Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Introduction to Bootstrapping: Examples

Bootstrapping is a method for estimating sampling variation (read: standard errors), confidence intervals, and other properties of statistics. It relies on repeated re-sampling—*with replacement*—from the observed data.

Bootstrapping the sample mean

The sample mean $\bar{x} = \sum_i x_i/n$ is an easy point of entry for understanding bootstrapping. It is not the kind of statistic for which bootstrapping is useful, however, since theory tells us much of what we need to know about the sampling distribution of $\bar{x}$. But it is useful for illustrating the logic of the bootstrap.

Suppose we take a random sample of size n from a population with mean μ and variance σ^2. The Central Limit Theorem tells us that, over repeated samples, $\bar{x}$ will have a mean of μ and a variance of σ^2/n. If the underlying population of x is normal, we know $\bar{x}$ will also have a normal distribution. If x is not normal, the distribution of $\bar{x}$ will be *approximately* normal when n is large. Normality of $\bar{x}$ allows us to use z-scores $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ and know that $z \sim N(0,1)$. If σ^2 is estimated with s^2, $t = (\bar{x} - \mu)/(s/\sqrt{n})$ has a t distribution with $n-1$ degrees of freedom.

Example 1: draw a random sample of $n = 20$ from the $N(0,1)$ population and calculate $\bar{x}$, its standard error, and a 90% confidence interval. Without doing any simulation, we expect that over repeated samples $\bar{x}$ will have a mean of 0 and a standard deviation of $1/\sqrt{20} = .224$.

```
. drawnorm x1, n(20)

. mean x1, level(90)

Mean estimation                   Number of obs   =         20

--------------------------------------------------------------
             |       Mean   Std. Err.     [90% Conf. Interval]
-------------+------------------------------------------------
          x1 |  -.1093778   .2121558     -.4762233    .2574677
--------------------------------------------------------------
```

The standard error $se(\bar{x}) = s/\sqrt{n}$ above is an estimate of the standard deviation of the sampling distribution, and at 0.212 is not far from 0.224. This sample is just one draw out of many possible samples, however.

Example 2: use the same sample but obtain a *bootstrap* standard error for $\bar{x}$ using 19 bootstrap samples, denoted $se^*(\bar{x})$. $se^*(\bar{x})$ is calculated as the variance of the $\bar{x}^*$ across the 19 bootstrapped samples. $\bar{x}^*$ is notation for a mean calculated from a bootstrap sample. In the Stata `bootstrap` options I have chosen to save the results from the 19 bootstrap samples. (`_b` is the parameter estimate and `se` is the standard error).

```
. bootstrap _b _se , reps(19) saving(example2, replace): mean x1, level(90)

Bootstrap results                                 Number of obs     =         20
                                                  Replications      =         19

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
b            |
          x1 |  -.1093778   .1808754    -0.60   0.545    -.4068913    .1881357
-------------+----------------------------------------------------------------
se           |
          x1 |   .2121558   .0452157     4.69   0.000     .1377825     .286529
------------------------------------------------------------------------------

. use example2, clear
. summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       _b_x1 |         19   -.1343204    .1808754  -.3871673   .2913764
      _se_x1 |         19     .192187    .0452157   .1098926   .2739723
```

The observed $\bar{x}$ and $se(\bar{x})$ are the same as Example 1. The bootstrap standard error differs but is not far from the $se(\bar{x})$ of 0.212. Compare this value to the `summarize` results: 0.181 is the standard deviation of the 19 estimated means $\bar{x}^*$. The "normal-based 90% confidence interval" is calculated as $\bar{x} \pm 1.6449 * se^*(\bar{x})$. It uses the observed $\bar{x}$, the bootstrapped standard error, and assumes the distribution of $\bar{x}$ is normal. (This is where 1.6449 comes from).

Example 3: repeat Example 2 but with $B = 499$ bootstrap replications.

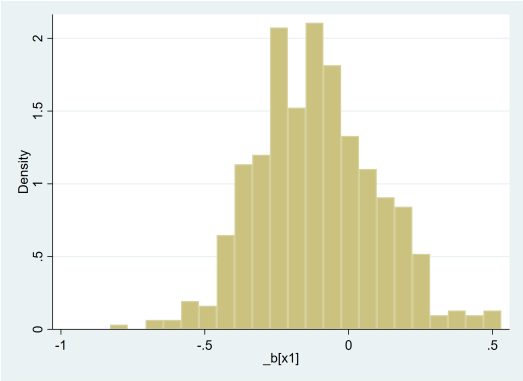
```
. bootstrap _b _se , reps(499) saving(example3, replace): mean x1, level(90)

Bootstrap results                                 Number of obs     =         20
                                                  Replications      =        499

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
b            |
          x1 |  -.1093778   .2112197    -0.52   0.605    -.4568033    .2380477
-------------+----------------------------------------------------------------
se           |
```

```
          x1 |   .2121558   .0420004     5.05   0.000     .1430713    .2812402
------------------------------------------------------------------------------
```

The results are shown above, and a histogram of the $\bar{x}^*$ from the 499 bootstrapped samples is below. The bootstrap standard error of 0.211 is remarkably close to the traditional $se(\bar{x})$ of 0.212 and the (known) population standard deviation of 0.224. I say "remarkably" because it is calculated from repeated samples from the same 20 observations.



An alternative to the reported normal-based confidence interval (which assumes a normal distribution) is to use percentiles from the distribution of $\bar{x}^*$. The interval is similar in this case (below), since the distribution of $\bar{x}^*$ here is somewhat normal. An advantage to this approach is that it uses the empirical distribution of the bootstrapped $\bar{x}^*$, which may be asymmetric. The Stata command `estat bootstrap, percentile` will produce the same results as those shown below (see Example 5).

```
. tabstat _b_x1, stat(p5 p95) save

    variable |        p5       p95
-------------+--------------------
       _b_x1 | -.4247129  .2362709
----------------------------------
```

Example 4: Examples 1-3 involved only one sample from the population. How does the bootstrap approach perform over repeated sampling? Repeat the following steps 1,000 times. (Based on Stine, 1989).

- Draw a sample of $n = 20$ from the $N(0, 1)$ population.
- Calculate $\bar{x}$, $se(\bar{x})$, the upper and lower bounds of the 90% CI, the width of the CI, and an indicator that $= 1$ if the CI contains the true population mean ($\mu = 0$).
- Obtain the bootstrap standard error $se^*(\bar{x})$, the upper and lower bounds of the 90% CI (both normal-based and percentile), the width of the CIs, and an indicator that $= 1$ if the CI contains the true population mean. Do this for $B = 19$ bootstrapped samples, $B = 99$, and $B = 499$.

The Stata code used to run this simulation is provided at the end of this document. The results are tabulated below, followed by a few observations.

	Mean over 1000 samples	SD over 1000 samples
$\bar{x}$	-0.008	0.227
$se(\bar{x})$	0.220	0.035
Width of classical t 90% CI	0.762	0.120
Coverage	0.887	
Bootstrap (19 replications)		
$se^*(\bar{x})$	0.212	0.050
Width of normal-based 90% CI	0.697	0.164
Coverage	0.850	
Width of percentile 90% CI	0.791	0.203
Coverage	0.868	
Bootstrap (99 replications)		
$se^*(\bar{x})$	0.213	0.036
Width of normal-based 90% CI	0.700	0.120
Coverage	0.854	
Width of percentile 90% CI	0.718	0.127
Coverage	0.858	
Bootstrap (499 replications)		
$se^*(\bar{x})$	0.215	0.035
Width of normal-based 90% CI	0.706	0.114
Coverage	0.859	
Width of percentile 90% CI	0.708	0.115
Coverage	0.858	

- The mean of $\bar{x}$ over the 1,000 samples is close to **0**. This is comforting, since $E(\bar{x}) = \mu$.
- The standard deviation of $\bar{x}$ over 1,000 samples is **0.227**. This is also comforting, since we know the (population) standard deviation of $\bar{x}$ is 0.224; see Example 1.
- On average, the width of the classical 90% CI is **0.762**. Given the (population) standard deviation of $\bar{x}$, we should expect a 90% CI to have a width of $2 * 1.729 * (1/\sqrt{20}) = 0.773$. The value of 1.729 comes from the t distribution with $\alpha/2 = 0.05$ and $df = 19$.

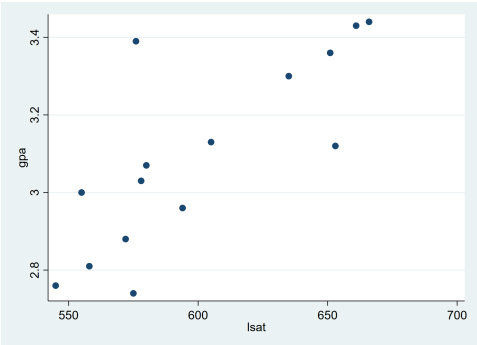
- The classical 90% CI included the true population mean in **88.7%** of the 1,000 samples. This is comforting, since a 90% CI is designed to include μ in 90% of random samples.
- Across repeated samples, the bootstrap standard error with 19 replications is (again, remarkably) close to $se(\bar{x})$ and the known standard deviation of 0.224. The number of replications has little effect on the average $se^*(\bar{x})$ here, but variability in this estimate falls with more replications.
- Similarly, the average width of the normal-based 90% CI for B =19, 99, and 499 is comparable to the expected width of 0.773. It is narrower than the classical t-based CI since it assumes normality. The number of replications decreases variability in the CI width across samples.
- The percentile-based CI tends to be wider than the normal-based CI. Since the real distribution of $\bar{x}$ with only $n = 20$ departs from normality, this makes sense and makes a case for the percentile-based interval.
- The bootstrap confidence intervals include the true population mean of 0 in **85-86%** of the 1,000 samples.

Bootstrapping the sample correlation coefficient

The sample correlation coefficient r_{xy} does not have as straightforward a sampling distribution as $\bar{x}$. The bootstrap can be used to calculate the standard error and confidence intervals for r_{xy}.

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Example 5: use Efron's sample dataset of LSAT scores and GPAs from 15 law schools to bootstrap the sampling distribution of r_{xy}. The scatter plot for these variables is shown below.



```
. bootstrap r(rho) , reps(1000) saving(example5, replace) level(90): corr gpa lsat

Bootstrap results                                 Number of obs     =         15
                                                  Replications      =      1,000

      command:  correlate gpa lsat
        _bs_1:  r(rho)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .7763744   .1348263     5.76   0.000     .5546048     .998144
------------------------------------------------------------------------------

. estat bootstrap, percentile

Bootstrap results                                 Number of obs     =         15
                                                  Replications      =       1000

      command:  correlate gpa lsat
        _bs_1:  r(rho)

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.  [90% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .77637442  -.0027943   .13482634    .5300619   .9495488   (P)
------------------------------------------------------------------------------
(P)    percentile confidence interval
```

The bootstrap standard error for r_{xy} is 0.134 and the 90% normal-based CI is (0.555, 0.998). The percentile confidence interval differs, at (0.540, 0.950). The histogram of the r^*_{xy}—correlation coefficients calculated from each bootstrap sample—is quite skewed, as shown below. This suggests the symmetric, normal-based CI would be inappropriate and makes a case for the percentile CI. In some settings the normal-based CI would produce an upper bound < 0 or > 1, which would not make sense for a correlation coefficient.