How to avoid specifying a model

Christopher A. Sims
Princeton University
sims@princeton.edu

February 4, 2021

Overview

- Bootstrap, randomization inference (RI), and design-based (DB) inference have in common that they postulate a model whose distribution for the data has discrete support.

- This may simplify the theory of inference, or its computational implementation.

- Also, standard parametric frameworks with only a few parameters may not provide good approximations to complicated distributions, whereas it may seem that discrete distributions, expanding in complexity automatically with sample size, can appproximate nearly any kind of complex distribution.

Bootstrap

- The full or direct bootstrap models the data as generated from random draws, with replacement, from $\{Y_i, i = 1, \ldots, N\}$. That is, it specifies $P[Y_i$ in the sample $] = 1/n$, for each i. Note that Y_i is generally a vector, This distribution is called the **sample distribution**. It converges in distribution to the distribution of Y if the obervations are i.i.d.

- Of course we almost never believe this model generated the observed data. In most applications each observation has a unique value of Y_i. If on every draw from the distribution of the data every i has probability $1/n$, even for modest values of n drawing all n values of Y exactly once each has low probability. (E.g, with $n = 8$, this probability is .0024.)

Implementing the bootstrap

- We have some function β of the distribution of Y that we want to estimate and we have an estimator $\hat{\beta}(\{Y_i\})$. We then make a large number M of computer-generated, pseudo-random i.i.d. draws $\left\{Y_1^j, \ldots, Y_N^j\right\}$, $j = 1, \ldots, M$, with replacement, from the sample distribution. These sample draws will always consist of observed Y_i's from the data, though in a given sample draw some observed Y_i's may not appear and some may appear more than once.

- We can then calculate from each randomly drawn sample $\left\{Y_i^j, i = 1, \ldots, N\right\}$ the $\hat{\beta}_j$ value implied by that sample distribution and treat the $\hat{\beta}_j$'s as if they were draws from the true distribution of $\hat{\beta}(Y_1, \ldots, Y_N)$

Parametric bootstrap

- We usually believe that in fact Y_i, or some elements of it, are continuously distributed, which the full bootstrap model denies.

- If we have a model $p(y \mid \theta))$ and we have an estimator $\hat{\theta}$ for θ, we can calculate our $\hat{\theta}$ from our sample, then generate M samples of size N from the $p(y \mid \hat{\theta})$ model.

- We can then treat construct $\widehat{\theta(Y_1^j, \ldots, Y_N^j)}$ for each of our pseudo-random draws of the data set, as in the full bootstrap.

- This makes sense only if $\hat{\theta}$ is quite accurate for a sampple of the size we have.

- It has the advantage over the full bootstrap that if Y is continuously distributed, the parameteric-bootstrap Y samples will not contain repeated values.

What does the bootstrap deliver?

- It delivers an artificial sample of M draws from the sample distribution of $\beta(\{\hat{Y}_i\})$, and hence allows us to calculate the mean and variance of it, or even plot histograms for it, showing the shape of its distribution.

- But this is the distribution of the estimator, across repeated samples, not the Bayesian concept of a distribution for β, given the data.

- So a left skewness in the distribution of $\hat{\beta}$ in the bootstrapped sample, for example, is not a reliable indicator that the most likely value of β is lower than the estimate we obtain from the original observed N values of Y_i. In fact, under some assumptions it is an indicator of the opposite.

Puzzle: Confidence or credible sets from the bootstrap?

- Suppose we have an estimator $\hat{\theta}$ of a parameter θ, and its pdf conditional on the true value of θ is $p(\hat{\theta} - \theta)$.

- Only the location of the distribution, not its shape, depends on θ, but the distribution may be asymmetric about θ.

- The parametric bootstrap should work well then: Even though our initial estimate $\hat{\theta}$ is slightly incorrect, the parametric bootstrap will give us an accurate picture of the shape of the pdf.

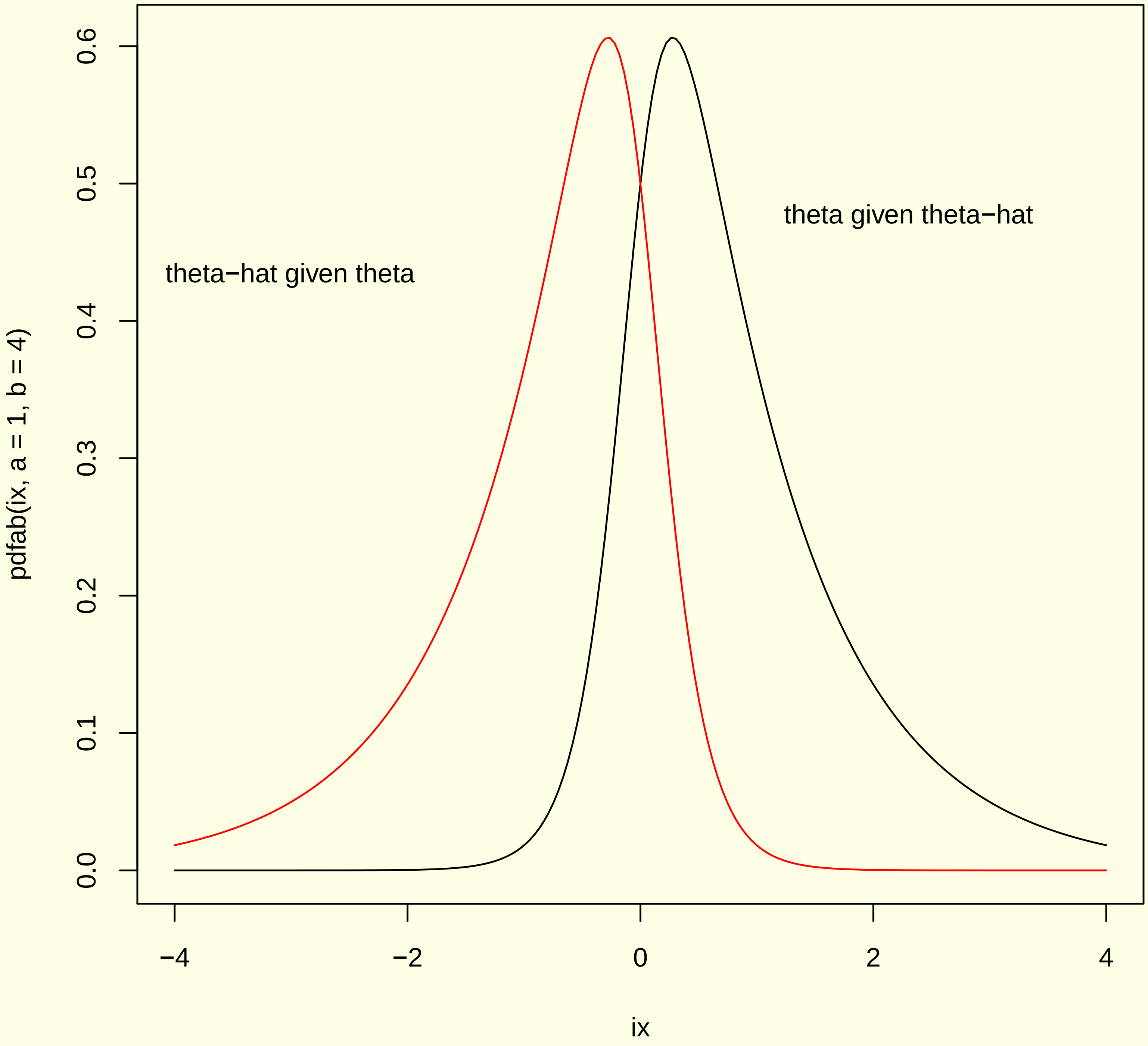
- The most commonly applied method to produce a confidence interval from bootstrap draws is to use the 2.5% and 97.5% quantiles of the bootstrap distribution as the confidence set.

- But the likelihood function for θ given $\hat{\theta}$ is $p(\hat{\theta} - \theta)$ *as a function of* θ, while the bootstrap has given us an estimate of the same thing as a function of $\hat{\theta}$.

- If the pdf of $\hat{\theta}$ is skewed left, that of θ is skewed right.

- It may help with the intuition to consider a case where $\hat{\theta}$ is downward biased. The naively bootstrapped CI will skew downwards, suggesting that the truth is more likely to be below than above the estimated value, whereas in fact, knowing the bias, we know the truth is more likely above the estimate.

- You may think there's an easy solution: Just flip the usual interval. But this is only reasonable in a pure location-shift case, and cases that are approximately pure location-shift and not at the same time approximately normal (and therefore symmetric) are not common.


theta−hat given theta
theta given theta−hat
pdfab(ix, a = 1, b = 4)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−4
−2
0
2
4
ix

Summary on bootstrap

- It is easy to implement and does not require an explicit model. (Though without a model, there is question whether you have a good reason for trying to estimate your β.)

- To characterize uncertainty by producing frequentist tests or confidence sets, one needs to add assumptions and rely on asymptotic arguments. In "regular" cases, basing inference on first and second moments of bootstrapped samples is justified as an asymptotic approximation.

Randomization inference

- It's an "automatic" way to generate a distribution for the data under a null hypothesis that you want to test, using the already-observed data rather than a fully specified model.

- We'll discuss it via one example: We have a sample $\{Y_i\}$ of size $2N$, of which N are "treated" and N "untreated". We'd like to test the null hypothesis that the means are the same for both groups.

- Obvious way to proceed: Find the sample mean and variance of Y_i within each group, treat the mean over the standard deviation as (asymptoticaly) independent $N(0,1)$, Form test statistic. This relies heavily on asymptotics.

The randomization test for the simple model

- Repeatedly allocate $\{Y_i\}$ randomly between the treated and untreated categories, calculate the resulting differences in means. This gives us $m_j, j = 1, \ldots, M$, an artifical sample of mean differences. Let m^* be the difference in means with the real data.

- (Note that if N is small, we may be able to exhaustively enumerate all possible ways to split the observations into two groups. In that case we don't need to make random draws; we can just enumerate the possibilities.)

- Reject the null at level α if the number of occurrences of $m_j > m^*$, divided by M, is less than α.

What does the randomization test deliver?

- It delivers a test of a single $H0$. The test is exact in finite samples under the null, so no appeal to asymptotics is required.
- It is easy to implement and somewhat intuitive.
- It cannot easily be used to produce confidence sets.
- It is a test of a null hypothesis that we would not in fact adopt as our beliefs if the test accepts the null.
- It is unclear what the test is powerful against. Constructing a test without explicit consideration of what the alternative hypothesis is can lead to nonsense.

Examples of problematic samples

$$\{1, 1, 1, 1, 5; .9, .9, .9, .9, .9\} \qquad \{.9, .9, .9, .9, 5; 1.1, 1.1, 1.1, 1.1, 1.1\}$$

In each of these, the first 5 are Y_i for the treated, with $N = 5$. In the first example, no way of reordering the 10 Y_i values can produce as large a mean difference as the one in the actual data, 0.9. So the probability of a mean difference as large or larger than that observed is one over the number of ways of allocating 10 objects into two equal size bins: 252 and the test rejects at the .004 level. The difference in means is extremely “significant”.

In the second example, every re-ordering that leaves the “5” in the treated category creates a greater mean difference than the actual $m^* = .72$, and of course half of all the possible orderings have the 5 in the treated group. So the probability of $m_i > m^*$ is .5; the difference in means is not at all significant.

What are the problems?

- The two samples are very close each other in Euclidean distance, yet the randomization test based on them has radically different conclusions.
- Any distribution we might have in mind for Y_i for the treated and Y_j for the untreated that had a continuous joint density for them would imply the two samples should have nearly the same implications.
- RI is sometimes seen as "conservative", because of its behavior in samples like the second, where it recognizes that if the sample contains some big outliers, the conclusions about differences in sample means could be determined mainly by these big outliers. But behavior like that in the first sample is also possible, where a more believable model would imply that it is harder to reject the null.

Reference on randomization inference vs Neyman (frequentist) inference vs model-based (Bayesian) inference

Imbens and Rubin (2015), chapters 5, 6, and 8. As in our example, this book focuses entirely on binary “treatments” (right-hand-side variables), though the basic ideas are more general. Athey and Imbens have recent work exploring design-based inference.

Design based inference

- Suppose we want to estimate the average height of all Princeton undergraduates in residence — not the average height in a "population" from which they are drawn, but the actual sample average over this finite number of people.

- We could of course in principle measure the heights of all of them and average. There would then be no "inference" to be done. We'd have the average height, a number. There would be no "standard error" on this.

- A more practical approach, if we didn't need the number down the the last millimeter, would be to draw randomly 100 from a list of the Princeton undergraduates in residence, and measure just them.

- The only “model” here is that asserting that our sampling is random. We would estimate the mean height as the sample average, and here a standard error based on asymptotic theory would be available.

- The usual regularity conditions, that the distribution we are drawing from be i.i.d. and with finite mean and variance are obviously satisfied.

An example where design-based inference would not work in practice

- Suppose that the heights of students in residence at Princeton are distributed like a draw of 5000 from a distribution with density $1/(1+h)^2$ on $0, \infty$.

- This density defines a distribution for which both mean and variance are infinite.

- Of course the actual 5000 student heights have a well-defined, finite mean and a well-defined, finite variance.

What goes wrong

- Despite the validity of the asymptotic theory, if their heights vary like draws from this density, the average heights from modest sized samples, say less than 1000, are likely to bear little relation to the population average.

- The only outliers are positive, and they are rare and large. So samples of modest size are likely not to contain outliers and to have means far below the population mean. In the rare instances where they do contain an outlier, the sample mean will be much too large.

How can the asymptotic theory still be valid?

- The asymptotic theory is still valid, since the repeated samples are i.i.d. and drawn from a finite-variance population,

- But it starts to become a good approximation only when the sample size greatly exceeds the size of the population! It starts to work only when most members of the population appear multiple times in the “sample”.

Is this example unrealistic?

- Yes, if we really are considering heights of people.

- But that's because we in fact know that the distribution of people's heights is unlikely to be extremely fat-tailed. We are invoking a priori knowledge about the nature of the distribution of the population.

- And if we were discussing not student heights, but instead sizes of firms, or student family incomes, or (as in an exercise in ECO517 last semester) sizes of cities, we actually know the distribution *is* likely to be fat-tailed, though we may be uncertain about how strongly fat-tailed it is.

- In that case, we can't avoid discussing what kind of "super-population" the population at hand is likely to have been drawn from.

- We should end up formulating and estimating a model for that distribution allowing for the rate of decline in the tails to be controlled by unknown parameters we estimate.

Bayesian bootstrap

- The bootstrap amounts to generating a sample from the distribution of the estimator, assuming that the sample distribution function is the actual distribution that generated the data.

- A small step toward realism would maintain the assumption that the N Y_i values that have occurred in sample are the full support of the distribution of the Y's, but that we don't know their probabilities.

A Dirichlet prior

- If we assume that their probabilities $\{p_1, \ldots, p_N\}$ have a prior distribution that is Dirichlet with the same parameter α_j for each p_j, then the posterior given the sample is

$$\prod_{j=1}^{N} p_j^{\alpha_j} \; .$$

- If some observations in the sample repeat n_j times, the corresponding term in the posterior density becomes $p_j^{\alpha_j + n_j - 1}$.

Choosing α

- Rubin, the inventor of the Bayesian bootstrap, suggested using $\alpha = 0$, which is an improper prior (and makes all the p_j's uniformly distributed in the posterior).

- A "Jeffreys prior" here (which approximates the idea of a "flat" prior) would make $\alpha = .5$; and there is some intuitive appeal to making our prior on each p_j flat, which corresponds to $\alpha = 1$

Drawing from the posterior

- We can easily make draws from this posterior. (If your computer package won't make draws from a Dirichlet with parameter vector $\vec{\alpha}$, draw independent gamma variates with shape parameter α and normalize them to sum to one.)
- For any such draw of the p's, we can compute any function of the distribution — e.g. the mean, or a regressioni coefficient, so that with M draws we can get a sample from the posterior distribution.

Why this instead of the standard bootstrap?

- As we have observed, the standard bootstrap gives you only the distribution of the estimator conditional on a particular value of the unkoown parameter. It thus can't directly provide confidence intervals or a posterior distribution. The Bayesian bootstrap directly delivers a sample from the posterior, from which credible sets can be formed.

*

References

IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference*. Cambridge University Press.