An Introduction to the Nonparametric Bootstrap

Charles J. Geyer

School of Statistics
Minnesota Center for Philosophy of Science
University of Minnesota

March 13, 2024

Slides for this talk:
http://users.stat.umn.edu/~geyer/boot.pdf

Relevant and Irrelevant Simulation

Statisticians love simulations that show their methods work great (at least on some toy problems).

IMHO, this is nonsense. Those simulations *prove* nothing. Those toy problems may be chosen (consciously or unconsciously) to make the methods look good.

And those toy problems are *very different* from statistical models you use for real data. So what could the simulation study possibly tell you that is relevant?

Relevant simulation — what everybody else calls the bootstrap — says that for each and every model you fit you do a simulation *for that model* not some irrelevant toy problem.

The Empirical Distribution

The *empirical distribution* is the probability distribution that puts equal probability at a set of data values if they are distinct. Otherwise puts probability proportional to the multiplicity of the data points.

A class of theorems called Glivenko–Cantelli theorems says the empirical distribution converges to the true unknown distribution of the data as the sample size goes to infinity for independent and identically distributed data.

This justifies using the empirical distribution as an estimator of the true unknown distribution.

The Bootstrap Metaphor

	Real World	Bootstrap World
true distribution	P	$\widehat{P}_n$
data	$X_1, \ldots, X_n$ IID P	$X_1^*, \ldots, X_n^*$ IID $\widehat{P}_n$
true parameter	$\theta = f(P)$	$\hat{\theta}_n = f(\widehat{P}_n)$
point estimate	$\hat{\theta}_n = g(X_1, \ldots, X_n)$	$\theta_n^* = g(X_1^*, \ldots, X_n^*)$
standard error	$\hat{s}_n = h(X_1, \ldots, X_n)$	$s_n^* = h(X_1^*, \ldots, X_n^*)$
pivotal quantity	$(\hat{\theta}_n - \theta)/\hat{s}_n$	$(\theta_n^* - \hat{\theta}_n)/s_n^*$

$\widehat{P}_n$ empirical distribution for $X_1, \ldots, X_n$.

IID independent and identically distributed (non-IID later)

pivotal quantity means sampling distribution does not depend on unknown parameters (exactly or approximately).

The Fundamental Principle of Statistics

Estimates are not the parameters they estimate. $\hat{\theta}_n$ is not θ.

Statistics tells us how to deal with this. It does not make uncertainty go away.

The bootstrap is just this taken up another level of abstraction. $\widehat{P}_n$ is not P.

Instead of a scalar parameter θ, what is unknown is the whole true distribution P of the data.

The Bootstrap Metaphor (cont.)

Because $\widehat{P}_n$ is not P, the bootstrap does the Wrong Thing.

But it is the best (nonparametric) thing to do (we should simulate from P but don't know P, so have to use $\widehat{P}_n$).

Hence — contrary to what many people seem to think — the bootstrap is **not** an exact small sample procedure like a t test.

It has only large sample (large n) validity (like many other statistical procedures). And n may need to be larger for the bootstrap than for many other large n procedures because P is more complicated to estimate than θ.

Moreover, the metaphor is not always good. There are many counterexamples where the bootstrap does not get close to the right answer for exceedingly large n (some fixes mentioned later).

Resampling

When you treat $X_1, \ldots, X_n$ as a finite population to sample from, and take an IID sample of size n, this is also called

- sampling with replacement from $X_1, \ldots, X_n$ or
- resampling.

But focus on these misses the point.

The point is we are using IID samples from *an estimate* of the true unknown distribution.

If we use some estimate other than $\widehat{P}_n$ (Efron's original proposal), and there is a lot of literature on various alternatives, then both of the bullet points above no longer describe it.

Pivotal Quantities

If $X_1, \ldots, X_n$ are (exactly) IID normal, and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

then

$$T_n = \frac{\overline{X}_n - \mu}{\hat{s}_n / \sqrt{n}}$$

has (exactly) a t distribution with $n - 1$ degrees of freedom.

We can use this to do (exact) hypothesis tests and confidence intervals (see intro stats).

There are *no unknown parameters* of this t distribution, so this is an (exact) *pivotal quantity*. And that is why this works!

Pivotal Quantities (cont.)

Pivotal quantities (exact or approximate) are what makes (frequentist) statistics work.

Without them you don't know the reference distribution for the procedure.

So even when we bootstrap, we need to bootstrap (approximate, at least) pivotal quantities. That is why we simulate the distribution of

$$T_n^* = \frac{\theta_n^* - \hat{\theta}_n}{s_n^*}$$

if we can assume the mean and variance of this (conditional on the original sample) don't depend (much) on unknown parameters, then we will have a good procedure. Otherwise not.

Bootstrap T

Simulate the sampling distribution of

$$T_n^* = \frac{\theta_n^* - \hat{\theta}_n}{s_n^*}$$

(formula repeated from preceding slide). For 95% confidence interval, look up 0.025 and 0.975 quantiles of this (bootstrap) simulated distribution. Call them c_1 and c_2.

Ignore Monte Carlo error (due to finite bootstrap sample size: number of simulations) and statistical error ($\widehat{P}_n$ is not P), so we pretend (the bootstrap metaphor) that above has the same distribution as

$$T_n = \frac{\hat{\theta}_n - \theta}{\hat{s}_n}$$

Bootstrap T (cont.)

Hence

$$\hat{\theta}_n - c_2 \hat{s}_n < \theta < \hat{\theta}_n - c_1 \hat{s}_n$$

is a 95% confidence interval for the true unknown θ.

Only approximate, not exact, but ....

Bootstrap T (cont.)

Suppose the t distribution had never been invented but the bootstrap had.

Then the bootstrap t procedure would *invent* the t distribution, because that would be the distribution of T_n^* that the bootstrap simulates.

Moreover the bootstrap does this for any population distribution (we do not need to assume normal data). So the bootstrap is better than theory.

Second-Order Correct

Call the difference between the actual coverage probability and the nominal coverage probability (0.95 for example) the *coverage error*.

For the asymptotics covered in most statistics courses and used by software that does not bootstrap, this error obeys square root law

$$\text{coverage error} \leq C / \sqrt{n}$$

where C is a constant.

The t distribution is exact (coverage error zero).

The bootstrap t distribution is second-order correct

$$\text{coverage error} \leq C / n$$

Not quite as good as exact, but way better than asymptotic theory.

Second-Order Correct (cont.)

Bootstrap t is not the only second-order correct procedure. Dozens in literature.

Find out about them and use them.

Bootstrap Percentile

The *bootstrap percentile* method is simpler than bootstrap t the endpoints of its 95% confidence intervals are the 0.025 and 0.975 quantiles of (the bootstrap sampling distribution of) θ_n^*.

These intervals are not second-order correct and not recommended by many experts.

Bootstrap Percentile (cont.)

When the sampling distribution of $\hat{\theta}_n$ is heavily skewed or biased — say $\hat{\theta}_n$ is usually below θ — then

- θ is usually above $\hat{\theta}_n$ so the confidence interval should extend farther above $\hat{\theta}_n$ than below. Bootstrap t and other second-order correct do this.
- Bootstrap percentile does not. Extends farther below $\hat{\theta}_n$ than above. Goes the wrong way!

Disclaimer: Hypothesis Tests

The (nonparametric) bootstrap does not naturally do hypothesis tests.

Reason: it samples (approximately, for large n) from the true unknown distribution of the data.

But hypothesis tests use the distribution *assuming the null hypothesis is correct* and often we are using the test to give evidence that is false.

So bootstrap hypothesis tests, naively done, have no power. They say $P \approx 0.5$ regardless of how far the data are from data from the null hypothesis.

Disclaimer: Hypothesis Tests (cont.)

Of course. one can always invert a bootstrap confidence interval to do a valid hypothesis test (which will have good power).

Disclaimer: Regression

The (nonparametric) bootstrap does not naturally do regression analysis.

Reason: it samples (approximately, for large n) from the true unknown **joint distribution** of the response and predictors.

But in the rest of statistics regression analysis is about the **conditional distribution** of the response given the predictors.

And that the (naive) nonparametric bootstrap cannot do.

Bootstrapping Cases versus Bootstrapping Residuals

Bootstrapping cases means you sample IID from the original data considering both **response** and **predictors** as random. This samples from the approximate joint distribution of response and predictors, which is usually **not** what is wanted.

When you have a regression model

$$y_i = x_i^T \beta + e_i$$

with estimated coefficients $\hat{\beta}$ and residuals

$$\hat{e}_i = y_i - x_i^T \hat{\beta}$$

then bootstrapping residuals means you sample IID from the residuals obtaining bootstrap residuals e_i^*, and then the bootstrap data are

$$y_i^* = x_i^T \hat{\beta} + e_i^*$$

Bootstrapping Cases versus Bootstrapping Residuals (cont.)

Bootstrapping residuals does sample from (approximately) the true unknown conditional distribution of response given predictors **but is no longer fully nonparametric** because it depends on the (parametric) regression model being correct.

One can use internally or externally standardized residuals to bootstrap sample from, but this doesn't change what was said above.

When the Bootstrap Does Not Work

Efron's original nonparametric bootstrap obviously does the wrong thing with dependent data (it mimics IID not dependence). So it does not work for time series, spatial statistics, network models, statistical genetics, and so forth.

It also does not work (for obscure technical reasons) when the square root law and asymptotic normality do not hold, that is when

$$\sqrt{n}(\hat{\theta}_n - \theta)$$

does not converge to a normal distribution.

One fix solves both.

Subsampling Bootstrap

What I call the subsampling bootstrap and its original authors call just subsampling is a different idea.

Suppose

$$\frac{n^\alpha(\hat{\theta}_n - \theta)}{\hat{s}_n} \xrightarrow{\mathcal{D}} Y, \qquad \text{as } n \to \infty,$$

for any rate $\alpha > 0$ and any distribution of Y. Trivially,

$$\frac{b^\alpha(\hat{\theta}_b - \theta)}{\hat{s}_b} \xrightarrow{\mathcal{D}} Y, \qquad \text{as } b \to \infty,$$

And we can use this to bootstrap at a different sample size b from the actual sample size n.

We only have to know or estimate the rate α.

Subsampling Bootstrap (cont.)

The subsampling bootstrap does not sample IID from the data considered as a population (like Efron's original nonparametric bootstrap).

Rather it samples *without replacement* from IID data and samples blocks of length b from time series data or analogous blocks from other dependent data.

This makes the subsampling data X^* have the true unknown distribution (for sample size b).

Subsampling Bootstrap (cont.)

Efron original nonparametric bootstrap

- samples from the wrong distribution $\widehat{P}_n$
- at the right sample size n.

The subsampling bootstrap

- samples from the right distribution P
- at the wrong sample size b.

Subsampling Bootstrap (cont.)

Now we are really relying on the asymptotics

$$\frac{n^{\alpha}(\hat{\theta}_n - \theta)}{\hat{s}_n} \xrightarrow{\mathcal{D}} Y, \qquad \text{as } n \to \infty,$$

even if we have no idea what the distribution of Y is (the subsampling bootstrap estimates that). Look up 0.025 and 0.975 quantiles of this (subsampling bootstrap) simulated distribution. Call them c_1 and c_2.

Now our 95% confidence interval for the true unknown θ is

$$\hat{\theta}_n - \left(\frac{b}{n}\right)^{\alpha} c_2 \hat{s}_n < \theta < \hat{\theta}_n - \left(\frac{b}{n}\right)^{\alpha} c_1 \hat{s}_n$$

so we do have to know that the asymptotics exists and the rate α (or we can estimate α by subsampling at different sample sizes following Chapter 8 of the subsampling book or these notes.

References

Angelo Canty and Brian Ripley (2022).
R package `boot`: Bootstrap R (S-Plus) Functions, version 1.3-28.1.
https:cran.r-project.org/package=boot.

Bradley Efron, and R. J. Tibshirani (1994).
An Introduction to the Bootstrap.
Chapman & Hall/CRC, Boca Raton, FL.

A. C. Davison and D. V. Hinkley (1997).
Bootstrap Methods and their Application.
Cambridge University Press, Cambridge.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf (1999).
Subsampling.
Springer, New York.

Rob Tibshirani and Friedrich Leisch (2019).
R package `bootstrap`: Functions for the Book *An Introduction to the Bootstrap*, version 2019.6.
https://CRAN.R-project.org/package=bootstrap.