Analyzing Small Sample Experimental Data

Session 4: Simulations, resampling, and more small sample applications

Dominik Duell (University of Essex)

July 16, 2017

1. Recap simulations
2. Bootstrapping
3. Applications (see problem set session 4)

Recap simulations

Setting up simulations

To evaluate test and estimators

1. Make assumptions about the world:
   - population distribution
   - characteristics of relationship between variables of interest
2. Simulate S samples according to assumptions with N number of observations
3. Apply test/estimator
4. Evaluate test/estimation output

Evaluation criteria

Evaluation criteria

- **Robustness:**
  - Unbiasedness of estimator
    - Simulation average of $\hat{\theta}$, $\bar{\hat{\theta}}$, is the estimate of $E(\hat{\theta})$
    - Accounting for simulation error, $\bar{\hat{\theta}}$ should be close to assumed value θ
  - Standard errors
    - Simulation variance of $\hat{\theta}$, $s^2_{\hat{\theta}}$ is the estimate of $\sigma^2_{\hat{\theta}}$
    - Standard deviation of simulated $\hat{\theta}$, $s_{\hat{\theta}}$ is estimate of $\sigma_{\hat{\theta}}$
    - Accounting for simulation error, $s_{\hat{\theta}}$ should be close to simulated standard errors $se(\hat{\theta})$
  - Distributions
    - shape of distribution of statistic should be close to assumed distribution of the test
    - Distribution of p-value: if assumed distribution is correct distribution for test, p-value is uniformly distributed on (0,1)

Evaluation criteria

- **Small type I error:** low probability of falsely rejecting H_0
  - **Size of test**
  - Estimated by proportion of simulations that lead to rejection of H_0
- Not discussed before: **Coverage probability:** actual probability that the confidence interval contains the true value of the statistic
- **High statistical power**

Resampling

Basics

- Uniform random selection of observations
    - **Bootstrap:** with replacement
    - **Permutation:** without replacement

Small sample issues

- Resampling does not help with generating generalizable statements per se because it only uses (limited) data at hand
- but, generalizations based on assumptions about parameters that are not met are worse
- there we cannot even learn about the data we have

Bootstrapping

Basics

Basics

- We will look at the non-parametric bootstrap
- Statistical inference by resampling without any assumptions about underlying population
- Applied to standard errors, confidence bounds, test statistics but also to check asymptotic behavior of estimators
- Also implemented with most standard estimation commands in your preferred software (e.g., Stata: `vce(bootstrap)`, `bootstrap`-option)
- In `bootstrap`, each resampling draws the same total number of observations (as in the original sample) but some observations may show up multiple times and others not at all
- Good for alternative estimations methods and diagnostics

Ideal and bootstrap world

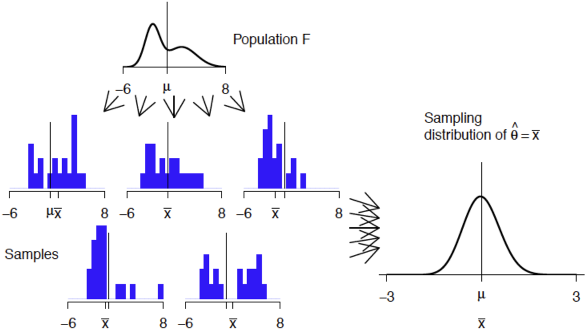


Figure 4: *Ideal world.* Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

Source: Hesterberg (2014), p.16

Ideal and bootstrap world

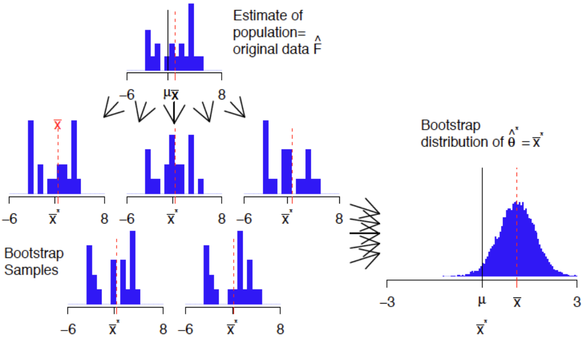


Figure 5: *Bootstrap world.* The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{x}$), not the parameter (μ).

Source: Hesterberg (2014), p.17

Why would we want to use the bootstrap?

- Let's treat our sample as the population and resample from it – after all, it is the only information we have about the population
- Very reasonable if sample is large and we just have problems to accurately estimate our quantity of interest

Why would we want to use the bootstrap

- What do we get out of it for small samples?
  - N^N bootstrap samples
  - **But,** if the sample is biased, resampling those biased observations makes them even more different from the population
  - **However**, already for $N = 10$, the number of distinct samples is 92,378, with $N = 20$ and 2000 repetitions, the probability that a bootstrap sample will be replicated is more than 0.95 (Hall 1992)
  - We get a sampling distribution of the sample statistic in question not an estimate of the population distribution!
  - Valuable if estimate of sampling distribution of the sample statistic hard to compute or inaccurate because of the small sample

Why would we want to use the bootstrap?

- What do we get out of it for small samples?
  - Should help us with improving asymptotic approximations in small samples – more accurate inferences
  - Mostly not helping in arriving at better estimates: e.g. all bootstrap samples will always be centered at the sample mean – estimates shape and spread of sampling distribution
  - In contrast to Monte Carlo, no assumptions about the distribution nor the true value of parameters

Procedure

- Say, we want to compute the standard error of test statistic $\hat{\theta}$
  1. Compute $\hat{\theta}$
  2. Take B samples from your sample with replacement
  3. Estimate of variance of $\hat{\theta}$:

$$\hat{var}_{boot}(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \overline{B^*})^2 \quad (1)$$

where $\hat{\theta}_i^*, \ldots, \hat{\theta}_B^*$ denote the test statistics and $\overline{\hat{\theta}^*} = 1/B\sum_{i=1}^{B}\hat{\theta}_i^*$

Procedure

- Bootstrap "standard error" is $se_{Boot}(\hat{\theta}) = \sqrt{\hat{var}_{boot}(\hat{\theta})}$
- Bootstrap bias estimate is $\overline{\hat{\theta}^*} - \hat{\theta}$

General pitfalls

- Are resampled observations independent? – use proper clustering and stratification of data when resampling
- bootstrap assumes that estimator is smooth ($\sqrt{N} - consistent$ and asymptotically normal)
- Don't forget to the set seed #
- Check default setting of number of repetitions of your preferred software when implementing the bootstrap – increase for results to be published and/or less well-behaved estimators
- How many repetitions? Efron/Tibshirani (1993) said 50 is mostly good enough ...
- Note, it is the more complicated estimators (more computationally intensive) that actually require more replications

All the different bootstraps

- Estimates:

```
bootstrap _b _se: reg var cat
```

- Other quantities of interest:

```
bootstrap diff = (r(mu_1) - r(mu_2)): ttest var, by(cat)
```

- Your own program:

```
bootstrap doodle = r(doodle): yourProgram
```

Example: bootstrap differences in means

```
ttest var, by(cat);

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       9    6.777778    1.769948    5.309844     2.69627    10.85929
       1 |       6        10.5    1.979057     4.84768    5.412672    15.58733
---------+--------------------------------------------------------------------
combined |      15    8.266667    1.367886    5.297798    5.332844    11.20049
---------+--------------------------------------------------------------------
    diff |           -3.722222    2.707443               -9.571297    2.126853
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.3748
Ho: diff = 0                                     degrees of freedom =       13

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0962         Pr(|T| > |t|) = 0.1924          Pr(T > t) = 0.9038
```

Example: bootstrap differences in means

```
. bootstrap diff = (r(mu_1) - r(mu_2)), seed(010101) nodots: ttest var, by(cat)

Warning:  Because ttest is not an estimation command or does not set e(sample), bootstrap has no
way to determine which observations are used in calculating the statistics and so assumes that all
observations are used.  This means that no observations will be excluded from the resampling
because of missing values or other reasons.
If the assumption is not true, press Break, save the data, and drop the observations that are to
be excluded.  Be sure that the dataset in memory contains only the relevant data.

Bootstrap results                               Number of obs      =        15
                                                Replications       =        50

      command:  ttest var, by(cat)
         diff:  r(mu_1) - r(mu_2)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |  -3.722222   2.648617    -1.41   0.160    -8.913416    1.468972
------------------------------------------------------------------------------
```

Evaluate and fix your bootstrap for small samples

- Number of replications
- Characteristics of statistic of interest
- Bias of the bootstrap
- Skewness of distribution
- Appropriateness of confidence intervals

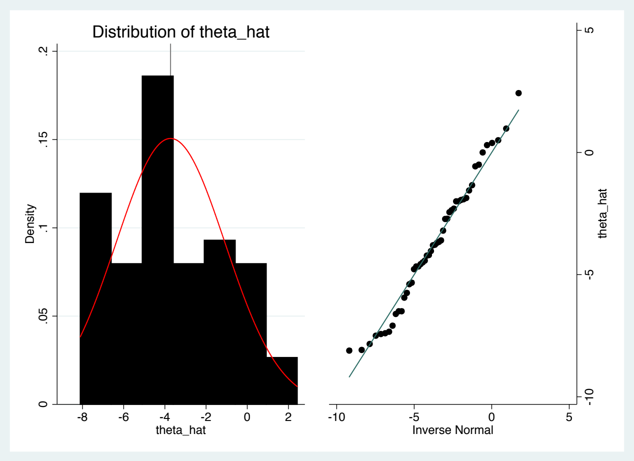
Number of replications

Example: bootstrap differences in means

```
. bootstrap diff = (r(mu_1) - r(mu_2)), seed(010101) nodots: ttest var, by(cat)

Bootstrap results                               Number of obs      =        15
                                                Replications       =        50

      command:  ttest var, by(cat)
         diff:  r(mu_1) - r(mu_2)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |  -3.722222   2.648617    -1.41   0.160    -8.913416    1.468972
------------------------------------------------------------------------------
```

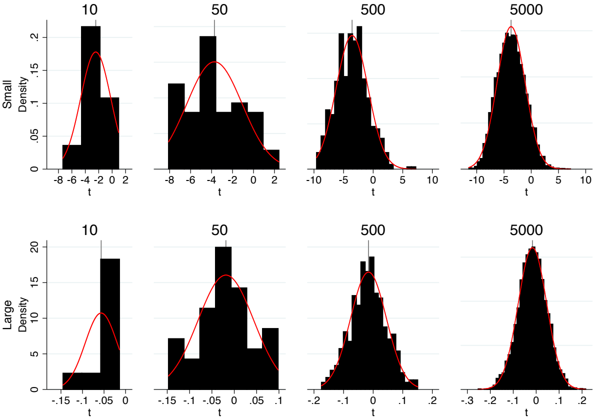
Number of replications may be too low.

Example: bootstrap differences in means



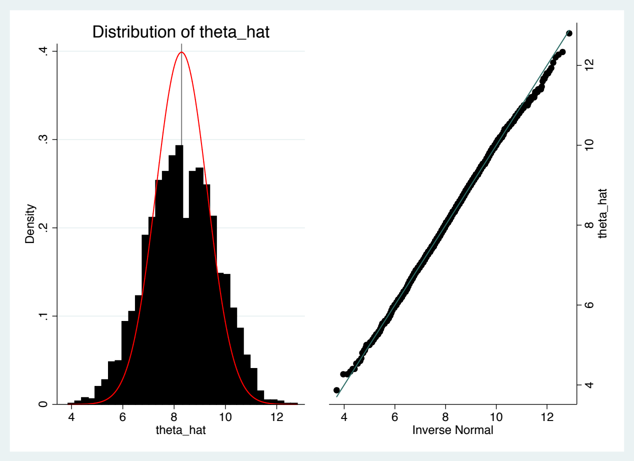
Example: bootstrap differences in means

Let's consider different B



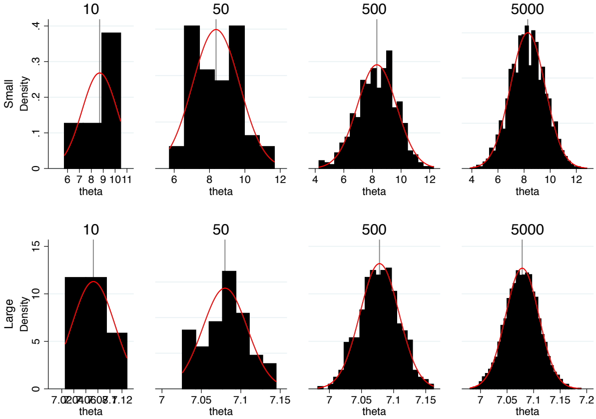
Better!

Example: bootstrap differences in means



Example: bootstrap differences in means

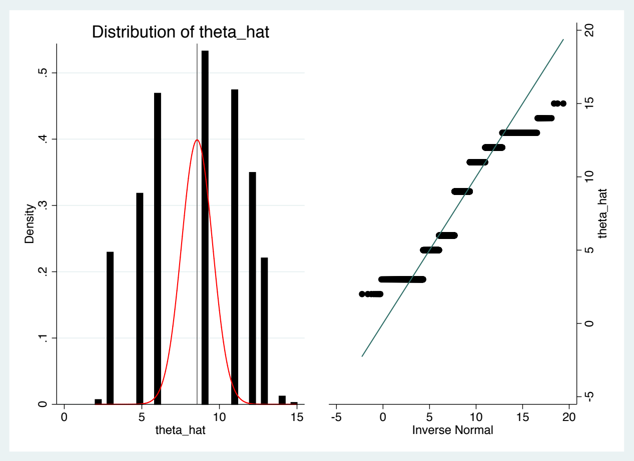
Also with different B



Much better!

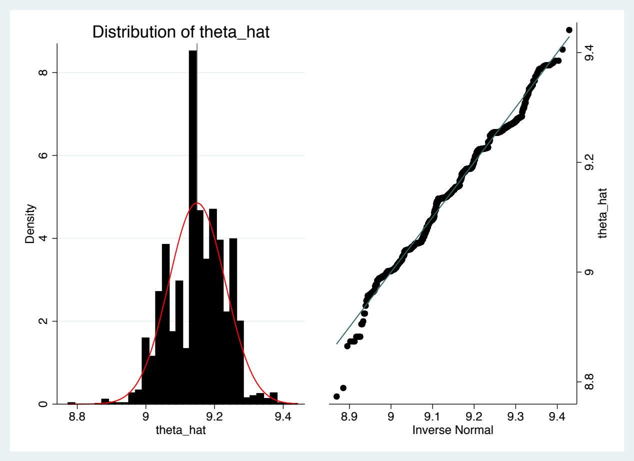
Bootstrap properties of statistics

Example: bootstrap the median



Often discontinuous empirical distribution function!

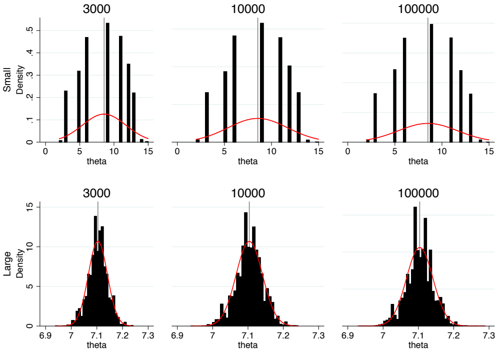
Example: bootstrap the median in a larger sample



Looks much better in a larger sample but is it the number of replications?

Example: bootstrap the median with more replications

Let's try different B



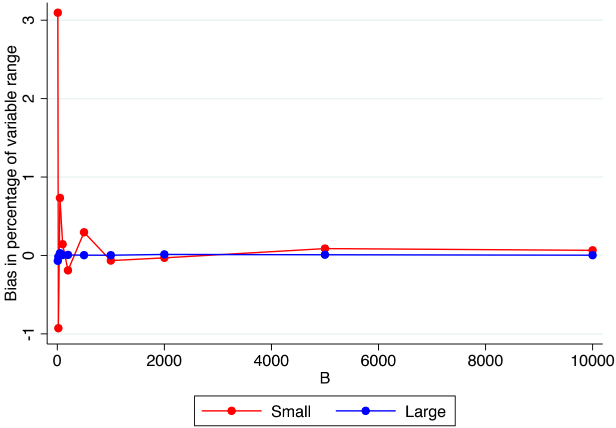
No improvement. Underlying variable not smooth enough, small sample provides too little variation.

Evaluating the bootstrap: Bias

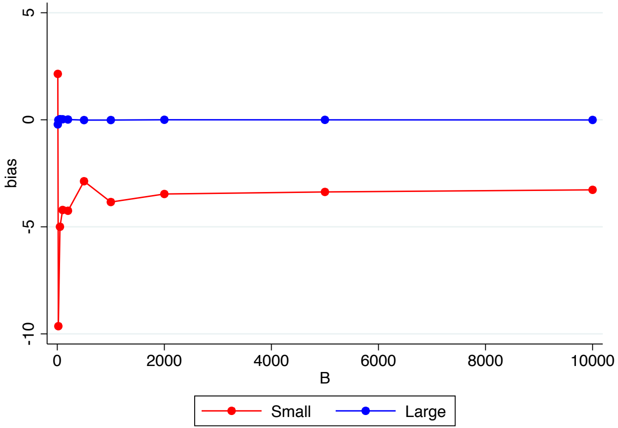
Basics

- Bootstrap bias: $\overline{\hat{\theta}^*} - \hat{\theta}$
- What produces bias:
  - Non-linear transformations of the statistic
  - Bootstrap procedure itself

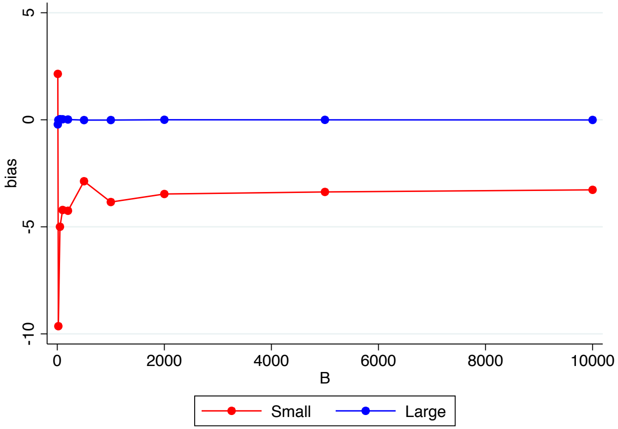
Example: bootstrap of the mean



Example: bootstrap of the median



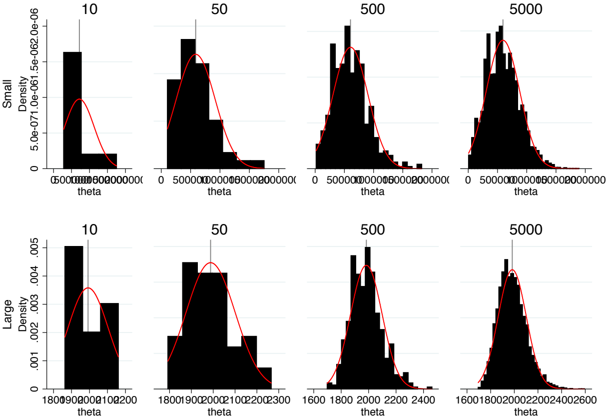
Example: bootstrap and bias correction



Bootstrapping from skewed distributions

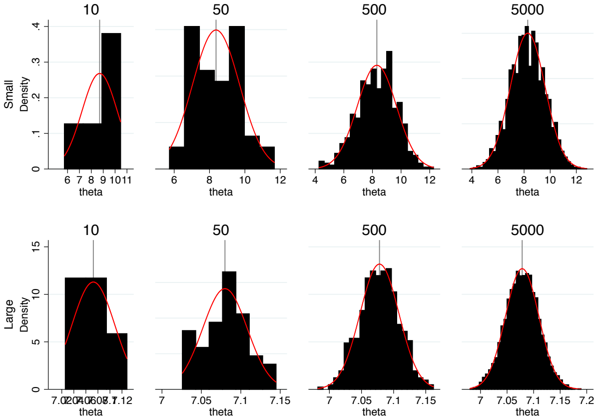
Example: bootstrap of mean

Which B does it take to account for the distortion?



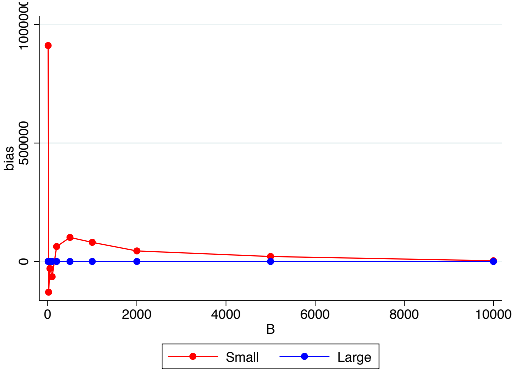
Example: bootstrap of in mean

Compare to less distorted raw data:



Example: bootstrap of in mean

What about bias with distorted raw data?



Oi, oi, oi . . .

Appropriate confidence bounds

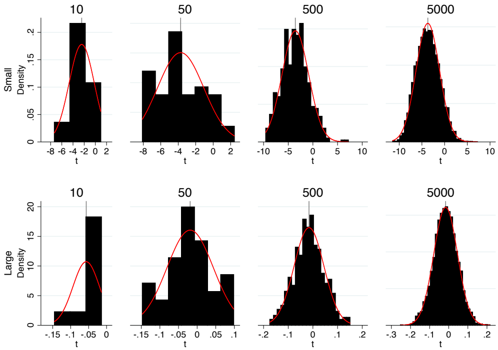
Example: bootstrap differences in means

```
. bootstrap diff = (r(mu_1) - r(mu_2)), seed(010101) nodots: ttest var, by(cat)

Bootstrap results                               Number of obs      =        15
                                                Replications       =        50

      command:  ttest var, by(cat)
         diff:  r(mu_1) - r(mu_2)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        diff |  -3.722222   2.648617    -1.41   0.160    -8.913416    1.468972
------------------------------------------------------------------------------
```

Bootstrap percentile confidence intervals probably too short for our small sample.

Example: bootstrap differences in means

Let's consider different B



Normal percentiles work for such well-behaved distribution function as produced by a bootstrap of means.

Example: bootstrap confidence intervals

Do we get more help from Stata? We do

- normal-based ci: $[\hat{\theta} - z_{1-\alpha/2}\hat{se}, \hat{\theta} + z_{1-\alpha/2}\hat{se}]$
- (empirical) percentile ci: $[\theta^*_{\alpha/2}, \theta_{1-\alpha/2}]$
- bias-corrected and accelerated method ci (bca): $[\theta^*_{p_1}, \theta^*_{p_2}]$

Example: bootstrap confidence intervals: bca

```
. bootstrap theta = r(mean), seed(010101) nodots reps(3000) bca: sum var;
Bootstrap results                               Number of obs      =        15
                                                Replications       =      3000

      command:  summarize var
        theta:  r(mean)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       theta |   8.266667   1.347879     6.13   0.000     5.624872    10.90846
------------------------------------------------------------------------------

. estat bootstrap, bca;
Bootstrap results                               Number of obs      =        15
                                                Replications       =      3000

      command:  summarize var
        theta:  r(mean)

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       theta |   8.2666667   .0236222   1.3478791          5.6       10.8 (BCa)
------------------------------------------------------------------------------
(BCa)  bias-corrected and accelerated confidence interval
```

Bootstrap confidence intervals: bca with skewed data

```
.       bootstrap theta = r(mean), seed(010101) nodots reps(3000) bca: sum varExp;
Bootstrap results                               Number of obs      =        15
                                                Replications       =      3000

      command:  summarize varExp
        theta:  r(mean)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       theta |   8.235248   1.788733     4.60   0.000     4.729396     11.7411
------------------------------------------------------------------------------

. estat bootstrap, bca;
Bootstrap results                               Number of obs      =        15
                                                Replications       =      3000

      command:  summarize varExp
        theta:  r(mean)

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       theta |   8.2352482   .0347594   1.7887332    5.023003   11.85562 (BCa)
------------------------------------------------------------------------------
(BCa)  bias-corrected and accelerated confidence interval
```

Bootstrap confidence intervals: bca

What is the bca-option doing?

- automatically adjusts for higher oder effects
- $[\theta^*_{p_1}, \theta^*_{p_2}]$ where
- $p_1 = \Phi\left\{z_0 + \frac{z_0+z_{1-\alpha/2}}{1-\alpha(z_0-z_{1-\alpha/2})}\right\}$ and
- $p_0 = \Phi\left\{z_0 + \frac{z_0+z_{1-\alpha/2}}{1-\alpha(z_0+z_{1-\alpha/2})}\right\}$
- where $z_0 = \Phi^{-1}\#(\hat{\theta}_i \leq \hat{\theta})/k$
- and $\alpha = \frac{\sum_{i=1}^{N}(\bar{\theta}_{(\cdot)}-\hat{\theta_{(i)}})^3}{6\sum_{i=1}^{N}(\bar{\theta}_{\cdot}-\hat{\theta}_{(i)})^{2^{3/2}}}$

Bootstrap confidence intervals

- How good are those confidence intervals in terms of accuracy (coverage probability)?
- Any other ideas for small samples with respect to confidence intervals?
  - Transformation of data to get a handle of the skewness or kurtosis? Which transformation?
  - Smoothed bootstrap:
    - bootstrap and then pertubate each estimate by a noise term
    - playing with noise term allows to simulate uncertainty we have about our small sample
  - Parametric bootstrap
    - specify a model of the world and resample from it
    - converges faster but potentially biased
    - again, interesting to built a counterfactual world

Summary of Small sample advice to assess bootstrap

- More bootstrap samples reduce variability of bootstrap distribution but does not fundamentally change it
- Know your statistic and whether those are sensitive to a few observations (see mean vs median example) Is the underlying data "too" discrete?
- Assess transformations, bias of the statistic, and skweness of the sampling distribution – what does it tell you about general performance of the bootstrap and number of replications?
- Think about adjustments to confidence interval
- Lock at smoothness or parametric bootstrap (look at Poi 2004)

References

Simulations

- Cameron and Trivedi (2009): Microeconomics using Stata, Stata Press, ch. 4 and 12
- Adkins and Gade: Monte Carlo Experiments using Stata: A Primer with Examples
- Davidson and MacKinnon (1997): Graphical Methods for Investigating the Size and Power of Hypothesis Tests

Bootstrap

- Cameron and Trivedi (2009): Microeconomics using Stata, Stata Press, ch. 13
- Poi (2004): From the help desk: Some bootstrapping techniques, The Stata Journal 4(3), pp.312-28
- Chernick and LaBudde (2011): An Introduction to Bootstrap Methods with Applications to R, Whiley
- Hall (1992): The Bootstrap and Edgeworth Expansion, Springer
- Hesterberg (2014): What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum