Lecture 10 - Key

The Bootstrap

Our final section takes a slight diversion from the textbook to focus on another important (re)sampling technique known as the bootstrap.

Situation: Let $y_1, y_2, \ldots, y_n$ be a SRS of size n taken from a distribution F that is unknown. Let θ be a parameter of interest associated with this distribution and let $\hat{\theta} = S(y_1, y_2, \ldots, y_n)$ be a statistic used to estimate θ. The notation $S(y_1, y_2, \ldots, y_n)$ denotes a statistic based on the data $y_1, y_2, \ldots, y_n$.

- Ex. The sample mean $\bar{y} = \hat{\theta} = S(y_1, y_2, \ldots, y_n)$ is a statistic used to estimate the true mean.

Similar to our other sampling procedures, with the point estimate $\hat{\theta}$ we are interested in determining:

1. the *standard error of the estimator,*

2. the *bias of* $\hat{\theta}$*,*

3. the *confidence interval for* $\hat{\theta}$*.*

Bootstrap methods are computer intensive methods for estimating these quantities using bootstrap samples.

A bootstrap sample is a SRS of size n taken with replacement from the original data $y_1, y_2, \ldots, y_n$.

We denote a bootstrap sample as $y_1^*, y_2^*, \ldots, y_n^*$ which consists of *members of the original data set* $y_1, y_2, \ldots, y_n$ *with some members appearing zero times, some appearing one time, some appearing twice,...*

A bootstrap sample replication of $\hat{\theta}$, denoted $\hat{\theta}^*$, is the value of $\hat{\theta}$ evaluated *using the bootstrap sample* $y_1^*, y_2^*, \ldots, y_n^*$.

The bootstrap algorithm requires that a large number (B) of bootstrap sample be taken. The bootstrap sample replication $\hat{\theta}^*$ is then calculated for each of the B bootstrap samples. *We will denote the* b^th *bootstrap replication as* $\hat{\theta}^*(b)$ *for* $b = 1, 2, \ldots, B$.

Example. Consider six data points that correspond to the inches of snowfall on your scheduled exam dates: $\{y_1 = 9, y_2 = 4, y_3 = 13, y_4 = 5, y_5 = 6, y_6 = 8\}$.

Using the snowfall dataset, write R pseudocode for taking bootstrap samples from the snowfall dataset.

sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
6	8	8	6	8	8
4	6	13	9	6	6
4	9	6	6	6	4
5	6	9	5	9	4
13	8	9	8	6	8
8	8	5	8	8	5
6	6	6	9	9	13
13	6	8	9	8	5
6	5	9	5	5	9
6	8	4	5	4	5
13	9	9	4	9	5
8	8	8	13	13	13
4	4	6	4	5	13
6	4	4	9	6	13
6	6	8	4	9	13
9	6	8	6	8	9
6	5	4	5	5	4
9	13	4	4	8	8
4	4	5	8	8	4
4	8	9	8	9	13

Bootstrap Estimate of Standard Error

The bootstrap estimate of the standard error of $\hat{\theta}$ is

$$SE_b(\hat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B}[\hat{\beta}^*(b) - \hat{\bar{\theta}}^*]^2}{B-1}},$$

where $\hat{\bar{\theta}}^* = \sum_{b=1}^{B} \hat{\theta}^*(b) \; B$ is the sample mean of the B bootstrap replications. We have actually done something very similar. *Recall our approximate sampling distributions constructed using repeated samples (Lab 2 - bird data). In a similar fashion the bootstrap procedure approximates the distribution for $\hat{\theta}$.*

Note that the previous equation for $SE_B(\hat{\theta})$ is *simply the standard deviation of the B bootstrap replications.*

Under many circumstances, as the sample size n increases, the sampling distribution of $\hat{\theta}$ becomes more normally distributed. Under this assumption, an approximate t-based bootstrap confidence interval can be generated using $SE_B(\hat{\theta})$ and a t-distribution:

$$\hat{\theta} \pm t^* SE_B(\hat{\theta}),$$

where t^* has $n-1$ degrees of freedom. *Note n corresponds to the number of original samples, not the number of bootstrap relicates*

Bootstrap Estimate of Bias

The bias of $\hat{\theta} = S(Y_1, Y_2, \ldots, Y_N)$ as an estimator of θ is defined as:

$$bias(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

The bootstrap estimate of the bias of $\hat{\theta}$ as an estimate of θ is calculated by replacing the distribution F with the empirical distribution function $\hat{F}$. In other words the expectation (with respect to F) is unknown, but can be estimated using the empirical CDF $\hat{F}$ using the bootstrap samples. This yields

$$\hat{bias}_b(\hat{\theta}) = \hat{\bar{\theta}}^* - \hat{\theta}.$$

Then, the bias-corrected estimate of θ is

$$\tilde{\theta}_B = \hat{\theta} - \hat{bias}_B(\hat{\theta}) = 2\hat{\theta} - \hat{\bar{\theta}}^*.$$

One suggestion is to center confidence intervals at $\tilde{\theta}$ such that bias corrected t-based confidence intervals can be expressed as $\tilde{\theta}_B \pm t^* SE_B(\hat{\theta})$.

Bootstrap Confidence Intervals

There are a few options for generating confidence intervals from bootstrap replications.

The first option uses the normal approximation. An approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm t^* SE_B(\hat{\theta}) \qquad \text{or} \qquad \hat{\theta} \pm z^* SE_B(\hat{\theta}),$$

where t^* is an $\alpha/2$ critical value from a t-distribution with $n-1$ degrees of freedom and z^* is the $\alpha/2$ critical value from a normal distribution.

Recall, the confidence intervals can also be centered at the bias corrected point estimate $\tilde{\theta}$.

For an approximate 95% confidence interval for θ to be useful, it is expected that 95% of the confidence intervals from this method would contain θ. The same principle holds for other confidence levels.

If the sample size is not large enough and the distribution sampled from is highly skewed (or not close to a normal distribution), *then the confidence interval stated above will not be reliable. That is the nominal confidence level is not close to the true confidence level.*

Another option for calculating bootstrap based confidence intervals uses the percentiles from the bootstrap samples.

If the sample size is *relatively small or it is suspected that the sampling distribution of $\hat{\theta}$ is skewed or non-normal, an alternative to confidence intervals using the normal distribution are preferred.*

The simplest alternative is to use percentiles from the B bootstrap replications.

The approximate bootstrap percentile-based confidence interval for θ is

$$(\hat{\theta}^*_L, \hat{\theta}^*_U)$$

where $\hat{\theta}^*_L$ and $\hat{\theta}^*_U$ are the lower $\alpha/2$ and upper $(1-\alpha/2)$ percentiles of the B bootstrap replications respectively.

Practically to find $\hat{theta}^*_L$ and $\hat{theta}^*_U$ you:

1. Order *the B bootstrap replications $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$ from smallest to largest.*

2. Calculate $L = B \times alpha/2$ *and* $U = B \times (1-\alpha/2)$.

3. Find *the L^{th} and U^{th} values in the ordered list of bootstrap replications.*

4. The *L^{th} value is the lower confidence interval endpoint $\hat{\theta}^*_L$ and the U^{th} value is the upper confidence interval endpoint $\hat{\theta}^*_U$.*

Note the function `quantile(theta.star,probs=c(alpha/2,(1-alpha/2))` in R will return $\hat{\theta}_L$ and $\hat{\theta}_U$.

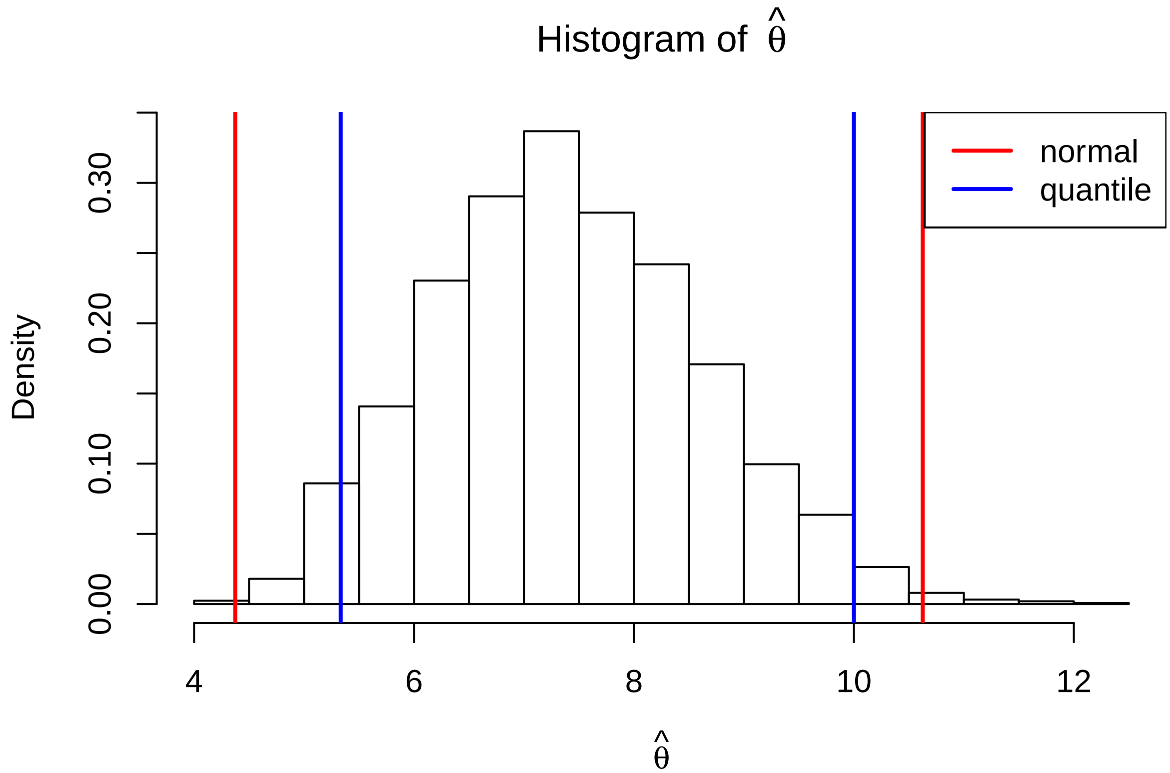
Bootstrap Examples in R

We will work through the code in class. Now continue with the snowfall dataset to create confidence intervals using the normal approach (t-dist) and the quantile-based approach.

```r
# Enter Data
snowfall <- c(9,4,13,5,6,8)
theta.hat <- mean(snowfall)
n <- length(snowfall)

B <- 5000 # number of bootstrap replicates
theta.boot <- rep(0,B)

# take bootstrap replications
for (i in 1:B){
  boot.samples <- sample(snowfall,replace=T)
  theta.boot[i] <- mean(boot.samples)
}

head(theta.boot,10)
```

```
##  [1] 7.833333 6.333333 7.166667 8.666667 6.833333 8.833333 5.333333
##  [8] 7.166667 6.666667 9.833333
```

```r
# visualize distribution from bootstrap draws
hist(theta.boot,probability=T,main=expression("Histogram of " ~ hat(theta)),
     xlab=expression(hat(theta)))

# compute standard error
se.b <- sd(theta.boot)

# compute bias corrected estimate
theta.boot.mean <- mean(theta.boot)
tilde.theta <- 2 * theta.hat - theta.boot.mean

# normality based confidence intervals
upper.norm <- theta.hat + qt(.975,n-1)* se.b
lower.norm <- theta.hat - qt(.975,n-1)* se.b

abline(v=upper.norm,col='red',lwd=2)
abline(v=lower.norm,col='red',lwd=2)

# percentile based confidence intervals
upper.per <- quantile(theta.boot,probs=.975)
lower.per <- quantile(theta.boot,probs=.025)

abline(v=upper.per,col='blue',lwd=2)
abline(v=lower.per,col='blue',lwd=2)

legend('topright',c('normal','quantile'),col=c('red','blue'),lty=1,lwd=2)
```


Histogram of $\hat{\theta}$
Density
0.00
0.10
0.20
0.30
4
6
8
10
12
$\hat{\theta}$
normal
quantile