Testing with one real parameter

Outline

1) Uniformly most powerful tests
2) Two-sided tests
3) p-Values
4) Confidence Regions

Uniformly most powerful tests

General setup: $\mathcal{P}, \Theta_0, \Theta_1$

Def If $\phi^*(x)$ has sig. level α, and for any other level-α test ϕ we have

$$\mathbb{E}_\theta \phi^* \geq \mathbb{E}_\theta \phi \qquad \forall \theta \in \Theta,$$

then ϕ^* is uniformly most powerful (UMP)

Typically only exist for 1-sided testing in certain 1-parameter families.

Def A model $\mathcal{P}$ is identifiable if

$\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2} \quad (\exists A : P_{\theta_1}(A) \neq P_{\theta_2}(A))$

Def Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has densities p_θ, and is identifiable. We say $\mathcal{P}$ has monotone likelihood ratios (MLR) if there is some statistic $T(X)$ s.t.

$\dfrac{p_{\theta_2}}{p_{\theta_1}}(X)$ is a nondecreasing function of $T(X)$, for any $\theta_1 < \theta_2$ [same $T(X)$ for all θ's]

$\left(\frac{c}{0} = \infty \text{ if } c > 0, \ \frac{0}{0} \text{ undef.}\right)$

Ex. Exp. fam: $e^{(\eta_1 - \eta_0)\sum T(x_i) - n(A(\eta_1) - A(\eta_0))} \nearrow$ in $\sum T(x_i)$

Theorem Assume $\mathcal{P}$ has MLR, test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$ at level $\alpha = (0,1)$

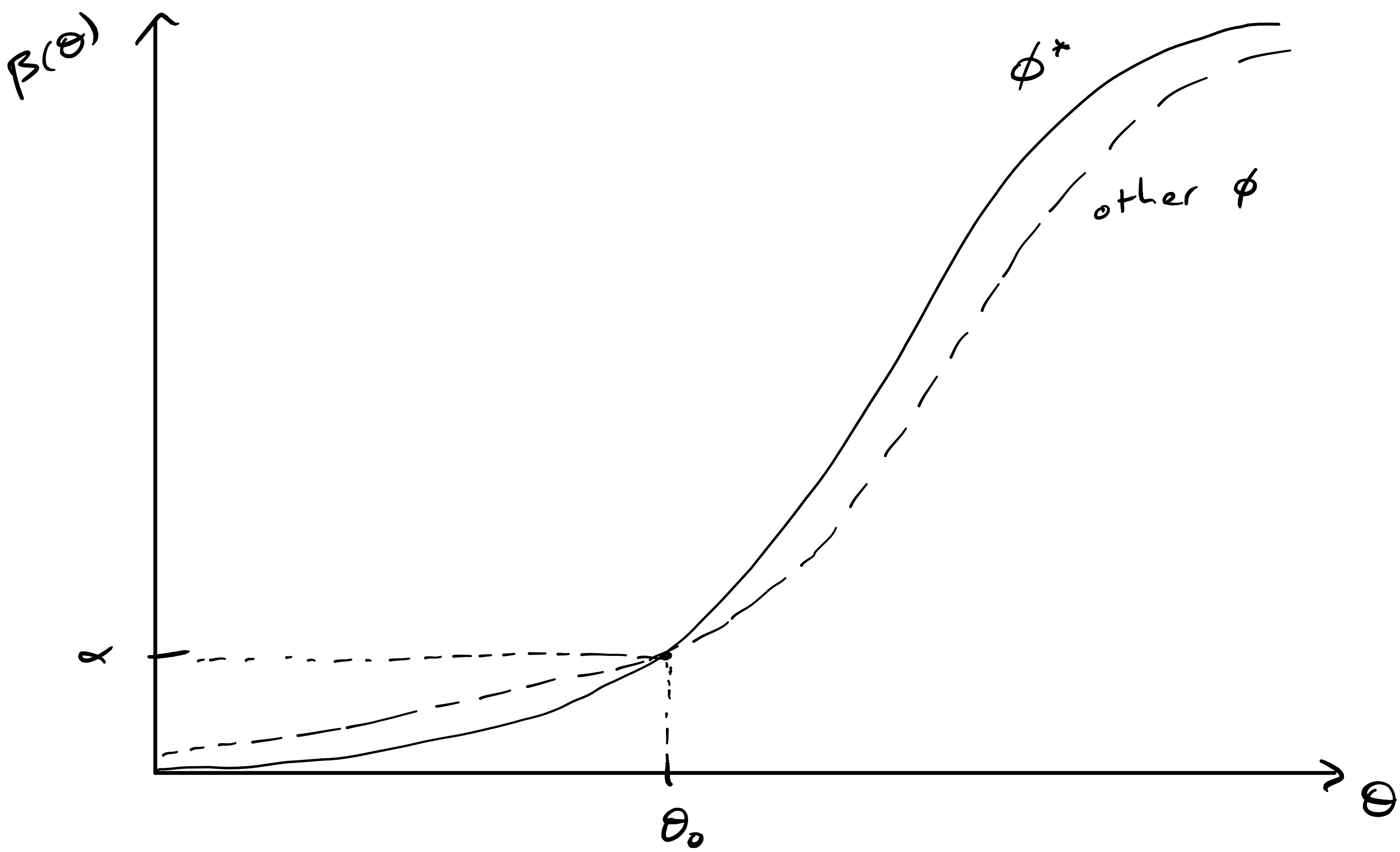
Let $\phi^*(x) = \begin{cases} 0 & T(x) < c \\ \gamma & T(x) = c \\ 1 & T(x) > c \end{cases}$,

with c, γ chosen so $\mathbb{E}_{\theta_0} \phi^*(X) = \alpha \in (0,1)$

a) ϕ^* is a UMP level-α test

b) $\beta(\theta) = \mathbb{E}_\theta \phi^*(X)$ is non-decreasing in θ, strictly incr. wherever $\beta(\theta) \in (0,1)$

c) If $\theta_1 < \theta_0$ then ϕ^* minimizes $\mathbb{E}_{\theta_1} \phi(X)$ among all tests with $\mathbb{E}_{\theta_0} \phi(X) = \alpha$



Proof

b) Suppose $\theta_1 < \theta_2$, then $\frac{p_{\theta_2}}{p_{\theta_1}}(x)$ is a nondecreasing function of $T(x)$

maybe not "the"

$\Rightarrow$ ϕ^* is a LRT for $H_0: \theta = \theta_1$ vs $H_1: \theta = \theta_2$ at level $\tilde{\alpha} = \mathbb{E}_{\theta_1} \phi^*(X)$

By Cor. 12.4, $\mathbb{E}_{\theta_2} \phi(X) \geq \mathbb{E}_{\theta_1} \phi(X)$, strict ineq. unless both $= 0$ or 1

a) Suppose $\theta_1 > \theta_0$ and $\tilde{\phi}$ has level $\leq \alpha$

$\Rightarrow$ $\mathbb{E}_{\theta_1} \phi^*(X) \geq \mathbb{E}_1 \tilde{\phi}(X)$ since ϕ^* is a LRT for $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$

c) $\theta_1 < \theta_0$, assume $\mathbb{E}_{\theta_0} \tilde{\phi}(X) = \mathbb{E}_{\theta_0} \phi^*(X) = \alpha$

Both $1-\phi^*$, $1-\tilde{\phi}$ are tests of $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ both have sig. level $1-\alpha$

$1-\phi^*$ is a LRT since $\frac{p_1}{p_0}(X)$ is non-incr. in $T(X)$

$\Rightarrow$ $\mathbb{E}_{\theta_1}(1-\tilde{\phi}) \leq \mathbb{E}_{\theta_1}(1-\phi^*) = 1-\alpha$ $\boxtimes$

Intuition ϕ^* is a LRT for $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ for any pair $\theta_0 < \theta_1$ (sig. level depends on θ_0)

[This lets us extend our simple vs. simple result to (a very special case of) composite vs comp.]

Two-sided Alternatives

Setup: $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$, $\theta_0 \in \Theta^\circ$

Test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

(Can be generalized naturally to $H_0 : \theta \in [\theta_1, \theta_2]$)

Def A real-valued statistic $T(X)$ is stochastically increasing in θ if

$$\mathbb{P}_\theta(T(X) \leq t) \text{ is non-incr. in } \theta, \ \forall t$$

Assume $T(X)$ is a stochastically increasing summary test statistic

Ex $X_i \overset{iid}{\sim} p(x-\theta)$ (location family)

$T(X) =$ sample mean / median

Ex $X_i \overset{iid}{\sim} \frac{1}{\theta} p(x/\theta)$ (scale family)

$T(X) = \sum X_i^2$ or $\text{median}(|X_1|, \ldots, |X_n|)$

Two-tailed test rejects when $T(X)$ is "extreme"

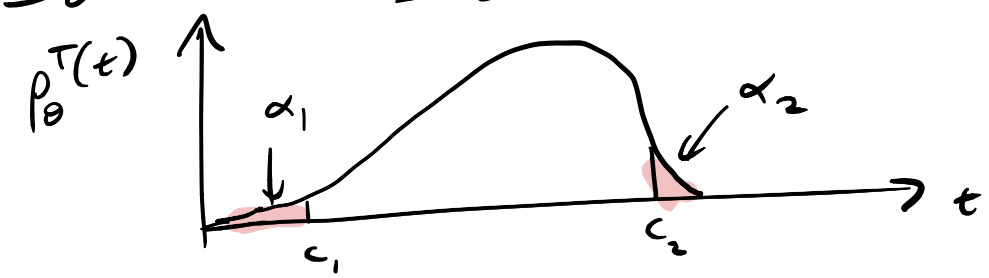
$$\phi(x) = \begin{cases} 1 & T(X) > c_2 \text{ or } T(X) < c_1 \\ 0 & T(X) \in (c_1, c_2) \\ \gamma_i & T(X) = c_i \end{cases}$$

Let $\alpha_1 = \mathbb{P}_{\theta_0}(T < c_1) + \gamma_1 \mathbb{P}_{\theta_0}(T = c_1)$

α_2 similar for upper tail

Need $\alpha_1 + \alpha_2 = \alpha$, but how to balance?

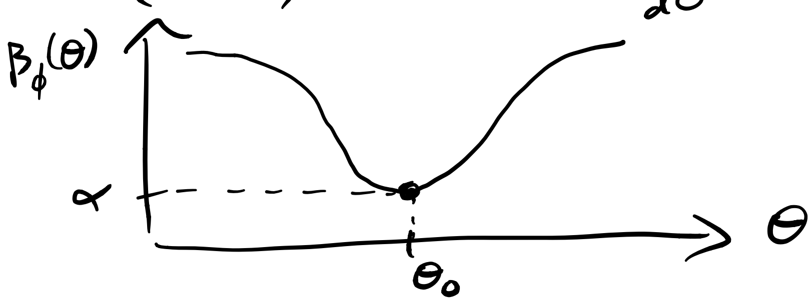
Idea 1: Equal-tailed test : $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$



Def $\phi(X)$ is unbiased if $\inf_{\theta \in \Theta_1} \mathbb{E}_\theta \phi(x) \geq \alpha$

Idea 2: Unbiased test : ensure $\min \beta_\phi(\theta) = \alpha$

(usually $\Longleftrightarrow$ $\frac{d\beta_\phi}{d\theta}(\theta_0) = 0$)



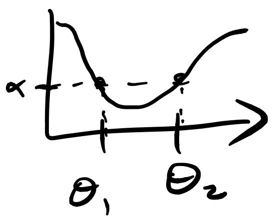
Theorem Assume $X_i \overset{iid}{\sim} e^{\theta T(x) - A(\eta)} h(x)$

$H_0 : \theta \in [\theta_1, \theta_2]$ vs $H_1 : \theta < \theta_1$ or $\theta > \theta_2$

(possibly $\theta_1 = \theta_2$)

Then

a) The unbiased test $\overbrace{\text{based on}}^{\text{(rejecting for extreme values of)}}$ $\sum T(X_i)$ with sig. level $= \alpha$ is UMP among all unbiased tests (UMPU)



b) If $\theta_1 < \theta_2$ the UMPU test can be found by solving for c_i, γ_i s.t. $\mathbb{E}_{\theta_1} \phi = \mathbb{E}_{\theta_2} \phi = \alpha$

c) If $\theta_1 = \theta_2 = \theta_0$ the UMPU test can be found by solving for c_i, γ_i s.t. $\mathbb{E}_{\theta_0} \phi(X) = \alpha$ and

$$\frac{d\beta_\phi}{d\theta}(\theta_0) = \mathbb{E}_{\theta_0}\left[\sum T(X_i)(\phi(x) - \alpha)\right] = 0$$

(Proof in Keener)

p-Values

<u>Informal definition:</u> Suppose $\phi(X)$ rejects for large values of $T(X)$.

$$p(x) = \text{"}\mathbb{P}_{H_0}\left(T(X) \geq T(x)\right)\text{"}$$

$$= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(T(X) \geq T(x)\right)$$

<u>Ex</u> $X \sim N(\theta, 1)$ $\quad H_0: \theta = 0$ vs. $H_1: \theta \neq 0$

Two-sided test rejects for large $T(X) = |X|$

$$\left(\Leftrightarrow \phi_\alpha(X) = 1\{|X| > z_{\alpha/2}\}\right)$$

The two-sided p-value is $p(X)$ where

$$p(x) = \mathbb{P}_0\left(|X| > |x|\right)$$

$$= 2\left(1 - \Phi(|x|)\right)$$

For $H_0: |\theta| < \delta$ vs. $H_1: |\theta| > \delta$:

$$p(x) = \mathbb{P}_\delta\left(|X| > |x|\right) \quad \left(= \mathbb{P}_{-\delta}(\text{ " })\right)$$

$$= 1 - \Phi(|x| - \delta) + \Phi(-|x| - \delta)$$

etc.-

Formal definition: $\mathcal{X}, \Theta_0, \Theta$.

Assume we have a test ϕ_α for each significance level, $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \phi_\alpha(X) \le \alpha$

(non-randomized case: $\phi_\alpha = 1\{x \in R_\alpha\}$)

Assume tests are monotone in α:
if $\alpha_1 \le \alpha_2$ then $\phi_{\alpha_1}(x) \le \phi_{\alpha_2}(x)$

(non-randomized: $R_{\alpha_1} \subseteq R_{\alpha_2}$)

Then $p(x) = \inf\{\alpha : \phi_\alpha(x) = 1\}$

$(= \inf\{\alpha : x \in R_\alpha\})$

(possible to define randomized p-value but not worth it)

Note $p(x) \le \alpha \iff \phi_{\tilde{\alpha}}(x) = 1 \quad \forall\, \tilde{\alpha} > \alpha$

For $\theta \in \Theta_0$, $\mathbb{P}_\theta(p(X) \le \alpha) \le \inf_{\tilde{\alpha} > \alpha} \underbrace{\mathbb{P}_\theta(\phi_{\tilde{\alpha}}(X) = 1)}_{\le \tilde{\alpha}} \le \alpha$

$\Rightarrow$ p-value <u>stochastically dominates</u> $U[0,1]$

If ϕ_α rejects for large $T(X)$, coincides with informal definition

Note the p-value depends on

- the model & null hyp.,
- the data, AND
- the choice of test

<u>Ex</u> $X \sim N_d(\theta, I_d)$ $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$

We can use $T_1(x) = \|X\|^2$ (χ^2 test)

or $T_2(X) = \|X\|_\infty$ (max test)

$= \max_i |X_i|$

<u>Very</u> different p-values / power if d large
(choice reflects belief about whether θ is sparse)

Confidence Sets

[Accept / reject decision only so interesting:
- usually we care how big θ is
- tiny p-value doesn't imply big θ
  (big p-value doesn't imply small θ either]

<u>Def</u> $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

$C(X)$ is a <u>$1-\alpha$ confidence set</u> for $g(\theta)$ if

$$\mathbb{P}_\theta\big(C(X) \ni g(\theta)\big) \geq 1-\alpha, \quad \forall \theta \in \Theta$$

subject verb object

We say $C(X)$ <u>covers</u> $g(\theta)$ if $C(X) \ni g(\theta)$

$\mathbb{P}_\theta(C(X) \ni g(\theta))$ is <u>coverage probability</u>

$\inf_\theta \mathbb{P}_\theta(C \ni g(\theta))$ is <u>conf. level</u>

<u>Notes</u>
- $C(X)$ is random, not $g(\theta)$
- Often misinterpreted as Bayesian guarantee
- Say "$C(X)$ has a 95% chance of covering"
  NOT "$g(\theta)$ has a 95% chance of being in C"
  NEVER "95% chance $g(\theta) \in [0.5, 1.5]$" (e.g.)

Duality of Testing & Confidence Sets

Suppose we have a level-α test $\phi(x;a)$
of $H_0: g(\theta) = a$ vs. $H_1: g(\theta) \neq a$, $\forall a \in g(\Theta)$

We can use it to make a confidence set for $g(\theta)$:

Let $C(X) = \{a : \phi(x;a) < 1\}$

$=$ "all non-rejected values of θ"

Then $\mathbb{P}_\theta(C(X) \not\ni g(\theta)) = \mathbb{P}_\theta(\phi(x; g(\theta)) = 1)$

$\leq \alpha \qquad \forall \theta$

Alternatively, suppose $C(X)$ is a $1-\alpha$ confidence set for $g(\theta)$.

We can use C to contruct a test $\phi(X)$ of
$H_0: g(\theta) = a$ vs. $H_1: g(\theta) \neq a$

$$\phi(X) = 1\{a \notin C(X)\}$$

For θ s.t. $g(\theta) = a$:

$$\mathbb{E}_\theta \phi(X) = \mathbb{P}_\theta(C(X) \not\ni g(\theta)) \leq \alpha$$

This is called <u>inversting the test</u>

Confidence Intervals / Bounds

If $C(X) = [C_1(X), C_2(X)]$ we say

$C(X)$ is a confidence interval (CI)

$C(X) = [C_1(X), \infty)$: lower conf. bd. (LCB)

$C(X) = (-\infty, C_2(X)]$: upper conf bd. (UCB)

We usually get LCB / UCB by inverting

a one-sided test in appropriate direction

Called uniformly most accurate (UMA) if test UMP

Get CI by inverting a two-sided test

Called UMAU if test is UMPU

Ex $X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta}$ $x > 0,\ \theta > 0$

CDF $\mathbb{P}_\theta(X \le x) = 1 - e^{-x/\theta}$

LCB: Invert test for $H_0: \theta \le \theta_0$

Solve $\alpha = \mathbb{P}_{\theta_0}(X > c(\theta_0)) = e^{-c(\theta_0)/\theta_0}$

$c(\theta_0) = -\theta_0 \log\alpha$ (>0)

$X \le c(\theta_0) \Rightarrow \theta_0 \ge \frac{X}{-\log\alpha}$

$C(X) = \left[\frac{X}{-\log\alpha}, \infty\right)$

UCB: Similar, $C(X) = \left(-\infty, \frac{X}{-\log(1-\alpha)}\right]$

Equal-tailed CI:

Invert equal-tailed test of $H_0: \theta = \theta_0$

$$\underbrace{\phi_\alpha^{2T}(X)}_{\substack{\text{2-tailed} \\ H_0:\theta=\theta_0}} = \underbrace{\phi_{\alpha/2}^{\ge\theta_0}(X)}_{H_0:\theta\ge\theta_0} + \underbrace{\phi_{\alpha/2}^{\le\theta_0}(X)}_{H_0:\theta\le\theta_0}$$

$$\Rightarrow C(X) = \left[\frac{X}{-\log\alpha/2}, \infty\right) \cap \left(-\infty, \frac{X}{-\log(1-\alpha/2)}\right]$$

$$= \left[\frac{X}{-\log\alpha/2}, \frac{X}{-\log(1-\alpha/2)}\right]$$

(Mis-) Interpreting Hypothesis Tests

Hypothesis tests ubiquitous in science

Common misinterpretations:

1) $p < 0.05$ therefore "there is an effect"
   or "the effect size = the estimate"

2) $p > 0.05$ therefore "there is no effect"

3) $p = 10^{-6}$ therefore "the effect is huge"

4) $p = 10^{-6}$ therefore "the data are signif."
   and everything about our model
   is correct in most naive interp.

5) Effect CI for men is $[0.2, 3.2]$,
   for women is $[-0.2, 2.8]$ therefore
   "there is an effect for men and not
   for women."

⋮

378) We rejected a specific, parametric null
   model therefore something interesting is happening

How to interpret testing

Learning about the world from data is not easy or automatic!

Hypothesis tests let us ask specific questions about specific data sets under specific modeling assumptions, using specific testing method.

All of these choices bear on the interpretation.

Top-tier medical journals let people publish claims, reporting p-values without saying what model was used or what test was employed

THIS IS ABSOLUTELY OUTRAGEOUS

Hyp. tests can be a good companion to critical thinking, never a substitute

"All models are wrong, some are useful" but need experience and theory to understand when assumptions do or don't cause real trouble

Conceptual Objections

<u>Q1:</u> Why should I test $H_0: \theta = 0$? No θ is ever exactly 0.

<u>A1:</u> a) Test $H_0: |\theta| \le \delta$ if you want!

If $s.e.(\hat{\theta}) = 10\delta$, not much difference.

b) Most two-sided tests justify <u>directional inference</u>:

"If $T > c_\alpha$ declare $\theta > 0$, if $T < c_1$ declare $\theta < 0$" with $\mathbb{P}(\text{false claim}) \le \alpha$

c) Harder to answer in non-parametric problems, e.g. $H_0: P = Q$ vs $H_1: P \ne Q$ for perm. test, but alternative frameworks like Bayes force very strong assumptions on us.

<u>Q2:</u> People only like frequentist results like p-values, CIs because they mistake them for Bayesian results.

95% chance $C(X) \ni \theta$ is misinterpreted as a claim about $p(\theta \mid X)$.

<u>A2:</u> a) True, but subjective Bayesian results often misinterpreted as "<u>the</u> posterior dist. of θ" when really should be "<u>my</u> posterior opinion about θ"

b) "Objective" Bayesian credible intervals are even worse: "nobody's posterior opinion about θ"

c) <u>Caveat:</u> in some simple, low-dim., high signal settings, can maybe say "any reasonable person's posterior opinion about θ." Then Bayes methods probably best!

<u>Q3</u>: p-values ignore $P(\text{Data} \mid H_1)$ and only look at $P(\text{Data} \mid H_0)$. Data might be <u>more</u> likely under H_0 but still reject.

<u>A3</u>: $P(\text{Data} \mid H_0)$, $P(\text{Data} \mid H_1)$ only make sense for simple null / alternative. Even in $N(\theta, 1)$ $H_0: \theta = 0$, what is $P(X = \ \mid H_1)$?

If H_1 is vague prior like $\theta \sim N(0, 10^6)$, then $X \sim N(0, 10^6 + 1)$ and $P(4 \mid H_1) \ll P(4 \mid H_0)$

Will scientists understand this??

Even bigger problems in high-dim, hierarchical, or nonparametric priors.

Q4: Scientists always misuse hypothesis testing, so we should switch to something else (Confidence intervals / Bayes / weird new idea)

A4: a) CIs great, but just a re-packaging of hypothesis tests. (Might still be good for staving off some common misinterpretations by naifs)

b) Bayes has its uses, but forcing scientists to make more choices / assumptions is not going to solve problem of scientist incentives / ignorance

c) Most weird new ideas have bigger issues but just haven't been criticized much yet b/c no one but proponents care.

d) Statistical inference will never be idiot-proof, b/c science / critical thinking are not idiot-proof. Engineers have to learn calculus, learning what a p-value means is not that hard. Suck it up, social scientists! (and ask for help)