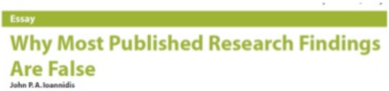
Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty

Dominik Rothenhäusler
Stanford University
joint work with Yujin Jeong

The replicability crisis

What is the problem?

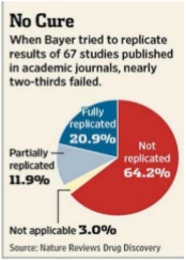


Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

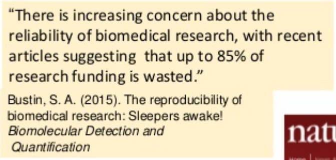
2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124



No Cure

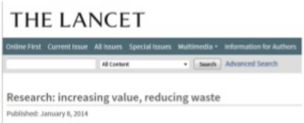
When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discovery



"There is increasing concern about the reliability of biomedical research, with recent articles suggesting that up to 85% of research funding is wasted."

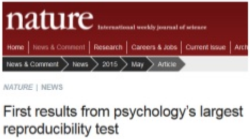
Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*



THE LANCET

Research: increasing value, reducing waste

Published: January 8, 2014



NATURE | NEWS

First results from psychology's largest reproducibility test

Source: Dorothy Bishop

What’s the reason for the crisis?

There are many reasons for the replication crisis: low power, p-hacking, convenience samples, correlated observations, distribution shift, . . .

Does statistical practice capture all relevant sources of variation??

Distributional uncertainty

Distribution shift, contaminations, confounding,... might lead to a sampling distribution that is different from the target distribution $\mathbb{P}^0$.

If the distributional perturbations have some (known) structure, we can address it via re-weighting, random effect modelling, sensitivity analysis or other statistical techniques. Here, we want to deal with unknown perturbations.

Distributional uncertainty

Running example: we are interested in a linear regression parameter

$$\theta(\mathbb{P}) = \arg\min_{\theta} \mathbb{E}_{\mathbb{P}}[(Y - X\theta)^2].$$

We observe i.i.d. data $(D_1, \ldots, D_n)$ from a perturbed distribution $\mathbb{P}^{\xi}$ and compute an estimator $\hat{\theta}(D_1, \ldots, D_n)$.

The error decomposes as

$$\hat{\theta} - \theta(\mathbb{P}^0) = \underbrace{\hat{\theta} - \theta(\mathbb{P}^{\xi})}_{\substack{\text{error due to}\\ \text{sampling}}} + \underbrace{\theta(\mathbb{P}^{\xi}) - \theta(\mathbb{P}^0)}_{\substack{\text{error due to}\\ \text{perturbation}}}.$$

Motivation

Some time ago, at a conference I presented a "distributional stability measure".

Audience member: "if you want this to be used widely, you have to integrate distributional stability and sampling uncertainty"

I agree! This would a) require little re-training for practitioners b) simplify decision-making c) integrate relatively easily with existing tools such as Bonferroni, FDR control . . .

Integrating sampling uncertainty and distributional uncertainty

Ideally, we would like to construct confidence intervals that cover the parameter of the target distribution $\theta(\mathbb{P}^0)$ (and not the contaminated parameter $\theta(\mathbb{P}^\xi)$).

Compared to sensitivity analysis in causal inference, we will NOT rely on user input how far $\mathbb{P}^\xi$ is from $\mathbb{P}^0$.

We will estimate the strength of the perturbations by evaluating model stability.

Related literature

- Many researchers recommend evaluating model stability to judge trustworthiness of statistical conclusions (Leamer 1993; Rosenbaum 2010; Yu 2013; Yu and Kumbier 2020; . . .)
- In causal inference, differently specified regressions are often used to estimate the size of omitted variable bias (Murphy and Topel 1990; Altonji, Elder, and Taber 2005a; Altonji et al. 2011; Oster, 2019)
- Stability principles have been used to infer causal relations based on heterogeneous data (Peters et al., 2016; Heinze-Deml et al., 2018; Bühlmann 2020)
- If the analyst chooses the final estimator out of a set of estimators in a data-driven fashion, inference has to account for the model selection step (Berk et al., 2013; Fithian et al., 2014;...)

What is the most generic distributional perturbation?

One can generate $\mathbb{P}^{\xi}$ by randomly up-weighting or down-weighting probabilities of events compared to the target distribution $\mathbb{P}^{0}$.

Example: the distributional perturbation model

For simplicity, we will focus on discrete distributions with $\mathbb{P}^0(X = x) = \frac{1}{m}$ for all $x \in \mathcal{X}$. Without loss of generality $\mathcal{X} = \{1, \ldots, m\}$.

Draw i.i.d. weights $\xi_k \geq 0$ with finite second moment. Set

$$\mathbb{P}^\xi(X = x) = \frac{\xi_x}{\sum_{k=1}^m \xi_k}$$

Draw $D_1, \ldots, D_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}^\xi$. Then, for all functions ψ

$$\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i) - \mathbb{E}_{\mathbb{P}^0}[\psi(D)])\right) = \left(1 + \frac{n}{m} \frac{\text{Var}(\xi)}{\mathbb{E}[\xi]^2}\right) \text{Var}(\psi(D)) + o(1),$$

where $D \sim \mathbb{P}^0$.

Our setting

Assumption (Simplified version)

Let $(D_1, \dots, D_n)$ be a data set such that for any bounded ψ with bounded total variation

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi(D_i) - \mathbb{E}_{\mathbb{P}^0}[\psi(D)]) \approx \mathcal{N}(0, \delta^2 Var(\psi(D))),$$

where $D \sim \mathbb{P}^0$ and $\delta > 0$ is unknown.

If the data is drawn i.i.d. from $\mathbb{P}^0$ this holds with $\delta = 1$. Thus, this can be seen as relaxing the i.i.d. assumption.

Assumption (Rigorous version)

Let $(D_1^n, \dots, D_n^n)$, $n \geq 1$ be a triangular array of random variables. For any bounded ψ with bounded total variation let

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi(D_i^n) - \mathbb{E}_{\mathbb{P}^0}[\psi(D)]) = \mathcal{N}(0, \delta^2 Var(\psi(D))) + o_p(1),$$

where $D \sim \mathbb{P}^0$ and $\delta > 0$ is unknown.

Since the data scientist only observes on data set $(D_1^n, \dots, D_n^n)$ for some fixed n, in the following for simplicity we just write $(D_1, \dots, D_n)$.

When does this assumption hold?

What sampling procedures satisfy Assumption 1? In our paper, we give several examples:

- Distributional perturbation model
- Drawing with replacement from an unknown subpopulation
- Sampling clusters of units with unobserved membership

"Alright, but I could've easily written down another perturbation model and would have gotten a different asymptotic behaviour!"

Result 1: Under a symmetry assumption, all distributional perturbations models are equivalent (in terms of second moments) to the one introduced above.

"Are there relationships to other statistical concepts?"

Result 2: The perturbation model induces correlated data, random confounding, and random sampling bias.

Details can be found in the manuscript.

Theorem (Characterization of isotropic distributional perturbations)

Let $(D, \xi) \sim \mathbb{P}^0$ *and assume that there exists a function* $h(\bullet)$ *such that* $h(D)$ *is uniformly distributed on* $[0, 1]$*. Assume that for any* D*-measurable events* A *and* B *with* $\mathbb{P}^0(A) = \mathbb{P}^0(B)$*,*

$$\mathrm{Var}(\mathbb{P}^\xi(A)) = \mathrm{Var}(\mathbb{P}^\xi(B)).$$

Furthermore, assume that for every sequence of D*-measurable events* A_j *with* $\mathbb{P}(A_j) \to 0$*,*

$$\mathrm{Var}(\mathbb{P}^\xi(A_j)) \to 0.$$

Then there exists $\delta_{dist} \geq 0$ *such that for all* $\phi \in L^2(\mathbb{P})$*, and* $D_i \overset{i.i.d.}{\sim} \mathbb{P}^\xi$

$$Var(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(D_i) - \mathbb{E}[\phi(D)]) = \delta^2 Var(\phi(D)),$$

for $\delta = 1 + n\delta_{dist}^2$*.*

Questions?

Inference

How NOT to do inference

If we use our standard variance formulas (or the bootstrap), we only estimate sampling uncertainty (not distributional uncertainty) and thus drastically underestimate uncertainty!

Assumptions

The statistician might have access to several estimators $\hat{\theta}^k$ that supposedly estimate a very similar quantity.

Assumption

The estimators $\hat{\theta}^k$, $k = 1, \ldots, K$ satisfy

$$\hat{\theta}^k - \theta(\mathbb{P}^0) = \frac{1}{n} \sum_{i=1}^{n} \phi^k(D_i) + o_p(\frac{1}{\sqrt{n}})$$

for some bounded ϕ^k with mean zero and bounded total variation.

In words: we assume that the estimators are asymptotically linear and that the estimators converge to the same quantity.

(If the latter assumption is violated, we will generally get overcoverage, more about that later...)

Example

On observational data, researchers often estimate a causal effect by running a regression of the outcome Y on the treatment T and confounders X. There may be many reasonable choices for the adjustment set.

$$\begin{aligned}
\hat{\theta}^1 &= \text{coef}(\text{lm}(Y \sim T + X_1))[2] \\
\hat{\theta}^2 &= \text{coef}(\text{lm}(Y \sim T + X_1 + X_2))[2] \\
\hat{\theta}^3 &= \text{coef}(\text{lm}(Y \sim T + X_1 + X_2 + X_3))[2] \\
\hat{\theta}^4 &= \dots
\end{aligned}$$

Other examples: Might want to estimate a causal effect via the instrumental variables approach, augmented inverse probability weighting, ...

How to do inference

Given multiple estimators $\hat{\theta}^1, \ldots, \hat{\theta}^K$, we recommend estimating δ^2 via

$$\hat{\delta}^2 = \frac{\sum_{k=1}^K n(\hat{\theta}^k - \frac{1}{K}\sum_j \hat{\theta}^j)^2}{\sum_{k=1}^K \frac{1}{n}\sum_{i=1}^n (\hat{\phi}^k(D_i) - \frac{1}{K}\sum_j \hat{\phi}^j(D_i))^2}$$
$$= \frac{\text{between-estimator-variation}}{\text{expected variation assuming i.i.d. sampling}}$$

The denominator is important! It's not the absolute between-estimator variation that counts, but the relative stability.

Let $\hat{\theta}$ be an estimator chosen by the data scientist.

Theorem (Calibrated inference)

Suppose Assumptions 1 and 2 hold. If $\hat{\phi}^k$ converge to ϕ^k, the estimators are uncorrelated and $K \to \infty$, under some regularity conditions

$$\mathrm{P}\left(\theta(\mathbb{P}^0) \in \left[\hat{\theta} \pm z_{1-\alpha/2} \cdot \hat{\delta}\sqrt{\frac{\widehat{Var(\phi)}}{n}}\right]\right) \to 1-\alpha.$$

Important: this confidence interval covers $\theta(\mathbb{P}^0)$ even in cases where the data might be drawn i.i.d. from $\mathbb{P}^\xi \neq \mathbb{P}^0$.

The scaling factor $\hat{\delta}$ takes care of the additional variation due to distributional perturbations.

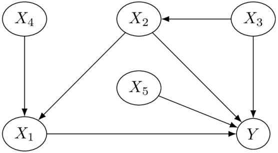
Questions?

Numerical examples

- Is the coverage of the proposed procedure approximately correct?
- Stability of rankings based on the proposed procedure

Evaluation of coverage

Define the distribution $\mathbb{P}^0$ via the following structural causal model.

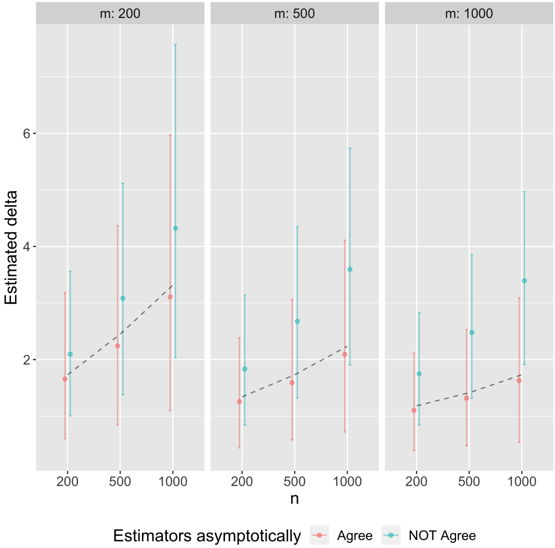


$$\begin{aligned}
&\epsilon, \epsilon_1, \epsilon_2, X_3, X_4, X_5 \overset{\text{i.i.d}}{\sim} N(0,1),\\
&X_2 \leftarrow X_3 + \epsilon_2,\\
&X_1 \leftarrow 0.5X_2 + X_4 + \epsilon_1,\\
&Y \ \ \leftarrow X_1 + 0.5X_2 + X_3 + X_5 + \epsilon
\end{aligned}$$

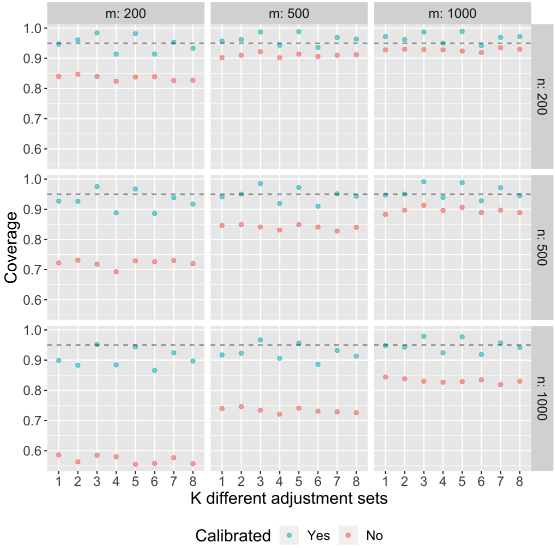
The data is drawn i.i.d. from $\mathbb{P}^\xi$, where $\mathbb{P}^\xi$ arises from perturbing $\mathbb{P}^0$ as in the random perturbation model. The strength of the perturbation is $\delta^2 = 1 + \frac{n}{m}$, where $m \in \{200, 500, 1000\}$ and $n \in \{200, 500, 1000\}$.

Goal: estimate the causal effect of X_1 on Y.

Can use different adjustment sets: $\{X_1, X_2\}$, $\{X_1, X_2, X_3\}$, ... leading to different estimators $\hat{\theta}^1, \ldots, \hat{\theta}^K$.



If we only use correct adjustment sets (red bars) then estimation of δ is almost unbiased. If we also use some incorrect adjustment sets (blue bars), then we overestimate δ.



Coverage of $\theta(\mathbb{P}^0)$ based on i.i.d. data from the perturbed distribution $\mathbb{P}^{\xi}$.

Stability of rankings

Ultimately, the goal of the proposed procedure is to increase stability and trustworthiness of decision-making.

We will see that the proposed procedure can increase stability even in situations without distribution shift.

Stability of rankings

We consider the data set (Cortez and Silva, 2008) about the relationship of final grades with 20 student-specific covariates. $n = 649$

The covariates include student grades, demographic, social and school-related features.

We consider 12 random covariate sets that include 7 binary covariates of interest.

Stability of rankings

- Method 1: The statistician randomly chooses one of the covariate sets, performs a linear regression, and ranks the effect sizes of 7 covariates.
- Method 2: The statistician employs the proposed method. They perform linear regressions with multiple covariate sets and for each covariate, average the estimators and compute its effect size in consideration of distributional perturbations.

Evaluating stability of rankings

We randomly split the data set into two, perform method 1 and method 2 on each split, and compare the rankings resulting from each split.

Stability measure: $|S_{1,k} \cap S_{2,k}|/K$, where
$S_{1,k} = \{\text{Top } k \text{ covariates by the effect size on split 1}\}$ and
$S_{2,k} = \{\text{Top } k \text{ covariates by the effect size on split 2}\}$

We repeat this procedure $N = 1000$ times and record the average set similarity measure.

ℓ	1	2	3	4	5	6	7
Method 1 ($K = 10$)	0.102	0.203	0.407	0.648	0.817	0.898	1.000
Method 2 ($K = 10$)	0.210	0.296	0.449	0.658	0.828	0.912	1.000

ℓ	1	2	3	4	5	6	7
Method 1 ($K = 20$)	0.090	0.203	0.417	0.659	0.817	0.893	1.000
Method 2 ($K = 20$)	0.235	0.313	0.445	0.679	0.845	0.912	1.000

Table: The stability of the ranking: The table above shows results with $K = 10$ adjustment sets and the table below shows results with $K = 20$ adjustment sets. Mean over $N = 500$ iterations of the computed set similarity measure between $S_{1,\ell}$ and $S_{2,\ell}$ for each $\ell = 1, \ldots, 7$ is provided for each method.

On this data set, the proposed method improves stability by more than 100%.

Pros & Cons

- Provides theoretical guarantees for a type of stability analysis that some researchers strongly advocate
- Yields p-values and confidence intervals (only little re-training is needed)
- Does not rely on practitioner specifying the strength of a confounder (as in sensitivity analysis)
- Common methods for multiplicity correction (Bonferroni, FDR,...) directly extend to distributional uncertainty
- Can be extended to more complex perturbation & dependency structures

Issues:

- Can be conservative
- Can be unstable (if all estimators have the same influence function)

Summary

Does current statistical practice capture all relevant sources of variation?

We propose to construct confidence intervals that account for both sampling uncertainty and distributional uncertainty.

Calibrated inference is not just a function of signal-to-noise ratio and thus leads to *different rankings* than *p*-values that capture sampling uncertainty.

`https://arxiv.org/abs/2202.11886`