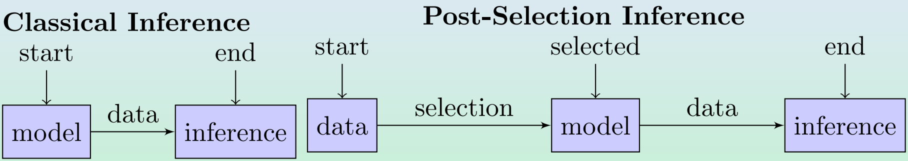


Post-Selection Inference

Todd Kuffner
Washington University in St. Louis

PhyStat ν 2016
Fermilab

Setting the mood

Cupuaçu and Octavia

Preliminary Comment: the p-value controversy (K-Walker 2016)

2015 *Basic and Applied Social Psychology* bans use of p-values (Trafimow & Marks, 2015)

- 'fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it'

Preliminary Comment: the p-value controversy (K-Walker 2016)

2015 *Basic and Applied Social Psychology* bans use of p-values (Trafimow & Marks, 2015)

- 'fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it'

2015 *International Society for Bayesian Analysis* doesn't gloat! (Bulletin, March 2015)

- 'it was inspired by a nihilistic anti-statistical stance, backed by an apparent lack of understanding of the nature of statistical inference, rather than a call for saner and safer statistical practice' (Christian P. Robert)

Preliminary Comment: the p-value controversy (K-Walker 2016)

2015 *Basic and Applied Social Psychology* bans use of p-values (Trafimow & Marks, 2015)

- ‘fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it’

2015 *International Society for Bayesian Analysis* doesn't gloat! (Bulletin, March 2015)

- ‘it was inspired by a nihilistic anti-statistical stance, backed by an apparent lack of understanding of the nature of statistical inference, rather than a call for saner and safer statistical practice’ (Christian P. Robert)

2016 *American Statistical Association* (Wasserstein & Lazar, 2016)

- ‘Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data ... would be equal to or more extreme than its observed value.’

What is a p-value? The setting

Given $X = \{X_1, \ldots, X_n\}$ (i.i.d.). Goal: test H

$S(X)$ =**sufficient statistic**; for simplicity assume $\dim(S(X)) = 1$
$\Rightarrow$ e.g. $S(X)$ takes values in $\mathbb{R}$

Two decisions: R or R^C
$\Rightarrow$ real line split into regions R and R^C

$$S(X) \in R \quad \text{or } S(X) \in R^C$$

Let $\alpha \in (0, 1)$, and define $R_\alpha \equiv R(\alpha)$; for simplicity, assume R_α of form $[c_\alpha, \infty)$
$\Rightarrow$ test rejects H if $S(X) \geq c_\alpha$

The Formal Definition of a p-value

p-value defined in setting where rejection regions are nested sets

$$\alpha < \alpha' \Rightarrow R_\alpha \subset R_{\alpha'}$$

Definition: The p-value, $p \equiv \hat{\alpha}$, is (Lehmann & Romano, 2005, §3.3)

$$\hat{\alpha} \equiv \hat{\alpha}_{S(X)} = \inf_{0<\alpha<1} \{\alpha : S(X) \in R_\alpha\}.$$

The p-value *function*, $\hat{\alpha} = f(S(X))$, for suitable map $f : S(X) \mapsto \hat{\alpha}$ is a

bijection from $\mathbb{R}$ to $(0, 1)$

A p-value is not *itself* defined as a probability, but rather takes values on the same scale as something formally defined as a probability.

For the curious

Goal: show that the p-value, $\hat{\alpha} = f(S(X))$ for suitable choice of map $f : S(X) \mapsto \hat{\alpha}$, is a bijection from $\mathbb{R}$ to $(0, 1)$. The actual form of $\hat{\alpha} = f(S(X))$ is specific to the model, hypothesis and test.

Write $S \equiv S(X)$ for simplicity.

Step 1 First, we note that the function $\hat{\alpha}$ is well-defined. That is, given two values S_1, S_2 such that $S_1 = S_2$, we have that $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$.

Step 2 Next, we require that $\hat{\alpha}$ is injective (one-to-one). To show this, we need that if $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$, then $S_1 = S_2$. Suppose that $S_1 \neq S_2$. If we can show that this implies $\hat{\alpha}_{S_1} \neq \hat{\alpha}_{S_2}$, this will establish injectivity. Without loss of generality, suppose $S_2 < S_1$. Then there exists an α' such that $S_1 \in R_{\alpha'}$ but $S_2 \notin R_{\alpha'}$. Therefore, if $S_2 < S_1$, it cannot be the case that $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$.

Step 3 Finally, we must have that $\hat{\alpha}$ is surjective (onto). For surjectivity, we require that for every $\beta \in (0, 1)$, there exists an $\tilde{S} \in \mathbb{R}$ such that

$$\hat{\alpha} = \inf_{0<\alpha<1} \{\alpha : \tilde{S} \in R_\alpha\} = \beta,$$

which is seen by choosing $\tilde{S} = \inf\{S : S \in R_\beta\}$.

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.
- Assume type I error probability α set <u>in advance</u>.

Example

Suppose $X = X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.
- Assume type I error probability α set <u>in advance</u>.
- $S(X) = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.
- Assume type I error probability α set <u>in advance</u>.
- $S(X) = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$
- p-value assigns value in $(0, 1)$ to each value in sample space of $\bar{X}$

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.
- Assume type I error probability α set <u>in advance</u>.
- $S(X) = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$
- p-value assigns value in $(0, 1)$ to each value in sample space of $\bar{X}$
- p-value is merely transformation of $\bar{X}$ $\Rightarrow$ p-value also sufficient for test

Example

Suppose $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

- Wish to test $H : \theta = 0$.
- Assume type I error probability α set <u>in advance</u>.
- $S(X) = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$
- p-value assigns value in $(0, 1)$ to each value in sample space of $\bar{X}$
- p-value is merely transformation of $\bar{X}$ $\Rightarrow$ p-value also sufficient for test

Rejecting use of p-value conceptually equivalent to rejecting use of $\bar{X}$

Source of the controversy: no decision rule

Three types of ‘testers’:

Source of the controversy: no decision rule

Three types of 'testers':

Tester 1 sets α before seeing data; computes observed value of p based on sample, rejects H if $p < \alpha$

Source of the controversy: no decision rule

Three types of 'testers':

- Tester 1 sets α before seeing data; computes observed value of p based on sample, rejects H if $p < \alpha$
- Tester 2 first computes observed value of p, then claims his/her α would have been bigger had he/she actually chosen one beforehand

Source of the controversy: no decision rule

Three types of 'testers':

Tester 1 sets α before seeing data; computes observed value of p based on sample, rejects H if $p < \alpha$

Tester 2 first computes observed value of p, then claims his/her α would have been bigger had he/she actually chosen one beforehand

Tester 3 first computes observed value of p, believes it is small, and subsequently rejects H; he/she believes α is actually the observed value of p, since following Tester 2's approach, *any* $\alpha > p$ will work; therefore, he/she argues: why not choose an α just above p and view that as the type I error probability?

Source of the controversy: no decision rule

Three types of 'testers':

- **Tester 1** sets α before seeing data; computes observed value of p based on sample, rejects H if $p < \alpha$
- **Tester 2** first computes observed value of p, then claims his/her α would have been bigger had he/she actually chosen one beforehand
- **Tester 3** first computes observed value of p, believes it is small, and subsequently rejects H; he/she believes α is actually the observed value of p, since following Tester 2's approach, *any* $\alpha > p$ will work; therefore, he/she argues: why not choose an α just above p and view that as the type I error probability?

For Testers 2 and 3, there is no decision rule; instead a *heuristic:* that small value of p is sufficient to reject the hypothesis.

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

Using the data to select the model introduces additional uncertainty

$\Rightarrow$ invalidates classical inference

The problem of post-selection inference

Classical inference assumes the model is chosen independently of the data.

Using the data to select the model introduces additional uncertainty

$\Rightarrow$ invalidates classical inference

Do you believe me?

Example

R. Lockhart, J. Taylor, Ryan Tibshirani, Rob Tibshirani (2014), 'A significance test for the lasso', *Annals of Statistics*.

Classical inference for linear regression: two fixed, nested models

Model A variable indices $M \subset \{1, \ldots, p\}$

Model B variable indices $M \cup \{j\}$

Goal: test significance of jth predictor in Model B

Compute drop in RSS from regression on $M \cup \{j\}$ and M

$$R_j = (\mathrm{RSS}_M - \mathrm{RSS}_{M \cup \{j\}})/\sigma^2 \quad \text{versus} \quad \underbrace{\chi_1^2}_{\text{for } \sigma^2 \text{ known}}$$

Post-selection inference: first use selection procedure, then do inference

- want to do the same test as above for Models A and B which <u>are not fixed</u>, but rather outputs of selection procedure

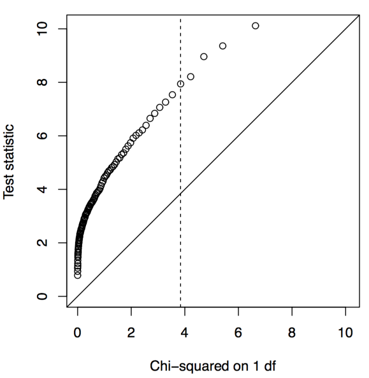
e.g. forward stepwise

- start with empty model $M = \emptyset$
- enter predictors one at a time: choose predictor j giving largest drop in RSS
- FS chooses j at each step to to maximize $R_j = (\mathrm{RSS}_M - \mathrm{RSS}_{M\cup\{j\}})/\sigma^2$
- each $R_j \sim \chi_1^2$ (under null)

$\Rightarrow$ max possible R_j stochastically larger than χ_1^2 under null

Illustration

Compare quantiles of R_1 in forward stepwise regression, i.e. chi-square for <u>first</u> predictor to enter versus those of χ_1^2 variable, when $\beta_k = 0 \ \forall k = 1, \ldots, p$.



$n = 100$, $p = 10$ (orthogonal); all true coefficients are zero; 1000 simulations of statistic R_1, versus χ_1^2 distribution; dotted line is 0.95 quantile of χ_1^2

At 0.05 level, using χ_1^2 quantile (3.84) has *actual* type I error probability of 0.39

Example: *File Drawer Effect* (Fithian, 2015)

Observe $X_1, \dots, X_n$ independently $\sim\mathcal{N}(\mu_i, 1)$

Suppose you focus on ‘apparently’ large effects, $|X_i| > 1$:

$$\hat{I} = \{i : |Y_i| > 1\}$$

Goal: test $H_{0,i} : \mu_i = 0$ for each $i \in \hat{I}$ at level 0.05.

- Usual approach: reject $H_{0,i}$ when $|Y_i| > 1.96$

Not Valid Due to Selection *Why?*

Seems counterintuitive: probability of falsely rejecting a given $H_{0,i}$ is *still* α, since most of the time $H_{0,i}$ is <u>not tested at all</u>.

Problem: for those hypotheses selected for testing, type I error rate is possibly much higher than α

Proof of concept

- let n_0 be # of true null effects
- assume $n_0 \to \infty$ as $n \to \infty$

Long-run fraction of errors among the true nulls we test:

$$\frac{\#\text{ false rejections}}{\#\text{ true nulls selected}} = \frac{\frac{1}{n_0}\sum_{i:H_{0,i}\text{ true}} 1\{i \in \hat{I},\ \text{reject } H_{0,i}\}}{\frac{1}{n_0}\sum_{i:H_{0,i}\text{ true}} 1\{i \in \hat{I}\}}$$

$$\to \frac{\mathbb{P}_{H_{0,i}}(i \in \hat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \hat{I})}$$

$$= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \hat{I})$$

For nominal test, this is $\Phi(-1.96)/\Phi(-1) \approx 0.16$

Why should particle physicists care?

We hate false discovery as much as you do.

- control of FDR, FCR, FWER are key desiderata in PSI
- applications are endless: need to formalize the informal ‘data snooping’ (adaptive selection) process to properly account for uncertainty

Possible problems?

- select minimum signal threshold, then do inference for selected signals
- selection of ‘events’
- data transformations based on data snooping

Broad Classification of PSI

1. **data splitting** (Cox, Wasserman) and **data carving** (Fithian)

   idea: the source of the problem is using the same data for selection and inference; solution: use some data for selection, the rest for inference
2. **high-dimensional inference** (the Swiss, signal processing, machine learning, econometrics):

   idea: ignore selection, view as single procedure followed by interval correction; not *really* PSI?
3. **simultaneous inference** (Benjamini, Yekutieli, Heller, Wharton)

   idea: control FDR for all models ever under consideration by selection procedure; solution: fix the confidence intervals
4. **selective inference** (Benjamini, Yekutieli, Stanford)

   idea: inference for selected hypotheses

Point of Contention

Suppose we have

$$\text{Full model:} \quad Y_i = \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i$$

Apply selection procedure, result is

$$\text{Selected/sub- model:} \quad Y_i = \sum_{k \in \hat{M}} \beta_{ik} x_{ik} + \gamma_i$$

with $\hat{M} \subseteq \{1, \ldots, p\}$.

Parameter spaces are not the same; should we do inference about full model parameters or submodel parameters?

More on Selective Inference

The selection of a model is a random event.

- helpful toy example: the set of selected variables in regression is a random set; hypotheses are only tested for selected variables, thus the hypotheses are random
- to condition on selection event, need to characterize this event in a manner suitable to uncertainty quantification

e.g. Lasso and forward stepwise partition $\mathbb{R}^n$ into convex polyhedra: if $y \in \mathrm{ConvPoly}_m$, then model m is selected

Are Bayesians immune?

Dawid (1994): selection should have no effect

> *Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data.*

Yekutieli (2012, 'Adjusted Bayesian inference for selected parameters,' *JRSSB*):

Actually, selection <u>can</u> affect Bayesian inference

Bayesian inference for parameters selected after viewing the data is a 'truncated' data problem.