Valid p-hacking via flipscores test

A. Vesely, P. Girardi, D. Lakens, G. Altoè, M. Pastore, A. Calcagnì, L. Finos

Statistical methods and models for complex data — Poster Session: 21 September 2022

Multiverse of analyses and results

Did people's willingness to vaccinate vary over the pandemic?

Each step of the analysis involves many justifiable choices:

- combining and transforming measurements
- handling missing data and outliers
- choosing a statistical model
- ...

Results with continuous predictors:

	contrast	sigma	t-stat	p-value	
LockDown - Pre	0.203	0.124	1.637	0.229	
Post - Pre	0.657	0.150	4.370	<0.001	***
Post - LockDown	0.454	0.139	3.278	0.003	**

Results after transforming predictors into categorical:

	contrast	sigma	t-stat	p-value	
LockDown - Pre	0.235	0.127	1.846	0.154	
Post - Pre	0.563	0.152	3.697	0.001	***
Post - LockDown	0.329	0.143	2.293	0.056	.

Etc: 81 plausible models from 3 different tansformations of 4 continuous predictors

p-hacking

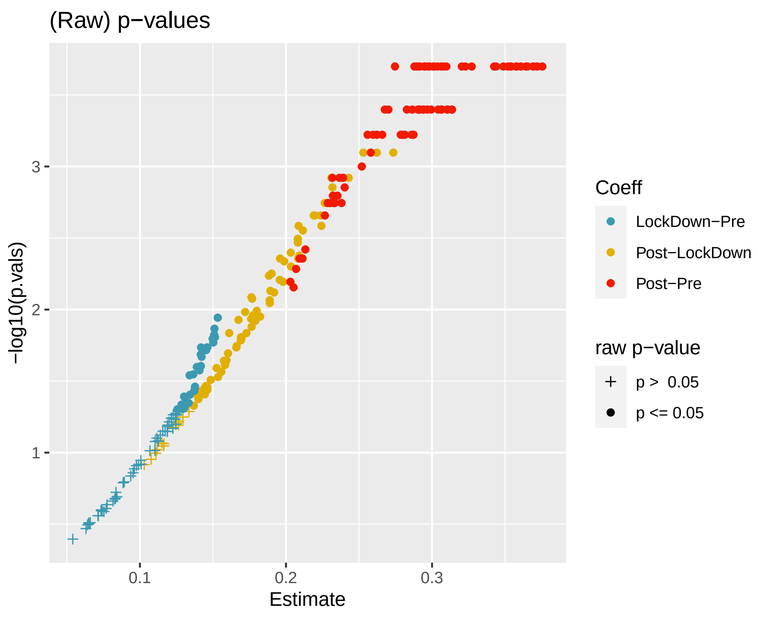
Performing many statistical tests and only reporting those that produce significant results

→ dramatically increases and undertakes the risk of false positives

→ replicability crisis

Multiverse analysis

A philosophy of reporting the outcomes of many different statistical analyses, showing how robust findings are



How to make formal inference in this framework?

Multiverse of models

Consider K plausible Generalized Linear Models, where for each model k

$$g_k(\mathbb{E}(y_{ki})) = \beta_k x_{ki} + \boldsymbol{\gamma}_k \mathbf{z}_{ki}, \qquad (i = 1, \dots, n)$$

- y_{ki}: response → outlier deletion or leverage point removal
- x_{ki} and $\mathbf{z}_{ki}$: transformed predictors → combination and transformation of variables
- β_k: parameter of interest
- $\boldsymbol{\gamma}_k$: nuisance

Does X have a non-null effect on the response...

- in at least one of the models?
- in how many models?
- in which models?

Hypothesis testing in a single model

Univariate hp: is there a non-null effect in model k?

$$H_{0k}: \ \beta_k = 0$$

Sign-flip score test

- Score test statistic:

$$T_k^1 = T_k^{\text{obs}} = \sum_{i=1}^{n} \nu_{ki}$$

where

$$\nu_{ki} = \frac{\partial}{\partial \beta_k} \log f_{\beta_k, \boldsymbol{\gamma}_k, x_{ki}, \mathbf{z}_{ki}}(y_{ki}) \mid_{\boldsymbol{\gamma}_k = \hat{\boldsymbol{\gamma}}_k, \beta_k = 0}$$

is the individual score contribution of observation i

- Permutation test statistics:

$$T_k^b = \sum_{i=1}^{n} \pm \nu_{ki} \qquad (b = 2, \dots, B)$$

from random sign flips

Under H_{0k}:

- $\mathbb{E}(T_k^{\text{obs}}) = \mathbb{E}(T_k^b)$
- $T_k^{\text{obs}} \overset{d}{=} T_k^b$ asymptotically

$$\text{p-value}_k = \frac{\#_b(T_k^b \geq T_k^{\text{obs}})}{B}$$

Hypothesis testing in the multiverse

Joint null hypothesis: is there a non-null effect in at least one of the plausible models?

$$H_0 : \bigcap_{k=1}^{K} H_{0k} : \ \beta_k = 0 \text{ for all } k = 1, \dots K$$

Sign-flip score test

- K score test statistics: $(T_1^{\text{obs}}, \dots, T_K^{\text{obs}})$
- Permutation test statistics: $(T_1^b, \dots, T_K^b)$ obtained by jointly flipping the signs of the K-variate contributions

$$\pm(\nu_{1i}, \dots, \nu_{Ki})$$

Global test statistic:

$$\begin{array}{c} \\ \text{sign flips} \end{array} \begin{array}{c} \text{models} \\ \begin{pmatrix} T_1^{\text{obs}} & \dots & T_K^{\text{obs}} \\ T_1^2 & \dots & T_K^2 \\ \vdots & & \vdots \\ T_1^B & \dots & T_K^B \end{pmatrix} \end{array} \xrightarrow{\ \psi\ } \begin{array}{c} \text{joint} \\ \begin{pmatrix} T^{\text{obs}} \\ T^2 \\ \vdots \\ T^B \end{pmatrix} \end{array}$$

ψ: suitable combining function, such as the (weighted) mean and the maximum

Refinements

- effective score → more powerful
- standardized effective score → 'almost' exact type I error in finite sample

Properties

- can be used whenever one can write a score test (GLMs and much more)
- asymptotically exact (exact, in practice)
- very robust to model – variance – misspecification, if the link function is correctly specified
- can be extended to the case of multiple parameters of interest

What is allowed

- any transformation of predictors and response
- any model
- any outlier deletion method

BUT all the models must be

- planned in advance
- valid (at least the right link)

Post-selection inference in multiverse analysis

Is there any non-null effect among the tested models?

Combine info from all models (e.g., $\psi = \max$)
→ overall p-value, weak FWER control

	stat	p-value	
LockDown - Pre	2.513	0.0382	*
Post - Pre	5.066	0.0002	***
Post - LockDown	3.894	0.0008	***

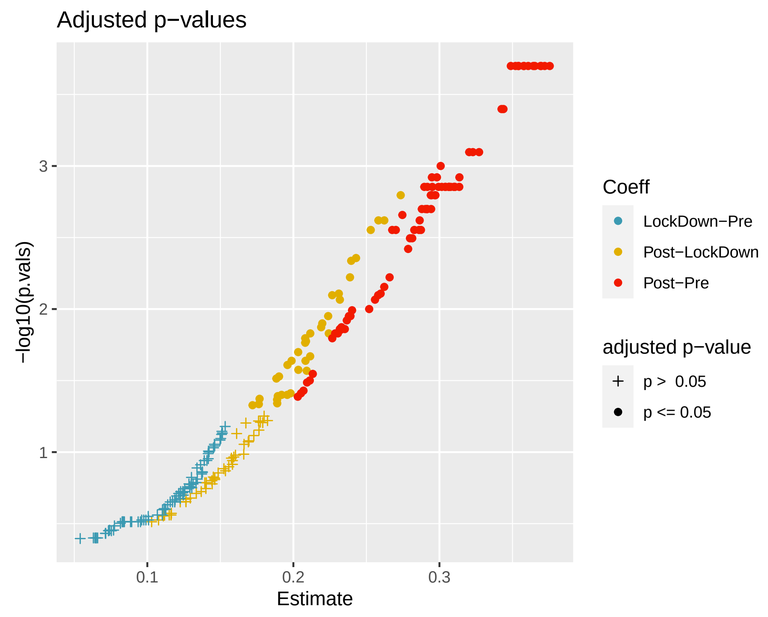
How many models are significant?

Closed testing (e.g., maxT-method)
→ lower $(1-\alpha)$-confidence bound for the True Discovery Proportion

	true discoveries	prop (%)
LockDown - Pre	0	0
Post - Pre	81	100
Post - LockDown	72	89

Which models are significant?

MaxT-method
→ adjusted p-values, strong FWER control



Results with continuous predictors:

	coeff	raw p-value	adj p-value	
LockDown - Pre	0.1006	0.1132	0.2814	
Post - Pre	0.3204	0.0002	0.0008	***
Post - LockDown	0.2198	0.0022	0.0126	*

Results after transforming predictors into categorical:

	coeff	raw p-value	adj p-value	
LockDown - Pre	0.1139	0.0754	0.2332	
Post - Pre	0.2305	0.0018	0.0148	*
Post - LockDown	0.116	0.0866	0.2678	

Etc.

Pick the model, choose the story to tell!

References

Finos, L. (2022). Jointest (R package). *github.com/livioivil/jointest*.

Hemerik, J., Goeman, J. J., Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *J. R. Statist. Soc. B*, **82(3)**, 841 – 864.

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*, **11(5)**, 702 – 712.

Contact information

Anna Vesely, postdoctoral research fellow
DPSS, University of Padova
anna.vesely@unipd.it
github.com/annavesely