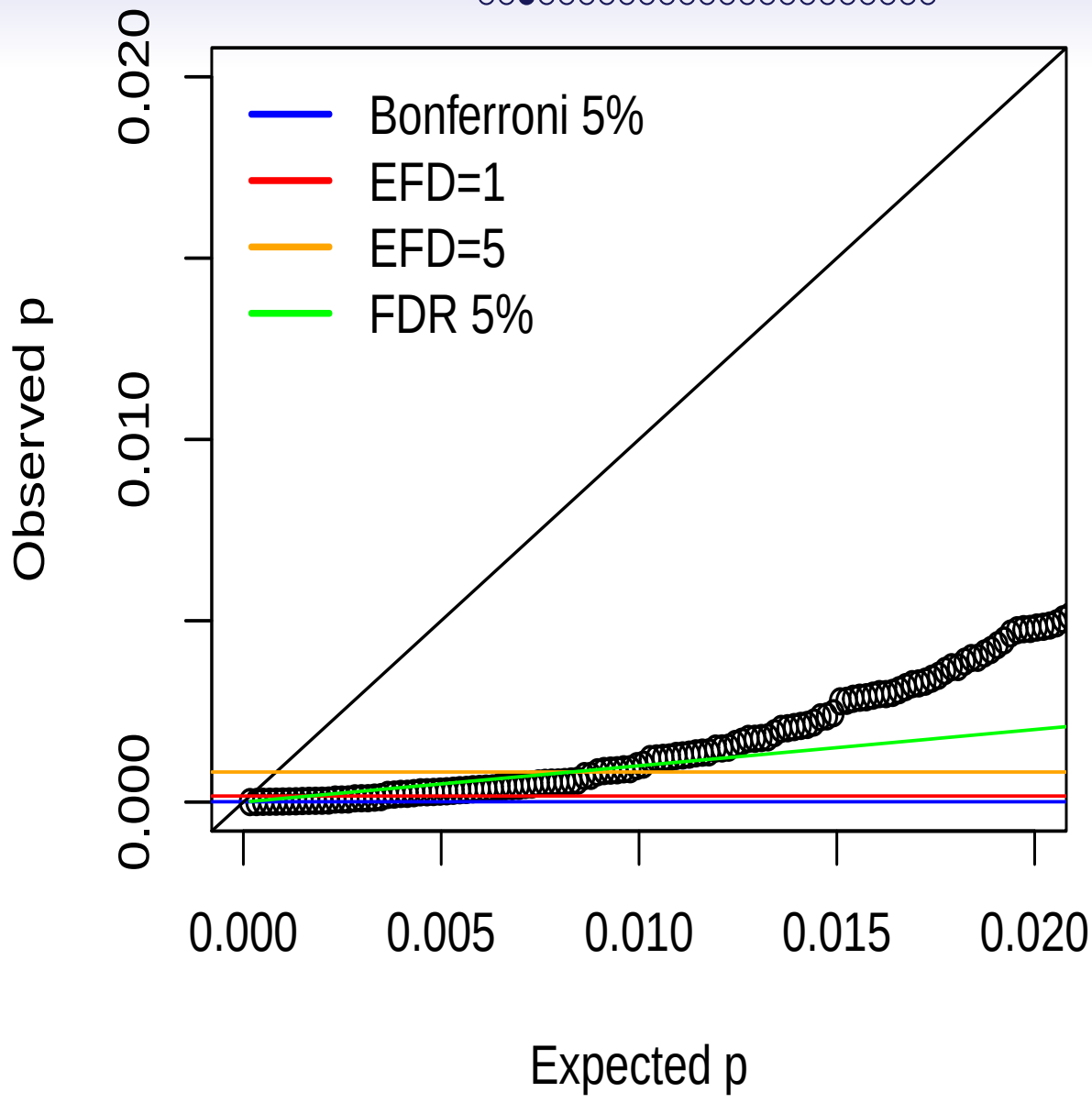



Figure 4: Observed versus expected *p*-values with truncated scale.

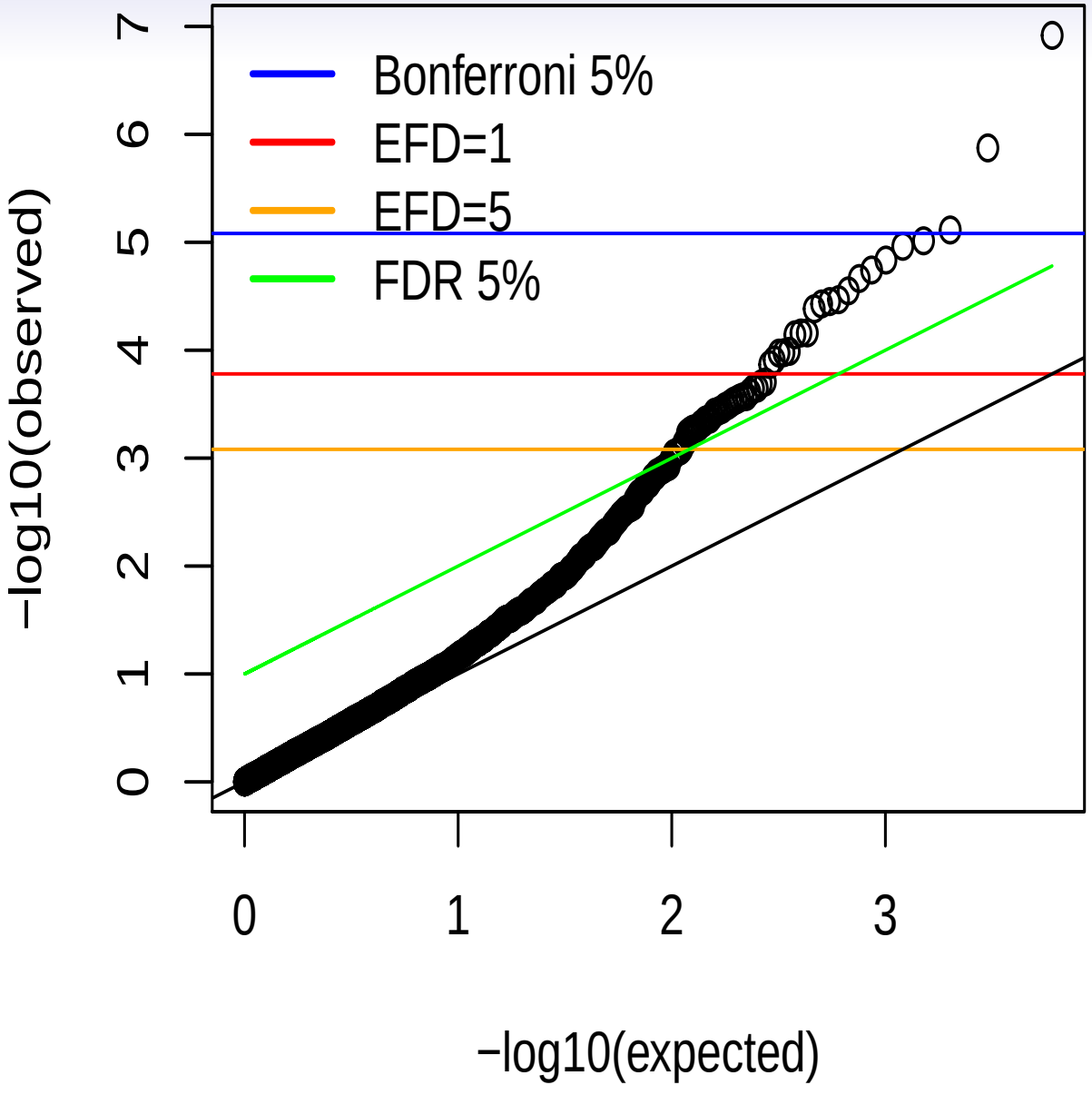



Figure 5: Observed versus expected *p*-values, on $-\log_{10}$ scale.

# 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 $\alpha$ and is slightly less conservative.

# 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[6].

[6] 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 $\Pr(H_0 |$ data $)$ is the probability of a false discovery.
- If we call a hypothesis not rejected then $\Pr(H_1 |$ 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

$$\text{Expected number of false discoveries} = \sum_{i=1}^{K} \Pr(H_{0i}|\ \text{data}_i)$$

$$\text{Proportion of false discoveries} = \frac{1}{K}\sum_{i=1}^{K} \Pr(H_{0i}|\ \text{data}_i)$$

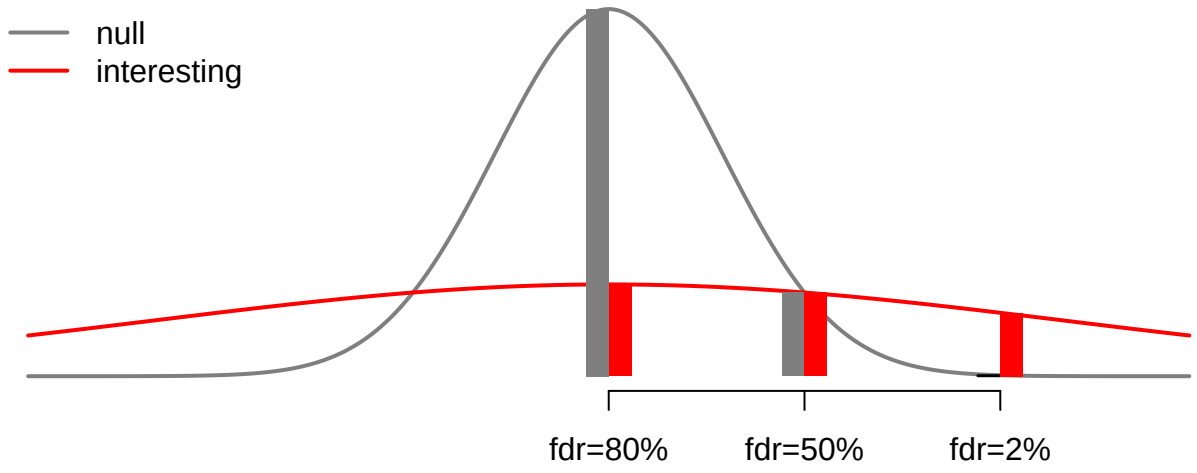
$$\text{Expected number of false non-discoveries} = \sum_{i=K+1}^{m} \Pr(H_{1i}|\ \text{data}_i)$$

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

In the frequentist approaches to the expected FDR is (as usual) with respect to infinite hypothetical identical situations; the above Bayesian approach we have posterior summaries (so they are dependent on the model).

# Empirical Bayes method

More complex; Efron's *local fdr* uses a two-groups model to estimates the proportion of null/signal as a function of $Z_i$;




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

# Local FDR Estimation

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

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

Note: not a tail area.

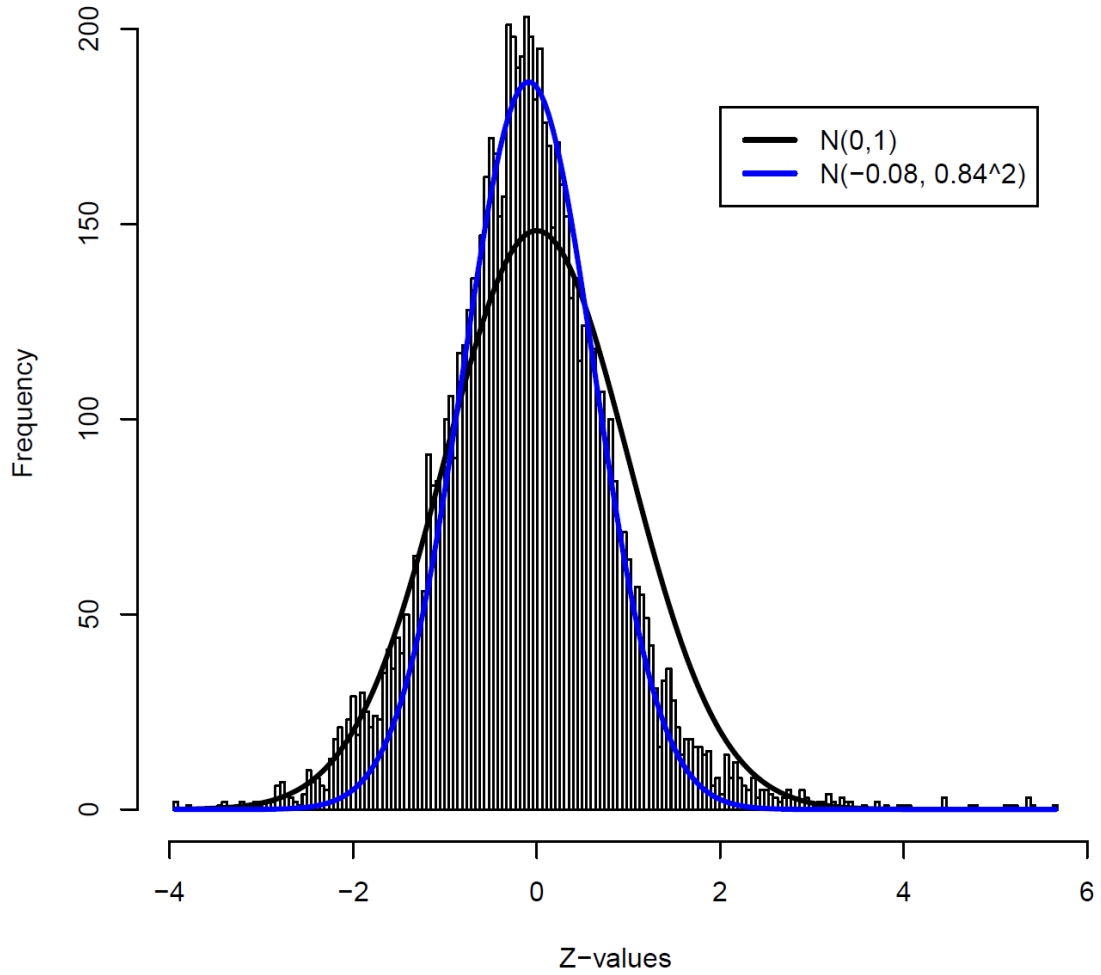
We have

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

In practice $f(z)$ is replaced by $\widehat{f}(z)$, which is estimated via a Poisson model with log mean taken as a polynomial in $z$ (so the $z$ values are binned).

# Empirical Bayes method

In this example, the 'nulls' don't look exactly N(0, 1);




We find 166/7860 gene expressions with $\widehat{\text{FDR}}(Z_i) < 0.1$ for association with HIV status.

# 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)[7], 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.

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

# Multiple testing: Does Bayes help?

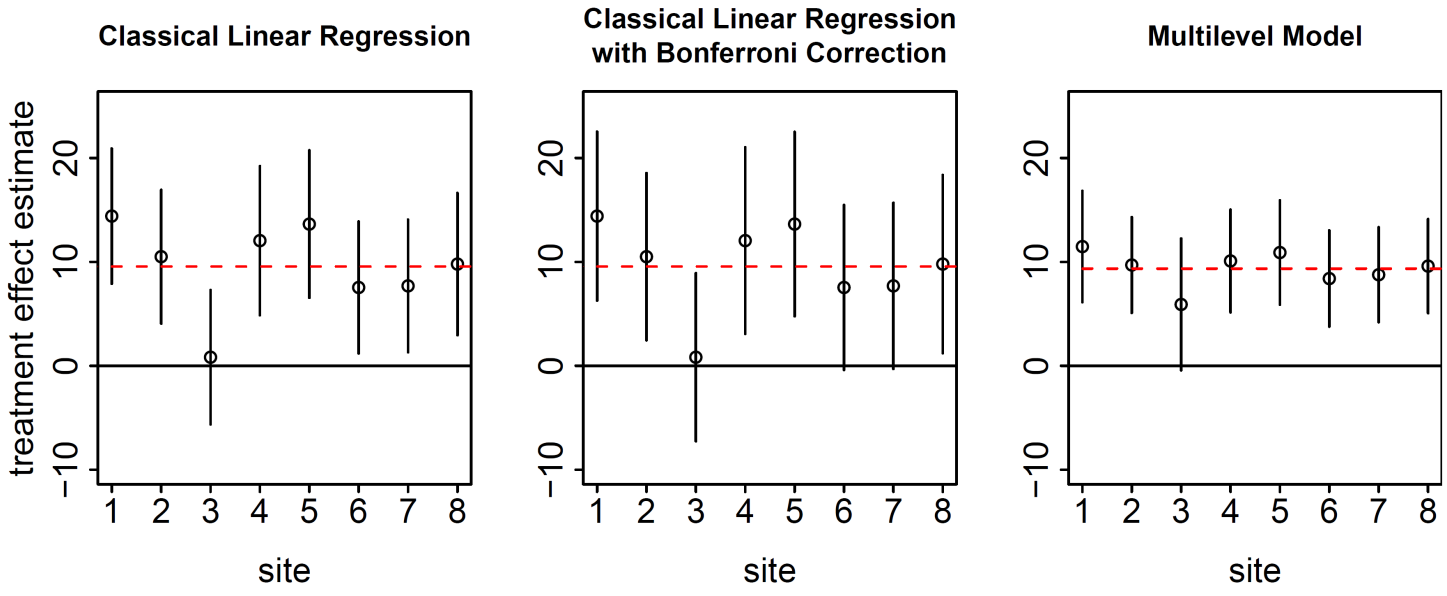



Figure 6: Point and 95% intervals, reproduction of Figure 1 from Gelman *et al.* (2012).

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 $\sigma_i^2$ 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 $\mu_i$ 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 $\delta$ and $\tau^2$.

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

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

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

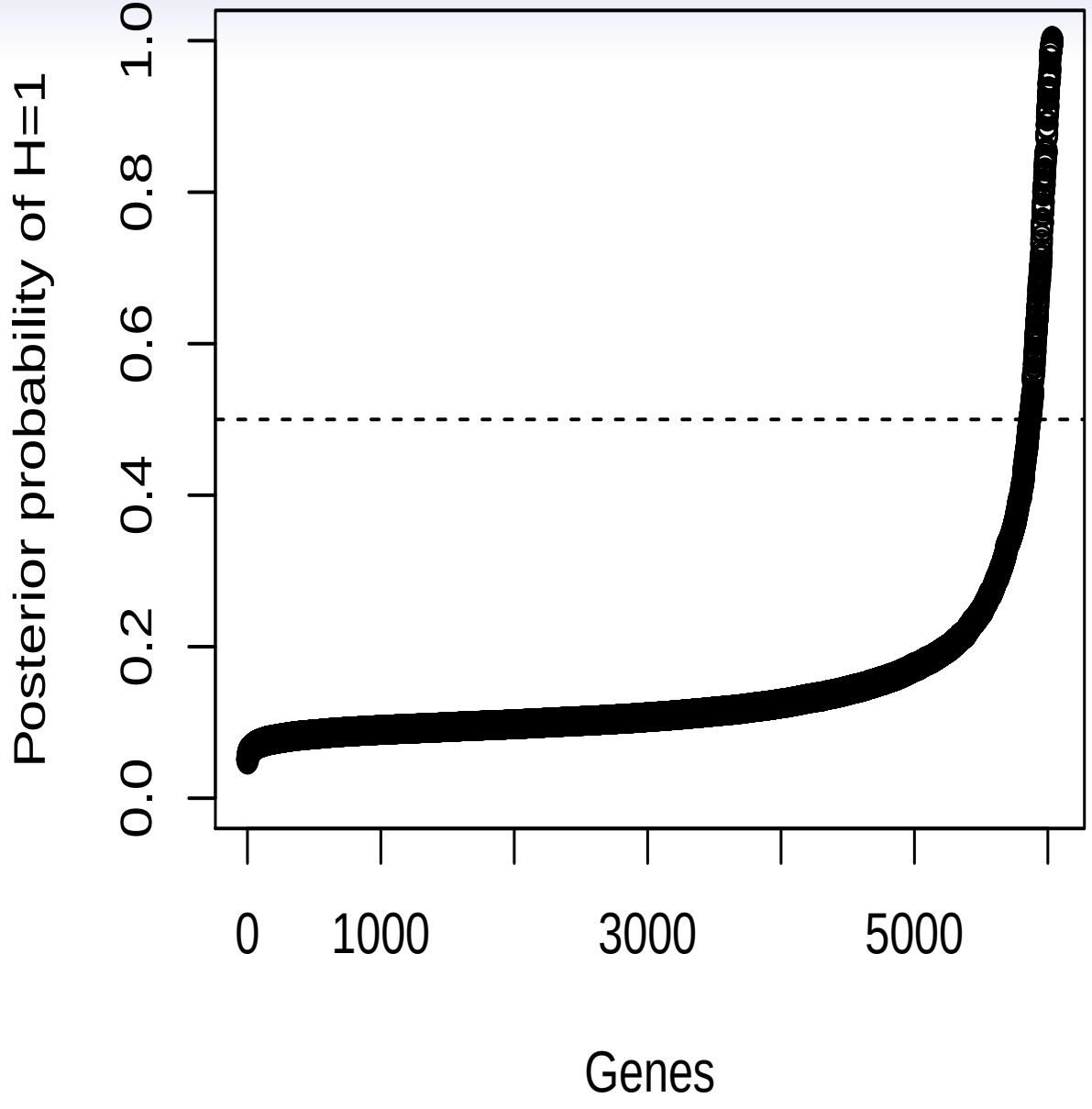



Figure 7: Posterior probability of alternative for prostate cancer.

# Local false sign rates

Stephens (2017) has recently proposed an approach building on previous ideas.

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

The posterior for $\beta$ 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 $\beta$ is assumed to be independent from a unimodal $g$, with

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

where $\delta_0(\cdot)$ is a point mass at 0.

# Local false sign rates

The approach centers on the local false sign rate $\text{LFSR}_i$ which is the probability that we would make an error in the sign of effect *i* if we forced to declare it either positive or negative (a Type S error):

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

Example: Suppose that

$$\begin{aligned}
\Pr(\beta_i < 0|\widehat{\pi}, \widehat{\boldsymbol{\beta}}, \boldsymbol{s}) &= 0.95 \\
\Pr(\beta_i = 0|\widehat{\pi}, \widehat{\boldsymbol{\beta}}, \boldsymbol{s}) &= 0.03 \\
\Pr(\beta_i > 0|\widehat{\pi}, \widehat{\boldsymbol{\beta}}, \boldsymbol{s}) &= 0.92
\end{aligned}$$

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

# Overall Treatment Effect

- We now describe the methodology for the VISP trial.
- Suppose we have two treatments $T = 0/1$ (e.g. low dose/high dose), a continuous response $Y$ and $n/2$ subjects in each treatment group, where $n$ is the number of trial participants.
- Let $Y_i$ be the response for the $i$-th individual and $T_i$ the treatment indicator.
- To estimate the overall treatment effect we fit the model

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

with $\text{var}(\epsilon_i) = \sigma^2$, so that $\beta$ is the parameter of interest.

- $H_0 : \beta = 0$ is the null of interest, i.e. no treatment effect?
- Test statistic:

$$Z = \frac{\widehat{\beta}}{\text{s.e.}(\widehat{\beta})} \sim \text{normal}(0, 1) \text{ under } H_0$$

where $\widehat{\beta} = \overline{Y}_{HI} - \overline{Y}_{LO}$ and $\text{s.e.}(\widehat{\beta}) = \widehat{\sigma}/\sqrt{n}$.

# Treatment-by-Marker Interactions

- Now consider the situation in which we wish to examine the treatment effect by marker.
- To be concrete, define the subgroups relative to a recessive model with $S$ being the number of minor alleles.
- At a generic SNP: $S = 0$ corresponds to:

*No Copies of the Minor Allele*

while $S = 1$ corresponds to:

*One or Two Copies of the Minor Allele*.

# Treatment-by-Marker Interactions

There are $m$ comparisons of interest, with summary data at marker $j$, as below:

| | **Group** | | |
|---|---|---|---|
| | $S = 0$ | $S = 1$ | **Sample Size** |
| $T = 0$ | $\overline{Y}_{00}$ | $\overline{Y}_{01}$ | $n/2$ |
| $T = 1$ | $\overline{Y}_{10}$ | $\overline{Y}_{11}$ | $n/2$ |
| | $n - n_s$ | $n_s$ | $n$ |

Table 1: **Summary data at a generic marker, under two treatments $T = 0/1$; there are $n$ individuals in total, of which $n_s$ are in the group of interest.**

# Treatment-by-Marker Interactions

- $S_i = 0/1$ is a group indicator for individual $i$ at a generic SNP.
- For the treatment effect and at each marker we fit the model

$$Y_i = \alpha + \beta T_i + \gamma S_i + \underbrace{\Delta}_{\text{Interaction}} T_i \times S_i + \epsilon_i$$

  with $\text{var}(\epsilon_i) = \sigma^2$.
- $H_0 : \Delta = \Delta_0$ is the null of interest, i.e. is there a differential treatment effect of a certain size at the SNP, e.g. $\Delta_0 = 0$, to compare to the marginal treatment effect.
- Test statistic

$$Z = \frac{\widehat{\Delta} - \Delta_0}{\text{s.e.}(\widehat{\Delta})} \sim \text{normal}(0, 1) \text{ under } H_0.$$

- To emphasize, the same 833/837 responses are used in each of the $m$ comparisons, but they are distributed into the four treatment $\times$ marker cells differently.
- Key Observation: Standard errors will vary considerably across SNPs.

# VISP Example

- After data cleaning, there were $m = 803,122$ SNPs on which data were available, with at least 5 individuals in each treatment $\times$ marker subgroup.
- Suppose we are interested in detecting marker subgroups for which there is an enhanced effect, i.e. an increased reduction over the marginal treatment effect.
- Figure 8 shows the standard errors in the VISP trial – large variability and so the power ranges considerably also.
- Now refresh memory on the Bayesian approach to testing.

# Computation of Bayes Factors

- Recall that

$$\begin{aligned} \widehat{\Delta}|\Delta &\sim N(\Delta, V) \\ \Delta &\sim N(\Delta_0, W). \end{aligned}$$

where $\sqrt{V}$ is the standard error of the estimator leads to a simple form for the Bayes factor:

$$BF = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2}\frac{W}{V+W}\right)$$

where

$$Z = \frac{\widehat{\Delta} - \Delta_0}{\sqrt{V}}.$$

# Bayesian Boundaries

- We again use the Bayes factors as a mechanism by which $Z$-score boundaries can be calculated, as a function of the standard error $\sqrt{V}$.
- The Bayesian $Z^2$ score threshold is:

$$Z^2 > z_B^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log\left(\frac{PO}{R}\right)\right\}$$

  to give a threshold which is an explicit function of $V$, $R$ and $PO$.
- If the prior odds $PO$ on the null increases, threshold increases: require more evidence.
- If cost of Type II to Type I errors $R$ increase, threshold decreases: require less evidence.

# Bayesian Boundaries

- The Bayesian boundary:

$$Z^2 > z_B^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log\left(\frac{PO}{R}\right)\right\}.$$

- Beyond a certain point, as $V$ decreases the Type I error decreases to zero.
- Specifically, let $n$ denote an appropriate measure of sample size and $V = \sigma^2/n$. Then, as $n \to \infty$,

$$z_B^2 \to \underbrace{\log\left(1 + \frac{nW}{\sigma^2}\right)}_{\to\infty} + 2\log\left(\frac{PO}{R}\right).$$

- Relative to a fixed boundary:
  - For small $n$/large standard error the Bayesian approach requires more evidence because of the low power.
  - For large $n$/small standard error the Bayesian approach requires more evidence because of the high power and the comparison with the distribution under $H_A$.

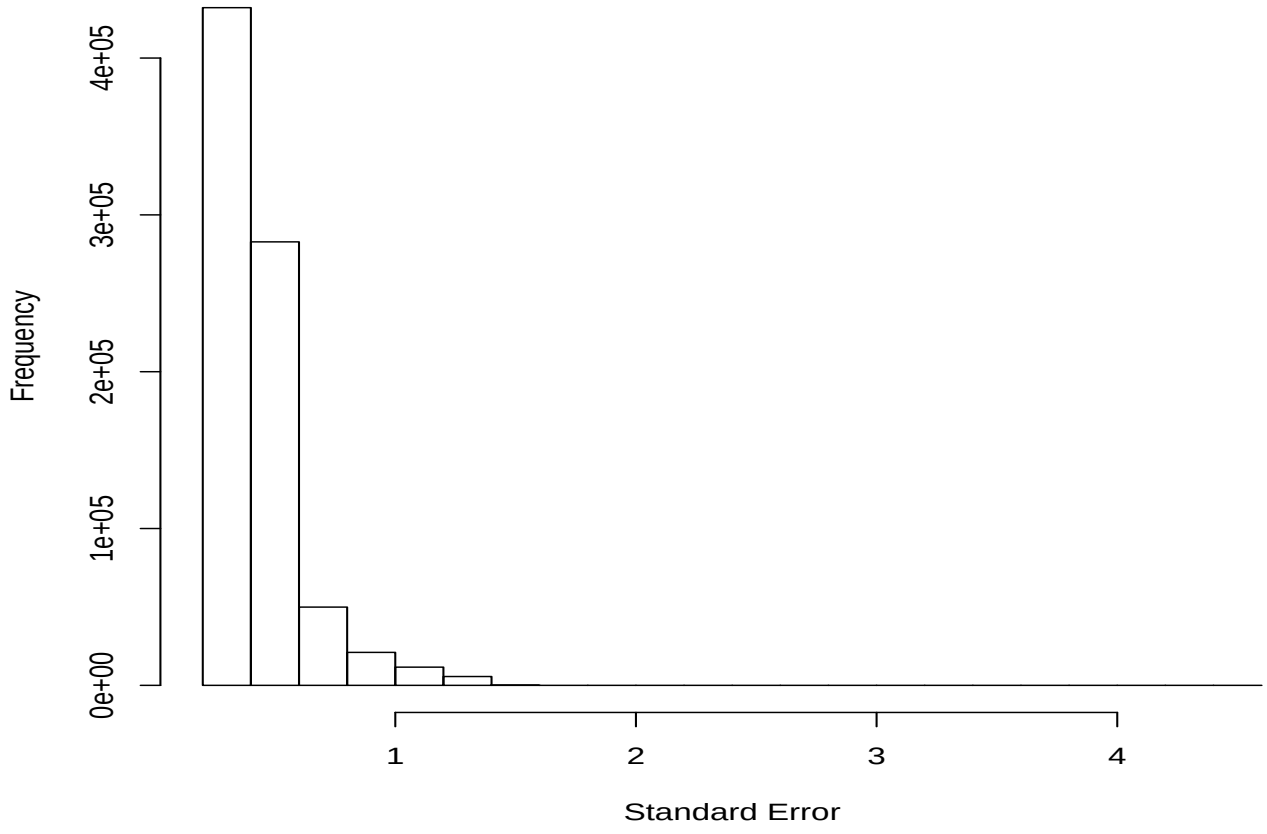



Figure 8: Histogram of standard errors of the interaction parameter estimates $\widehat{\Delta}$ in the VISP study.

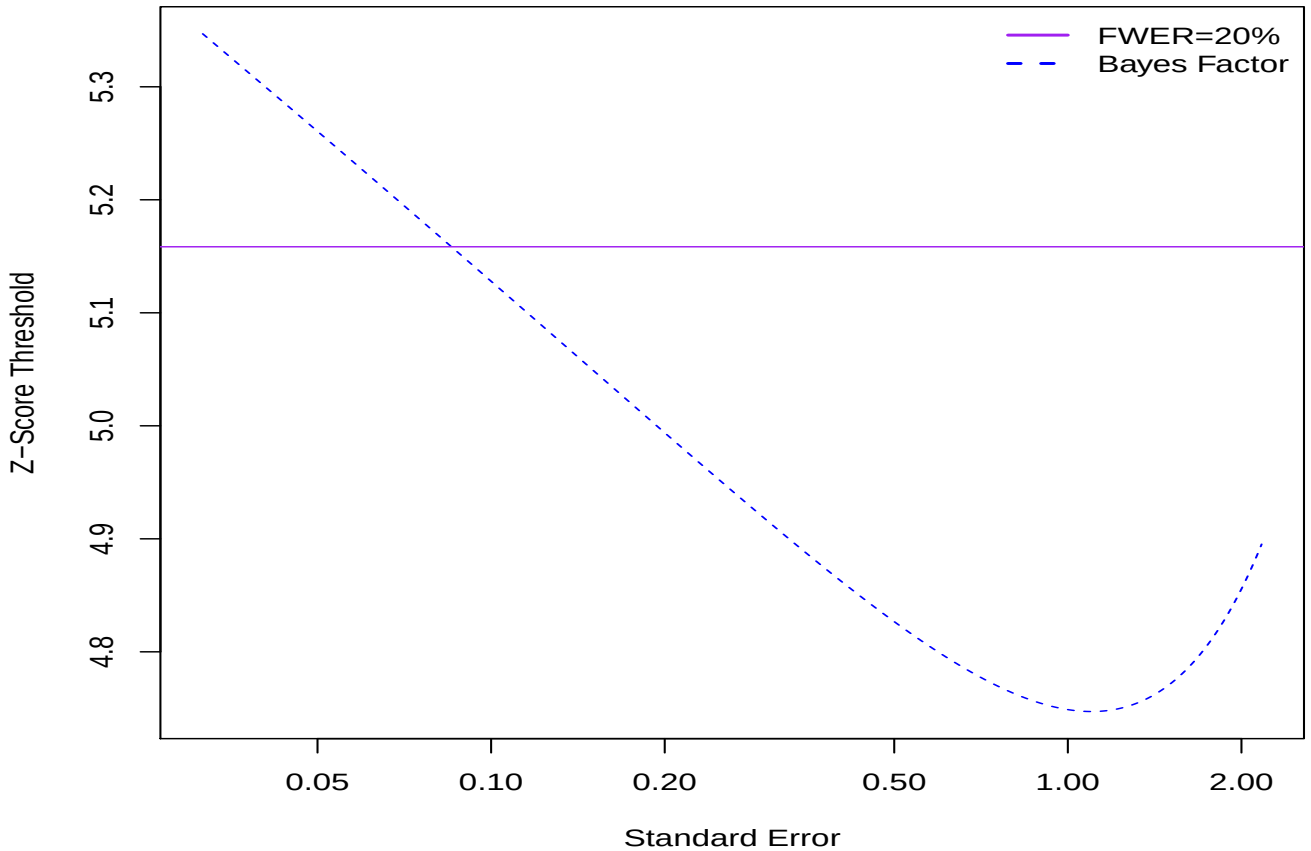



Figure 9: Bayesian $Z$-score threshold as a function of the standard error. The Bayesian threshold is based on a prior on the alternative of 0.0001, $R = 1$ and a prior standard deviation on the interaction effect size of $\sqrt{W} = 5.1$; this prior gives a 95% interval on $\Delta$ of (-10,10).

# A Priori Operating Characteristics

- Ranking is straightforward with Bayes factors, since the only choice is the prior on the effect parameter ($W$), and inference is relatively insensitive to this value.
- There is much greater sensitivity to the ratio of costs $R$ and the prior odds $PO$.
- Deciding upon values for $R$ and $PO$ is not straightforward, but only the ratio $PO/R$ is needed.
- We assume $R = 1$ (equal costs of type I and type II errors) and $\pi_1 = 0.001, 0.0001, 0.00001$.
- For $m = 803,122$ SNPs this corresponds to expecting 803, 80 and 8 non-null interactions, respectively.
- These signals will not reflect 803, 80, 8 different causal variants since typically multiple SNPs will tag each causal variant.
- Figure 10 plots various useful operating characteristics.

# Operating Characteristics

- To determine the EFD and ETD we require specification of the number of null and non-null signals, which we label as $m_0$ and $m_1$, respectively (so that $m = m_0 + m_1$).
- We take the true number of signals as $m_1 = 50$ so that there are $m_0 = 803,072$ null signals.
- Then,

$$\begin{aligned} \text{EFD} &= m_0 \times \alpha \\ \text{ETD} &= m_1 \times (1 - \beta) \end{aligned}$$

  where $\alpha$ and $\beta$ are the type I and type II errors [8].
- We emphasize that in a GWAS in which the fraction of non-null associations is close to zero, the ETD is highly sensitive to the choice of $m_1$ (in contrast to EFD, which is insensitive, because it depends on $m_0$)

[8] These can be worked under the Bayesian approach as well, and will vary as a function of the standard error

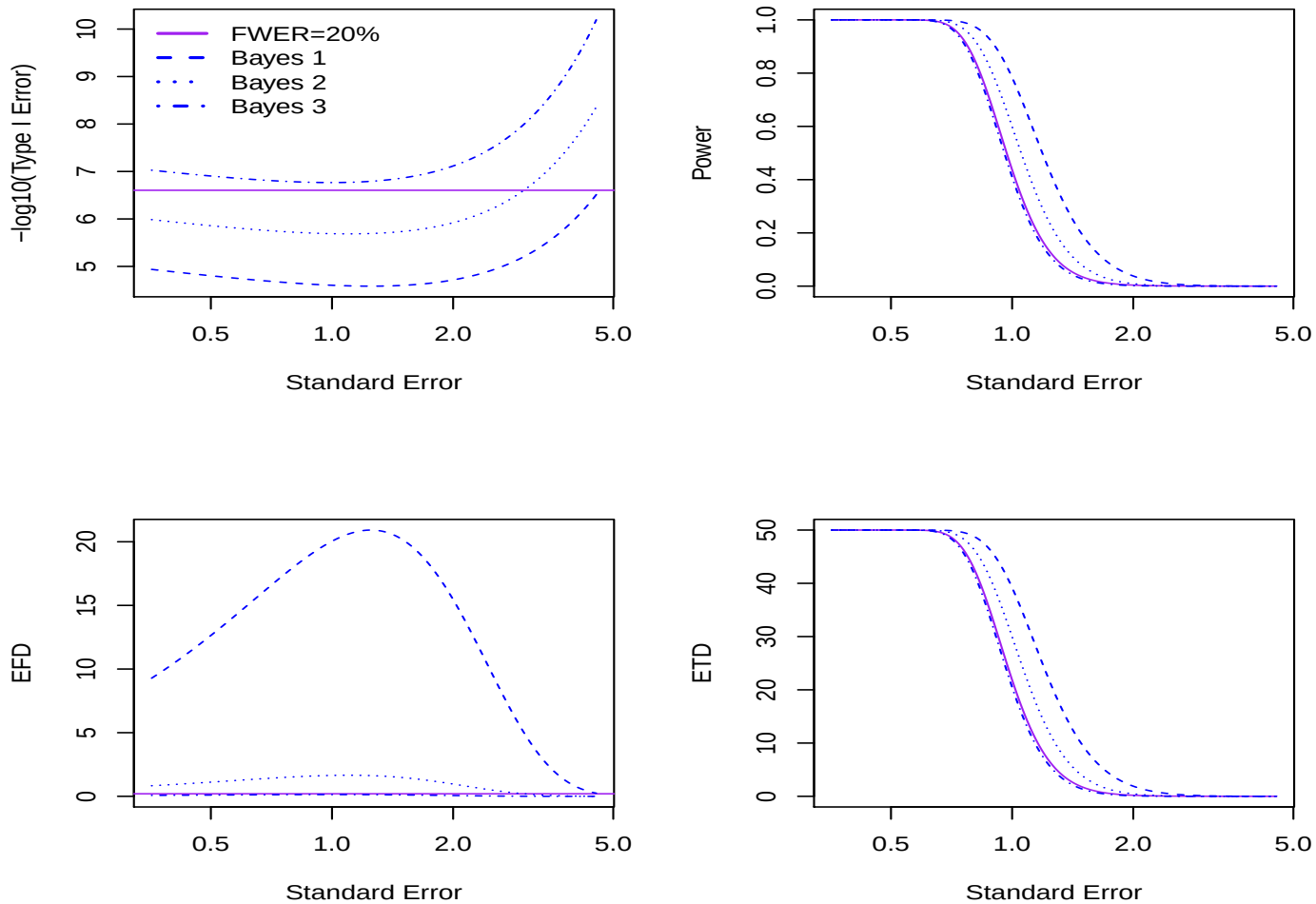



Figure 10: Operating characteristics of Bayes/Bonferroni. For Bayes boundaries R $= 1$ and “Bayes 1”, “Bayes 2”, “Bayes 3” correspond to priors of $\pi_1 = 0.001, 0.0001, 0.00001$. Power is to detect a drop of 5 units. For EFD/ETD we set $m_1 = 50$.

# Operating Characteristics

- The most liberal prior of $\pi_1 = 0.001$ produces a large number of type I errors (around 20 for standard errors in the mid-range) and might be judged to give unacceptably poor performance.
- The most sceptical prior is more conservative than Bonferroni (with a FWER of 20%) and the prior with $\pi_1 = 0.0001$ is a compromise for this choice of $m_1$.
- For example: For a standard error of 1, around 2 false discoveries would be expected (as in the lower left panel) but with around 10 more true signals being detected (as seen in the lower right panel), which seems a reasonable trade-off.
- Note, however, that if we think the number of true signals is smaller than $m_1 = 50$ then the number of true signals will fall proportionally.
- For example: At a standard error of 1, if $m_1 = 5$ then we would only expect to detect a single additional signal, when compared to the use of Bonferroni.
- Armed with this information we move to an analysis of the VISP data.

# Motivating Homocysteine Example

- We fitted the interaction model with adjustment for age and gender.
- The genetic subgroups are defined as having at least one copy of the minor allele as compared to two copies of the major allele.
- The number in the former subgroup ranges between 21 and 1,564 across SNPs.
- We choose $W$ to give a 95% prior interval for the interactions $\Delta$ of $\pm 10$.
- Figure 11 plots the $Z$-scores versus the standard error, along with boundary corresponding to a FWER of 20%.
- For both the most conservative prior and the Bonferroni approach (with a FWER of 20%, which gives a $p$-value threshold of $2.5 \times 10^{-7}$) two SNPs are flagged.
- With a FWER of 5% the Bonferroni threshold is $6.2 \times 10^{-8}$ and results in a single SNP being deemed significant.
- With the more optimistic prior of $\pi_1 = 0.0001$, a further signal is flagged (and these are not significant using Bonferroni).

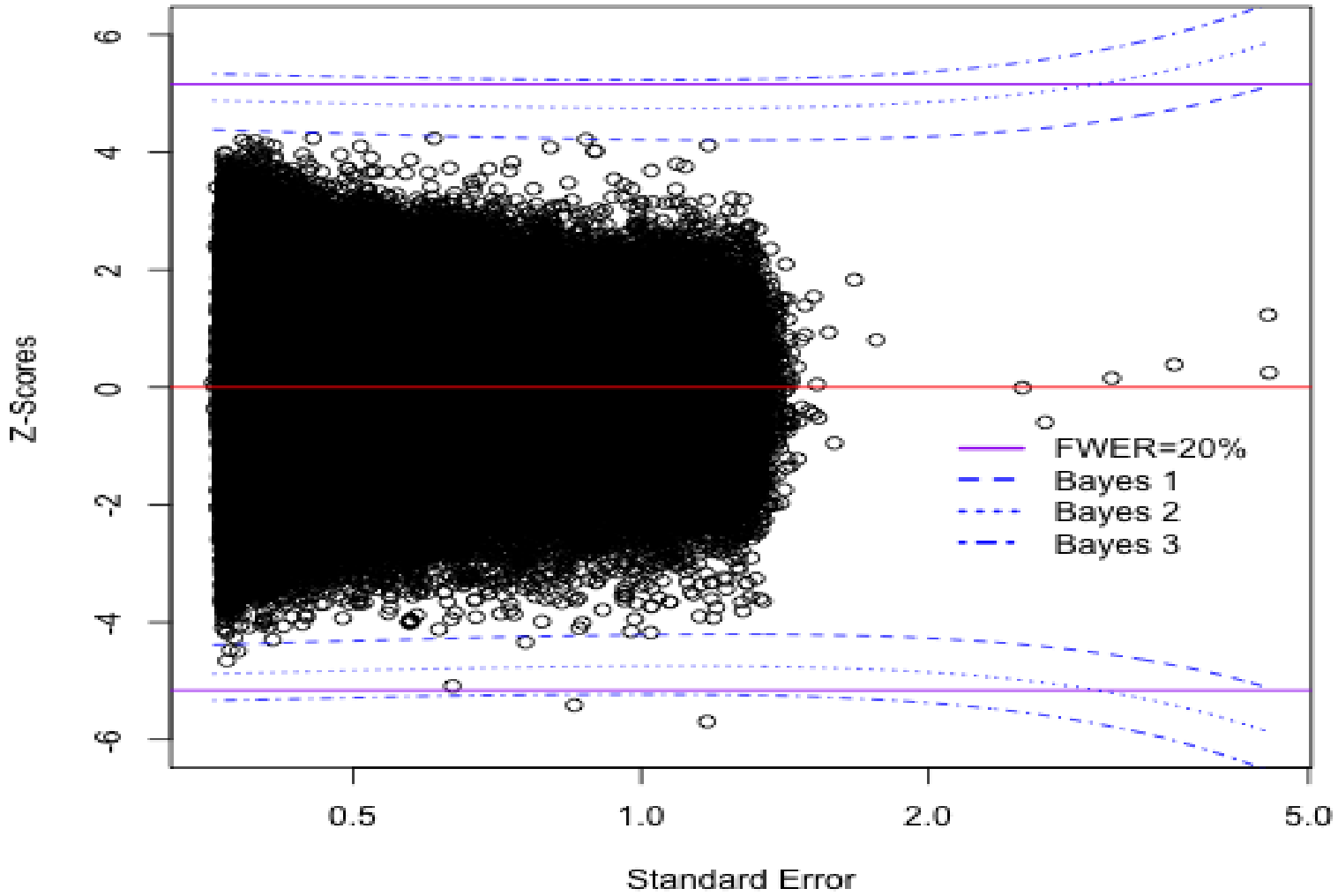



Figure 11: $Z$-score threshold as a function of the standard error for the VISP data, ratio of costs of type II to type I errors R $= 1$ and varying priors on the alternative of $\pi_1 = 0.001, 0.0001, 0.00001$ (to give Bayes 1, Bayes 2, Bayes 3 boundaries).

# Flagged Signals

| SNP ID | Chrom | $\widehat{\Delta}$ | $\widehat{\text{s.e.}}(\widehat{\Delta})$ | $p$-value | Bayes Factor | Post Prob$H_1$ of |
|---|---|---|---|---|---|---|
| rs3736238 | 17 | -6.68 | 1.38 | $1.5 \times 10^{-8}$ | $9.3 \times 10^{-7}$ | 0.99 |
| rs16893296 | 6 | -4.61 | 0.85 | $7.1 \times 10^{-8}$ | $3.9 \times 10^{-6}$ | 0.96 |
| rs1739317 | 6 | -3.23 | 0.64 | $4.0 \times 10^{-7}$ | $2.3 \times 10^{-5}$ | 0.81 |
| rs11819196 | 10 | -1.72 | 0.37 | $3.5 \times 10^{-6}$ | $2.9 \times 10^{-4}$ | 0.26 |

Table 2: The SNPs in the VISP study that had posterior probabilities on the alternative of greater than 0.25 ($R = 3$), with a prior on the alternative of $\pi_1 = 0.0001$ and under the equal variances recessive genetic model.

# VISP Results

- Figure 12 plots the posterior probabilities of the alternative hypothesis (with $\pi_1 = 0.0001$) versus chromosomal position (this is similar to a Manhattan plot in which $-\log_{10}$ p-values are plotted against position).
- The 3 SNPs that fall outside of the boundary in Figure 11 are highlighted.
- The strongest signal is for SNP rs3736238 on chromosome 17. For this SNP there are 42 individuals in the $M = 1$ subgroup, of which 24 and 18 are in the low and high dose groups, respectively.
- The probability of this signal being a false discovery is 0.01 *under* our assumed prior.

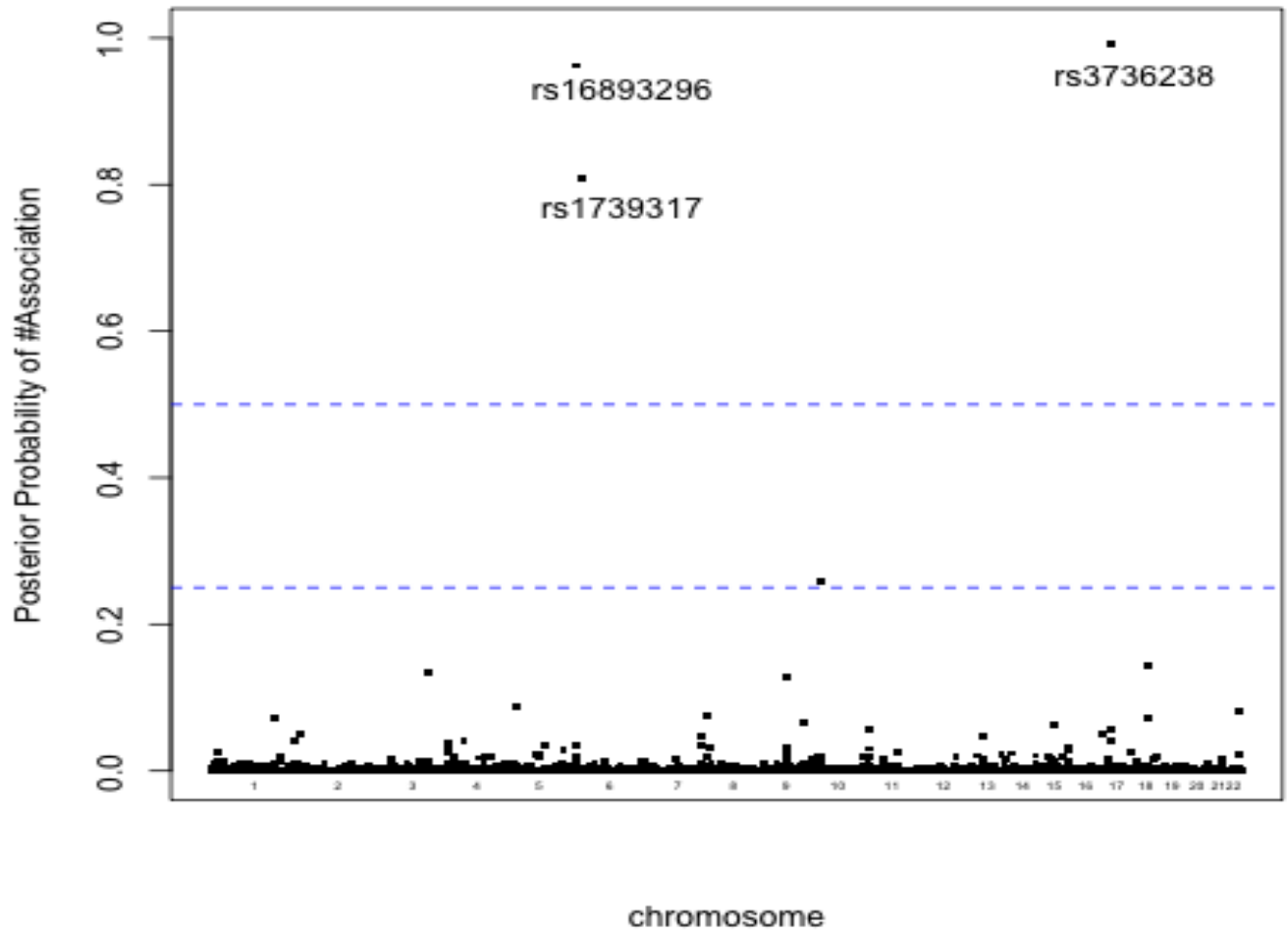



Figure 12: Posterior probability on the alternative plotted versus genomic position for the VISP data. The prior on the alternative is $\pi_1 = 0.0001$.

# VISP Example

- Figure 13 shows that the *p*-values and Bayes factors differ in their rankings due to the differing sample sizes/standard errors.
- The points are color-coded by the size of the standard error and we see that the points with larger standard errors are consistently ranked as giving greater evidence for the alternative under the Bayesian approach.
- This behavior occurs here because of the association between the $Z^2$ boundary and the standard error for these priors, as shown in Figure 9.
- Specifically, the majority of the signals occur in that portion of the latter curve in which the Bayes boundary lies below the FWER boundary.
- Figure 14 shows an example in which distinctly different behavior occurs.

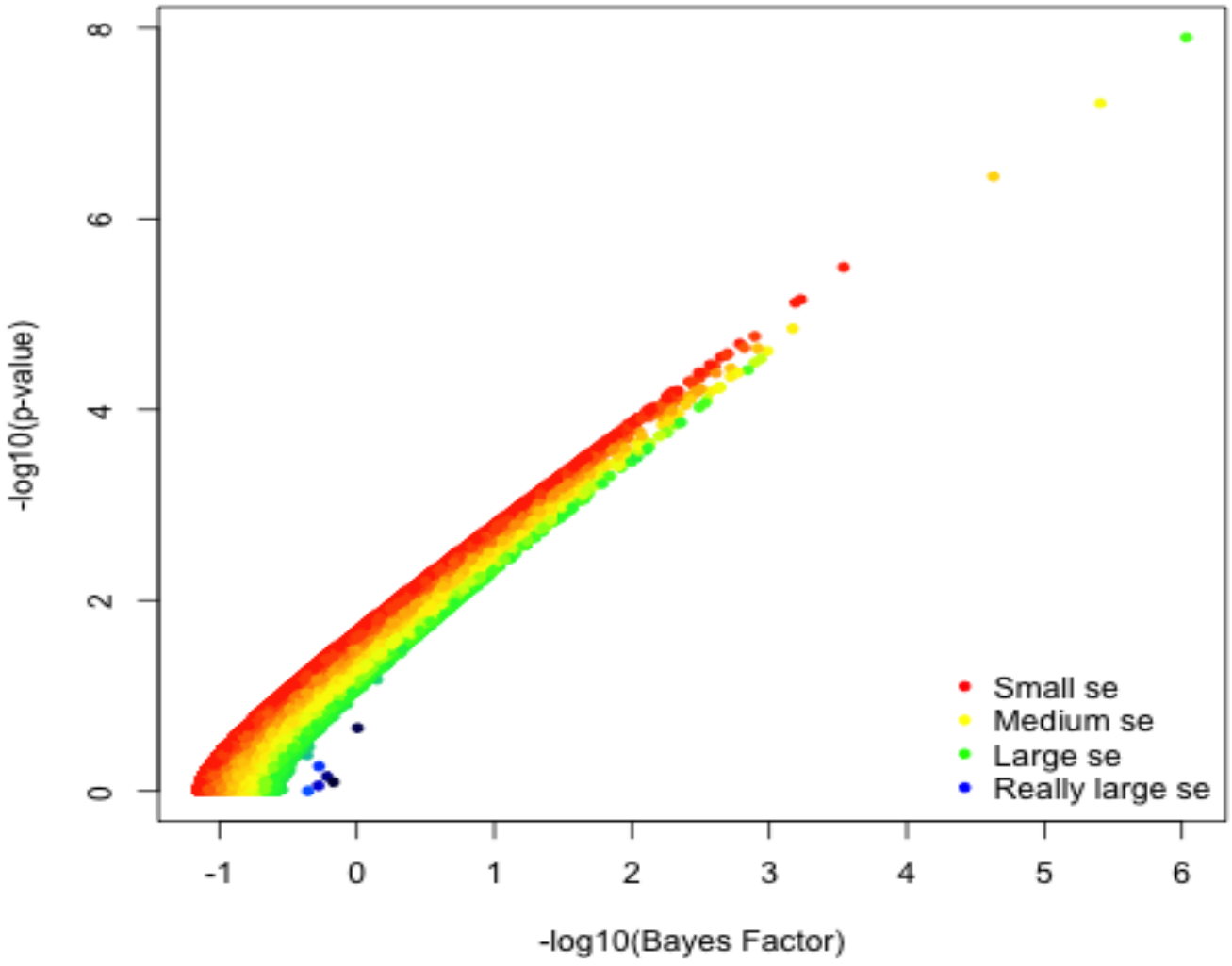



Figure 13: -$\log_{10}$BFs vs -$\log_{10}$ $p$-values, color-coded by standard error with $W = 10$.

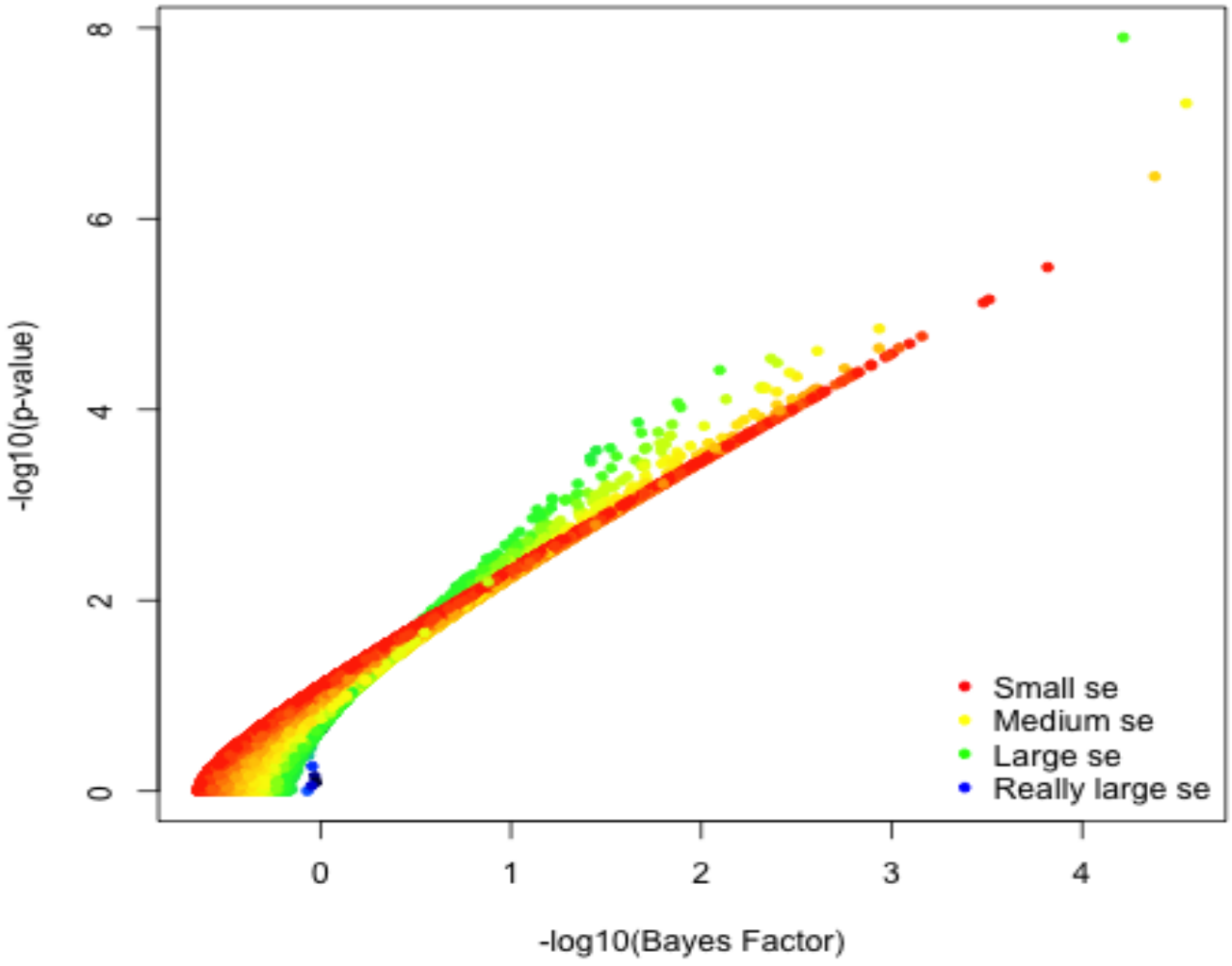



Figure 14: -$\log_{10}$BFs vs -$\log_{10}$ $p$-values, color-coded by standard error with $W = 3$.

# VISP Example

- A related interesting exercise is to simulate the distribution of observed effect sizes under our assumed priors (on both the proportion of non-null signals and the effect sizes), using the observed distribution of standard errors.
- The distribution of effect sizes is $N(\Delta, V + W)$ for the non-null signals and normal$(0, V)$ for the null signals.
- We can then evaluate the power, and hence determine the number of signals we would expect to detect given our prior assumptions.

# VISP Example

- For the VISP data, with a proportion of non-null signals $\pi_1 = 0.0001$, R= 1 and 95% range for the effect sizes of $\pm 10$, we would expect to see 52 true positives and one false positive.
- Given we only observed three non-null signals, this implies that either the range of effect sizes (as defined through $W$) was too wide or, more probably, that our estimate of $\pi_1$ was optimistic.
- Repeating this exercise with $\pi_1 = 0.00001$ gives 5 true positives and close to 0 false positives, which is more consistent with that which was observed.
- Figure 15 gives the posterior probabilities for this prior.

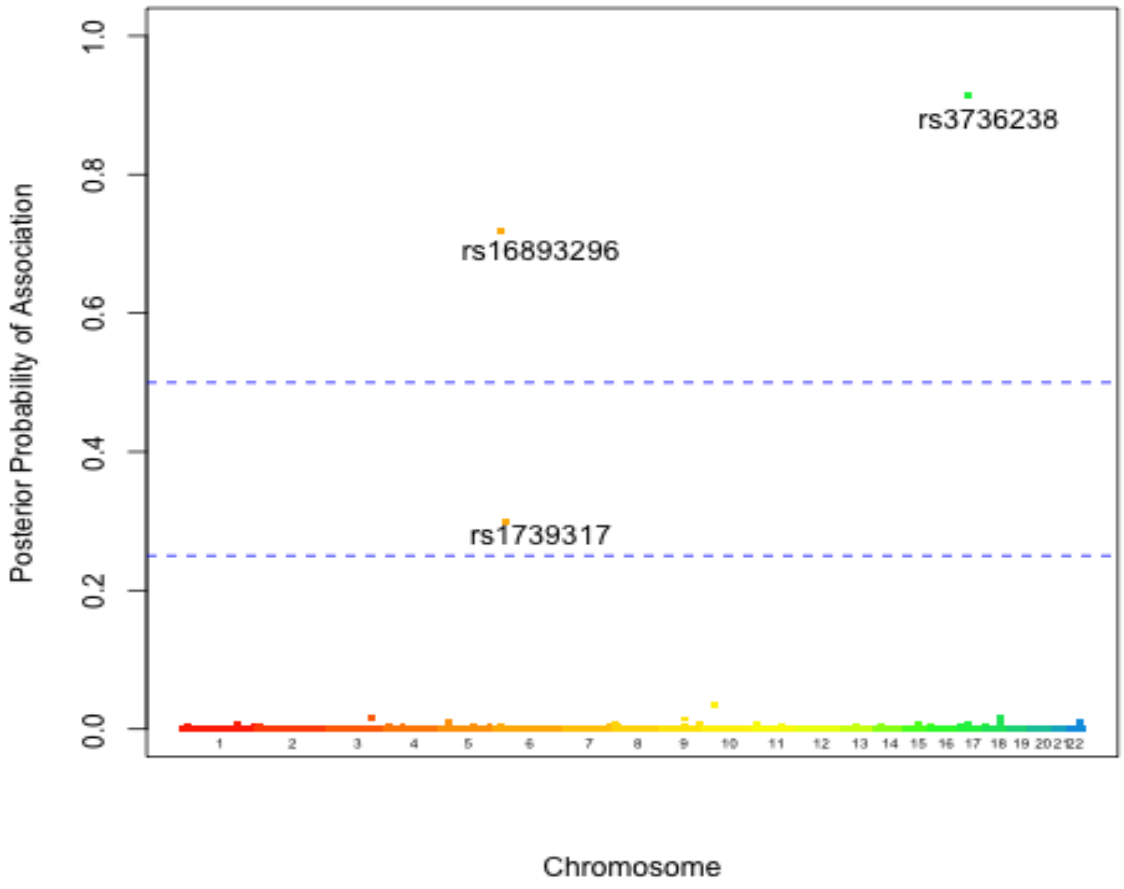



Figure 15: Posterior probability on the alternative plotted versus genomic region for the VISP data. The prior on the alternative is the more conservative choice of $\pi_1 = 0.00001$.

# VISP Discussion

- We chose the value $\pi_1 = 0.0001$ by examining frequentist summaries before the real data analysis was performed.
- We define $\pi_1$ as the proportion of SNPs that would be associated with the disease, if the power were 1.
- After the data are analyzed we can, for those SNPs declared as null (i.e. all but 3 SNPs in the VISP trial), sum up the posterior probabilities of being non-null, and this gives the expected number of false non-discoveries .
- For the VISP data, this expected number is 24.6 so that we are missing a large number of signals, with lack of power being the major issue.

# VISP Discussion

- For the three significant signals, at the 0.5 threshold, the probabilities of the null being true are 0.01, 0.04 and 0.19, so that the expected number of false discoveries is 0.24.
- Taking the threshold of significance as 0.25 gives an additional SNP as being declared significant.
- The sum of the posterior probabilities of the null is 0.98 in this case and so, under this prior, we would expect one of the reports signals to be a false discovery.

# Sensitivity to $\pi_1$

- The posterior probability of the alternative is highly dependent on the choice of prior on the null $\pi_0$, and a sensitivity analysis is always warranted.
- Ideally, rather than fix $\pi_0$ as we have done, one would estimate of $\pi_0$ from the totality of data (i.e. over all $m$ SNPs), but this is difficult because in a GWAS the proportion of detectable null signals is typically very close to 1; there may be many thousands of small but non-zero effects, but the power to detect these signals is low, with the usual sample sizes.

# Sensitivity to $\pi_1$

- In other contexts, such as the analysis of gene expression data (Storey and Tibshirani, 2003), the data can be used to estimate $\pi_0$ more reliably.
- If the same prior on the null is used for all the tests, the rankings based on the Bayes factor will remain the same as the ranking based on posterior probabilities.
- However, calibrating the Bayes factors to the probability scale requires prior probabilities.
- Within a sensitivity exercise one may include an analysis in which any available information on particular SNPs may be included.

# An Alternative Approach to Significance

- The posterior probability (and the $Z$-score threshold) is equally sensitive to R as to $\pi_1$.
- The form of the latter suggests that all we need to do is to fix PO/R.
- As mentioned above, in the VISP analysis we selected $\pi_1$ by examining the frequentist operating characteristics.
- An alternative method (Wakefield, 2012) for obtaining PO/R is to specify a value for the $Z^2$ boundary, $z_B^2$, at a particular $V$ (for example, at a MAF and sample size that one is familiar with) and then solve for $U = \log(\text{PO}/\text{R})$ via

$$\widehat{U} = \frac{z_B^2 \times W}{2(V+W)} - \frac{1}{2}\log\left(\frac{V+W}{V}\right).$$

- With this value of $\widehat{U} = \text{PO/R}$ one can then proceed to use

$$Z^2 > z_B^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log(\widehat{U})\right\}$$

across the observed range of standard errors.

# Incorporating Prior Information

- We now briefly describe a method for assigning priors to SNPs based on substantive information.
- In collaboration with scientists at IARC and at the Department of Computer Science at Sheffield University a method had been developed that searches through PubMed abstracts for pre-assigned keywords and key concepts.
- More details in Johansson *et al.* (2012).
- This information is used to assign prior probabilities of association with the phenotype for each SNP of interest.

# Incorporating Prior Information

- Three prior groups were assigned, depending on the number of hits.
- The priors can subsequently be incorporated with the association results of GWAS using the previously described Bayesian framework.
- The method has acronym: Adjusting Association Priors with Text (AdAPT).
- Details of the method can be found in Johansson *et al.* (2012).
- The `AdAPT` software is available here:
  http://services.gate.ac.uk/lld/gwas/service/config

# Incorporating Prior Information in a GWAS

- SNPs are assigned to a group, based on the number of keywords that were found to be associated with this SNP.
- For the priors, keywords were ranked by priority: In the:
  - 1st group $G_1$: were words specific to lung cancer (eg, smoking, lung carcinoma).
  - 2nd group $G_2$: were more general words specifcally relevant to lung cancer (smoking, nicotine, non-small cell carcinoma),
  - 3rd group $G_3$: were more general words (carcinogen, DNA damage).
- Each SNP was then placed in one of three prior categories:
  1. $C_1 = \{\text{not } G_1, \text{not } G_2, \text{not } G_3\}$.
  2. $C_2 = \{$ at least one of $G_1, G_2, G_3$ but not all $\}$.
  3. $C_3 = \{G_1, G_2, G_3\}$.

# Incorporating Prior Information in a GWAS

- We then assigned prior odds (PO) to $\Pr(H_0|C_j)/\Pr(H_1|C_j)$. Specifically for the three categories, the PO was set to 7874 ($C_1$), 899 ($C_2$) and 224 ($C_3$).
- These were used in the analysis to obtain the posterior odds on $H_0$:

$$\frac{\Pr(H_0|y, C_j)}{\Pr(H_1|y, C_j)} = \frac{\Pr(y|H_0)}{\Pr(y|H_1)} \times \frac{\Pr(H_0|C_j)}{\Pr(H_1|C_j)}.$$

# Incorporating Prior Information

- First, the power was evaluated for the three categories, see Figure 16.
- The method was tested by comparing rankings of known susceptibility alleles in a previous lung cancer GWAS of 1989 cases and 2625 controls in 6 central European countries.
- The rankings of 6 SNPs that have been independently replicated in multiple studies were calculated.

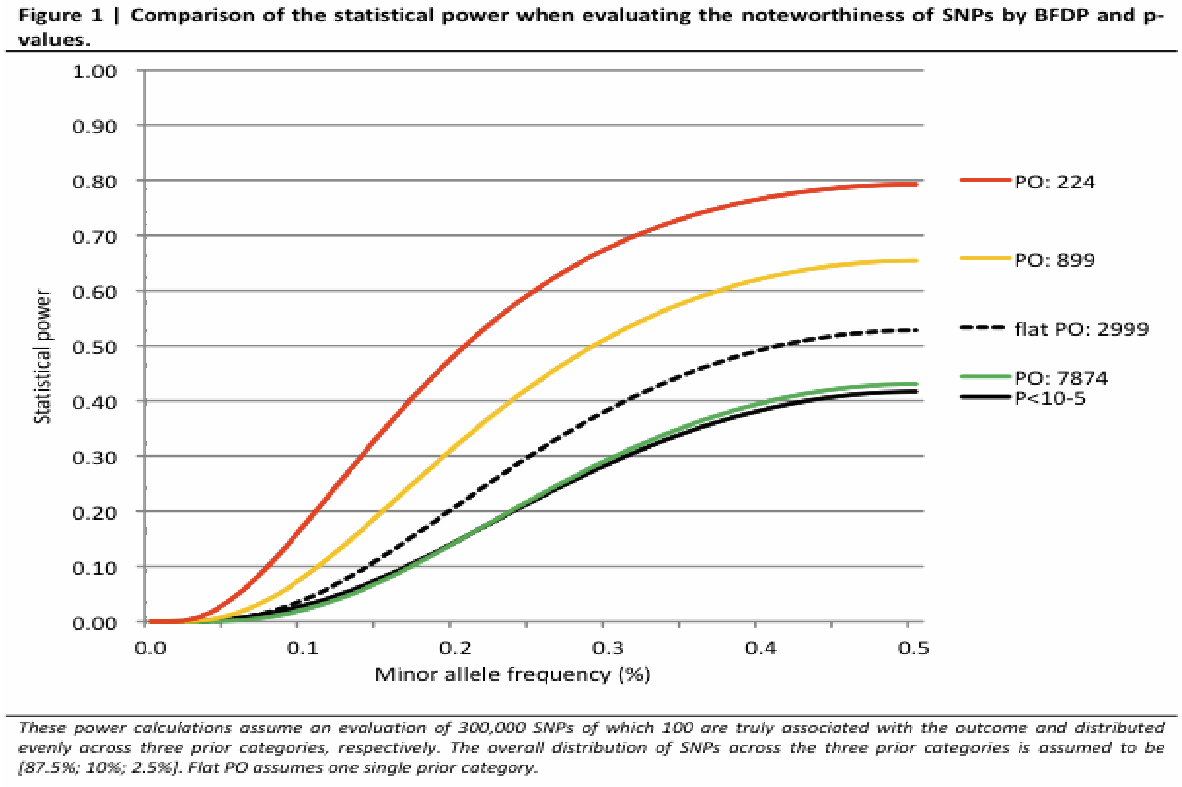



Figure 16: Power as a function of MAF, for three prior categories, for a single prior, and for a *p*-value approach.

# Incorporating Prior Information: Proof of Principle Results

- The results below show that known susceptibility SNPs were ranked more highly by AdAPT BFDPs than by p-values.
- Rankings based on initial data with informative priors for the Bayes rankings:

| SNP | *p*-value ranking | Bayes ranking |
|---|---|---|
| rs8034191 | 1 | 1 |
| rs1051730 | 2 | 2 |
| rs4324798 | 4 | 5 |
| rs401681 | 73 | 30 |
| rs2736100 | 76 | 32 |
| rs3117582 | 121 | 34 |

# Incorporating Prior Information: New Study Results

- Subsequently, the method was applied on a novel two phase GWAS of oral cancer, with 791 cases and 7,012 controls included in the discovery phase.
- A Bayes threshold on the null of 0.8 was assigned and 6 SNPs passed this test.
- One of these was already replicated, the replication was carried out for the remaining 5 AdAPT ranked SNPs in 1,046 cases and 2,131 controls from 4 case-control studies.
- rs991316, located in the ADH gene region of 4q23, displayed a statistically significant association with oral cancer risk in the replication phase (per-rare-allele log additive *p*-value $=2.5 \times 10^{-3}$).
- This SNP was ranked 76th in the *p*-value list and so would not have been selected to carry forward, but was ranked 4th in the BFDP list.
- The combined odds ratio associated with having one additional rare allele was 0.84 (95% CI: 0.75–0.94).

# Conclusions

- Bayesian analysis is attractive in a multiple testing context, but the results are very sensitive to the prior on the proportion of nulls, $\pi_0$.
- Fast methods are required for large $m$ (e.g. in a GWAS context) of tests, which is still a drawback for many Bayesian approaches.
- Priors can be made a function of characteristics of the SNP (e.g. non-synonymous, previously implicated,...). See Johansson *et al.* (2012) for an example.
- 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.
- Investigators have a vested interest in doing as little as possible about the problem – but the good ones recognize this, and try to produce good (but not *perfect*) science.
- 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.

Daly, A. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics*, **11**, 214–246.

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

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.

Johansson, M., Roberts, A., Chen, D., Li, Y., Delahaye-Sourdeix, M., Aswani, N., Greenwood, M., Benhamou, S., Lagiou, P., Holcáová, I., Richiardi, L., Kjaerheim, K., Agudo, A., Castellsagué, X., Macfarlane, T., Barzan, L., Canova, C., Thakker, N. S., ans A. Znaor, D. C., Healy, C., Ahrens, W., Zaridze, D.,

Szeszenia-Dabrowska, N., Lissowska, J., Fabiánová, E., Mates, I., Bencko, V., Foretova, L., Janout, V., Curado, M., Koifman, S., Menezes, A., W unsch-Filho, V., Eluf-Neto, J., Boffetta, P., Franceschi, S., Herrero, R., Garrote, L., Talamini, R., Boccia, S., Galan, P., Vatten, L., Thomson, P., Zelenika, D., Lathrop, M., Byrnes, G., Cunningham, H., Brennan, P., Wakefield, J., and Mckay, J. (2012). Using prior information from the medical literature in GWAS of oral cancer identifies novel susceptibility variant on chromosome 4 – the AdAPT method. *PLoS One*, **7**.

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.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, **4**, 422–438.

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.

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

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

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*, **33**, 79–86.

Wakefield, J. (2012). Commentary: Genome-wide significance thresholds via bayes factors. *International Journal of Epidemiology*, **42**, 286–291.

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

Wakefield, J., Skrivankova, V., Hsu, F.-C., Sale, M., and Heagerty, P. (2014). Detecting signals in pharmacogenomic studies. *The Pharmacogenomics Journal*. Published online, January 2014.

Westfall, P., Johnson, W., and Utts, J. (1995). A Bayesian perspective on the bonferroni adjustment. *Biometrika*, **82**, 419–427.