Bayesian Statistics for Genetics
Lecture 10a: Decision Theory

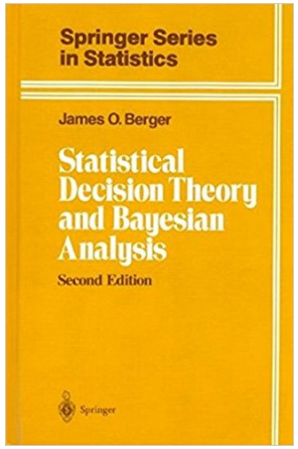
Ken Rice

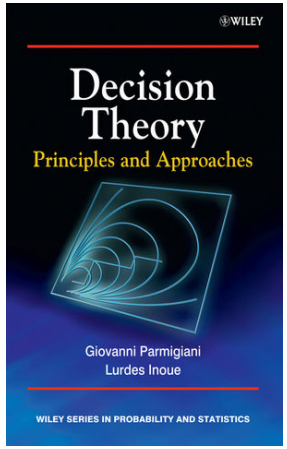
UW Dept of Biostatistics

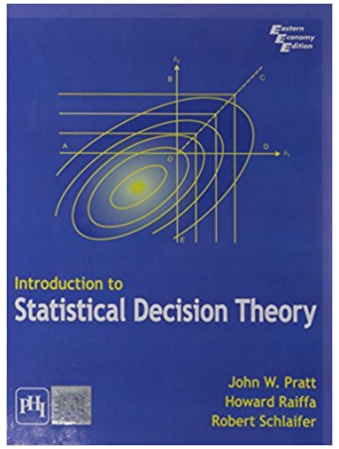
July, 2017

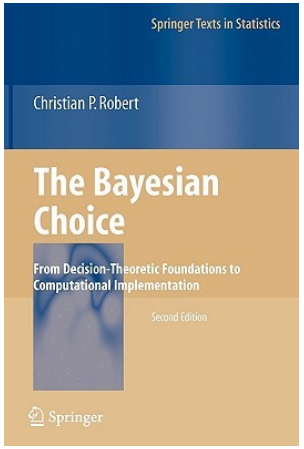
Overview

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









- Decision theory — saying what we're interested in
- Testing (again!) and how to motivate it as a decision problem

Note: while in genetics research, you may not have to make 'active' decisions when reporting results – e.g. who to treat – ideas from decision theory are helpful for deciding *which* posterior samples are useful, or not, for others.

Toolkit so far (& what it can't do)

Important tools/concepts;

- **Prior**: what You know about all parameters, external to study data
- **Likelihood:** what the data tells You about those parameters
- **Posterior:** what You know about the parameters, combining those resources
- **Model choice:** which sub-models You have more/most support for
- **Model checking:** how/whether the data and prior don't line up

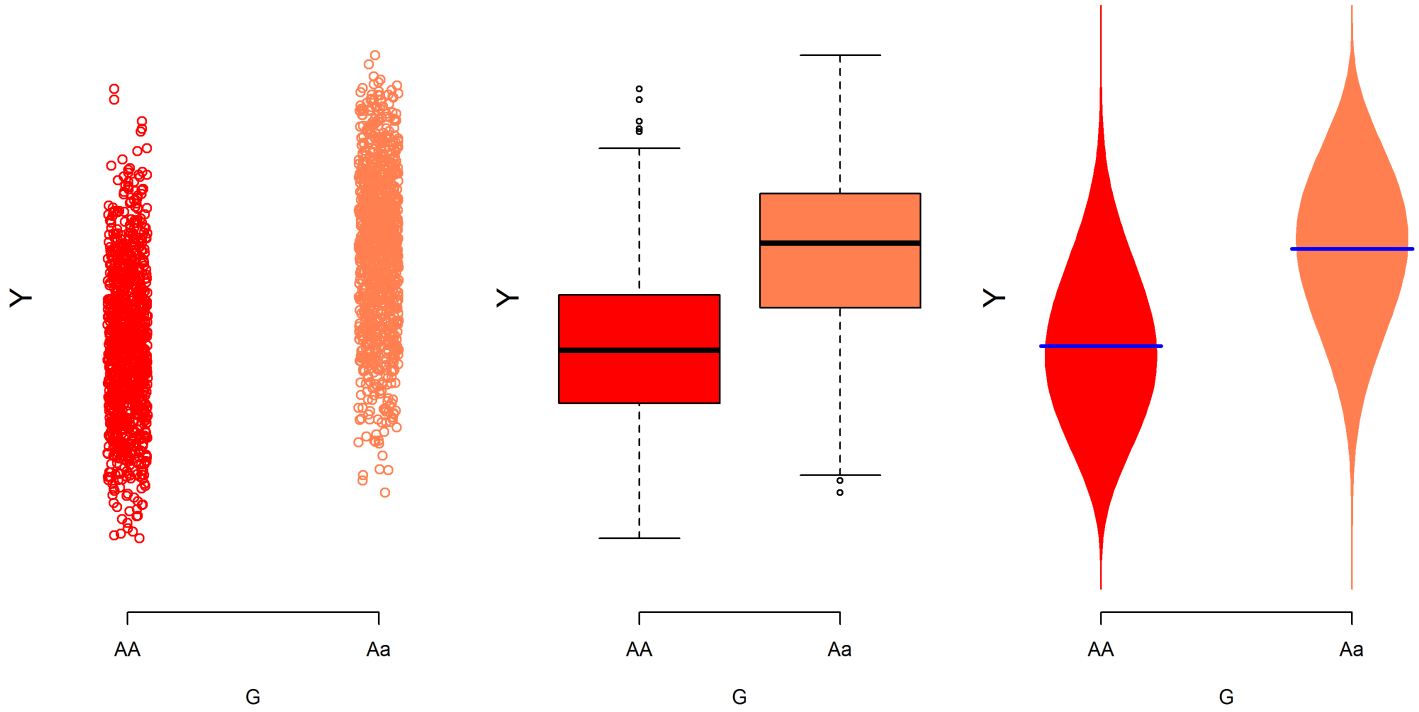
But these don't help us to answer;

- Which parameter(s) is/are of particular interest?
- How to choose summaries of Your knowledge about parameters?

These are common issues! For example **why** 95% intervals? **Why** the posterior mean? **Why** shrink some parameters to zero?

Example: inference for associations

Association means that the distribution of trait (e.g. Y=height) differs with genotype (SNP G=AA or Aa, or 0/1 copies of 'a')



- One measure is difference is the mean,
$$\beta = \mathbb{E}[\,Y|G = \text{Aa}\,] - \mathbb{E}[\,Y|G = \text{aa}\,]$$
- Interpret as difference in average Y in those with Aa vs AA
- But could also examine difference in median, i.e. difference in Y for 'average' AA people versus 'average' Aa people

Example: inference for associations

Other concerns about parameter choice, here;

- Is it causal? Confounding by ancestry a concern, also finding association at a variant in LD with the causal variant(s)
- Can it be estimated robustly-enough? (Means are very sensitive to extreme values, medians are not)
- Can it be estimated easily-enough? (Linear regression – which can be implemented quiakly – estimates means. Median regression is feasible but much less commonly-used)
- Will anyone else be able to understand it?

For now, we assume that $\beta = \mathbb{E}[\,Y|G = \mathsf{Aa}\,] - \mathbb{E}[\,Y|G = \mathsf{aa}\,]$ is a reasonable choice, and that we'll report its posterior.

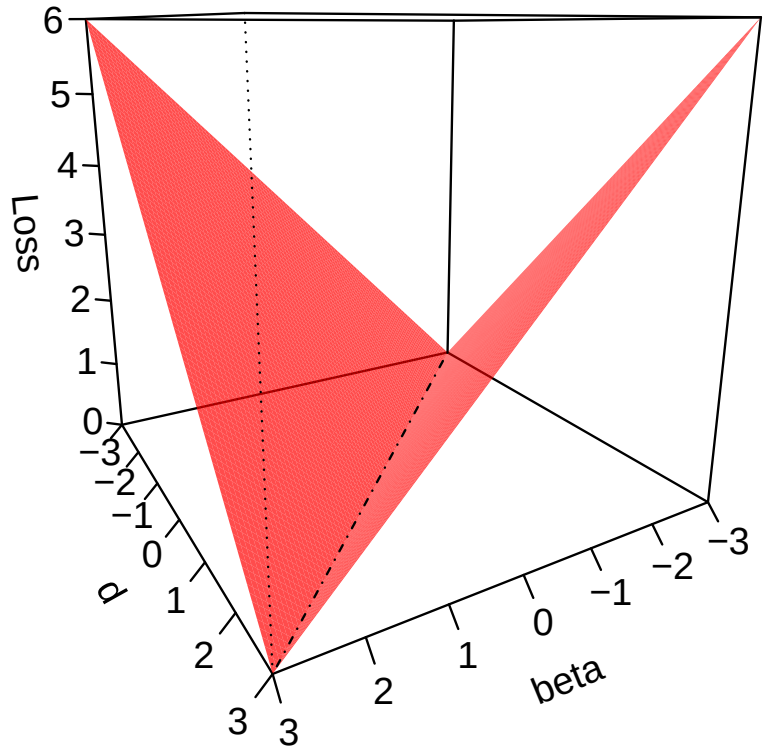
To address what might be a reasonable posterior summary of β, it's helpful to consider how bad each summary would be, if/when it's wrong.

Loss functions: intro

'How bad it would be' is captured by a loss function – how bad summary decision d would be, if the truth were β;

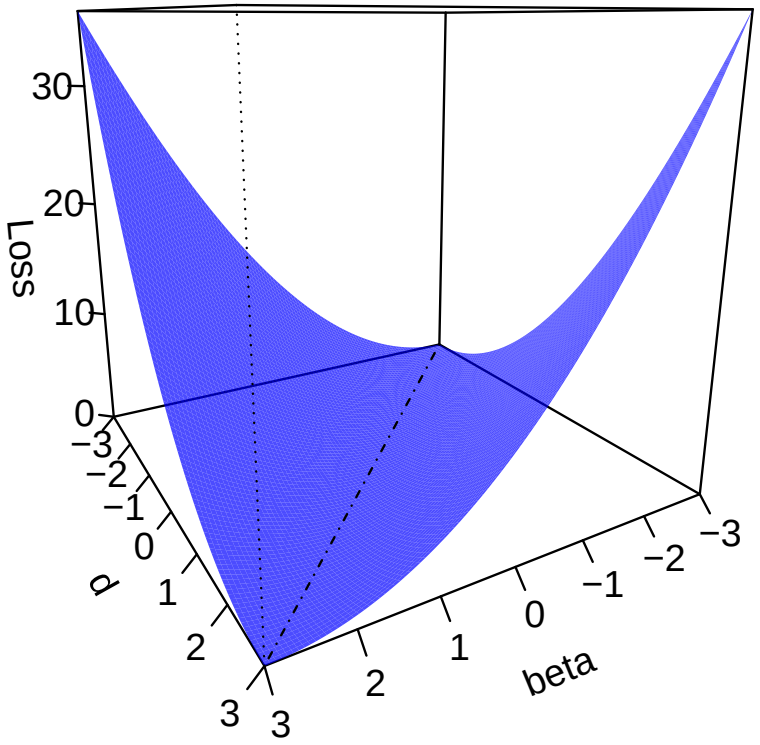
Absolute loss

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



Quadratic loss

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

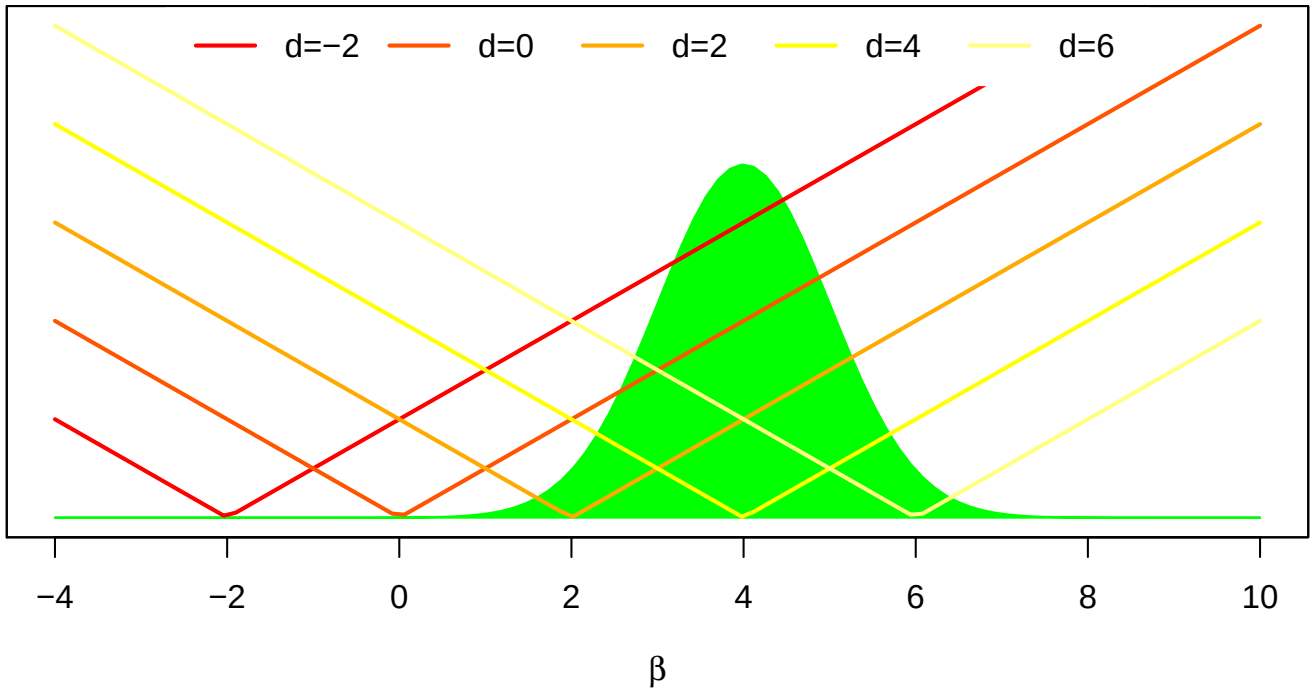


Loss functions: intro

We only get to choose d — and β is uncertain. Bayesian decision theory chooses d by minimizing the *expected posterior loss*;

$$d^B = \underset{d}{\text{argmin}}\, \mathbb{E}[\, L(\beta, d)|\mathbf{Y}\,].$$

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



This optimal decision d^B is called the *Bayes rule*.

Loss functions: intro

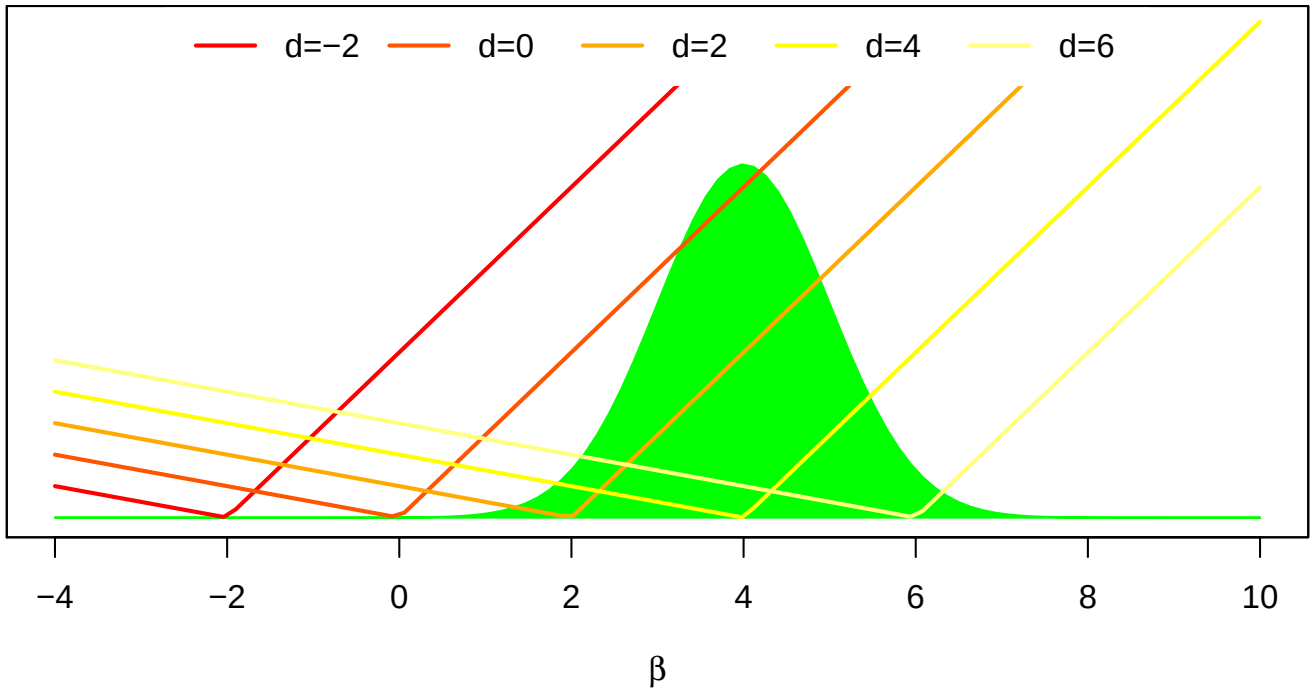
Why minimize the expected posterior loss?

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

If these seem unconvincing/esoteric, also note that any method of choosing between decisions will **have** to define and operate on $L(\beta, d)$, somehow – so it's useful to consider different choices.

Loss functions: intro

A more complex loss function; suppose under-estimates of β are worse than over-estimates – reasonable if you are predicting e.g. number of disease cases for resource planning.



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

Loss functions: intro

The math of this; (previous example assumes α close to 1)

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

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

- To set α, ask how much worse it would be to have e.g. one disease case with no resources, versus wasting resources by being prepared for one too many cases
- For α=1/2, the scaling is identical for over-estimates and under-estimates, so we're back to *absolute loss*, seen earlier
- More generally, scaling $L(\beta, d)$ by a positive constant doesn't affect the loss, neither does adding a constant – only **relative** losses matter
- (...neither does adding any function of β alone)

Loss functions: intro

For two decisions, we can add together two loss functions;

$$L(\beta, d_{\text{lo}}, d_{\text{hi}}) = \begin{cases} \frac{\alpha}{2}|\beta - d_{\text{lo}}|, & d_{\text{lo}} < \beta \\ (1 - \frac{\alpha}{2})|\beta - d_{\text{lo}}|, & d_{\text{lo}} > \beta \end{cases} + \begin{cases} (1 - \frac{\alpha}{2})|\beta - d_{\text{hi}}|, & d_{\text{hi}} < \beta \\ \frac{\alpha}{2}|\beta - d_{\text{hi}}|, & d_{\text{hi}} > \beta \end{cases}$$

and for $d_{\text{lo}} < d_{\text{hi}}$, this can be re-written as

$$L(\beta, d_{\text{lo}}, d_{\text{hi}}) = \frac{\alpha}{2}(d_{\text{hi}} - d_{\text{lo}}) + \begin{cases} |\beta - d_{\text{lo}}|, & d_{\text{lo}} > \beta \\ |\beta - d_{\text{hi}}|, & d_{\text{hi}} < \beta \end{cases}$$

... i.e. **trading off** the width of the interval for a costlier penalty for any distance to values of β outside the interval $(d_{\text{lo}}, d_{\text{hi}})$.

You may have guessed the Bayes rule already;

$$(d^B_{\text{lo}}, d^B_{\text{hi}}) = (\alpha/2, 1 - \alpha/2) \text{ posterior quantiles of } \beta$$

The tradeoff rate α/s justifies the level of the interval.

- Much more directly than considering replicate studies
- Also beats using $\alpha = 0.05$ just because everyone else does!

Which question?

As we've seen, decision theory forces us to look carefully at what our analysis is for – even beyond modeling and other prior assumptions, which only describe what the truth is, or might be.

Statistical tests can also be understood this way. Some possible tradeoffs, when considering whether $\theta = 0$, or $\theta < 0$ or $\theta > 0$;

- 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 (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

A reminder of the ingredients for decision theory;

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

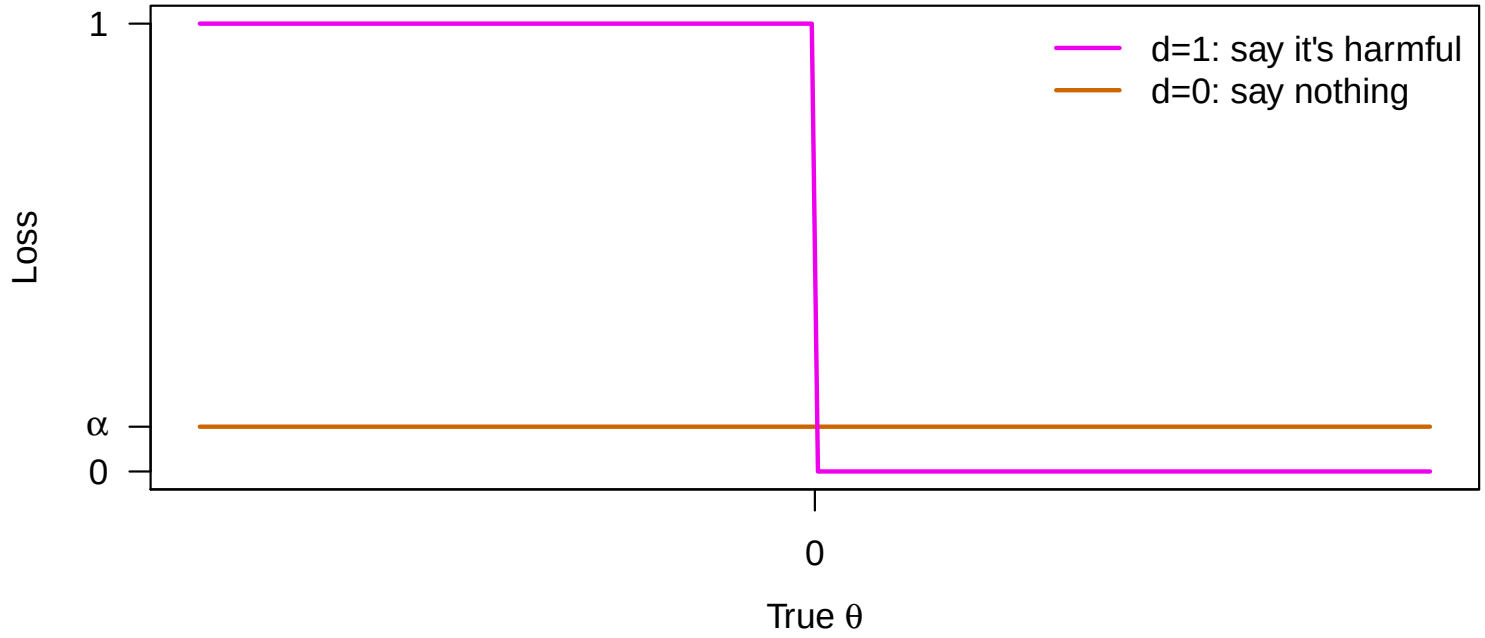
The Bayes rule is the decision d that minimizes $\mathbb{E}[\, L(\theta, d)\,]$ – but for testing, d is 0 or 1, so this means checking whether

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

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

Testing: first example

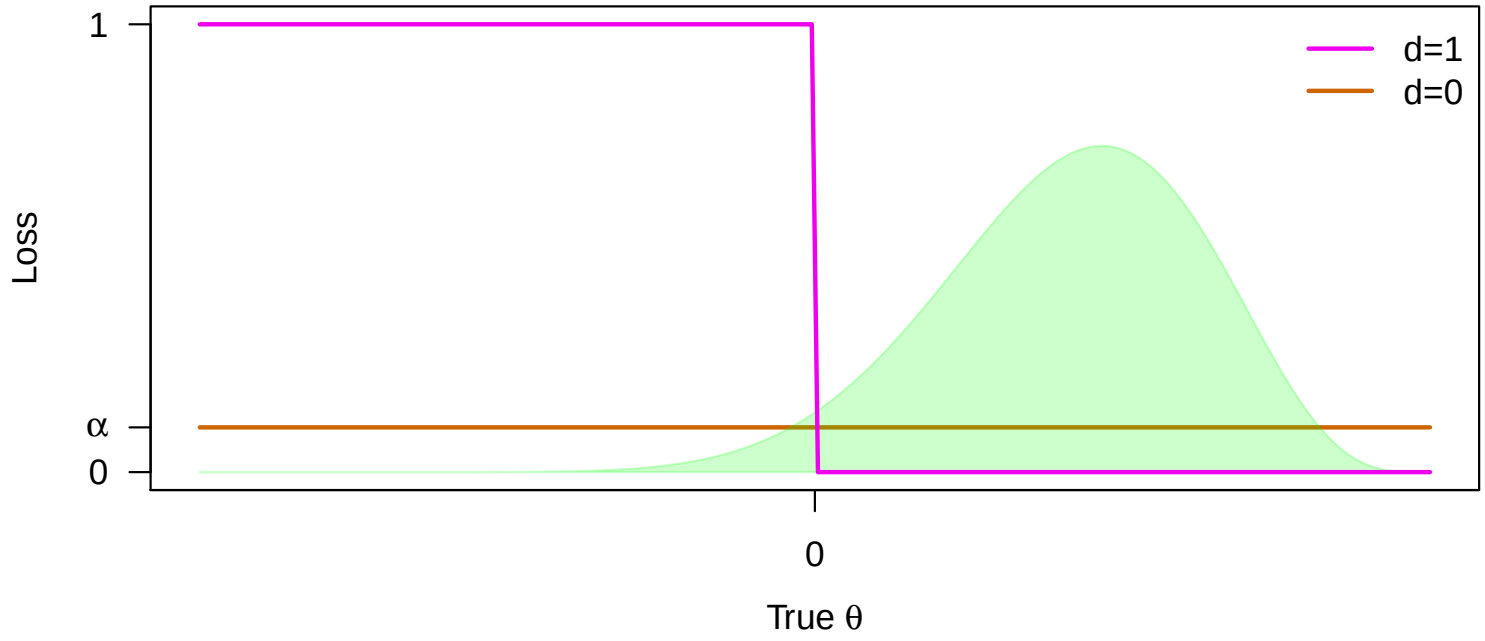
Suppose we are assessing whether a variant is harmful – it has $\theta > 0$ – suppose we either state ($d = 1$) that it is or say nothing at all ($d = 0$) about θ;



- $L(\theta, d = 1) = 1$ 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;



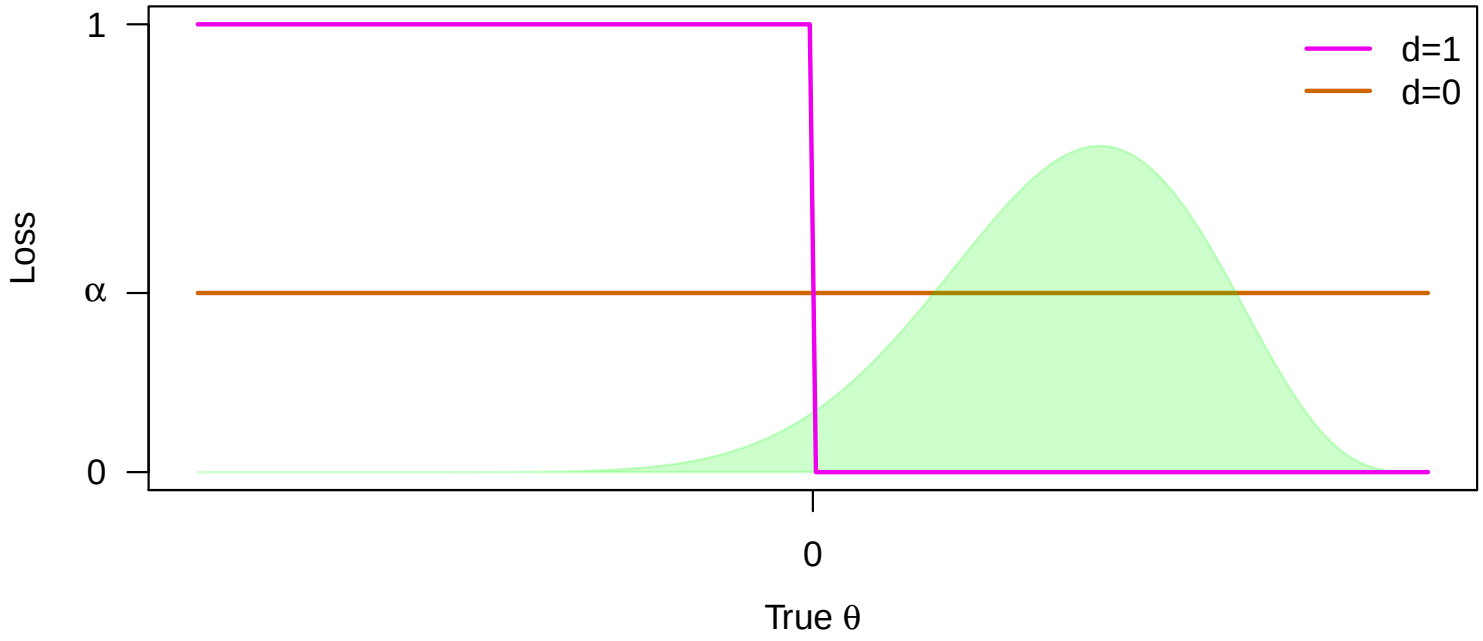
The expected posterior loss is

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

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

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

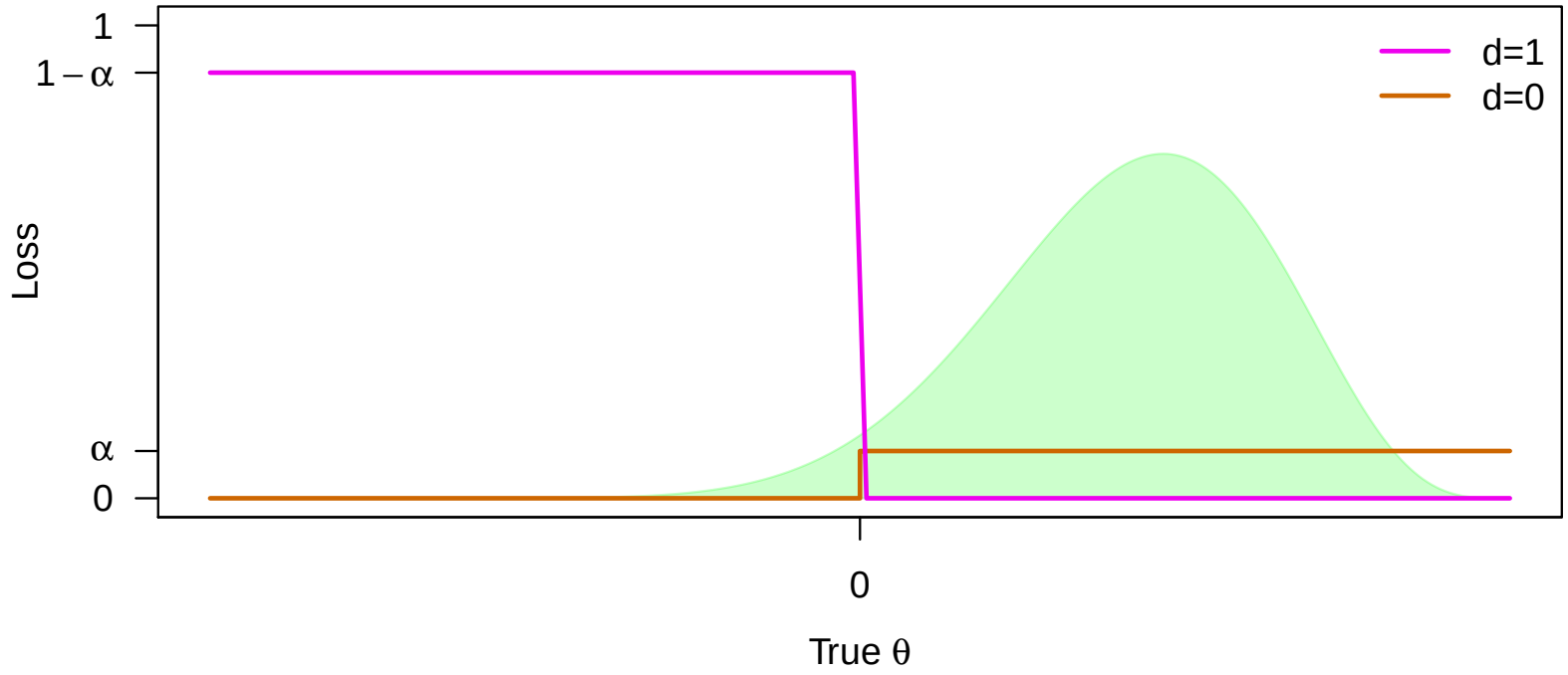
Now suppose we assess the sign of the variant's effect; and let $d = 1$ decide $\theta > 0$, and $d = 0$ for $\theta \leq 0$?

		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*
- α determines **relative** cost of tradeoffs
- 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. So, p-values are not unBayesian (but they're also not BFs)
- Not (yet!) making decisions that θ is exactly zero, or any other specific value... so don't conclude this without more assumptions

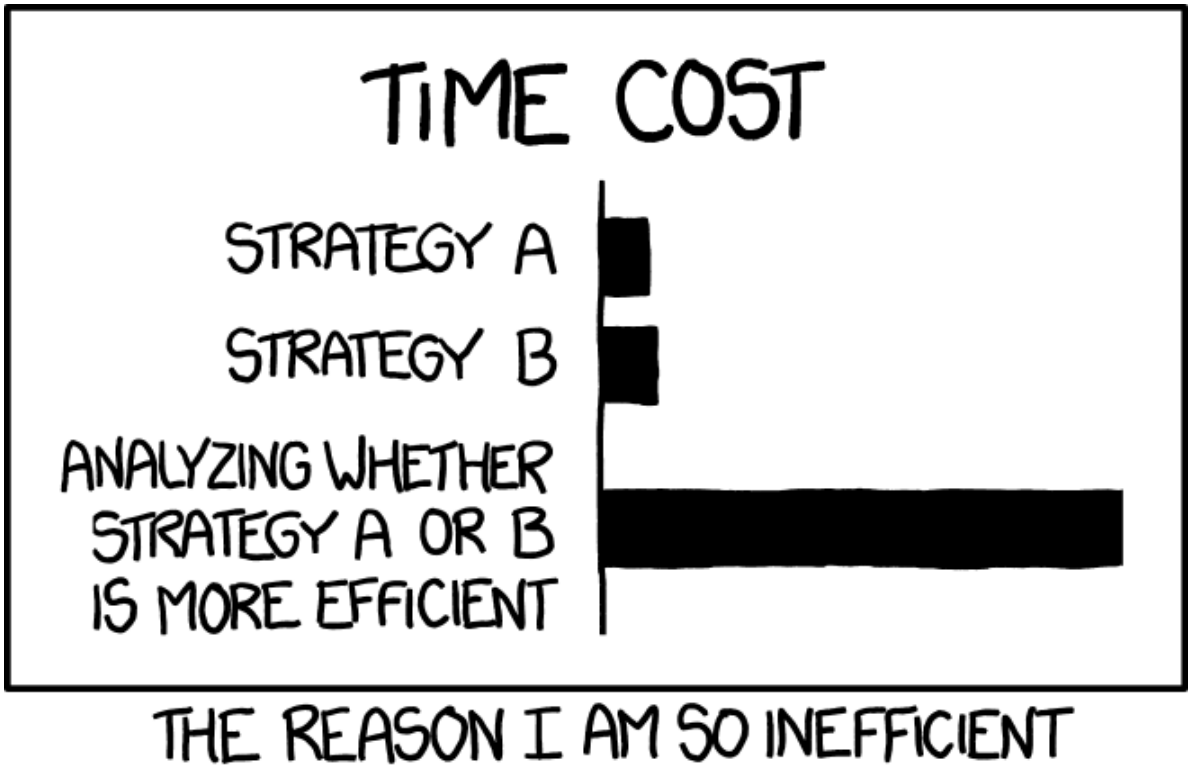
Testing

Never distinguished significance vs hypothesis tests? It may help to consider the **three** verdicts in 'Scots Law';

Verdict	Significance test	Hypothesis test
Guilty	Reject H_0	Reject H_0
Not proven	No conclusion	no analog
Not guilty		Accept H_0

Testing

XKCD on loss functions;



Testing: doing two tests at once

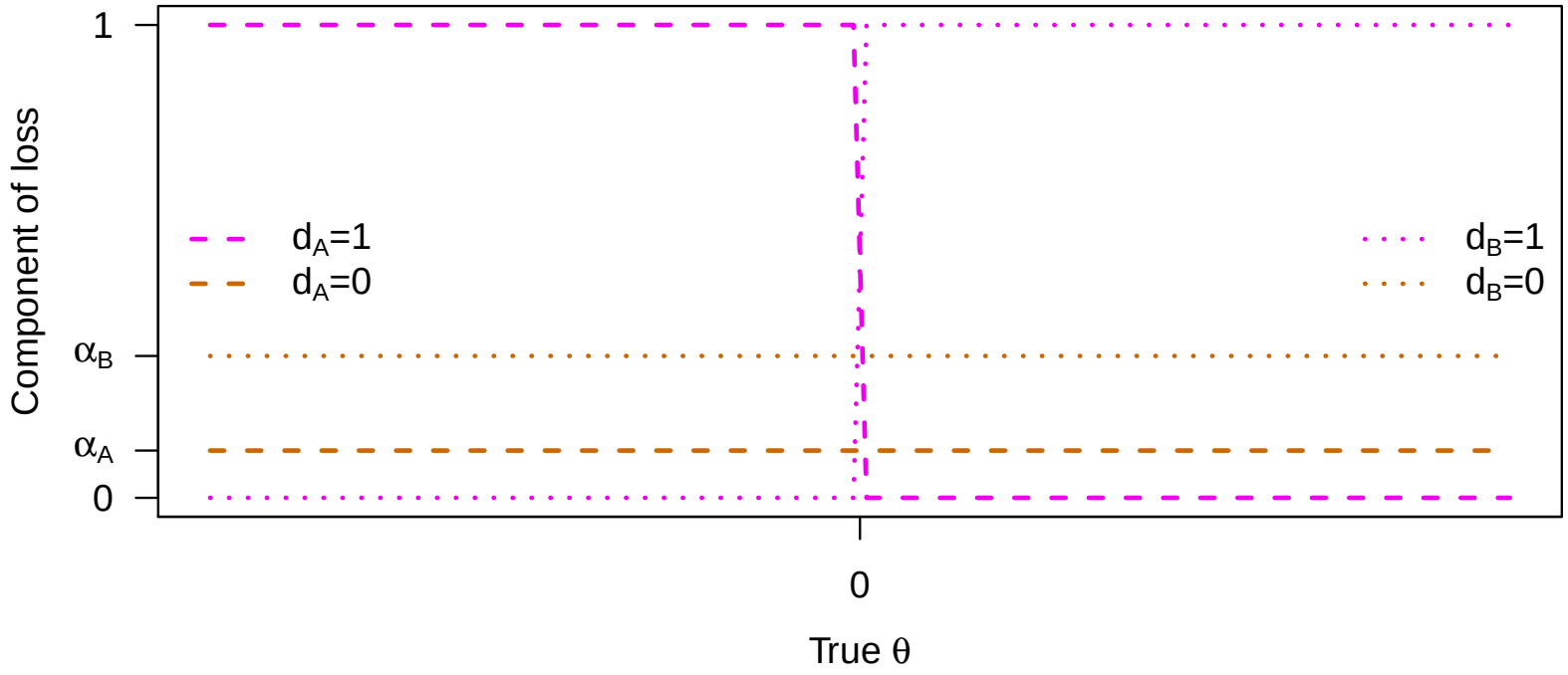
Back to the significance test, i.e. say something vs nothing – but now let's do **two** one-sided tests, that decide if θ is **A**bove 0 or **B**elow 0;

	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?

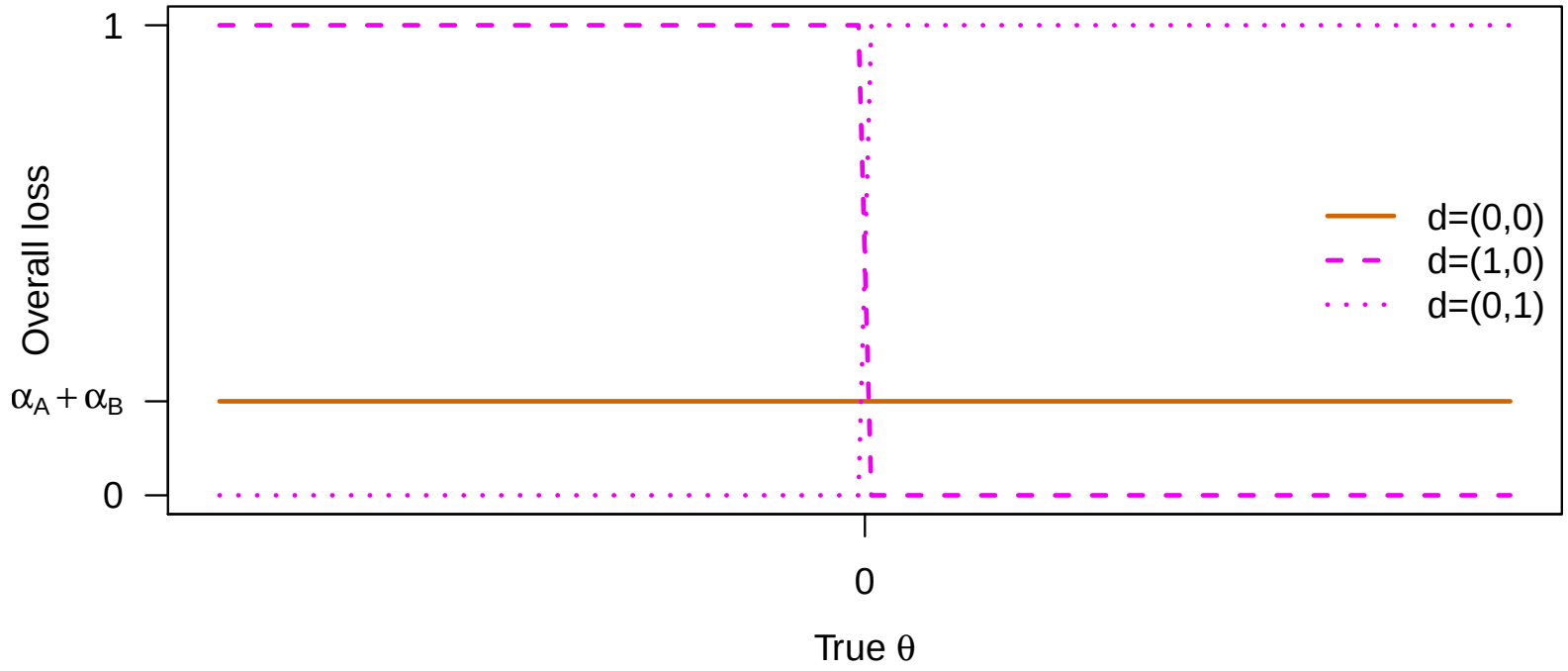
	$d_B = 0$	$d_B = 1$
$d_A = 0$	$\alpha_A + \alpha_B$	$\alpha_A + \mathbb{P}[\theta > 0]$
$d_A = 1$	$\alpha_B + \mathbb{P}[\theta < 0]$	1

- 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*.

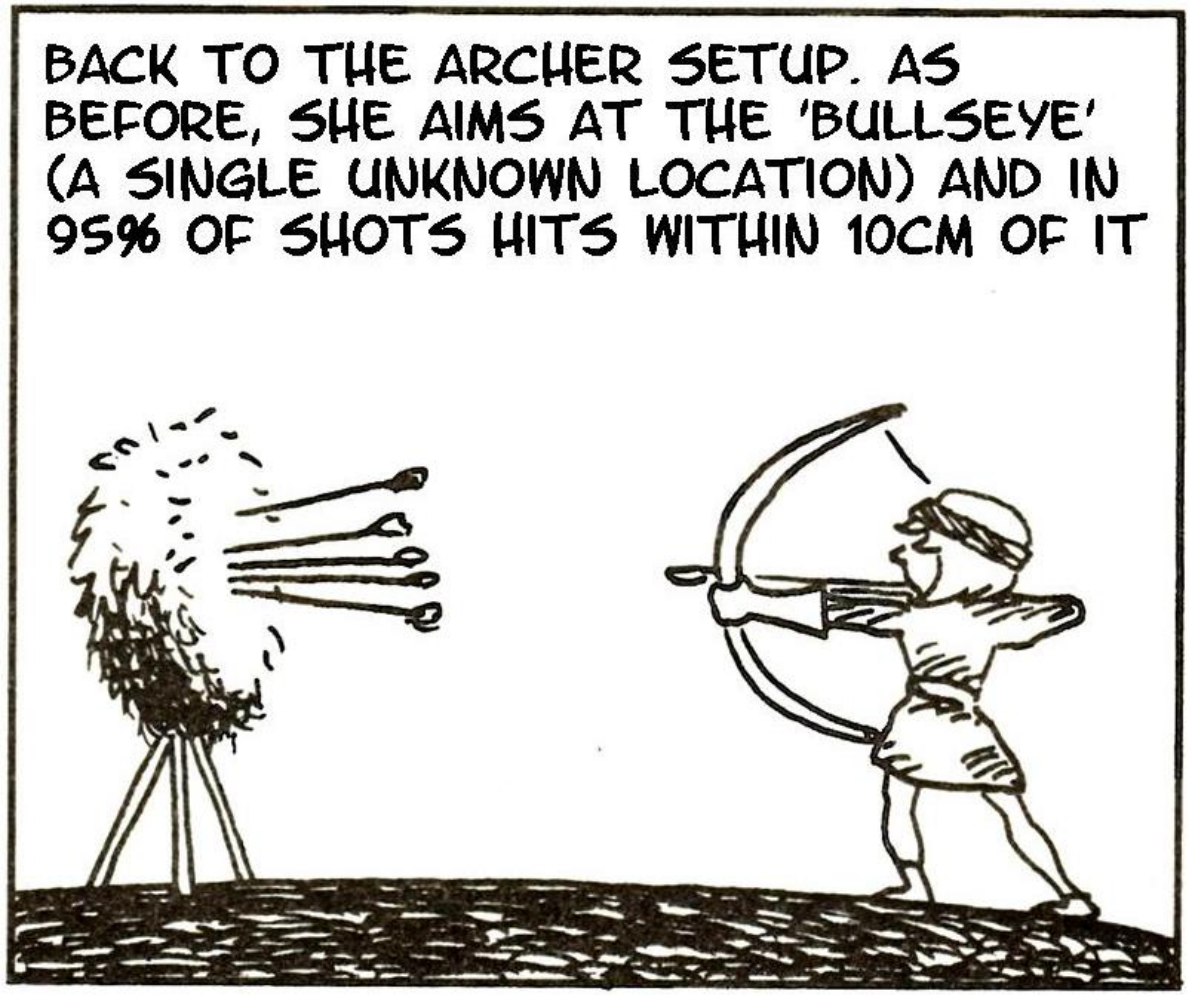
Testing: doing two tests at once

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;



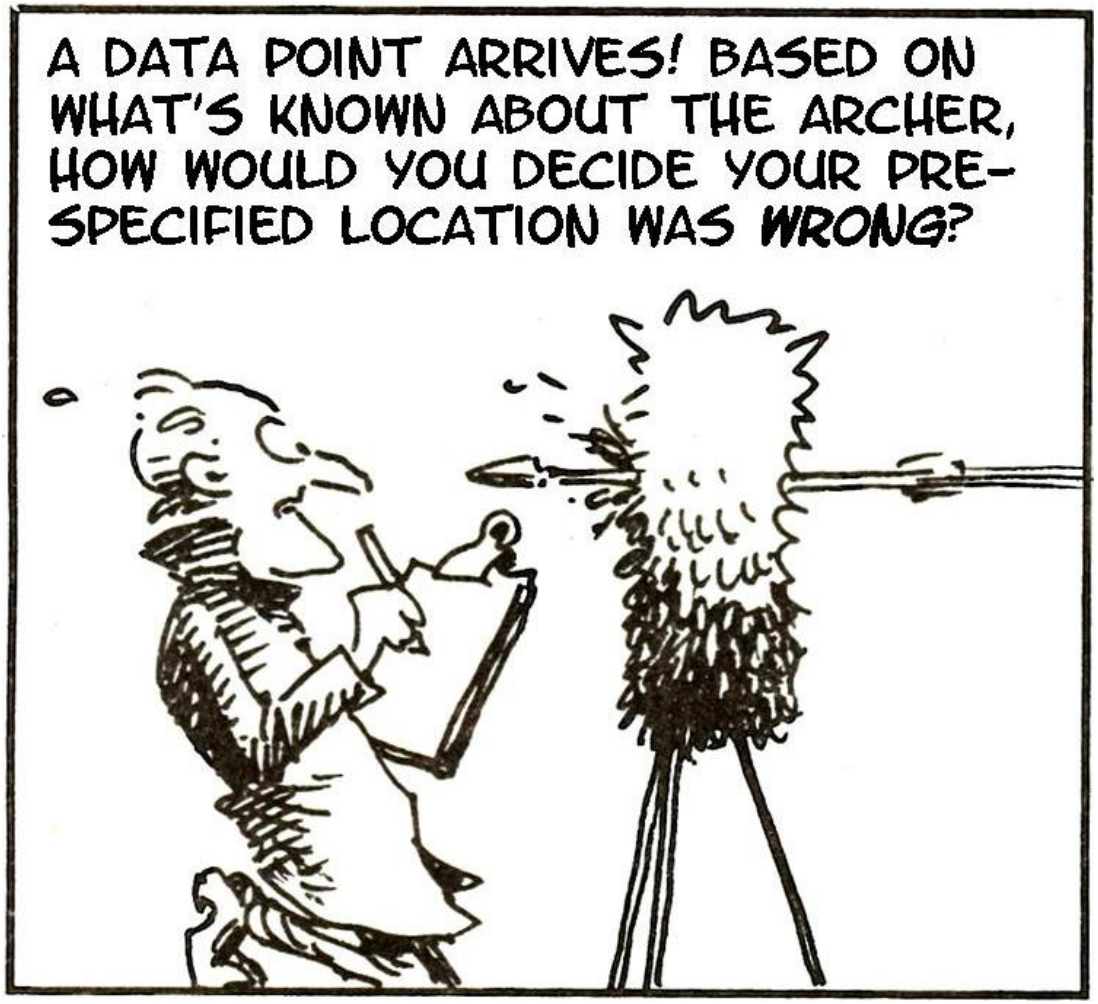
Adapted from Gonick & Smith, *The Cartoon Guide to Statistics*

Testing: frequentist tests

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

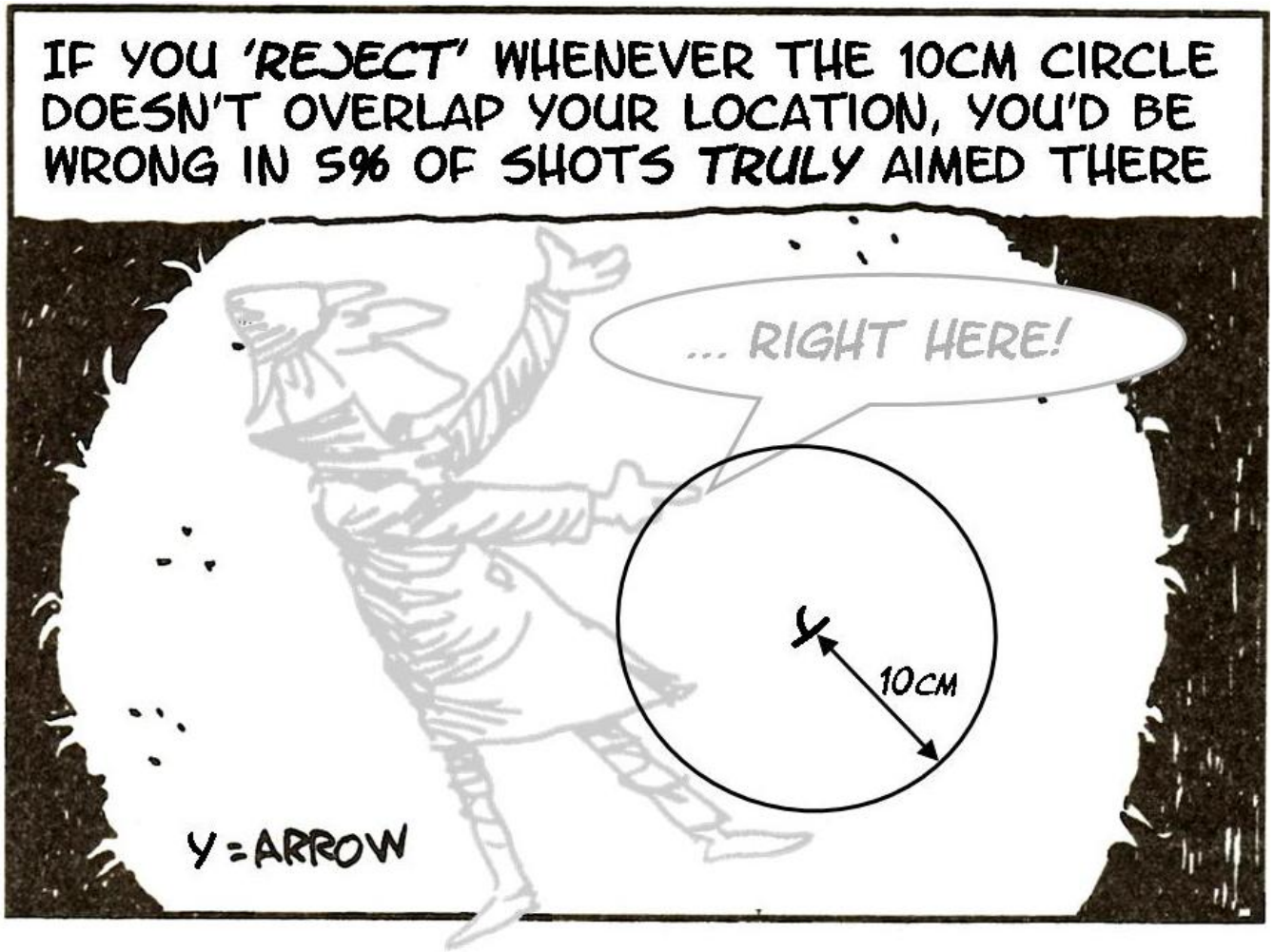
Testing: frequentist tests

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



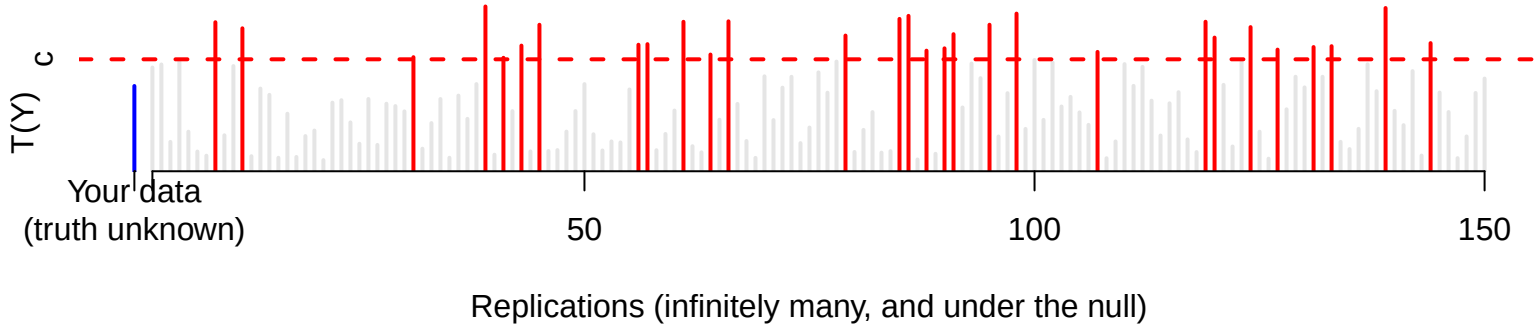
Testing: frequentist tests

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



Testing: frequentist tests

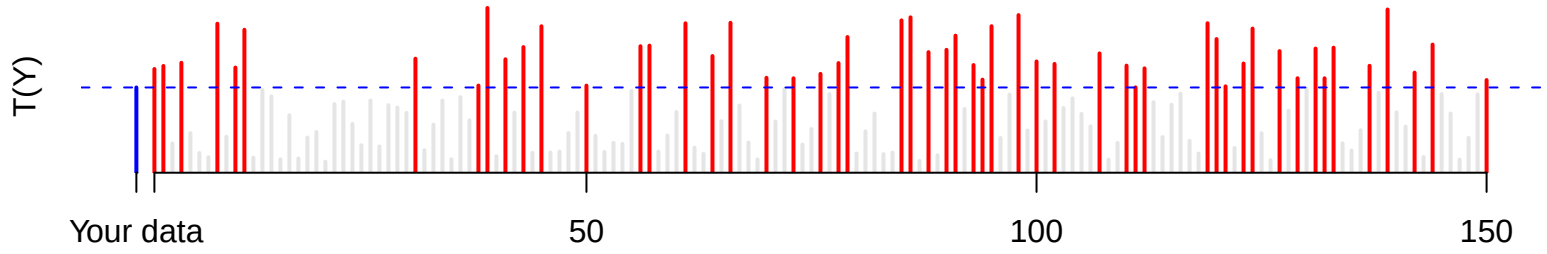
Performing the test means assessing whether our data beats some pre-specified measure of extremity;



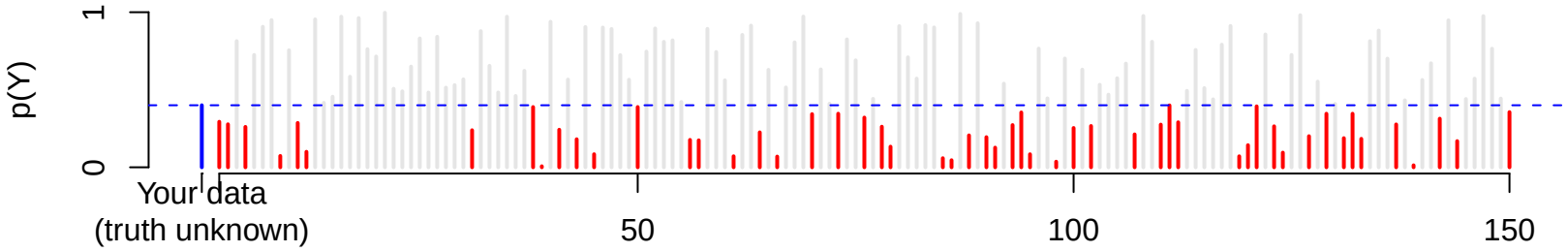
... where the threshold c is chosen so that, under the null, a fixed proportion α of datasets would be that extreme.

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:

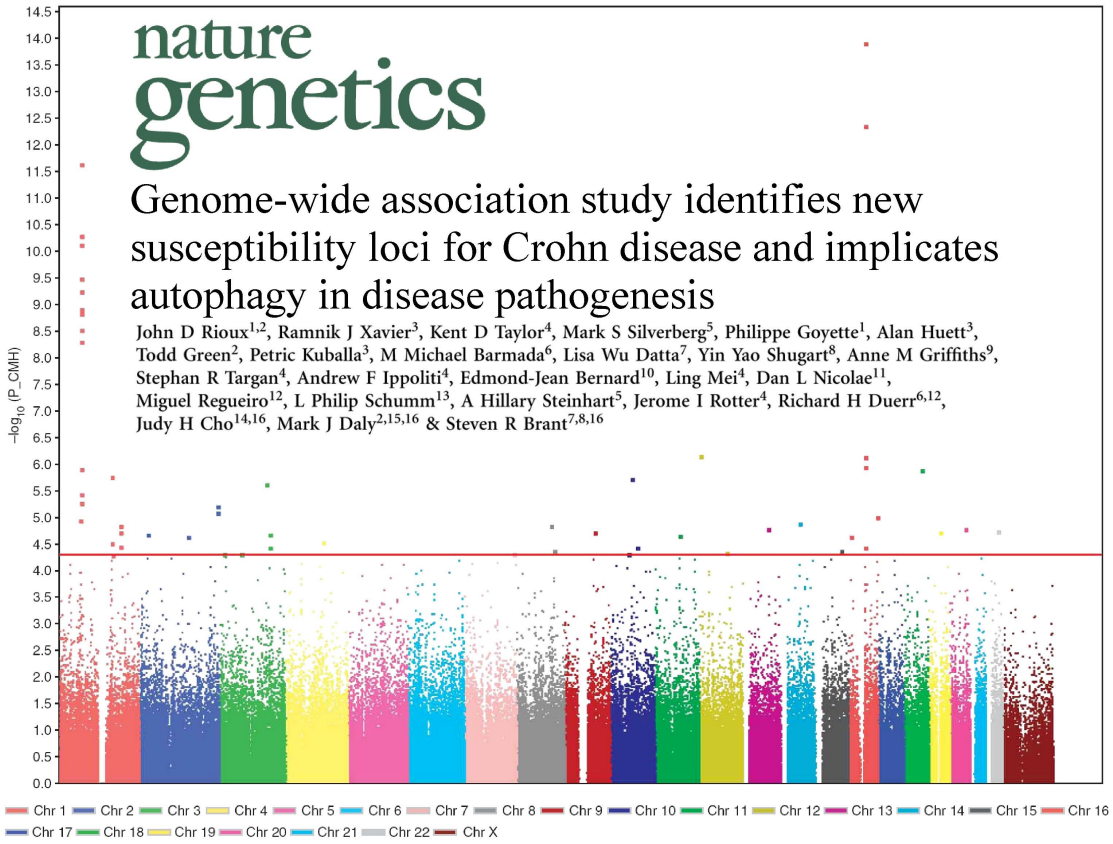
- This form of loss returns $d = 1$ if the posterior odds of positive θ, i.e. $\frac{\mathbb{P}[\theta>0|Y]}{1-\mathbb{P}[\theta>0|Y]}$ are more than B_0 times bigger than the prior odds of positive θ^*
- The ratio of the odds is known as the *Bayes factor* – usually denoted B. It does not depend on 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

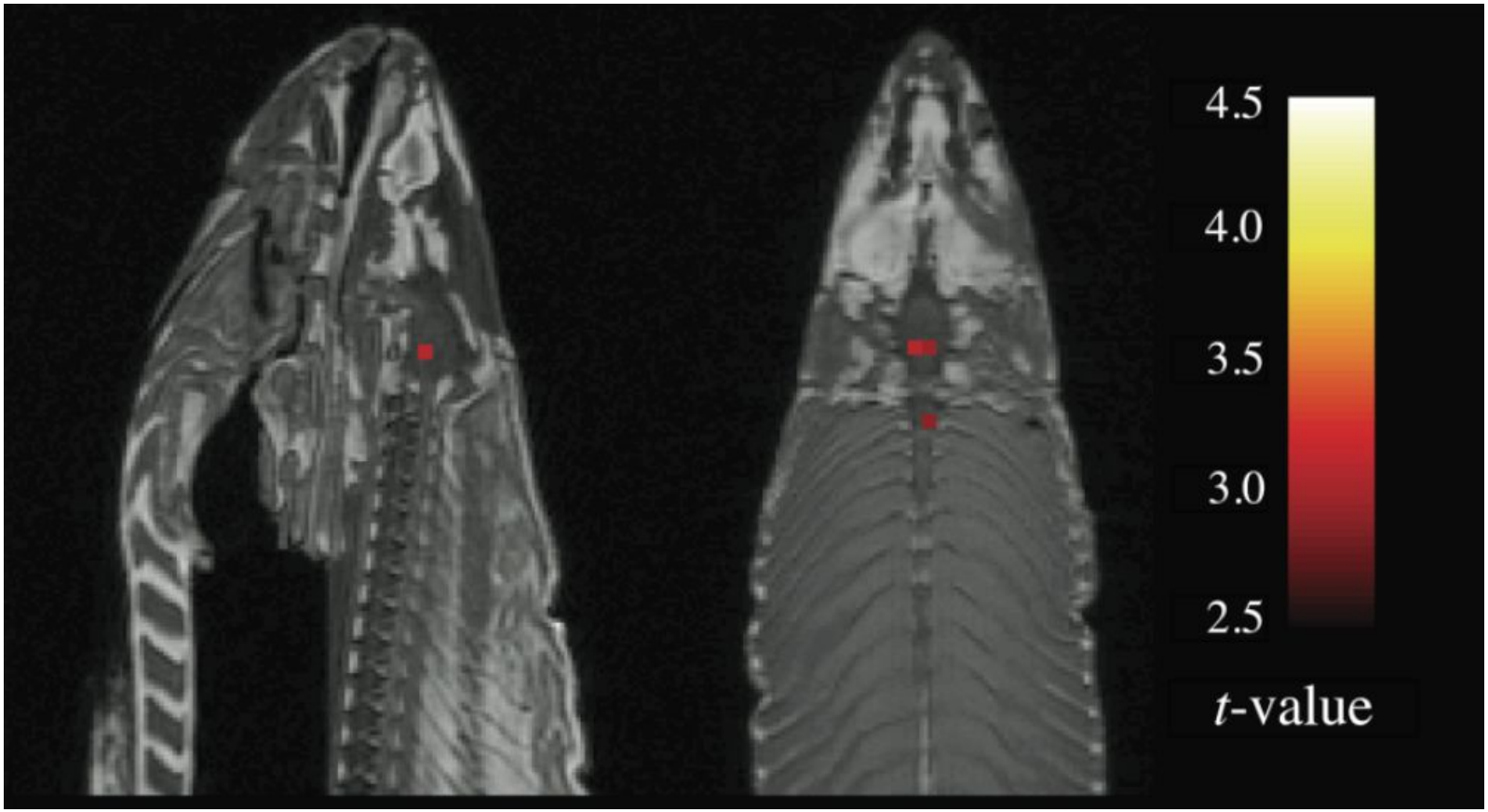
From one test to many;



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

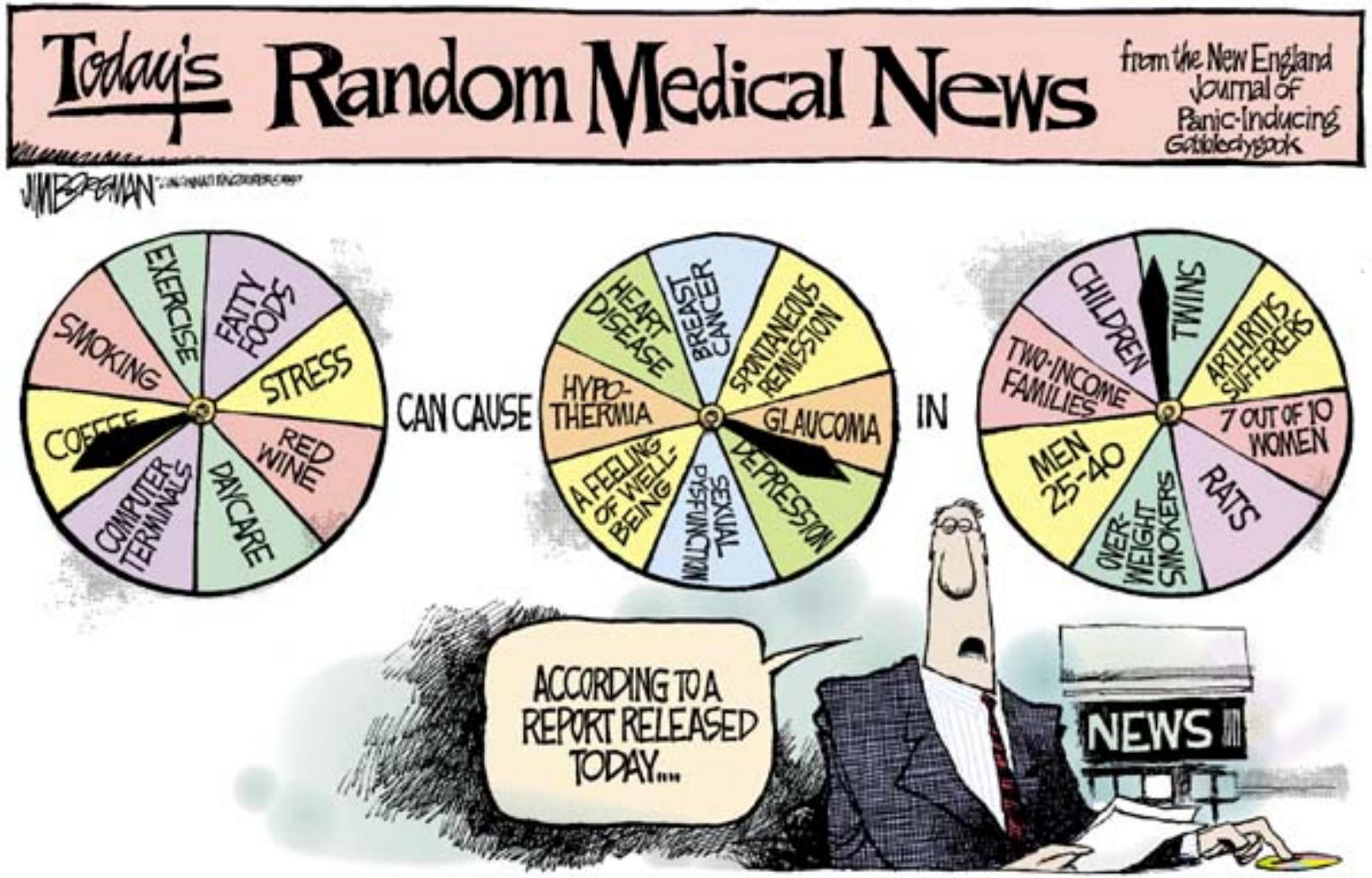
But big data has not always been covered in statistical glory;



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?

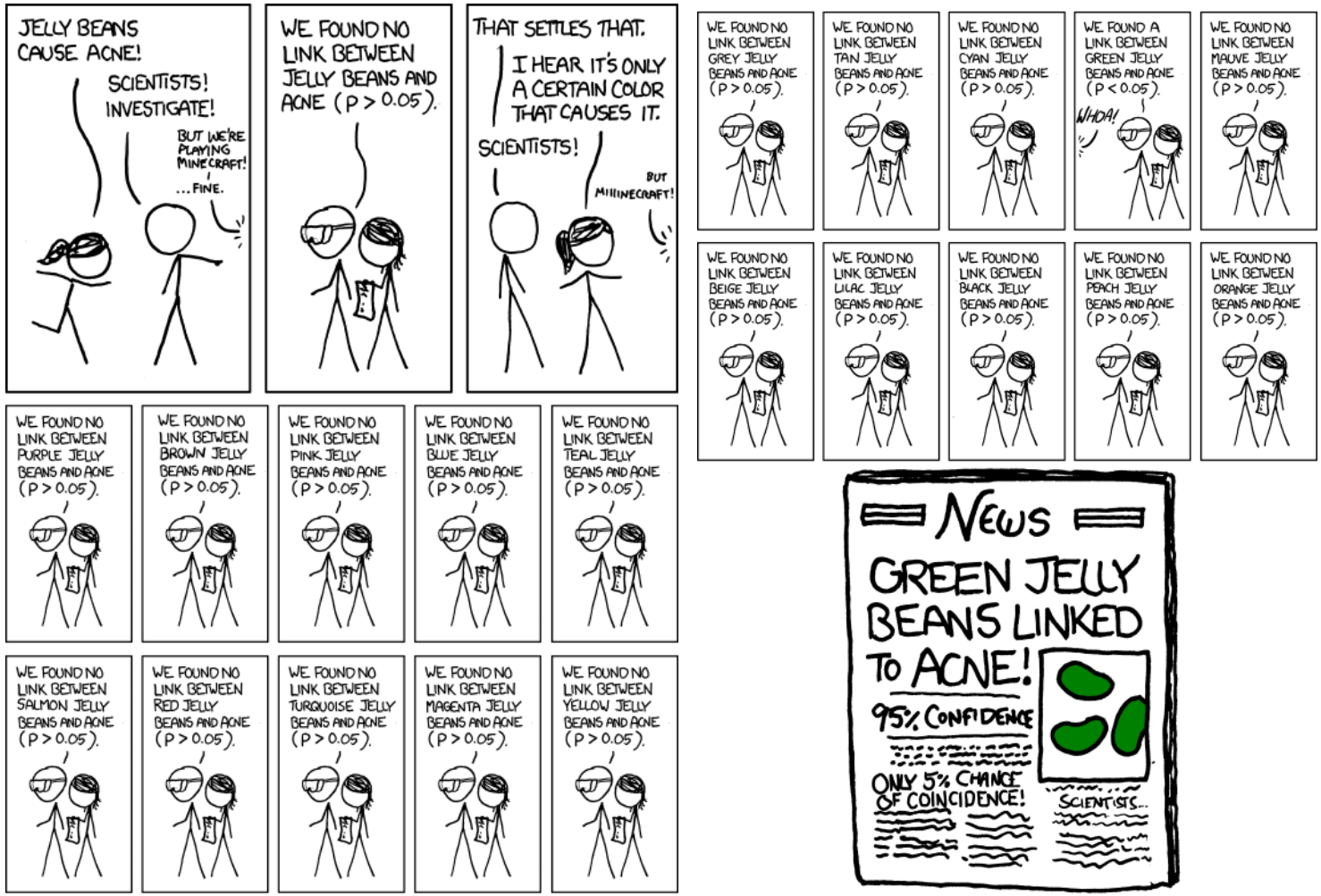
Multiple testing: background

...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.

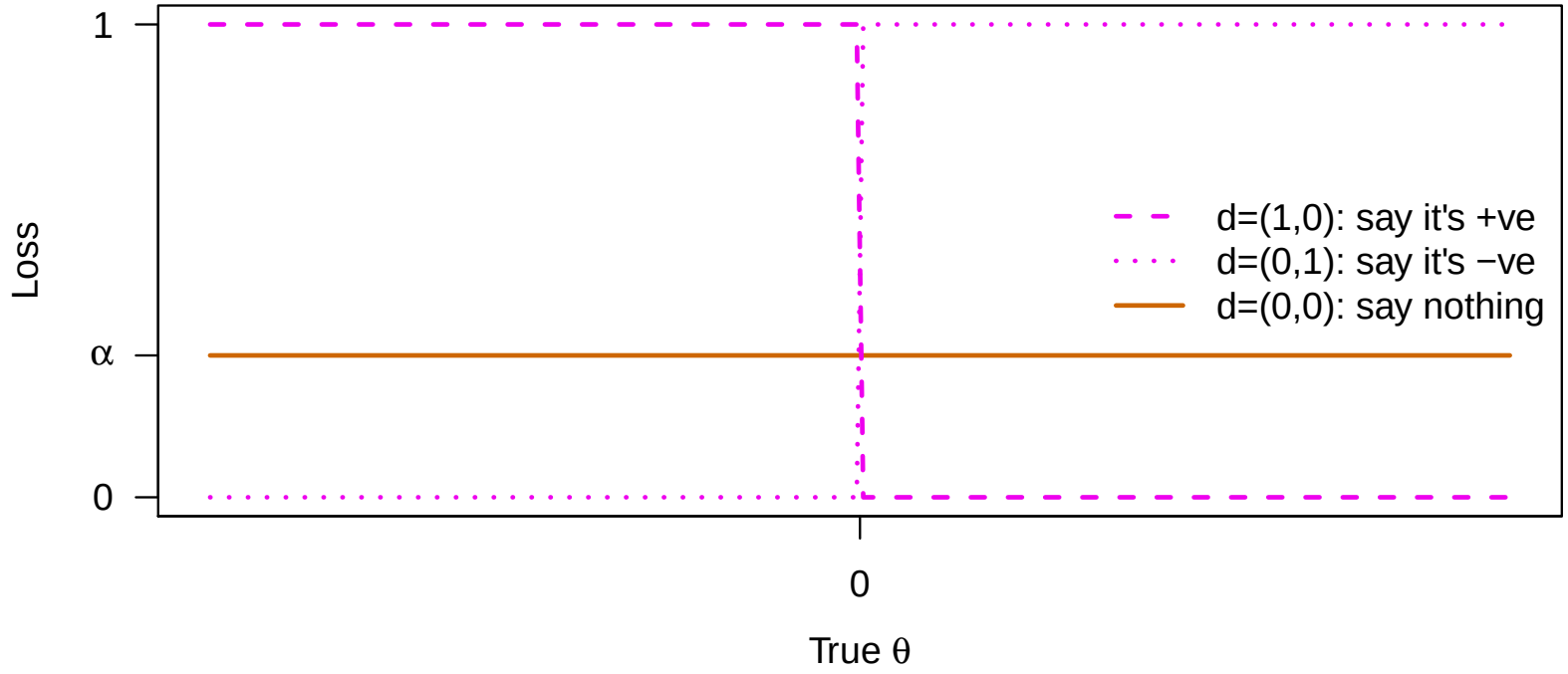
It is thus time to recognize of the extent of multiple comparison problems in everyday epidemiology and deploy modern methods toward their resolution.

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? To ensure that we have *some* chance of saying nothing at all, need to ensure that $\sum_{i=1}^m \alpha_i = \alpha$, for some $\alpha < 1$ – also known as *alpha-spending*. Of course, one easy way to do this is set each $\alpha_i = \alpha/m$, giving a Bayesian analog **Bonferroni correction** of the significance levels.

The 'Bonferroni-corrected decisions' would set $d_i = 1$ for each parameter for which $\min(\mathbb{P}[\theta_i < 0], \mathbb{P}[\theta_i > 0]) < \alpha/2m$. This is a conservative approximation to the Bayes rule $\boldsymbol{d}^B$ here – Bonferroni may set more $d_i = 0$ than the exact Bayes rule.

Multiple testing: many decisions

Rather more simply, consider a different 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
- The Bayes rule is exactly Bonferroni correction – set $d_i = 1$ when tail areas are below α/m

Summary

- Decision theory is a language for saying what we want from analyses – and how much we want it
- Any good analyst does this informally – but for difficult problems, the formal language can help
- Why does anyone use p-values? Perhaps because they only care about the sign of β?
- Multiple testing corrections are not unBayesian. Decision theory can help state what they do...
- ...or at least provide alternative justifications