Sampling Variability and the Bootstrap

Stats200, Winter 2018
Recently, we saw that Maximum Likelihood Estimators are asymptotically normal and asymptotically efficient. (Given a distributional form and enough data, no unbiased estimator can estimate $\theta_0$ with fewer observations.)
Recently, we saw that Maximum Likelihood Estimators are asymptotically normal and asymptotically efficient. (Given a distributional form and enough data, no unbiased estimator can estimate $\theta_0$ with fewer observations.)

What about finite (and even small) samples? What if the distribution that generated the data is unknown?
Recently, we saw that Maximum Likelihood Estimators are asymptotically normal and asymptotically efficient. (Given a distributional form and enough data, no unbiased estimator can estimate $\theta_0$ with fewer observations.)

What about finite (and even small) samples? What if the distribution that generated the data is unknown?

The bootstrap, a versatile point and interval estimator, resamples so researchers can learn about small samples with few assumptions.
Origin

- The phrase “pull yourself up by your bootstraps” means “make something of yourself” or “you can succeed even if you have little to start with.”
- To “boot” a computer is another play on bootstrap that originally meant “load a little software at a time.”
- Bootstrapping was introduced in 1979 by Bradley Efron (Stanford) by way of comparison to another resampling technique, the jackknife. That paper currently has over 16,000 citations on Google Scholar.
Example 1: Stanford undergraduate population

We know that the proportion of “Black or African American” at Stanford is 0.064. Suppose we were to take 10,000 samples of size 50 and calculate the sample means in each of them...
How much we are off on average?

If we could average all the 10,000 samples, we would be fine

```r
mean(sample_p)
[1] 0.063982
```

The “problem” is that we get to see only one sample. The samples have a sizeable standard deviation:

```r
sd(sample_p)
[1] 0.03439392
```
How much we are off on average?

If we could average all the 10,000 samples, we would be fine

```r
mean(sample_p)
[1] 0.063982
```

The “problem” is that we get to see only one sample. The samples have a sizeable standard deviation:

```r
sd(sample_p)
[1] 0.03439392
```

Also, there's too much skew for the Central Limit Theorem...

```r
library(moments)
skewness(sample_p)
[1] 0.487662
```
Intuitions behind bootstrapping

- As a “best guess” for the population we can use the sample we actually observed
- Repeated sampling with replacement for the population defined by the sample
Bootstrapping with R

```r
N <- 50  # sample size
B <- 10000  # bootstrap replicates

a_sample <- sample(UGRace, N, replace = FALSE)
p <- function(x){
    mean(x == "Black or African American, non-Hispanic")
}

b_sample_p <- matrix(nrow = B)
for(i in 1:B){
    b_sample_p[i] <- p(sample(a_sample,N,replace=TRUE))
}
```
Bootstrapping the Distribution

10,000 samples from population

10,000 re-samples from sample

null device

1
Comparing Estimates

\[
\text{sd(sample}_p) \quad \# \; 10,000 \; \text{samples from full population} \\
[1] \; 0.03439392 \\
\text{sd(b_sample}_p) \quad \# \; 1 \; \text{sample, 10,000 bootstrap replicates} \\
[1] \; 0.04585346 \\
\text{skewness(sample}_p) \\
[1] \; 0.487662 \\
\text{skewness(b_sample}_p) \\
[1] \; 0.2986171
\]
Bootstrap approach to Sampling Variance

Sample: \((X_1, \ldots, X_n)\);

Sample summary: \(S(X_1, \ldots, X_n)\) which we use to estimate the corresponding value in the population \(S_{pop}\).

- Resample with replacement to obtain bootstrap samples \((X^b_1, \ldots, X^b_n)\), \(b = 1, \ldots, B\)
- Calculate the summaries of the bootstrap samples \(S_b = S(X^b_1, \ldots, X^b_n)\), \(b = 1, \ldots, B\)
- The variance of the bootstrap samples is an estimate of the variance of \(S(X_1, \ldots, X_n)\) across all possible samples

\[
\text{Standard Error}(S(X_1, \ldots, X_n)) \approx \sqrt{\sum_{b=1}^{B} \frac{(S_b - \bar{S})^2}{B}}
\]
Interval estimation

Rather than just a point estimate, report a range of possible values that you expect to cover the true value (95% of the time, given repeated experiments).

Normal \( \left( \bar{X} - 2SD(X_1, \ldots, X_n)/\sqrt{n}, \ ar{X} + 2SD(X_1, \ldots, X_n)/\sqrt{n} \right) \)

\[
xbar_a <- \text{mean}(p(a\_sample))
\]
\[
sd_a <- \text{sd}(\text{grepl}("Black", a\_sample)) \ # N = 50
\]
\[
c(xbar_a - 2*sd_a, xbar_a + 2*sd_a)/(N^.5)
\]
[1] -0.07587559 0.10981672

Bootstrap \( (\text{Quantile}_{\text{BootSample}}(0.025), \text{Quantile}_{\text{BootSample}}(0.975)) \)

\[
\text{quantile}(b\_sample\_p, c(0.025, .975))
\]
2.5% 97.5%
0.04 0.22
Example 2: DNA correlates of Lefthandedness

Claridge data

Variability of sample correlation

Propensity to use left hand

DNA variant

n

1

2

3

4

5

Correlation in Bootstrap sample

count

−0.5

0.0

0.5

0

500

1000

1500
Bootstrapped Statistics

Sample Correlation Standard Error (Bootstrap)
0.509 0.204

And two confidence intervals...

2.5% 97.5%
-0.030 0.754
5% 95%
0.066 0.724
Maximum Likelihood is one of the most popular methods of estimation for regression, particularly when the outcome of interest is not normally distributed. MLE requires that researchers specify a function that links observable outcomes \( y \) to probability distributions, as well as explanatory variables \( x_1, x_2, ..., x_p \).

MLE allows you to estimate many quantities of interest such as slopes i.e. partial derivatives \( \frac{\delta y}{\delta x_1}, \frac{\delta y}{\delta x_2}, ..., \frac{\delta y}{\delta x_p} \).
Probit Regression

For binary outcomes, probit is a popular choice (which is similar to logistic regression). The model assumes that each Bernoulli outcome has a slightly different chance of occurring. That latent (unobservable) probability can be represented in terms of a dichotomized normal distribution.

\[ P(Y = 1|X) = \Phi(X'\beta) \]

where \( \beta = \frac{\delta y}{\delta x_1}, \frac{\delta y}{\delta x_2}, \ldots, \frac{\delta y}{\delta x_p} \), \( X \) is corresponding data, and \( \Phi \) is the cumulative normal distribution. The model also assumes an i.i.d. error term \( \epsilon \sim N(0, 1) \).
The Log Likelihood

Substituting into a Bernoulli with \( p = P(Y = 1|X) \) yields:

\[
\ln L(\beta) = \sum_{i}^{N} \{ y_i \ln(\Phi(x_i'\beta)) + (1 - y_i) \ln(1 - \Phi(x_i'\beta)) \}
\]

Using the multivariate generalizations of the techniques we used in the univariate case, these MLE coefficient estimates can be shown to achieve the Cramer-Rao lower bound asymptotically.
Example 3: Satisfaction with Direction of the Country (Survey Data)

In January 2016, Pew Research asked: “All in all, are you satisfied or dissatisfied with the way things are going in this country today?”

FALSE TRUE
753 259

In general, political scientists consider questions like these to be predictive of elections because they reveal whether the public is satisfied with the incumbent party.
MLE in R

```r
est <- glm(satisfied ~ age + sex + party + educ + racecmb,
           data = pew,
           family = binomial(link = "probit"))
```

|               | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | -0.663   | 0.354      | -1.876  | 0.061    |
| Age           | -0.007   | 0.003      | -2.555  | 0.011    |
| Female        | -0.176   | 0.095      | -1.854  | 0.064    |
| Democrat      | 1.379    | 0.147      | 9.385   | 0.000    |
Bootstrapping Regression’s Uncertainty Estimates

Suppose we are not sure if we have enough data for MLE and so are not sure whether to trust the confidence intervals around the slope.

```r
library(boot)

slope <- function(data, indices, which_slope) {
  d <- data[indices, ]  # indices allows boot to resample
  est <- glm(satisfied ~ age + sex + party + educ + racecmb,
             data = d, family = binomial(link = "probit"))
  return(coef(est)[which_slope])
}

# bootstrapping with 500 replications
results <- boot(data=pew, statistic=slope,
                R=1000, which_slope=3)
```
Results

Bootstrap Replicates: Are Women Less Satisfied with Direction of Country?

Data: Pew Research January 2016 (own calculations)

MLE Slope Estimate across 1000 Simulations

Frequency

−0.5 −0.4 −0.3 −0.2 −0.1 0.0 0.1

0 50 150

2.5% 97.5%

-0.377502827 0.006578967

MLE Slope Estimate across 1000 Simulations
Data: Pew Research January 2016 (own calculations)
Conclusion

- Bootstrapping is an effective method for estimating confidence intervals, particularly for finite data or unknown data generating processes (distributions).
- Whether bootstrapping produces intervals similar to those from parametric approaches depends on the data at hand.
- In R, the library `boot` provides convenience functions to minimize the amount of coding and help manage replicates.