Statistical Inference with Screening and Selection

Isaiah Andrews

Tuesday, May 23

Screening and Selection

- Frequentist statistical guarantees control performance under repeated sampling
    - That is, if we could draw the data multiple times in a given situation, certain properties would hold on average across data realizations
- The way in which statistical procedures are commonly applied often doesn't match the sampling thought experiment
    - We might only write up/publish certain findings
    - We might choose the target for inference based on the data
- In such cases, standard statistical procedures can yield non-standard behavior

Example: Randomized Trial

- To illustrate, imagine that we conduct a randomized trial of a job training program
- Usual statistical procedure:
  - Compare average outcomes in treatment and control groups
  - Conclude that treatment has an effect if the difference in average outcomes is large relative to the standard error
- Analogously, we can analyze the effect of treatment within observable subgroups of the sample, e.g. based on location or prior employment history
- Usual statistical guarantee: under regularity conditions, if treatment in fact has no effect, we will mistakenly conclude that there is an effect at most α (e.g. 5%) of the time

Screening and Selection

- The usual statistical guarantee fixes a procedure (e.g. run the experiment and take the difference-in-means) and asks how it performs over repeated draws of the data
- Empirical practice often differs from this idealized description
    - Our experimental result may only be written up or published if it is positive and statistically significant $\Rightarrow$ screening/publication bias
    - We might focus our analysis on the subgroup with the largest effect $\Rightarrow$ selection bias
- There is an active literature in statistics and related fields which aims to correct for these issues
    - My goal today: provide a brief review of this literature, some of the tools it suggests, and some of the questions that it raises

Decision Theory Without Screening or Selection

Decision Theory Without Screening or Selection

- Suppose we observe $X \sim F(\mu)$ for an unknown parameter μ
- For today, will specialize to $X \sim N(\mu, \Sigma)$ for $\mu \in \mathbb{R}^J$ and Σ known
  - For X a vector of estimates based on underlying, potentially non-normal observations, justified in many contexts by the central limit theorem
- Suppose we are interested in a parameter $\theta = v'\mu \in \mathbb{R}$
- For a set of possible actions $\mathcal{A}$, and a loss function L, we want to choose a decision rule $\delta(X)$ to achieve a low expected loss

  $$E_\mu \left[L(\delta(X), \theta) \right].$$

  We may also require this rule to satisfy some additional constraints

Example: Randomized Trial

- In the job-training experiment, the vector X could collect treatment-control differences across J demographic subgroups
  - So long as the number of trial participants in each subgroup is large, the central limit theorem justifies the approximation

    $$X \sim N(\mu, \Sigma)$$

    for Σ a diagonal matrix
- We can consider different target parameters θ in this context
  - For ω_j the population share of group j and $v = (\omega_1, ..., \omega_J)'$,

    $$\theta = v'\mu = \sum_j \omega_j \mu_j$$

    captures the average treatment effect in the population
  - For $v = e_j = (0, .., 0, 1, 0, ..., 0)$ the jth standard basis vector,

    $$\theta = v'\mu = \mu_j$$

    captures the average treatment effect in subgroup j

Loss Functions and Constraints

- For estimation we can take $\mathcal{A} = \mathbb{R}$ and consider squared-error loss $L(a, \theta) = (a - \theta)^2$, so

$$E_\mu \left[L\left(\delta\left(X\right), \theta\right)\right] = E_\mu \left[\left(\delta\left(X\right) - \theta\right)^2\right]$$

  corresponds to mean squared error

- We may further impose unbiasedness or median-unbiasedness,

$$E_\mu \left[\delta\left(X\right)\right] = \theta \text{ or } Med_\mu \left(\delta\left(X\right) > \theta\right) = \frac{1}{2} \text{ for all } \mu$$

- To quantify uncertainty we might focus on confidence intervals, taking $\mathcal{A}$ to be the set of closed intervals in $\mathbb{R}$, define $L(a, \theta) = |a|$ as the length of a, and impose a coverage constraint,

$$Pr_\mu \left\{\theta \in \delta\left(X\right)\right\} \geq 1 - \alpha \text{ for all } \mu$$

  potentially along with other constraints

Optimal Decision Rules

These problems have well-known solutions

- The maximum likelihood estimator

  $$\hat{\theta} = v'X$$

  is the best (median-)unbiased estimator for θ, in the sense that for any other (median-)unbiased estimator $\tilde{\theta}$,

  $$E_\mu\left[\left(\hat{\theta} - \theta\right)^2\right] \leq E_\mu\left[\left(\tilde{\theta} - \theta\right)^2\right] \text{ for all } \mu$$

- Similarly, for $\sigma_{\hat{\theta}} = \sqrt{v'\Sigma v}$ the standard error of $\hat{\theta}$, confidence intervals of the form

  $$\left[\hat{\theta} \pm c \cdot \sigma_{\hat{\theta}}\right]$$

  are optimal in various senses
  - $c = 1.96$ for gives the standard 95% confidence interval.

Screening and Selection

- The decision-theoretic setup above made two assumptions
  1. We care about performance on average across all realizations of X
  2. The target parameter $\theta = v'\mu$ is the same for all realizations of X

  which can fail in practice
- I'll refer to failures of (1) as **screening** and failures of (2) as **selection**
- Warning: useful shorthand for today, but these are not consistently adopted terms in literature

Screening

Screening Problems

- Above, we averaged performance over all realizations of X
- Sometimes, however, we may only care about performance over a subset of data realizations
- To formalize this, suppose that for a screening variable S
  1. The conditional distribution of $S|X$ does not depend on μ
  2. We only care about behavior conditional on $S = 1$, e.g.

$$E_{\mu}\left[L\left(\delta\left(X\right),\theta\right)|S=1\right],\ E_{\mu}\left[\delta\left(X\right)|S=1\right],\ Pr_{\mu}\left\{\theta\in\delta\left(X\right)|S=1\right\}$$

Example: Randomized Trial

- In the randomized trial example, suppose θ corresponds to the average treatment effect over the population
- Randomization ensures that $\hat{\theta} = v'X$ is unbiased for θ on average across data realizations
  - But not all estimates $\hat{\theta}$ are equally likely to be published
  - An extensive literature expresses concern about, and provides evidence of, publication bias
  - A few recent examples include Open Science Collaboration (2015), Bruns and Ioannidis (2016), and Camerer et al. (2016)

Example: Randomized Trial

- To study publication bias in this example, let $S = 1$ be an indicator for the event that a given estimate $\hat{\theta}$ gets published,

$$S = 1\left\{\text{Estimate } \hat{\theta} \text{ gets published}\right\}$$

  - We assumed that that the conditional distribution $S|X$ does not depend on θ
  - Means that publication decisions depend only on estimates, and not on the underlying parameters *once we hold the experimental results fixed*
- Publication decisions could depend on many factors
  - Preference for positive results
  - Preference for surprising results
  - Preference for results consistent with previous literature

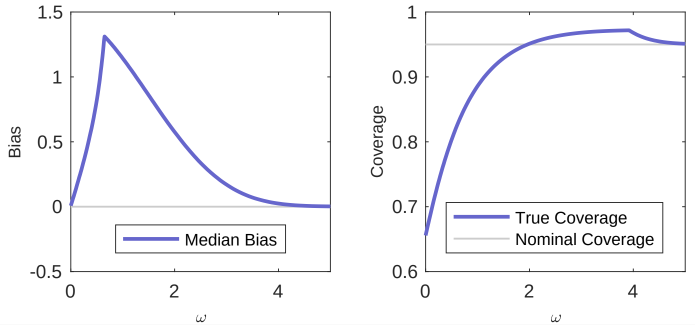
Example: Randomized Trial

- Here, I'll focus on an example from Andrews and Kasy (2019)
    - Preference for statistically significant results: many papers in the literature point to this
    - Specifically, let $Z = \hat{\theta}/\sigma_{\hat{\theta}} \sim N(\omega, 1)$ denote the estimate, standardized to have variance one
    - Suppose results that are significantly different from zero at the 5% level (i.e $|Z| > 1.96$) are 10 times more likely to be published than are statistically insignificant results

$$Pr\left\{S = 1|X\right\} \propto \begin{cases} 1 & \text{if } X \text{ implies } |Z| > 1.96 \\ 0.1 & \text{otherwise} \end{cases}$$

- What is the effect of such screening on the distribution of published results?

Example: Randomized Trial



Screening Problems

- While I’ve discussed screening in terms of publication bias, note that it doesn’t matter who’s doing the screening
  - e.g. authors choosing not to write up results vs. journals choosing not to publish
  - Hence, “screening” as discussed here covers many forms of so-called “p-hacking”
- Moreover, some practices which are recommended on other grounds generate the same issues
  - e.g. dropping models where a specification test suggests the model is incorrect

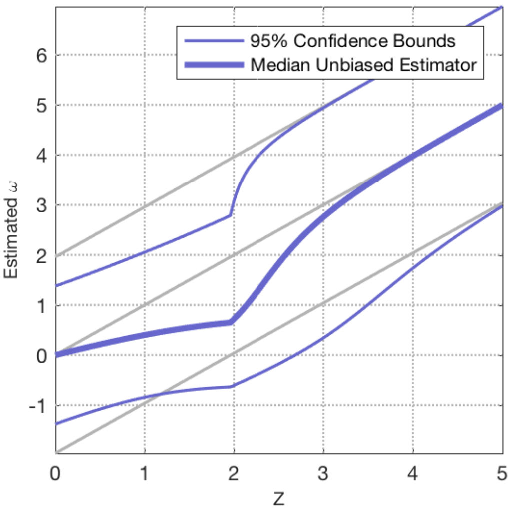
Corrections for Screening

- Fortunately, some results in statistics provide powerful tools to correct for screening
- Let $f_X(x|\mu)$ denote the density of X without screening. Bayes rule implies that the conditional density of X given $S = 1$ is

$$f_{X|S=1}(x|\mu) = \frac{E[S|X = x]\, f_X(x|\mu)}{E_\mu[S]}$$

- This implies that if $f_X(x|\mu)$ has exponential family structure (as is true for the normal distribution) then $f_{X|S=1}(x|\mu)$ does as well
- Results in statistics then deliver optimal median-unbiased estimators, optimal confidence intervals for $\theta = v'\mu$

Example: Randomized Trial



Selection

Selection Problems

- In screening problems, we only cared about some values of X, but always cared about the same target θ
- In selection problems, by contrast, we want to conduct inference on $\theta_X = v(X)' \mu$
    - Hence, we may have a different target for inference for different values of X
- Selection problems of this sort have been extensively studied in the recent statistics literature
    - Motivated by model selection concerns: let M_X be the model selected when realized data are X, and define θ_X as the target parameter under this model
    - e.g. Berk et al. (2013), Lee et al (2016), Fithian et al. (2017)
- Selection problems also arise outside the context of model selection, however

Example: Randomized Trial

- Recall that in this example, X records the treatment-control differences over J different subgroups
- We might be interested in the effect of treatment on the group for whom treatment appears most effective,

$$\theta_X = e_{\hat{j}}'\mu,\ \hat{j} = \arg\max X_j$$

- Alternatively, we might be interested in the average effect of treatment across those subgroups where it appears helpful,

$$\theta_X = v(X)'\mu,\ v_j(X) = \frac{\omega_j 1\{X_j > 0\}}{\sum_j \omega_j 1\{X_j > 0\}}$$

Example: Randomized Trial

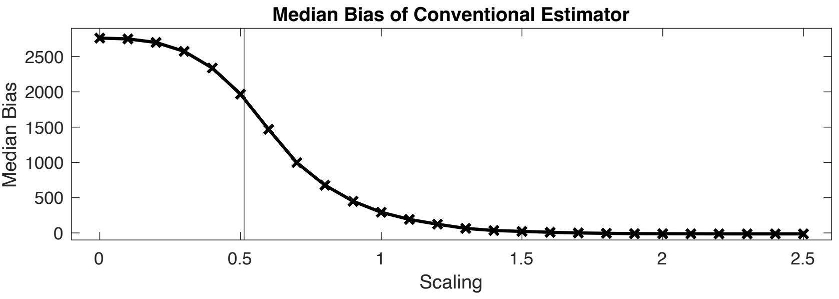
- As with screening, selection generally invalidates conventional inference approaches
- To illustrate, use an example from Andrews et al. (2023)
- Calibrate a simulation based on data from a randomized trial of job-training programs, conducted at 13 different sites
    - JOBSTART Experiment, conducted by US Department of Labor
    - Program at one site appeared most effective
    - This site was subsequently selected as the model for a subsequent replication study conducted at new sites
    - Results in replication turned out to be disappointing

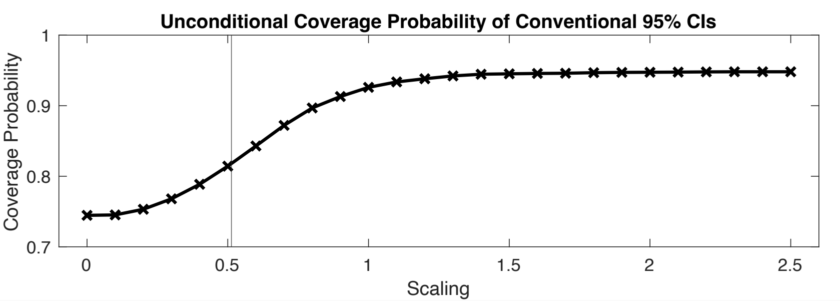
Example: Randomized Trial

Simulations take the experimental results as starting point

- Scale up/down to vary the heterogeneity in effect sizes across sites
- For a given effect size, simulate draws of experimental results, and ask how selecting the site with the largest estimated effect impacts inference
- Vertical line shows scaling to match unbiased estimate for variance of effects across sites.

Example: Randomized Trial





Selection Problems

- We see that, like screening, selection can invalidate conventional inference procedures
  - Estimated effect sizes are biased upwards
  - Conventional confidence intervals under-cover
- The selection problem introduces a new wrinkle relative to our analysis so far: the target parameter θ_X is now random
  - Different target parameters for different data realizations
- This suggests two possible routes forward:
  - We could condition on $v(X)$ to remove this randomness...
  - ... or we could not

Conditional Inference

- By conditioning on $v(X)$, I mean requiring conditional median unbiasedness or conditional coverage

$$\begin{array}{c} Med_{\mu}\left(\delta\left(X\right)|v\left(X\right)=\tilde{v}\right)=\tilde{v}'\mu \\ Pr_{\mu}\left\{\tilde{v}'\mu\in\delta\left(X\right)|v\left(X\right)=\tilde{v}\right\}=1-\alpha \end{array} \quad \text{for all } \mu,\tilde{v}$$

- However, this immediately returns us to the selection case by defining $S=1\left\{v\left(X\right)=\tilde{v}\right\}$
  - Hence, we know how to construct optimal estimators and confidence sets once we condition on the target parameter
- This route was advocated by Fithian et al (2017):

  Our guiding principle is: The answer must be valid, given that the question was asked.

Unconditional Inference

- Alternatively, we could focus just on unconditional bias and coverage, requiring that

$$\begin{array}{c} Med_{\mu}\left(\delta\left(X\right)-v\left(X\right)\right)=0 \\ Pr_{\mu}\left\{v\left(X\right)'\theta\in\delta\left(X\right)\right\}\geq 1-\alpha \end{array} \quad \text{for all } \mu$$

- Unconditional inference is less demanding
    - Any procedure that is conditionally valid for all $\tilde{v}$ is also unconditionally valid by the law of iterated expectations
    - This also means that the class of unconditionally valid procedures is larger $\Rightarrow$ may be able to obtain better performance

Unconditional Inference

- Berk et al. (2013)'s initial proposal for unconditional inference amounts to forming a joint confidence set for μ, that is, a set $CS_\mu = CS_\mu(X)$ such that

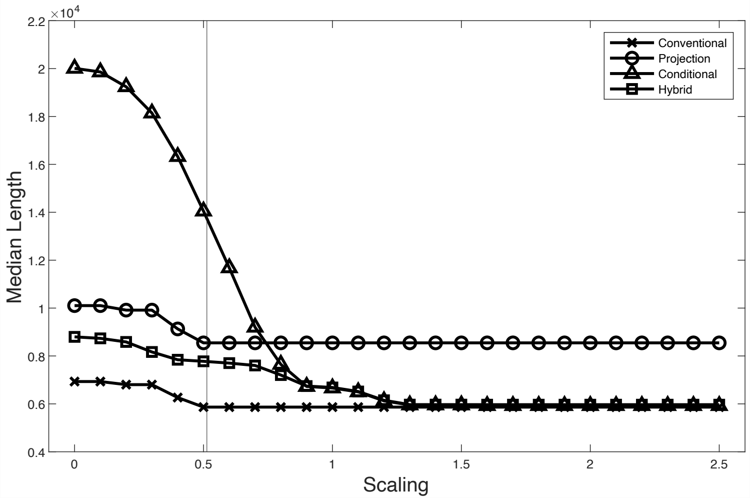
$$Pr_\mu \{\mu \in CS_\mu\} \geq 1 - \alpha \text{ for all } \mu,$$

  and then forming a confidence set for θ_X as

$$\delta(X) = \left\{v(X)' \mu : \mu \in CS_\mu\right\}$$

  - i.e. take the projection of CS_μ on the dimension of interest
- This ensures (unconditional) coverage, but it can result in confidence sets that are much longer than necessary
- On the other hand, an advantage of this approach is that we don't need to know the function $v(\cdot)$ to implement it - suffices to know $v(X)$
- When $v(\cdot)$ is known, Andrews et al. (2023) propose a hybrid approach that combines projection and conditioning

Example: Randomized Trial



Open Questions

Can We Relax Information Requirements?

- The available techniques to correct for screening and selection impose substantial information requirements
  - For screening, need to know $Pr\left\{S=1|X\right\}$
  - For selection, need to know either $v\left(\cdot\right)$ (for conditional and hybrid inference) or the set of possible target parameters θ_X (for projection inference)

  In many contexts, this is too demanding: we do not have an explicit description of what guides our choices
- In some contexts, we may be able to estimate screening or selection rules based on observed choices
  - Andrews and Kasy (2019) do this in the case of publication bias
- In other contexts, we may resort to sample-splitting
  - Screen or select based on part of the data, and use the remainder for inference
- Are there better options?

How to Think About Screening?

- Screening invalidates conventional inference
  - Motivates suggestions to reduce screening, e.g. pre-analysis plans, registered reports (i.e. pre-result peer review)
- However, this isn't the only option: once the form of screening is known we can correct for it
- Moreover, there are cases where screening seems to be helpful
  - Frankel and Kasy (2022) show that screening in favor of suprising results can be optimal for a journal seeking to inform readers
  - Screening based on specification tests is a common (implicit or explicit) suggestion

Inference After Selection

When the target parameter is θ_X, open questions include:

- Should we condition on the target parameter selected?
- If not, what's the right framework for optimal inference when the target parameter is random?

Thanks very much!

References

- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251)
- Bruns, S. B. and Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. PLoS One, 11(2)
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280): 1433-1436.
- Andrews, I. and M. Kasy (2019). “Identification of and Correction for Publication Bias,” American Economic Review, 109 (8): 2766-9.

References

- Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. Annals of Statistics, 41(2): 802-831.
- Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. Annals of Statistics, 44(3): 907-927.
- Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. arXiv.
- Andrews, I., T Kitagawa, and A. McCloskey (2023). Inference on Winners. arXiv.
- Frankel, A. and M. Kasy (2022). Which Findings Should Be Published? American Economic Journal: Applied, 14(1), 1-38.