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

Dominik Rothenhäusler
rdominik@stanford.edu
joint work with Yujin Jeong

Thank you, Peter!

How can we draw trustworthy scientific conclusions?

Some answers:

- Careful design, relevant data
- Reasonable assumptions
- Appropriate quantification of statistical uncertainty
- Replication by independent teams
- ...

Are we missing something?

Repeating experiments is not enough

Marcus R. Munafo & George D. Smith (Nature, 2018)

"If a study is skewed and replications recapitulate that approach, findings will be consistently incorrect or biased.

An essential protection against flawed ideas is (...) the strategic use of multiple approaches to address one question.

Results that agree across different methodologies are less likely to be artefacts."

STATISTICAL MODELS AND SHOE LEATHER

*David A. Freedman**

Regression models have been used in the social sciences at least since 1899, when Yule published a paper on the causes of pauperism. Regression models are now used to make causal arguments in a wide variety of applications, and it is perhaps time to evaluate the results. No definitive answers can be given, but this paper takes a rather negative view. Snow's work on cholera is presented as a success story for scientific reasoning based on nonexperimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, this paper suggests that statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings.

"The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use."

It's expensive to run multiple studies that correspond to different lines of evidence.

Can we make "strategic use of multiple approaches" on one data set?

Leamer (1983)

"Sometimes I include observations from the decade of the fifties sometimes I exclude them, sometimes the equation is linear and sometimes nonlinear.

The professional audience (...) withholds belief until an inference is shown to be adequately insensitive to the choice of assumptions."

Yu and Kumbier (2020)

"For example, the biologist studying gene regulation must choose both how to normalize raw data and what algorithm(s) will be used in analysis.

When there is no principled approach to make these decisions, the knowledge data scientists can extract from analyses is limited to conclusions that are stable across appropriate choices."

Suggestions: change pre-processing, specification of regressions, changing the model for the errors

These researchers recommend evaluating several modelling choices for one single data set.

- How should we choose multiple approaches?
- How should we aggregate different approaches?
- How should we report statistical uncertainty?
- Main question: if several similar regression return similar coefficients, what's the criterion that tells us whether we should be concerned or not?

Main challenge: we do not know the structure and strength of the biases. Can't just use random effect models!

We'll come back to this. Let's talk about our model.

Distributional uncertainty

Batch effects, contaminations, confounding, sampling bias,. . . 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 effects modelling, robust methods, sensitivity analysis or other statistical techniques.

Here, we want to deal with unknown non-adversarial 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

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

How to deal with distributional perturbations

Idea 1: Use worst-case bounds to control the distributional error (similar to sensitivity analysis or robust statistics).

Idea 2: Model distributional perturbations as random and strive for marginally valid confidence intervals.

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 and robust statistics, we will NOT rely on user knowledge 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)
- In the classical robustness literature, one considers estimation in the presence of outliers (Huber 1964, Hampel 1968,. . .)
- In the modern robustness literature, one considers prediction under worst-case distributional perturbations (Duchi and Namkoong, 2018; Sinha et al., 2018; . . .)

How to model distributional perturbations?

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

Three options for the asymptotic regime

- Sampling uncertainty is of higher order than distributional uncertainty
- Sampling uncertainty is of the same order as distributional uncertainty
- Sampling uncertainty is of lower order than distributional uncertainty

How to model distributional perturbations?

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

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

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, the assumption can be seen as relaxing the i.i.d. assumption.

Assumption (Rigorous version)

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

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

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

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

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 a subpopulation
- Sampling clusters of units with unobserved membership

"Alright, but I could've easily written down another perturbation model with a different asymptotic behaviour!"

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

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 $\psi \in L^2(\mathbb{P})$, and $D_i \overset{i.i.d.}{\sim} \mathbb{P}^{\xi}$

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

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

"Are there relationships to other statistical concepts?"

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

(details: see manuscript)

Violation of the i.i.d. assumption

Draw real-valued random variables $D_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}^\xi$, where $\mathbb{P}^\xi$ is generated as above. Let σ^2 denote the variance of D under $\mathbb{P}^0$. Then, marginally,

$$\mathsf{Cor}(\psi(D_i), \psi(D_j)) = \delta^2_{\mathsf{dist}}$$

for some constant $\delta_{\mathsf{dist}} \geq 0$. Thus, under the random perturbation model the observations are marginally correlated.

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!

How NOT to do inference

Example: estimation of the mean. Let $\hat{\theta} = \frac{1}{n}\sum_i D_i$. Under Assumption 1,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \overline{D}))^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2 - (\overline{D})^2 \approx \text{Var}(D).$$

However,

$$\text{Var}(\hat{\theta}) \approx \delta^2 \frac{\text{Var}(D)}{n}$$

Thus, if we just use our standard variance formulas (or the bootstrap), we might drastically underestimate uncertainty!

Without additional assumptions, it is impossible to estimate δ consistently (δ is not identifiable).

Assumptions

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

Assumption (Asymptotic linearity)

The estimators $\hat{\theta}^k$, $k = 1, \ldots, K$ are asymptotically linear, i.e.

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

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

We show that this usually holds for M-estimators in low-dimensional settings (in particular, maximum likelihood estimators).

Assumption (Consistency)

We assume that $\theta^k = \theta(\mathbb{P}^0)$ *for all* $k = 1, \ldots, K$.

If this 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.

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

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

How to do inference

If we consider the difference $\hat{\theta}^1 - \hat{\theta}^2$, by Assumption 1 and 2

$$n(\hat{\theta}^1 - \hat{\theta}^2)^2 \overset{d}{\approx} \delta^2 \mathrm{Var}(\phi^1(D) - \phi^2(D))Z^2,$$

where $D \sim \mathbb{P}^0$ and where Z is a standard Gaussian random variable.

How to do inference

If we consider the difference $\hat{\theta}^1 - \hat{\theta}^2$, by Assumption 1 and 2

$$n(\hat{\theta}^1 - \hat{\theta}^2)^2 \overset{d}{\approx} \delta^2 \mathrm{Var}(\phi^1(D) - \phi^2(D))Z^2,$$

where $D \sim \mathbb{P}^0$ and where Z is a standard Gaussian random variable.

Under regularity assumptions, we can estimate the variance term and obtain

$$\frac{n(\hat{\theta}^1 - \hat{\theta}^2)^2}{\frac{1}{n}\sum_{i=1}^n (\hat{\phi}^1(D_i) - \hat{\phi}^2(D_i))^2} \overset{d}{\approx} \delta^2 Z^2,$$

This term has high variance; this variance can be reduced by averaging over multiple estimators.

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.

Often, researchers assure themselves that different estimators give similar conclusions, by comparing very similar estimators.

"As if someone were to buy several copies of the morning newspaper to assure himself that what it said was true." (Wittgenstein)

Takeaway: absolute stability is not the right criterion; relative stability is

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

Theorem (Yujin intervals)

Suppose Assumptions 1, 2 and 3 hold. If $\hat{\phi}^k$ converge to ϕ^k, the estimators are only weakly correlated and $K \to \infty$, under some regularity conditions

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

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.

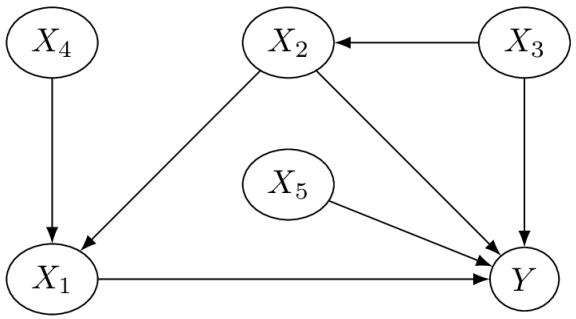
Questions?

Numerical examples

- Is the coverage of Yujin intervals 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.

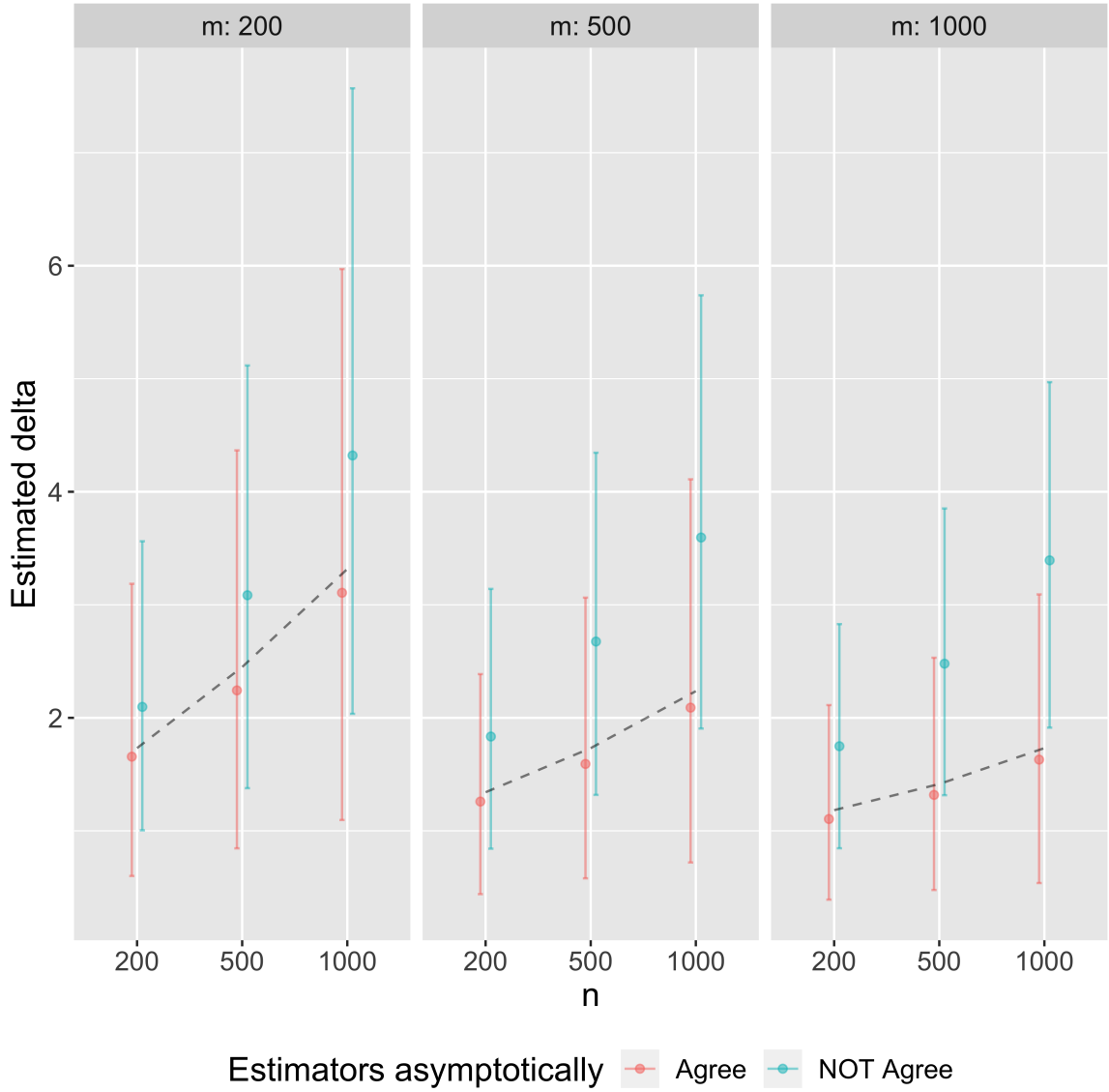


$$\epsilon, \epsilon_1, \epsilon_2, X_3, X_4, X_5 \overset{\text{i.i.d}}{\sim} N(0,1),$$
$$X_2 \leftarrow X_3 + \epsilon_2,$$
$$X_1 \leftarrow 0.5X_2 + X_4 + \epsilon_1,$$
$$Y \leftarrow X_1 + 0.5X_2 + X_3 + X_5 + \epsilon$$

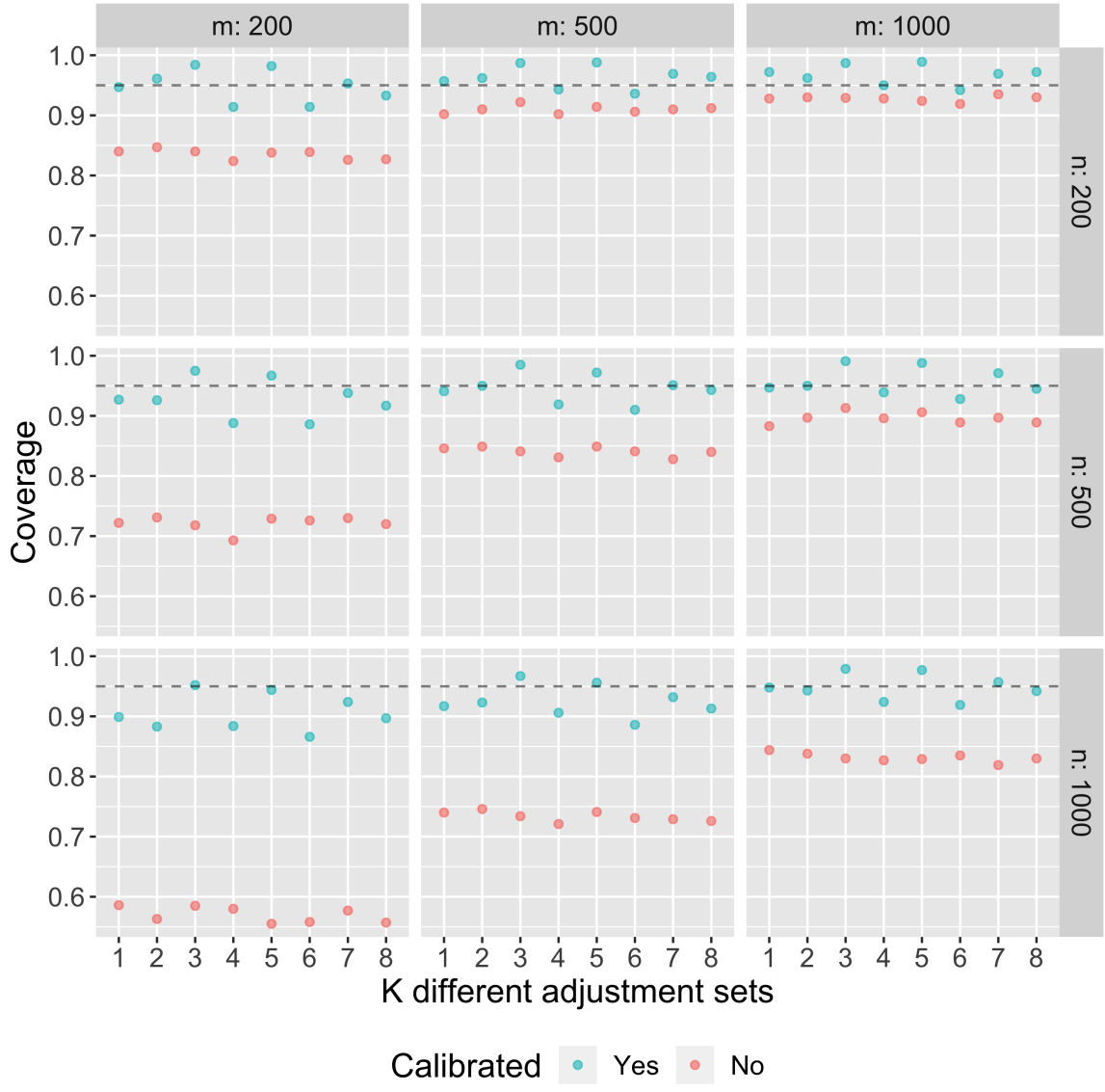
The data is drawn i.i.d. from $\mathbb{P}^\xi$, where $\mathbb{P}^\xi$ arises from perturbing $\mathbb{P}^0$ as in the random perturbation model. The strength of the perturbation is $\delta^2 = 1 + \frac{n}{m}$, where $m \in \{200, 500, 1000\}$ and $n \in \{200, 500, 1000\}$.

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%.

Some frequently asked questions. . .

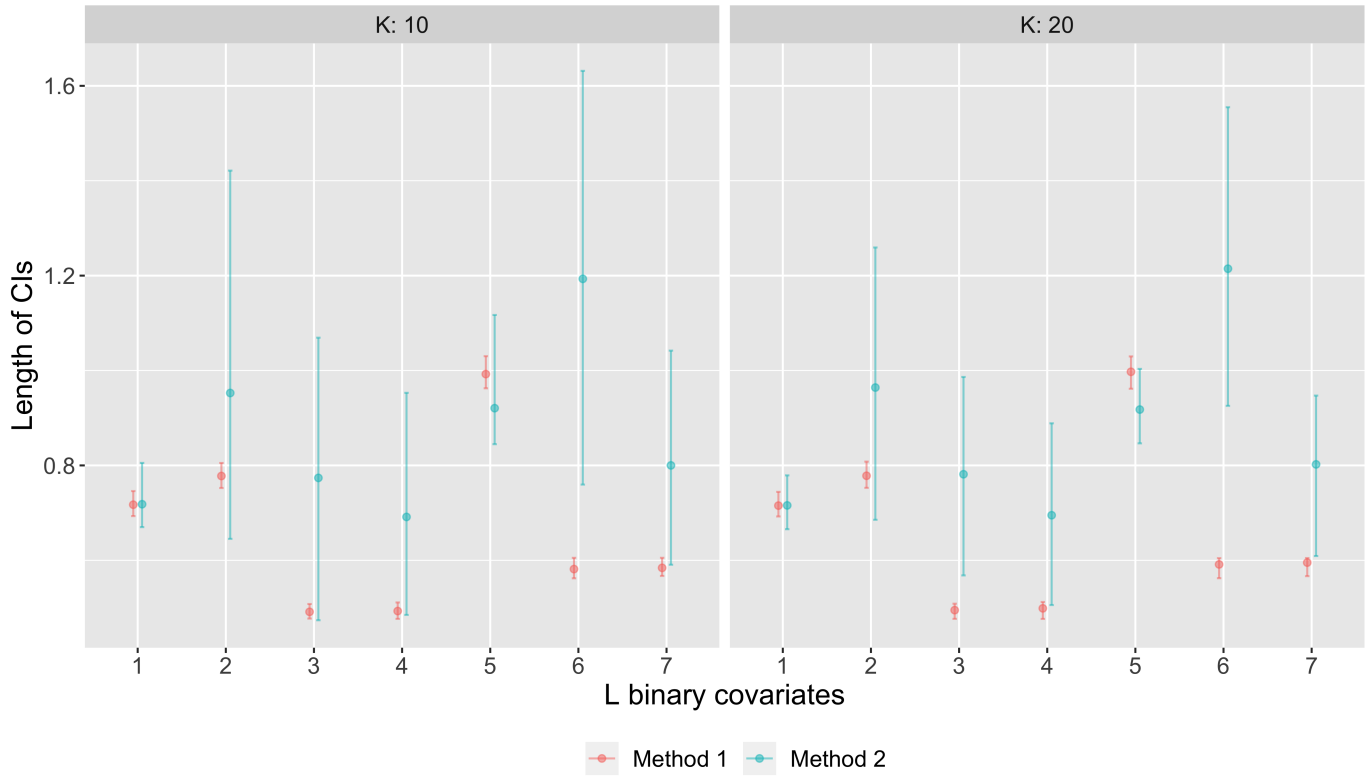
Q: The isotropic perturbation model makes very strong assumptions!

We agree! However, please note that it is weaker than assuming that the data is drawn i.i.d. from the target distribution. In this sense, it can be seen as an extension of the most common inferential strategy in statistics.

The isotropic perturbation model can be generalized (see later).

Q: How large are calibrated confidence intervals compared to ordinary ones? Are they extremely conservative?

Answer 1: In our application, it was reasonable.



Answer 2: Sometimes, calibrated confidence intervals will be quite large.

Under the assumptions outlined above, this is an indication that distributional uncertainty is of higher order than sampling uncertainty.

If distributional uncertainty is high, standard confidence intervals might be useless?

In some sense, this is a feature, not a bug.

Q: Does this really work?

Some practitioners have advocated to use different estimation strategies on a single data set for decades (Leamer 1983, Rosenbaum 2010, Yu and Kumbier 2020).

Our theory provides rigorous guarantees for a version of this practice & some guidance how to do it.

Looking ahead.

Extensions: Generalizing the isotropic perturbation model

For example, one can perturb the distribution of X and $Y|X$ differently, leading to

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi(D_i^n) - \mathbb{E}_{\mathbb{P}^0}[\psi(D)])\right)\\ &= \delta_1^2\mathrm{Var}(\mathbb{E}_{\mathbb{P}^0}[\psi(D)|X]) + \delta_2^2\mathrm{Var}(\psi(D) - \mathbb{E}_{\mathbb{P}^0}[\psi(D)|X])\end{aligned}$$

We can estimate δ_1 and δ_2 similarly as above. In words, we would calibrate a multidimensional uncertainty model.

Extensions: Other ways to calibrate inference (that means, estimate δ)

So far, we have discussed how to use several estimators $\hat{\theta}^1, \ldots, \hat{\theta}^K$ to estimate δ.

There are lots of other ways to estimate δ. In principle, can use any types of moment equations either within data sets or across data sets.

Examples:

- Negative controls in causal inference (negative outcomes or negative exposures)
- Knowledge of population quantities (for example, if $\mathbb{E}_{\mathbb{P}^0}[X]$ is known)
- Multiple data sets drawn from $\mathbb{P}^{\xi_1}, \ldots, \mathbb{P}^{\xi_E}$.

Getting started

We're writing an R package `calinf` available under `github.com/rothenhaeusler/calinf`.

Sampling from the distributional perturbation model

Sampling uncertainty

```
x <- rnorm(1000)
y <- 2*x + rnorm(1000)
```

Sampling from the distributional perturbation model

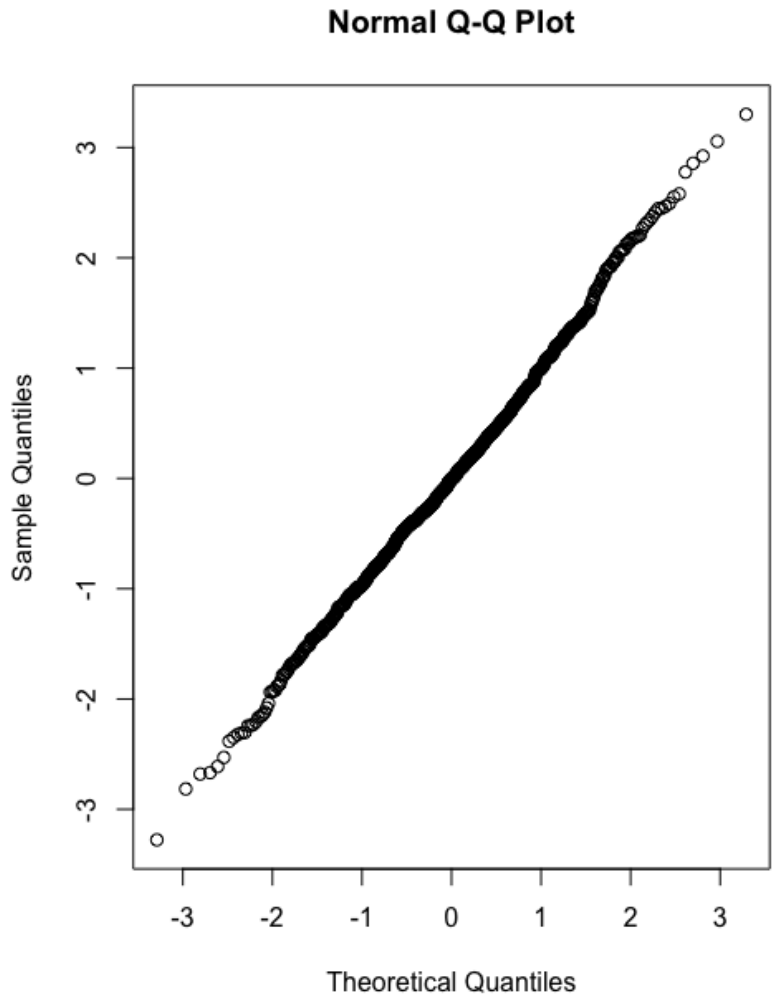
Sampling uncertainty

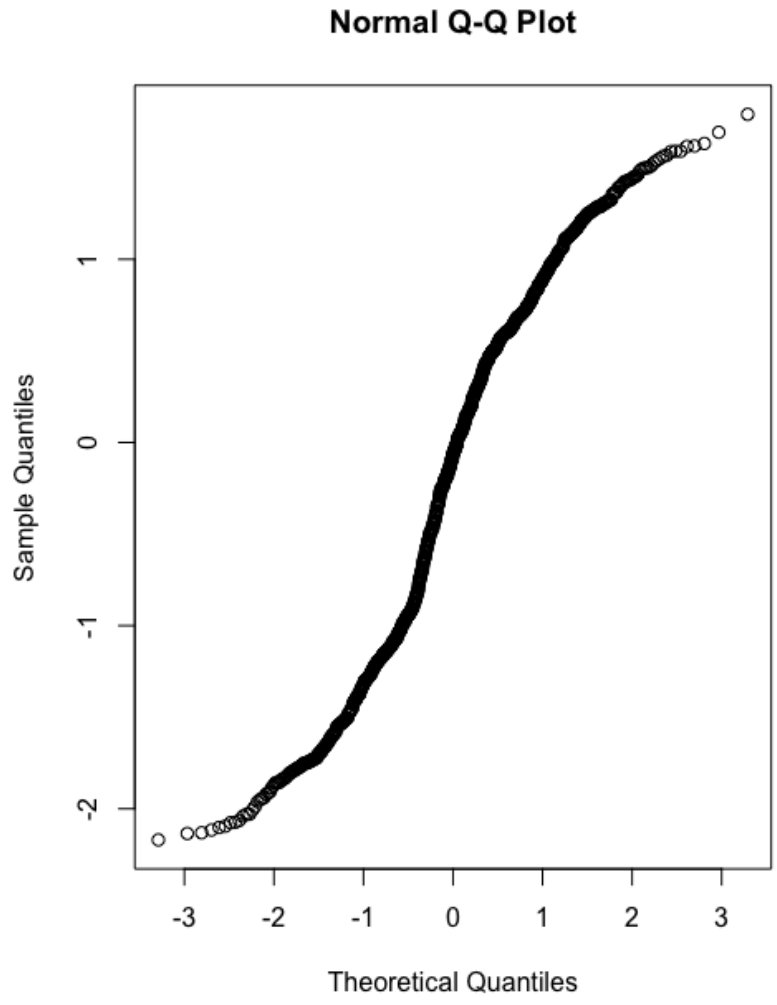
```
x <- rnorm(1000)
y <- 2*x + rnorm(1000)
```

Sampling uncertainty and distributional uncertainty

```
d_seed <- distributional_seed(n=1000,delta=5)
x <- drnorm(d_seed)
y <- 2*x + drnorm(d_seed)
```

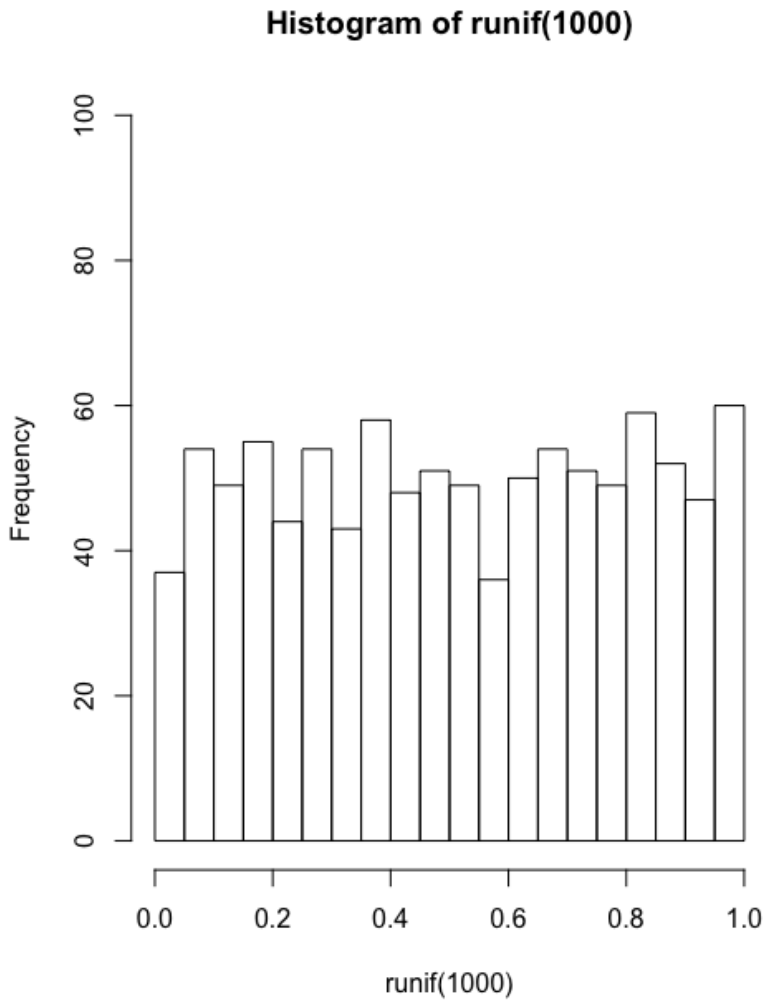
drnorm()

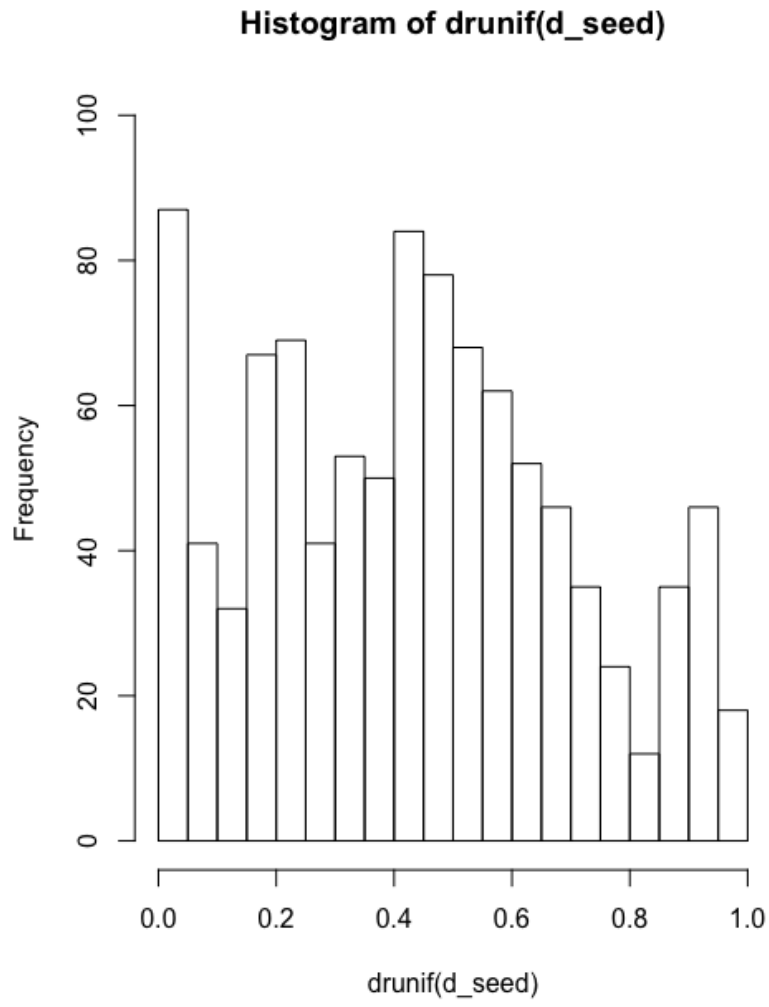




Left: $\delta = 1$; Right: $\delta = 5$. In both cases $n = 1000$.

drunif()





Left: $\delta = 1$; Right: $\delta = 5$. In both cases, $n = 1000$.

Calibrated inference

```
> formulas <- list(
                  Y ~ X,
                  Y ~ X + I(X^2),
                  Y ~ X + Z1,
                  Y ~ X + Z2 + I(X^2),
                  Y ~ X + Z1 + Z2 + I(X^2)
                  )
> calm(formulas,data=data, target="X")

Quantification of both distributional and sampling uncertainty

  Estimate Std. Error Pr(>|z|)
X   1.0244     0.0159        0
```

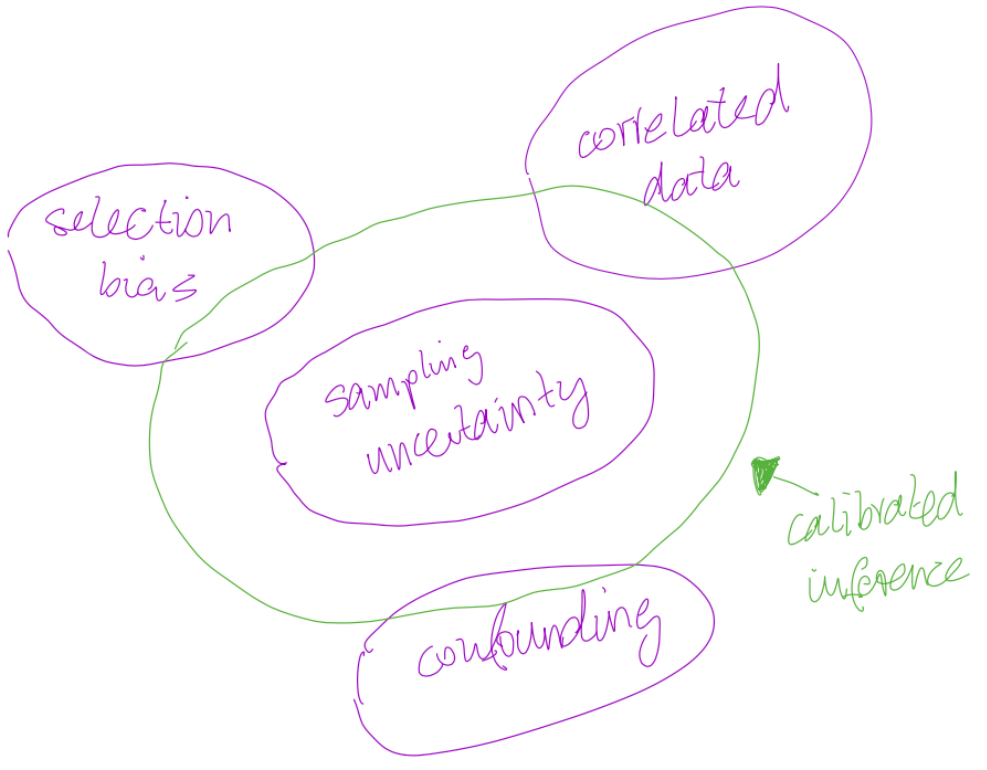
Tools

We have functions

1. to generate distributional seeds and draw from perturbed distributions (drnorm, drunif, drbinom, drchisq,...)
2. to conduct inference in generalized linear models (calm, caglm)

Some more philosophical remarks

High-level intuition



Calibrated inference can quantify uncertainty in settings where due to isotropic distributional perturbations, there is confounding, correlated data, or selection bias.

In Statistics, as $n \to \infty$, we often report that uncertainty goes to zero.

```
Call:
lm(formula = ROLL ~ UNEM + HGRAD + INC, data = datavar)

Residuals:
      Min        1Q    Median        3Q       Max
-1148.840  -489.712    -1.876   387.400  1425.753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.153e+03  1.053e+03  -8.691 5.02e-09 ***
UNEM         4.501e+02  1.182e+02   3.809 0.000807 ***
HGRAD        4.065e-01  7.602e-02   5.347 1.52e-05 ***
INC          4.275e+00  4.947e-01   8.642 5.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 670.4 on 25 degrees of freedom
Multiple R-squared: 0.9621, Adjusted R-squared: 0.9576
F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

Your inference is only as good as your assumptions

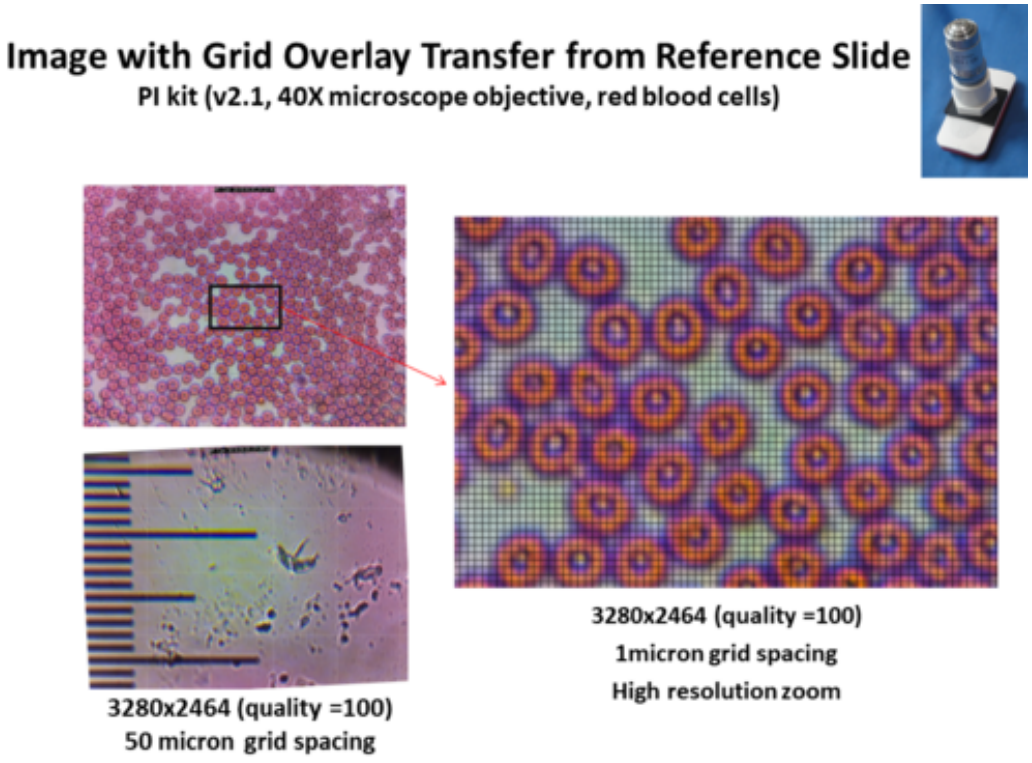
For $n \to \infty$, assuming that the estimators are uncorrelated, the estimated variance is

$$\frac{\hat{\tau}^2 \hat{\sigma}^2}{n} \sim \sum_{k=1}^{K} w_k (\theta^k - \frac{1}{K} \sum_{j=1}^{K} w_j \theta^j)^2,$$

for some weights $w_i \geq 0$.

Consequences:

- Assumptions almost correct $\rightarrow$ precise inference
- Assumptions grossly violated $\rightarrow$ high uncertainty



Even for $n \rightarrow \infty$, uncertainty will not go to zero. Uncertainty is lower bounded by the quality of your assumptions.

What is the third number?

In Statistics, there is consensus that we report 1) the estimate and 2) the standard error. Should we report a third number?

Three versions of the future:

1. (pessimistic) No.
2. (optimistic) We'll report other numbers which measure issues stability under distribution shift, sensitivity to outliers, method stability, ...
3. (in between) Integrate additional sources of uncertainty in the variance estimate.

Summary

- Provides theoretical guarantees for a type of stability analysis that some researchers strongly advocate
- Yields p-values and confidence intervals (easy to interpret, integrates with FWER, FDR control)
- Can be extended to more complex perturbations

Limitations:

- Can lead to large confidence intervals
- Can be unstable (if all estimators have the same influence function)

Thank you for your attention!

Draft is on arXiv.

Package is available under `github.com/rothenhaeusler/calinf`.