STAT 111, Week 13 - Bootstrap Confidence Intervals (Final Presentation)

Leia Donaway — Swarthmore College

April 24, 2025

I based my presentation off of an Advanced Topic Outline that Professor Everson posted for our use on Moodle. It can be found, edited for length, at the end of this document, along with R code and outputs that I will reference below. This presentation makes use of discussions of Bootstrap Confidence Intervals found in Rice 8.5, Rice 10.4.6, and Hesterberg 5.4-5.5, as well as resources made available on Moodle, which include Phil's grant data.

The name "bootstrap"
From the phrase "pull yourself up by your bootstraps," which is used to encourage self-reliance in a tough situation. Logically speaking, this is impossible, so the name as it is used in this technique implies that we are getting something from nothing because the bootstrap algorithm makes use of the variability in your sample to estimate the sampling distribution of a sample statistic.

A) We will start by discussing the parametric bootstrap, which makes use of simulation to approximate the sampling distributions of parameters from a known data distribution (Read more in Rice Chapter 8).

We are going to talk about the procedure in the context of finding interval estimates for the parameters α and λ for random variables

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \lambda)$$

We can get crude "method of moments" estimates for α and λ, the shape and rate parameters. We know the mean $\text{E}[X] = \frac{\alpha}{\lambda}$ and variance $\text{Var}(X) = \frac{\alpha}{\lambda^2}$ of a Gamma(α, λ), so we can solve for λ from the mean and plug this into our variance formula to find both estimates.

$$\bar{x} = \frac{\alpha}{\lambda} \implies \lambda = \frac{\alpha}{\bar{x}}$$

$$s_x^2 = \frac{\alpha}{\lambda^2} = \frac{\alpha}{(\frac{\alpha}{\bar{x}})^2} = \frac{\bar{x}^2}{\alpha} \implies \hat{\alpha} = \frac{\bar{x}^2}{s_x^2}$$

$$\hat{\lambda} = \frac{\alpha}{\bar{x}} = \frac{\bar{x}}{s_x^2}$$

Note that these are not MLEs because α and λ do not have closed-form solutions in general. (We can imagine writing out the log-likelihood for Gamma, setting equal to zero and attempting to solve for λ in terms of α.) The method of moments estimates are not extremely informative, but they are something we have access to from only our sample, without tons of other assumptions.

So, we'll describe the procedure for generating bootstrap samples. Our goal is to approximate the sampling distribution of $\hat{\alpha}$ and $\hat{\lambda}$ to construct confidence intervals.

1. Estimate the parameters from observed data. For our Gamma setup we compute $\hat{\alpha}$ and $\hat{\lambda}$ using MOM formulas.

2. Simulate bootstrap samples. Ask R to generate many bootstrap datasets (say $B = 10,000$) of size n from our distribution.

$$\hat{X}_1^{*(b)}, \hat{X}_2^{*(b)}, \ldots, \hat{X}_n^{*(b)} \sim \text{Gamma}(\hat{\alpha}, \hat{\lambda}), \text{ for } b = 1, \ldots, B$$

3. Recalculate estimates. For each bootstrap sample, compute $\hat{\alpha}^*$ and $\hat{\lambda}^*$ using the same MOM formulas.

4. Construct the confidence interval(s). Use either the percentile method or the reverse percentile method.

The distinction between the bootstrap percentile method and the reverse bootstrap percentile method for confidence intervals (Read more in Hesterberg 5.4):

- Bootstrap Percentile Interval
  - Take the empirical quantiles of the bootstrap parameter estimates ($\hat{\alpha}^*$ and $\hat{\lambda}^*$)
  - For any parameter θ, 5th and 95th percentiles of the bootstrap distribution of $\hat{\theta}$ would give 90% CI endpoints $[\hat{\theta}^*_{(.05)}, \hat{\theta}^*_{(.95)}]$
  - This is sort of intuitive, we can think: "most bootstrap estimates fall between these values"

- Reverse Bootstrap Percentile Interval
  - Instead, we will use the formula $[2\hat{\theta} - \hat{\theta}^*_{(.95)}, 2\hat{\theta} - \hat{\theta}^*_{(.05)}]$ for our 90% CI.
  - This uses the idea of estimating the distribution of the estimation error, $\hat{\theta}^* - \hat{\theta}$
  - This reflects symmetry of error around the observed estimate, which is a big assumption
  - This isn't good for skewed data, it over-estimates in the opposite direction of the skew. This method is sort of a retired pedological method, one we wouldn't use in practice.

From here on out, we will be using the Bootstrap Percentile Interval.

Example: (Parametric) Swarthmore College Women's Soccer data in 2015:
We are given that they played $n = 23$ games, which resulted in a mean $\bar{x} = 15.63$ minutes before scoring their first goal, with a sample sd of $s = 16.2$ minutes ($s^2 = 262.44$ sample variance).
We can use the parametric bootstrap to obtain a 90% confidence interval for the estimate of α and λ, assuming the waiting times were iid Gamma(α, λ). Our method of moments estimates are as follows.

$$\hat{\alpha} = \frac{\bar{x}^2}{s_x^2} = \frac{15.63^2}{262.44} = 0.93$$

$$\hat{\lambda} = \frac{\bar{x}}{s_x^2} = \frac{15.63}{262.44} = 0.0596$$

See the R code appended on the last page (LHS) for the full procedure carried out by the code. The output for α is a 90% CI: [0.5708, 1.848]. From the problem, we know that a simple model has the goals following a Poisson(λ) process, meaning the X_is would be $\overset{iid}{\sim}$Exponential(λ)=Gamma(1,λ). We want to consider whether $\alpha = 1$ seems plausible. We see that 1 is within our 90% CI, so we conclude that this is a plausible model.

B) The nonparametric bootstrap does not assume the data follow a specific distribution. Instead, it makes use of the empirical CDF as an approximation to the actual CDF (usually assuming iid draws have been obtained). Not surprisingly, larger sample are required when the distribution itself is an unknown parameter. We'll go over the general procedure in this case:

1. Start with original data: $X_1, \ldots, X_n$, observed, assumed iid
2. Generate bootstrap samples: resample (with replacement) n values from the original data to get a bootstrap sample (repeat B times)

$$X_1^*, \ldots, X_n^*$$

3. Calculate statistic (same as before)
4. Form the confidence interval (same as before)

Example: (Nonparametric) Swarthmore College Women's Soccer data in 2015:
We're going to use bootstrap to find a confidence interval for the method of moments estimates of the soccer data. Since we don't have the raw data, we use a resampling technique (after simulating the original data using rnorm). See the right hand side of the final page for R code procedure and output. To compare to our parametric results, the 90% CI for α is [0.577,3.861]. We see that 1 is still included in our interval, so we'd make the same conclusion as in part A. We also notice with this interval as well as with the one for λ that the intervals are more variable, especially higher in the upwards direction. This makes sense, because we are assuming we know less about our sampling distribution.

Example: (Nonparametric) Grant PI funding data (grants.txt):
We're going to use the bootstrap to find a confidence interval for the mean grant value using the gender differences grant data.
The original mean is $274,952 with n=200. See the code on the following page to see the bootstrap simulation procedure. We get an output 90% CI of [241288, 311933]. We see that our original mean is somewhat centered within this interval. This is not the best estimate, but it is a relatively good one considering that it doesn't require us to make assumptions about the distribution our data come from.

Closing Remarks:
There are other nonparametric bootstrap applications to use, such as the bootstrap t, and the difference between means procedures, which can be read more about in Hesterberg Bootstrap textbook.

12. **Bootstrap Confidence Intervals**

The name "bootstrap" comes from the phrase "pull yourself by your bootstraps", which became a common expression for self-reliance. Of course it is impossible to pull yourself up by your bootstraps, and the statistical technique takes the name partly because it conveys a sense of getting something for nothing. The bootstrap algorithm makes use of the variability in your sample to estimate the sampling distribution of a sample statistic.

a) The *parametric* bootstrap makes use of simulation to approximate the sampling distributions of parameters from a known data distribution (Rice pp. 284-285). Describe this procedure in the context of finding interval estimates for α and λ based on $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \lambda)$. We can get crude "method of moments" estimates:

$$\hat{\alpha} = \frac{\bar{X}^2}{s_x^2} \quad \text{and} \quad \hat{\lambda} = \frac{\bar{X}}{s_x^2}$$

where s_x^2 is the usual sample variance (you could also use $\hat{\sigma}_x^2$). Note that these parameter values would imply $E(X) = \bar{X}$ and $\text{Var}(X_i) = s_x^2$. Describe the procedure for generating bootstrap samples and discuss the distinction between the bootstrap percentile method and the 'reverse bootstrap percentile' interval (Rice p. 285, Hesterberg 5.4).

As an example, in their $n = 23$ matches in 2015, Swarthmore's women's soccer team averaged $\bar{x} = 15.63$ minutes before their first goal, with a standard deviation $s = 16.2$ minutes. Use the parametric bootstrap to obtain 90% interval estimates for α and λ assuming the waiting times are iid Gamma(α, λ). A simple model has goals following a Poisson(λ) process, meaning the X_i's would be iid Exponential(λ), or Gamma$(1, \lambda)$. Does $\alpha = 1$ seem plausible?

b) The *non-parametric* bootstrap (Rice 10.4.6 and Hesterberg) does not assume the data follow a specific distribution. Rather, it makes use of the empirical CDF as an approximation to the actual CDF (usually assuming iid draws have been obtained). Not surprisingly, larger samples are required when the distribution itself is an unknown parameter. The actual dollar values for the grant data, along with the genders of the PI's, are contained in the file grants.txt. Use the bootstrap to find a confidence interval for the mean grant value.

```
> grants <- read.table("/Users/Leia/Downloads/grantsSub.txt", header=TRUE,sep=",")
> grant_values <- grants$Value
>
> B <- 10000
> n <- length(grant_values)
> bootstrap_means <- numeric(B)
>
> for (b in 1:B) {
+   sample_b <- sample(grant_values, size = n, replace = TRUE)
+   bootstrap_means[b] <- mean(sample_b)
+ }
> # Percentile confidence interval (90%)
> ci_percentile <- quantile(bootstrap_means, probs = c(0.05, 0.95))
>
> # Output results
> cat("Original Sample Mean:", round(mean(grant_values), 2), "\n")
Original Sample Mean: 274952
> cat("90% Percentile Bootstrap CI: [", ci_percentile[1], ",", ci_percentile[2], "]\n")
90% Percentile Bootstrap CI:
[ 241288.4 , 311933.6 ]
```

Soccer (Parametric)

```
> n <- 23
> x_bar <- 15.63
> s <- 16.2
> s2 <- s^2
>
> # Method of moments estimates
> alpha_hat <- x_bar^2 / s2
> lambda_hat <- x_bar / s2
>
> alpha_boot <- numeric(10000)
> lambda_boot <- numeric(10000)
>
> for (i in 1:10000) {
+     sample <- rgamma(n, shape =
alpha_hat, rate = lambda_hat)
+
+     # Sample mean and variance
+     x_bar_star <- mean(sample)
+     s2_star <- var(sample)
+
+     # Recalculate MOM estimates from the
bootstrap sample
+     alpha_boot[i] <- x_bar_star^2 / s2_star
+     lambda_boot[i] <- x_bar_star / s2_star
+ }
>
> # 90% Confidence Intervals using
percentiles
> alpha_CI <- quantile(alpha_boot, c(0.05,
0.95))
> lambda_CI <- quantile(lambda_boot,
c(0.05, 0.95))
>
> # Rounded Output
> round(alpha_CI, 4)
   5%    95%
0.5708 1.8480
> round(lambda_CI, 4)
   5%    95%
0.0337 0.1357
```

Soccer (Nonparametric)

```
> # Simulate Data
> n <- 23
> sim_data <- rnorm(n, mean = 15.63, sd =
16.2)
>
> # Set up bootstrap
> B <- 10000
> alpha_hat <- numeric(B)
> lambda_hat <- numeric(B)
>
> for (b in 1:B) {
+   sample_b <- sample(sim_data, size = n,
replace = TRUE)
+   mean_b <- mean(sample_b)
+   var_b <- var(sample_b)
+
+   # Method of moments estimators:
+   alpha_hat[b] <- mean_b^2 / var_b
+   lambda_hat[b] <- mean_b / var_b
+ }
>
> # Confidence intervals
> ci_alpha <- quantile(alpha_hat, probs =
c(0.05, 0.95))
> ci_lambda <- quantile(lambda_hat, probs
= c(0.05, 0.95))
>
> # Output
> cat("Nonparametric 90% CI for alpha: [",
round(ci_alpha[1], 3), ",", round(ci_alpha[2],
3), "]\n")
Nonparametric 90% CI for alpha:
[ 0.577 , 3.861 ]
> cat("Nonparametric 90% CI for lambda: [",
round(ci_lambda[1], 3), ",",
round(ci_lambda[2], 3), "]\n")
Nonparametric 90% CI for lambda:
[ 0.049 , 0.22 ]
```