2016 SISG MODULE 17: Bayesian Statistics for Genetics
Lecture 7: Bayesian and Frequentist Multiple Testing

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction

- With the advent of high-throughput technologies in genomics there is now the possibility of carrying out millions of tests, and so the implications of such multiple testing must be carefully considered.
- In this lecture we will review the rationale for p-values.
- We then explore the connection between p-values and Bayes factors.
- Multiple testing will be reviewed and a Bayesian perspective presented.
- An example in the context of a pharmacogenomics GWAS will be presented.
- The use of substantive prior information will also be demonstrated.

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.
- 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 μmol/L in the low dose group versus -2.36 μmol/L in the high dose group, i.e., a difference of -1.99 μmol/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 number in this subgroup ranges between 21 and 1564 across SNPs.

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).

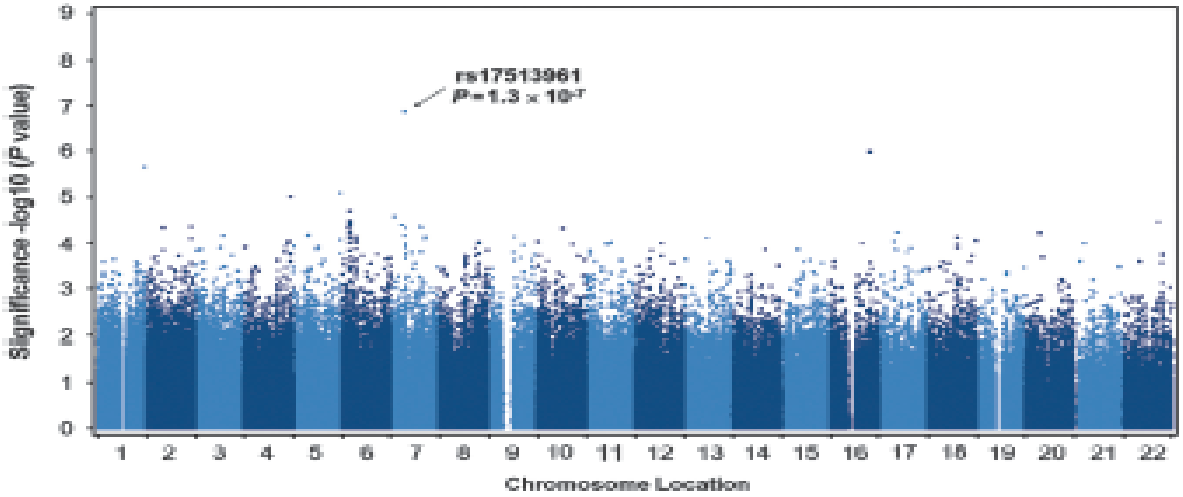


Figure 1 Representative Manhattan plot from GWAS in the entire cohort for genotype BB vs not BB. All possible combinations of genotypes representing a dominant, heterozygous and recessive inheritance were evaluated. Three pairs of comparisons AA vs not AA, AB (heterozygous) vs not AB and BB vs not BB 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

- 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.
- 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_{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 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, $Y_1, ..., Y_n$ of size n from normal$(\theta, 1)$,

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

  Obvious that we should reject H_0 for $\overline{Y} > k$, a constant.

- 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
1	0.01	0.91	6.21
25	0.01	0.0038	0.42
100	0.01	8×10^{-15}	2.5×10^{-12}

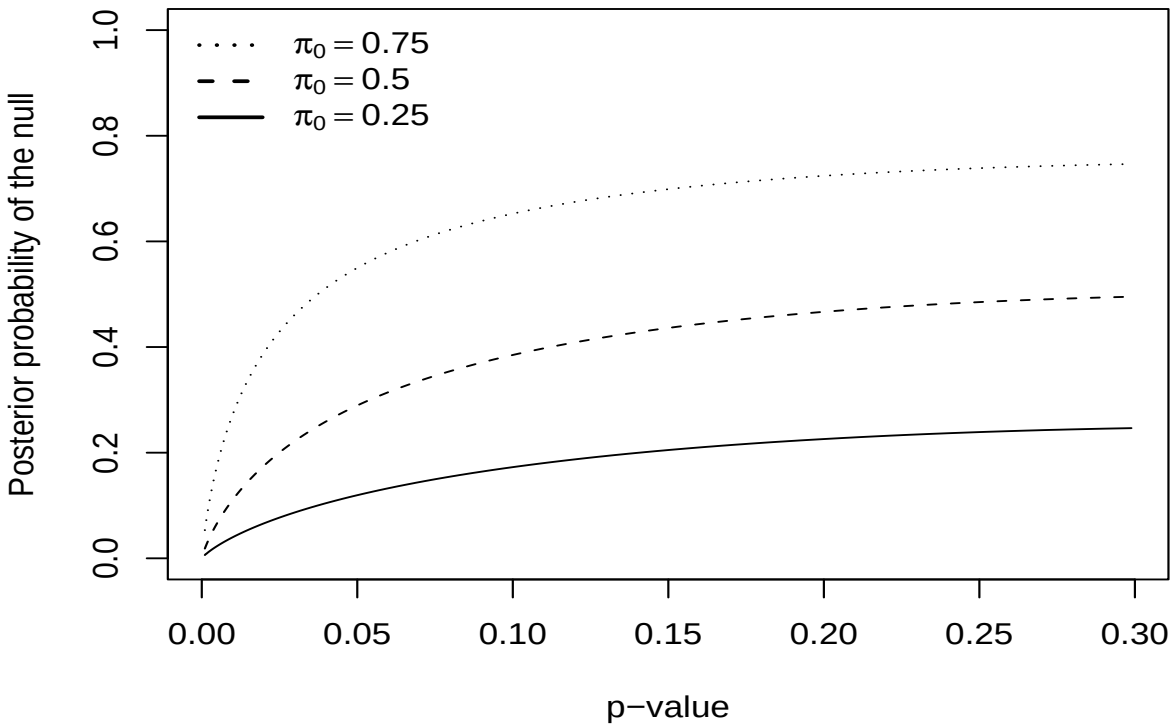
- Also: Statistical versus practical significance.
- For both p-values and α levels we need thresholds that decrease as a function of the sample size n. Pearson (1953, p. 68), "...the quite legitimate device of reducing α as n increases".

Calibrating *p*-values

- With $\pi_0 = \Pr(H_0)$, Sellke, Bayarri and Berger (2001) show that:

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

- 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 $\Pr(H_0|$ data $) > 0.29$.
  - p-value $= 0.01$ gives $\Pr(H_0|$ 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 $\Pr(H_0|$ 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,

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

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

A Sanity Check via a Simple Example

- The model:

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

  $i = 1, ..., n$.

- The distribution of the MLE is:

$$\widehat{\theta} = \overline{Y} \sim \text{normal}(\theta, V)$$

  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 $\theta \sim \text{normal}(0, W)$.
- Recall from Lectures 3 and 4, that the evidence in the data for a pair of hypotheses is summarized in the Bayes factor.
- The Bayes factor is

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

  where dnorm is shorthand for the density of a normal random variable.
- 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 R this gives the decision rule to reject H_0:

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

- 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

Interesting question: When do Bayes and frequentist *p*-value inference coincide?

- Consider an approximate Bayes approach. We have parameter of interest θ with
  - Data: MLE $\widehat{\theta}$ and standard error $\sqrt{V}$ to give likelihood $\widehat{\theta}|\theta \sim \text{normal}(\theta, V)$,
  - Prior: $\theta \sim \text{normal}(0, W)$.
- The null and alternative hypotheses of interest are

$$H_0 : \theta = 0 \qquad H_1 : \theta \neq 0.$$

- This leads to the approximate Bayes factor (ABF)

$$\begin{aligned} \text{ABF} &= \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right) \\ &= \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2}\frac{W}{V+W}\right) \end{aligned}$$

  where $r = W/(V+W)$ and $Z = \widehat{\theta}/\sqrt{V}$.
- Here we write explicitly as a function of Z, with $T = Z^2$.

Calibration with *p*-values

- $\text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2}\frac{W}{V+W}\right)$, which depends on n, because V does.
- Recall we reject if ABF $\times$ PO $>$ threshold R.
- We are trying to find a Bayes factor that does not depend on n, to correspond with a p-value rule.
- We can reverse engineer a version of ABF that does not depend on n by taking the prior variance $W = K \times V$, where K is a constant.
- Then we have approximate Bayes factor

$$\text{ABF}^p = \sqrt{1+K} \exp\left(-\frac{Z^2}{2}\frac{K}{1+K}\right).$$

- Important point: No dependence on n, i.e. it depends on Z only, and therefore on the p-value only.
- If we use the above prior and Bayes factor in multiple tests, then the rankings of p-values and ABF^p will be identical. The problem is that this prior makes no sense.

Calibration with *p*-values

- The ABF with W not depending on n is consistent (you get the right answer with a lot of data), 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 π_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

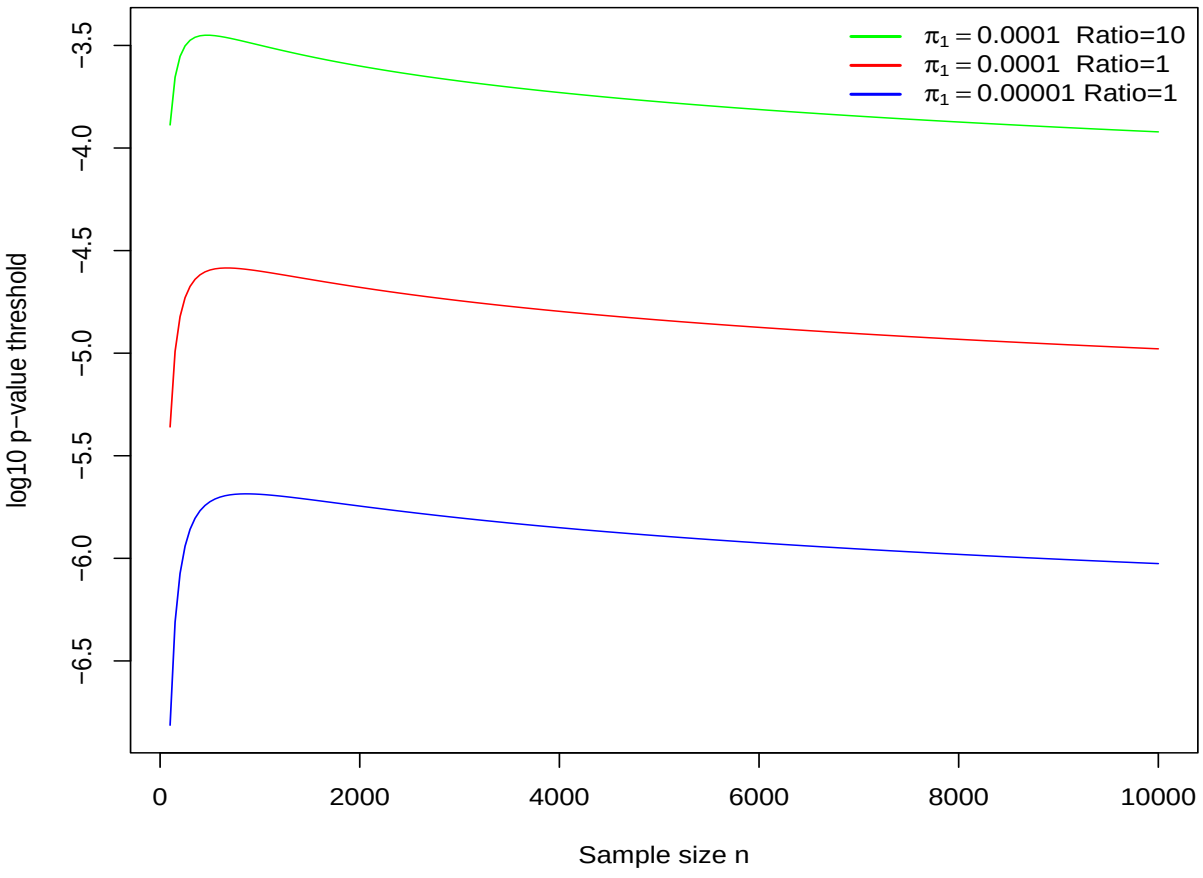


Figure 1 : Threshold for rejection, on the $\log_{10}(p)$-value scale, versus sample size. Notice how the threshold is decreasing with increasing sample size.

Multiple Testing

The problem: m tests being carried out, often (in a GWAS context, for example) with a tiny probability of any association being non-null.

We describe:

- Family-wise error: Bonferroni and Sidak.
- Control of expected number of false discoveries.
- Control of false discovery rate.
- A Bayesian perspective.

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 $m = 500K$ 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 α should depend on n, in partcular should decrease as n gets larger and larger).

Sidak Correction

- If all the tests are independent:

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

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

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

Expected Number of False Discoveries

- We describe an alternative criterion.
- For $i = 1, ..., m$ 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

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

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 α/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 $P_{(1)} < ... < P_{(m)}$ 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 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

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

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.

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

Simulated Example

- We begin by plotting, in Figure 2 the oberved p-values versus those expected under the null, i.e. $i/(m+1)$ for $i = 1, ..., m$.
- Hard to interpret, so we truncate the scales in Figure 3.
- Finally we stretch the scale in Figure 4 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 ($p = 0.05/100 = 5 \times 10^{-4}$, or $-\log_{10}(p) = 3.30$) and only flags one test as being significant.
- The EFD=1 gives a p-value threshold of 0.01, or $-\log_{10} p = 2$ and picks up all 5 signals.
- The FDR control at 5% gives the green diagonal line and rejects 3 tests.

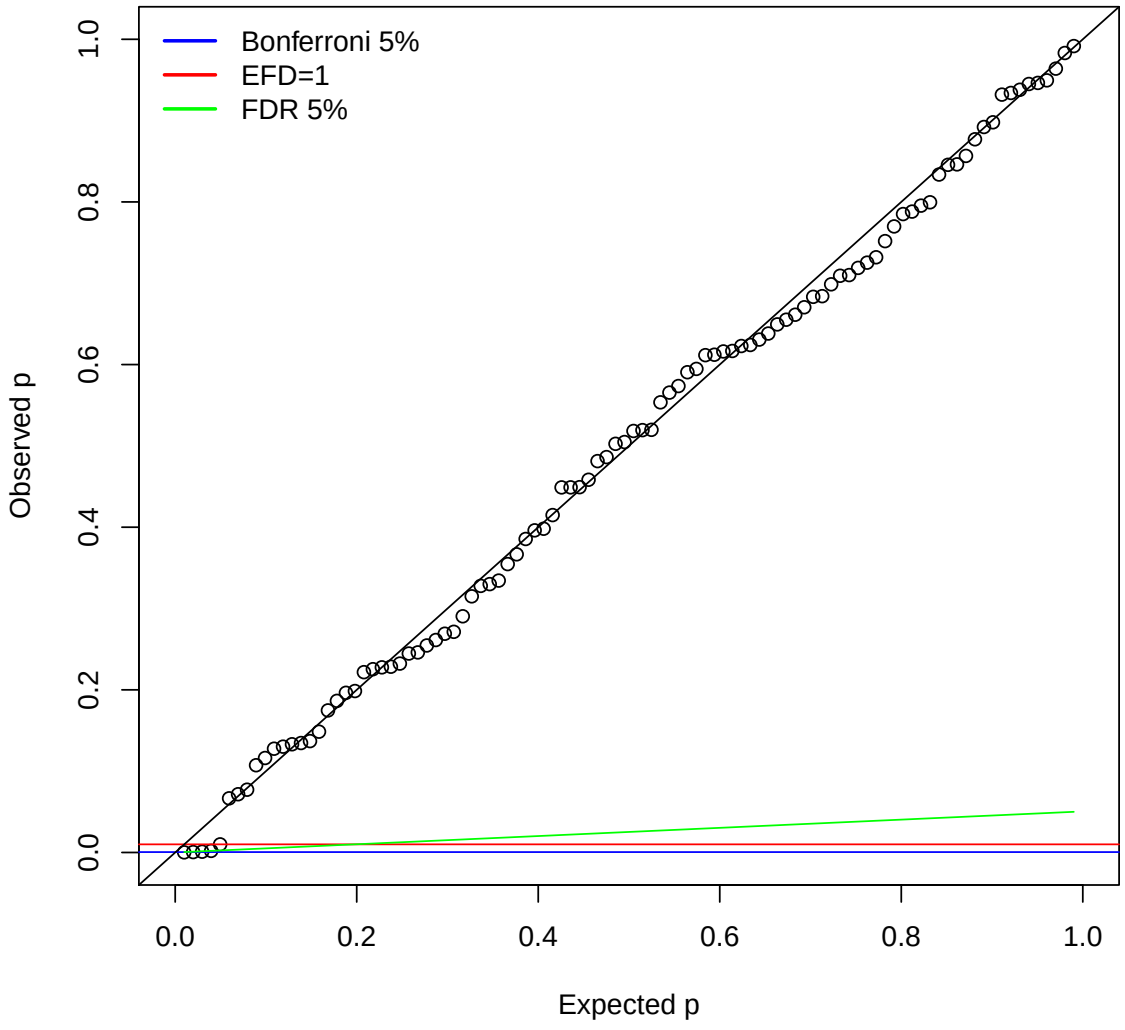


Figure 2 : Observed versus expected *p*-values.

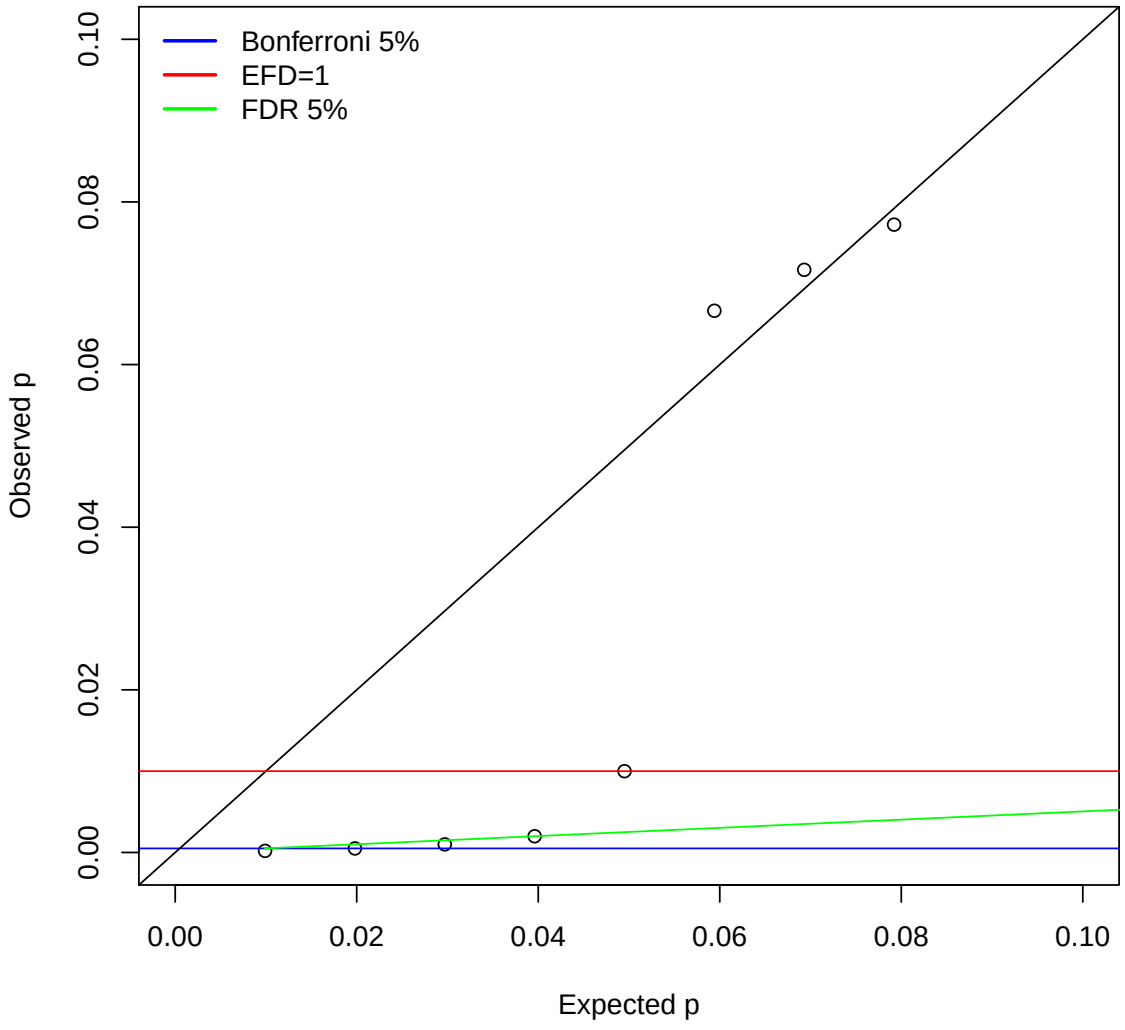


Figure 3 : Observed versus expected *p*-values with truncated scale.

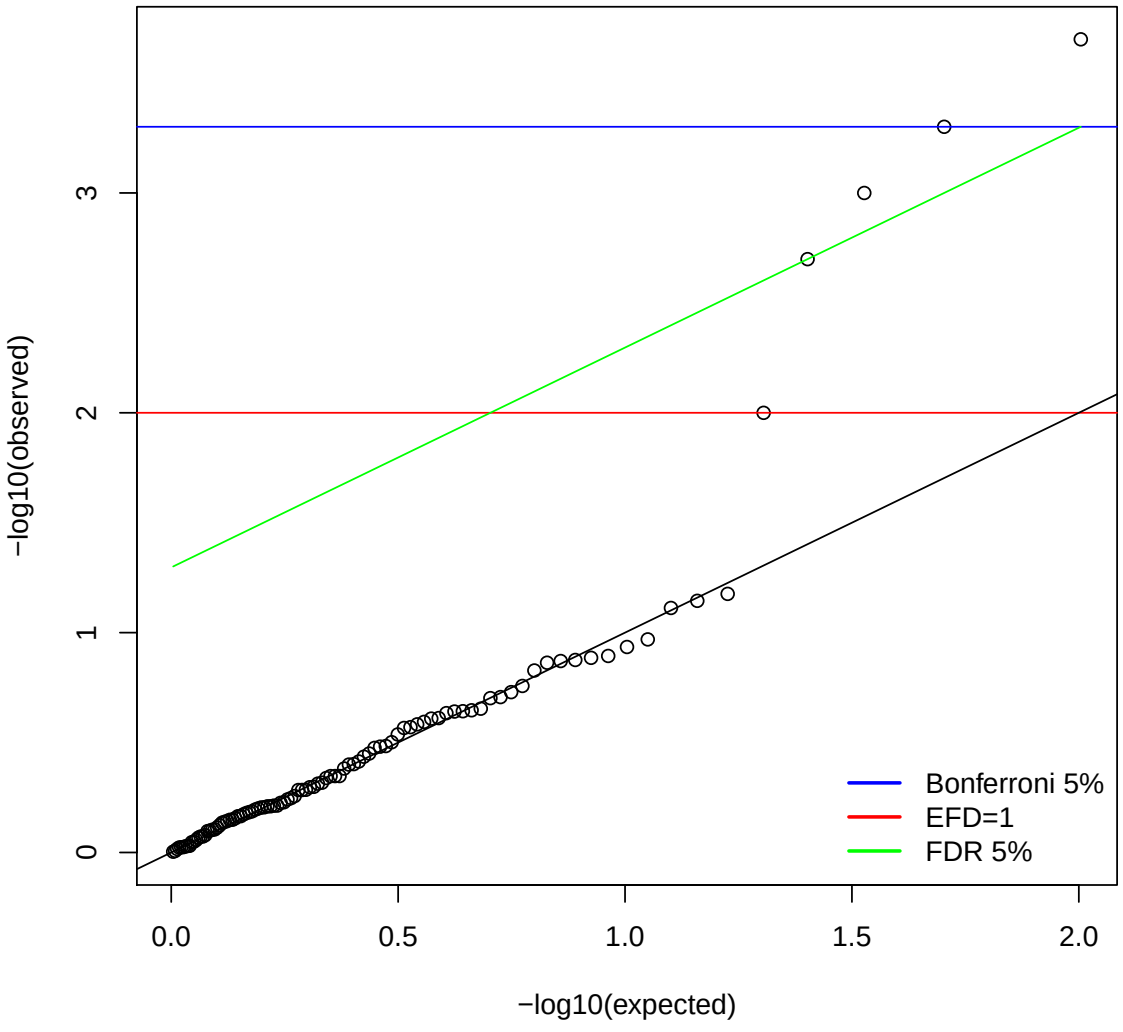


Figure 4 : Observed versus expected *p*-values, on $-\log_{10}$ scale.

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[2].

[2] 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.
- 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.
- In a multiple-hypothesis testing situation (and assuming ordered so the first K are rejected), we have

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

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

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).

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 β 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 so that at a generic SNP we have $S = 0$:

  No Copies of the Minor Allele

  and $S = 1$:

  One or Two Copies of the Minor Allele.
- There are therefore m comparisons of interest, with summary data at marker j, as below:

	Marker		
	$S = 0$	$S = 1$	**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 possess the marker of interest.

Treatment-by-Marker Interactions

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

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

  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: The standard error 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 5 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.

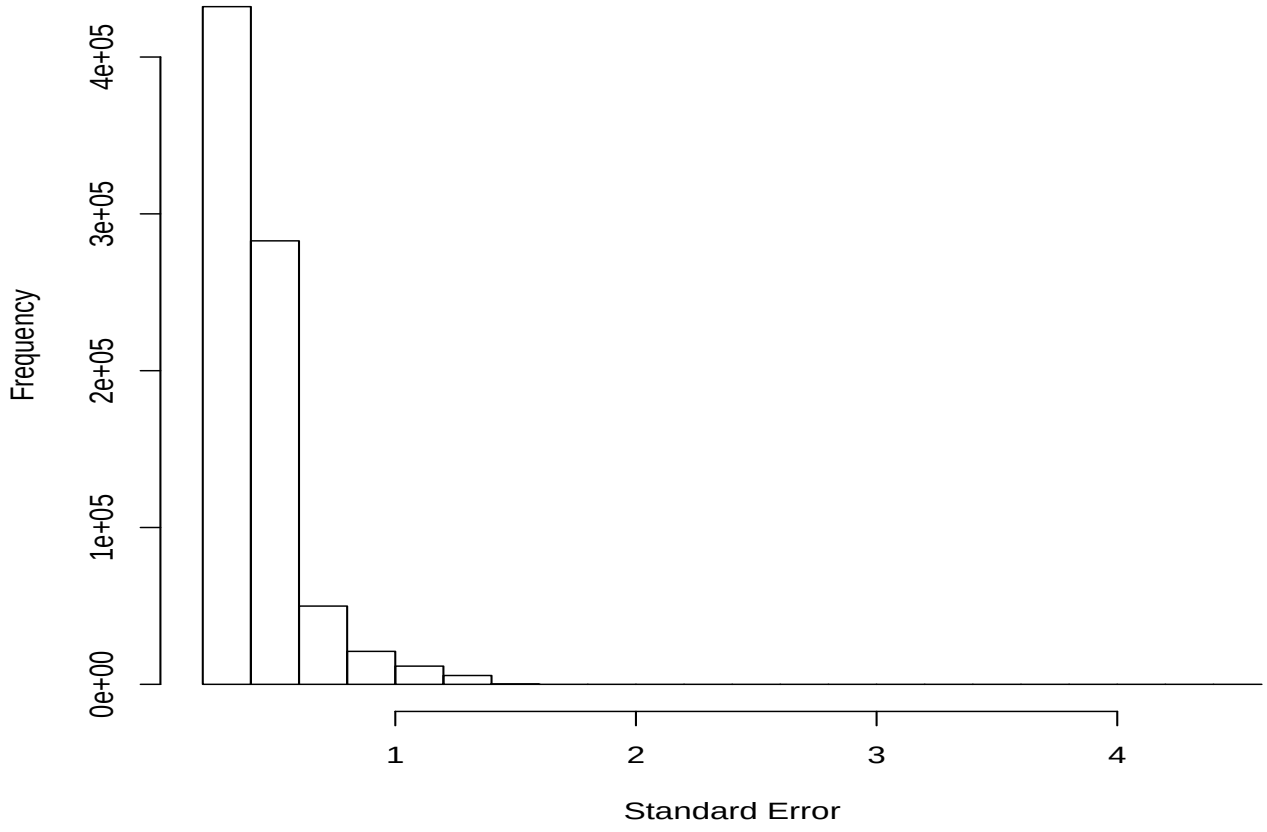


Figure 5 : Histogram of standard errors of the interaction parameter estimates $\widehat{\Delta}$ in the VISP study.

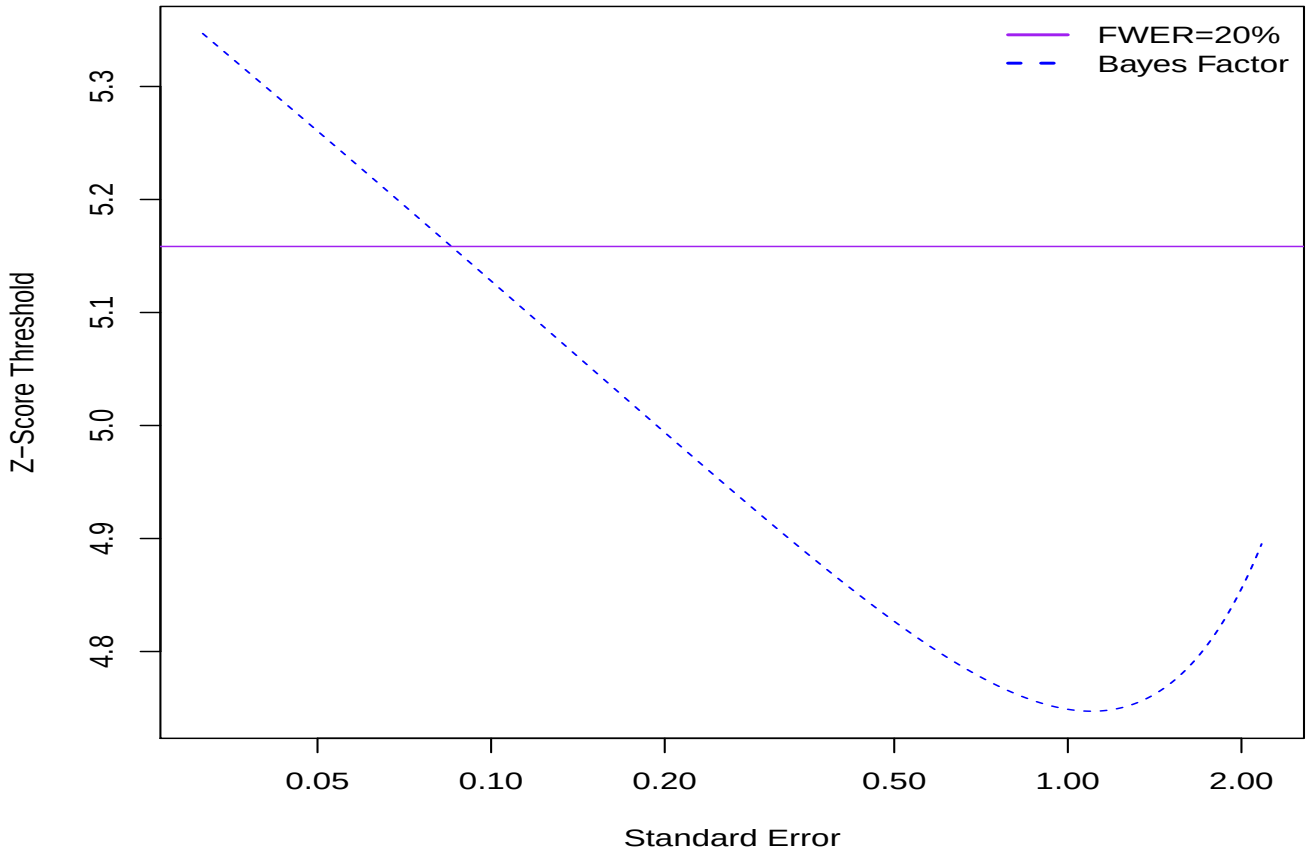


Figure 6 : Bayesian Z-score threshold as a function of the standard error. The Bayesian threshold is based on a prior on the alternative of 0.0001, $R = 1$ and a prior standard deviation on the interaction effect size of $\sqrt{W} = 5.1$; this prior gives a 95% interval on Δ of (-10,10).

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 $\pi_1 = 0.001, 0.0001, 0.0001$.
- 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 7 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 α and β are the type I and type II errors.
- 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)

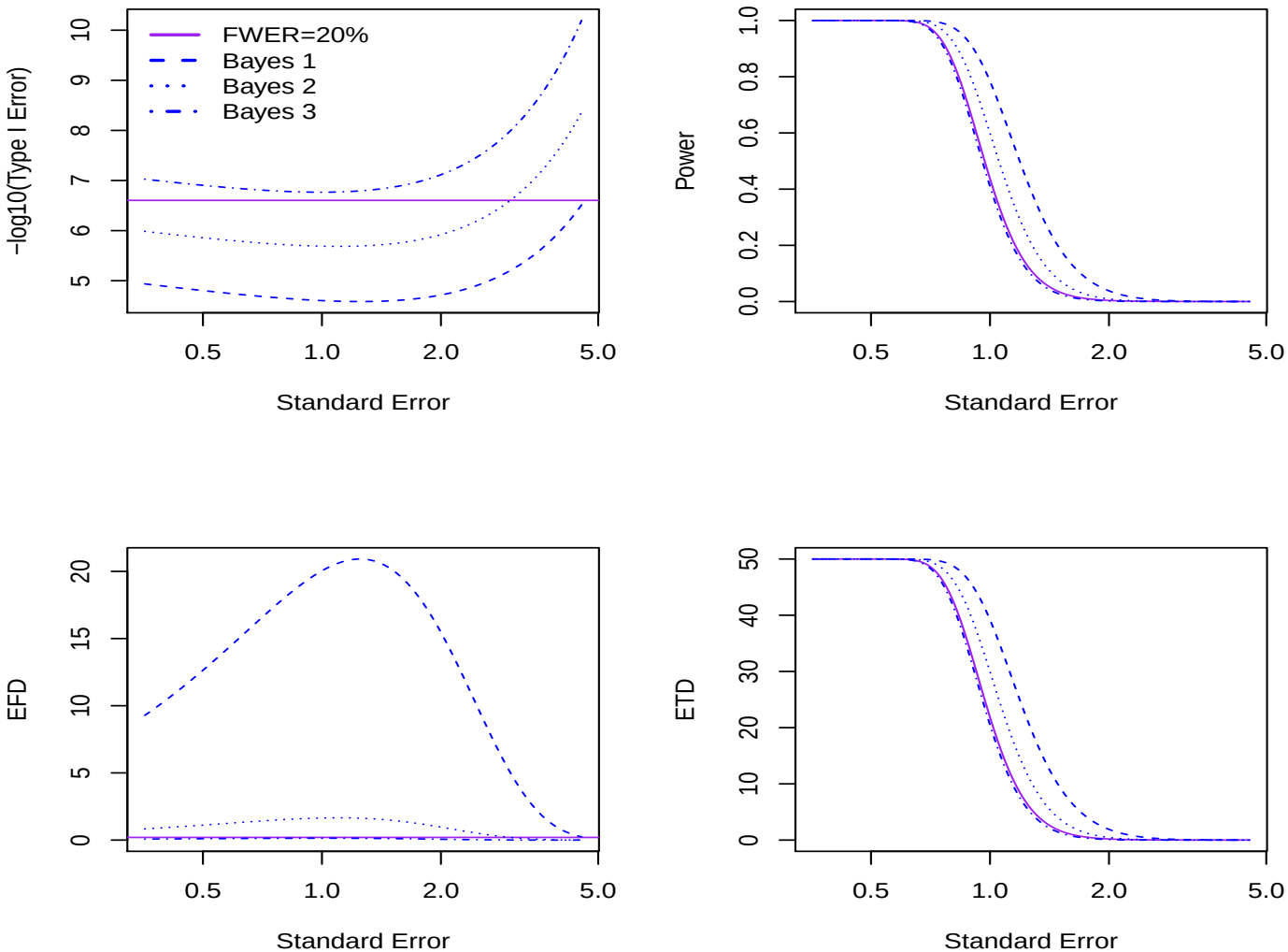


Figure 7 : Operating characteristics of Bayes/Bonferroni. For Bayes boundaries R = 1 and "Bayes 1", "Bayes 2", "Bayes 3" correspond to priors of $\pi_1 = 0.001, 0001, 00001$. Power is to detect a drop of 5 units. For EFD/ETD we set $m_1 = 50$.

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 Δ of ± 10.
- Figure 8 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 2.5×10^{-7}) two SNPs are flagged.
- With a FWER of 5% the Bonferroni threshold is 6.2×10^{-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).

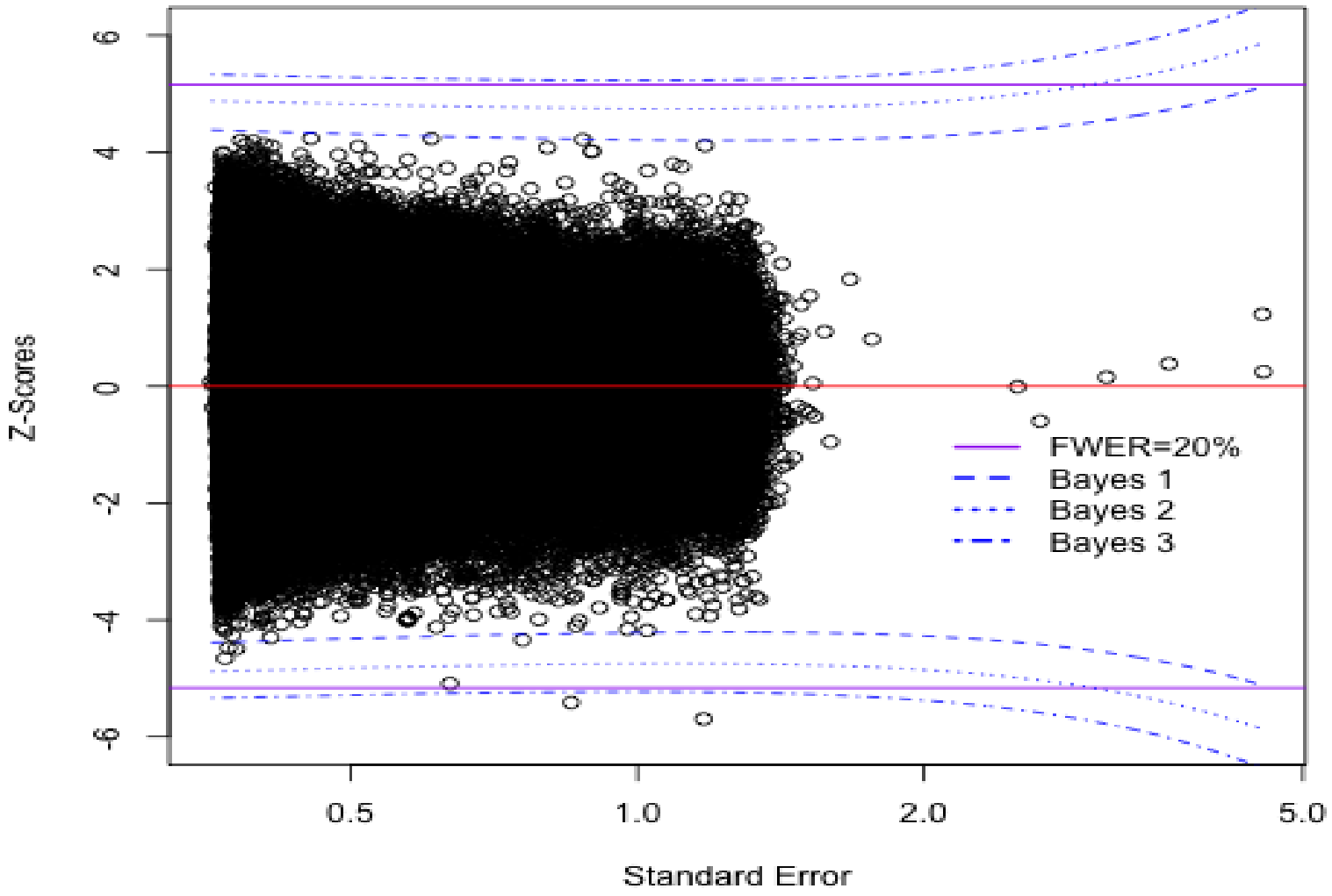


Figure 8 : Z-score threshold as a function of the standard error for the VISP data, ratio of costs of type II to type I errors R = 1 and varying priors on the alternative of $\pi_1 = 0.001, 0001, 00001$ (to give Bayes 1, Bayes 2, Bayes 3 boundaries).

Flagged Signals

SNP ID	Chrom	$\widehat{\Delta}$	$\widehat{\text{s.e.}}(\widehat{\Delta})$	*p*-value	Bayes Factor	Post Prob
rs3736238	17	-6.68	1.38	1.5×10^{-8}	9.3×10^{-7}	0.99
rs16893296	6	-4.61	0.85	7.1×10^{-8}	3.9×10^{-6}	0.96
rs1739317	6	-3.23	0.64	4.0×10^{-7}	2.3×10^{-5}	0.81
rs11819196	10	-1.72	0.37	3.5×10^{-6}	2.9×10^{-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 9 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 8 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.

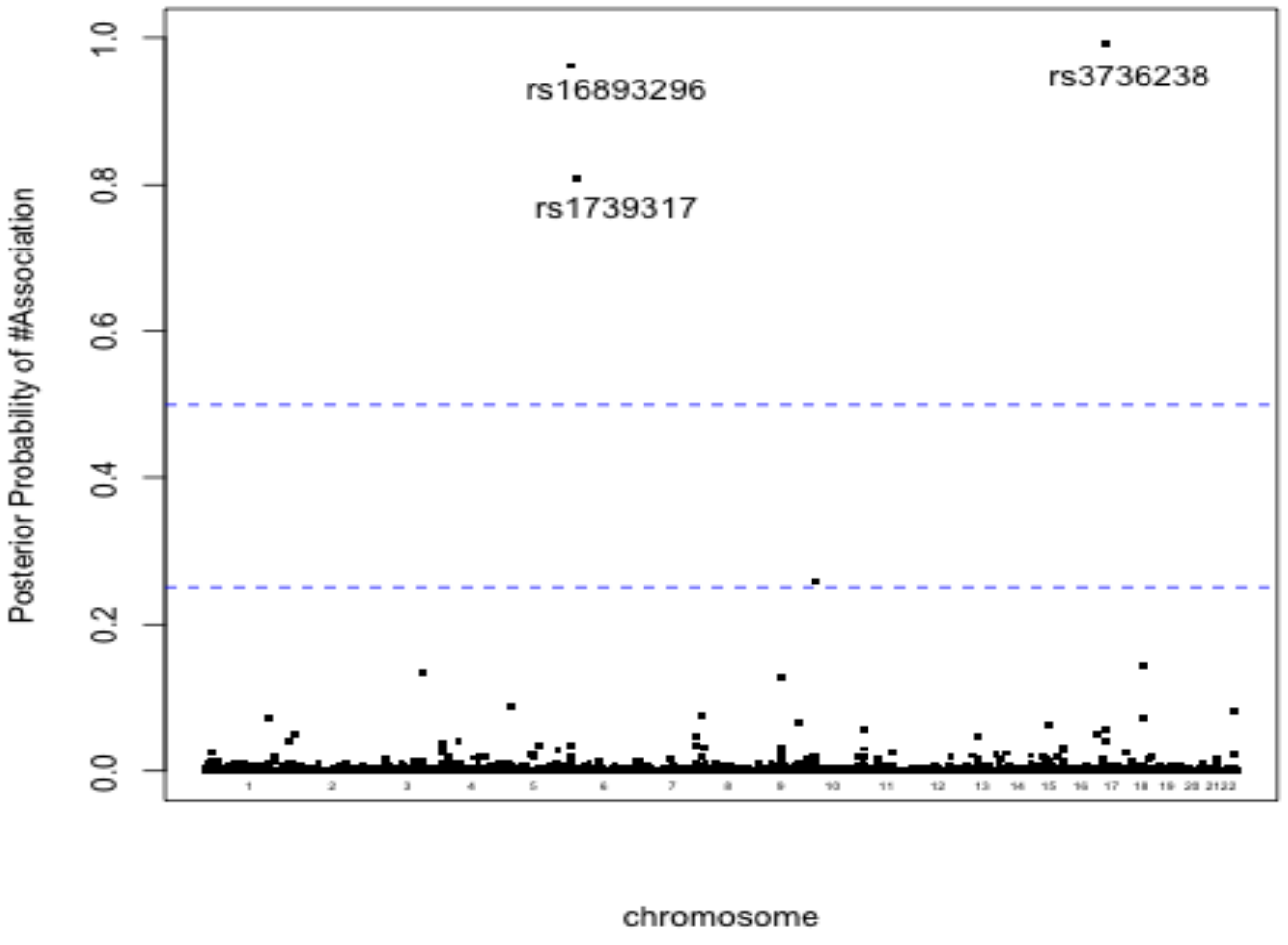


Figure 9 : Posterior probability on the alternative plotted versus genomic position for the VISP data. The prior on the alternative is $\pi_1 = 0.0001$.

VISP Example

- Figure 10 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 6.
- 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 11 shows an example in which distinctly different behavior occurs.

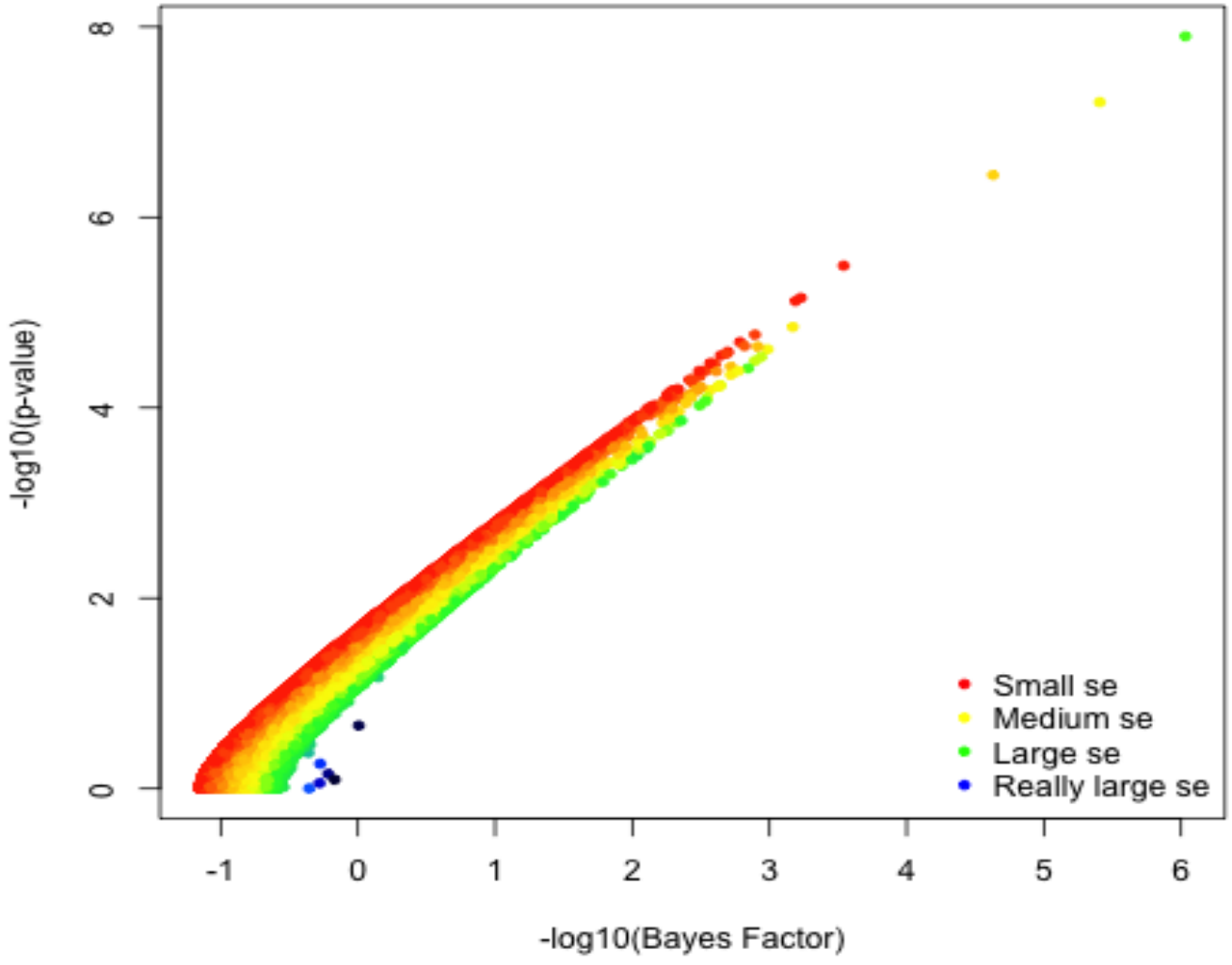


Figure 10 : -$\log_{10}$BFs vs -$\log_{10}$ p-values, color-coded by standard error with $W = 10$.

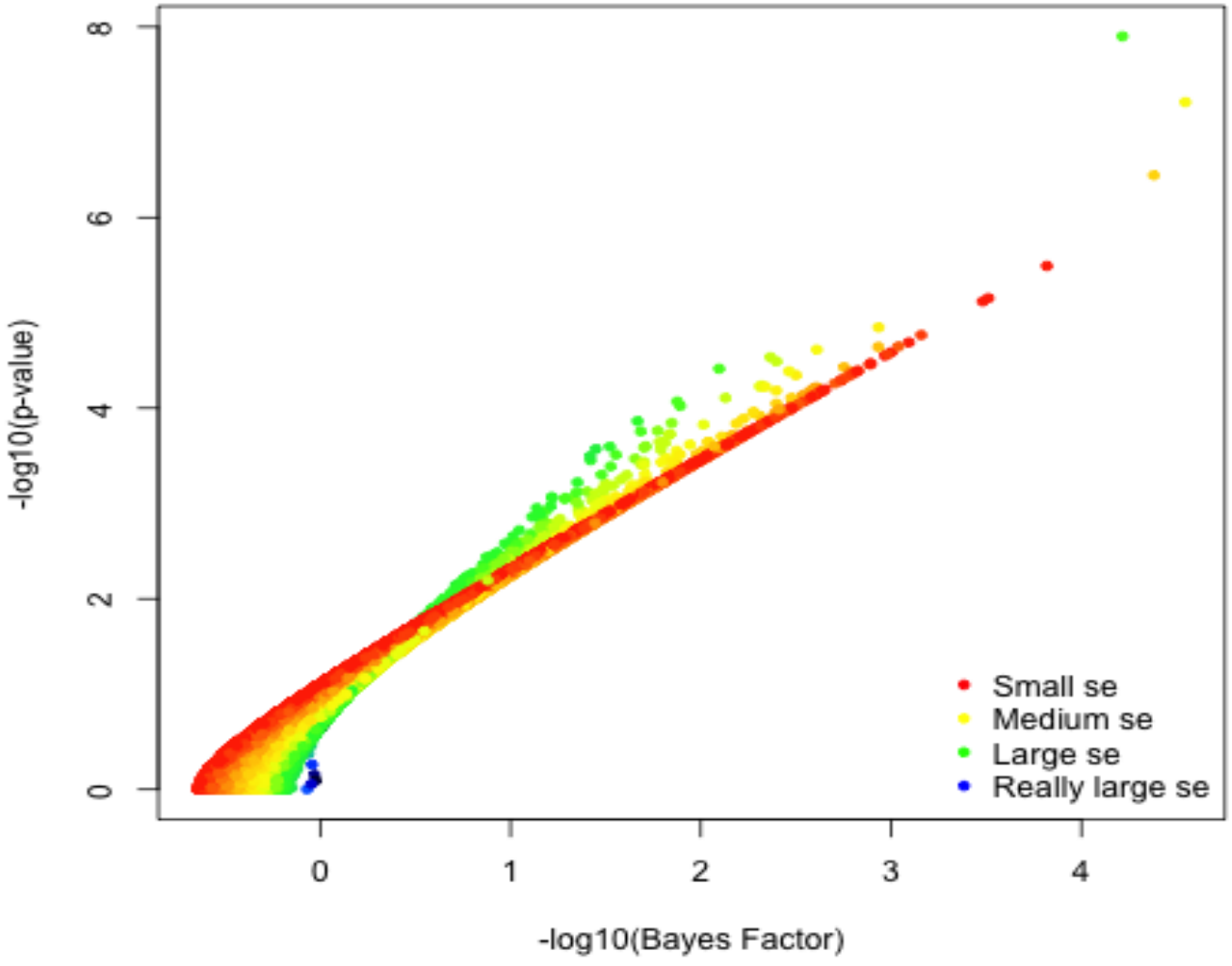


Figure 11 : -$\log_{10}$BFs vs -$\log_{10}$ p-values, color-coded by standard error with $W = 3$.

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.
- 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 ± 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 π_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 12 gives the posterior probabilities for this prior.

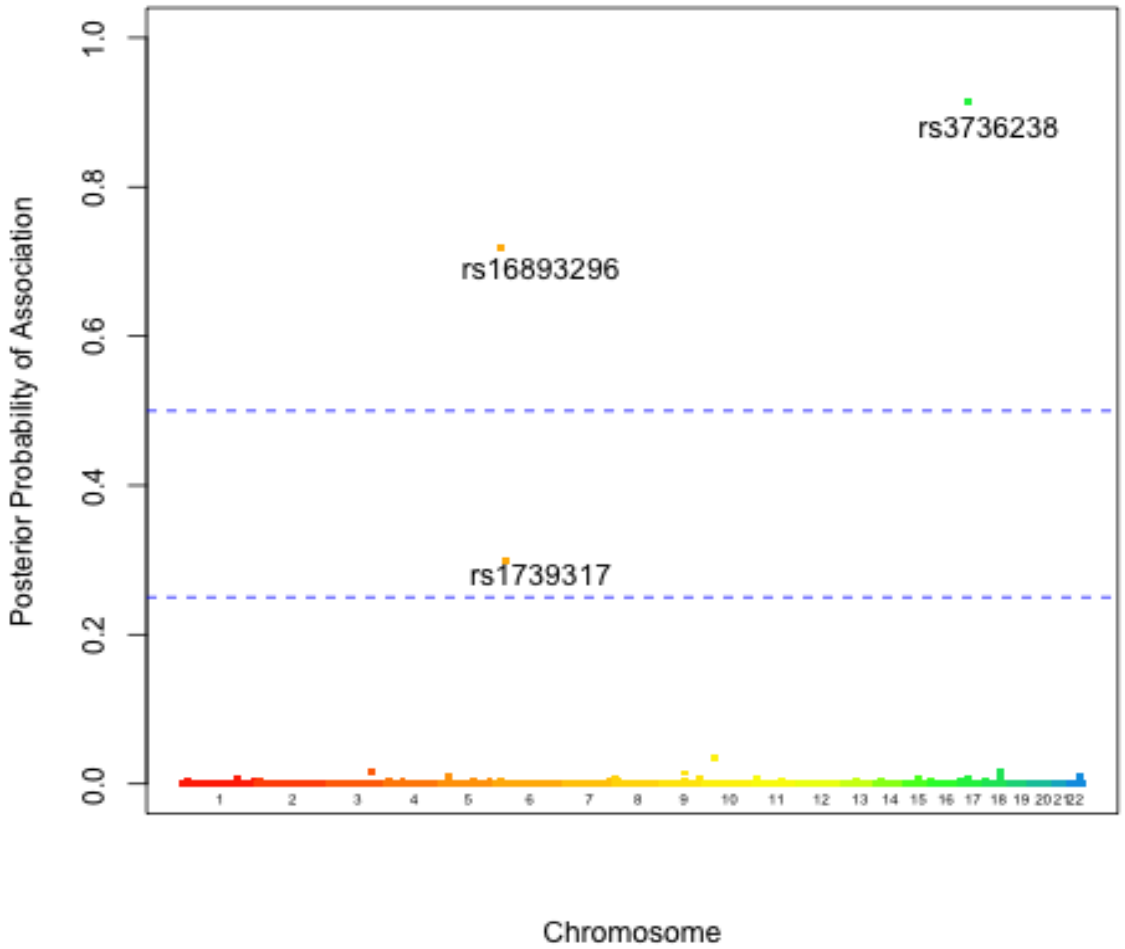


Figure 12 : Posterior probability on the alternative plotted versus genomic region for the VISP data. The prior on the alternative is the more conservative choice of $\pi_1 = 0.00001$.

VISP Discussion

- We chose the value $\pi_1 = 0.0001$ by examining frequentist summaries before the real data analysis was performed.
- We define π_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 a major issue.
- 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 π_1

- The posterior probability of the alternative is highly dependent on the choice of prior on the null π_0, and a sensitivity analysis is always warranted.
- Ideally, rather than fix π_0 as we have done, one would estimate of π_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.
- In other contexts, such as the analysis of gene expression data (Storey and Tibshirani, 2003), the data can be used to estimate π_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 π_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 π_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 $U = \log(\text{PO/R})$ 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.

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.
- 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.
- In genetics, journals are sympathetic to Bayes analyses (not true in all disciplines).
- A huge GWAS enterprise used p-values and Bayes factors to assess significance (Wellcome Trust Case Control Consortium, 2007).
- Stephens and Balding (2009) provide a review of Bayesian approaches in GWAS.

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.

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.

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.

Stephens, M. and Balding, D. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, **10**, 681–690.

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

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Bayesian Statistics for Genetics
Lecture 7: Testing and Multiple Testing

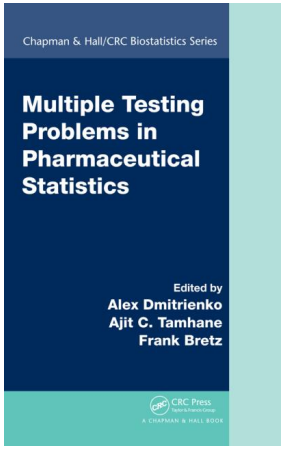
Ken Rice

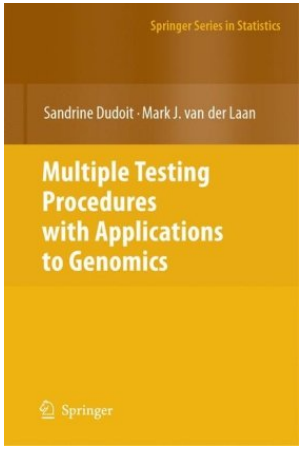
UW Dept of Biostatistics

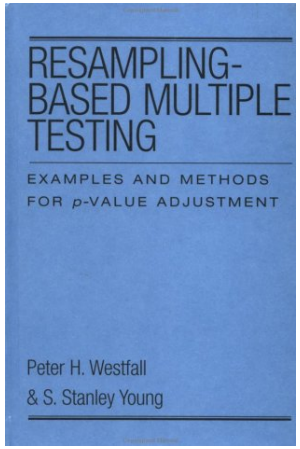
July, 2016

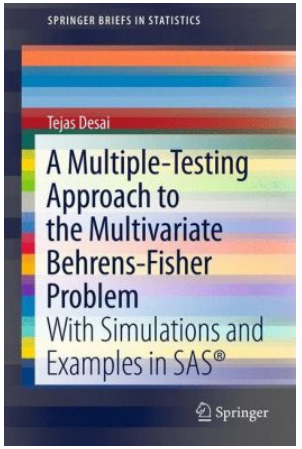
Overview

Rather than trying to cram another book's-worth of material into a single session...







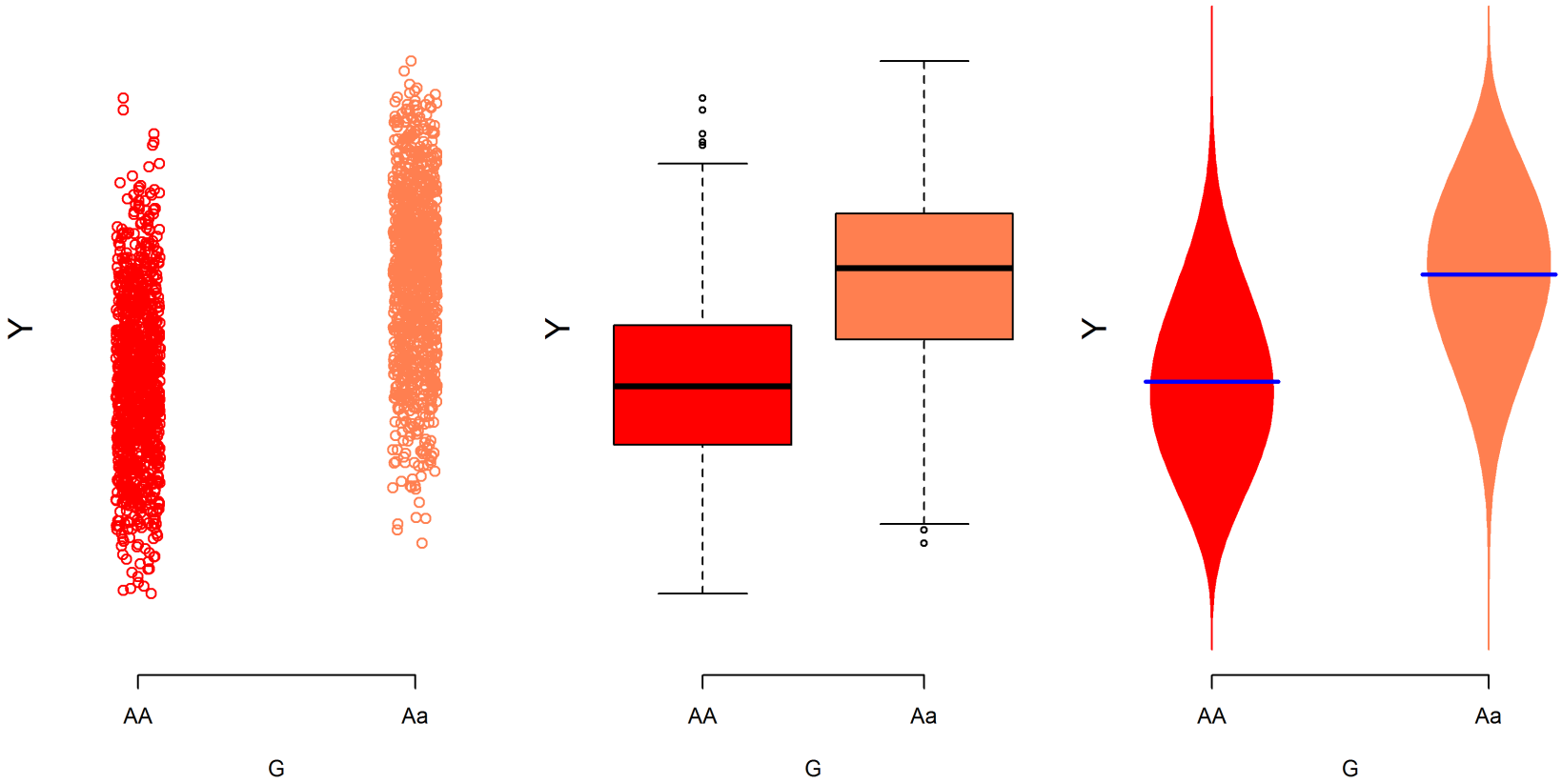


- Testing — what we do and what it (may) mean
- Multiple testing — primarily for high-throughput settings

NB Testing is the statistical area with most divergence between *default* Bayesian and non-Bayesian methods — and the foundations of both are contentious — so this focus in this session is more statistical than usual.

Testing

Typical motivation; is the SNP variant (G) associated with phenotype (Y, e.g. height)?



- Is the association positive? negative? zero?
- Is the variant causal – or just associated via LD?
- Given what You know, is the association worth mentioning?

Testing

To answer whether an association is "worth mentioning", this term must be defined — and there are many ways we could measure worth;

- Strong enough belief that θ is positive to outweigh saying it's negative
- Strong enough belief about θ's direction to outweigh saying nothing about direction
- Strong enough belief that θ is non-zero to outweigh saying that it's zero
- Strong enough belief (based on the data, and relevant to the prior) that θ is non-zero to outweigh saying that it's zero
- Strong enough belief about θ's distance from zero to outweigh saying nothing about its value

These can all give different answers, depending on the details – and the data.

Testing: decision theory

We introduce some concepts from Bayesian *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)
- Expected posterior loss $\mathbb{E}[\, L(\theta, d)\,]$ – loss for some decision d averaged over posterior uncertainty

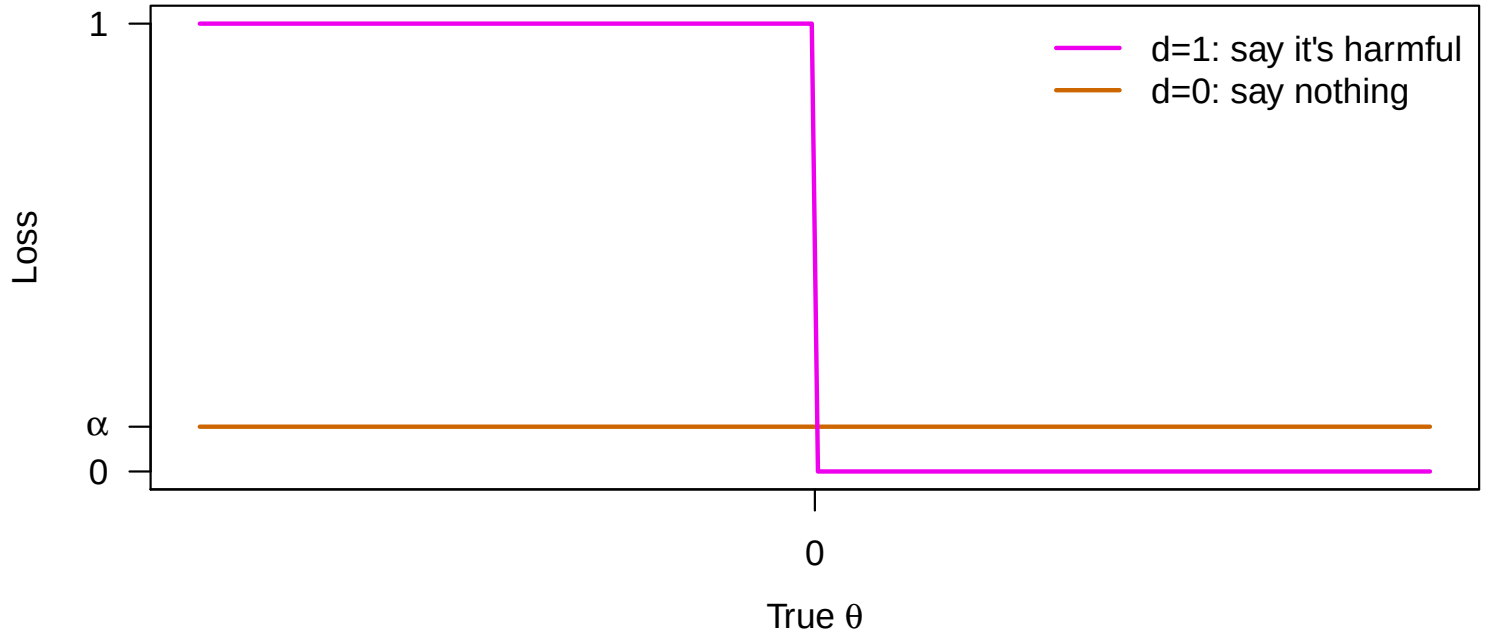
Theory (and intuition) states that Your best decision – the *Bayes rule* — is the decision d that minimizes $\mathbb{E}[\, L(\theta, d)\,]$.

For testing, d is 0 or 1, so this means checking whether

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

Testing: first example

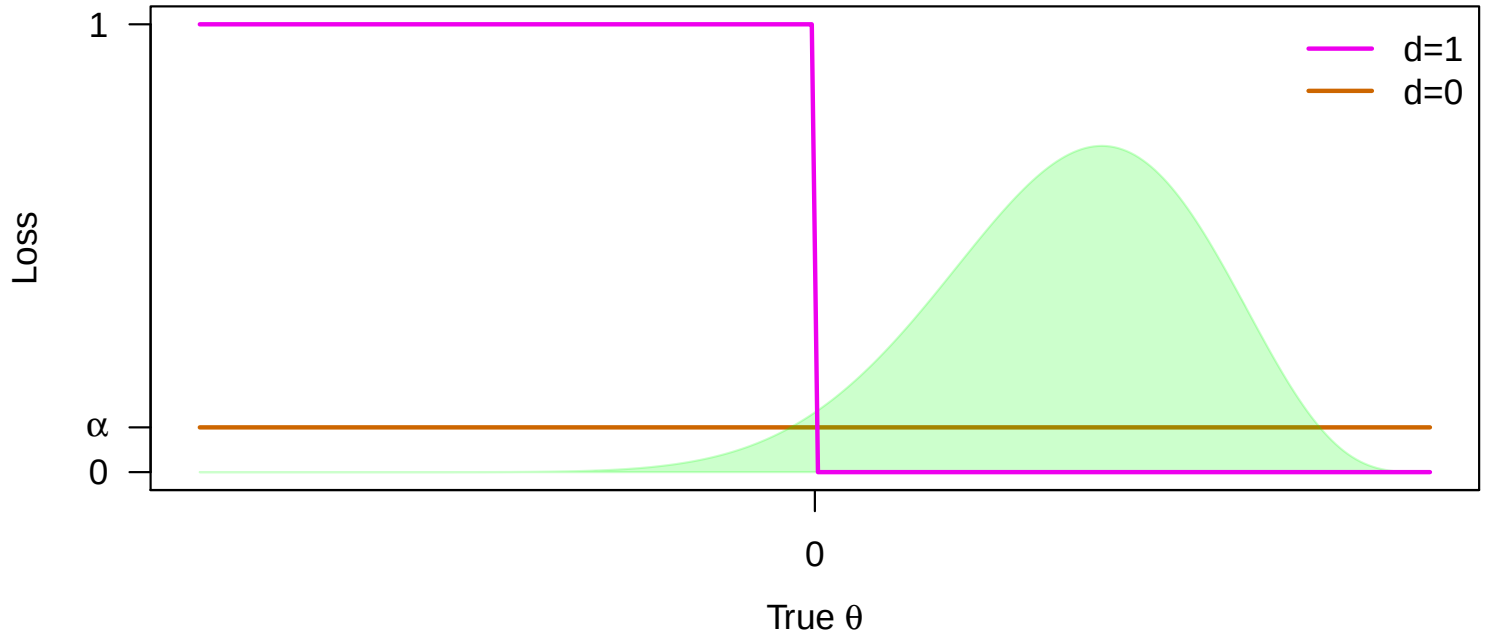
A first example: assessing whether a variant is harmful – it has $\theta > 0$ – suppose we either state ($d = 1$) that it is or say nothing at all ($d = 0$) about θ;



- $L(\theta, d = 1) = 1$ if $\theta \leq 0$, i.e. large cost for getting it wrong
- $L(\theta, d = 1) = 0$ if $\theta > 0$, i.e. no cost for getting it right
- $L(\theta, d = 0) = \alpha$: small cost of saying nothing, regardless of the true value of θ

Testing: first example

Averaging over a green posterior;



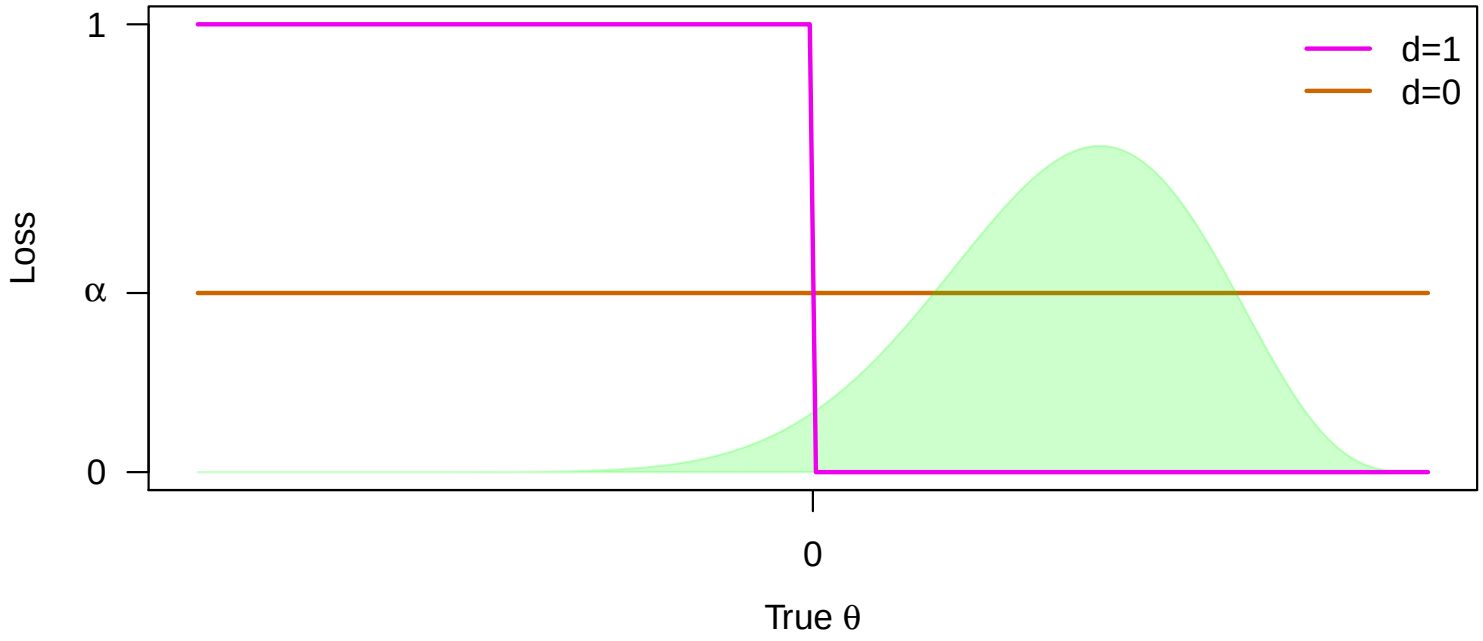
The expected posterior loss is

$$\mathbb{E}[\,L(\theta,d)\,] = \begin{cases} \alpha, & d=0 \\ \mathbb{P}[\,\theta<0|Y\,], & d=1 \end{cases},$$

... so the Bayes rule sets $d=1$ if $\mathbb{P}[\,\theta<0|Y\,]<\alpha$ here.

Testing: first example

At a higher α, its 'easier' to get $d = 1$;



If more than α of the posterior is in the tail below zero, the Bayes rule is to say nothing, i.e. return $d = 0$.

Testing: first example revisited

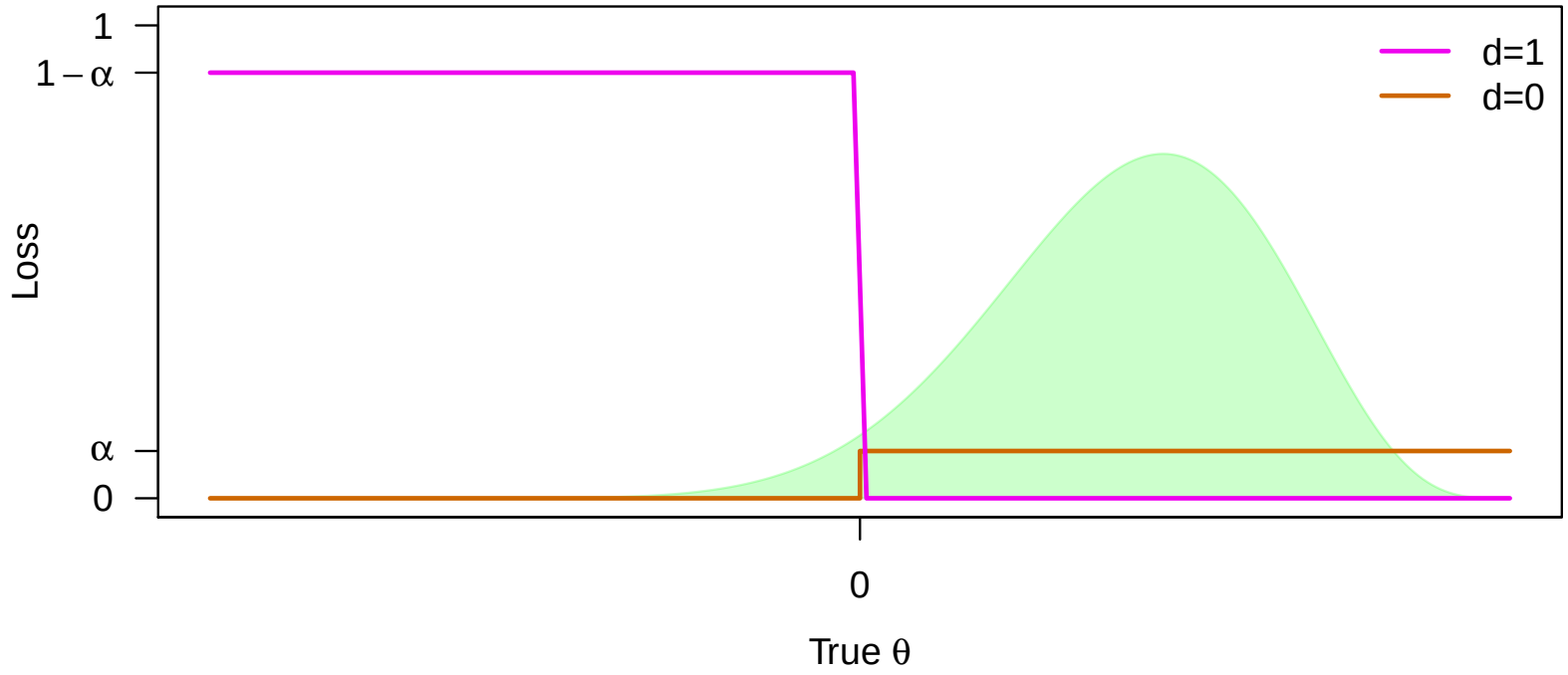
Now suppose we assess the sign of the variant's effect; and let $d = 1$ decide $\theta > 0$, and $d = 0$ for $\theta \leq 0$?

		Truth	
		$\theta \leq 0$	$\theta > 0$
Decision	$d = 0$	0	α
	$d = 1$	$1 - \alpha$	0

- No cost for getting the answer right (a *proper* loss function)
- Small penalty for incorrectly saying $\theta > 0$
- Large penalty for incorrectly saying $\theta \leq 0$

Testing: first example revisited

As a picture;



And working out the posterior loss;

$$\mathbb{E}[\,L(\theta,d)\,] = \begin{cases} \alpha\mathbb{P}[\,\theta > 0|Y\,], & d=0 \\ (1-\alpha)\mathbb{P}[\,\theta < 0|Y\,], & d=1 \end{cases},$$

... so – again! – the Bayes rule sets $d=1$ if $\mathbb{P}[\,\theta < 0|Y\,] < \alpha$.

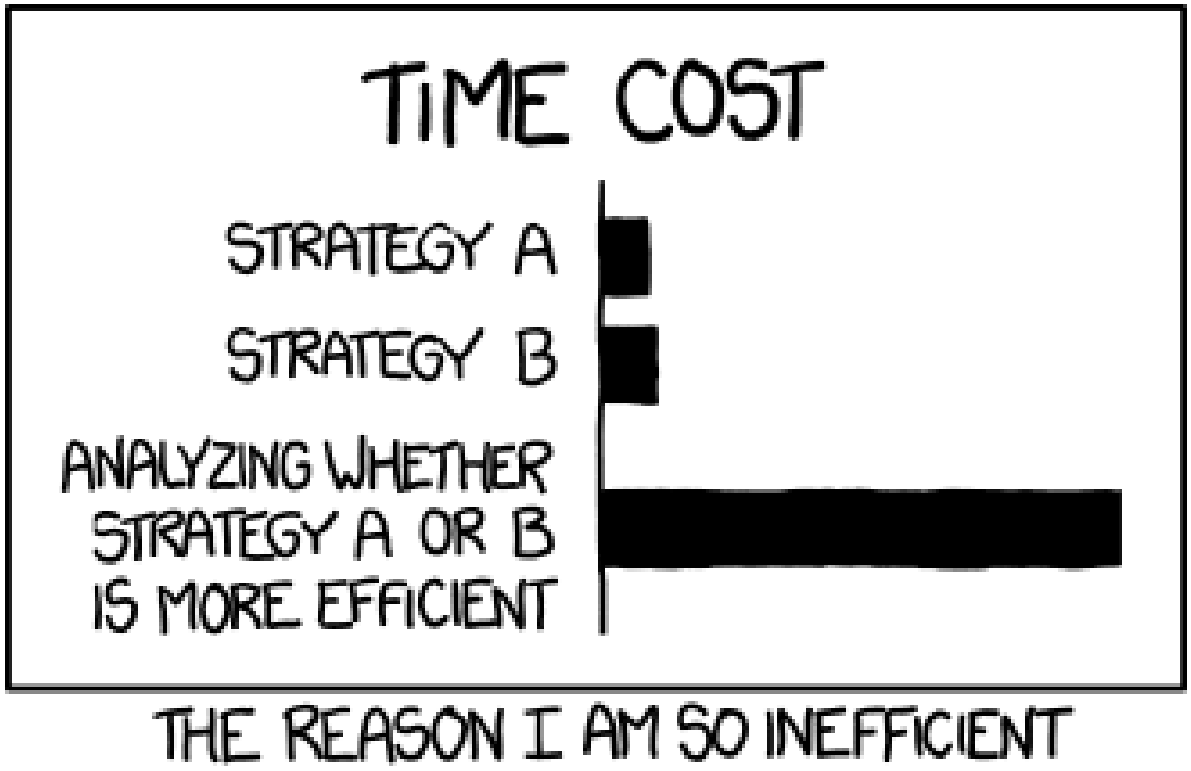
Testing

Some notes so far;

- These are *one-sided* tests, of the null hypothesis that $\theta < 0$
- "Reject the null *vs* say nothing" is a *significance test*
- "Reject the null *vs* accept the null" is a *hypothesis test*
- The test have different decisions, even though both just look at whether tail area $< \alpha$.
- This is also true for one-sided frequentist significance/hypothesis tests – in which p-values are approximately our tail areas, in large samples, if likelihood dominates prior
- Not (yet!) making decisions that θ is exactly zero, or any other specific value... so don't conclude this without more assumptions

Testing

XKCD on loss functions;



Testing: doing two tests at once

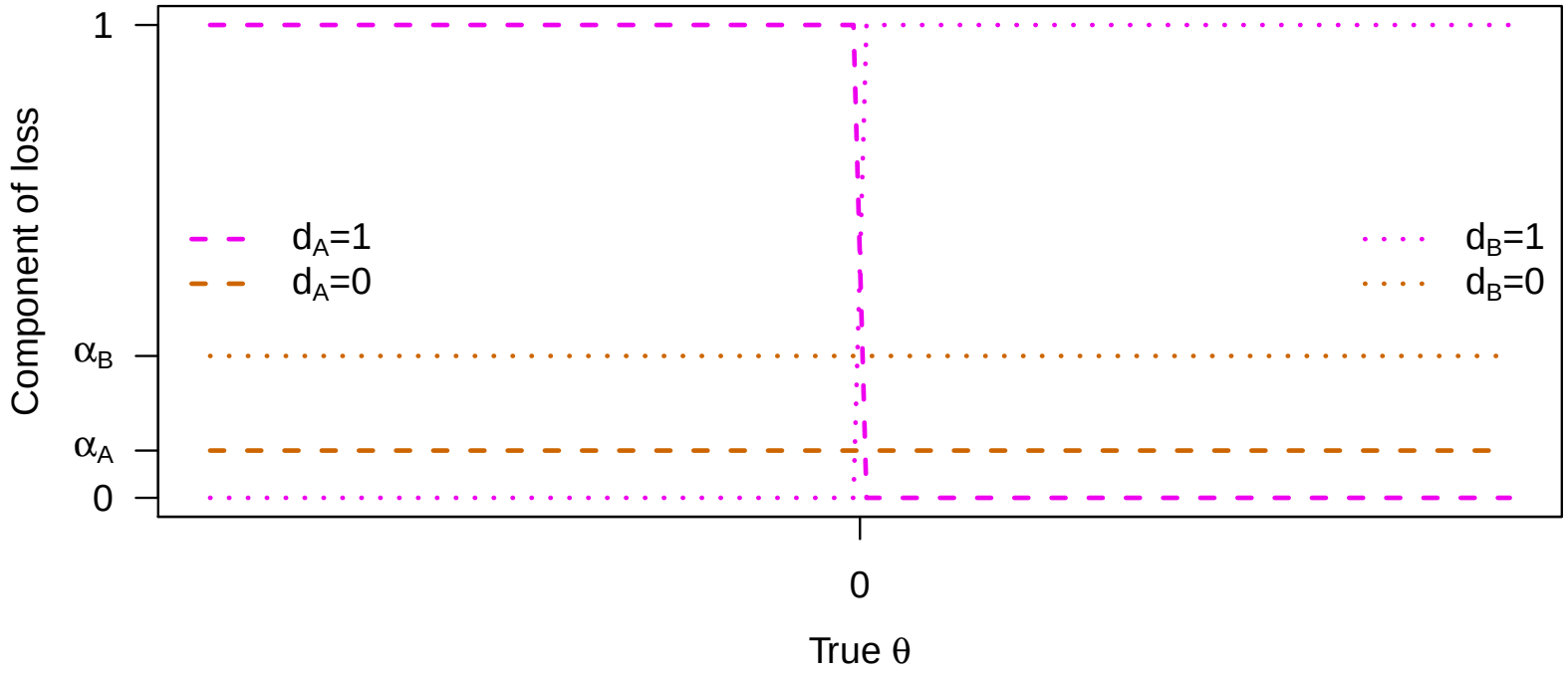
Back to the significance test, i.e. say something vs nothing – but now let's do **two** one-sided tests, that decide if θ is **A**bove 0 or **B**elow 0;

	Decision	Truth	Loss
d_A	0		α_A
	1	$\theta > 0$	0
	1	$\theta \leq 0$	1
d_B	0		α_B
	1	$\theta < 0$	0
	1	$\theta \geq 0$	1

... where we get $L(\boldsymbol{d}, \theta)$ by adding the two components.

Testing: doing two tests at once

As a picture – d_A as dashed lines, d_B as dotted;



Here are the possible overall posterior losses;

	$d_B = 0$	$d_B = 1$
$d_A = 0$	$\alpha_A + \alpha_B$	$\alpha_A + \mathbb{P}[\theta > 0]$
$d_A = 1$	$\alpha_B + \mathbb{P}[\theta < 0]$	$\mathbb{P}[\theta < 0] + \mathbb{P}[\theta > 0] = 1$

Testing: doing two tests at once

Which option is best?

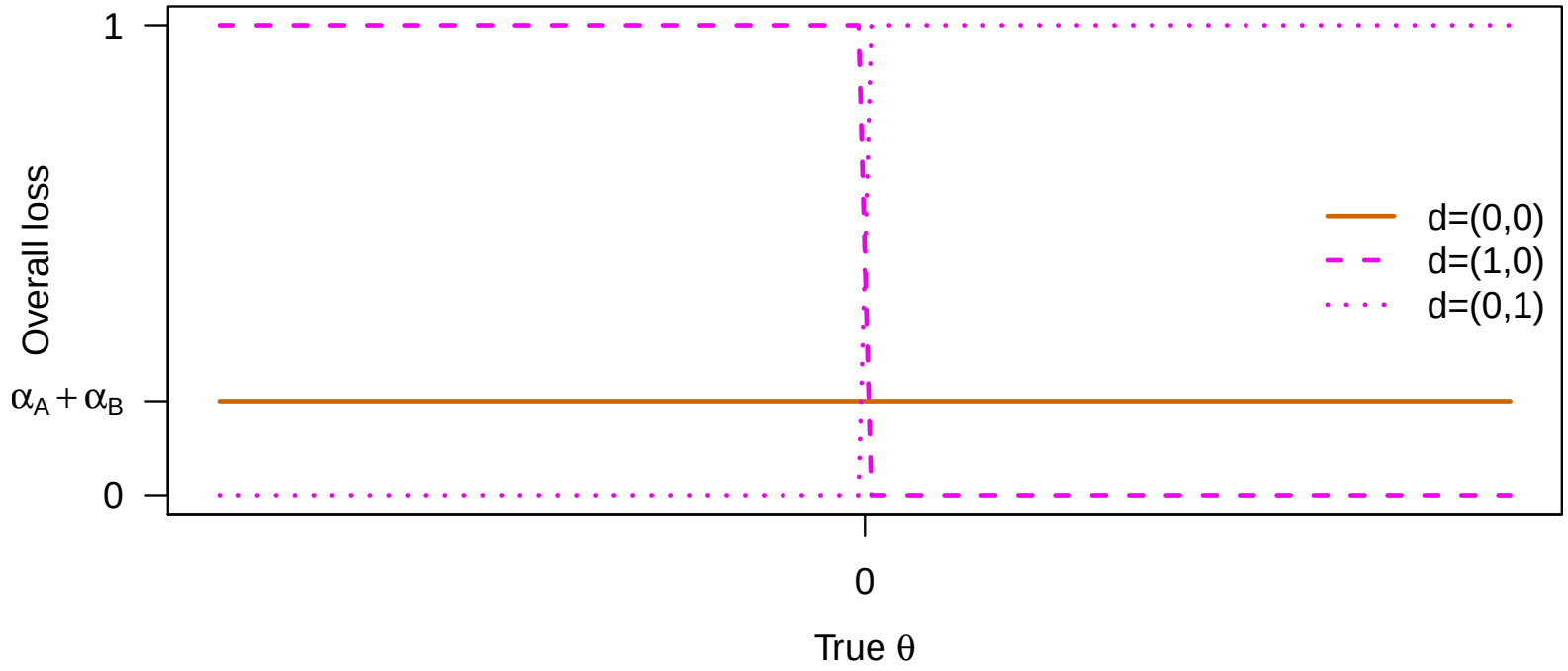
	$d_B = 0$	$d_B = 1$
$d_A = 0$	$\alpha_A + \alpha_B$	$\alpha_A + \mathbb{P}[\theta > 0]$
$d_A = 1$	$\alpha_B + \mathbb{P}[\theta < 0]$	1

- Assuming $\alpha_A + \alpha_B < 1$, we **never** choose $\boldsymbol{d} = (d_A, d_B) = (1, 1)$
- If $\mathbb{P}[\theta < 0] < \alpha_A$, then $(1, 0)$ beats $(0, 0)$. And because $\mathbb{P}[\theta < 0] > 1 - \alpha_A$ it also beats (0,1) $\Rightarrow$ choose $\boldsymbol{d} = (1, 0)$
- If $\mathbb{P}[\theta > 0] < \alpha_B$, then $(0, 1)$ beats $(0, 0)$. And because $\mathbb{P}[\theta < 0] > 1 - \alpha_B$ it also beats (1,0) $\Rightarrow$ choose $\boldsymbol{d} = (0, 1)$
- If $\mathbb{P}[\theta < 0] > \alpha_A$ and $\mathbb{P}[\theta > 0] > \alpha_B$, $\alpha_A + \alpha_B$ is the best option, $\Rightarrow$ choose $\boldsymbol{d} = (0, 0)$

... so we 'say nothing' unless at least one tail is small. When one tail is small, the Bayes rule gives the corresponding statement about the sign of θ.

Testing: doing two tests at once

Overall loss functions for the three decisions we consider;



To keep the ratio of costs for 'say nothing' versus 'say something' the same α : 1 ratio as in the one-sided test, we need to put $\alpha_A + \alpha_B = \alpha$. One obvious way to do this is setting $\alpha_A = \alpha_B = \alpha/2$ – known as using *equal tails*.

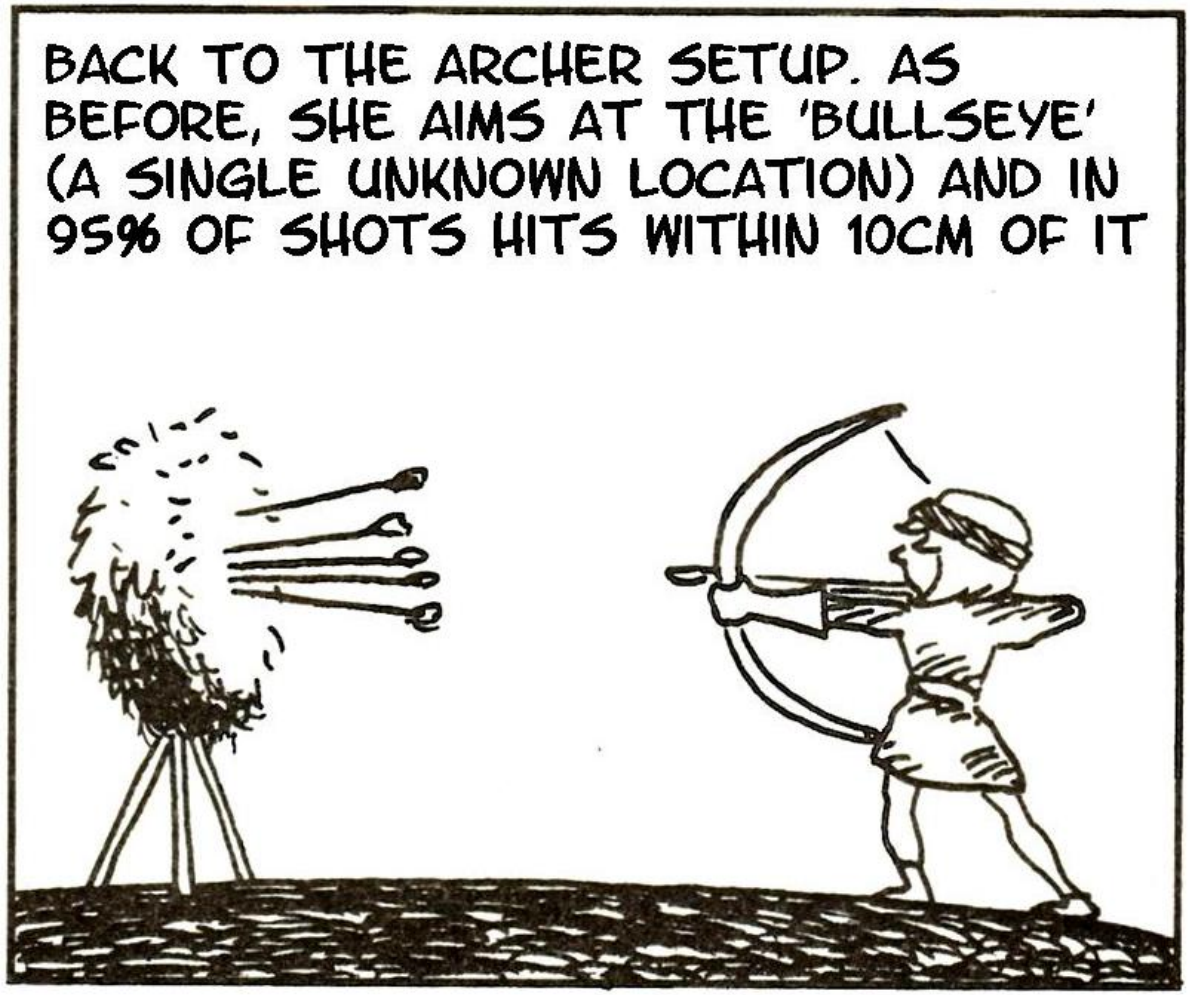
Testing: doing two tests at once

More notes:

- This is a Bayesian analog of a standard two-sided frequentist test. In large samples, they will give the same reject/don't reject decisions (with non-spiky priors)
- For two-sided tests, using anything except equal tails is unusual, in Bayesian or frequentist work
- Still not declaring that $\theta = 0$!
- Modifications of much the same argument can cope with multivariate $\boldsymbol{\theta}$ – where $d = 1$ trades off error in estimates of $\boldsymbol{\theta}$ versus inaccuracy saying ($d = 0$) that $\boldsymbol{\theta} = \mathbf{0}$. But the result is equivalent to checking $p < \alpha$.
- Here, α interpreted as how much You value saying nothing vs saying something – which is highly context-specific, but a lot easier than frequentist arguments...

Testing: frequentist tests

Recall our frequentist archer, from Session 1;



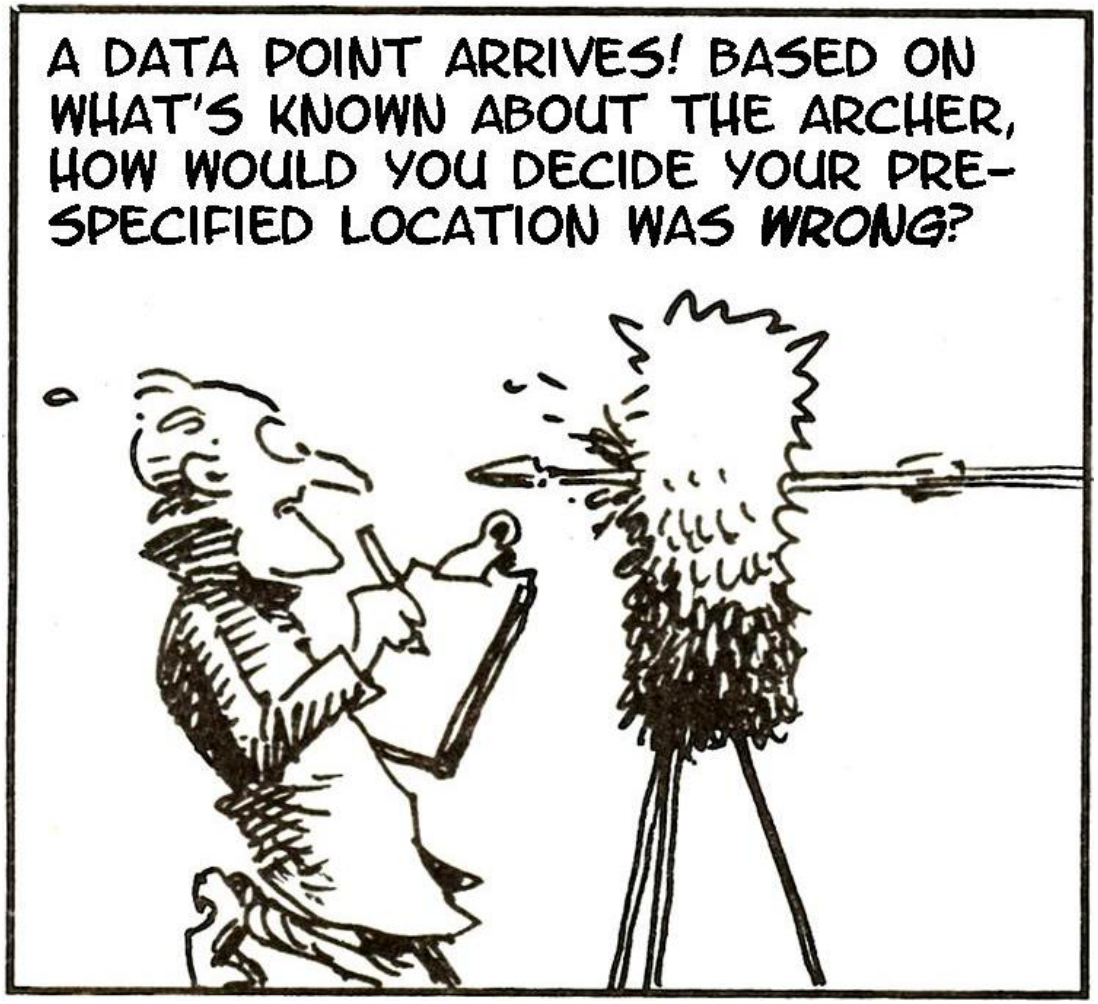
Adapted from Gonick & Smith, *The Cartoon Guide to Statistics*

Testing: frequentist tests

Let's do some more 'target practice', for frequentist testing;

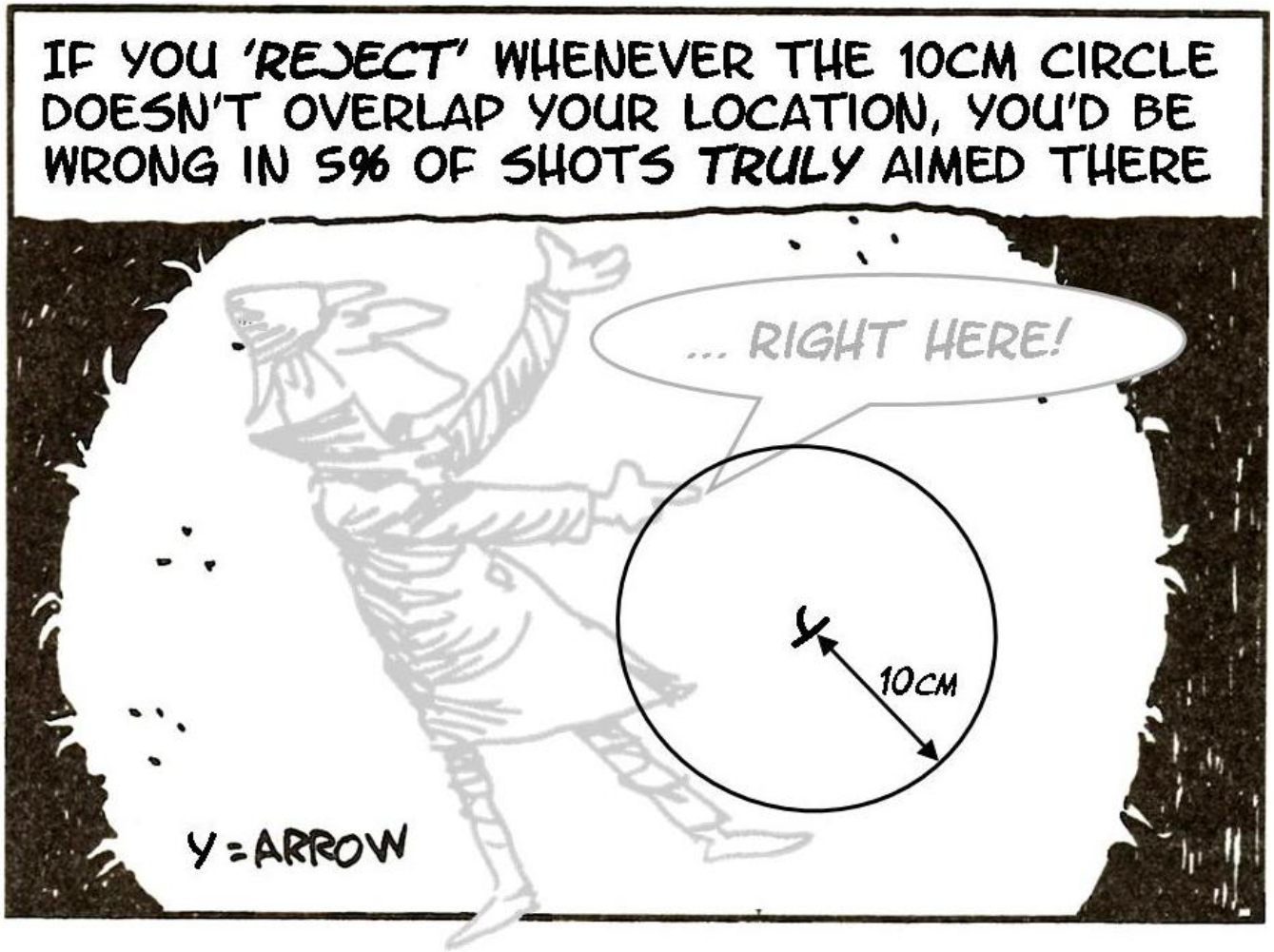
Testing: frequentist tests

Let's do some more 'target practice', for frequentist testing;



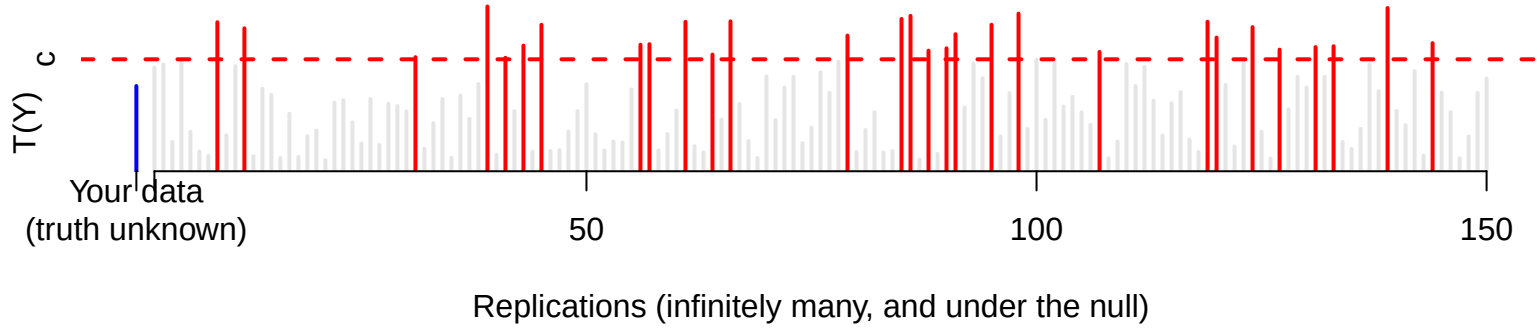
Testing: frequentist tests

Let's do some more 'target practice', for frequentist testing;



Testing: frequentist tests

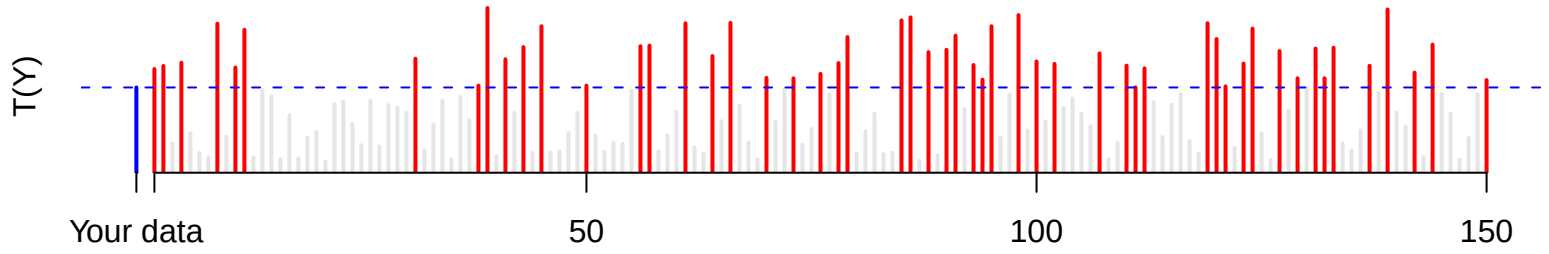
Performing the test means assessing whether our data beats some pre-specified measure of extremity;



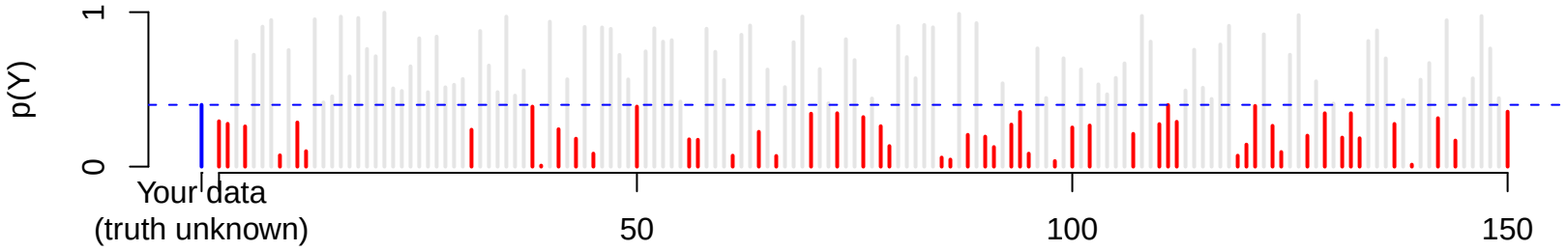
... where the threshold c is chosen so that, under the null, a fixed proportion α of datasets would be that extreme.

Testing: frequentist tests

For any measure $T(Y)$, we can also obtain the p-value the proportion of datasets we might observe at least as extreme as that observed, under the null;



Replications (infinitely many, and under the null)



Replications (infinitely many, and under the null)

... and then directly assess whether $p < \alpha$.

Testing: frequentist tests

- Frequentist testing is convoluted; at minimum, it requires comparison against many hypothetical replications
- Which test statistic to use is subjective; good choices can optimize power* for given Type I error rate α, but these may not be known
- One silly no-data example: throw a 20-sided dice and if you get 20 reject
- In practice – in genetics and elsewhere – controlling Type I error rates is a heavy focus, and power comes second

* NB power = probability of seeing a significant result, given that one is present

Testing: losses featuring the prior

We've considered only the sign of θ – giving no posterior to prior comparison. To fix this, we introduce θ^*, a parameter with the same prior as θ, but which is **not** updated by the data.

A loss examining how the signs of θ and θ^* compare;

	$\theta^* < 0$		$\theta^* > 0$	
	$\theta < 0$	$\theta > 0$	$\theta < 0$	$\theta > 0$
$d = 0$	l_N	1	0	l_P
$d = 1$	l_N	0	B_0	l_P

- If signs agree, d doesn't matter
- No penalty for $d = 1$ if $\theta^* < \theta$, or for $d = 0$ if $\theta^* > \theta$
- Small penalty (1) if $d = 0$ but $\theta^* < \theta$
- Large penalty (B_0) if $d = 1$ but $\theta^* > \theta$
- Bayes rule returns $d = 1$ if

$$B_0 \mathbb{P}[\,\theta^* > 0\,]\mathbb{P}[\,\theta < 0|Y\,] \;<\; \mathbb{P}[\,\theta^* < 0\,]\mathbb{P}[\,\theta > 0|Y\,],$$

i.e.

$$\frac{\mathbb{P}[\,\theta > 0|Y\,]}{1 - \mathbb{P}[\,\theta > 0|Y\,]} \;>\; B_0 \frac{\mathbb{P}[\,\theta^* > 0\,]}{1 - \mathbb{P}[\,\theta^* > 0\,]}.$$

Testing: losses featuring the prior

Notes:

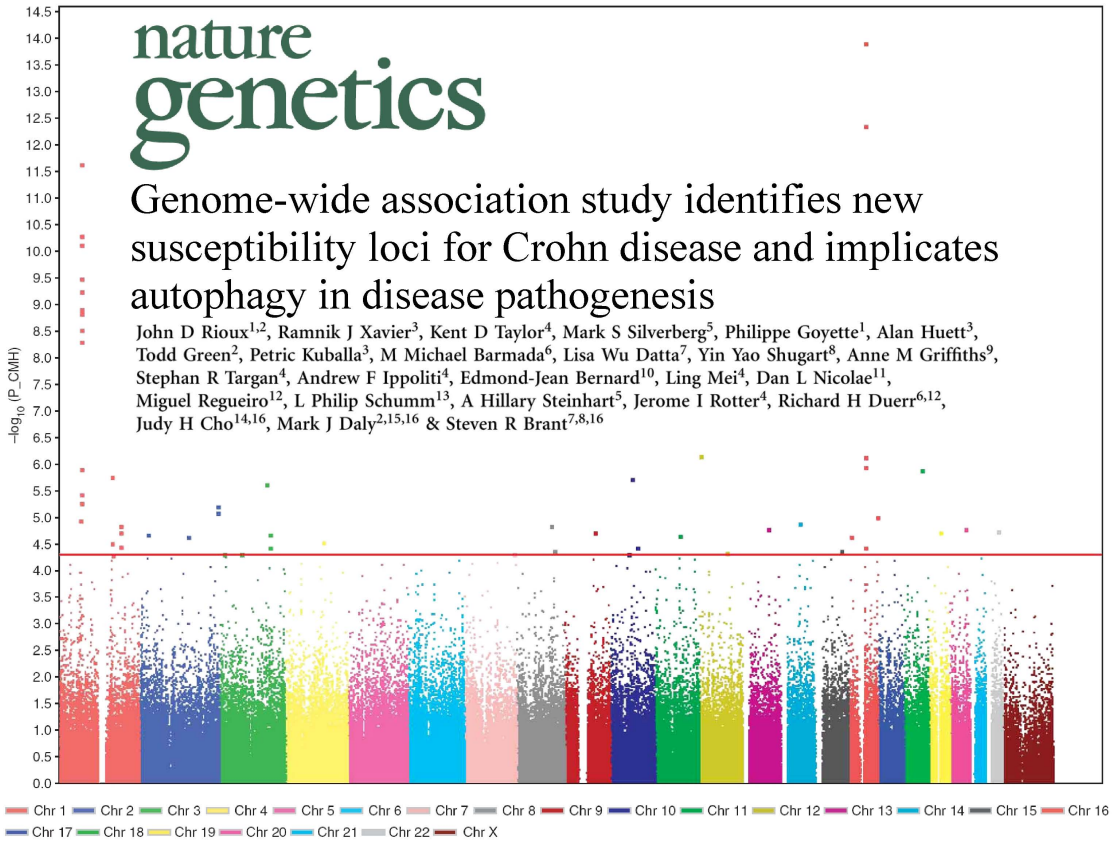
- This form of loss returns $d = 1$ if the posterior odds of positive θ, i.e. $\frac{\mathbb{P}[\theta>0|Y]}{1-\mathbb{P}[\theta>0|Y]}$ are more than B_0 times bigger than the prior odds of positive θ^*
- The ratio of the odds is known as the *Bayes factor* – usually denoted B. It does not depend on the prior support for $\theta^* > 0$
- We have compared sign ($\theta > 0$ and $\theta < 0$) but any two sets would do, e.g. $\theta = 0$ and $\theta \neq 0$.

T-shirt sizes for Bayes Factors > 1; (Kass & Raftery 1995)

B	Evidential meaning
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
>150	very strong

Multiple testing

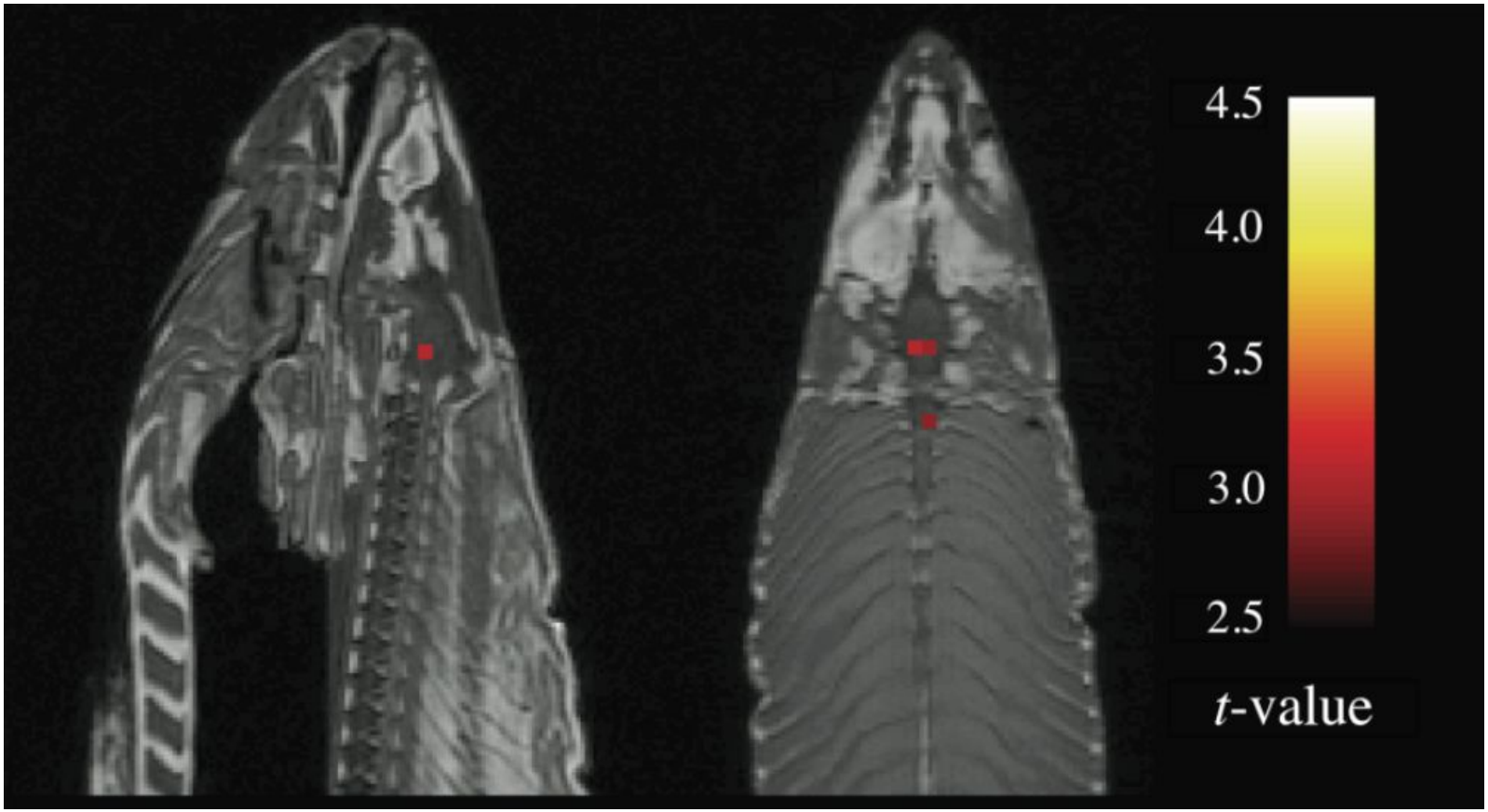
From one test to many;



Just finding "hits" is okay, as no-one will understand a "big"-dimensional posterior, and exact size of association (beyond positive/negative) doesn't matter.

Multiple testing

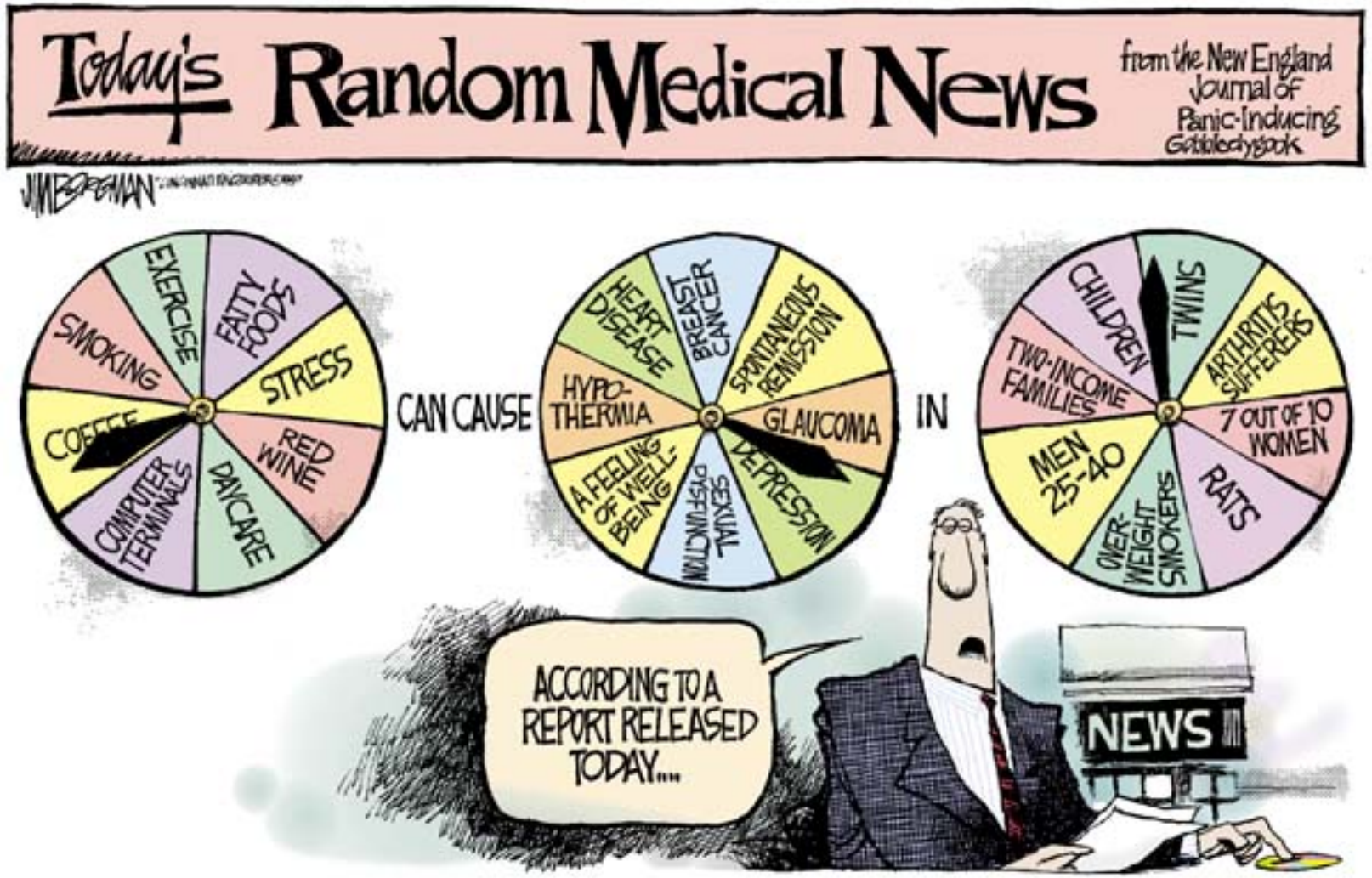
But big data has not always been covered in statistical glory;



Bennett *et al* exposed a salmon to two different stimuli, measuring brain activity in 8064 voxels. Standard methods show 16 differential-response 'hits'. Any problems?

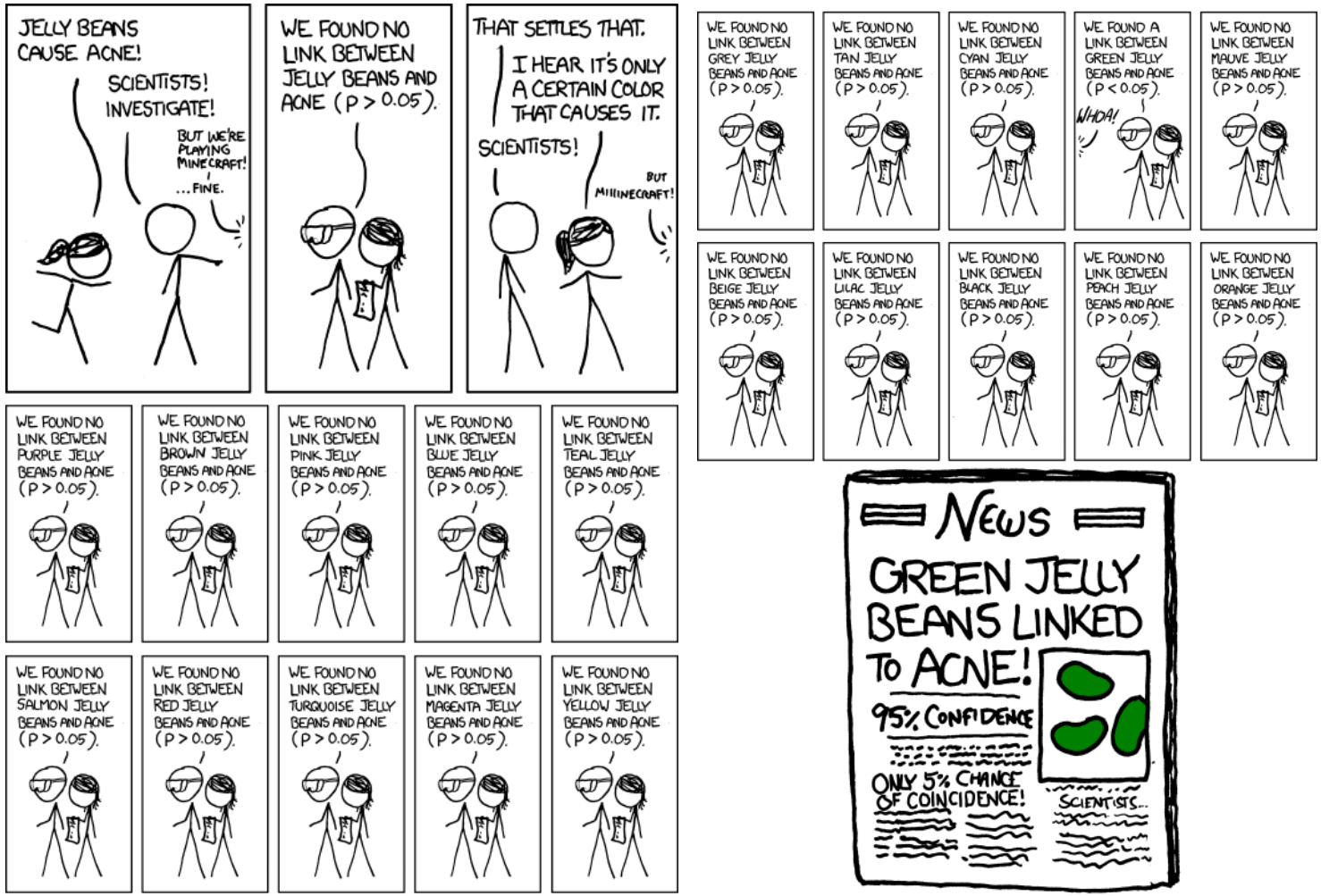
Multiple testing: background

...resulting in skepticism (and panic-inducing gobbledygook)



Multiple testing: background

And yes, XKCD knows about it;



Multiple testing: background

What statisticians should **do** with more than one test is an old problem;

The topic of multiple comparisons is sorely in need of clarification ... we do not really understand what its purpose is ... The statistical literature is full of [multiple testing] methods and techniques but quite devoid of a basic rationale and clearly stated purpose, and there still are many who doubt if the topic has any relevance at all.

K Ruben Gabriel, JASA 73:363 1978

In my view multiple comparison methods have no place at all in the interpretation of data.

John Nelder, JRSSB, 1971 33, 244–246
Re-iterated (!) in JRSSD, 1999 48, 257–269

Multiple testing: background

Another old-timer, in olden-times;

The theoretical basis for advocating routine adjustment for multiple comparisons is the 'universal null hypothesis' that 'chance' serves as the first-order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations ... Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings.

Ken Rothman
No adjustments are needed for multiple comparisons
Epidemiology 1990, 1:43–6

But with no penalty for leads being wrong, logically we have to investigate *everything*. In highly-restricted settings one can do this – e.g. small factorial designs – but that's all.

Multiple testing: background

Genetic epidemiology to the rescue!

The emergence of genetic epidemiology, with its staggering number of associations to explore, has brought multiple-inference concepts into the mainstream of epidemiology and biostatistics.

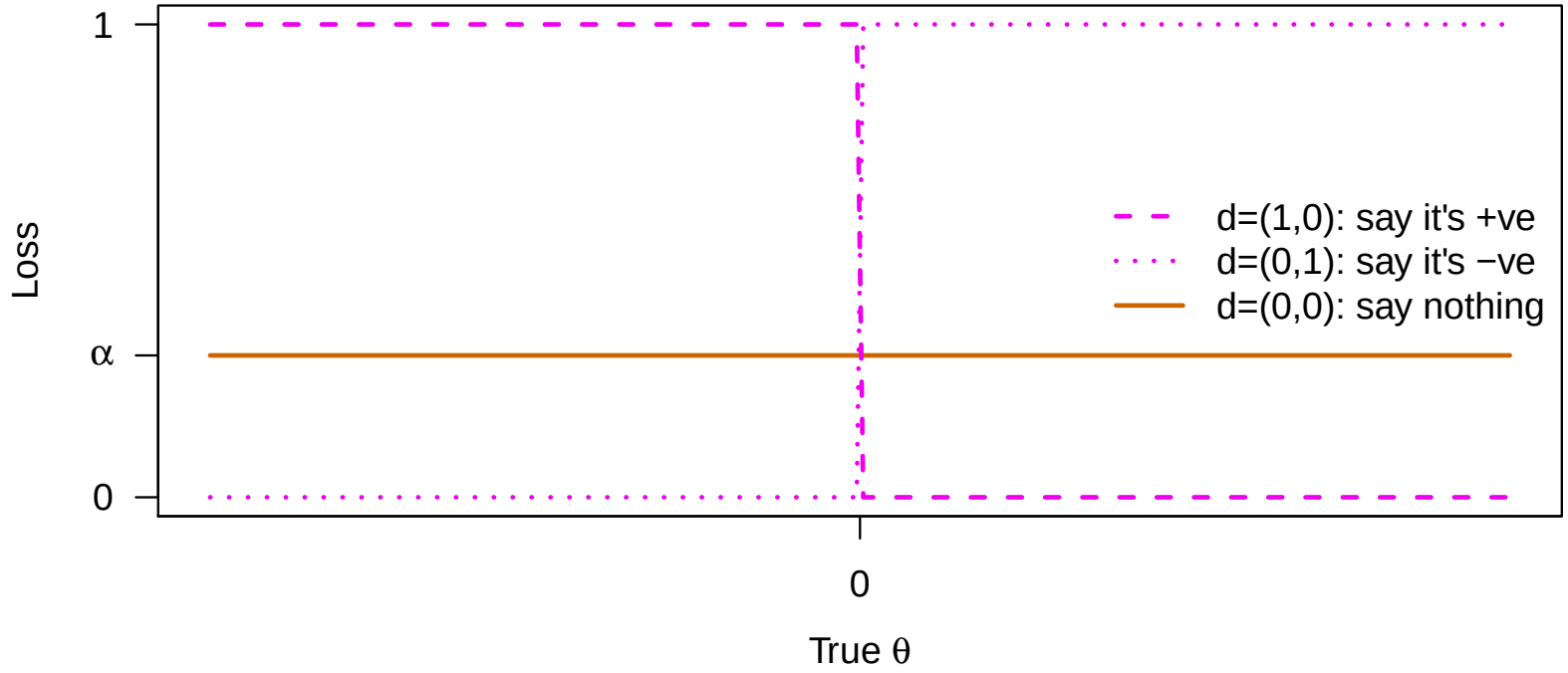
It is thus time to recognize of the extent of multiple comparison problems in everyday epidemiology and deploy modern methods toward their resolution.

Sander Greenland (discussing Jon's work)
International Journal of Epidemiology 2008;37:430–434

Rothman & Greenland are co-editors of a very popular Epidemiology textbook. In the latest edition (2008) Rothman has *considerably* moderated his earlier views.

Multiple testing: many decisions

Back to deciding the sign of a single θ;



Written in terms of indicator functions, this is

$$L(\theta, d) = \alpha 1_{\{\text{say nothing}\}} + 1_{\{\text{say something, wrong sign}\}},$$

which emphasises α is a *tradeoff rate*; how much cheaper is it to say nothing than to get the wrong sign?

Multiple testing: many decisions

Just as we combining two one-sided tests, for testing m multiple parameters θ_i, we can add the loss functions;

$$L(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_i^m L(\theta_i, d_i) \;\; = \;\; \sum_{i=1}^m \alpha_i 1_{\{\text{say nothing about } \theta_i\}} + \sum_{i=1}^m \; 1_{\{\text{say something, wrong sign for } \theta_i\}}.$$

But what α_i to use? Again following two-sided tests, suppose we want to have a ratio of at least α between the loss saying nothing about any θ_i, and the loss making m decisions but getting some wrong. To do this, we need to consider

Decision	Loss	Situation
$d = (0, 0, ..., 0)$	$\sum_{i=1}^m \alpha_i$	Say nothing about any θ_i
$d = (1, 1, ..., 1)$	1	Get wrong sign for one θ_i

... and we see using $\sum_{i=1}^m \alpha_i = \alpha$ gives this $\alpha : 1$ ratio.

Multiple testing: many decisions

Or, if you prefer, just have loss function

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{d}) &= \alpha \text{Prop(non-decisions, out of } m) + \#\{\text{wrong signs}\} \\ &= \alpha/m \#\{\text{non-decisions}\} + \#\{\text{wrong signs}\} \\ &= \sum_{i=1}^{m} \alpha_i 1_{\{\text{say nothing about } \theta_i\}} \\ &\quad + \sum_{i=1}^{m} 1_{\{\text{say something, wrong sign for } \theta_i\}}, \end{aligned}$$

if we use $\alpha_i = \alpha/m$.

- This is a conservative criterion – trading off an average against a sum
- Frequentist version of using $\alpha_i = \alpha/m$ is *Bonferroni correction* (see Session 3) which controls *Family-wise Error Rate* at level α (more later)
- Using loss functions, Bonferroni is exactly the right answer

Summary so far

- Multiple testing corrections are not unBayesian. But state carefully what they mean, and why they are used
- Does $L(\theta, d)$ reflect Your loss? Maybe not, if each additional sign error is not equally bad...
- Faced with many studies on many phenotypes, frequentist ideas may help pick an α level