Thesis statement: TBD

Cool points we want make:

There is a kind of “sampling” uncertainty (not really sampling, more like sensitivity) that is distinct from SEs that does not disappear asymptotically.

The sensitivity is high when there is high signal to noise ratio i.e. large N does not solve low SNR ratio according to this metric

for linear regression, this uncertainty tends to be large when the scale of the residuals is large and when there are observations with high leverage (you need both)

results / analyses which are simple and/or seem solid according to their standard errors may be revealed to be unstable or lack robustness according to the other metric

An alternative idea: just order the influence functions. But we provide and recommend a summary: how many data points you should remove before you change your decision/claim substantially

When and when not to use:

- When not to use:
    - If you actually care about (a) the quantity you’re measuring and (b) the variability is due only to random sampling.
    - If you intend to repeat the precise intervention you applied (to repeat the exact experiment being studied), and only care about the summary statistic you’re computing.
    - Example: polling, vaccine trial, agricultural trials
- When to use:
    - When you are computing a summary statistic/proxy, intended to describe something about the population as a whole. What we really want is e.g. the full distribution (of e.g. causal effects). Hard to deal with full distribution and also not identifiable -- so convenient, etc.
    - If you’re trying to learn a general truth about interventions qualitatively similar to your intervention
    - Example: causal inference for profit

Example of times when standard error is not enough:

- Distributions change across experiments
    - Example: microcredit
- Proxy not same as thing I care about: and in that case nonrobustness means there is an obvious wedge between the proxy and the policy-relevant parameter or claim because the latter cannot depend on 5 people, it is supposed to be more general than that / more

robust to heterogeneity than that (maybe there is a simple example of how under homogeneity of individual effects our thing is the same as an SE but they diverge under heterogeneous treatment effects…)

When we do a data analysis, we want to be sure that our conclusions would not change had our data been different. One way for the data to be different is for it to be randomly sampled from the same distribution that generated our data in the first place. In certain rarefied situations this is enough. In other circumstances, particularly when our experiment is a proxy for a more general or hard to precisely articulate set of interventions or contexts, we may consider other types of data perturbation to be "small", such as adversarially removing some small percentage of data points. We examine one such perturbation and provide a tool for measuring sensitivity to it.

Note: "our method" refers to the linear approximation

Results we have now:
- The linear approximation does not disappear asymptotically
- The error between the linear approximation and the true sensitivity to the perturbation does not disappear asymptotically
- standard error and the linear approximation to the sensitivity are not necessarily ordered in the same way
- We are using asymptotic analysis to approximate a combinatorially difficult problem. We do not care about the actual limiting quantity (which is as alpha goes to zero). We care about finite alpha and finite N. We will analyze finite but small alpha and large N to provide intuition.
- We provide something extra relative to AGS 2017:
    - if as they claim a parameter's sensitivity to a sample moment is a signal of the moment's role in identification / driving the conclusion, then it is important to know the moment's sensitivity to small perturbations in the data, which is what we provide.
    - theirs depends on taking the model seriously as an exact representation of the objects of policy interest
    - ours applies more generally. (theirs is for GMM.) Ours is very intuitive
- one can in theory compute the exact sensitivity we're interested in. the only reason we don't is it's prohibitively expensive, so we propose an approximation
- In our OHIE example, the IV was not substantially more sensitive than the OLS, which is not what we would expect per Young's work.
- In M estimators in general the sensitivity is determined by two things: one is the SNR and the other is the shape parameter that cannot vary that much. Note, it's NOT heavy tails.
- In the standard gaussian data set up with a binary treatment and a linear regression to estimate TE, then unless the SNR is very low, you cannot substantially change results/claims by removing small % of the data.
    - corollary: we're not in that world

- we apply our method to linear regression and illuminate a number of points. We are therefore able to formalize the ideas in Chatterjee and Hadi 86, and provide intuition
- Non-significance is non-robust. Significance doesn't mean much by itself; you must look at effect sizes.

Grand Plan

Introduction: lay out general plan of action. Statistics for physics is different than statistics for social science. A lot of statistics is designed for physics (or for rarified situations in which we know the model of the world and we can draw many samples from the same population which is truly the only population of interest). We say: Let's start from first principles to approach statistics for social science. ← This is an ambitious plan! We're just going to start from one concrete proposal; we suspect and hope it won't be the end of this discussion.

- "Evaluative statistics" (for e.g. physics) vs "narrative statistics" (for social sciences)

The Method

<u>Our proposal for a base method</u>:
"Largest influence fraction" (LIF)
Concrete thing we propose: what happens when you leave out some of your data points? Would you be worried if just a few data points totally changed your significance? Or became significant in the other direction? Specifically, we look at (each contingent on any particular claim):

- In what follows, a "data point" could be a single data point, a pair of data points, data points constrained to both be in a particular group, &c (whenever we introduce this, we should be clear applies to all cases/claims)
- 1. How many data points to change the significance status of the result
- 2. How many data points to change the sign of the parameter?
- 3. How many data points to generate a significant result of the opposite sign?
- (But you could think about another claim! Not restricted to these three! Some brief examples here and then point to experiment(s) where we consider another option)
    - E.g. Bayesians can look at how posterior probabilities change
    - E.g. change amount of money in microcredit REG experiment to something else substantively different
    - E.g. change the order of magnitude of the hypervariance in the microcredit VB

<u>Our proposal for approximation of the base method</u>:
"Approximate largest influence fraction" (ALIF)
OK, now we have to actually compute this in practice. Turns out that's too costly! We propose an approximation: linear approximation / influence functions. (More info later!)

- Derive / Write out the approximation (give just the minimal info needed to understand why this is even a non-silly approximation)

- Autodiff is key; see software package below
- By using the chain rule can approximation sensitivity to functions of the point estimate too

Illustration of the total method (base method + approximation):
Microcredit (regression version) example to blow minds.

- Motivate the use of our new type of statistics / our metric
- Linear regression worked out as example; it will recur later as “REG EXAMPLE”
- (Variational Bayes version later -- at end in experiments.)

Implementation:
We have an open-source software package!

- Automatic differentiation makes everything easy
- We can handle regression and IV (and weighted versions and robust/clustered standard errors) automatically. Other M estimators are a bit more involved; see appendix for discussion. See appendix for discussion of relation between clustered and robust standard errors
- Code snippet (to illustrate how easy it is!)

What’s going on?

- Justification and discussion of the approximation. Solving the exact problem is combinatorially difficult, so we form an asymptotic approximation, in the sense from physics.
    - Discussion of what asymptotic approximation means:
        - We are not concerned with asymptotic quantities; we are using asymptotics to approximate finite-sample quantities.
        - Under regularity conditions, we expect that the approximation scales linearly in alpha, the error quadratically in alpha. Thus, for small alpha, the error will be much smaller than the effect we’re trying to quantify.
        - Relation to AGS 2017: we are in a general framework of which AGS is an example.
            - Do weights explicitly here (i.e. in main text)
            - Do general parameter case in appendix
            - Move absolutely everything not necessary to the immediate point to the appendix
    - The influence function ends up being the linear approximation [maybe write out the linear approximation here, define derivative wrt weights].
        - Write out the approximation
        - You already know and love the influence function. We’re going to use the influence function for our thing. Some notes about how you already know and love it even more:
            - The L2 norm is the ordinary standard error.
            - The sup is the Huber notion of robustness.

            - REG EXAMPLE: the linear regression robust SEs are also just another norm on the influence function!
    - We are using a linear approximation to the thing we actually care about. What is its error? In the case of IV and linear regression, we can write closed-form, computable, finite-sample(!) bounds for the error and analyze its asymptotic behavior under clearly articulated regularity assumptions.
        - REG EXAMPLE
            - Examples of our linear approximation working
            - Example with a large change for which the linear approximation is not good (illustration and discussion of where our approximation can fail)
        - Point to experiments
        - Proofs to appendix. The HOIJ paper provides a general framework but computing the constants may be difficult.
    - In the case that it's not too expensive to re-compute the metric for your proposed reweighting, we provide a lower bound on the effect, since we have in the worst case found a sub-optimal ordering of the data.
        - REG EXAMPLE (brute force ten or so datapoints maybe)
- What's really driving what this metric is telling us?
    - Illustrative SIMULATION (introduce after the conjectures)
    - Some conjectures you might have going in about what's driving this include: amount of data, heavy tails, SNR/scale/width
        - Plan for this section: Write out formula. Then use it to hit each point in turn (amount of data, heavy tails, leverage, SNR/scale/width). Refer to REG EXAMPLE and SIMULATION with each point.
            - This sensitivity is unlikely to be resolved or minimized by getting larger N (this is shown by the fact that the sensitivity exists within the asymptotic approximation)
            - What matters is the scale (SNR) and the shape
            - The shape doesn't matter as much
            - The shape in fact matters in a counterintuitive way ("heavy tails" are good") [math + SIMULATION and helps interpret REG EXAMPLE]
            - It's not because means are not robust in a huber sense, means can be robust in our sense if the SNR is high.
    - Leverage (separate from other ideas above because takes more development/setup)
        - Influence is not the same thing as leverage: In the special case of linear regression, the influence function for any data point turns out to be the product of the leverage score and the regression error (formalizing the Chatterjee Hadi)
- Corollary of bullet above: shows that non-significance is non-robust.
    - This would be true of any metric that does not vanish asymptotically

    - This also implies that significant results will only be robust when the point estimate is large relative to the size of the sensitivity metric

Experiments

- Cash transfers
    - Motivate the use of our new type of statistics / our metric
    - Our linear approximation to the exact thing is a good approximation
    - Compare raw removal to paired removal
    - Note the difference in robustness across the direct vs indirect analyses
    - The authors already removed “outlying” Y data points and we did the analysis following their protocol so outlier-removal does not fix this problem
        - This suggests that dropping/winsorizing is not necessarily that great: "since you'd think trimming would help with this problem AND YET IT DOES NOT, maybe trimming is in fact a subtle and potentially confusing practice, and we might accidentally do something bad with it, let's be careful"
- Oregon Medicaid experiment
    - Motivate the use of our new type of statistics / our metric. (Ryan’s opinion: arguably, the Oregon experiment could play both roles: evaluative for expanding Medicaid in Oregon, or narrative for expanding Medicaid everywhere)
    - Our linear approximation to the exact thing is a good approximation
    - IV is just as robust as OLS on this problem (cf. Young’s paper)
    - Compare raw removal to paired removal
- Microcredit: full, VB
    - Motivate the use of our new type of statistics / our metric
    - Our linear approximation to the exact thing is a good approximation
    - Compare raw removal to paired removal
    - Shouldn’t be that surprised because this stuff is not significant anyway
    - BHM doesn’t rescue the TEs, they remain non-robust (not that surprising since everything is small)
    - However the hyper-variances are relatively robust (i.e. you can change the size of the TEs a lot relative to the size of their standard errors but you can’t change the size of them relative to each other)

Discussion/Conclusion

- We discussed that there are different paradigms for statistics other than the current dominant one.
- It turns out that you get a very different story in the two paradigms.
- Once again, the intuition is not the same thing as that which underlies general huber / robust statistics: we have shown that conditional means can perform well if SNR is high.
- We’ve shown a way to identify issues. There needs to be work on solutions (but we don’t do that here). Refer back to the general framework of thinking about statistics for social

science; we encourage these new solutions to be in the narrative framework rather than the evaluative framework.