Bootstrap

References:

Applied:

Efron B & Tibshirani RJ (1983) An Introduction to the Bootstrap

Theory:

Hall P (1992) The Bootstrap and Edgeworth Expansion

Use:

Bias, variance, confidence intervals

There are two basic approaches to the bootstrap: Nonparametric and Parametric

Focus on Nonparametric first

Empirical CDF:

Let $x_1, x_2, \ldots, x_n \overset{iid}{\sim} F$

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$$

$$= \frac{1}{n}\{\#\, x_i \le x\}$$

$$P_{F_n^*}[x \in A] = \frac{1}{n}\{\#\, x \in A\}$$

As you may have seen before, both are unbiased estimates of $F(x)$ and $P_F[x \in A]$ respectively. In addition

$$nF_n^*(x) \sim \text{Bin}(n, F(x))$$

$$nP_{F_n^*}[x \in A] \sim \text{Bin}(n, P_F[x \in A])$$

Plug-in estimators

Parameter: $t(F)$

Any functional of a distribution, not necessarily just the classical parameters of a distribution (e.g. μ & σ^2 for a normal, α & β for a Beta, etc)

Examples:

$$\mu_k(F) = \int x^k dF(x)$$

$$\omega_k(F) = \int \left(x - \mu_1(F)\right)^k dF(x)$$

$$\xi_p(F) = \inf\left\{x : F(x) \geq p\right\}$$

Estimator: $T(\mathbf{x})$

Natural estimators of these quantities are

$$\hat{\mu}_k(\mathbf{x}) = \mu_k\left(F_n^*\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$\hat{\omega}_k(\mathbf{x}) = \omega_k\left(F_n^*\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_1\left(F_n^*\right)\right)^k$$

$$\hat{\xi}_p(\mathbf{x}) = \xi_p\left(F_n^*\right) = \inf\left\{x : F_n^*(x) \geq p\right\}$$

In all cases, these estimators obey

$$T(\mathbf{x}) = t\left(F_n^*\right)$$

and are known as “plug-in” estimators. While many common estimators used are plug-in, not all are.

For example, the usual sample variance estimator

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x-\bar{x}\right)^2 = \frac{n}{n-1}\omega_2\left(F_n^*\right)$$

is not a plug-in estimator.

We want to learn about the properties of these estimators without having to make distributional assumptions about the data.

Standard Error estimation

For $\bar{x}$ getting a standard error is easy, as

$$se\left(\bar{x}\right) = \frac{s}{\sqrt{n}}$$

is a consistent estimate for any distribution F (assuming that the variance exists).

However for parameter estimates, such as the sample median, the distribution its difficult.

$$se(\text{Med}) \approx \frac{1}{2f(\xi_{0.5})\sqrt{n}}$$

where $f(\xi_{0.5})$ is the density evaluated at the population median. As density estimation is hard to do accurately, even in large samples, another approach is needed.

The bootstrap is a general procedure that can be used to estimate standard errors, biases, confidence intervals, etc for "any" estimator.

We can do this by looking at properties of these estimators when we sample from the distribution F_n^*, the empirical distribution.

Bootstrap sample:

$$\mathbf{x}^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$$

is obtained by sampling n items, **with replacement**, from the original data points $\mathbf{x} = \left(x_1, x_2, \ldots, x_n\right)$.

If n =5, two possible bootstrap samples are

$$\mathbf{x}^{*1} = \left(x_4, x_1, x_4, x_2, x_5\right)$$
$$\mathbf{x}^{*2} = \left(x_2, x_3, x_3, x_2, x_4\right)$$

The underlying idea behind the bootstrap is that the distribution of

$$T\left(\mathbf{x}\right) - t\left(F\right)$$

is similar to the distribution of

$$T\left(\mathbf{x}^*\right) - t\left(F_n^*\right)$$

While getting the exact sampling properties of $T\left(\mathbf{x}^*\right)$ is difficult, this second distribution is easy to deal with as it is easy to simulate from (or deal with exactly when n is small).

Bootstrap for Estimating Standard Errors

1) Select B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$.

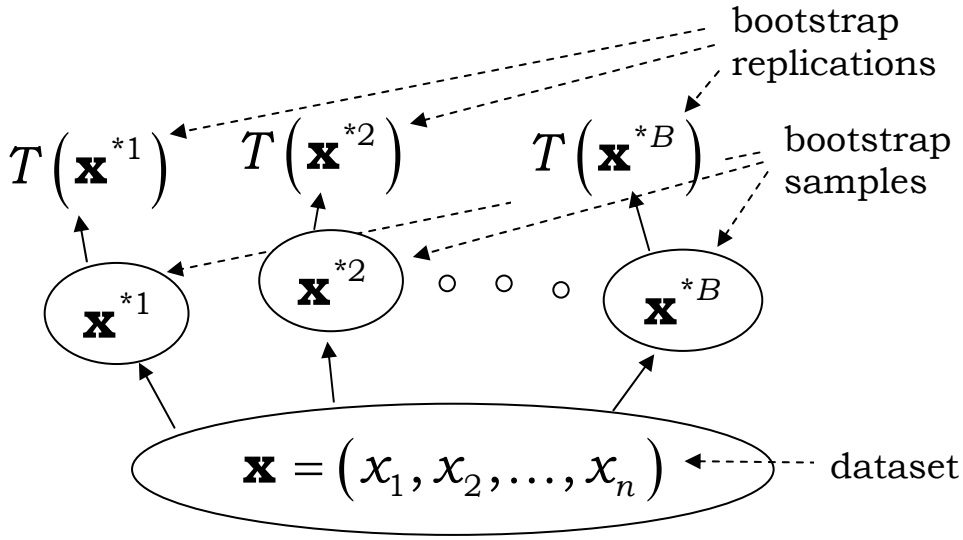
2) Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$T\left(\mathbf{x}^{*b}\right); \quad b = 1, \ldots, B$$

3) Evaluate the standard error $se\left(T\right)^{*}$ by

$$\widehat{se}\left(T\right)^{*} = \sqrt{\frac{\sum_{b=1}^{B}\left(T\left(\mathbf{x}^{*b}\right) - \hat{E}\left[T\right]^{*}\right)^{2}}{B-1}}$$

where $\hat{E}\left[T\right]^{*} = \sum_{b=1}^{B} T\left(\mathbf{x}^{*b}\right) \Big/ B$



Example: 1973 Law School Admissions

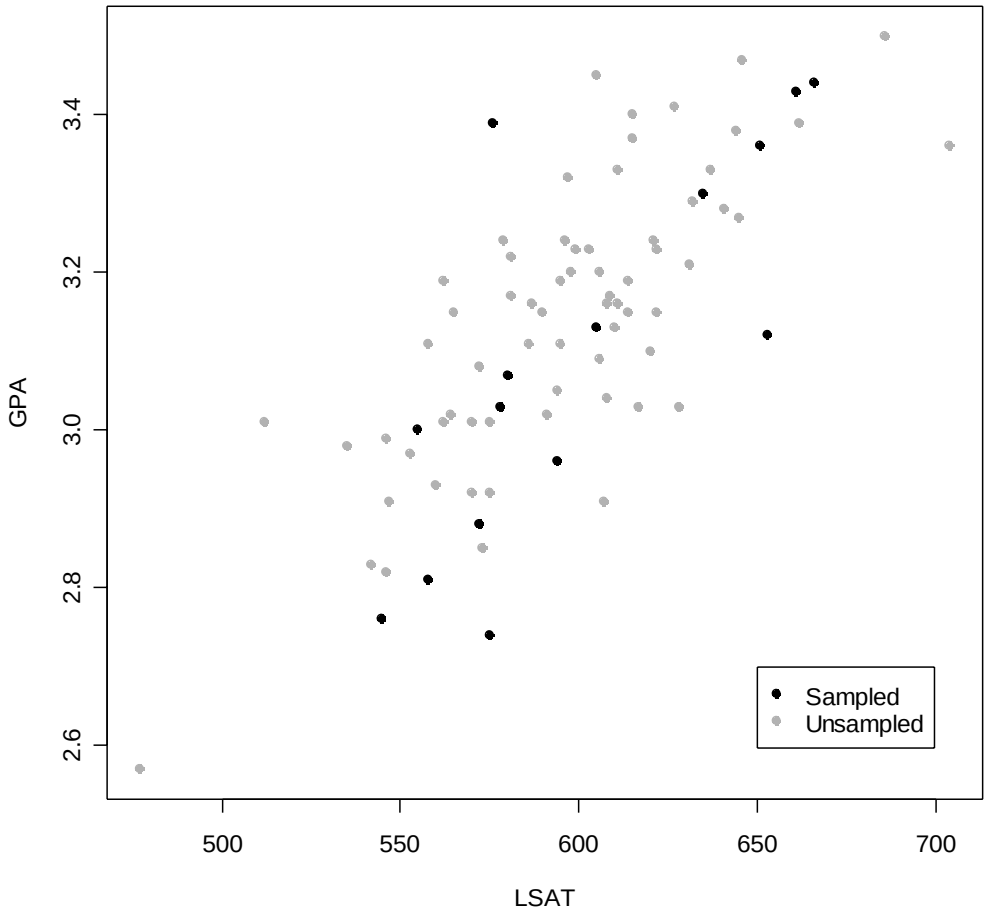
82 Law Schools

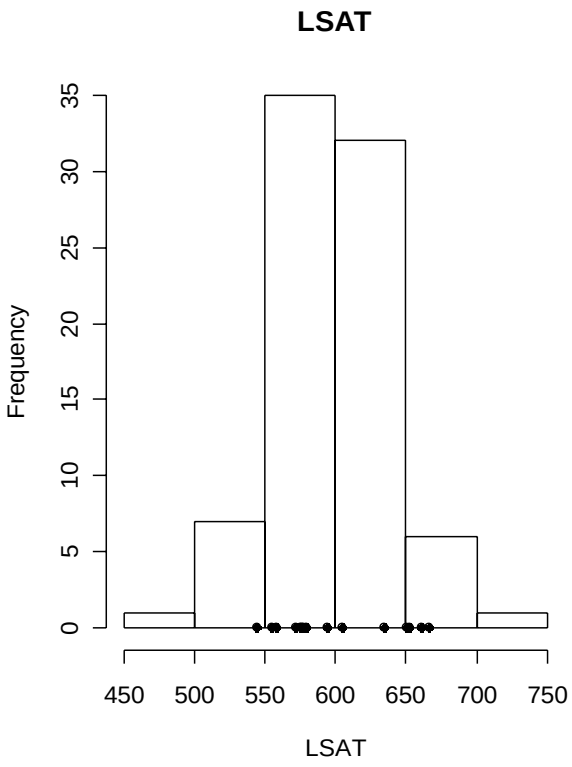
2 Variables:

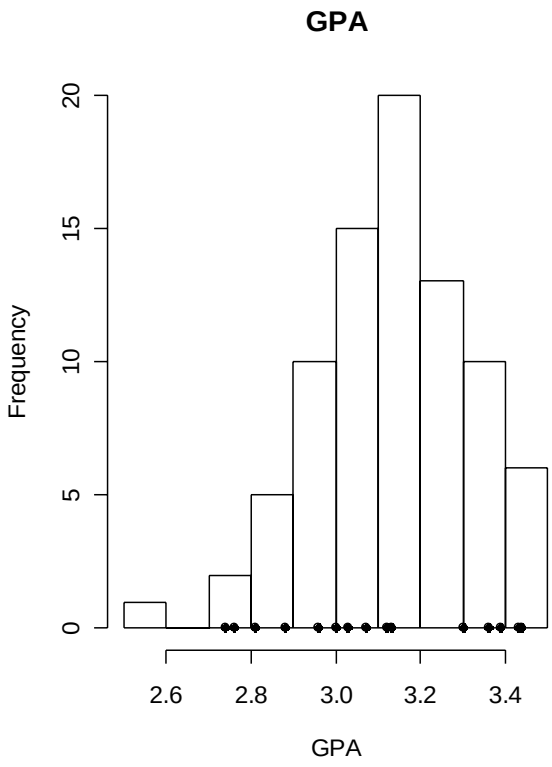
LSAT: average LSAT for incoming class

GPA: average GPA for incoming class

Sampled 15 schools







Want to examine the standard errors for $\bar{x}$, Median, and s for both LSAT and GPA and $r\left(LSAT, GPA\right)$, the correlation between the two variables.

True parameter values (N = 82)

	LSAT	GPA
Mean	597.549	3.135
Median	597.50	3.15
Std Dev	38.253	0.188

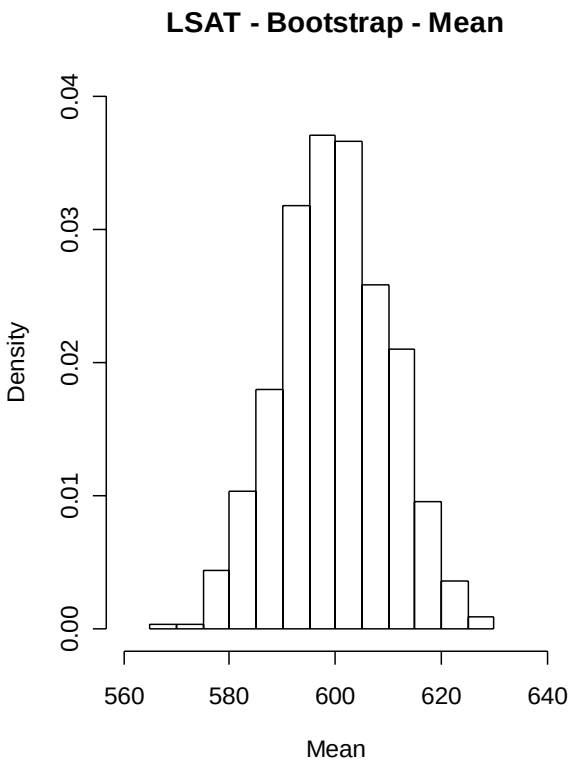
$$\rho\left(\text{LSAT}, \text{GPA}\right) = 0.760$$

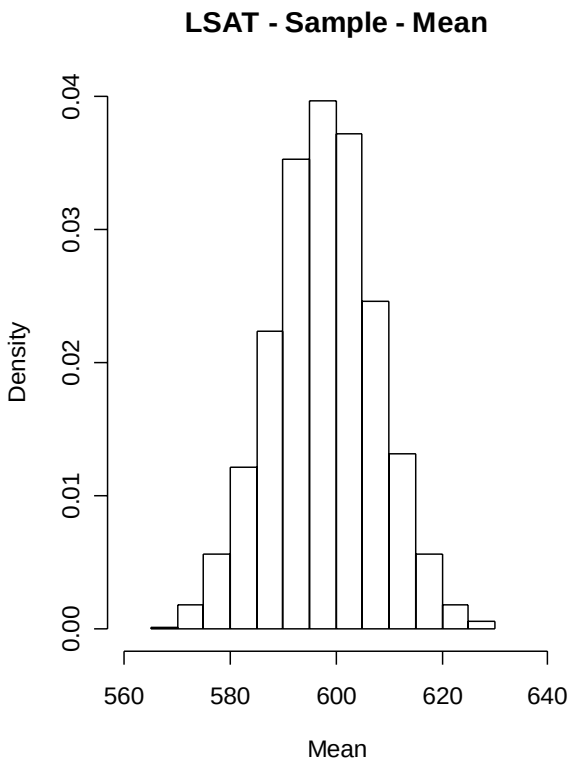
Parameter estimates from sample (n = 15)

	LSAT	GPA
Mean	600.267	3.095
Median	580.00	3.07
Std Dev	38.488	0.189

$$r\left(\text{LSAT}, \text{GPA}\right) = 0.776$$

In what follows, the histograms are based on 1000 samples. For the "Sample" histograms, samples of 15 observations (without replacement) from the 82 law schools in the population.

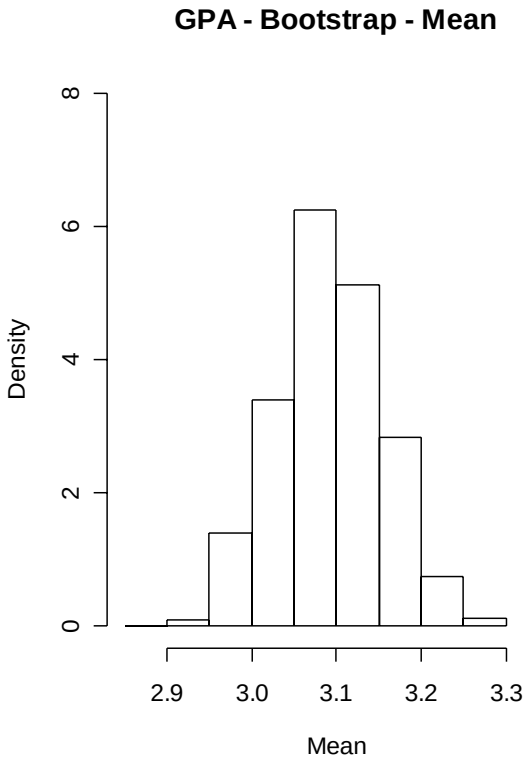


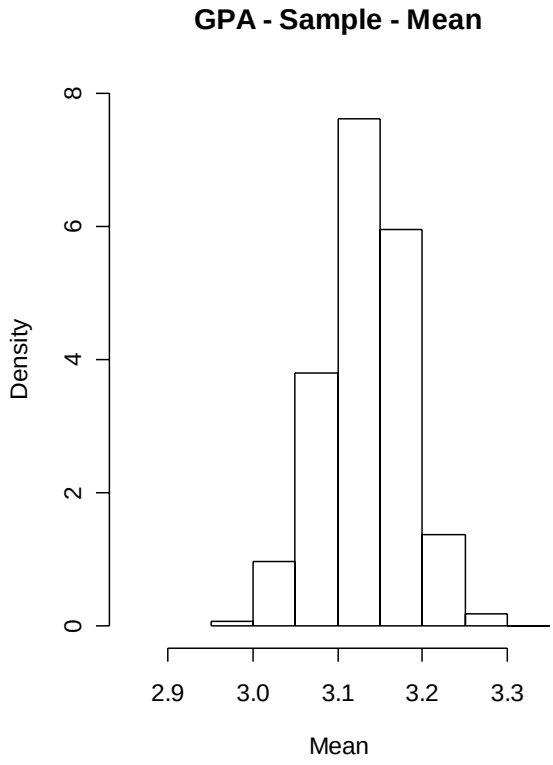


$$se(\bar{x}) = 9.877$$

Bootstrap estimates

B	50	100	150	200
$se(\bar{x})$	9.547	9.459	9.808	9.589
B	250	500	750	1000
$se(\bar{x})$	9.496	10.224	10.358	10.258

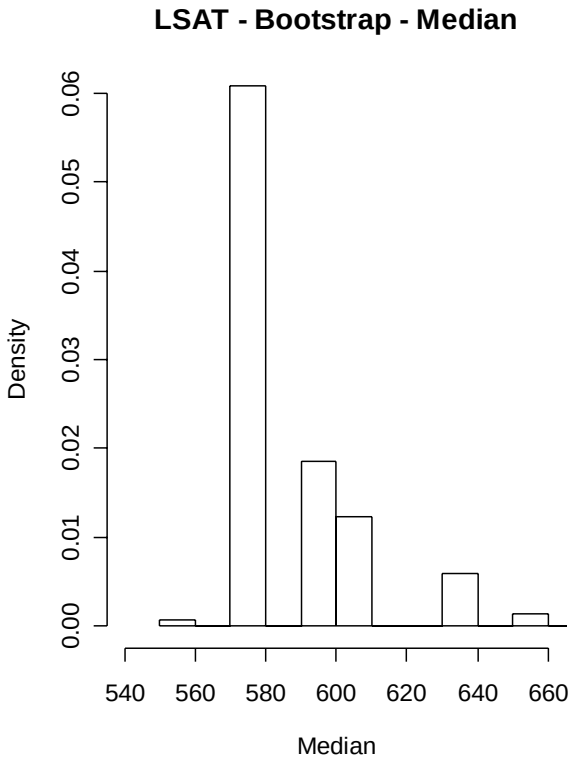


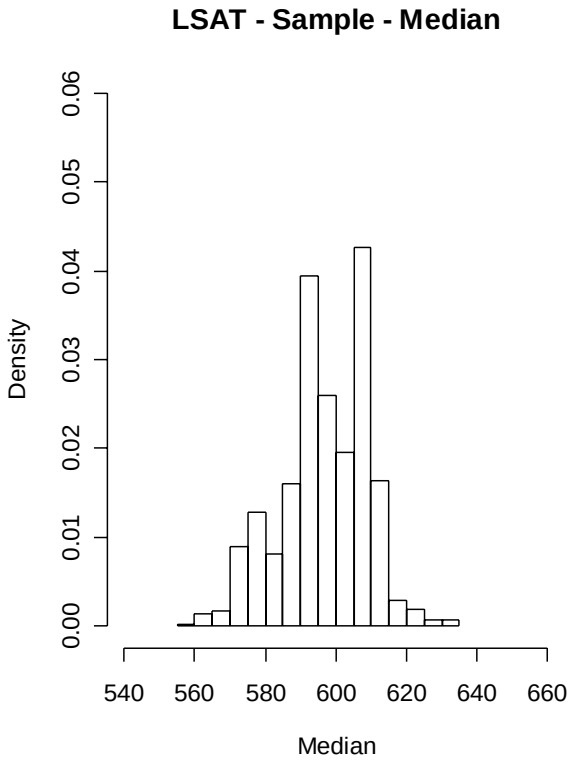


$$se(\bar{x}) = 0.0486$$

Bootstrap estimates

B	50	100	150	200
$se(\bar{x})$	0.0571	0.0575	0.0591	0.0622
B	250	500	750	1000
$se(\bar{x})$	0.0618	0.0632	0.0624	0.0632

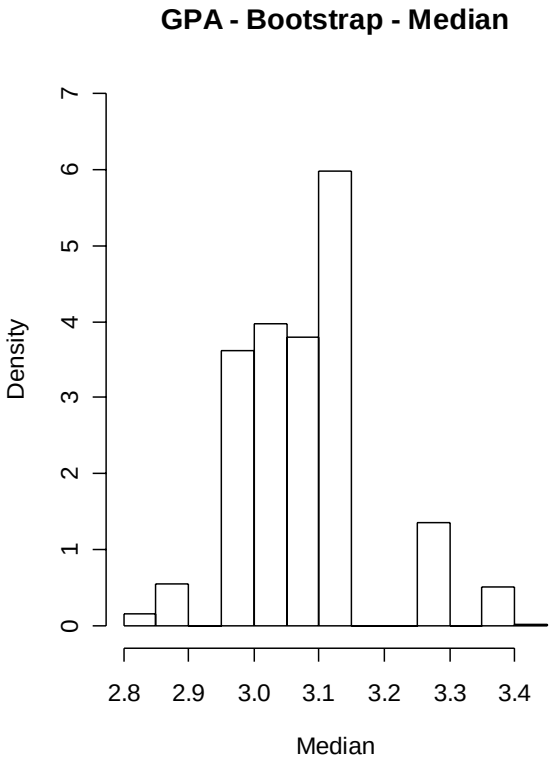


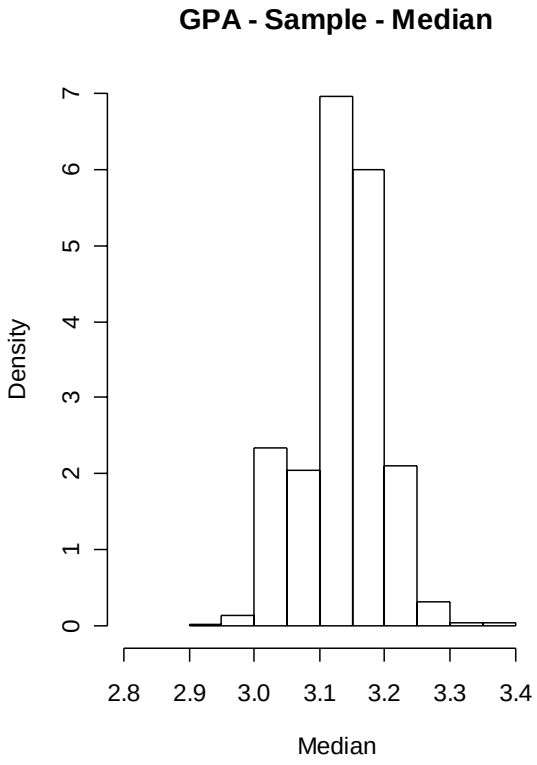


$se(\text{Med}) = 12.384$ (by Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$se(\text{Med})$	16.546	17.889	17.085	16.611
B	250	500	750	1000
$se(\text{Med})$	16.549	17.408	17.449	17.663

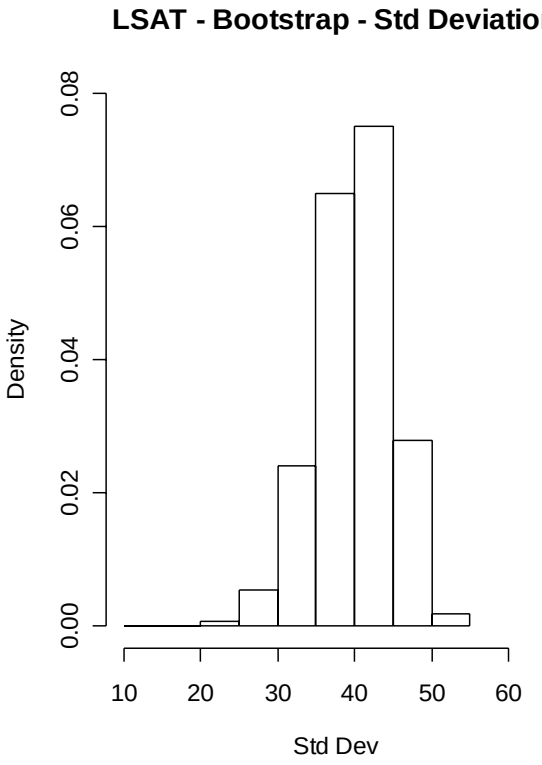


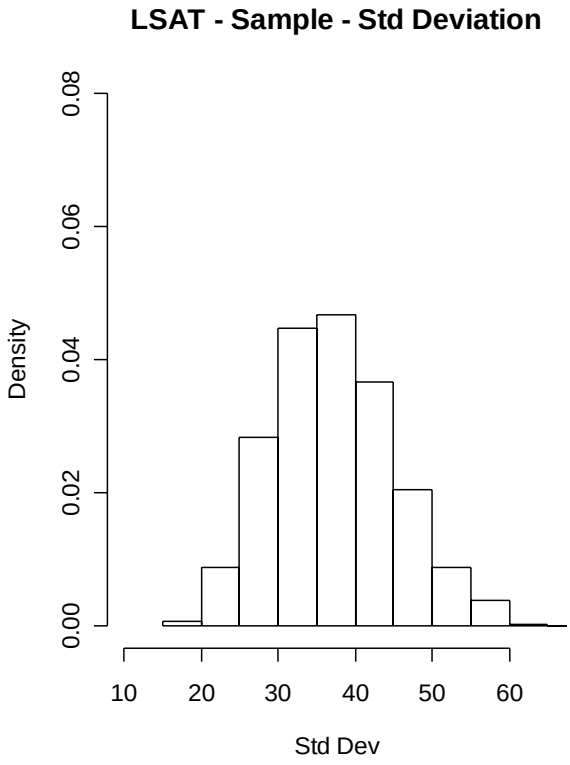


$se(\text{Med}) = 0.0619$ (by Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$se(\text{Med})$	0.0838	0.0931	0.0922	0.0977
B	250	500	750	1000
$se(\text{Med})$	0.0989	0.1008	0.0990	0.0992

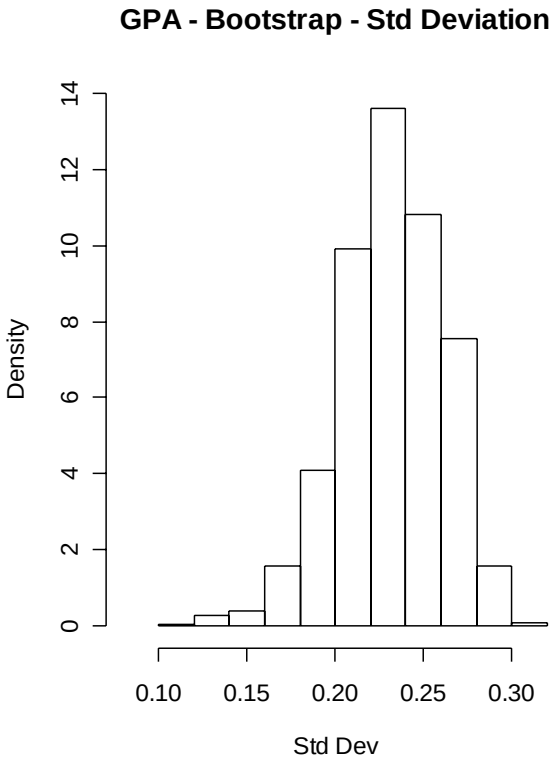


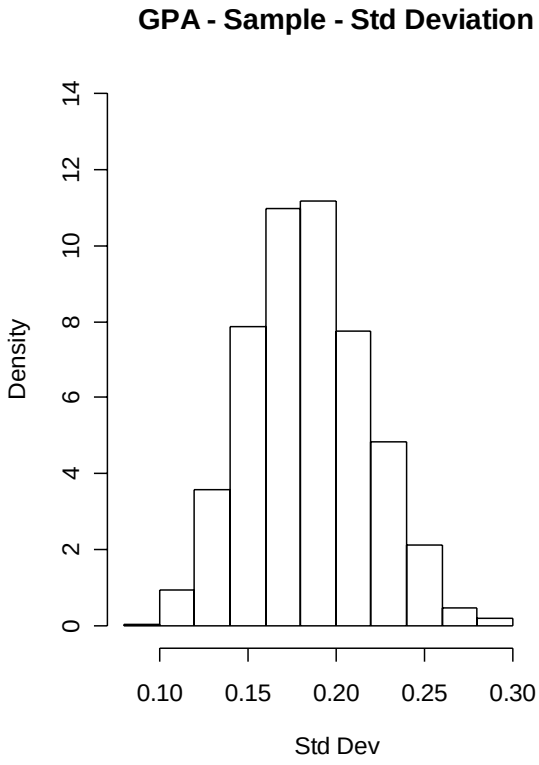


$se(s) = 8.025$ (by Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$se(s)$	4.738	4.587	4.801	4.848
B	250	500	750	1000
$se(s)$	4.802	4.887	4.958	4.936

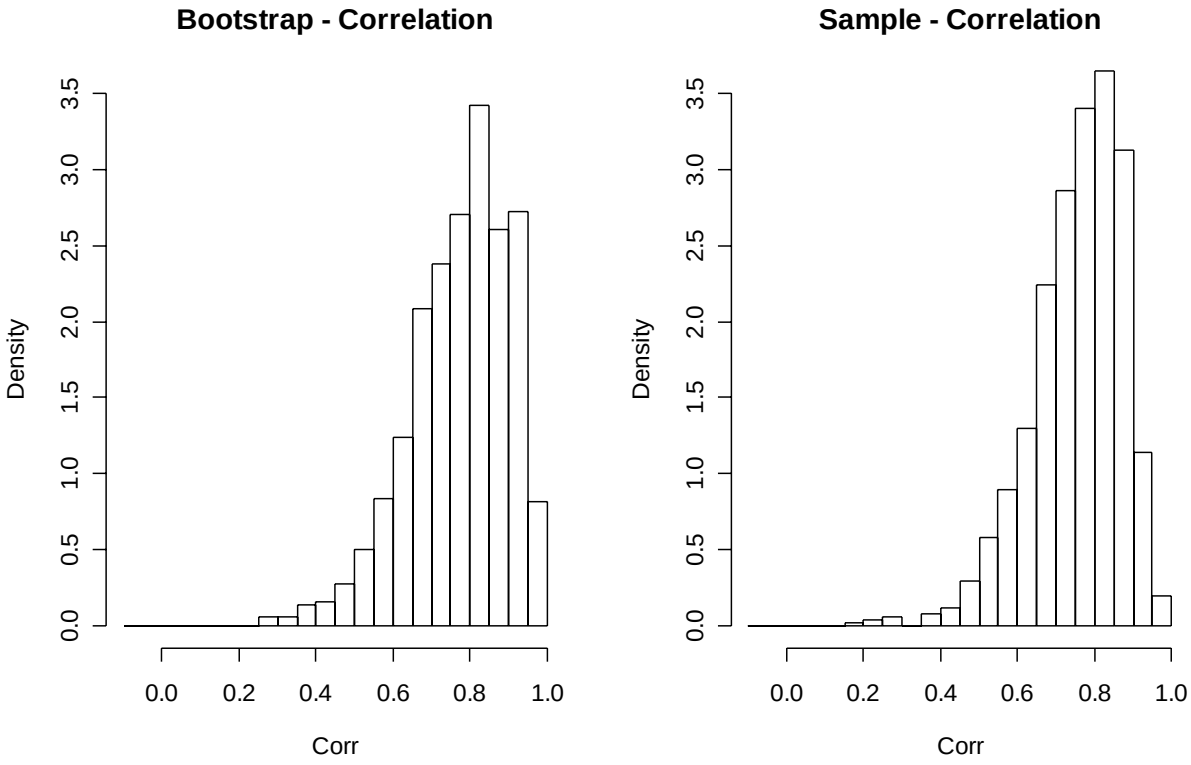




$se(s) = 0.0339$ (by Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$se(s)$	0.0275	0.0271	0.0268	0.0275
B	250	500	750	1000
$se(s)$	0.0272	0.0296	0.0293	0.0292



$se(r) = 0.118$ (by Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$se(r)$	0.126	0.129	0.137	0.138
B	250	500	750	1000
$se(r)$	0.135	0.131	0.129	0.130

What should B be?

For determining standard errors, B in the range of 25 to 100 is usually adequate.

Rarely should $B \geq 200$ be needed.

One way of determining B is based on the formula

$$cv\left(\widehat{se}_B\right) \approx \sqrt{cv\left(\widehat{se}_\infty\right) + \frac{E\left[\Delta\right] + 2}{4B}}$$

where Δ is a parameter that measures how long tailed the distribution of $T\left(\mathbf{x}^{*b}\right)$ is. Its 0 for the normal, has a minimum of -2 and can get arbitrarily large.

For usual values of Δ, $cv\left(\widehat{se}_B\right)$ is not much more than $cv\left(\widehat{se}_\infty\right)$ for $B \geq 200$.

However for other problems, such as determining CIs, B may need to be much larger. Booth and Sarkar argue for B = 800 for this problem

How well the bootstrap works for a particular problem depends on how smooth the distribution of the statistic of interest is.

The distribution of the sample median is not as smooth as that of the sample mean, particularly for a discrete distribution as we have here.

Estimating Bias:

$$\begin{aligned} Bias_F &= E_F\left[T(\mathbf{x}) - t(F)\right] \\ &= E_F\left[T(\mathbf{x})\right] - t(F) \end{aligned}$$

Can approximate this by

$$\begin{aligned} Bias_{F_n^*} &= E_{F_n^*}\left[T(\mathbf{x}^*) - t(F_n^*)\right] \\ &= E_{F_n^*}\left[T(\mathbf{x}^*)\right] - t(F_n^*) \end{aligned}$$

Thus it is easy to estimate the bias since we know how to approximate $E_{F_n^*}\left[T(\mathbf{x}^*)\right]$ by Monte Carlo.

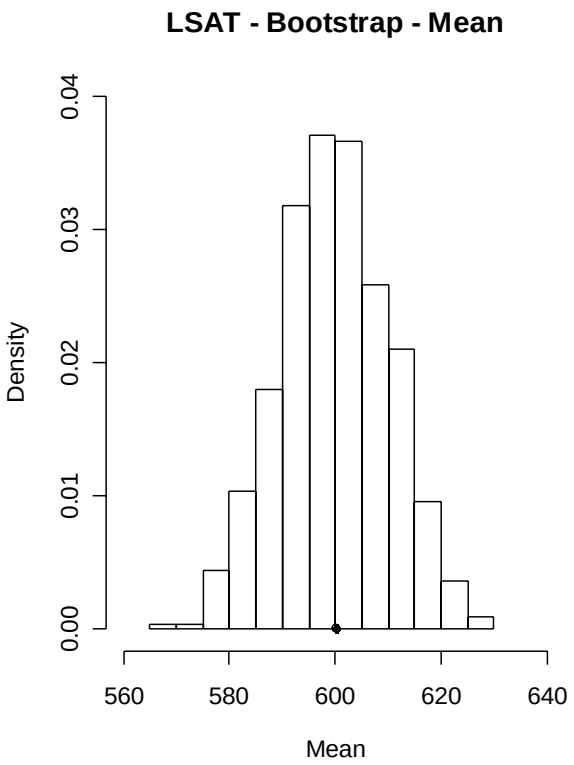
In fact we needed to estimate this as part of the standard error calculation as

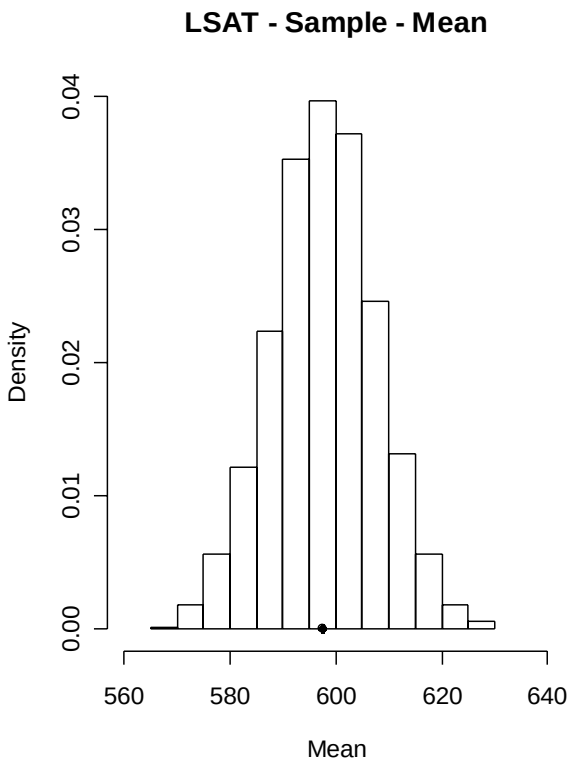
$$\hat{E}\left[T\right]^{*} = \sum_{b=1}^{B} T\left(\mathbf{x}^{*b}\right) \Big/ B$$

is an estimate of this quantity.

Therefore

$$\begin{aligned} \widehat{Bias}_B &= \hat{E}\left[T\right]^{*} - t\left(F_n^{*}\right) \\ &= \hat{E}\left[T\right]^{*} - T\left(\mathbf{x}\right) \end{aligned}$$

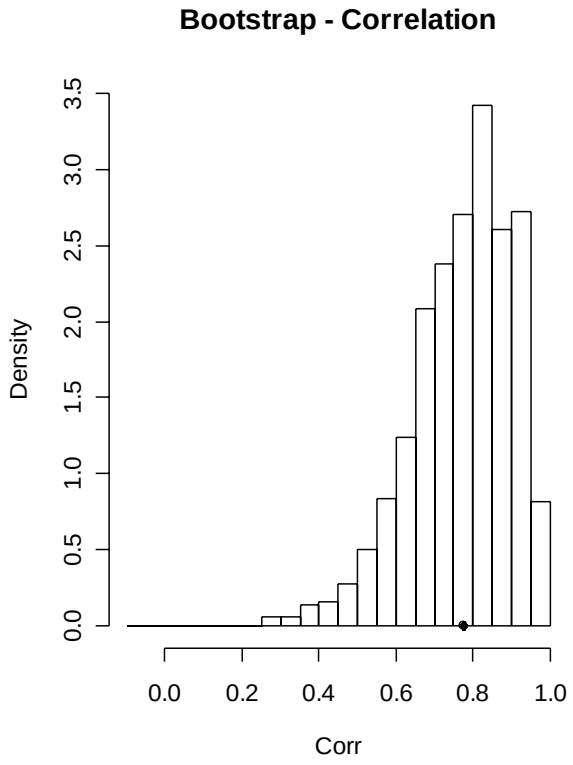


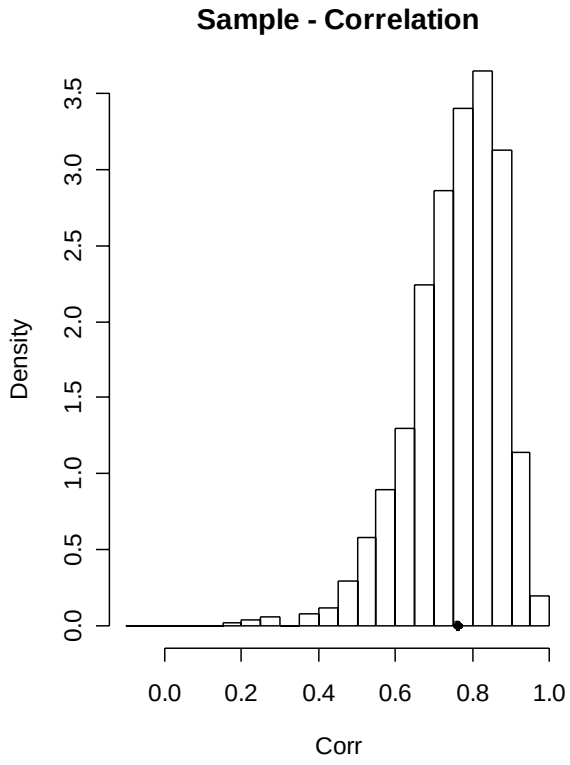


$$Bias(\bar{x}) = 0$$

Bootstrap estimates

B	50	100	150	200
$Bias(\bar{x})$	-0.052	0.364	-0.200	0.229
B	250	500	750	1000
$Bias(\bar{x})$	0.196	-0.591	-0.642	-0.351





$Bias(r) = -0.004827$ (Monte Carlo)

Bootstrap estimates

B	50	100	150	200
$Bias(\bar{x})$	0.001762	0.001063	-0.004509	0.002188
B	250	500	750	1000
$Bias(\bar{x})$	-0.008018	-0.003889	-0.003044	-0.003662

Bias Correction

Since we can estimate the bias of an estimator, we can use this to correct for it.

$$\begin{aligned} Bias_F &= E_F\left[T(\mathbf{x}) - t(F)\right] \\ &= E_F\left[T(\mathbf{x})\right] - t(F) \end{aligned}$$

Can approximate this by

$$\begin{aligned} Bias_{F_n^*} &= E_{F_n^*}\left[T(\mathbf{x}^*) - t(F_n^*)\right] \\ &= E_{F_n^*}\left[T(\mathbf{x}^*)\right] - t(F_n^*) \end{aligned}$$

One approach to correcting for bias is based on figuring out the expectation of an estimator.

For example, lets estimate μ_1^2 by $\bar{x}^2$

$$\begin{aligned} E\left[\bar{x}^2\right] &= E\left[\left(\mu_1 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)\right)^2\right] \\ &= \mu_1^2 + 2\mu_1 E\left[x - \mu_1\right] + \frac{\omega_2}{n} \\ &= \mu_1^2 + \frac{\omega_2}{n} \end{aligned}$$

Thus a less biased estimator is

$$\bar{x}^2 - \frac{\hat{\omega}_2}{n} = \bar{x}^2 - \frac{s^2}{n}$$

This approach assumes that the expectations can be determined.

As this is often difficult to do, the bootstrap gives us an easy way to approximate this as

$$\begin{aligned}\widehat{Bias}_B &= \hat{E}\left[T\right]^* - t\left(F_n^*\right) \\ &= \hat{E}\left[T\right]^* - T\left(\mathbf{x}\right)\end{aligned}$$

So the estimator

$$T\left(\mathbf{x}\right) - \widehat{Bias}_B$$

usually will have lower bias than $T\left(\mathbf{x}\right)$

Parametric Bootstrap

Assume that the data comes from some parametric family F_θ.

For example, with the Law School example, we could assume that the data is bivariate normal ($\theta = (\mu, \Sigma)$).

The common approach for determining standard errors in the parametric setting is to use the delta rule or some other asymptotic approximation.

For example

$$se(r) = \frac{1 - r^2}{\sqrt{n - 3}}$$

So for the Law School example, r = 0.776 which gives $se(r)$ = 0.115 (which is similar to the nonparmetric bootstrap value of 0.130 (B = 1000)).

Instead of using the textbook asymptotic formula, we can use the parametric bootstrap instead.

Parametric Bootstrap for Estimating Standard Errors and Bias

1) Estimate the parameter given the assumed distributional form (call it $\hat{\theta}$)

2) Select B independent parametric bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, each consisting of n data values drawn from the distribution $F_{\hat{\theta}}$.

3) Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$T\left(\mathbf{x}^{*b}\right); \quad b = 1, \ldots, B$$

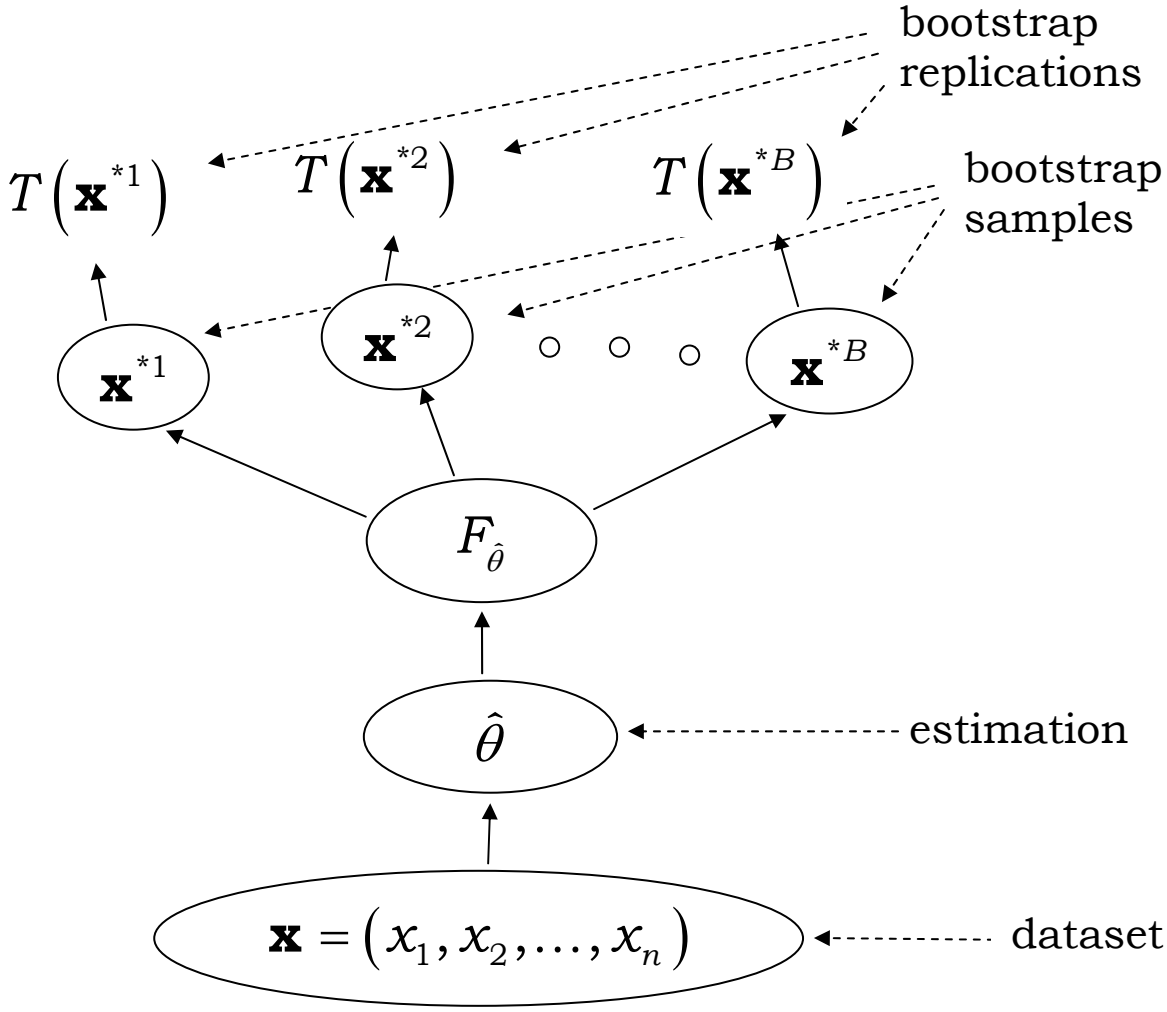
4) Evaluate the standard error, $se\left(T\right)^*$ by

$$\widehat{se}\left(T\right)^* = \sqrt{\frac{\sum_{b=1}^{B}\left(T\left(\mathbf{x}^{*b}\right) - \hat{E}\left[T\right]^*\right)^2}{B-1}}$$

where $\hat{E}\left[T\right]^* = \sum_{b=1}^{B} T\left(\mathbf{x}^{*b}\right)\Big/B$

5) Evaluate the bias $\widehat{Bias}_B$ by

$$\widehat{Bias}_B = \hat{E}\left[T\right]^* - T\left(\mathbf{x}\right)$$



For the Law School example, lets set $\hat{\theta} = \left(\hat{\mu}, \hat{\Sigma}\right)$ where

$$\hat{\mu} = \begin{bmatrix} \bar{y} & \bar{z} \end{bmatrix}^T$$

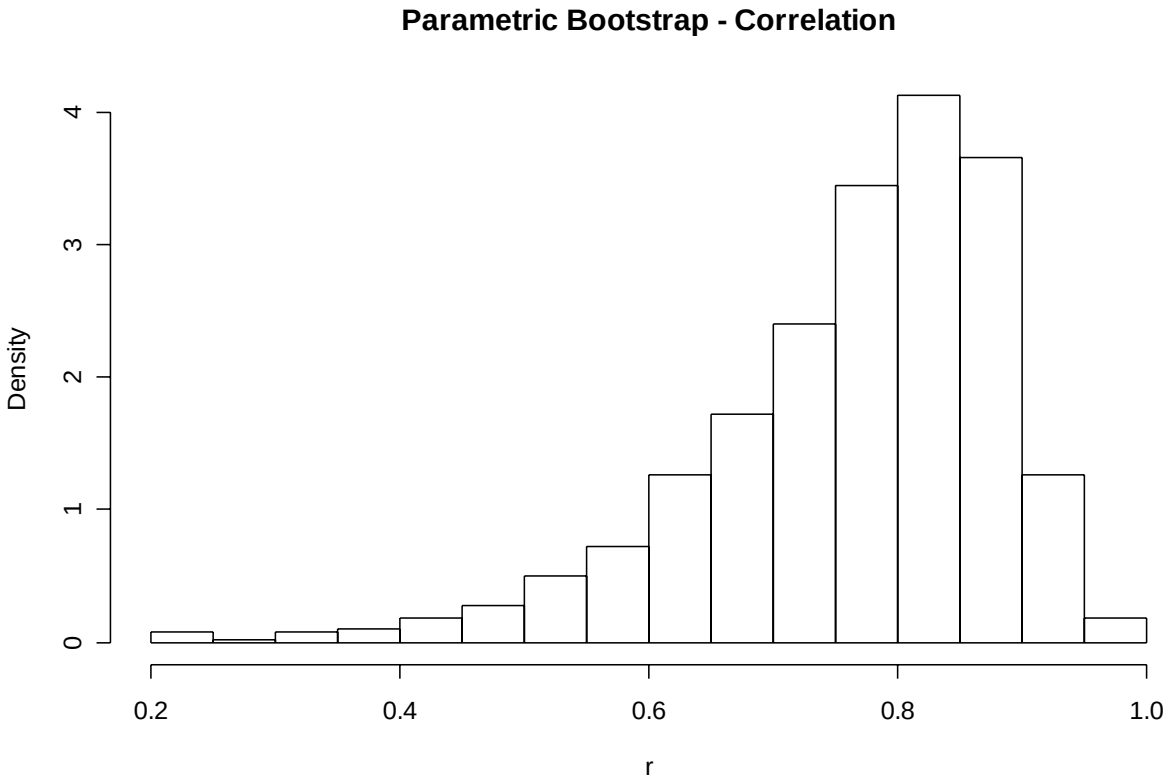
$$\hat{\Sigma} = \frac{1}{14}\begin{bmatrix} \sum (y_i - \bar{y})^2 & \sum (y_i - \bar{y})(z_i - \bar{z}) \\ \sum (y_i - \bar{y})(z_i - \bar{z}) & \sum (z_i - \bar{z})^2 \end{bmatrix}$$

are the usual parameter estimates (y: LSAT, z: GPA). For the example, these are

$$\hat{\mu} = \begin{bmatrix} 600.267 & 3.095 \end{bmatrix}^T$$

$$\hat{\Sigma} = \begin{bmatrix} 1746.781 & 7.902 \\ 7.902 & 0.0593 \end{bmatrix}$$

$$r = 0.776$$



$se(r)$ = 0.115 (asymptotic formula)

$\widehat{se}(r)$ = 0.122 (parametric bootstrap)

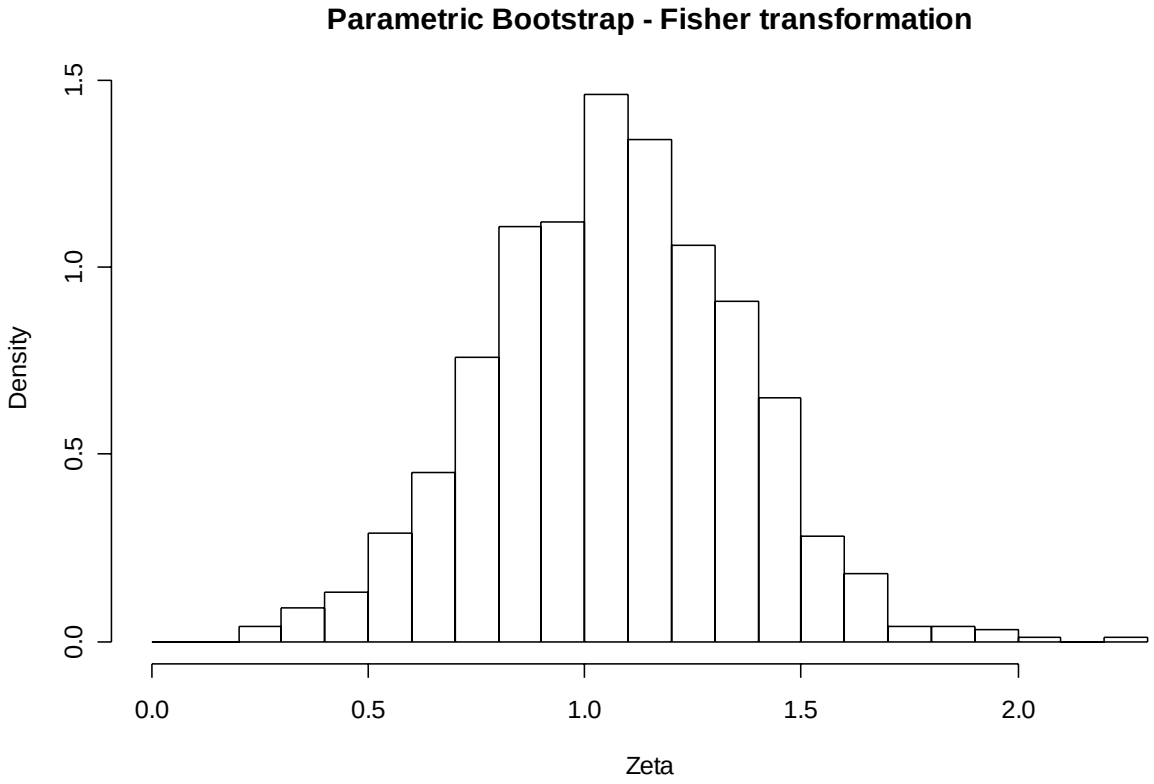
$\widehat{Bias}$ = -0.0125

For the correlation, Fisher's transformation is often used since it better distributional properties. Fisher showed that

$$\xi = \frac{1}{2}\frac{1+r}{1-r} \sim N\left(\frac{1}{2}\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

approximately.

Lets transform our bootstrap sample to see how well this works



$$\frac{1}{2}\frac{1+r}{1-r} = 1.036$$

$se(\xi) = 0.289$ (asymptotic formula)

$\widehat{se}(\xi) = 0.292$ (parametric bootstrap)

$\widehat{Bias} = 0.0319$

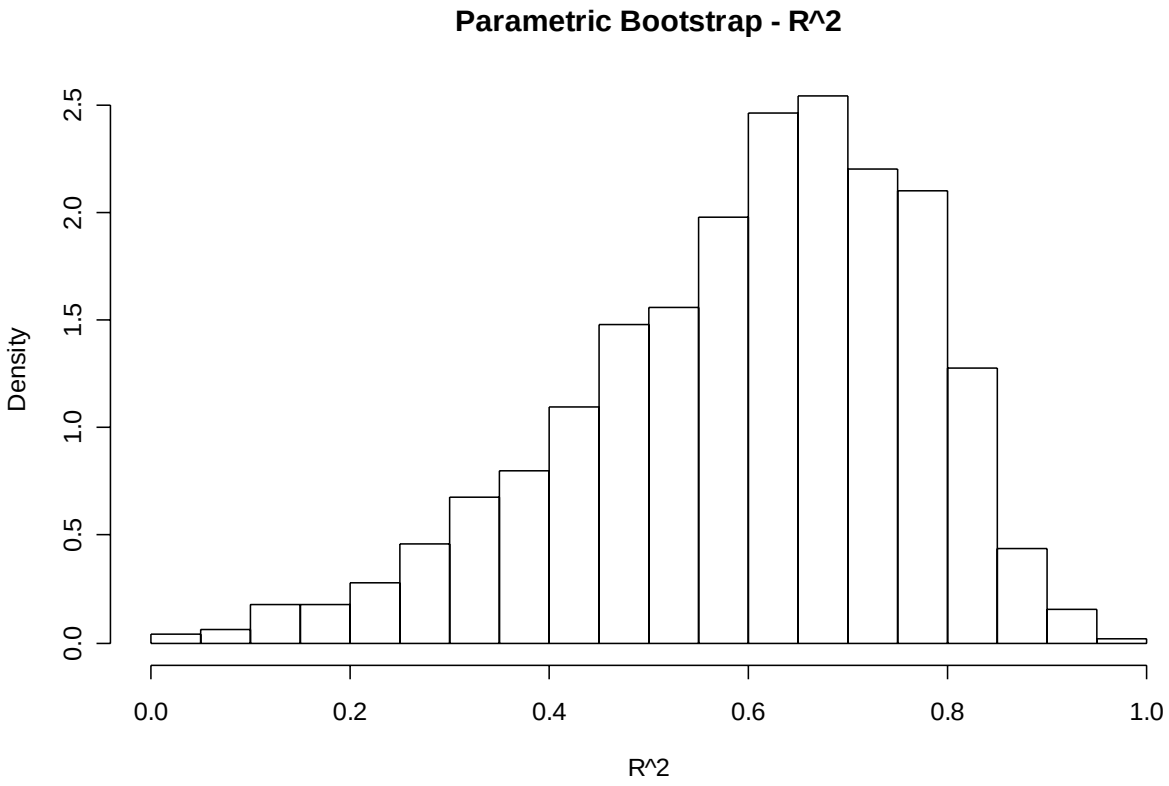
For both of these examples, the parametric bootstrap estimates of the standard error agree well with the asymptotic formula.

This is usually the case. So why bother with the parametric bootstrap?

It can provide more accurate answers. Some of the text book formulas only work well when the sample size is large. An example of this is $se(RR)$ from a 2x2 table.

It can be used when asymptotic formulas for standard errors are unknown.

For example, what is $se\left(r^2\right)$?



r^2 = 0.603

$se\left(r^2\right)$ = ??? (asymptotic formula)

$\widehat{se}\left(r^2\right)$ = 0.169 (parametric bootstrap)

$\widehat{Bias}$ = -0.00445

Actually an asymptotic formula for $se\left(r^2\right)$ is probably known since

$$r^2 = \frac{F}{F + n - 2}$$

where F is the usual ANOVA F-test for $\rho = 0$

Confidence Intervals

As the bootstrap is used to approximate the sampling distribution, it can be used to generate confidence intervals (and for hypothesis testing as well).

There is a wide range of bootstrapping approaches to this problem. Which of these to use depends on the form of the bootstrap distribution.

In the examples, I'll use a nonparametric bootstrap, but parametric bootstrap can also be used.

Also in all the examples, I'll be using $1 - 2\alpha$ percentile intervals.

Notation: $\hat{\theta}^*\left(b\right) = T\left(\mathbf{x}^{*b}\right)$

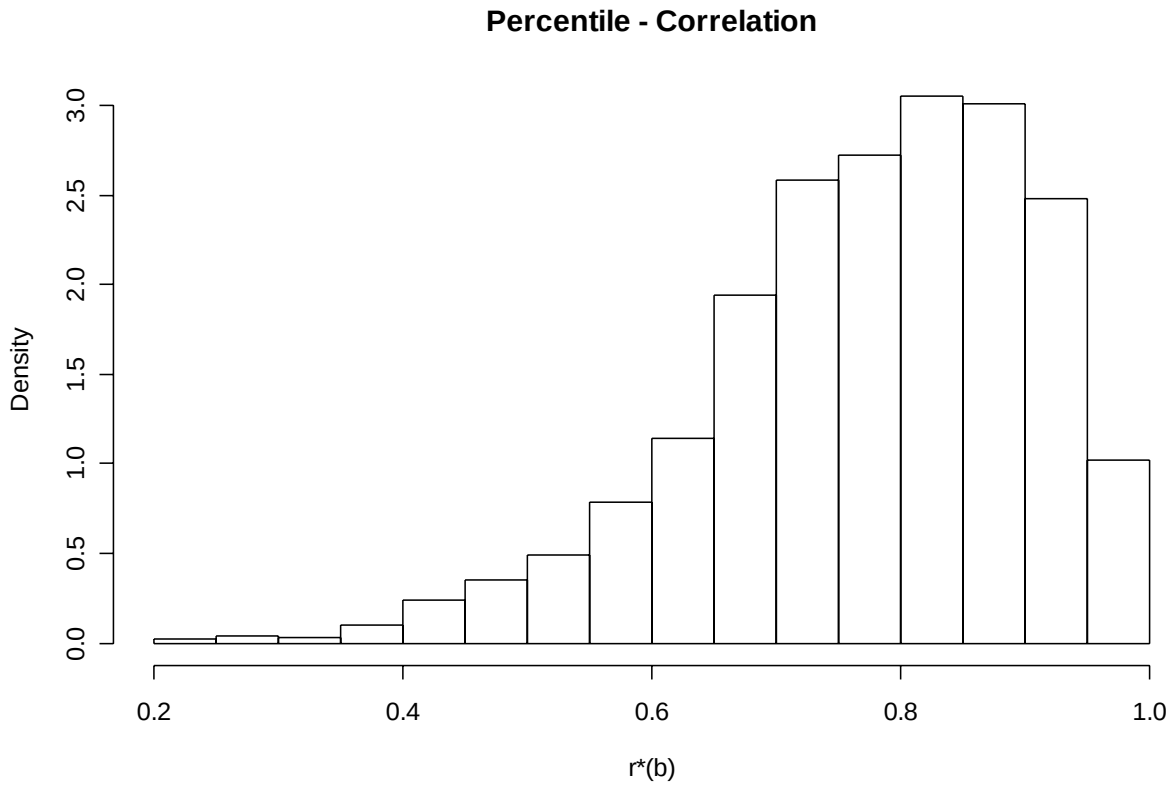
Percentile Interval

Easy interval to generate

$$\left[\hat{\theta}_{lo}, \hat{\theta}_{up}\right] = \left[\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}\right]$$

where $\hat{\theta}_B^{*(\alpha)}$ is the 100αth empirical percential of the $\hat{\theta}^*(b)$ values.

So if B = 2000 and α = 0.05, we need the 100th and 900th ordered values of the $\hat{\theta}^*(b)$'s



For the correlation example, confidence intervals for different confidence levels are

Level	Lower	Upper
90%	0.525	0.951
95%	0.458	0.964
99%	0.367	0.980

To get these confidence interval, B needs to be fairly large since we need to determine the tail properties of the sampling distribution. Efron and Tibshirani recommend B being at least 1000 for reasonable choices of α.

Bootstrap-t interval

Based on the standard t based confidence interval

$$\left[\hat{\theta} - t^{(1-\alpha)} se\left(\hat{\theta}\right), \hat{\theta} - t^{(\alpha)} se\left(\hat{\theta}\right)\right]$$

which is based on

$$z = \frac{\hat{\theta} - \theta}{se\left(\hat{\theta}\right)}$$

having an approximate t distribution.

The idea behind the bootstrap-t interval is to use the bootstrap to approximate the distribution of z.

$$\text{Let } Z^*\left(b\right) = \frac{\hat{\theta}^*\left(b\right) - \hat{\theta}}{\widehat{se}^*\left(b\right)}$$

Then the $\hat{t}^{(\alpha)}$ is the 100αth empirical percential of the $Z^*\left(b\right)$.

The bootstrap-t interval is

$$\left[\hat{\theta} - \hat{t}^{(1-\alpha)} se\left(\hat{\theta}\right), \hat{\theta} - \hat{t}^{(\alpha)} se\left(\hat{\theta}\right)\right]$$

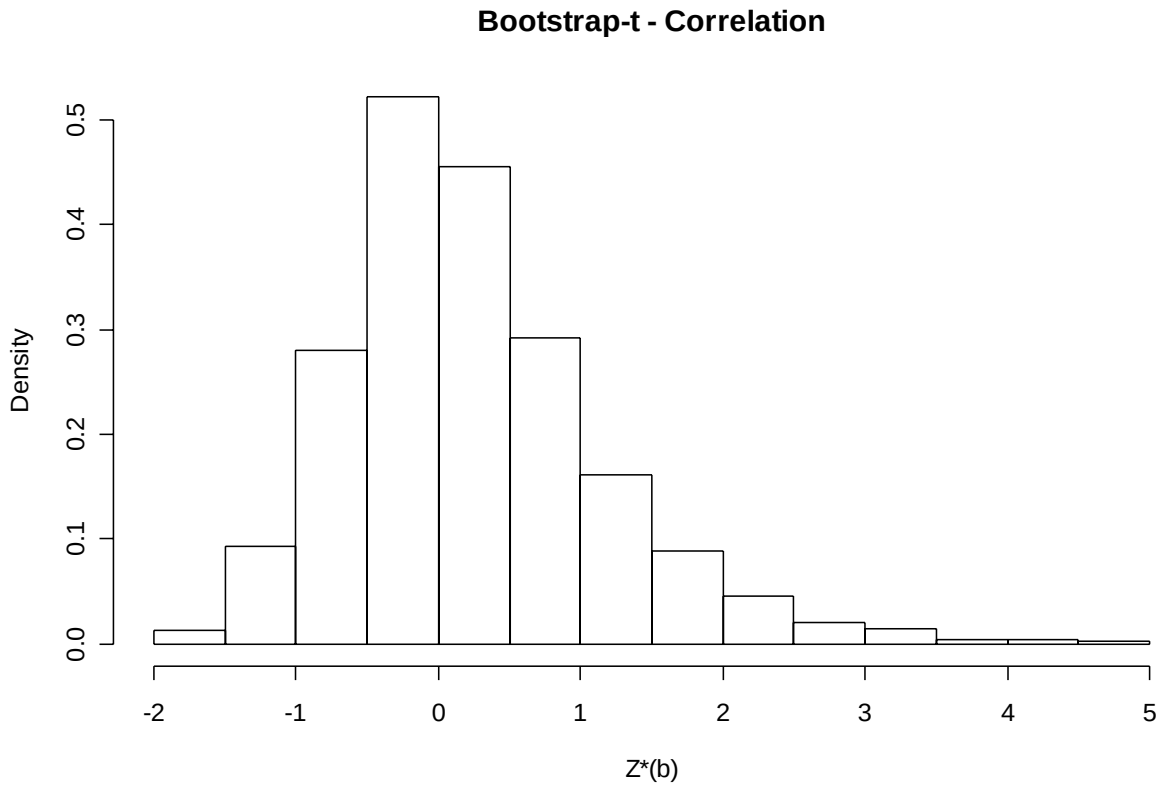
One complication to this procedure is that a standard error, $\widehat{se}^*\left(b\right)$ is needed for each bootstrap sample.

For the correlation example, you could use the textbook formula assuming normality

$$se(r) = \frac{1 - r^2}{\sqrt{n - 3}}$$

Another approach, not used here, is to use a second level of bootstrapping to estimate the standard error.

This requires $B_1 B_2$ total bootstrap samples, where B_1 is the number of bootstrap samples to get the distribution of $Z^*(b)$ and B_2 is the number of bootstrap samples for each B_1 sample to get the standard error.



Level	Lower	Upper
90%	0.522	0.910
95%	0.459	0.938
99%	0.311	0.975

This approach often works better when the distribution of $Z^*(b)$ is roughly pivotal (the distribution doesn't depend on the parameters of interest.

For the correlation example, Fisher's transformation can be used.

Get a CI for ξ and then transform back to get one for ρ.

In this case the back transformation is

$$\rho = \frac{e^{2\xi} - 1}{e^{2\xi} + 1}$$

So for this case, if a confidence interval for ξ is $\left[\xi_{lo}, \xi_{up}\right]$, a confidence interval for ρ is

$$\left[\frac{e^{2\xi_{lo}}-1}{e^{2\xi_{lo}}+1},\frac{e^{2\xi_{up}}-1}{e^{2\xi_{up}}+1}\right]$$

For the example, the intervals are

Level	Lower	Upper
90%	-0.020	0.926
95%	-0.221	0.941
99%	-0.555	0.955

These intervals are based on the asymptotic variance formula for ξ and can be replaced by a bootstrap estimate. This gives much different intervals than the other 2 procedures.

While it didn't occur with this example, it is possible for an end point, to be outside the range [-1, 1] with either of the bootstrap-*t* approaches.

Bootstrap-*t* based intervals are not transformation respecting.

However the Percentile intervals are, assuming of course you aren't doing something stupid with your estimation procedures.

There are other bootstrapping approaches to confidence intervals.

The most common 2 are

- BC_a: Bias-corrected and accelerated
- ABC: Approximate bootstrap confidence.

ABC is an approximation to BC_a which reduces the number of bootstrap samples.

These two approaches tend to give better intervals than those discussed earlier.

They are both transformation respecting and tend to have more accurate coverage probabilities.