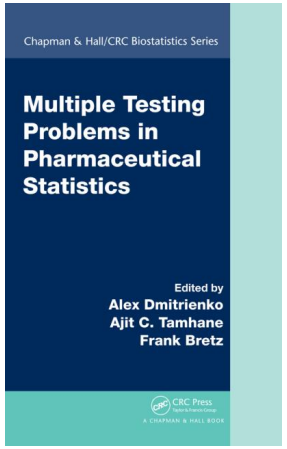
Bayesian Statistics for Genetics

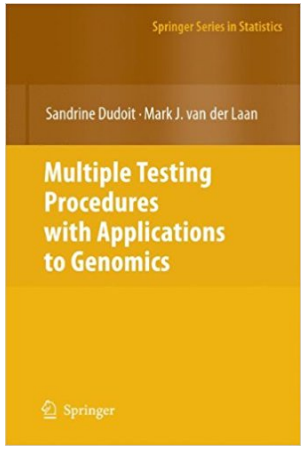
Lecture 9: Testing and Multiple Testing

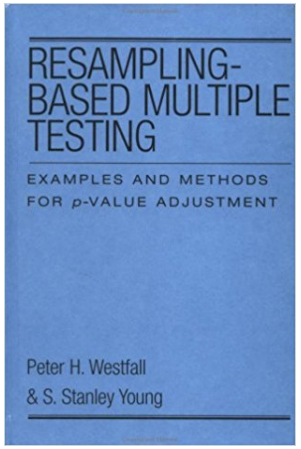
July, 2023

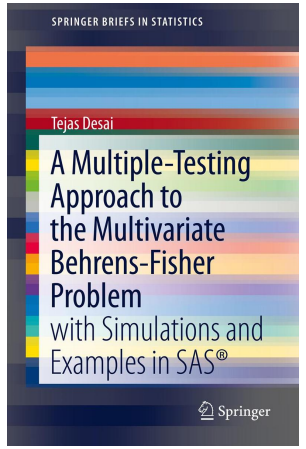
Overview

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









- More on Bayes Factors, for point null hypotheses
- Decision theory – how to calibrate
- Two-sided tests as optimal Bayes decisions
- Connections with FDR, and more

Bayes Factors, again

Recall the Bayes Factor for two models/hypotheses is

$$BF = \frac{\mathbb{P}[\,\boldsymbol{y}|H_0\,]}{\mathbb{P}[\,\boldsymbol{y}|H_1\,]} = \frac{\mathbb{P}[\,H_0|\boldsymbol{y}\,]}{\mathbb{P}[\,H_1|\boldsymbol{y}\,]} \Big/ \frac{\mathbb{P}[\,H_0\,]}{\mathbb{P}[\,H_1\,]}$$

Large BF values indicate support for the null.

- For one-sided tests (as we've seen) results are typically little different from using p-values
- With large samples/sane priors, posterior probability of the null $\approx$ p-value from a one-sided test. (Casella & Berger 1987).
- **But** particularly in high-throughput studies (e.g. GWAS) we don't want one-sided tests – just an indicator that 'something interesting is going on', i.e. that $\theta \neq 0$. Which hypotheses are low-hanging fruit, ready for further studies?

Bayes Factors, again

Testing in this way, it's natural to use *two-sided tests*, of hypotheses

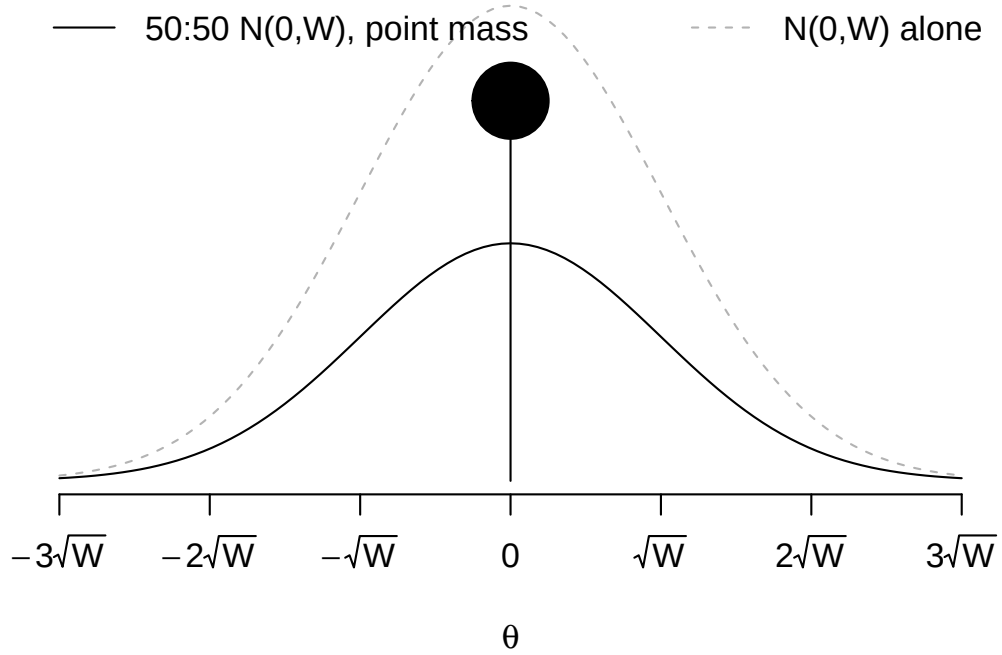
- $H_0 : \theta = 0$, i.e. **exactly** nothing going on
- $H_1 : \theta \neq 0$, i.e. **something** going on (but we're not saying what)

- Adapting the frequentist test is easy; just double the smaller p from two one-sided tests
- Or equivalently use $p < 0.025$ (not 0.05) as a threshold, i.e. $|Z| > 1.96$ (not 1.64) to identify the significant results

Warning: No such neat relationship holds between the Bayes Factors used in one-sided and two-sided tests.

Bayes Factors, again

This may not be intuitive – but the one-sided version has a smooth prior, versus the two-sided's *lump and smear* — here with a $N(0, W)$ 'smear':part



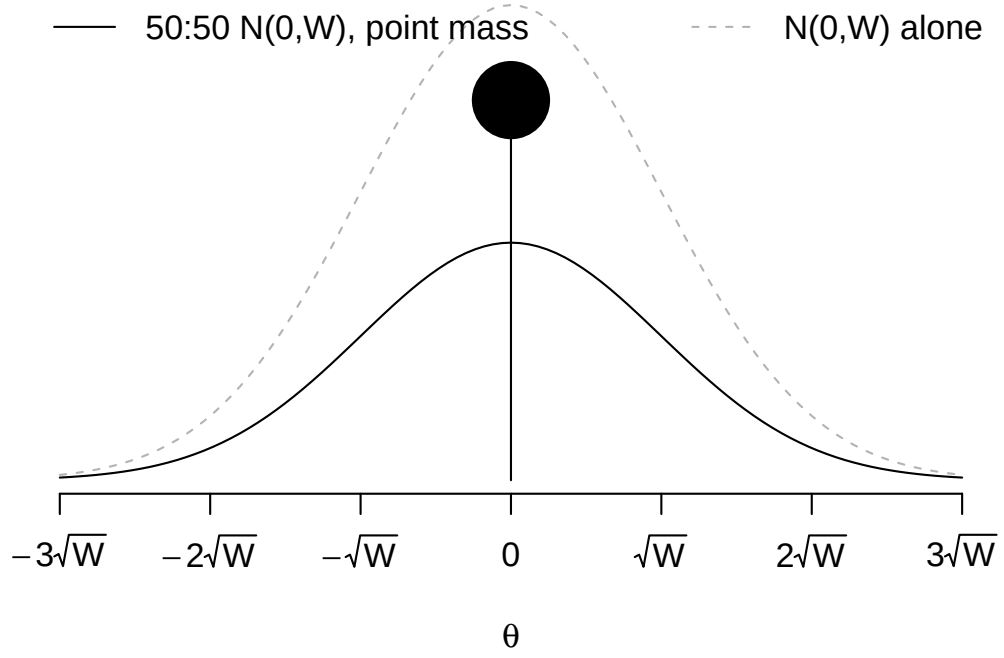
To a good approximation (Wakefield 2009), the Bayes Factor is

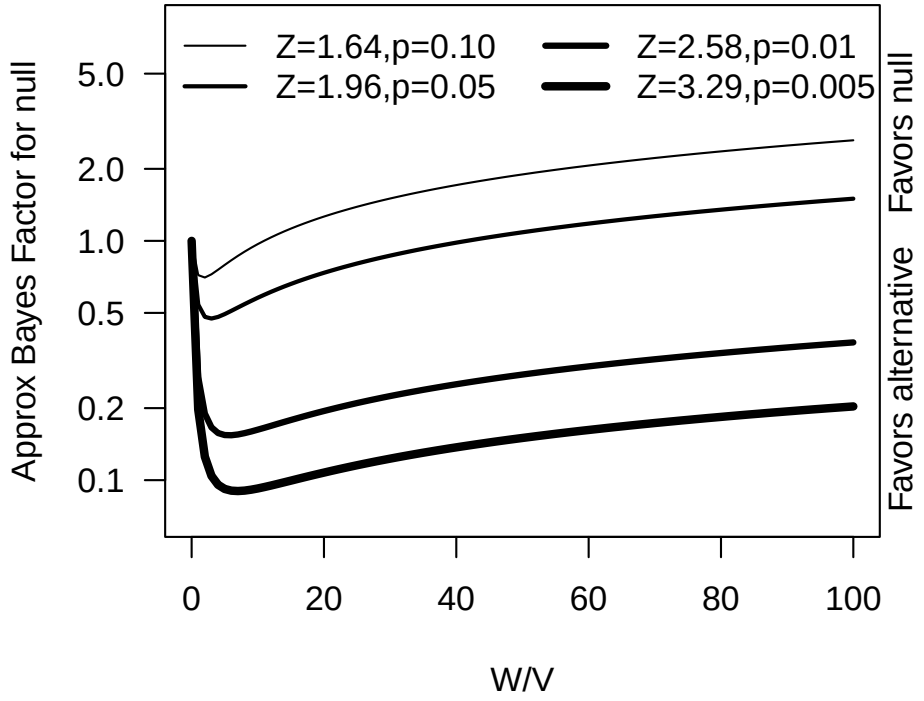
$$\sqrt{\frac{V+W}{V}}e^{\frac{Z^2}{2}\frac{W}{W+V}} = \sqrt{(1+W/V)}e^{-\frac{Z^2}{2}\frac{W/V}{1+W/V}},$$

where V is the large-sample variance estimate of $\widehat{\theta}_{MLE}$.

Bayes Factors, again

Making the prior more diffuse, eventually this happens:





- With W huge, any data we observe is massively unlikely under H_1, so the BF points **strongly** to H_0, **completely contradicting the classical test (!!!)**
- Known as the *Jeffreys-Lindley paradox*. BFs are **sensitive** to the 'smear' prior

Bayes Factors, again

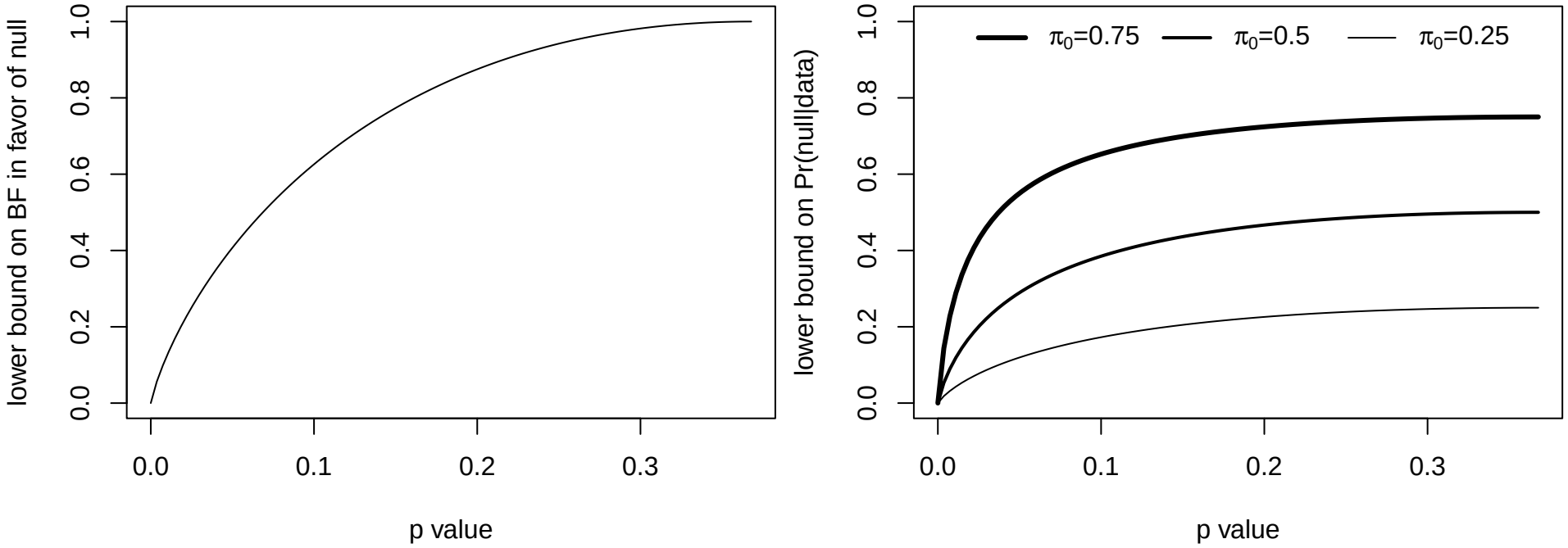
With $BF \approx \sqrt{(1+W/V)}e^{-\frac{Z^2}{2}\frac{W/V}{1+W/V}}$, we also see that the BF varies with n for fixed Z – because V shrinks with $1/n$

- BF fans can motivate them as classical test where α changes with n — not keeping $\alpha = 0.05$, or $\alpha = 5 \times 10^{-8}$. (Specifically, having α shrink with $1/\sqrt{n \log n}$ — see e.g. Wakefield (2009))
- Broadly, bigger studies do look for smaller effects. But it's hard to motivate any particular formula when effective n is due to e.g. imputation quality
- Conversely, Sellke *et al* (2001) use two-sided p-values in **lower bounds** on the BF and posterior probability of the null: (with prior $\mathbb{P}[\,H_0\,]$ denoted π_0)

$$\begin{aligned} BF &\geq -ep\log(p) \\ \mathbb{P}[\,H_0|\boldsymbol{y}\,] &\geq \frac{1}{1-\frac{1}{ep\log p}\times\frac{1-\pi_0}{\pi_0}}, \text{ for } p < 1/e \approx 0.368 \end{aligned}$$

Bayes Factors, again

Illustrating those bounds:



If you believe in a 'lump' at zero, a small p-value need **not** provide strong evidence to overwhelm that lump. This is one argument to redefine statistical significance as $p \leq 0.005$.

Decision theory

Decision theory is (formally) how **statisticians make decisions!**

The **decision** *of whether or not a vaccine is safe and effective, that is made by a completely independent group, not by the federal government, not by the company. It's made by an independent group of scientists, vaccinologists, ethicists,* **statisticians**.

How much worse do we believe **other** decisions are — those we *could have made*?

Decision theory

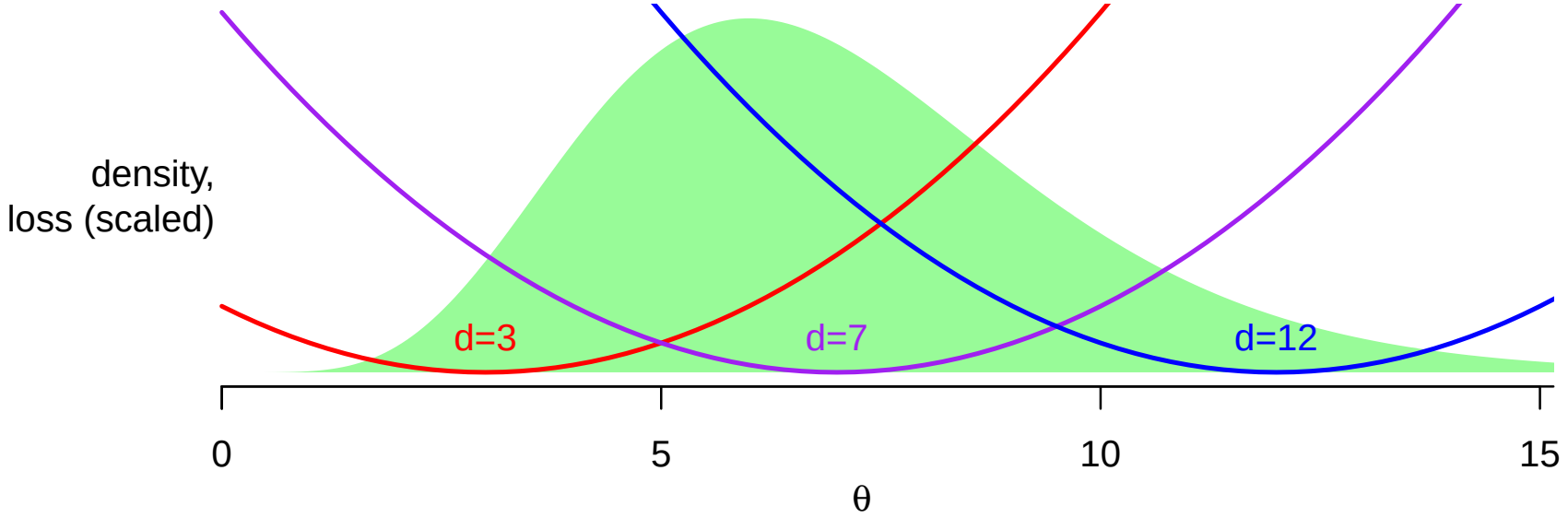
Extending our taxonomy:

- Prior distribution: statement of everything we know about θ **outside** of the current data
- Likelihood: statement of how plausible the observed data is under different values of θ
- Posterior distribution: updated prior, everything we know about θ including the current data
- **Loss function**: for true parameter value θ, **how bad it would be** if we make decision d

The costs of getting it wrong depend on d and θ, but **not** sample size, prior belief, etc.

Decision theory

Before we get to testing, for *quadratic loss* $(\theta - d)^2$, first compare some decisions:



- The **expected loss**, i.e. the loss averaged over our posterior uncertainty about θ, is $\mathbb{E}[\,(\theta - d)^2\,] = \mathrm{Var}[\,\theta\,] + (\mathbb{E}[\,\theta\,] - d)^2$
- The choice of d with smallest expected loss (the *Bayes rule*, i.e. best decision) is the posterior mean — so d=7, here
- With absolute loss $|\theta - d|$, the posterior median is the Bayes rule

Decision theory: for tests

To make it work for statistical tests, we borrow some nuance from 'Scots Law', which has *three* possible verdicts – guilty, not guilty and **not proven**:

How do the verdicts overlap with test-based decisions?

Verdict	Hypothesis test (Neyman-Pearson)	Significance test (Fisher)
Guilty	Reject H_0	Reject H_0
Not proven	no analog	No conclusion
Not guilty	Accept H_0	no analog

Decision theory: for tests

"Three-decision" problems (is $\theta > 0$? $\theta < 0$? not saying?) must have this loss:

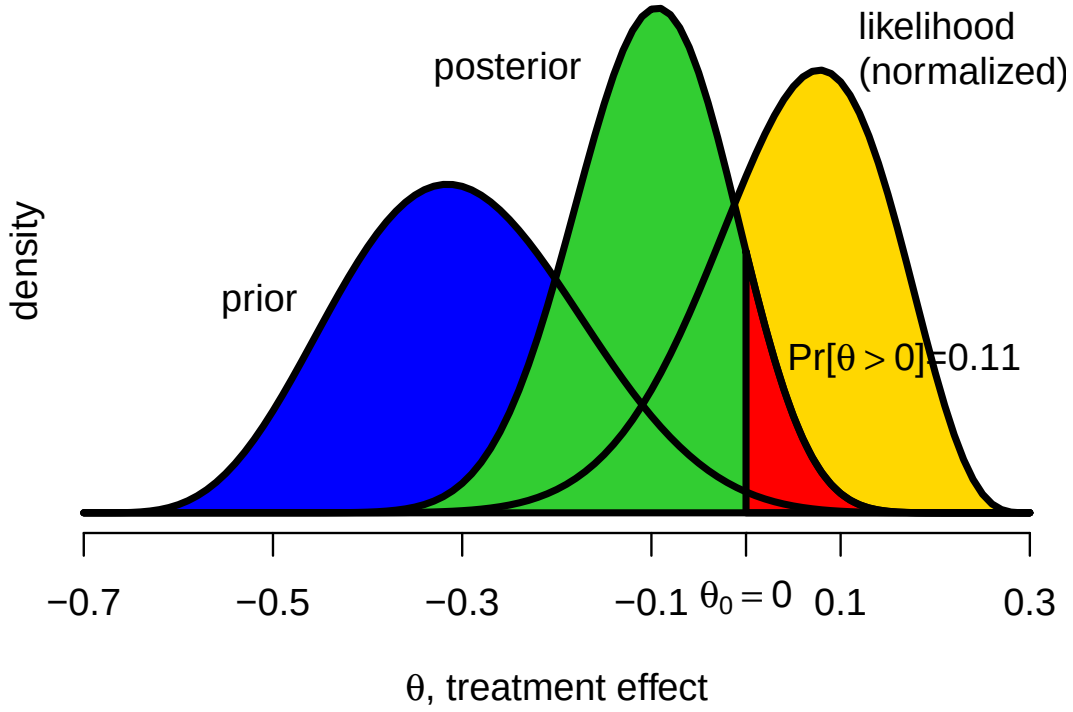
		Decision (what do we assert?)		
		Above	**No Decision**	Below
Loss when	$\theta > 0$	l_{TA}	l_{NA}	l_{FB}
	$\theta < 0$	l_{FA}	l_{NB}	l_{TB}

With any non-decision equally bad, coherence conditions & sign-symmetry, get;

		Decision		
		Above	No Decision	Below
Loss when	$\theta > 0$	0	$\alpha/2$	1
	$\theta < 0$	1	$\alpha/2$	0
Bayes rule: do this iff		$\mathbb{P}[\theta < 0] < \alpha/2$	Otherwise	$\mathbb{P}[\theta > 0] < \alpha/2$

... i.e. a Bayesian sided test — $\alpha/2$ is the **ratio of costs** for making **any** no-decision vs a **wrong** sign-decision. (See Rice *et al* (2020) for more.)

Three-decision problems: transparent example



- With $\alpha = 0.05$, sign errors are $\times$**40 worse** than making no decision
- ...so only make sign decision if $2\min(\mathbb{P}[\theta < 0], \mathbb{P}[\theta > 0]) < 0.05$.
- Making sign decisions around other θ_0 works similarly

Multiple decisions

Informally, we could write the sign-testing loss as

$$\text{Loss} = \frac{\alpha}{2} 1_{d=N} + 1_{\text{sign error}}$$

... where $\alpha < 1$ prevents us saying $d \neq N$ without even seeing the data.

For m multiple decisions, if we simply add loss functions for individual losses, i.e.

$$\text{Loss} = \sum_{j=1}^{m} \text{Loss}_j(\theta_j, d_j)$$

then overall Bayes rule $\mathbf{d}_B$ just collects the individual Bayes rules $\{d_{1B}, d_{2B}, ..., d_{mB}\}$.

This seems trivial* – but note that to account for multiple tests we **must**, somehow, say how one result affects how we value other results.

*But frequentist methods don't do it (!!!) Famously, under squared error losses and simple Normal locations $\theta_1, \theta_2, ...\theta_m$, then the sample mean $\bar{y}_1, \bar{y}_2, ..., \bar{y}_m$ is worse (on average) than estimates that shrink together the components. This is Stein's paradox.

Multiple sign tests: Bonferroni/FWER

For $j = 1, 2, ...m$ tests, we trade off the **sum** of the non-decision losses for a **single** sign error:

$$\text{Loss} = \sum_{j:d_j=N} \alpha_j/2 + 1_{\text{any sign error}}$$

- Must constrain $\sum_j \alpha_j < 1$, or would never decide all $d_j = N$
- With this constraint and symmetry wrt θ_j, set each $\alpha_j = \alpha/m$ for $\alpha < 1$. A (mildly) conservative approximation to the Bayes rule makes sign decisions iff

$$2 \min(\mathbb{P}[\,\theta < 0\,], \mathbb{P}[\,\theta > 0\,]) < \alpha/m$$

  ...i.e. **Bonferroni correction!**
- Classical Bonferroni correction uses $p < \alpha/m$ to control *family-wise error rate*, i.e. the $\mathbb{P}[$ any false positive $]$, at or below level α. FWER is a conservative criterion – its control by Bonferroni is usually mildly conservative

Multiple sign tests: Bonferroni/EFP

Alternatively: just add m copies of the 3-decision loss, with all $\alpha_j = \alpha/m$:

$$\text{Loss} = \frac{\alpha}{2m}\#\{\text{non-decisions}\} + \#\{\text{sign errors}\}$$

- Each θ_j in its own sign error/non-decision tradeoff
- Bonferroni-corrected 2-sided tests are the **exact** Bayes rule – not a conservative approximation
- Classical Bonferroni using $p < \alpha/m$ controls the *expected number of false positives* (EFP) at α – not very conservatively, and regardless of any correlation between the test statistics. (Gordon *et al* 2007)
- No automatic reason to constrain $\alpha < 1$, but EFP $\gg 1$ will usually be undesirable

Multiple sign tests: Benjamini-Hochberg/FDR

Lewis & Thayer (2009), in our notation, use

$$\text{Loss} = \underbrace{\frac{\#\{\text{sign errors}\}}{1 \vee \#\{\text{sign decisions}\}}}_{\text{Prop(wrong sign}|\text{decide sign)}} + \frac{\alpha}{2} \underbrace{\frac{\#\{\text{non-decisions}\}}{m}}_{\text{Prop(no decision}|\text{decision possible)}},$$

and a conservative approximation to the Bayes rule is a *step up procedure*: ordering by smaller tail area, keep making signs until $2\times$ tail areas exceeds $\alpha j/m$

This is a Bayesian analog of the famous Benjamini-Hochberg algorithm, that rejects ordered p-values until $p_{[j]} < \alpha j/m$, which controls the frequentist *False Discovery Rate*,

$$FDR = \mathbb{E}\left[\frac{\#\{\text{false positives}}{1 \vee \#\{\text{positives}\}}\right],$$

at pre-specified level α. (For 'nice' patterns of inter-test correlation)

Decision theory: lumps versus smears

When we have a lump and smear model, losses for decisions that $\theta = 0$ (exactly!) make more sense;

	Decision	
	Accept lump	Accept smear
True $\theta = 0$	0	L_1
True $\theta \neq 0$	L_2	0

We accept the alternative 'smear' if and only if

$$L_1 \mathbb{P}[\, H_0 | \boldsymbol{y} \,] < L_2 \mathbb{P}[\, H_1 | \boldsymbol{y} \,]$$

i.e. when the **posterior odds** of the alternative exceeds L_1/L_2

- If Type I errors are worse than Type II, $L_1 > L_2$ and this threshold is high
- The **relative** costs of Type I versus Type II errors determine the threshold; compare this to frequentist focus on controlling Type I error rate and **only then** worry about power, or equivalently Type II error rate.

Decision theory: lumps versus smears

For a given prior $\mathbb{P}[\theta = 0]$, the L_1/L_2 ratio can be turned into a threshold on the Bayes Factor. Alternatively use a *clone* parameter θ^* with the same prior as θ, **not** updated by the data, and use this loss:

		Decision on θ	
		Accept lump	Accept smear
$\theta^* = 0$	$\theta = 0$	l_{00}	l_{00}
	$\theta \neq 0$	L_2	0
$\theta^* \neq 0$	$\theta = 0$	0	L_1
	$\theta \neq 0$	l_{11}	l_{11}

We accept the alternative 'smear' if and only if

$$L_1\mathbb{P}[\,H_0|\boldsymbol{y}\,]\mathbb{P}[\,H_1\,] < L_2\mathbb{P}[\,H_1|\boldsymbol{y}\,]\mathbb{P}[\,H_0\,]$$

i.e. when the **Bayes Factor** in favor of H_1, i.e. $\frac{\mathbb{P}[\,H_1|\boldsymbol{y}\,]}{\mathbb{P}[\,H_0|\boldsymbol{y}\,]}/\frac{\mathbb{P}[\,H_1\,]}{\mathbb{P}[\,H_0\,]}$ exceeds L_1/L_2.

... so can calibrate BF via relative costs of Type I/II error **when true θ and clone θ^* disagree** – and if we **don't care about decisions when θ, θ^* agree**.

Summary

- Bayes provides various forms of tests: to choose between them, it helps to state how bad right/wrong answers would be
- There is some interplay between prior on θ and how we test ideas about θ: using sign tests makes less sense if $\theta = 0$ has a 'lump'
- Calibration of tests — and multiple tests — is easiest via ratios of (specific!) costs
- Yes, Bayesians may need to worry about multiple tests
- Ask 'which question are we answering?' and answer carefully!
- If no threshold can be agreed, report the summaries (plural) that make decisions possible, and don't *actually* do any tests