PHS2000B Lab 6

Bootstrap

03/09/20

Contents

Background **1**
- Identification versus estimation . . . 1
- The superpopulation model . . . 2
- What is the sampling distribution? . . . 3
- What is a standard error? . . . 4
- Classical approaches to asymptotic standard errors . . . 4

The bootstrap **4**
- The fundamental idea . . . 4
- Inference based on the bootstrap . . . 5
  - Standard errors . . . 5
  - Confidence intervals . . . 6
  - Hypothesis testing and p-values . . . 6
  - The bootstrap-t . . . 6

Bootstrap variants **7**
- Cluster/Block bootstrap . . . 7
- Residual bootstrap . . . 7
- Wild bootstrap . . . 8
- Examples of when the bootstrap might perform better than classic approaches . . . 8

Bootstrap myths and misconceptions **10**

Implementation in `R` **10**

References **10**

Background

The bootstrap is a widely used *resampling technique* for obtaining standard errors and confidence intervals for complex estimators or when the parametric assumptions underlying traditional methods fail. It is a simple but powerful tool that is an essential part of any applied statistician's toolbox. However, to truly understand why we need the bootstrap and how it works, we need to briefly revisit some of the foundational statistical principles we learned in PHS2000A.

Identification versus estimation

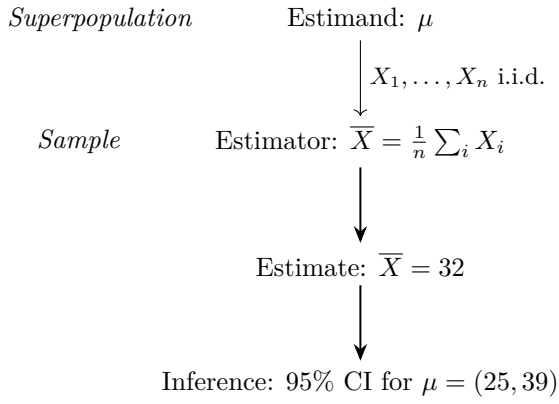
Recall that a fundamental task in the population sciences is collecting observations on a subset of individuals and using them to draw inferences about population quantities of interest. These quantities could be descriptive, like the proportion of people infected with the novel coronavirus, or casual, like the effect of the distribution of insecticed-treated bednets on the incidence of malaria, but regardless our project is generally to attempt to draw general conclusions based on the data. The inference problem can be usefully divided into two components: the problem of identification and the problem of estimation.

- The problem of *identification* refers to whether our target quantity could ever be unbiasedly determined from our data, e.g. if, for instance, we had an infinite sample size.
- The problem of *estimation* refers to the weaker set of conclusions that can be drawn about our target given that we only observe a finite set of random samples, even if it is identified.

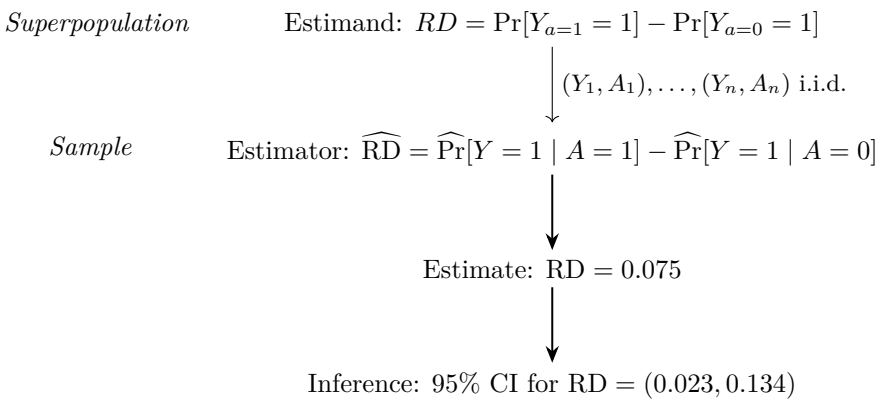
You can think of the problem of identification as relating to systematic error in our inferences while the problem of estimation relates to inferences based on the random variability inherent in the sampling process. When the target of interest is a causal effect, identification is based on the conditions we've discussed: exchangeability, consistency, and positivity (or alternatively no bias due to confouding, selection, or measurement). On this basis, it seems logical that identification should proceed estimation as a lack of identifiability implies that estimation is fruitless (we couldn't get the right answer even if we had infinite data). However, most of the inferential tools at our disposal —hypothesis tests, confidence intervals, and p-values— are only useful to characterize uncertainty in estimation, and can fail spectacularly[1] when the underlying target is not identifiable. Likewise the bootstrap is just another technique for characterizing uncertainty in estimation and as such only makes sense when the target is identifiable. For the purposes of this lab we will just assume from here on out that the identifiability conditions hold, but note that they are always a necessary condition for any estimation of uncertainty to be valid.

The superpopulation model

In order to characterize how likely an observed result is under random variability, we make a couple of useful assumptions. We assume there's a near-infinite and well-defined superpopulation and our data is a random sample from this superpopulation. Our goal is to make inferences about identifiable superpopulation quantities. We call this population target quantity the *estimand* (e.g. mean height of women under 30 in the population, the population average treatment effect on infection risk if everyone washed their hands versus if no one did). An *estimator* is a rule/recipe we use for taking the data from a sample and producing a numerical value for the the estimand. The numeric value for the estimand in a particular sample is an *estimate*. Finally we use statistical theory to quantify how compatible our estimate is with different values of the estimand, for instance by calculating a confidence interval.

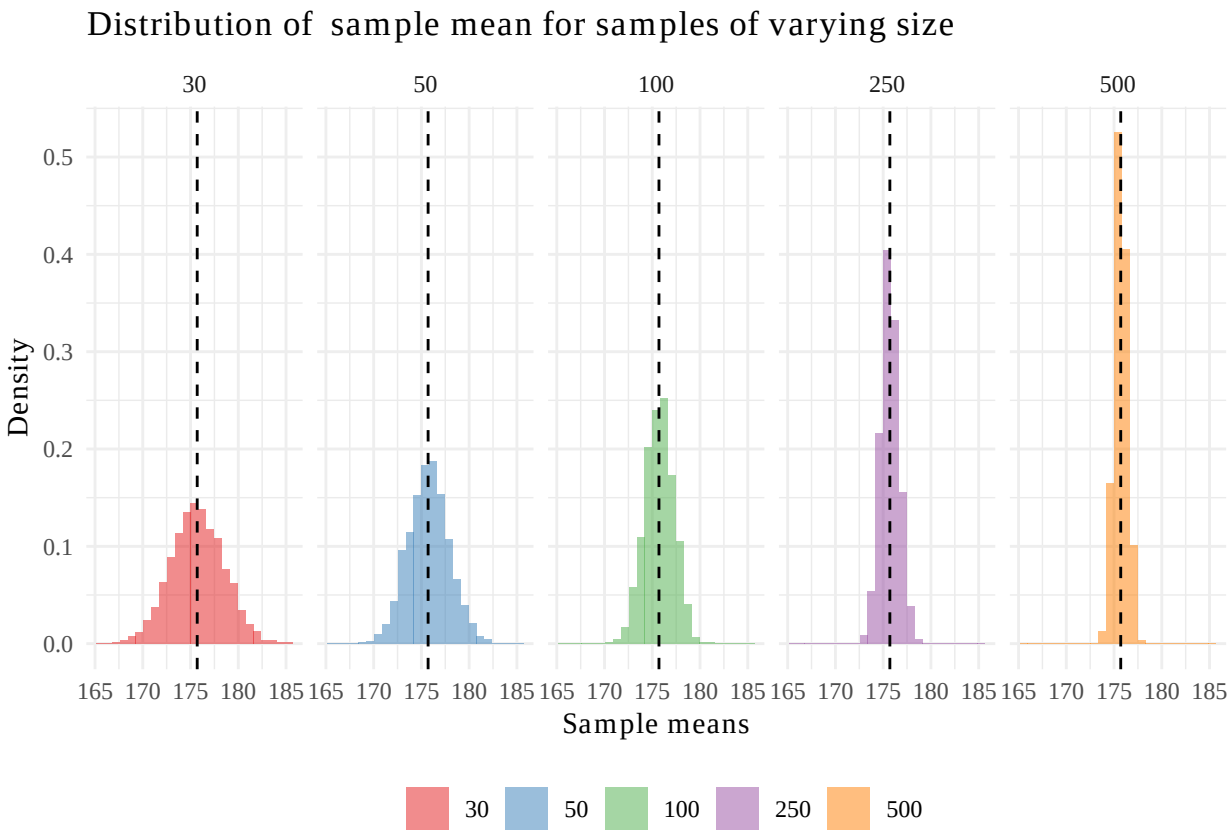


[1] This is what Miguel emphasized during his lecture, that p-values and confidence intervals no longer retain their original meaning if the underlying identifiability assumptions no longer hold



What is the sampling distribution?

A useful concept in the quantification of uncertainty in our estimates is **the sampling distribution**. The sampling distribution is simply the distribution of estimates across all possible samples of the same size from the same population. It's useful because it tells us how likely a specific estimate value is to occur due simply to random variability during sampling. In general, the sampling distribution is a function of the sample size and the intrinsic variability of the estimand/target quantity in the population. As the sample size increases the sampling distribution for a consistent estimator will get tighter and tighter around the true value, reflecting additional confidence in the likely values of the population quantity.



What is a standard error?

A standard error is simply the standard deviation of the sampling distribution. If we were to take a number of samples from the population and get an estimate from each one (repeating the sampling and estimation methods exactly each time), we could literally take the standard deviation of those values to estimate the standard error. Recall that the standard error decreases as the sample size increases: e.g., the standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$. For example, if you are estimating mean height in the U.S., you will get a more precise estimate if you have 1000 subjects instead of 100. What about the σ in the numerator? Intuitively, if we are estimating the mean of a random variable that has little variablility in the population, we will get more precise estimates than if we're estimating the mean of something that's highly variable.

Classical approaches to asymptotic standard errors

For the majority of estimators that we care about (e.g., sample means, regression coefficients, etc.), someone has already used statistical theory to derive an exact or approximate standard error. For example, $\widehat{SE} = \frac{\sigma}{\sqrt{n}}$ for the sample mean is an approximation based on the Central Limit Theorem. Sometimes these theoretically-derived standard errors are valid only when the data fulfill certain distributional assumptions.

What if you're interested in a more exotic statistic – like the marginal risk difference – or have data that don't fulfill distributional assumptions?

There are a few options:

1. Try to mathematically derive or approximate the standard error for your specific situation.
2. Rob a bank. Use the money to repeat your experiment 10,000 times. Calculate your exotic statistic each time. Take the standard deviation of the estimates.
3. Instead of robbing a bank, use your single experiment to *approximate* what would happen if you had actually repeated the experiment 10,000 times. We can do this using **bootstrapping**.

The bootstrap

The bootstrap is a very general algorithm for doing inference. You can bootstrap standard errors, confidence intervals, and hypothesis tests. All of it can be done without complicated variance derivations, distributional assumptions, or large sample theory.

The fundamental idea

The basic idea behind the bootstrap is quite elegant: given that we already assume that our data are a random sample from the population why not pretend that the population distribution that we are sampling from looks exactly like the sample that we have (a reasonable assumption under random sampling). Then we can just draw samples from the empirical distribution in our dataset to simulate the sampling distribution of our estimator and use that to calculate confidence intervals and p-values.

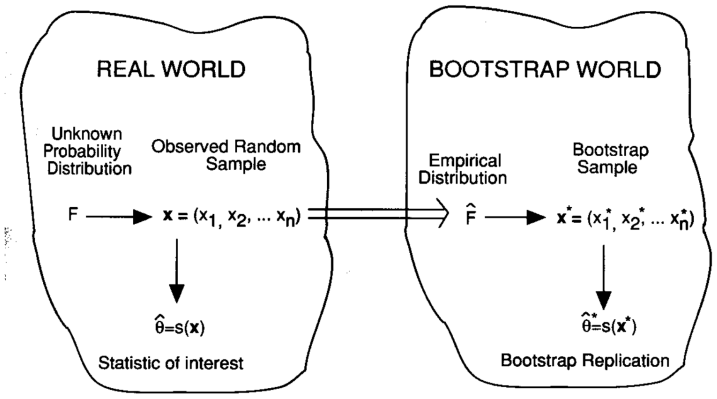


Figure 8.1. *A schematic diagram of the bootstrap as it applies to one-sample problems. In the real world, the unknown probability distribution F gives the data $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ by random sampling; from $\mathbf{x}$ we calculate the statistic of interest $\hat{\theta} = s(\mathbf{x})$. In the bootstrap world, $\hat{F}$ generates $\mathbf{x}^*$ by random sampling, giving $\hat{\theta}^* = s(\mathbf{x}^*)$. There is only one observed value of $\hat{\theta}$, but we can generate as many bootstrap replications $\hat{\theta}^*$ as affordable. The crucial step in the bootstrap process is "$\Longrightarrow$", the process by which we construct from $\mathbf{x}$ an estimate $\hat{F}$ of the unknown population F.*

1. Draw sample of size n with replacement from the observed data
2. Calculate our statistic of interest in the simulated bootstrap sample $\widehat{\theta}_b$
3. Repeat steps 1 and 2 a total of B times
4. Use the resulting collection of B bootstrap estimates $\{\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_B\}$ as an estimate of the sampling distribution we would have observed under repeated sampling from superpopulation.
5. Use simulated sampling distribution to calculate inferential quantities (e.g. confidence intervals, p-values, etc.)

Bootstrapping assumes that the sample at hand reflects the relative distribution of the underlying variables in the population – fully random representative sample. If the relative frequency of each "type" of observation is the same in the sample as it is in the population, we can generate "artificial samples" that are also representative of the underlying population.

By estimating the relation of interest in the bootstrapped samples, we can learn about the empirical distribution of point estimates without making any parametric assumptions. The closer the sample distribution of X to the true population distribution, the closer the in-sample bootstrapping gets to the "true" population distribution, and the closer the distribution of the bootstrapped standard errors will get to the distribution of the standard errors in independent random samples of the same size. The more diverse the underlying population, and the smaller the sample at hand, the larger the differences between the true population distribution and the distribution generated by bootstrapping.

Inference based on the bootstrap

Standard errors

As the bootstrap is meant to simulate the sampling distribution for our estimator, we can estimate the standard error by just taking the standard deviation of our bootstrap estimates $\{\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_B\}$.

$$SE_{boot}(\widehat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B}(\widehat{\theta}_b - \overline{\theta}_b)}{B-1}}$$

Confidence intervals

There are two general approaches to using the bootstrap to calculate confidence intervals. The first may look very familiar. Under the normal approximation, we could just take the estimate of the standard error we found by taking the standard deviation of the boostrap distribution from above, multiply it by 1.96 and add/subtract it from our original estimate to get a 95% confidence interval.

$$95\%\ CI_{boot}(\theta) = \left(\widehat{\theta} - Z_{0.975}SE_{boot}(\widehat{\theta}),\ \widehat{\theta} + Z_{0.975}SE_{boot}(\widehat{\theta})\right)$$

This assumes that the sampling distribution at this sample size is at least approximately normal.

However another method that does not rely on any assumption about the underlying distribution is to just calculate the 2.5th and 97.5th percentiles of the bootstrap distribution and use those to define our interval. This goes back to the fundamental concept of what a confidence interval is (i.e. an interval that will contain the true value of the estimand in 95% of repeated samples).

$$95\%\ CI_{boot}(\theta) = \left(q_{2.5}(\widehat{\theta}_b), q_{97.5}(\widehat{\theta}_b)\right)$$

Hypothesis testing and p-values

We can perform hypothesis tests of the form

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

by either again using a normal approximation

$$\frac{\widehat{\theta} - \theta_0}{SE_{boot}(\widehat{\theta})} \sim N(0,1)$$

or by using the equivalence between confidence intervals and hypothesis tests to conduct a test.

$$\text{Reject } H_0 \text{ if } \theta_0 \in \left(q_{2.5}(\widehat{\theta}_b), q_{97.5}(\widehat{\theta}_b)\right)$$

The bootstrap-t

Notice that, under certain regularity conditions, the bootstrap procedure can work for any function statistic or function of the observed data, e.g. the mean, the median, a β coefficient, the marginal risk difference, etc. This can be useful when we want to get standard errors or confidence intervals for some complex statistic or combination of regression coefficients where there is no other available method for calculating asymptotic properties. However, in some circumstances some statistics may also just perform better under boostrap than others. For instance, a commonly used variant of the bootstrap collects the t-statistic in each bootstrap sample. This variant tends to have better properties in finite samples than just bootstrapping the regression coefficient for instance because the t-statistic is "asymptotically pivotal" (i.e. does not depend on any unknown information). This modified bootstrap-t procedure is as follows:

1. Draw sample of size n with replacement from the observed data
2. Calculate our statistic of interest in the simulated bootstrap sample $t_b = \frac{\widehat{\theta}_b}{SE(\widehat{\theta}_b)}$

3. Repeat steps 1 and 2 a total of B times
4. Use the resulting collection of B bootstrap estimates $\{t_1, t_2, \ldots, t_B\}$ as an estimate of the sampling distribution of our observed t-statistic if we repeatedly sampled from superpopulation.
5. Use simulated sampling distribution to calculate a p-value by $P(|t^*| \geq q_{0.975}(t_b))$ where t^* is observed t-statistic in real data.

While the bootstrap-t is a really powerful procedure for hypothesis testing/calculating p-values, a downside of the bootstrap-t is that it cannot be used to calculate standard errors or confidence intervals.

Bootstrap variants

Cluster/Block bootstrap

What if our data are not i.i.d. samples from a superpopulation but instead might be correlated; for example, what if our data contain multiple measurements on the same individual over time or are randomly sampled within geographic clusters? Well as before we can use the insight that our bootstrap procedure is an insample stand in for the sampling/data generation mechanism to tweak the basic bootstrap procedure to accomodate correlated data. The basic idea is to resample blocks or clusters of possibly correlated observations rather than resampling individual observations.

1. If data consists of n observations from C clusters, block sample C of the original clusters with replacement from the observed data
2. Calculate our statistic of interest in the simulated bootstrap sample $\widehat{\theta}_b$
3. Repeat steps 1 and 2 a total of B times
4. Use the resulting collection of B bootstrap estimates $\{\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_B\}$ as an estimate of the sampling distribution we would have observed under repeated sampling of clusters from superpopulation.
5. Use simulated sampling distribution to calculate inferential quantities (e.g. confidence intervals, p-values, etc.)

Residual bootstrap

In the regression context sometimes it makes more sense to view the Xs as fixed and resample the errors/residuals.

1. fit $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ to the real data, predict $\widehat{Y}_i$ and $\widehat{\varepsilon}_i$
2. in each bootstrap iteration, $b = 1, 2, \ldots, B$
   (a) assign each observation, i, a new ε^*_{ib} drawn randomly from the estimated residuals from observed data $\{\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \ldots, \widehat{\varepsilon}_N\}$
   (b) calculate a new outcome by adding resampled residual to predicted outcome from observed data $Y^*_{ib} = \widehat{Y}_i + \varepsilon^*_{ib}$
   (c) fit $Y^*_{ib} = \beta_0 + \beta_1 X_i + \varepsilon_{ib}$ to obtain $\widehat{\beta}_b$ or t_b.
3. Use the resulting collection of B bootstrap estimates $\{\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_B\}$ as an estimate of the sampling distribution we would have observed under repeated sampling from superpopulation.
4. Use simulated sampling distribution to calculate inferential quantities (e.g. confidence intervals, p-values, etc.)

The residual method is more natural/efficient in the regression context; however while it makes no assumptions about the exact distribution of the residuals it does assume that they are homoscedastic and therefore doesn't make sense to use in cases where you expect heteroscedasticity.

This method can be adapted for clustered data by resampling blocks of residuals instead of individual residuals however it does require equal cluster size (otherwise it's unclear how to assign resampled residuals).

Wild bootstrap

An alternative regression-based bootstrap technique is the so-called wild bootstrap. Unlike the residual method, the wild bootstrap preserves the $\{\varepsilon_i, X_i\}$ relationship and therefore can be used when there is suspected heteroscedasticity. It also does not require equal cluster/group sizes when using the cluster variant. The basic procedure is as follows:

1. fit $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ to the real data, predict $\widehat{Y}_i$ and $\widehat{\varepsilon}_i$
2. in each bootstrap iteration, $b = 1, 2, \ldots, B$
   (a) calculate a new outcome by randomly drawing

$$Y^*_{ib} = \begin{cases} \widehat{Y}_i + \widehat{\varepsilon}_i & \text{with prob. } 0.5 \\ \widehat{Y}_i - \widehat{\varepsilon}_i & \text{with prob. } 0.5 \end{cases}$$

   (b) fit $Y^*_{ib} = \beta_0 + \beta_1 X_i + \varepsilon_{ib}$ to obtain $\widehat{\beta}_b$ or t_b.
3. Use the resulting collection of B bootstrap estimates $\{\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_B\}$ as an estimate of the sampling distribution we would have observed under repeated sampling from superpopulation.
4. Use simulated sampling distribution to calculate inferential quantities (e.g. confidence intervals, p-values, etc.)

The wild bootstrap is more versatile than the residual bootstrap has been shown to perform better in small sample sizes. However, it does assume that errors are mean independent, i.e. $\mathbb{E}[\varepsilon_i \mid X_i] = \mathbb{E}[\varepsilon_i]$.

Examples of when the bootstrap might perform better than classic approaches

The following tables are from Cameron, Gelbach, and Miller (2008), which I highly recommend. They show the simulated performance of the cluster variants of the different bootstrap procedures from this lab and compares them with traditional standard errors as well as those heteroscedasticity and cluster robust standard errors we talked about in 2000A. In general the bootstrap performs quite well. The wild boostratp-t in particular retains appropriate rejection rates even in data sets with very few clusters.

TABLE 3.—1,000 SIMULATIONS FROM DGP WITH GROUP-LEVEL RANDOM ERRORS AND HETEROSKEDASTICITY
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

Estimator #	Method	Number of Groups (G)					
		5	10	15	20	25	30
1	Assume i.i.d.	0.302	0.288	0.307	0.295	0.287	0.297
		(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
2	Moulton-type estimator	0.261	0.214	0.206	0.175	0.174	0.180
		(0.014)	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)
3	Cluster-robust	0.208	0.118	0.110	0.081	0.072	0.068
		(0.013)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
4	CR3 residual correction	0.138	0.092	0.086	0.070	0.062	0.062
		(0.011)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)
5	Pairs cluster bootstrap-se	0.174	0.111	0.109	0.085	0.074	0.070
		(0.012)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
6	Residual cluster bootstrap-se	0.181	0.169	0.183	0.157	0.149	0.163
		(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)
7	Wild cluster bootstrap-se	0.019	0.041	0.057	0.040	0.038	0.043
		(0.004)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
8	Pairs cluster bootstrap-BCA	0.183	0.103	0.099	0.082	0.070	0.064
		(0.012)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)
9	BDM bootstrap-t	0.181	0.108	0.110	0.090	0.070	0.068
		(0.012)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
10	Pairs cluster bootstrap-t	0.079	0.067	0.074	0.058	0.054	0.053
		(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
11	Pairs CR3 bootstrap-t	0.064	0.062	0.072	0.057	0.050	0.048
		(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
12	Residual cluster bootstrap-t	0.066	0.057	0.066	0.049	0.043	0.047
		(0.008)	(0.007)	(0.008)	(0.007)	(0.006)	(0.007)
13	Wild cluster bootstrap-t	0.053	0.056	0.058	0.048	0.041	0.044
		(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)
	T_distribution(G-2)	0.145	0.086	0.072	0.066	0.062	0.060

The first table shows the simulated rejection rates for tests of nominal size 0.05. The rows are different methods for calculating standard errors and test statistics and the columns represent simulations for datasets with different numbers of total clusters. If the procedure is performing correctly we expect that it should falsely reject the null when the null is in fact true just 5% of the time. What we observe is that for low numbers of clusters many techniques reject more than the nominal alpha level meaning that they are **anti-conservative** (BAD).

TABLE 4.—1,000 SIMULATIONS FROM DIFFERENT DGPS (SEE TEXT) AND $G = 10$ GROUPS
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

Estimator #	Method	Column Number	Main—from Table 2	Reject based on T (8 dof)	Cluster Size = 2	Cluster Size = 10	Cluster Size = 100	4 RHS Variables	Xs are Constant Within Group	Xs Are i.i.d.	Unbalanced Group Sizes (10, 50)
			1	2	3	4	5	6	7	8	9
1	Assume i.i.d.		0.491		0.106	0.268	0.679	0.687	0.770	0.054	0.524
			(0.016)		(0.010)	(0.014)	(0.015)	(0.015)	(0.013)	(0.007)	(0.016)
2	Moulton-type estimator		0.092	0.044	0.095	0.098	0.088	0.089	0.125	0.061	0.129
			(0.009)	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.008)	(0.011)
3	Cluster-robust		0.129	0.082	0.137	0.126	0.115	0.129	0.183	0.103	0.183
			(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.013)	(0.010)	(0.012)
4	CR3 residual correction		0.090	0.054	0.094	0.086	0.077	0.080	0.090	0.086	0.091
			(0.009)	(0.007)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
5	Pairs cluster bootstrap-se		0.120	0.071	0.100	0.114	0.120	0.128	0.063	0.122	0.138
			(0.010)	(0.008)	(0.009)	(0.010)	(0.010)	(0.010)	(0.008)	(0.010)	(0.011)
6	Residual cluster bootstrap-se		0.058	0.013	0.069	0.068	0.060	0.057	0.054	0.080	
			(0.007)	(0.004)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.009)	
7	Wild cluster bootstrap-se		0.028	0.006	0.048	0.044	0.032	0.030	0.036	0.053	0.019
			(0.005)	(0.002)	(0.007)	(0.006)	(0.006)	(0.005)	(0.006)	(0.007)	(0.004)
8	Pairs cluster bootstrap-BCA		0.111		0.125	0.112	0.109	0.112	0.100	0.134	0.140
			(0.010)		(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.011)	(0.011)
9	BDM bootstrap-t		0.119		0.086	0.115	0.112	0.119	0.121	0.097	0.128
			(0.010)		(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.011)
10	Pairs cluster bootstrap-t		0.096		0.085	0.083	0.086	0.090	0.066	0.079	0.120
			(0.009)		(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.010)
11	Pairs CR3 bootstrap-t		0.090		0.075	0.077	0.081	0.084	0.050	0.082	0.110
			(0.009)		(0.008)	(0.008)	(0.009)	(0.009)	(0.007)	(0.009)	(0.010)
12	Residual cluster bootstrap-t		0.055		0.052	0.056	0.050	0.043	0.043	0.065	
			(0.007)		(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.008)	
13	Wild cluster bootstrap-t		0.055		0.064	0.056	0.048	0.052	0.045	0.064	0.061
			(0.007)		(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
	T_distribution(8)		0.086								

The second table is like the first however now we add some extra real-world scenarios. In column 7 the intraclass correlation is is increased to 1 so that observations within clusters are perfectly correlated. In column 8 the authors consider what would happen if we falsely assume clustering/correlation if the Xs are in fact i.i.d. In column 9 the authors use unbalanced clusters.

Bootstrap myths and misconceptions

The bootstrap is wonderful, but it is not a panacea. Beware of the following myths:

1. **Myth:** The bootstrap is a cure for small sample sizes.

People often use the bootstrap for sample sizes smaller than we would need for standard asymptotic inference (e.g., confidence intervals based on the Central Limit Theorem). However, the bootstrap itself requires similar asymptotics to hold. Intuitively, the bootstrap uses the observed data to stand in for the true distribution that generated those data, and this only holds as n (the sample size in the original dataset) becomes large. Sometimes the bootstrap works actually does work better than standard methods for small sample sizes, but this isn't necessarily the case.

2. **Myth:** No matter what estimator you have in mind, you can use bootstrapping.

For example, say we want a confidence interval for the maximum value in the sample. Can we use the bootstrap? Unfortunately, no. The bootstrap works for estimators with certain "smoothness" properties. As a rule of thumb, "smooth" estimators are things like sums, ratios, etc., of well-known estimators like sample means, and – unlike the maximum value or median – they usually do not depend on specific data values. Bootstrapping can fail specatularly with non-smooth estimators.

Implementation in `R`

References

1. Efron,B. (1979). Bootstrap methods: Another look at jackknife. *Annals of Statistics*, 7:1-26.
2. Efron B, Tibshirani R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1(1):54-77.
3. Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics*, 14: 1431-1452.
4. Cameron, Gelbach & Miller (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics* 90(3):414-427.
5. For historical review and recent developments, see special edition of *Statistical Science*, Silver Anniversary of the Bootstrap, Vol. 18, No. 2, May, 2003