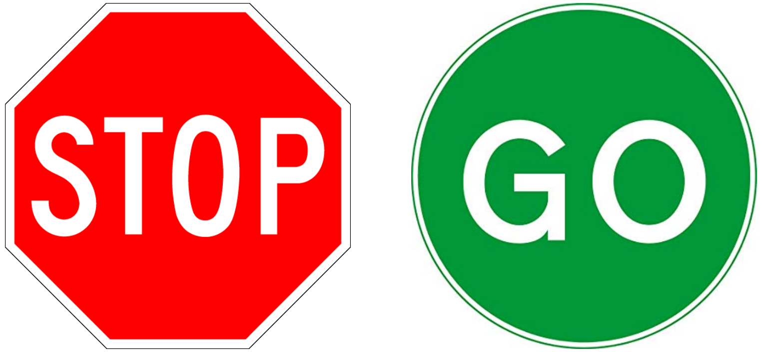


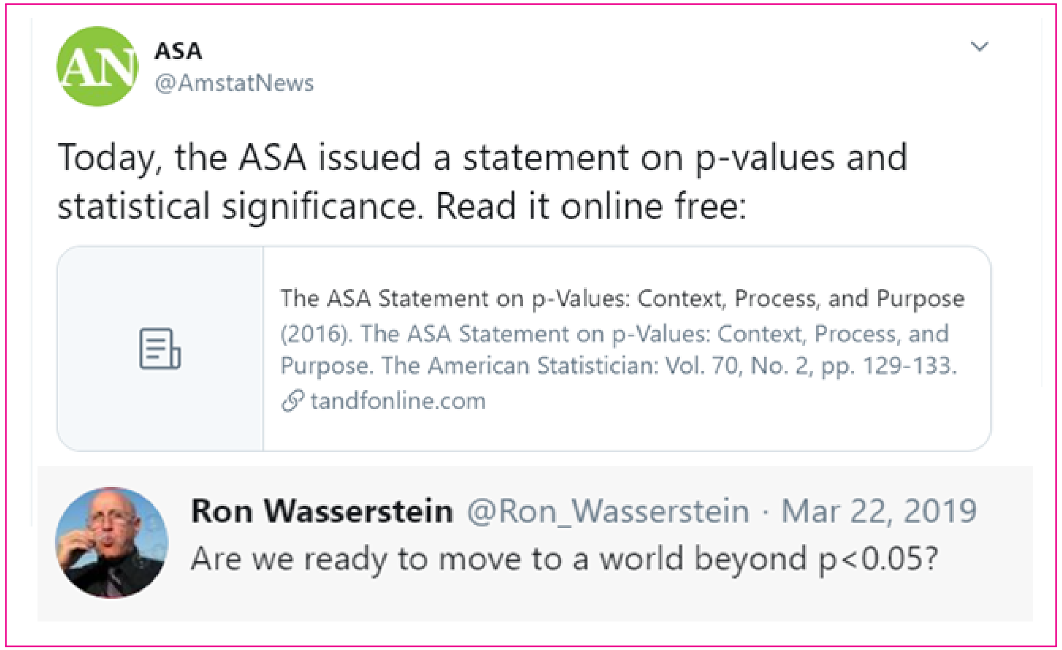
Knowing the signs:
a sensible formulation of tests, and multiple tests

Ken Rice, University of Washington
Joint work with Tyler Bonnett, Chloe Krakauer & Spencer Hansen

tinyurl.com/knowsignsMCP

Motivation: should we eat our p's?

Yes! (2016–19)



No! (2021, with Yoav B!)

- Also recommended: Megan Higgs' thoughtful discussion
- This mess is bad, multiple tests even more acrimonious

Motivation: what would a good solution look like?

What *do* we want/not want from testing methods, for 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???

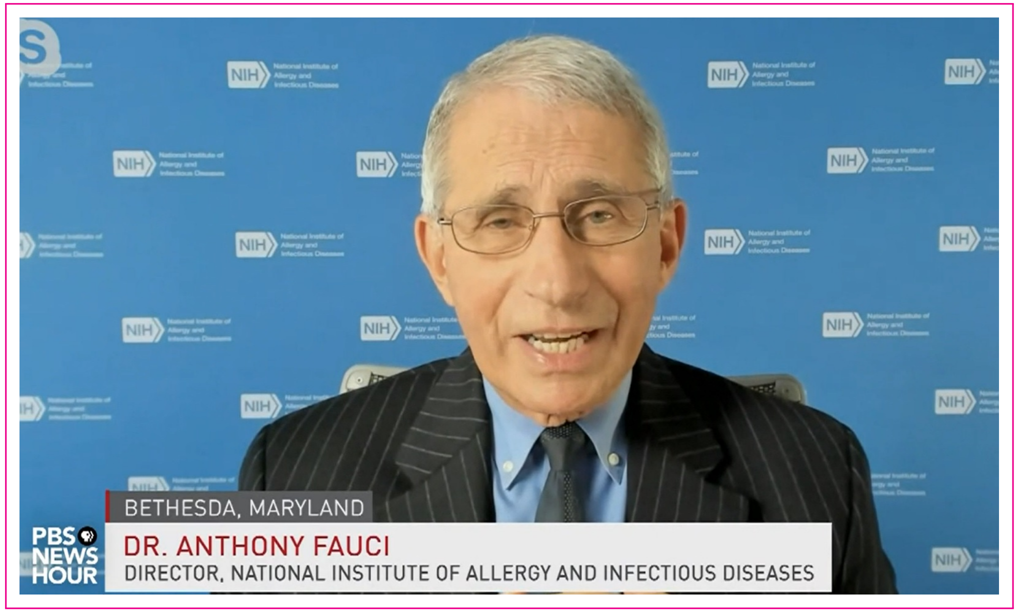
Unlike most statistical tests, 'Scots Law' has *three* possible verdicts – guilty, not guilty and **not proven**:

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 analog	No conclusion
Not guilty	Accept H_0	no analog

Why decision theory?

We develop statistical tests as decisions – because **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**.

Considering *hypothetical* decisions is a reasonable way to prep for the real thing.

Three-decision problems: how bad can it be?

Losses for "three-decision" problems (is $\theta > 0$? $\theta < 0$? not saying?) are limited!

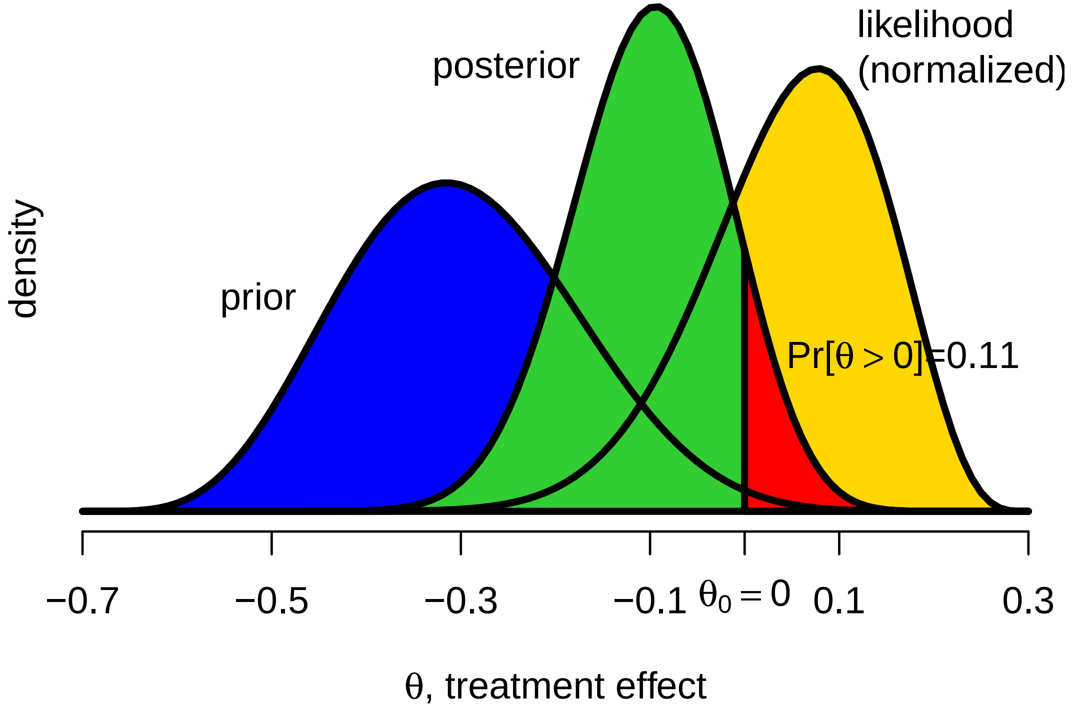
		Decision (what do we assert?)	
	Above	No Decision	Below
Loss when $\theta > 0$	l_{TA}	l_{NA}	l_{FB}
$\theta < 0$	l_{FA}	l_{NB}	l_{TB}

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

		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 analog of 2-sided testing via p-values**

Three-decision problems: can they be transparent?



- With $\alpha = 0.05$, sign errors are $\times$**40 worse** than making no decision
- ...so only make sign decision if $2\min(\mathbb{P}[\theta < 0], \mathbb{P}[\theta > 0]) < 0.05$.
- Here, $2\mathbb{P}[\theta > 0] = 0.22$, make no decision – and incur loss $0.05/2$

Three-decision problems: notes

- Tukey (2000) viewed the 3-decision setup as a "sensible formulation" of tests
- Known much earlier, e.g. Cox (1982) notes unknown sign is "perhaps most common" hypothesis

- Under 3-decision setup, p-value based tests are **basically inevitable** – no Jeffreys-Lindley paradox/embarrassment
- Frequentist Type I error rate control at α, with large n (Bernstein-Von Mises)
- In our 3-decision setup, α is a **fixed ratio of costs**, and we minimize

$$\text{risk} = \text{Rate}_\theta[\,\text{sign error}\,] + \frac{\alpha}{2}\text{Rate}_\theta[\,\text{no decision}\,]$$

... i.e. a weighted sum of Type III and Type II error rates

For references/review, see Rice *et al* (2019, JRSSA) and discussion

Three-decision problems: how to explain them?

Main points for communicating with non-statisticians:

- When testing we **assert** that $\theta > 0$, $\theta < 0$ – or make no decision
- This is crude! But so are tests!
- Less prone (I think) to over-interpretation than usual accepting/rejecting implausible point null

- Normative: 3-decision approach gives 'best' test via one criterion without UMPU-ness, asymptotic efficiency, exponential familes...
- Yes, priors matter—perhaps a lot—but may be needed. No, this approach won't fix all problems, e.g. outright fraud, or data-dredging

Three-decision problems: what else do we get?

Details at tinyurl.com/knowsignsMCP, but simple extensions give:

- Two-sided p-values
- Intervals
- Bayes Factors
- Why *post hoc* power calculations tell you nothing new
- Prior sensitivity checks (reverse-Bayes)
- Coherent tests of interval nulls (Bayes and frequentist)
- 80% power as default (it means study is low risk, i.e. $\times 5$ smaller risk than do-nothing $\alpha/2$)
- $p < 0.005$ a 'next-level' threshold (it means we make sign decision AND have $>$50% belief study was low risk)

… and of course **multiple testing**

Multiple sign tests

For $j = 1, 2, ...m$ tests, **tempting** to 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!**

- The loss is simply

$$\text{Loss} = \frac{\alpha}{2m} \#\{\text{non-decisions}\} + 1_{\text{any sign error}}$$

Gives FWER analog, but α enters **only as a ratio of costs**

Multiple sign tests: can it be more realistic?

But one sign error $\neq$ all m sign errors! **Better** to instead add m copies of the 3-decision loss, with all $\alpha_j = \alpha/m$:

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

- Each θ_j in its own sign error/non-decision tradeoff
- Bonferroni-corrected 2-sided tests are the **exact** Bayes rule!
- Analog of using α as expected number of false positives (EFP), see e.g. Gordon et al 2007
- No automatic reason to constrain $\alpha < 1$ (but EFP $\gg 1$ usually undesirable)
- Distinguishes 'conservative' control from 'conservative' criterion

Note: making no decisions for any θ_j, we **know** loss=$\alpha/2$.

Multiple sign tests: what else does this give?

Some (nice!) extensions:

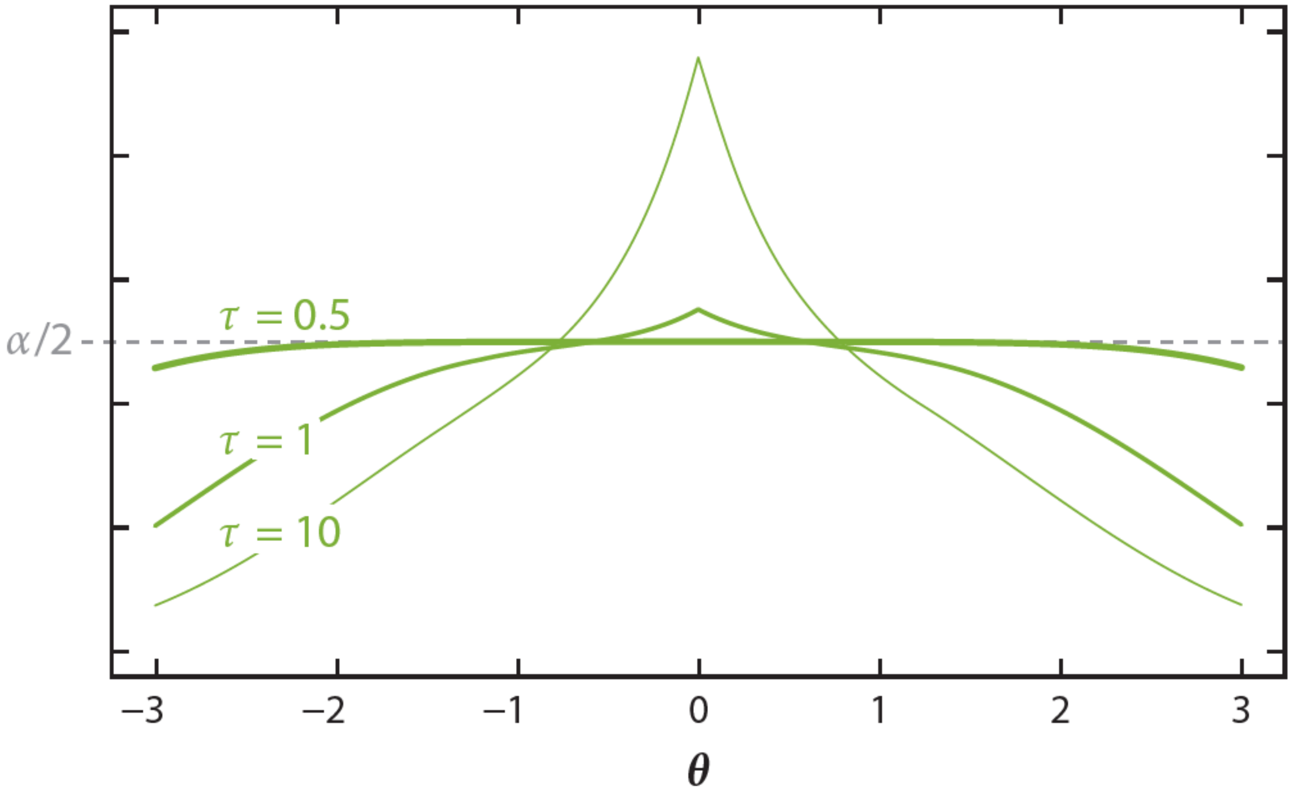
- Lewis & Thayer (2013), following Sarkar and Zhou (2008), show how using 'simpler' loss controls expectation of $\frac{\#\{\text{sign errors}\}}{1\vee\#\{\text{sign decisions}\}}$ wrt **both** prior and sampling uncertainty – controlling the *Bayesian directional false discovery rate*
- Lewis & Thayer (2009) use

$$\text{Loss} = \frac{\#\{\text{sign errors}\}}{1 \vee \#\{\text{sign decisions}\}} + \frac{\alpha}{2}\frac{\#\{\text{non-decisions}\}}{m}$$

to motivate Bayesian analog of Benjamini-Hochberg algorithm: step-up procedure comparing $\times 2$ tail areas to $\alpha j/m$

Futility

Briefly back to a single test; for simple $Y \sim N(\theta, 1)$ location problem with $\theta \sim N(0, \tau^2)$ prior, frequentist **risk** of Bayes rule, at different θ:



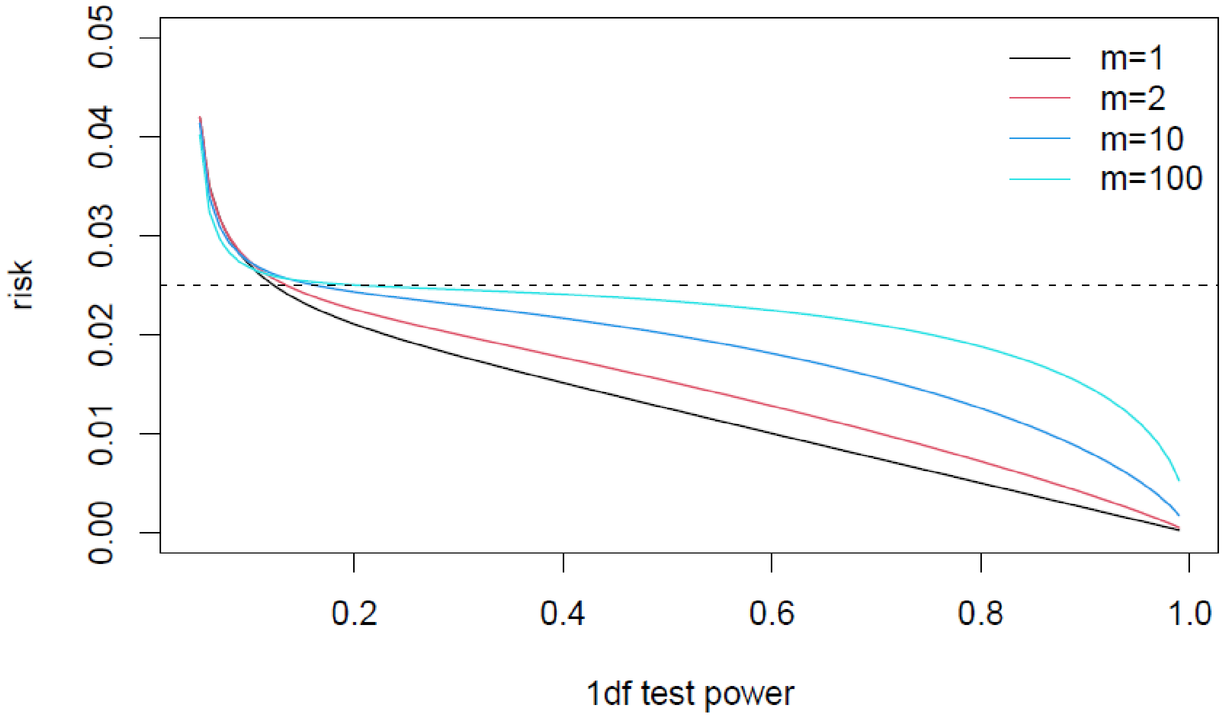
- For $\theta \approx 0$ making no decision **regardless of data** (loss $\equiv \alpha/2$) is better
- For Z-tests/flat prior, *futility* occurs with $< 12.2\%$ power ...which can be realistic!

Multiple tests: can they be futile?

Using the better Bonferroni-correction loss, study is futile if

$$\mathbb{E}[\,\#\{\text{sign errors}\}\,] + \frac{\alpha}{2m}\mathbb{E}[\,\#\{\text{non-decisions}\}\,] > \alpha/2$$

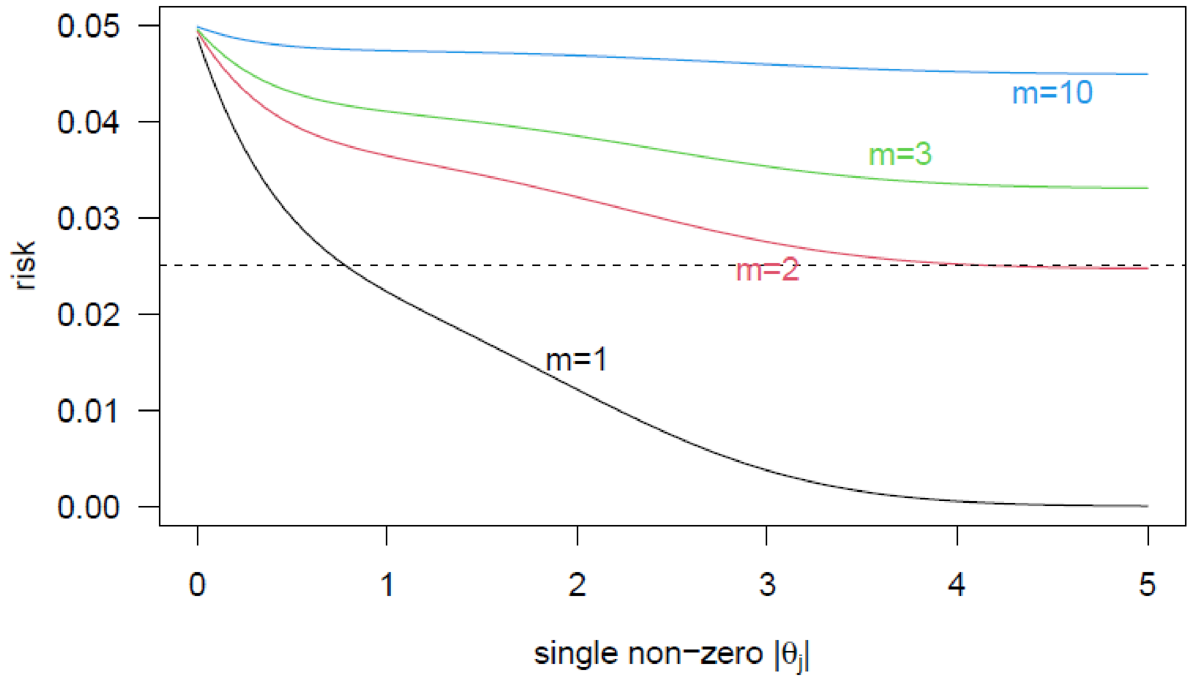
For independent $Y_j \sim N(\theta_j, 1)$, flat priors & **all** θ_j **equal** doesn't look too bad: study is futile if 1df tests have power between 12.2% (m=1) and 19.8% (m=100) – threshold is $\approx$ log-linear in m.



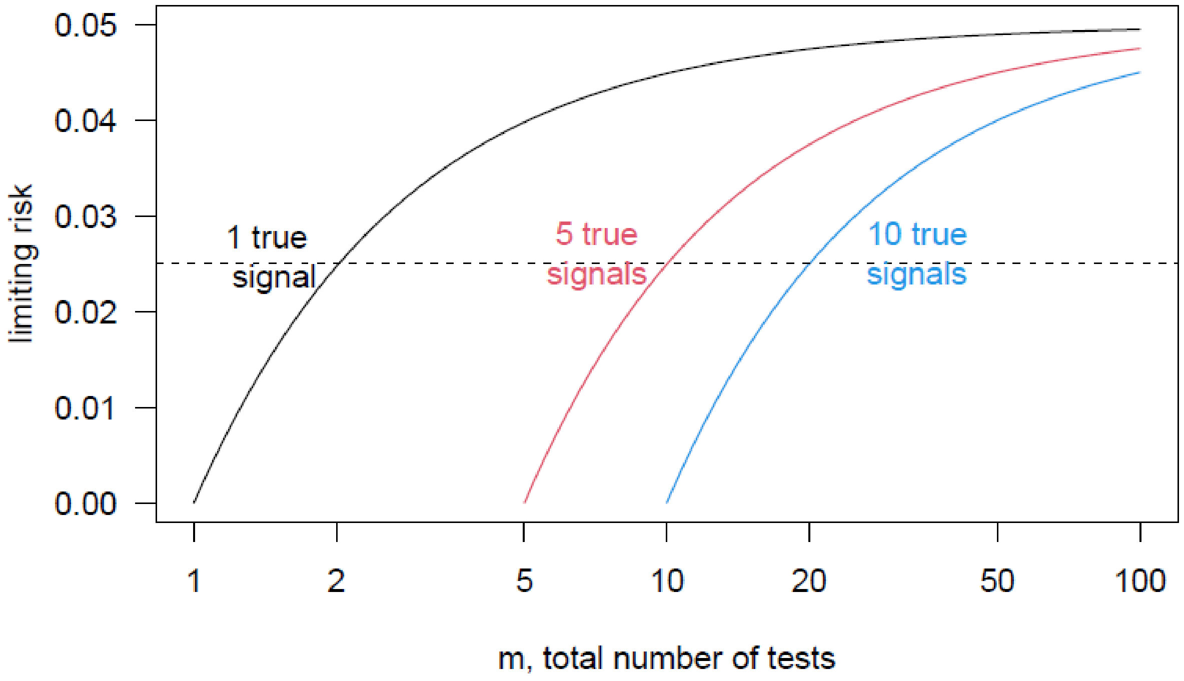
Multiple tests: can they be futile?

But elsewhere some alarming properties: (flat priors/classic Z tests)

With one signal θ_j,
the other $m-1$ pure noise:



Limiting risk
when all non-noise $|\theta_j| \to \infty$



Multiple tests: can they be futile?

Work in progress: with the better loss,
Bayesian Bonferroni is admissible, but classic Bonferroni is **not**.

Strong hints of this Stein-type behavior:

- With enough near-zero θ_j, must be optimal to **heavily** shrink borderline sign decisions to non-decisions
- Futile parameter space is bounded for $m = 1, 2$ only – classic Stein paradox kicks in at dimension ≥ 3
- The loss is penalized OLS: writing decision d_j=-1, +1 or 0 (for no decision)

$$\text{Loss} = \underbrace{\frac{1}{4}\sum_j (\text{sign}(\theta_j) - d_j)^2}_{\text{squared error}} + \underbrace{\left(\frac{\alpha}{2m} - \frac{1}{4}\right)\sum_j (1 - d_j^2)}_{\text{discourage decisions}}$$

- Better rules will work much like Storey's ODP (2007)

Are you going to stop now?

Key points:

- Sign-decisions provide a simple, general system by which we can understand *and criticize* tests and multiple tests
- Optimize a single criterion, **not** optimizing one while another is controlled (over what θ? under what modeling assumptions?)
- Bayes/frequentism pluralism (basically!)
- Don't like these loss functions? What is *your* definition of a good/bad answer?

For forthcoming Annual Reviews paper, links, etc see

tinyurl.com/knowsignsMCP

Thank you!

This work would not be possible without;

Tyler Bonnett
(now at NIH)

Chloe Krakauer
(now at Kaiser)

Spencer Hansen
(now at UW CHSCC)

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