The Bootstrap (I)

Suppose that $X_1, \ldots, X_n$ is a random sample from some distribution.

We can estimate the population mean μ by $\bar{X}$, and the standard error (estimated standard deviation) is $S/\sqrt{n}$, where [S is the sample standard deviation.]

Similarly, if we want to estimate a proportion p using n independent trials, the estimate is $\hat{p}$, with standard error $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

The estimate and the SE can be combined to give asymptotic confidence intervals as: estimate $\pm z_{\alpha/2} SE$.

⇒ But what do we do for other parameters?

比如 median 和其他类的而不是等
或者 mean, proportion

The Bootstrap (II)

定义：

The bootstrap is a way to find an SE for any estimate or a confidence interval for any parameter.

It doesn't work well in every situation, but tends to work well in many situations, especially those where asymptotic normality holds.

Key Idea: Learn about the variability in an estimate over repeated sampling by using sampling from the sample to approximate sampling from the population.

Why "bootstrap"?: It conveys the idea that we're getting extra information that shouldn't be there in some sense by using the sample in more than one way.

Bootstrap Samples

Given $X_1, \ldots, X_n$, a bootstrap sample is a sample of size n drawn with replacement from the sample.

We get a bootstrap sample by sampling from the distribution w/ distribution function $\hat{F}$ (the EDF).

Note: This is the nonparametric bootstrap. It is also possible to do a parametric bootstrap by assuming a particular parametric family like the normal family.

Ex: If the data are $X_1 = 1$, $X_2 = 4$, list all possible bootstrap samples and their probabilities.

Solution: $2^2 = 4$

Sample	Prob.	Sample	Prob
1, 1	1/4	4, 1	1/4
1, 4	1/4	4, 4	1/4

← Ordered samples →

Small Bootstrap Example

Ex: Find the bootstrap distribution of the sample median (median) of the data are $X_1 = 3$, $X_2 = 5$, $X_3 = 7$. $(n = 3)$.

Solution: There are $3^3 = 27$ possible samples.

	median		median		median
333	3	533	3	733	3
335	3	535	5	735	5
337	3	537	5	737	7
353	3	553	5	753	5
355	5	555	5	755	5
357	5	557	5	757	7
373	3	573	5	773	7
375	5	575	5	775	7
377	7	577	7	777	7

sample median	prob.
3	7/27
5	13/27
7	7/27

The Bootstrap Via Simulation (I)

(When the sample is larger listing is not feasible.)

Instead, we do a simulation study.

Suppose we are estimating a parameter θ using $\hat{\theta}$. ← our estimator

Procedure: ① Draw B different bootstrap samples for

B = 500, 1000, or more.

② Compute the bootstrap estimates $\hat{\theta}_{b,1}, \ldots \hat{\theta}_{b,B}$.

boot $\hat{\theta}$ is sample mean. you gonna get sample mean for all those bootstrap example

③ Obtain the estimated MSE (mean squared error)

as $\widehat{MSE} = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_{b,i} - \hat{\theta})^2$ and the

note: $\hat{\theta}$ plays the role of true promotion (실제) value for ex: the mean for the sample you got

SE as $\sqrt{\widehat{MSE}}$.

Estimates $E[(\hat{\theta}-\theta)^2]$.

The Bootstrap Via Simulation (II)

We can also estimate other quantities.

Bias: The bias is $B(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$.

Estimate this as $\hat{B}(\hat{\theta}) = \underbrace{\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i}}_{\hat{E}} - \hat{\theta}$

$\hat{\theta}$ plays the role of θ.

为实际参数

Variance: Estimate this as

$$\hat{V}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_{b,i} - \hat{E})^2$$

$\hat{E}$ estimated expected value.

Note: Some references use $\sqrt{\hat{V}(\hat{\theta})}$ as the bootstrap SE. Using $\sqrt{\widehat{MSE}(\hat{\theta})}$ allows us to account for bias, too. ($MSE = \text{Variance} + \text{Bias}^2$).

Bootstrap Example

Ex: Consider the data 3, 4, 4, 5, 7, 9, 11, 13, 14, 19.

$n = 10$

目的: Using the bootstrap with $B = 1000$ bootstrap samples, estimate the MSE, the bias, & the variance associated with

(a) Using the sample median to estimate the population median and

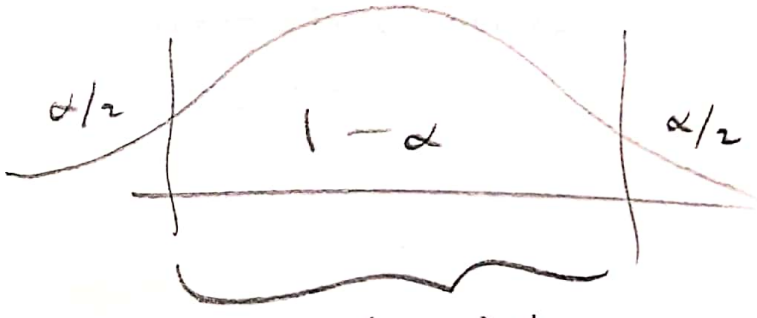
(b) Using the sample standard deviation to estimate the population standard deviation.

In each case, also give the point estimate and SE.

Bootstrap Percentile Confidence Intervals

One of the simplest methods for obtaining bootstrap confidence intervals is called the percentile method.

We find the B bootstrap point estimates $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,B}$. We then put these point estimates in order, and our $100(1-\alpha)\%$ confidence interval for θ goes from the $(\alpha/2)$ quantile of the bootstrap $\hat{\theta}$ values to the $1-\alpha/2$ quantile.



Note: These CIs do not necessarily perform well, particularly if the distribution for $\hat{\theta}$ is skewed.

Percentile Confidence Intervals, Example

Using the 10 data values from earlier, find 95% bootstrap percentile confidence intervals for the population mean, median, and standard deviation.

One helpful R function: quantile.

Using quantile(x, c(.025, .975)) gives a 95% bootstrap percentile confidence interval.

- x: bootstrap $\hat{\theta}$ values
- c(.025, .975): quantiles of interest

How well do the intervals work?

↳ { coverage probability
short CI

Assuming normal data and using B = 1000,

n = 5, 10, 20, and confidence level 95%,

estimate the true coverage probability for 95%

bootstrap percentile confidence intervals for the

population mean.

Then repeat the same study using exponential

data to assess the impact of skewness.

n = 5 79%
n = 10 88%
n = 20 94%
n = 40 91.4%

n is sample size:

for normal data

n = 5 ⇒ 0.781
n = 10 0.9
n = 20 0.94
n = 40 0.934

Close 95%

not so great for small sample size

Better CIs by Pivoting

A pivotal quantity is a random variable with a distribution that doesn't depend on any unknown parameters.

Ex: If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then both (population mean)

① $t = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ ($\bar{X}$: sample mean) and ② $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$

are pivotal quantities. ∵ no matter what μ & σ's are

Their distributions?

① t w/ $n-1$ df

② $\chi^2(n-1)$.

} free of μ and σ^2

Pivotal Quantities Lead to CI's

(I)

Ex: Show how to derive a CI for a normal mean μ by using the fact that $t \sim t_{n-1}$.

Soln: $$P\left(-t_{\alpha/2} \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t_{\alpha/2}\right) = 1-\alpha \quad \leftarrow t$$

$$\Rightarrow P\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow \boxed{\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \text{ is a } 100(1-\alpha)\% \text{ CI for } \mu.}$$

Pivotal Quantities Lead to CIs (II)

Ex: Show how to find a CI for a normal variance σ^2 by using the fact that $X^2 \sim \chi^2_{n-1}$.

Soln: $P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2} \right) = 1-\alpha$ ← In upper tail.

$$\Rightarrow P\left(\frac{1}{\chi^2_{n-1,\alpha/2}} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{\chi^2_{n-1,1-\alpha/2}} \right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right) = 1-\alpha.$$

$$\Rightarrow \left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right) \text{ is a } 100(1-\alpha)\% \text{ CI for } \sigma^2.$$

Bootstrap Pivotal CIs

The idea is to use the bootstrap to estimate the appropriate critical values for non-normal data.

Suppose that the data are non-normal, but do come from a location scale family so that for any distribution

$$g(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$ is the pdf.

↑ rescaled version of a base pdf f

Then t & X^2 are still pivotal, but they no longer have the t_{n-1} and χ^2_{n-1} distributions respectively.

$\Rightarrow$ We need different critical values.

Pivotal CIs for the mean μ

Let $t_{b,.025}$ and $t_{b,.975}$ satisfy

$$P(t_{b,.025} \le t \le t_{b,.975}) = 0.95.$$

Instead use t critical value for this formula. We gonna use critical value from bootstrap dist for t

Then $$P\left(t_{b,.025} \le \frac{\bar{x} - \mu}{s/\sqrt{n}} \le t_{b,.975}\right) = 0.95$$

$$\Rightarrow P\left(t_{b,.025}\frac{s}{\sqrt{n}} \le \bar{x} - \mu \le t_{b,.975}\frac{s}{\sqrt{n}}\right) = 0.95$$

$$\Rightarrow P\left(-\bar{x} + t_{b,.025}\frac{s}{\sqrt{n}} \le -\mu \le -x + t_{b,.975}\frac{s}{\sqrt{n}}\right) = 0.95$$

$$\Rightarrow P\left(\bar{x} - t_{b,.975}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} - t_{b,.025}\frac{s}{\sqrt{n}}\right) = 0.95$$

$$\Rightarrow \left(\bar{x} - t_{b,.975}\frac{s}{\sqrt{n}},\ \bar{x} - t_{b,.025}\frac{s}{\sqrt{n}}\right) \text{ is a 95\% CI for } \mu.$$

usually negative

Pivotal CIs for the variance σ^2

By a similar argument, the 95% CI will be

$$\left(\frac{(n-1)s^2}{\chi^2_{b,.975}} \;\; \frac{(n-1)s^2}{\chi^2_{b,.025}} \right)$$

$\chi^2_{b,.975}$ ← 97.5th percentile

$\chi^2_{b,.025}$ ← 2.5th percentile

We estimate the needed percentiles for the distributions of t and X^2 using the bootstrap.

Ex: Find bootstrap 95% pivotal confidence intervals for μ and σ^2 using the data from the earlier $n=10$ example.

See next page first.

Important Clarification

When we do the bootstrap to estimate the distribution for $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ or $X^2 = \frac{(n-1)s^2}{\sigma^2}$, the sample plays the role of the population, and the bootstrap sample plays the role of the sample.

t: $t_b = \frac{\bar{x}_b - \bar{x}}{s_b/\sqrt{n}}$ ← from sample

X^2: $X_b^2 = \frac{(n-1)\, s_b^2}{s^2}$

Comparison Example

Compare the performance of (normal-theory), (percentile bootstrap), and (bootstrap pivotal confidence intervals for the population variance.)

In particular, how well do 95% confidence intervals maintain their level?

Use ① normal data and ② exponential data.

Use sample sizes of 5, 10, 20, 40.

Q: What if we have uniform data?

Extending the Basic Bootstrap

There are many different versions of the bootstrap for different kinds of data.

What can we do if we have bivariate or multivariate data?

Ex: Here's a sample of heights for randomly selected mother/daughter pairs. The units are inches.

Mother	Daughter
70	67
69	64
65	62
64	64
66	69
65	70
64	65
66	66
60	63
70	74

$n = 10$ pairs

Using the bootstrap, find

(a) r, (b) an SE for r,

(c) an estimate of the bias of r,

and

(d) a bootstrap 95% CI for ρ, the population correlation.

Extending the Basic Bootstrap (II)

Ex: Ten subjects were randomly assigned to each of two treatments. Find

(a) A bootstrap SE for $\bar{X}_1 - \bar{X}_2$.

(b) A bootstrap 95% CI for $\mu_1 - \mu_2$

(c) A bootstrap 95% CI for σ_1^2 / σ_2^2.

How would we do this?

Note: There are multiple possible approaches that one might use. A key thing for success of the bootstrap is to mimic the original sampling as much as possible.

Two- Sample Example

Treatment	Values				
1	9	12	12	14	17
	19	21	22	26	31
2	8	9	10	11	13
	13	19	21	22	24