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 $F(x)$ 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 $\mathbf{x}$.

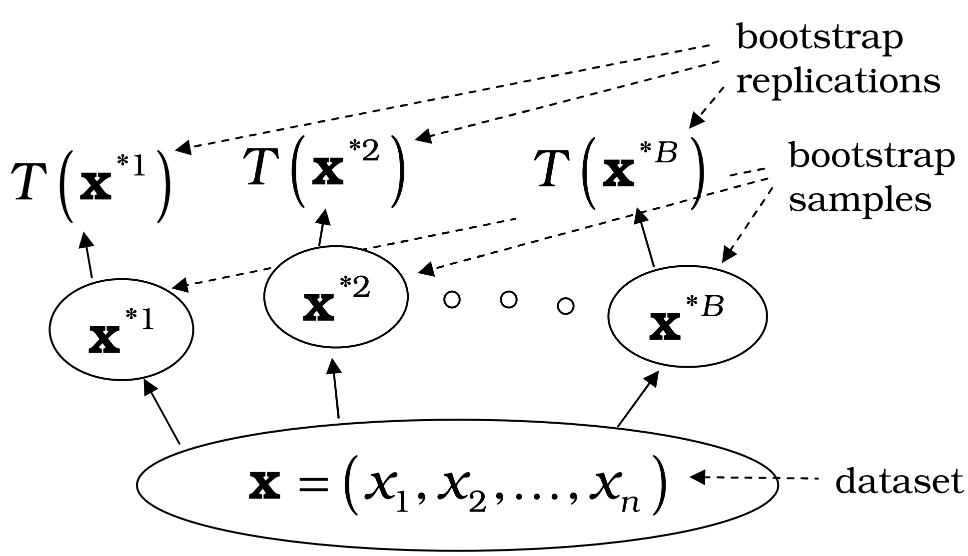
2) Evaluate the bootstrap replication corresponding to each bootstrap sample,

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

3) Evaluate the standard error $se(T)^*$ by

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

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



Example: 1973 Law School Admissions

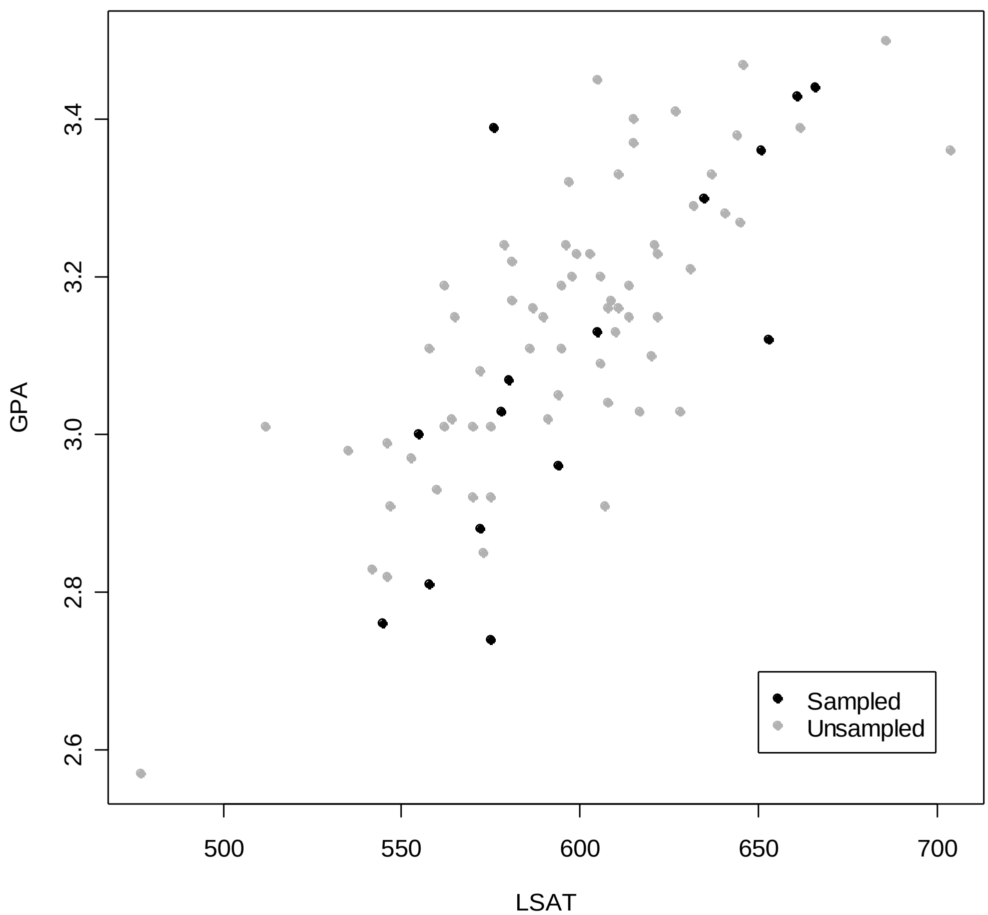
82 Law Schools

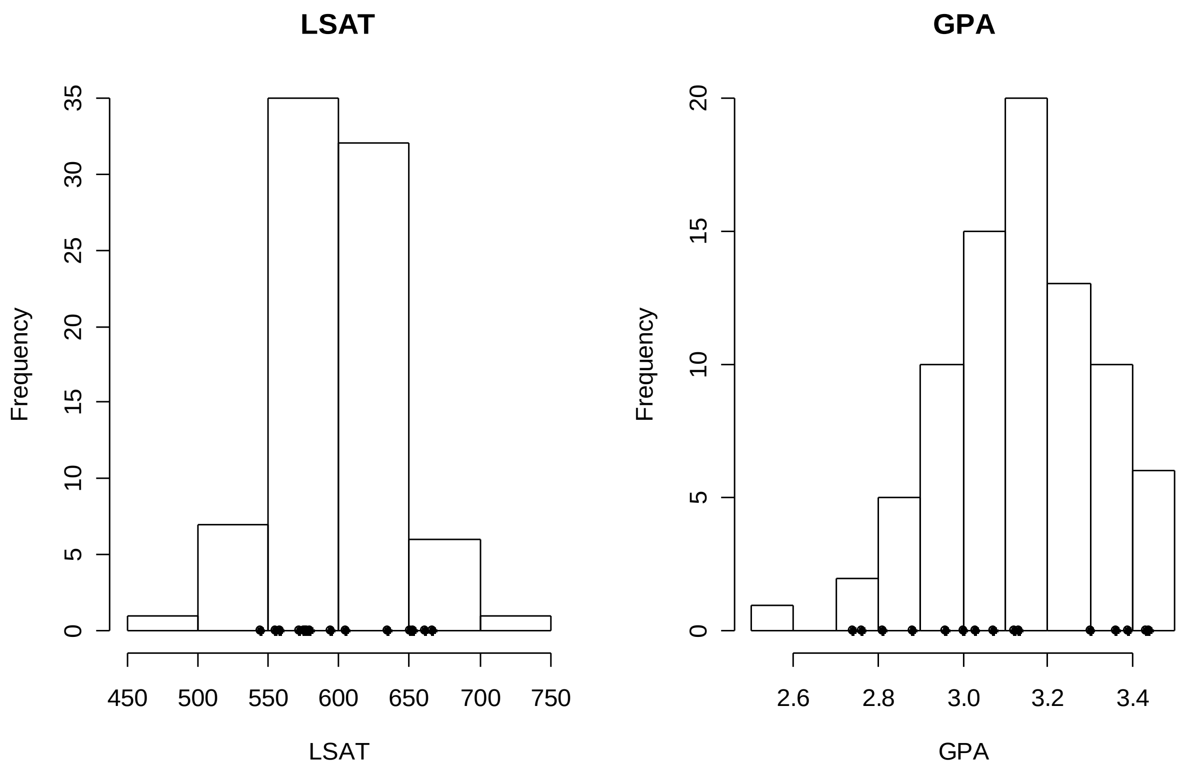
2 Variables:

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

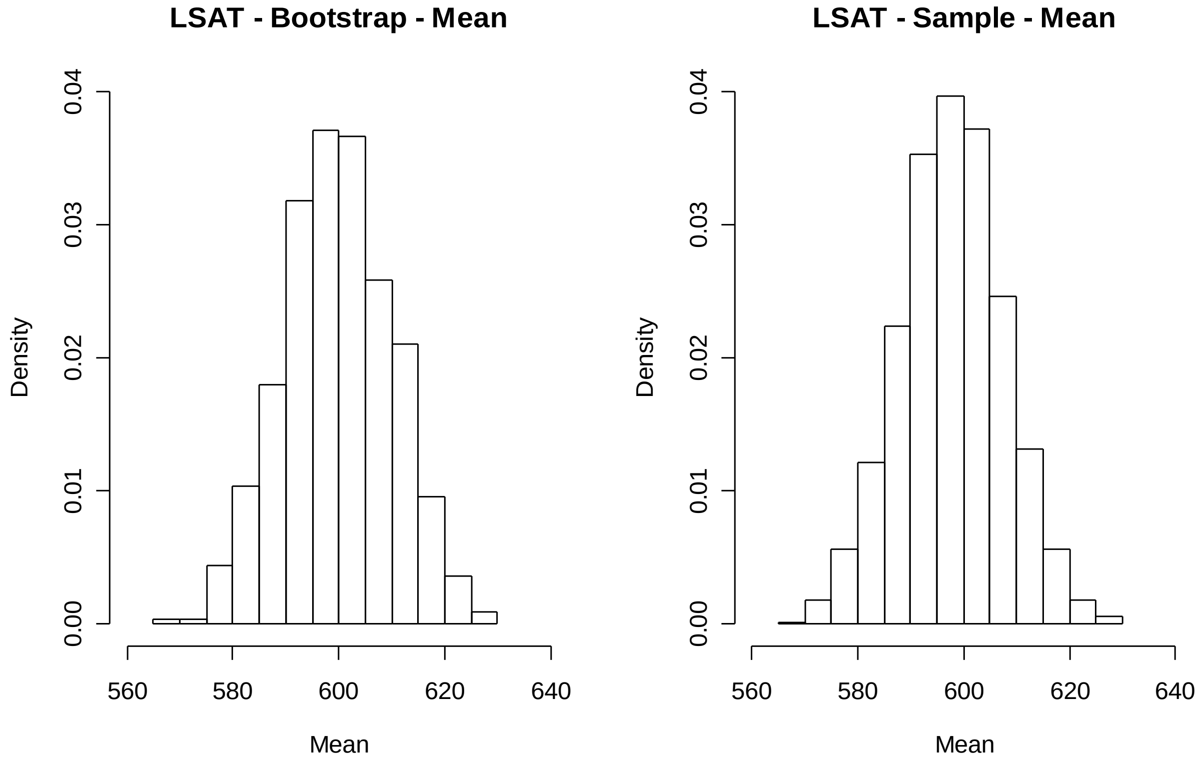
$$\rho(\text{LSAT}, \text{GPA}) = 0.760$$

Parameter estimates from sample (n = 15)

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

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

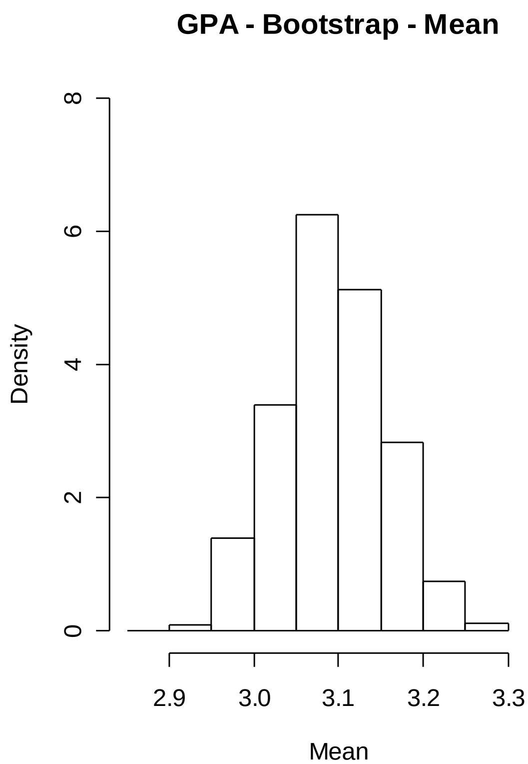
In what follows, the histograms are based on 1000 samples. For the "Sample" histograms, samples of 15 observations (without replacement) from the 82 law schools in the population.

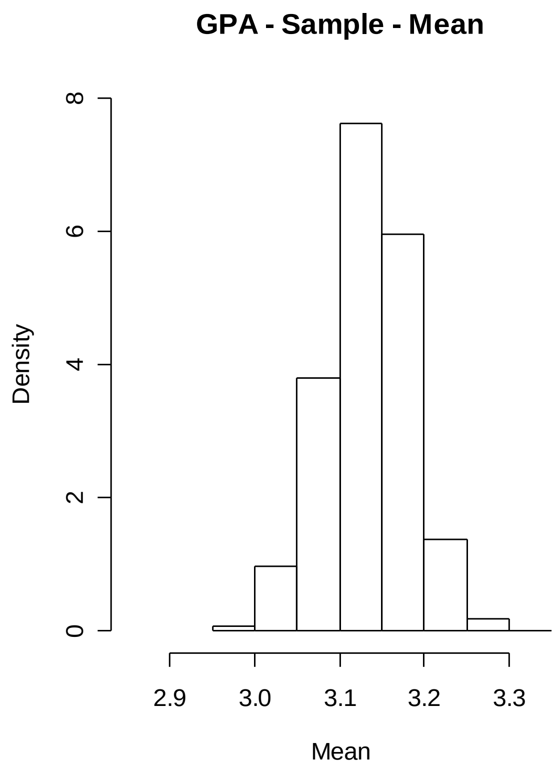


$$se(\bar{x}) = 9.877$$

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

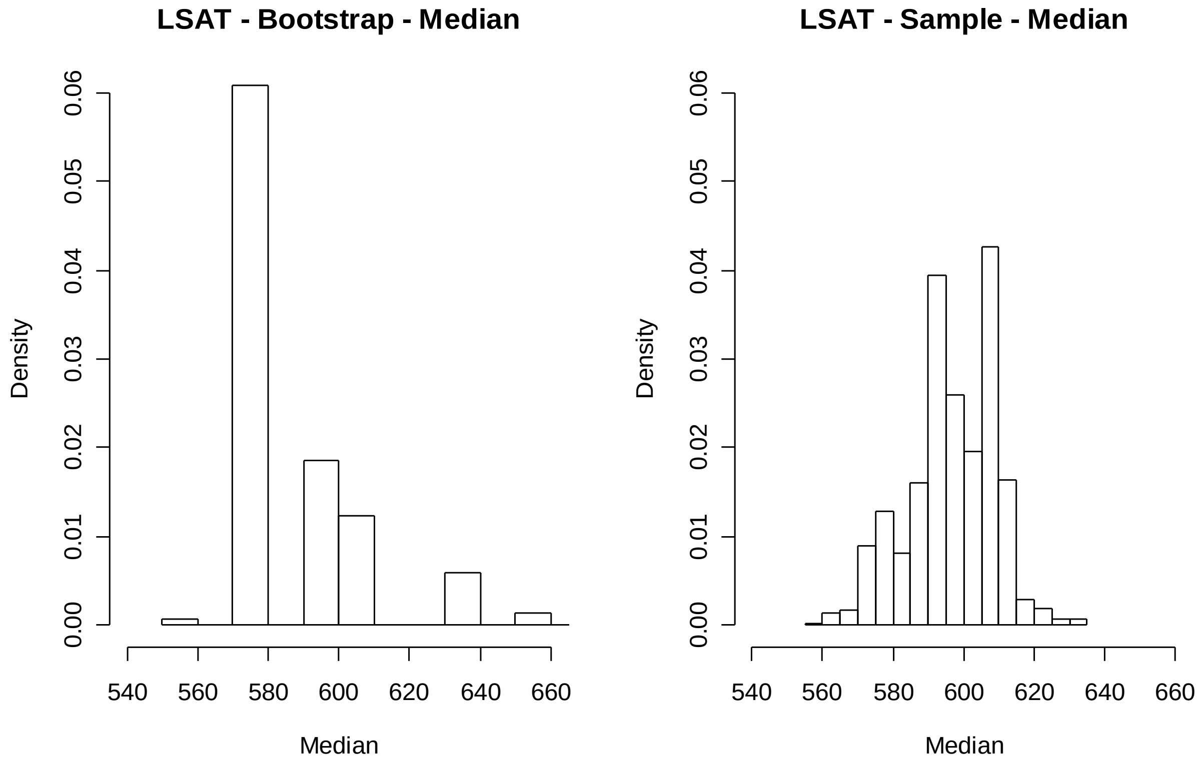




$$se(\bar{x}) = 0.0486$$

Bootstrap estimates

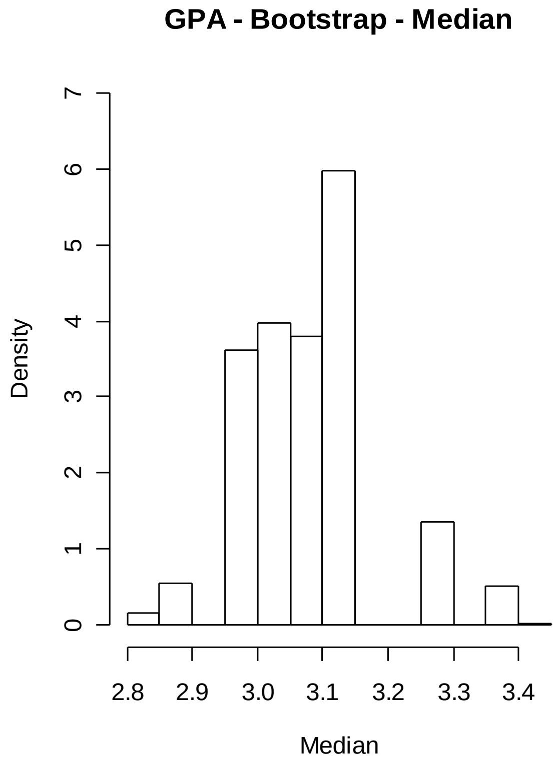
B	50	100	150	200
$se(\bar{x})$	0.0571	0.0575	0.0591	0.0622
B	250	500	750	1000
$se(\bar{x})$	0.0618	0.0632	0.0624	0.0632

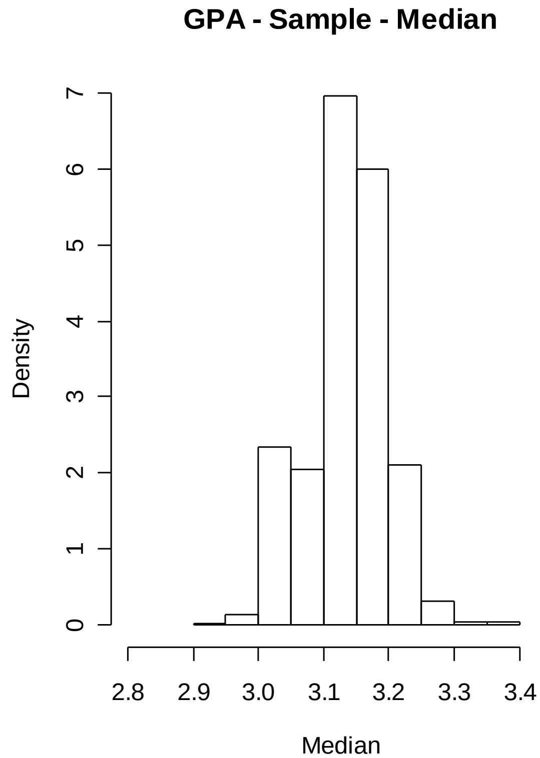


$$se(\text{Med}) = 12.384 \text{ (by Monte Carlo)}$$

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

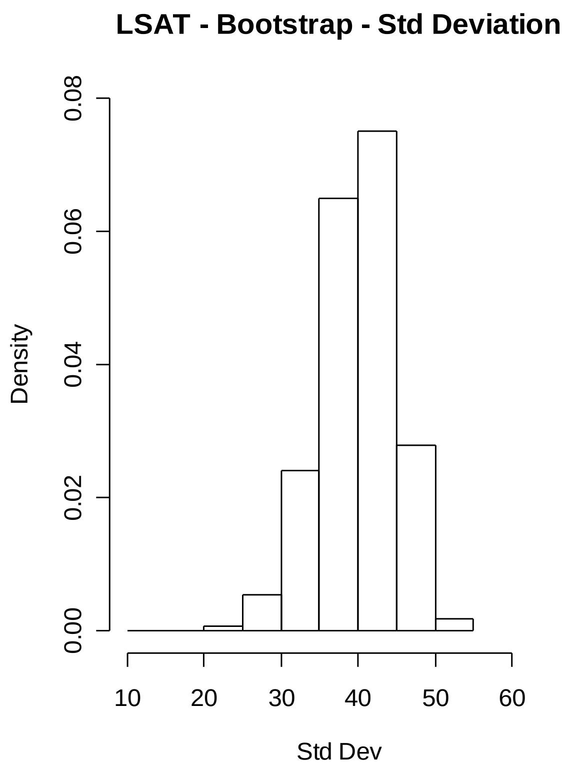


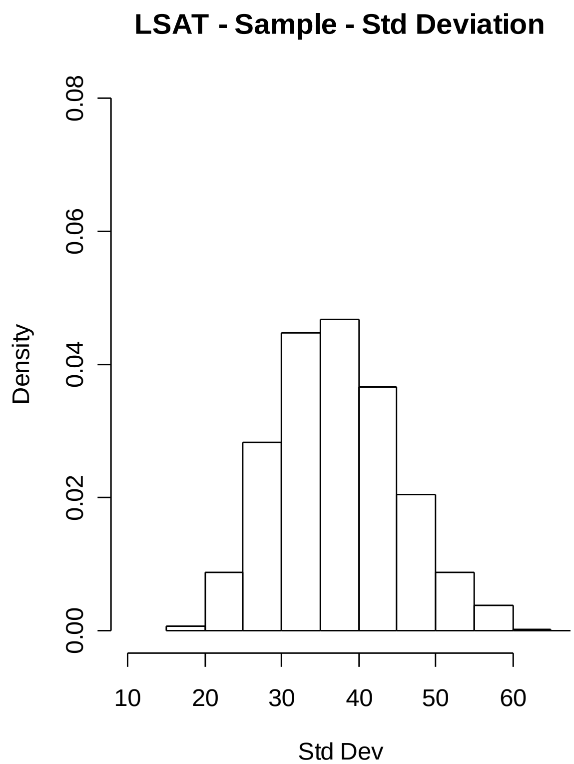


$se(\text{Med}) = 0.0619$ (by Monte Carlo)

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

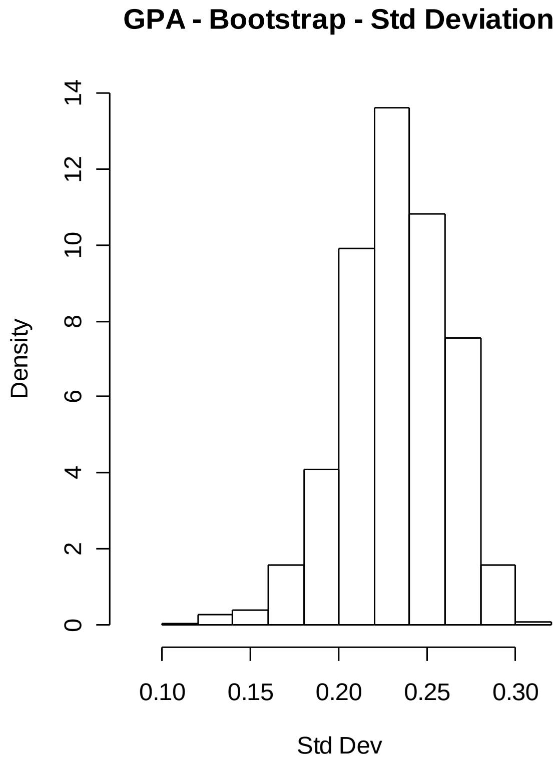


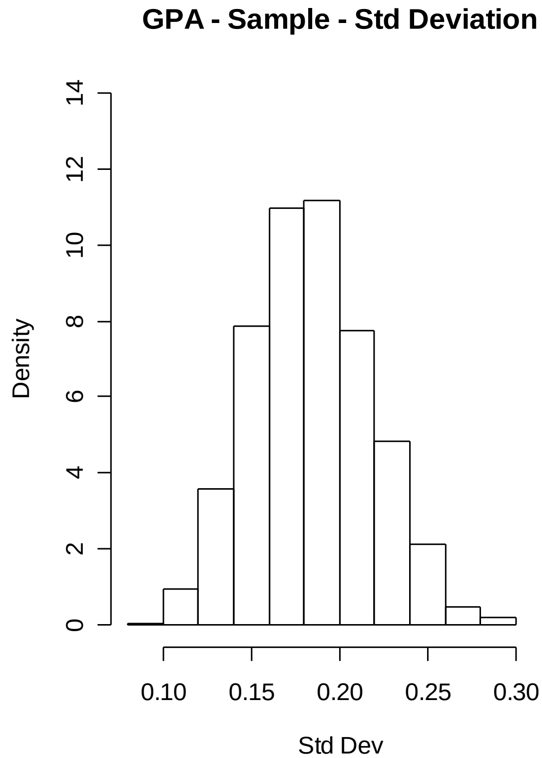


$se(s) = 8.025$ (by Monte Carlo)

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

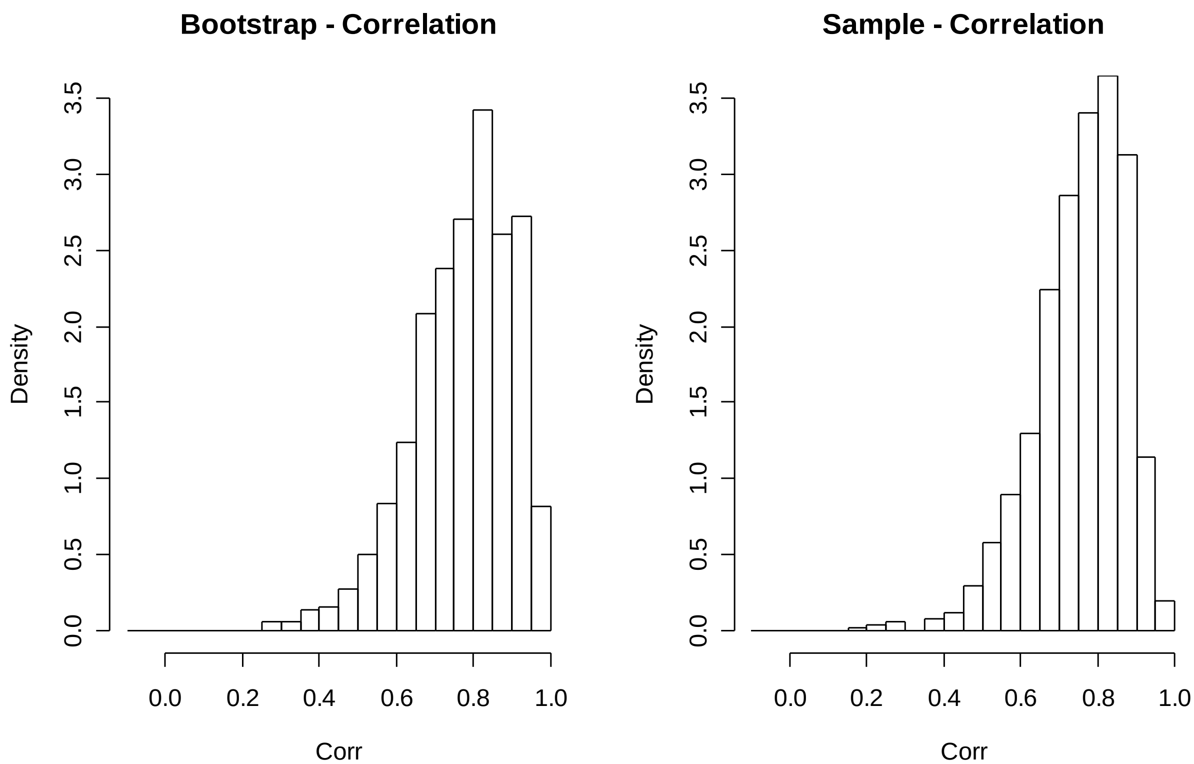




$se(s) = 0.0339$ (by Monte Carlo)

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



$$se(r) = 0.118 \text{ (by Monte Carlo)}$$

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

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

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

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

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

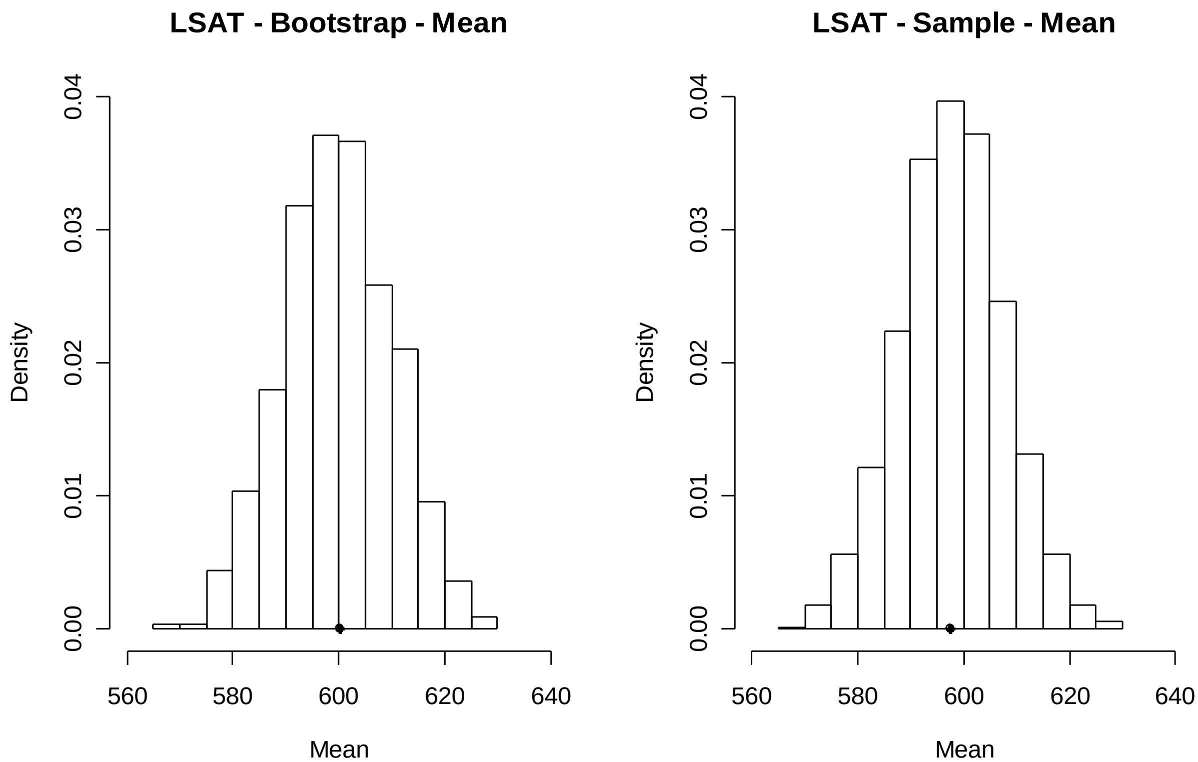
In fact we needed to estimate this as part of the standard error calculation as

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

is an estimate of this quantity.

Therefore

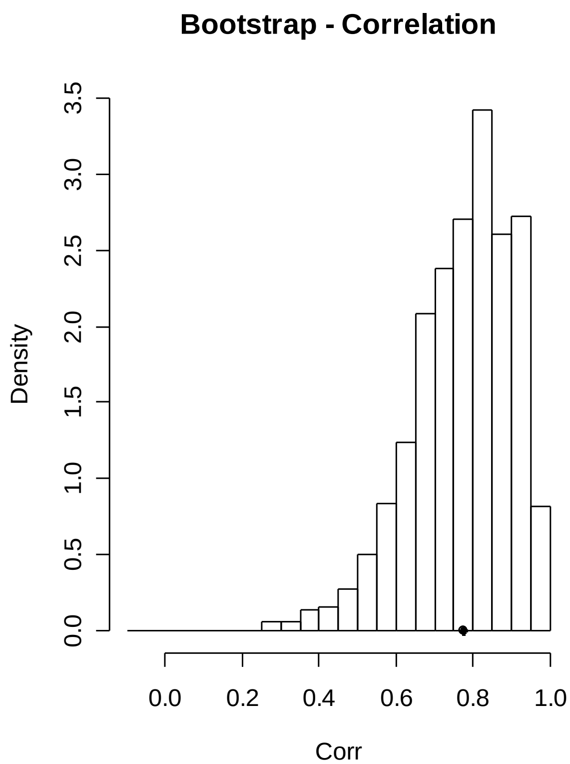
$$\begin{aligned} \widehat{Bias}_B &= \hat{E}[T]^* - t\left(F_n^*\right) \\ &= \hat{E}[T]^* - T(\mathbf{x}) \end{aligned}$$

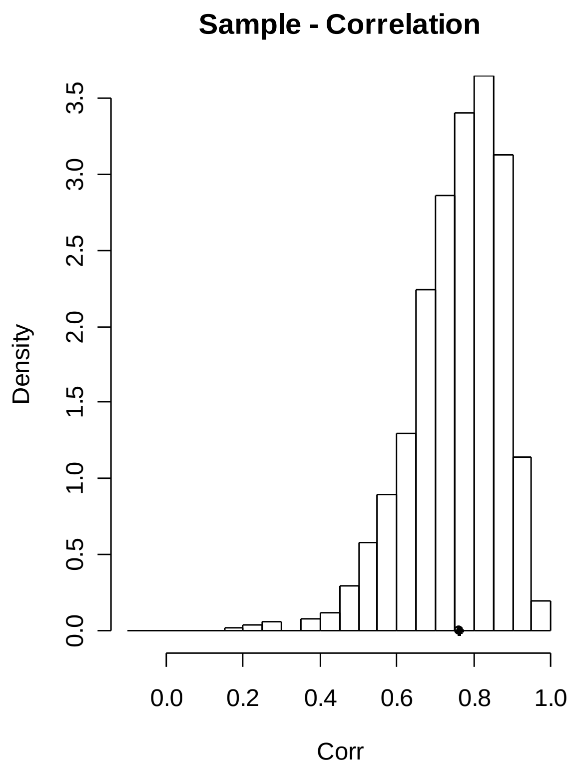


$$Bias(\bar{x}) = 0$$

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