Review of Sampling Distributions, Bootstrap, and boot package

David M. Vock

PubH 7485/8485

Key Terminology

- Population: Individuals or units to which we would like to learn about
- Parameter (of interest)/Estimand: Summary measure of some characteristic (i.e., variable) of the population. Typically use Greek letters to denote a parameter
- Sample: Subset of the population of interest on which we collect data
- Estimate: Best guess of the parameter of interest using the sample data we collect
- Statistic/Estimator: A function or algorithm of the sample data to produce an estimate.

Sampling Distribution

Experimental set-up is usually as follows:

1) Identify population of interest
2) Take a sample from that population
3) Calculate a statistic
4) Use that statistic to infer something about population (more on this shortly)

If I repeated steps 2-4, I would obtain a different sample, calculate a different value for the statistic which may affect my inferences in step 4.

A Toy Example

- Suppose I am interested in the mean systolic blood pressure (SBP) of undergraduate students here at UMN (μ)
- Take a simple random sample of 10 students and get the following data: 105, 110, 112, 115, 118, 120, 125, 128, 130, 140 mmHg. $\hat{\mu} = \overline{x} = 120.3$ mmHg
- Throughout, we will use X to denote the SBP a randomly selected undergraduate student from UMN

A Toy Example Terminology

- Population: Undergraduate students at UMN
- Parameter (of interest)/Estimand: (population) Mean
- Sample: Simple random sample of 10 students
- Estimate: 120.3 mmHg
- Statistic/Estimator: sample average $\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$

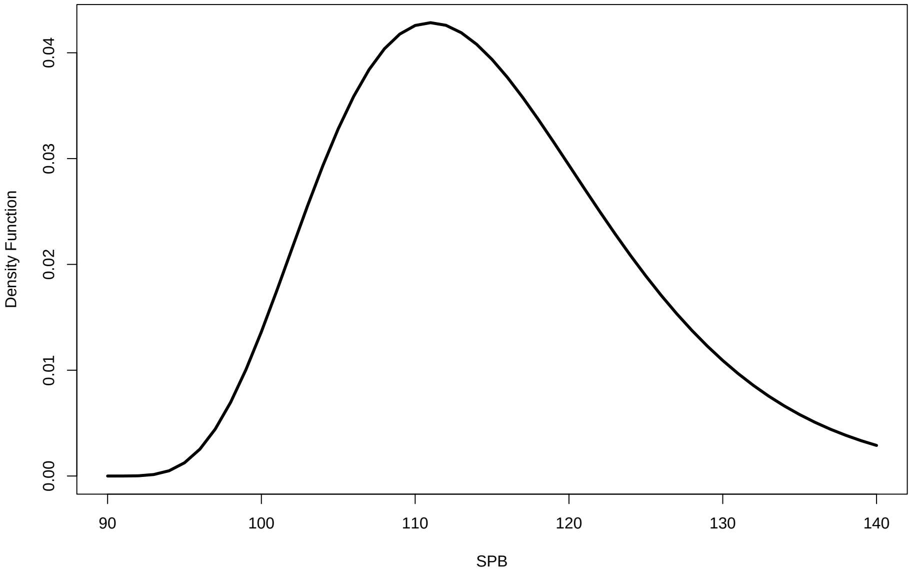
Sampling Distribution

- Distribution: A function which gives the probability of different values for a random variable
- Sampling distribution: "The probability distribution of a statistic/estimator is sometimes referred to as its sampling distribution. This emphasizes how the statistic varies in value across all samples that might be selected." (Devore and Beck)
- Sampling distribution is important because our uncertainty in our estimate can be characterized through the sampling distribution. In other words, standard errors, confidence intervals, and hypothesis testing are all derived from the sampling distribution
- The sampling distribution depends (at least exactly) on the distribution in the population, the sample size, and sampling mechanism
- If we knew population distribution then one way of approximating the sampling distribution is through simulation

Simulating the Sampling Distribution

- Suppose in the SBP example, the distribution of SBP among undergraduates at UMN follows a three-parameter gamma distribution with shape parameter 6.25, scale parameter 4, and threshold or shift paramater 90
- If we wanted to learn about the sampling distribution of an estimator $\hat{\mu}$ for a given sample size *n* and hypothetically had access to the entire population, we would take repeated samples of size *n* from the population, compute $\hat{\mu}$, and then assess the (properties of the) distribution of the distribution of $\hat{\mu}$

Distribution of the Population



Simulating the Sampling Distribution

{ In the following code, I have

- Sampled 10 subjects from this population.
- Computed the mean in the sample (i.e., my estimate of the mean in the population given this sample, $\hat{\mu}_1$. The subscript 1 indicates that this this is the estimated parameter from the first sample).

```
print("Sample of 10 Individuals")
```

```
## [1] "Sample of 10 Individuals"
```

```
samp.1 <- rgamma3(10, alpha, scale = 1/beta, thres= 90)
print(sort(round(samp.1)))
```

```
##  [1] 104 108 112 114 115 120 121 125 132 141
```

```
print("mu hat")
```

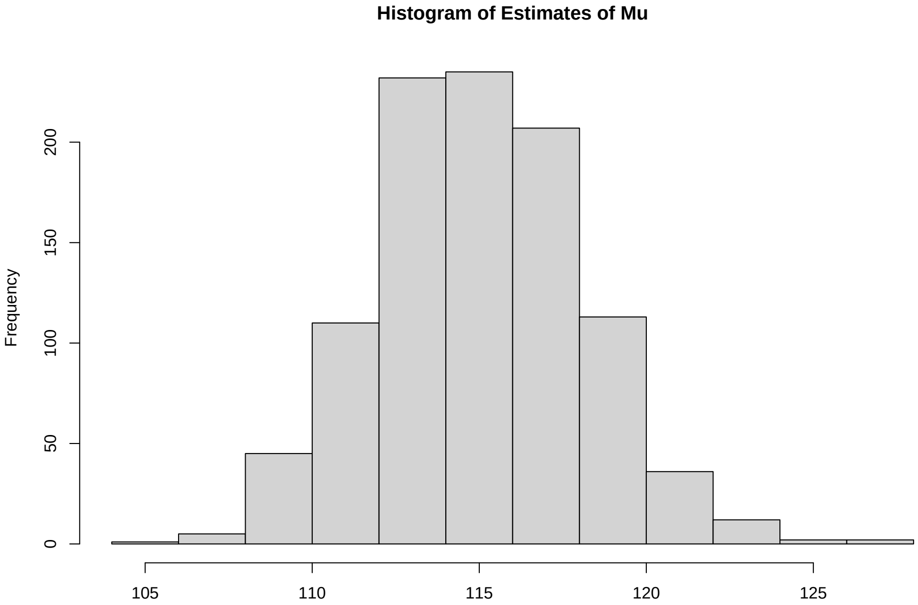
```
## [1] "mu hat"
```

```
print(round(mean(samp.1), 1))
```

Simulating the Sampling Distribution

- Note that I can repeat the above process B times
- Compute the parameter estimate in each sample. End up with B estimates of the parameter μ (i.e., $\hat{\mu}_1, \ldots, \hat{\mu}_B$).
- $\hat{\mu}_1, \ldots, \hat{\mu}_B$ are random samples from the sampling distribution and summary measures (e.g., standard deviation) of these B estimates are summary measures of the sampling distribution (plus Monte Carlo error)

Histogram of Sampling Distribution



Standard Error

- Standard error: standard deviation of the sampling distribution
- Because we have B samples from the sampling distribution, the standard deviation of $\hat{\mu}_1, \ldots, \hat{\mu}_B$ is the standard error (plus Monte Carlo error)

```
## [1] "Standard Deviation of Estimated Mu"
```

```
## [1] 3.1
```

Bootstrap

- Key problem with implementing the above in practice is that we do not know the distribution of X
- Re-sampling-based method invented by Bradley Efron in 1979
- Fundamental idea is that we can simulate the sampling distribution of an estimator using only one sample from the population

Resampling From the Sample

- Basic idea is that the original sample is a "pretty good" representation of the population
- Pretend that 105, 110, 112, 115, 118, 120, 125, 128, 130, 140 is the population. More formally we pretend that the population has pmf given by

{

x	105	110	112	115	118	120	125	128	130	140
$\hat{p}(x)$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

}

Empirical Distribution

- Estimating $P(X = x) = p(x)$ by $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = x$ is known as the empirical distribution
- Empirical distribution is a consistent, non-parametric estimator of the population pmf/pdf

Resampling From the Sample

In the following code, I have

- Sampled 10 subjects from this (estimated) population distribution. This sample is called a bootstrap sample (to distinguish from the original sample). Each observation in the bootstrap sample has a $1/10$ chance of being 105, 110, 112, 115, 118, 120, 125, 128, 130, or 140. This is equivalent to taking a sample of size 10 with replacement from original sample.
- Computed the mean in the sample (i.e., my estimate of the mean in the population given this sample, $\hat{\mu}_1$. The subscript 1 indicates that this this is the estimated parameter from the first sample).

Approximately Simulating the Sampling Distribution

- Example of taking 10 bootstrap samples and computing $\hat{\mu}$ using these bootstrap samples on the next slide

```
##      V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 mu_hat
## 1  120 118 130 125 118 125 125 130 140 105  123.6
## 2  128 140 112 105 110 128 110 110 118 120  118.1
## 3  125 120 128 125 128 140 115 120 130 120  125.1
## 4  120 125 105 130 118 118 115 115 115 112  117.3
## 5  105 140 112 110 128 115 140 115 112 128  120.5
## 6  118 140 130 130 120 105 112 125 112 128  122.0
## 7  125 115 118 140 125 115 130 140 120 118  124.6
## 8  112 120 110 110 110 140 120 115 140 110  118.7
## 9  112 125 118 118 128 118 130 128 128 120  122.5
## 10 125 140 118 118 130 118 105 120 118 140  123.2
```

Approximately Simulating the Sampling Distribution

- Note that I can repeat the above process B times
- Compute the parameter estimate in each sample. End up with B estimates of the parameter μ (i.e., $\hat{\mu}_1, \ldots, \hat{\mu}_B$).
- $\hat{\mu}_1, \ldots, \hat{\mu}_B$ are APPROXIMATELY random samples from the sampling distribution. The approximation is because I had to estimate the population distribution instead of using the TRUE distribution
- Summary measures (e.g., standard deviation) of these B estimates are summary measures of the sampling distribution (plus Monte Carlo error AND error from estimating the population distribution)

General Idea: Standard Error

- Idea is that the standard deviation of the bootstrap estimates of $\hat{\mu}$ should be close to the standard deviation of the sampling distribution of $\hat{\mu}$

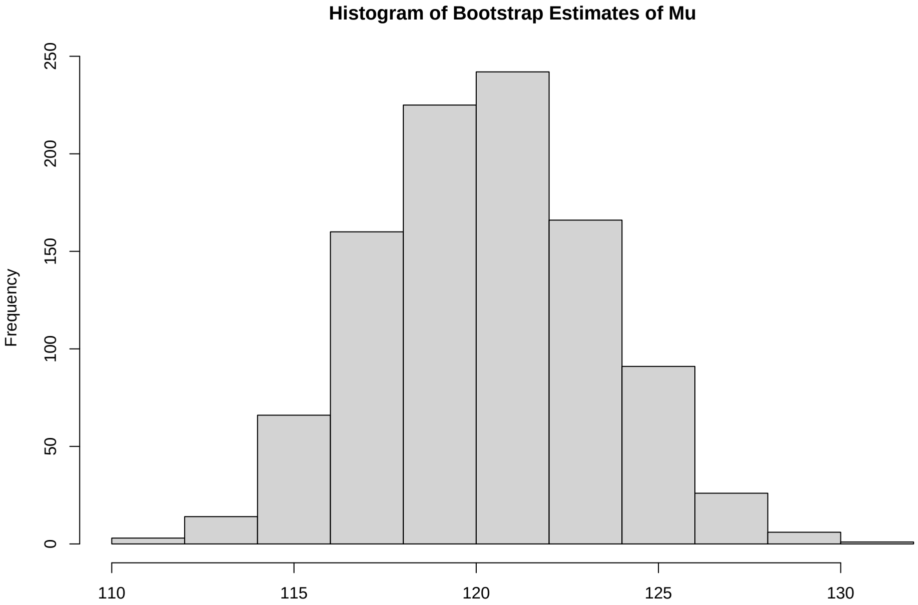
General Idea: Confidence Intervals

- Percentile Bootstrap CI: 2.5th and 97.5th percentile of the bootstrap estimates of $\hat{\mu}$ should form a valid 95% confidence interval for μ
- Normal-based Bootstrap CI: Estimate ± critical value × bootstrap se
- Many other flavors of bootstrap CIs: basic bootstrap, studentized bootstrap, bias corrected (BC), bias corrected and accelerated (BC_a),
- Percentile Bootstrap is by far the most intuitive and widely used

General Idea: Hyptothesis Testing

- Much harder to do directly because resampling with replacement "puts no restriction on the data and thus does not generate an approximation to the null distribution of the test statistic"
- Could generate a test statistic as (estimate-null value)/(bootstrap se)

Bootstrap Sampling Distribution



Bootstrap versus Typical SE

Bootstrap SE

```
## [1] 3.1
```

Typical SE $= \frac{s}{\sqrt{n}}$ where s is the sample standard deviation

```
## [1] 3.3
```

Bootstrap versus Typical CI

Bootstrap CI

```
## [1] 114.3
```

```
## [1] 126.3
```

Typical CI $= \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

```
## [1] 112.7365
```

```
## [1] 127.8635
```

Bootstrap versus Typical Test Statistic

Let's consider a null hypothesis of $\mu = 120$

Bootstrap Test Statistic

```
## [1] 0.096
```

Typical Test Statistic

```
## [1] 0.09
```

Why Use Bootstrap

- The toy example and estimand/estimators used throughout there were "nice" (approximate) formula for the standard error, confidence interval, test statistics, etc.
- BUT lots of estimators that we will consider will not be "nice" and will be complicated functions of other estimated parameters. We could derive the standard errors using the multivariate Delta theorem and M-estimation theory but that is tedious!!!

A Word of Caution

- A LOT of researchers think that the bootstrap is magical or more precisely does not require making any distributional assumptions OR relying on the sample size to be large (asymptotic approach).
- We did not make any distributional assumptions.
- BUT the bootstrap is premised on the assumption that the sample is a good approximation of the population distribution (i.e., the empirical distribution is good). That requires a “large” sample to be true.
- The bootstrap is an asymptotic procedure!

A Note About Distributional Assumptions

- Wald-type test-statistic: (parameter estimate - null value)/(standard error)
- As long as the standard error is "correct" (i.e., consistent), then under suitable regularity conditions the test-statistic follows a standard normal distribution under the null, at least asymptotically
- Note that we usually do not have have to make distributional assumptions so long as the standard error estimator is consistent (think t-test)
- For linear regression, in order for the standard error estimates of regression parameters which come from software to be consistent (1) the standard deviation of the residual error must not depend on the covariates (homoskedastic) and (2) model must be correctly specified
- Sandwich or robust standard errors do not make this assumption; nonparametric bootstrap is a consistent estimator of the robust standard errors

Computational Challenge

- A key challenge of the bootstrap is computational bandwidth
- Parallel computing can help reduce the computational burden
- boot package in R allows fast implementation of parallel computing

Example IHDP Data

- The Infant Health and Development Program (IHDP) targeted low-birth-weight, premature infants.
- The study, conducted from 1985-1988, was a randomized trial. The treatment group received intensive high-quality child care and home visits.
- We are interested in studying only those from the treatment group who were sufficiently compliant with the intervention. This is, of course, a nonrandom subset of the treatment group in the population so there are important prognostic differences among the control group and this subset of treatment.
- Outcome is child's IQ at 36 months
- Data source: Hill JL. "Bayesian Nonparametric Modeling for Causal Inference" Journal of Computational and Graphical Statistics 20(1):217-240 DOI:10.1198/jcgs.2010.08162

Summarize Baseline Covariates

```
##                                     Stratified by treat
##                                      0                1                SMD
##   n                                      561              67
##   IQ at 36 mo. (mean (SD))             84.55 (19.94)   101.37 (15.94)   0.932
##   Treatment group (mean (SD))           0.00 (0.00)      1.00 (0.00)      Inf
##   Birth weight (mean (SD))           1789.26 (465.09) 1774.15 (449.19)  0.033
##   Mother's age (mean (SD))             25.01 (6.13)     25.94 (5.80)    0.155
##   Neo-natal health index (mean (SD))   99.61 (15.68)   101.09 (15.96)   0.094
##   Birth order (mean (SD))               1.96 (1.17)      1.82 (1.09)    0.121
##   Parity (mean (SD))                    1.91 (1.14)      1.79 (1.11)    0.102
##   Premature births (mean (SD))          0.25 (0.61)      0.18 (0.49)    0.131
##   Cigarettes/day (mean (SD))            3.89 (7.07)      4.85 (8.36)    0.124
##   Drinks/week (mean (SD))               0.35 (1.43)      0.82 (3.60)    0.172
##   Mother's PPVT score (mean (SD))      81.00 (21.05)    82.93 (18.85)   0.097
##   Female (mean (SD))                    0.51 (0.50)      0.51 (0.50)    0.001
##   Twins (%)                                                             0.285
##      0                                   497 (88.6)        54 (80.6)
##      1                                    35 ( 6.2)        10 (14.9)
##      2                                    29 ( 5.2)         3 ( 4.5)
##   Marital status (%)                                                    0.115
##      1                                   279 (49.7)        35 (52.2)
##      2                                   237 (42.2)        26 (38.8)
##      3                                    43 ( 7.7)         6 ( 9.0)
##      4                                     2 ( 0.4)         0 ( 0.0)
##   Living status (%)                                                     0.074
##      1                                   335 (59.7)        41 (61.2)
##      2                                   164 (29.2)        18 (26.9)
##      3                                     5 ( 0.9)         1 ( 1.5)
##      4                                    57 (10.2)         7 (10.4)
##   Primary language (%)                                                  0.276
##      1                                   529 (94.3)        65 (97.0)
##      2                                    18 ( 3.2)         0 ( 0.0)
##      3                                     2 ( 0.4)         0 ( 0.0)
```

Adjusted Treatment Effect

- Some imbalance among groups for key covariates
- Could estimate the ATE (causal treatment effect) by fitting a model which adjusts for the other potential confounders. We will only include main effects and linear terms. (NB: we will discuss the merits of this later)
- We will collapse values of some categorical covariates with small frequencies

Adjusted Treatment Effect

```
summary(m1)
```

```
##
## Call:
## lm(formula = iqsb.36 ~ . - mlt.birtF - b.marryF - languageF,
##     data = ihdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.297  -8.931   0.322   8.991  52.645
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.505028   7.218858   7.827 2.26e-14 ***
## treat           13.658437   1.970802   6.930 1.08e-11 ***
## bw               0.002970   0.001314   2.260  0.02417 *
## momage           0.039073   0.137615   0.284  0.77656
## nnhealth         0.109941   0.039512   2.782  0.00556 **
## birth.o         -0.366990   1.186503  -0.309  0.75720
## parity          -1.977900   1.183381  -1.671  0.09516 .
## moreprem        -1.349465   1.120616  -1.204  0.22898
## cigs            -0.002597   0.091897  -0.028  0.97746
## alcohol         -0.079484   0.357161  -0.223  0.82397
## ppvt.imp         0.119932   0.044709   2.683  0.00751 **
## female           1.023101   1.225492   0.835  0.40413
## livwhoF2         0.750597   2.254144   0.333  0.73926
## livwhoF3         1.098846   6.557446   0.168  0.86698
## livwhoF4        -2.429416   2.700213  -0.900  0.36863
## whenprenF1       1.579942   3.309679   0.477  0.63327
## whenprenF2       0.422621   3.332991   0.127  0.89914
## whenprenF3       0.116226   4.223073   0.028  0.97805
## momed4F2        -0.089615   1.650703  -0.054  0.95672
## momed4F3         5.389108   2.001691   2.692  0.00729 **
## momed4F4        11.421857   2.819704   4.051 5.78e-05 ***
```

Standard Error of the ATE

- IF we believe that the homoskedastic assumption (and independence), then the standard error that is part of the software output is valid
- BUT let's estiamte a robust standard error using the bootstrap
- Demonstrate how to use boot package

First create function which returns ATE

```
ate.stat <- function(data, indices){
  data.boot <- data[indices,]

  m1.boot <- lm(iqsb.36 ~ .
    - mlt.birtF - b.marryF - languageF,
    data = data.boot)
  return(coef(m1.boot)[2])
}
```

Call boot package

```
set.seed(1101985) # bootstrapping is random - don't forget to set a seed!
start <- proc.time()
results <- boot(data=ihdp, statistic=ate.stat, R=1000)
results; head(results$t)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = ihdp, statistic = ate.stat, R = 1000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 13.65844 -0.01656407    1.748839
```

```
##          [,1]
## [1,] 10.53010
## [2,] 15.33117
## [3,] 13.82295
## [4,] 16.61262
## [5,] 14.99163
## [6,] 13.54609
```

```
proc.time() - start
```

```
##    user  system elapsed
##   5.182   0.135   5.336
```

Parallelizing with boot

```r
set.seed(1101985) # bootstrapping is random - don't forget to set a seed!
start <- proc.time()
boot(data=ihdp, statistic=ate.stat, R=1000, parallel="multicore", ncpus=2) # on Mac OS
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = ihdp, statistic = ate.stat, R = 1000, parallel = "multicore",
##     ncpus = 2)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 13.65844 -0.01656407    1.748839
```

```r
# boot(data=ihdp, statistic=ate.stat, R=1000, parallel="snow", ncpus=2) # on Windows OS
proc.time() - start
```

```
##    user  system elapsed
##   7.226   0.514   3.927
```