Knowing the Signs:
Decision theory for significance tests

Ken Rice, Tyler Bonnett & Chloe Krakauer, Univ of Washington, Seattle, USA

Motivation

What do we want/not want from testing methods, for a real-valued θ?

Based on my applied work in high-throughput genetics...

Must not have	Can live with	Must be
Prior 'spikes' at $\theta = 0$	1D parameters	Simple to explain
Conclusions that $\theta = 0$	Parametric models	Optimal, somehow
	Only specifying sign of θ	Connected to p's
		Scottish!

Scottish???

Unlike most statistical tests, ‘Scots Law’ 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 analogue	No conclusion
Not guilty	Accept H_0	no analogue

Decision theory for hypothesis tests

Loss functions deciding signs (is $\theta > 0$? $\theta < 0$?) are *very* limited.

Doing one-sided hypothesis tests we can *only* have:

		Decision	
		$d =$ Above	$d =$ Below
Loss when	$\theta > 0$	l_{TA}	l_{FB}
	$\theta < 0$	l_{FA}	l_{TB}

And with proper loss functions this is wlog:

		Decision	
		$d =$ Above	$d =$ Below
Loss when	$\theta > 0$	0	α
	$\theta < 0$	$1 - \alpha$	0

... for some $0 \leq \alpha \leq 1$. The Bayes rule sets

$$d = \text{Above} \iff \mathbb{P}[\,\theta < 0 | \text{data}\,] < \alpha.$$

—acts like 1-sided p's with large n, but no 2-sided 'double the smallest tail'.

Decision theory for 1-sided significance tests

How we translate 'not proven' into a loss function:

		Decision	
		$d =$ Above	$d =$ **No Decision**
Loss when	$\theta > 0$	0	α
	$\theta < 0$	1	α

- 'Proper' loss fixes the single zero entry, and $0 \leq \alpha \leq 1$ ordering
- We *also* assume "no decision" is equally bad *regardless* of truth

Different decision, same Bayes rule:

$$d = \text{Above} \Longleftrightarrow \mathbb{P}[\,\theta < 0|\text{data}\,] < \alpha$$

—acts like one-sided p's with large n (*cf* Casella & Berger 1987)

Decision theory for 2-sided significance tests

Using proper losses and "no decision equally bad" idea for 2-sided decisions:

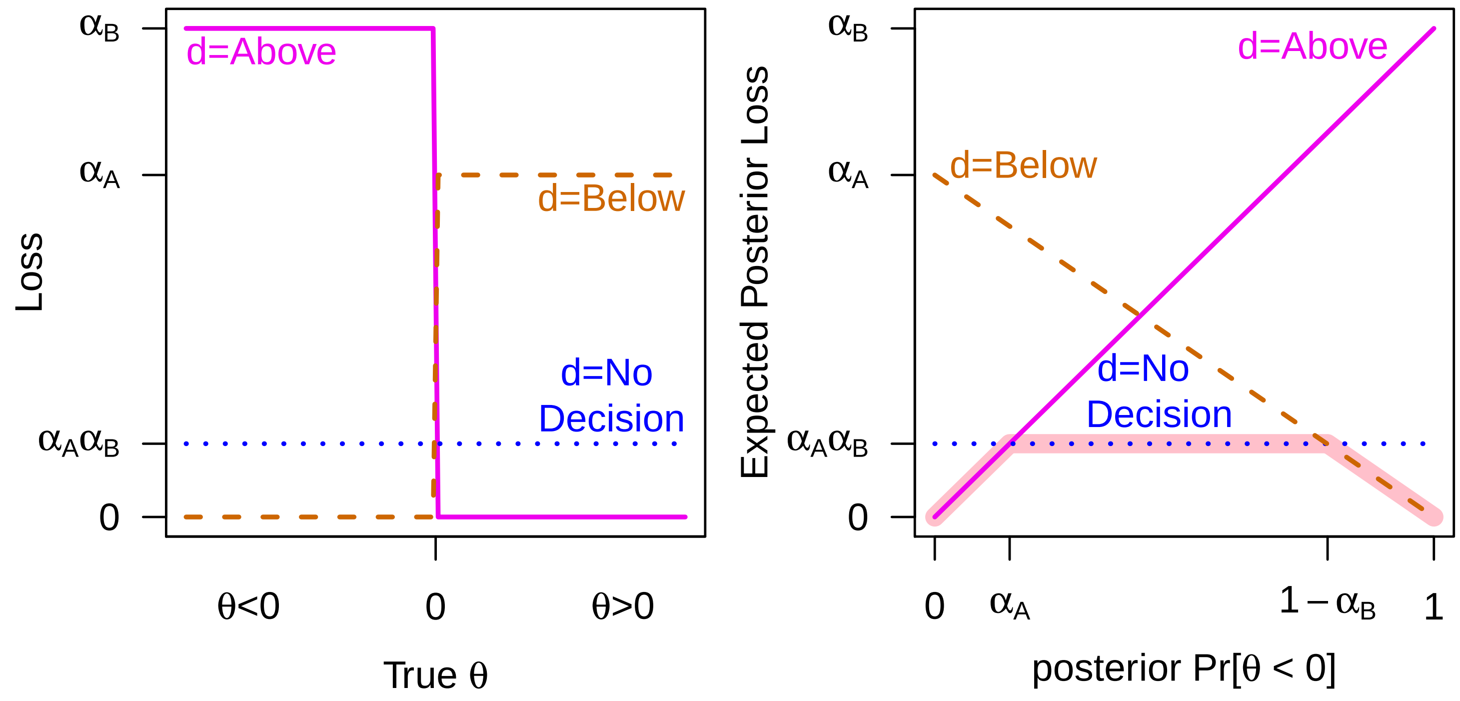
		Decision	
	$d =$ Above	d=No Decision	$d =$ Below
Loss when $\theta > 0$	0	$\alpha_A \alpha_B$	α_A
$\theta < 0$	α_B	$\alpha_A \alpha_B$	0
Bayes rule: do d iff	$\mathbb{P}[\theta < 0] < \alpha_A$	Otherwise	$\mathbb{P}[\theta > 0] < \alpha_B$

...and force $\alpha_A + \alpha_B \leq 1$ by insisting that 'Otherwise' can happen *sometimes*.

With symmetry, get a **Bayesian analog of usual two-sided tests:**

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

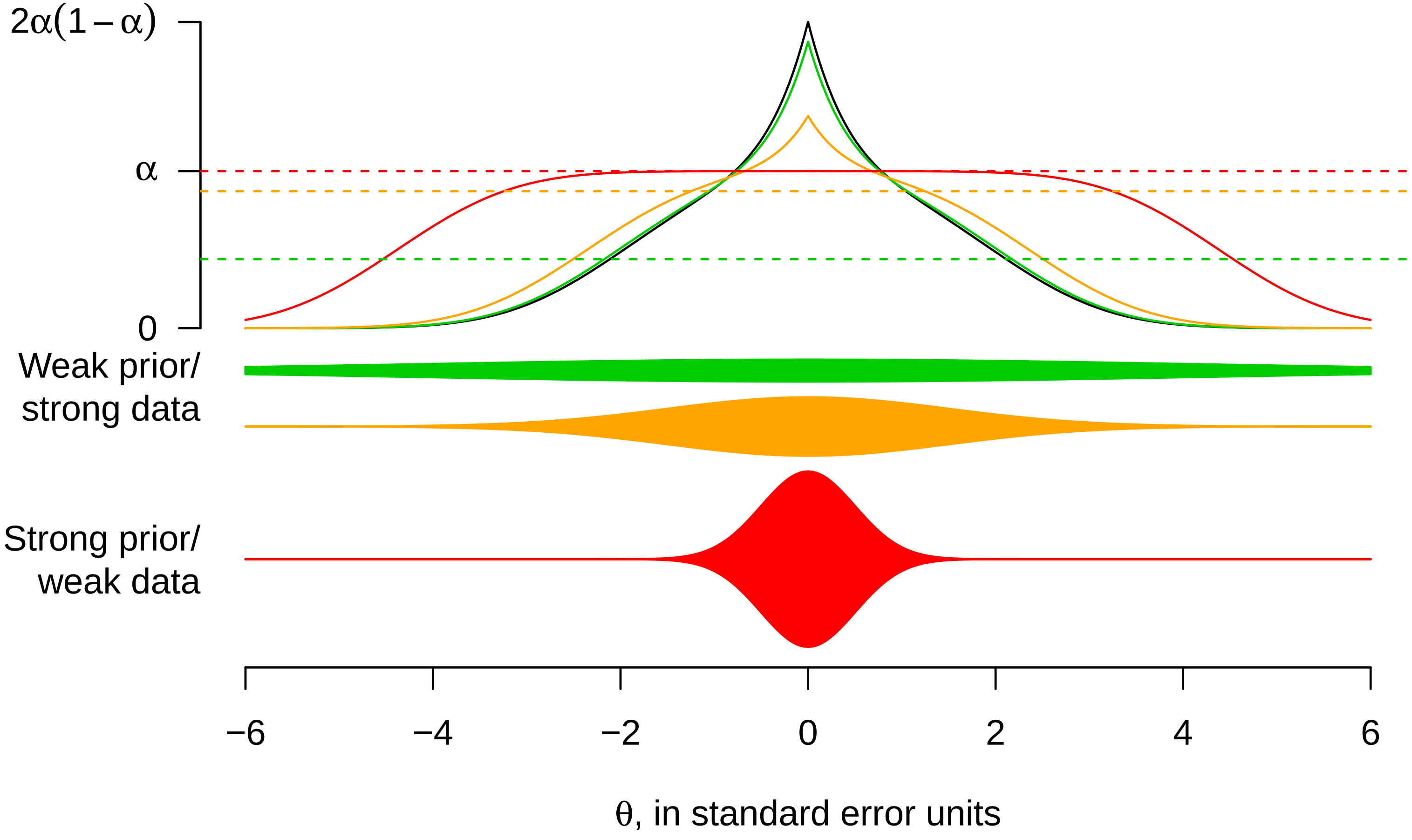
Decision theory for 2-sided significance tests



- Two-sided significance tests are a close (large n) approximation of a Bayes rule for choosing signs – and up to 'proper' conditions, *no other losses/decisions are available*
- With symmetry, expected posterior loss = min$\{P, \alpha\}$ for P=2×minimum tail area – so significant p-value tells us risk of optimal sign-decision

How risky is it?

Risk (——) and also Bayes risk (-----) using Bayes rules for the Normal location problem, $Y \sim N(\theta, \sigma^2)$, with σ^2 known:



Black curve gives risk of classic non-Bayes Z-test – same for any prior.

How risky is it?

- When $\alpha = 0.05$ and there is $\leq 12\%$ power, the classic Z test has *worse* risk than fixing d=No Decision *regardless of the data*. With very low power two-sided Z tests are **futile**
- Using full Bayes rule in those situations is better, but still *almost* futile – risk and Bayes risk very close to α
- Bayes risk goes to zero as prior becomes improper – a reason to not use that prior! To say how much risk is involved, some prior (i.e. contextual) knowledge is **required**

Not shown:

- Similar behavior for other models/parameters
- Connections to *severity* assessments of what is/isn't warranted from a test (but see later)

Extensions: intervals

Written as a function, the 2-sided loss with 'null' value θ_0 is

$$\alpha_B 1_{d=\text{Above}} 1_{\theta<\theta_0} + \alpha_A \alpha_B 1_{d=\text{No Decision}} + \alpha_A 1_{d=\text{Below}} 1_{\theta>\theta_0}$$

Making one decision for *each* possible null value θ_0, and adding the loss functions wrt non-negative measure π on Θ, get loss

$$\alpha_B \pi\left(\mathcal{A} \cap \{\theta : \theta > \theta_0\}\right) + \qquad \alpha_A \alpha_B \pi(\mathcal{N}) \qquad + \alpha_A \pi\left(\mathcal{B} \cap \{\theta : \theta < \theta_0\}\right)$$

for set-valued decisions $\mathcal{A}, \mathcal{B}, \mathcal{N}$.

- *Regardless* of exact π used, Bayes rule sets:
  - $\mathcal{A}$ to be all θ_0 below low α_A quantile of posterior
  - $\mathcal{B}$ to be all θ above high α_B quantile of posterior
  - $\mathcal{N}$ to be the rest, i.e. the **credible interval**
- **Bayesian analog of confidence interval** as "set of all θ_0 that wouldn't be rejected", similarly respects transformations – and often has similar value
- Want to compare intervals? Choose a π and calculate!

Extensions: multiple testing

Recall loss for a single θ: (one-sided for simplicity)

$$L(d,\theta) = 1_{d=\text{Above}} 1_{\theta<0} + \alpha 1_{d=\text{No Decision}}$$

For m different $\theta_j/d_j/\alpha_j$, conservatively trade total N-loss for a *single* wrong sign:

$$L(\boldsymbol{d},\boldsymbol{\theta}) = \left(\sum_{j:d_j=N} \alpha_j \right) + 1_{\cup\{j:d_j=A \text{ and } \theta_j<0\}}$$

and to avoid never setting all $d_j = N$, set $\sum_{j=1}^m \alpha_j = \alpha < 1$ for some α.

- With all α_j equal, do this by setting $\alpha_j = \alpha/m$, i.e. **Bayesian Bonferroni correction of α**. More generally, motivates **Bayesian alpha-spending**
- A conservative *approximation* to the Bayes rule here rejects null when $\mathbb{P}[\,\theta_j|\text{data}\,] < \alpha/m$, i.e. **Bayesian Bonferroni correction of decisions**

Extensions: multiple testing

Trading total α_j for *any number* of wrong signs answers a conservative question. Instead, trading an average of weighted "No Decision" losses against the sum of losses for sign errors, loss is

$$\frac{1}{m}\sum_{j=1}^{m}\alpha_j 1_{d_j=N} + \sum_{j=1}^{m} 1_{d_j=A}1_{\theta_j<0}.$$

- *Exact* Bayes rule sets $d_j = A$ for $\mathbb{P}[\,\theta_j|\text{data}\,] < \alpha/m$, i.e. Bayesian Bonferroni, again – but much simpler than 'classical' version
- A Bayesian analog of **Bonferroni's non-conservative motivation** via control of Expected False Positives (Gorden *et al* 2007)
- Similar trade-offs provide **Bayesian Benjamini-Hochberg algorithm** (Lewis & Thayer 2009)

Extensions: Bayes Factors*

Bayes Factors (BFs) compare posterior to prior – so are not available from losses that use only θ. More Scottish inspiration...

Dolly the Sheep (1996–2003), first mammal cloned from an adult cell – at the University of Edinburgh, Roslin Institute

- We consider a *clone parameter* θ^*: same prior as θ, but *not* updated by data
- Decide if Sign(θ) $>$ Sign(θ^*)? Sign(θ) $<$ Sign(θ^*)? Or make no decision?

* ...Ba-a-a-a-ayes Factors?

Extensions: Bayes Factors

To get Bayes Factors as 1-sided signif'ce test rule for $\theta > \theta^*$, *must* have loss

Truth		Decision, d: $d =$ Above	$d =$ No Decision
$\theta^* < 0$	$\theta < 0$	l_b	l_b
	$\theta > 0$	0	$\frac{1}{1+B}$
$\theta^* > 0$	$\theta < 0$	1	$\frac{1}{1+B}$
	$\theta > 0$	l_a	l_a
Bayes rule: do d iff		$\frac{\mathbb{P}[\theta>0]}{\mathbb{P}[\theta<0]}\frac{\mathbb{P}[\theta^*<0]}{\mathbb{P}[\theta^*>0]} > B$	$\frac{\mathbb{P}[\theta>0]}{\mathbb{P}[\theta<0]}\frac{\mathbb{P}[\theta^*<0]}{\mathbb{P}[\theta^*>0]} < B$

... for l_b, l_a and B all > 0.

- Provides **Bayesian interpretation** of cutoff values for B — not "rough descriptive" guidelines where B=1/3.2/20/150 means S/M/L/XL
- Exactly the same as earlier significance tests, now with prior-dependent threshold $\alpha = \frac{\mathbb{P}[\theta^*<0]}{B\mathbb{P}[\theta^*>0]+\mathbb{P}[\theta^*>0]}$

Extensions: tail area as decision

Significance loss trades-off Above/Below/No Decisions:

$$L(d,\theta) = 2 \times 1_{d=\text{Above}} 1_{\theta<0} + \alpha 1_{d=\text{No Decision}} + 2 \times 1_{d=\text{Below}} 1_{\theta>0}$$

A dual problem: decide the optimal price for *making* tradeoffs between these functions of θ:

$$L(s,a,\theta) = \frac{1}{\sqrt{a}}(2s1_{\theta<0} \qquad + a \qquad + 2(1-s)1_{\theta>0})$$

... for binary s and $0 \le a \le 1$. Note we *heavily* penalize tradeoffs that make No Decision cheap. Bayes rule sets:

- $s = 0/1$ depending if left/right tail is smaller
- $a = 2\times$minimum tail area

Decision a is a **Bayesian analog of two-sided p-value**, and (with direction of smallest tail) tells us about the *process* of choosing signs.

Extensions: tail area as decision

For one-sided tests, need no s decision,

$$\begin{aligned} \text{Loss}: &\quad \frac{1}{\sqrt{a}}(a + 1_{\theta<\theta_0}) \\ \text{Bayes Rule}: &\quad a = \mathbb{P}[\,\theta < \theta_0\,] \\ \text{Minimized expected posterior loss}: &\quad 2\sqrt{\mathbb{P}[\,\theta < \theta_0\,]} \\ \text{risk}: &\quad \mathbb{E}\left[\sqrt{\mathbb{P}[\,\theta < \theta_0\,]} + \frac{1_{\theta<\theta_0}}{\sqrt{\mathbb{P}[\,\theta<\theta_0\,]}}\right] \end{aligned}$$

- Outer expectation in risk is wrt data (but also used to calculate Bayes risk)
- Proportion of risk where a from replicate data is more extreme than a from actual data behaves somewhat like *severity* (Mayo & Spanos, 2006)
- ...can also motivate as proportion of risk coming from replicates with more extreme minimized expected posterior loss
- Can view severity as assessing *how bad the process of choosing signs would be*, in replicate studies, relative to that process with observed data

Conclusions/Questions

Where learning signs is all we'll do, there are simple Bayesian arguments for testing via p-values, and many related methods.

- Not the only Bayesian way to motivate p-values, but could be useful for introducing them
- Prompts users to usefully ask *is the loss relevant?*— does the analysis match scientific goals?

- Normative aspect also helpful: can argue an analysis is 'best' without recourse to UMPU etc
- Yes, priors matter—perhaps a lot—but may be needed. No, this version of p won't fix all problems, e.g. outright fraud, or saying what "evidence" means

Acknowledgements

This work would not be possible without two University of Washington students;

Tyler Bonnett (now at NIH) Chloe Krakauer

Thanks also to: Thomas Lumley, Lurdes Inoue, Jon Wakefield, Leonard Held, the excellent JRSSA referees and editors, and the organizers of this session.
Funding: National Institutes of Health Contract No. HHSN261200800001E

Bonus track: more nuanced decisions

Truth	Decision, d				
	Above	Suggest Above	No Decision	Suggest Below	Below
$\theta > 0$	l_{AA}	l_{Aa}	l_N	l_{Ab}	l_{AB}
$\theta < 0$	l_{BA}	l_{Ba}	l_N	l_{Bb}	l_{BB}

- Bayes rule determined by posterior tail area, again
- 'Proper' conditions on losses $\Longrightarrow$ means decision $A/a/N/b/B$ follows monotonically in left tail area
- Bayesian analog of recent Art Owen/Andrew Gelman work – counterintuitively to some, need **more** significant p-value to declare significance **and** sign of θ.