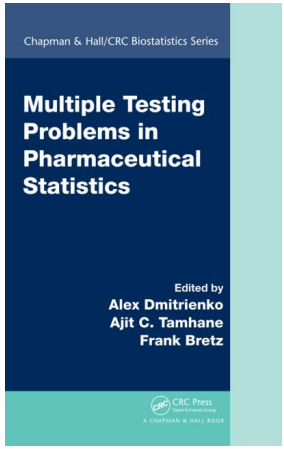
# Bayesian Statistics for Genetics

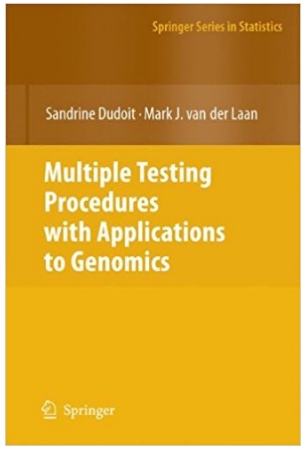
# Lecture 9: Testing and Multiple Testing

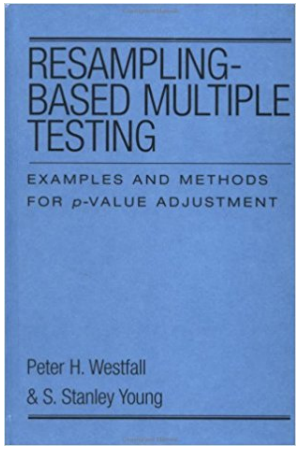
*June, 2025*

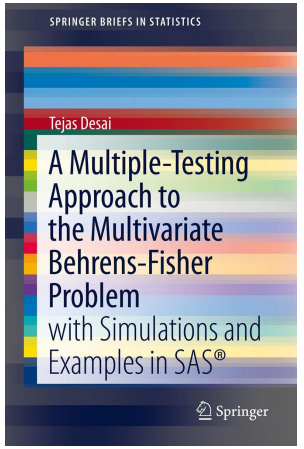
# Overview

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




- Testing as model selection
- 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

# Testing as model selection

Suppose we have *some* prior belief that a $\beta_j = 0$; a model allowing this specifies $\beta_j = z_j \times b_j$, where $z_j \in \{0, 1\}$ and $b_j \in \mathbb{R}$.

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i.$$

For example, in Session 4's FTO experiment,

$$\begin{aligned}
\mathbb{E}[\,Y|\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{z} = (1,0,1,0)\,] &= b_1 x_1 + b_3 x_3 \\
&= b_1 + b_3 \times \text{age} \\
\mathbb{E}[\,Y|\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{z} = (1,1,0,0)\,] &= b_1 x_1 + b_2 x_2 \\
&= b_1 + b_2 \times \text{group} \\
\mathbb{E}[\,Y|\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{z} = (1,1,1,0)\,] &= b_1 x_1 + b_2 x_2 + b_3 x_3 \\
&= b_1 + b_2 \times \text{group} + b_3 \times \text{age}.
\end{aligned}$$

Can **also** think of each value of $\boldsymbol{z} = (z_1, \ldots, z_p)$ representing a different model.

# Testing as model selection

But easier to implement thinking of $z_j$ as unknown components in one (big) model – written informally as;

$$\begin{aligned} z_j &\overset{i.i.d.}{\sim} Bern(0.5) \\ b_j &\sim p(b_j) \\ \epsilon_i &\overset{i.i.d.}{\sim} N(0, \sigma^2) \\ \sigma^2 &\sim p(\sigma^2) \\ y_i &= z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i \end{aligned}$$

Each of the $2^p$ possible values of of $\boldsymbol{z}$ has a posterior probability. (In the prior we treat them as a 'coin toss', equally likely to be 'in' or 'out'.)

# Bayesian model comparison

The posterior probability of the submodels is obtained from

$$p(\boldsymbol{z}|\boldsymbol{y},\mathbf{X}) = \frac{p(\boldsymbol{z})p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z})}{p(\boldsymbol{y}|\mathbf{X})}$$

To compare submodels $a$ and $b$, usually consider the odds of each, and how they compare:

$$\begin{array}{rcl} \dfrac{p(\boldsymbol{z}_a|\boldsymbol{y},\mathbf{X})}{p(\boldsymbol{z}_b|\boldsymbol{y},\mathbf{X})} & = & \dfrac{p(\boldsymbol{z}_a)}{p(\boldsymbol{z}_b)} \times \dfrac{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X},\boldsymbol{z}_b)} \\ \text{posterior odds} & = & \text{prior odds} \times \text{“Bayes factor”} \end{array}$$

Importantly, the Bayes Factor (BF) does not depend on the prior for $\boldsymbol{z}$ – so the ‘coin toss’ prior is not crucial for this approach.

# Parsimony

In the linear regression model, the formula for $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})$ is complex, but

$$\frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_a)}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{z}_b)} = (1+n)^{(p_{z_b} - p_{z_a})/2} \left(\frac{s^2_{z_a}}{s^2_{z_b}}\right)^{1/2} \times \left(\frac{s^2_{z_b} + \mathrm{SSR}_g^{z_b}}{s^2_{z_a} + \mathrm{SSR}_g^{z_a}}\right)^{(n+1)/2}.$$

where $\mathrm{SSR}_g$ denotes a form of sum of squared residuals.

So a model $\boldsymbol{z}_a$ is penalized if;

- it is too complex (number of covariates $p_A$ is large)
- it doesn't fit well ($\mathrm{SSR}_g^a$ is large)

# FTO example

$$\begin{aligned}\mathbb{E}[\,Y_i|\boldsymbol{\beta},\boldsymbol{x}_i\,] &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \\ &= \beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{grp}_i \times \text{age}_i\,.\end{aligned}$$

effect of group $\Leftrightarrow$ one of more of $\beta_2, \beta_4$ not zero

| $\boldsymbol{z}$ | model | $\log p(\boldsymbol{y}\|\mathbf{X},\boldsymbol{z})$ | $p(\boldsymbol{z}\|\boldsymbol{y},\mathbf{X})$ |
|---|---|---|---|
| (1,0,0,0) | $\beta_1$ | $-71.82$ | 0 |
| (1,1,0,0) | $\beta_1 + \beta_2 \times \text{grp}_i$ | $-70.04$ | 0 |
| (1,0,1,0) | $\beta_1 + \beta_3 \times \text{age}_i$ | $-67.04$ | 0 |
| (1,1,1,0) | $\beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i$ | $-61.19$ | 0.63 |
| (1,1,1,1) | $\beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{grp}_i \times \text{age}_i$ | $-61.72$ | 0.37 |

$$\begin{aligned}\mathbb{P}[\,\beta_2 \text{ or } \beta_4 \neq 0\,] &= 0.60 \\ \mathbb{P}[\,\beta_2 \text{ or } \beta_4 \neq 0|\boldsymbol{y},\mathbf{X}\,] &\approx 1\end{aligned}$$

# FTO example: using JAGS

Using the conjugate g-prior is a little artificial here;

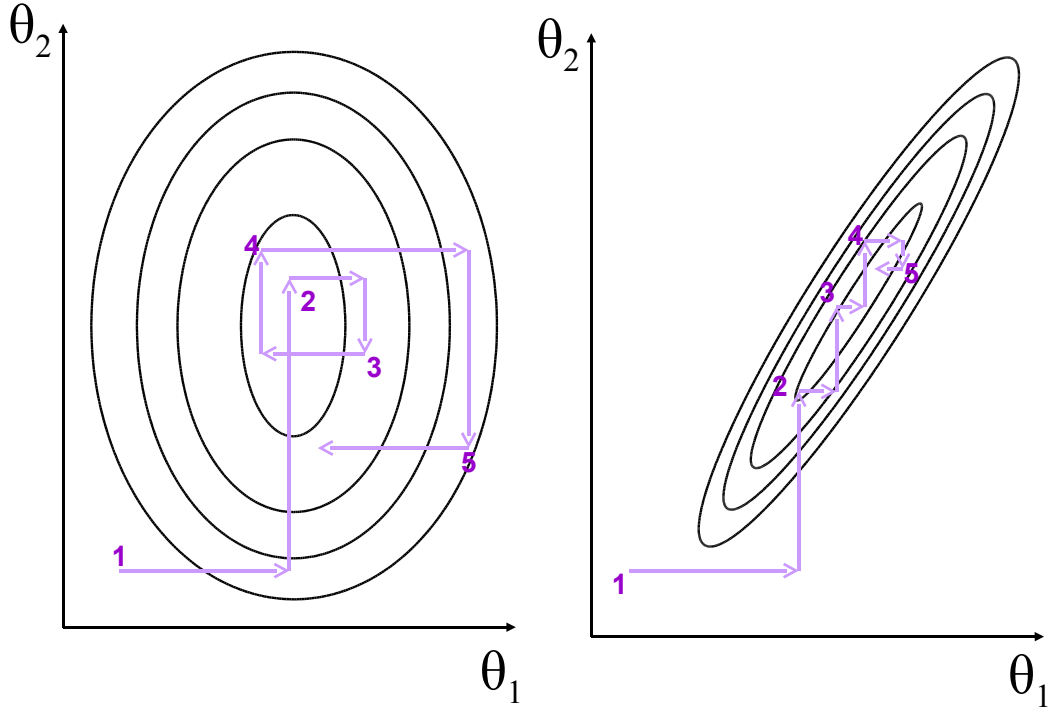
- Each sub-model has a prior that corresponds to one observation's information, but those observations are not the same.
- It's strange to support the model with all $\beta_j = 0$, i.e. where $\mathbb{E}[\, Y_i | x_i \,]$ is exactly zero for everyone

So we'll instead use a general-purpose Gibbs sampler for the same model, but with $z_1 = 1$ (forcing an intercept) and

$$\begin{aligned} z_j &\overset{i.i.d.}{\sim} Bern(0.5) \\ b_j &\sim N(0, 10), \text{ for } j = 2, 3, 4 \\ \epsilon_i &\overset{i.i.d.}{\sim} N(0, \sigma^2) \\ 1/\sigma^2 &\sim \Gamma(0.5, 1.839) \text{ ... as in Lec 4} \\ y_i &= z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i \end{aligned}$$

# Reminder: Gibbs sampler

For a couple of 2D examples; the same idea also works for binary parameters



# FTO example: using JAGS

Stan can't handle discrete parameters (yet?) so we'll use JAGS – *Just Another Gibbs Sampler*. The JAGS model and data:

```
library("rjags")
# first, write the model as to a text file
cat(file="linearprog2.txt", "model{
   for(j in 1:p){
      b[j]~dnorm(0, 0.1)   }
   z[1] <- 1  # fix the intercept to be in the model
   for(j in 2:p){
      z[j] ~ dbern(0.5)   }
inv.sigma2 ~ dgamma( 0.5, 1.839 )
   sigma <- sqrt(1/inv.sigma2)
   for(i in 1:n){
mu[i] <- x[i,1]*b[1]*z[1] + x[i,2]*b[2]*z[2] + x[i,3]*b[3]*z[3] + x[i,4]*b[4]*z[4]
y[i] ~ dnorm(mu[i], inv.sigma2) }
}")
# compile code based on model and data, then run chain
jags1 <- jags.model("linearprog2.txt", data=list(y=y,x=X, n=nrow(X), p=ncol(X)) )
update(jags1, 50000) # initial iteraions
```

# FTO example: using JAGS

And some of the output;

```
> jags1.out <- coda.samples(jags1, c("b","inv.sigma2", "z"), n.iter=100000)[[1]]
> summary(jags1.out)
Iterations = 50001:150000
Number of chains = 1
Sample size per chain = 1e+05
1. Empirical mean and standard deviation for each variable & std err of the mean:
             Mean      SD  Naive SE Time-series SE
b[1]       0.7593 1.26609 0.0040037      0.0184052
b[2]       1.2431 2.71152 0.0085746      0.0300475
b[3]       2.6202 0.39962 0.0012637      0.0057575
b[4]       2.1791 0.62138 0.0019650      0.0091990
inv.sigma2 0.2676 0.09069 0.0002868      0.0004338
z[1]       1.0000 0.00000 0.0000000      0.0000000
z[2]       0.5604 0.49634 0.0015696      0.0058886
z[3]       1.0000 0.00000 0.0000000      0.0000000
z[4]       0.9928 0.08431 0.0002666      0.0015052
```

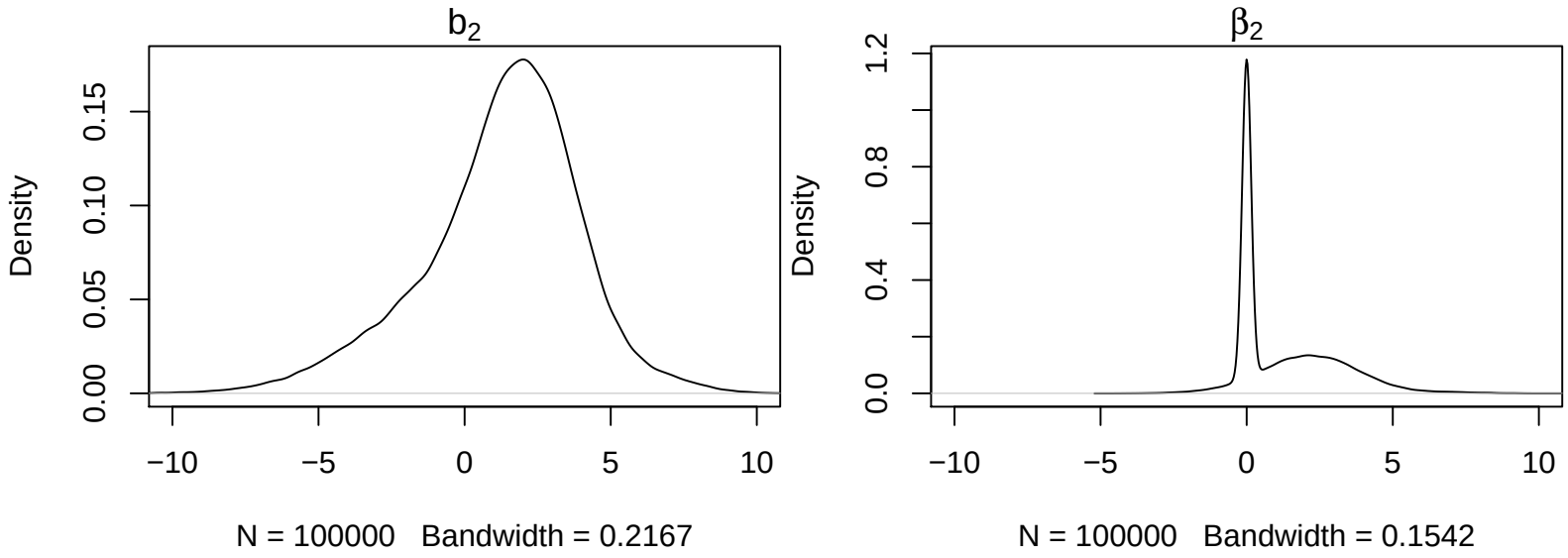
The coefficient of genotype is $\neq 0$ with 56% posterior support; the interaction term being $\neq 0$ has 99% support. The chain never moved from supporting age term $\neq 0$, so it has (approximately) 100% support.

# FTO example: using JAGS

All 100,000 steps in the chain are stored, so we can assess posterior for other terms – for example the support for each set of included/excluded variables;

```
> table(apply( jags1.out[,c("z[1]","z[2]","z[3]","z[4]")], 1, paste, collapse="") )/100000
   1011    1110    1111
0.43851 0.00693 0.55456
```

And comparing the posteriors for $b_2$ to the posterior to the actual genotype coefficient, $\beta_2 = b_2 \times z_2$;



# FTO example: using JAGS

To computing the Bayes Factor for whether any $\beta_j = b_j z_j = 0$;

- Compute compute $p_j = \mathbb{P}[\, b_j = 0\,]$ (which may be of interest on its own)
- Divide $p_j/(1-p_j)$ by prior odds of the null

Note it's straightforward to test multiple parameters, e.g. that $\beta_2 = \beta_4 = 0$ – just compute the relevant prior and posterior probabilities.

**But** this doesn't scale well with $p$, for tests that rely on the sampler exploring $2^p$ submodels. (Sensitivity to the prior on $b_j$ also a problem)

# FTO example: using JAGS

Using MCMC, we have to start the 'chain' somewhere – but this arbitrary choice shouldn't affect analysis, if we run the chains for long enough.

- After running long enough, the chains from any two starting points should *converge* to cover the posterior in the same way
- Less formally, after running long enough, chains forget where they started
- It's pragmatic (but not perfect) to run chains from a few different starting points, and check they give similar answers

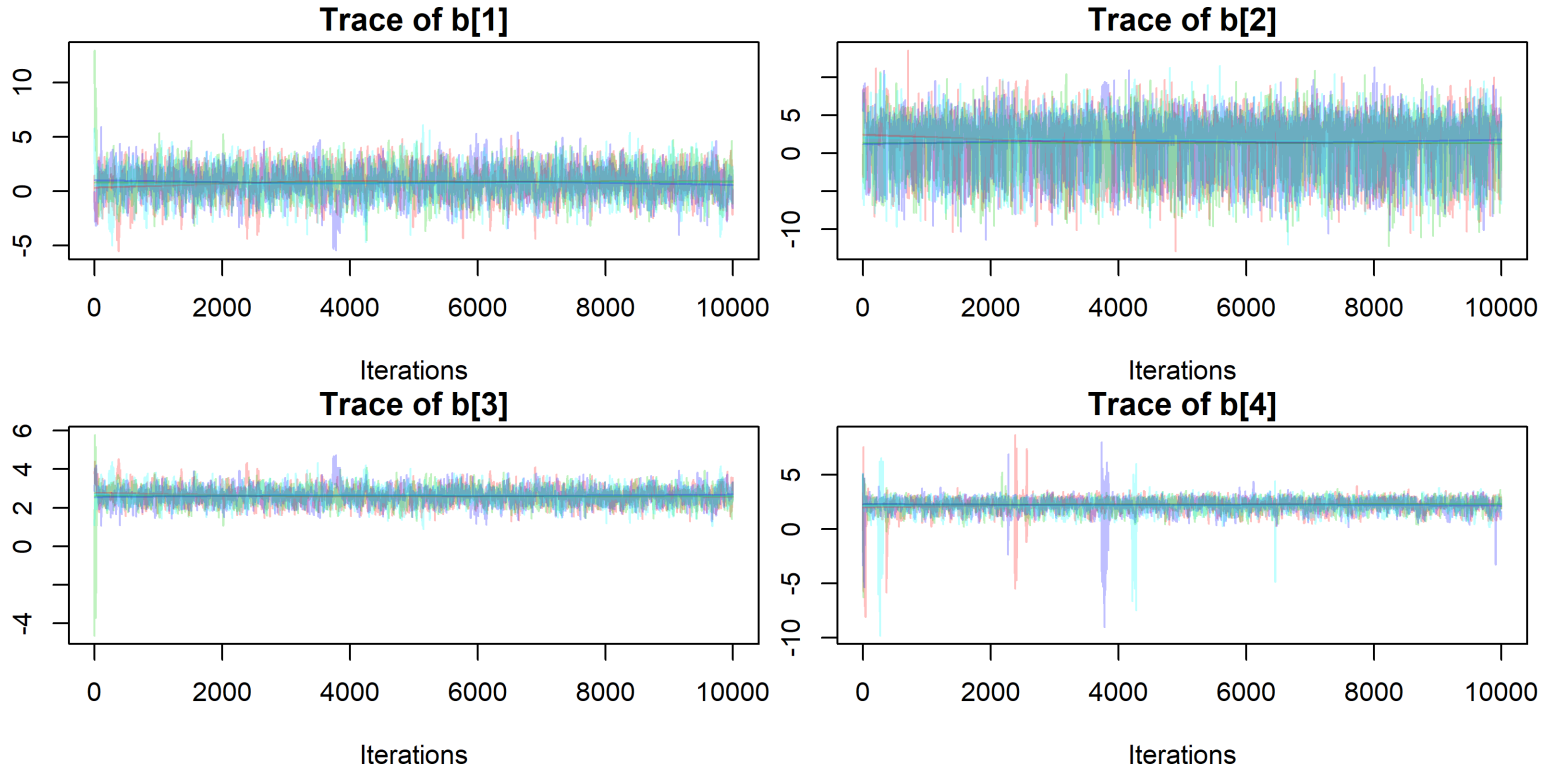
JAGS makes this fairly painless – here for 4 short chains;

```
set.seed(4)
inits1 <- list( b=rnorm(4,0,1),inv.sigma2=0.5,z=c(NA,0,1,0))
inits2 <- list( b=rnorm(4,0,1),inv.sigma2=0.5,z=c(NA,0,0,0))
inits3 <- list( b=rnorm(4,0,1),inv.sigma2=0.5,z=c(NA,1,1,0))
inits4 <- list( b=rnorm(4,0,1),inv.sigma2=0.5,z=c(NA,1,1,1))
jags2 <- jags.model("linearprog2.txt", data=list(y=y,x=X, n=nrow(X), p=ncol(X)),
inits=list( inits1, inits2, inits3, inits4), n.chains=4 )
jags2.out <- coda.samples(jags2, c("b"), n.iter=10000)
```

# FTO example: using JAGS

An informal way to check for convergence is to look for differences in each chain's traceplot; (no issues seen here)

```
plot(jags2.out, trace=TRUE, density=FALSE, auto.layout=FALSE, col=adjustcolor(2:5, alpha.f=0.25), lty=1)
```



# FTO example: using JAGS

To more formally check convergence of the chains for individual parameters, the *Gelman-Rubin diagnostic* compares within-chain variance ($W$) to between-chain variance ($B$), using tools from mixed models. For a converged chain their ratio $R = W/B$ should be $\approx 1$...

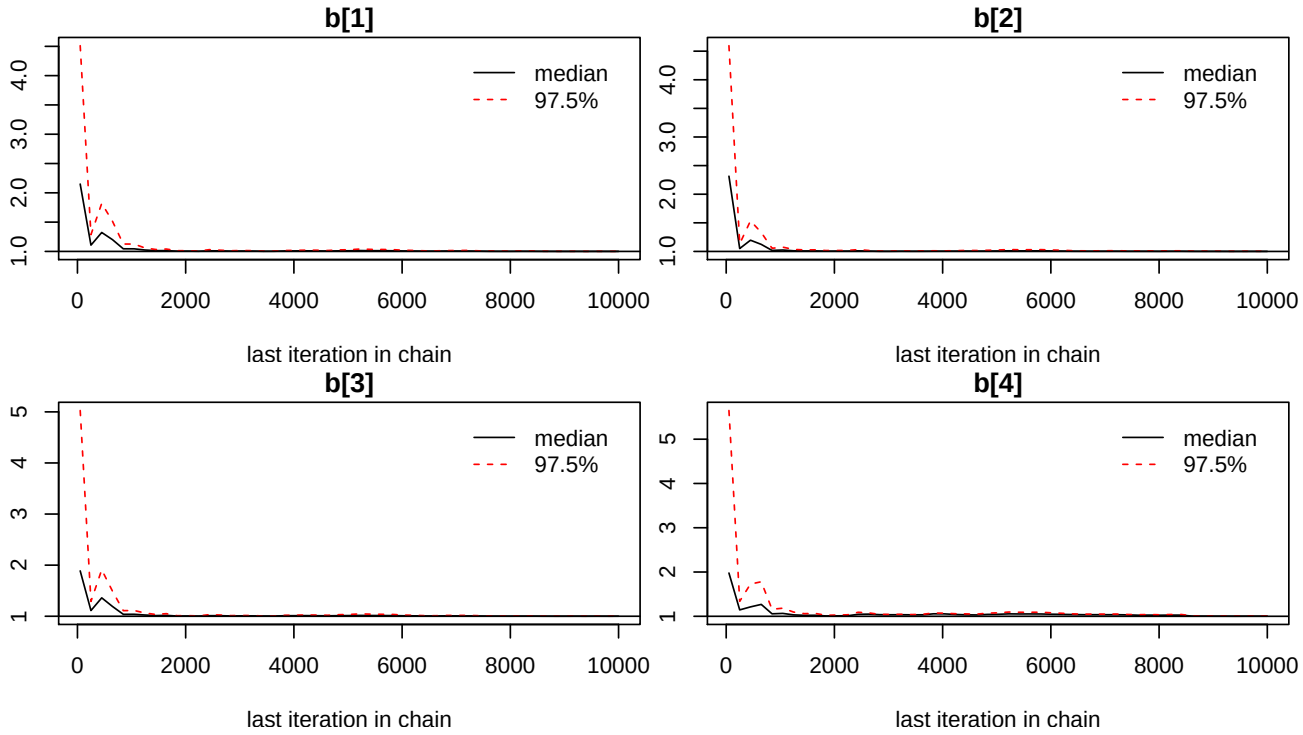
```
> gelman.diag(jags2.out)
Potential scale reduction factors:
     Point est. Upper C.I.
b[1]          1       1.00
b[2]          1       1.00
b[3]          1       1.00
b[4]          1       1.01
```

Similar ideas provide the effective sample size, i.e. roughly how big a simple random sample from the posterior is represented by the (auto-correlated) chain

```
> effectiveSize(jags2.out)
    b[1]     b[2]     b[3]     b[4]
1860.972 3057.044 1898.274 1586.170 # each from 40,000 iterations
```

# FTO example: using JAGS

`gelman.plot(jags2.out)` shows how $W/B$ evolves over iterations;




Ideally, don't start using the chain output until it looks like it converged – & even then, use as long a chain as you can manage. Thin it, if memory is an issue.

# 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 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

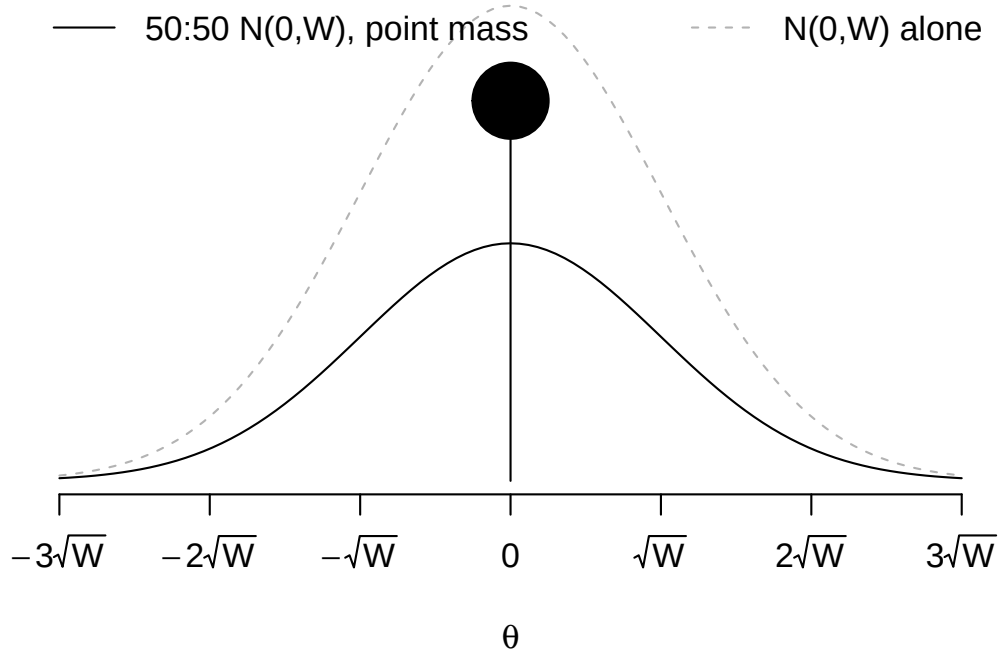
- $H_0 : \theta = 0$, i.e. **exactly** nothing going on
- $H_1 : \theta \neq 0$, i.e. **something** going on (but we're not saying what)

- 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



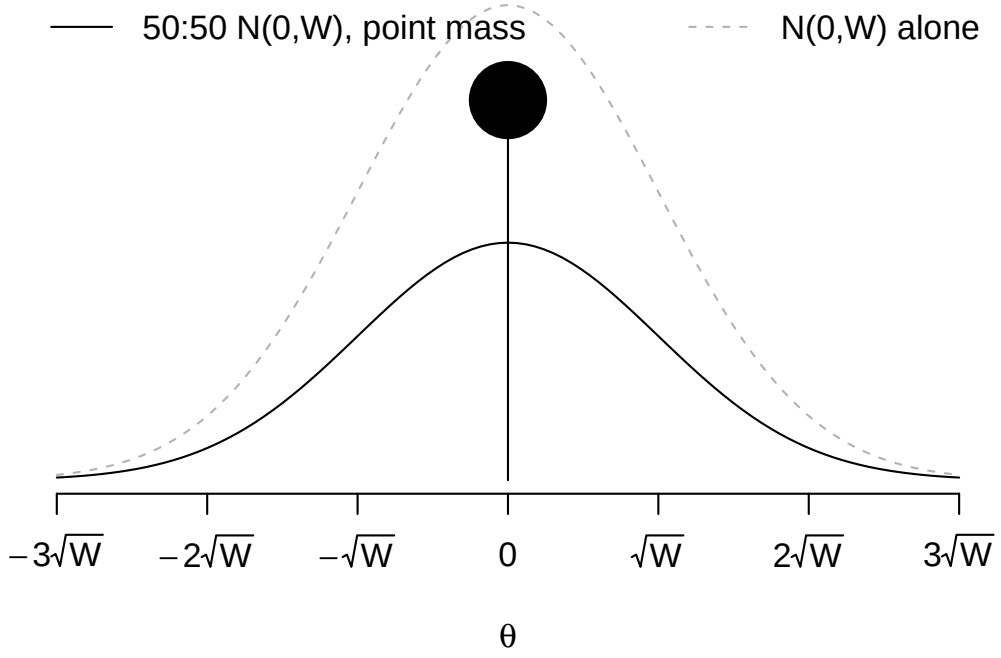

To a good approximation (Wakefield 2009), the Bayes Factor is

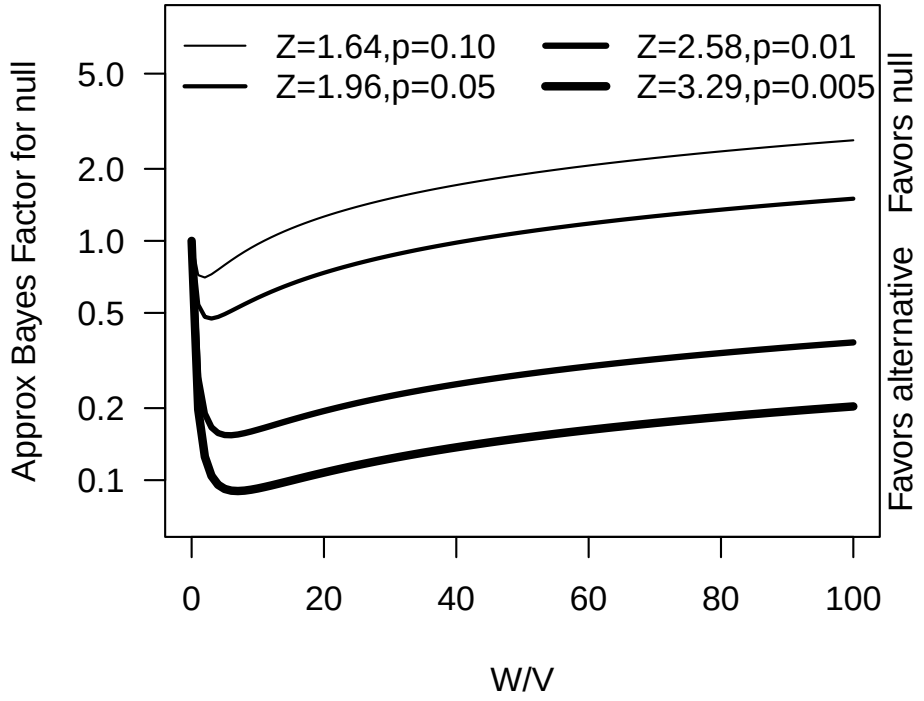
$$\sqrt{\frac{V+W}{V}}e^{\frac{Z^2}{2}\frac{W}{W+V}} = \sqrt{(1+W/V)}e^{-\frac{Z^2}{2}\frac{W/V}{1+W/V}},$$

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

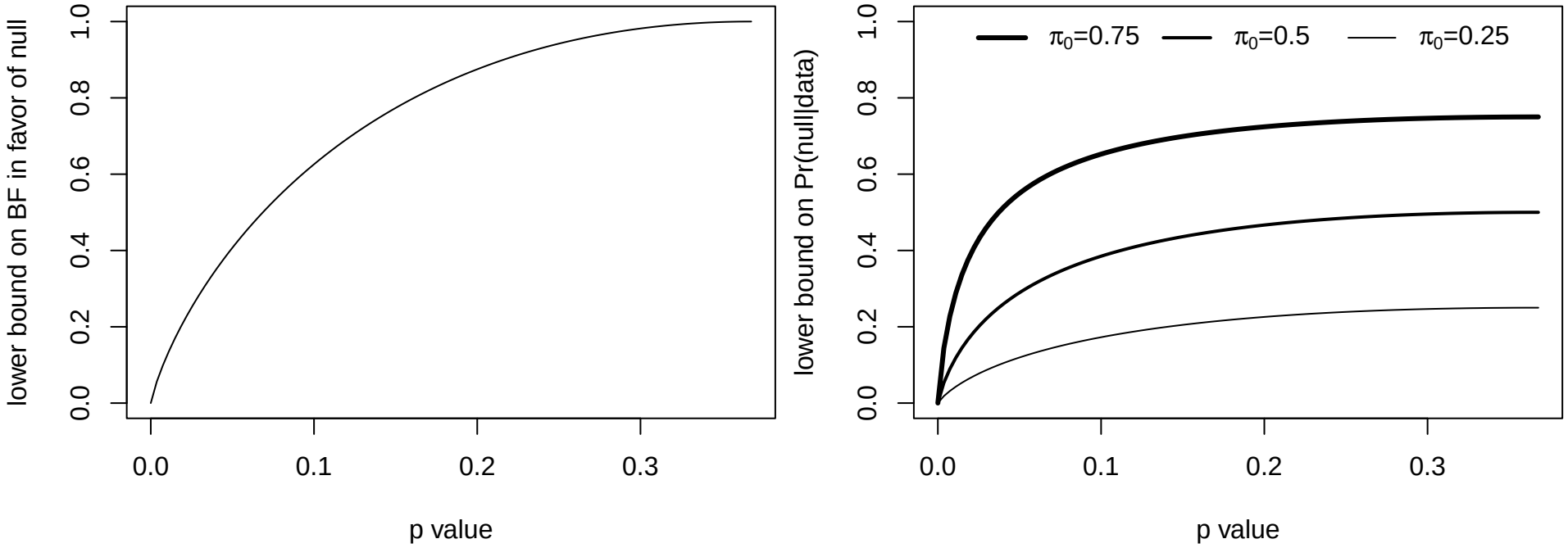
With $BF \approx \sqrt{(1+W/V)}e^{-\frac{Z^2}{2}\frac{W/V}{1+W/V}}$, we also see that the BF varies with $n$ for fixed $Z$ – because $V$ shrinks with $1/n$

- BF fans can motivate them as classical test where $\alpha$ changes with $n$ — not keeping $\alpha = 0.05$, or $\alpha = 5 \times 10^{-8}$. (Specifically, having $\alpha$ 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 $\pi_0$)

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

# 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

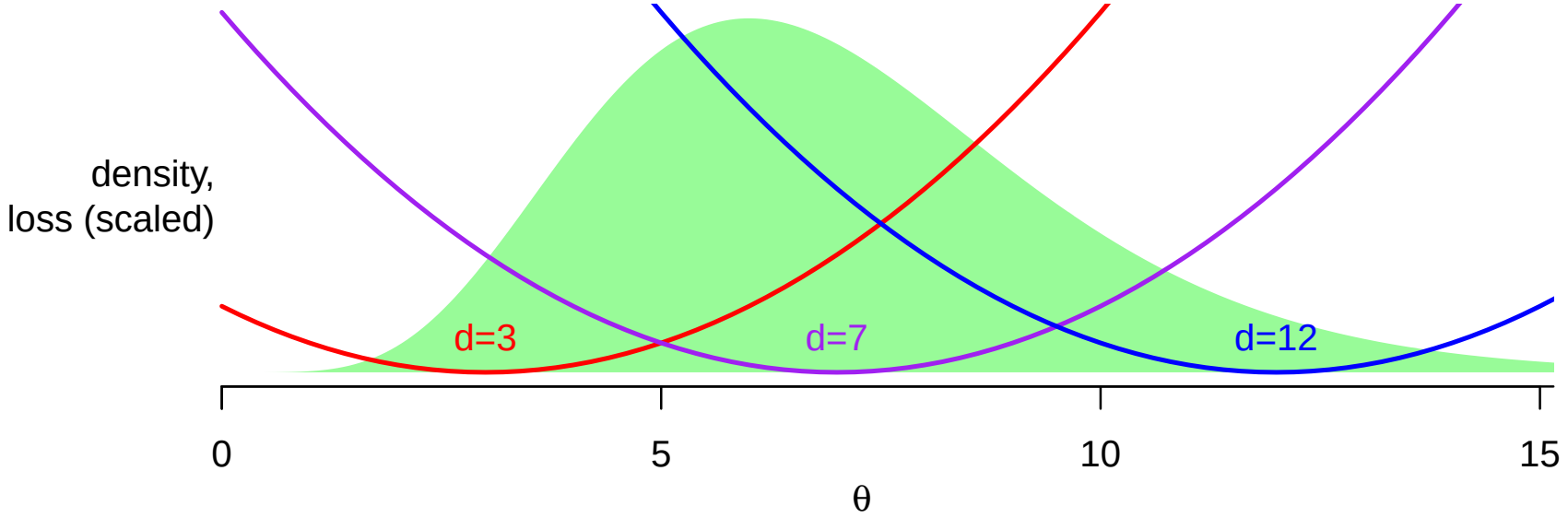
Extending our taxonomy:

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

The costs of getting it wrong depend on $d$ and $\theta$, 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 $\theta$, 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:

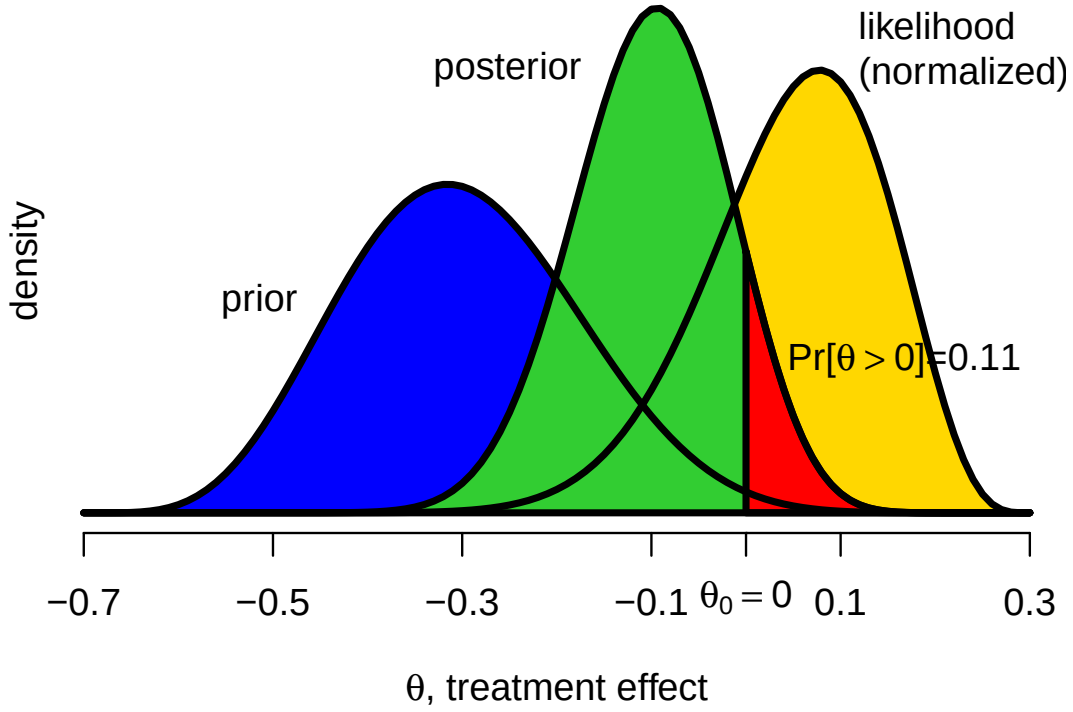
| | | Decision (what do we assert?) | | |
|---|---|---|---|---|
| | | Above | **No Decision** | Below |
| Loss when | $\theta > 0$ | $l_{TA}$ | $l_{NA}$ | $l_{FB}$ |
| | $\theta < 0$ | $l_{FA}$ | $l_{NB}$ | $l_{TB}$ |

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 $\theta_0$ 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 $\theta_j$, 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 $\alpha$. 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 $\theta_j$ 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 $\alpha$ – 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 $< 1$ is desirable in application where we don't expect to find overwhelming numbers of 'hits'

# 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 $\alpha$. (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 $\theta^*$ with the same prior as $\theta$, **not** updated by the data, and use this loss:

| | | Decision on $\theta$ | |
|---|---|---|---|
| | | 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 $\theta$ and clone $\theta^*$ disagree** – and if we **don't care about decisions when $\theta, \theta^*$ 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 $\theta$ and how we test ideas about $\theta$: 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