Week 6: Bootstrap

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-02-12

Announcements

- Anonymous mid-semester feedback survey: https://forms.office.com/r/r6kXw2GzTD
- Study Halls:
  - February 20th
  - February 25th
- Midterm: February 27th

The Bootstrap

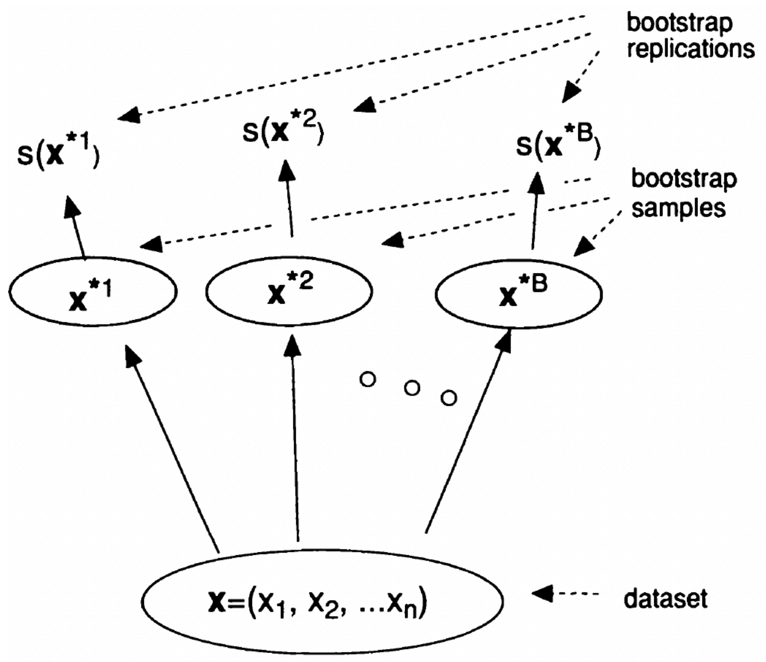
Bootstrap Definition

Let $X_1, \ldots, X_n \sim F$ be a random sample.

A *bootstrap sample* $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ is obtained by sampling n times, *with replacement*, form the original sample $X_1, \ldots, X_n$. This implies that

$$X_1^*, \ldots, X_n^* \sim \hat{F}_n.$$

This is repeated B times to get B surrogate samples $X^{*1}, \ldots, X^{*B}$ from the distribution $\hat{F}_n$.



Bootstraping the sample average

Consider the sample average statistic $\bar{X}_n$, from which we can define the quantity $X_n - \mu$. How does this quantity deviate? To answer this question, we can use the bootstrap.

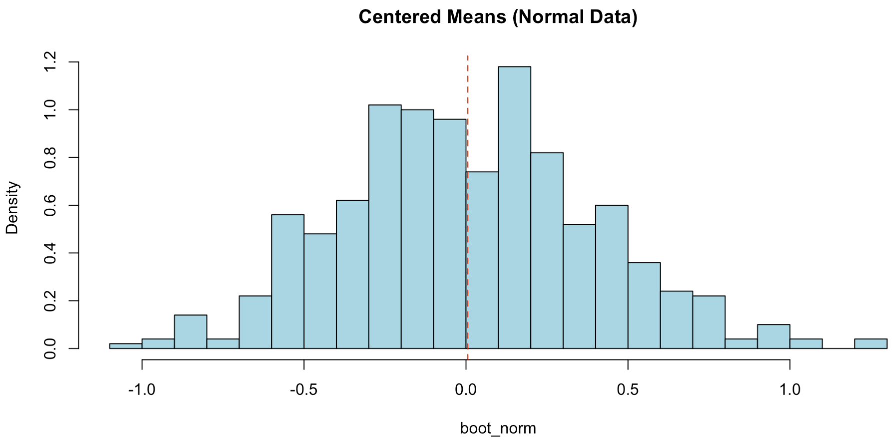
1. Sample n observations $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ from $\hat{F}_n$.
2. Compute $\bar{x}^* - \mu^*$.

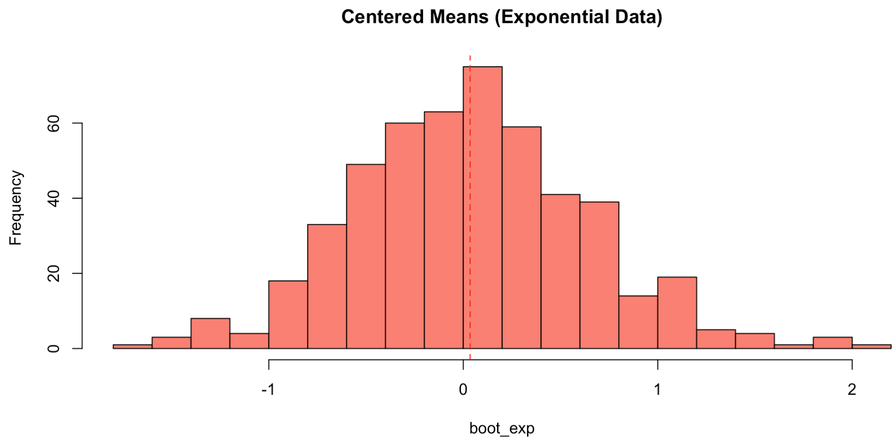
$\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$, $\mu^* = \mathbb{E}_{\hat{F}_n}[X_1^*] = \bar{X}_n$

Repeat the previous two steps B many times. Note that we aren't using any information about the original distribution F that we don't already have in $\hat{F}_n$.

Bootstraping the sample average: Examples

Here are two examples of what the *bootstrap distribution* of $\bar{X} - \mu$ looks like when F is a normal distribution $\mathcal{N}(10, 3^2)$ or an exponential distribution $\mathrm{Exp}(0.2)$. This simulation was performed on samples of size $n = 50$ and the number of bootstrap repetitions used was $B = 500$.





$\bar{X}^* - \mu^*$

Example: centered median

We will now use the bootstrap procedure taking as our statistic $s(\cdot)$ the sample median $M_n = \mathrm{Med}(x_1, \ldots, x_n)$. $\hat{F}_n$ is defined by $x_1, \ldots, x_n$

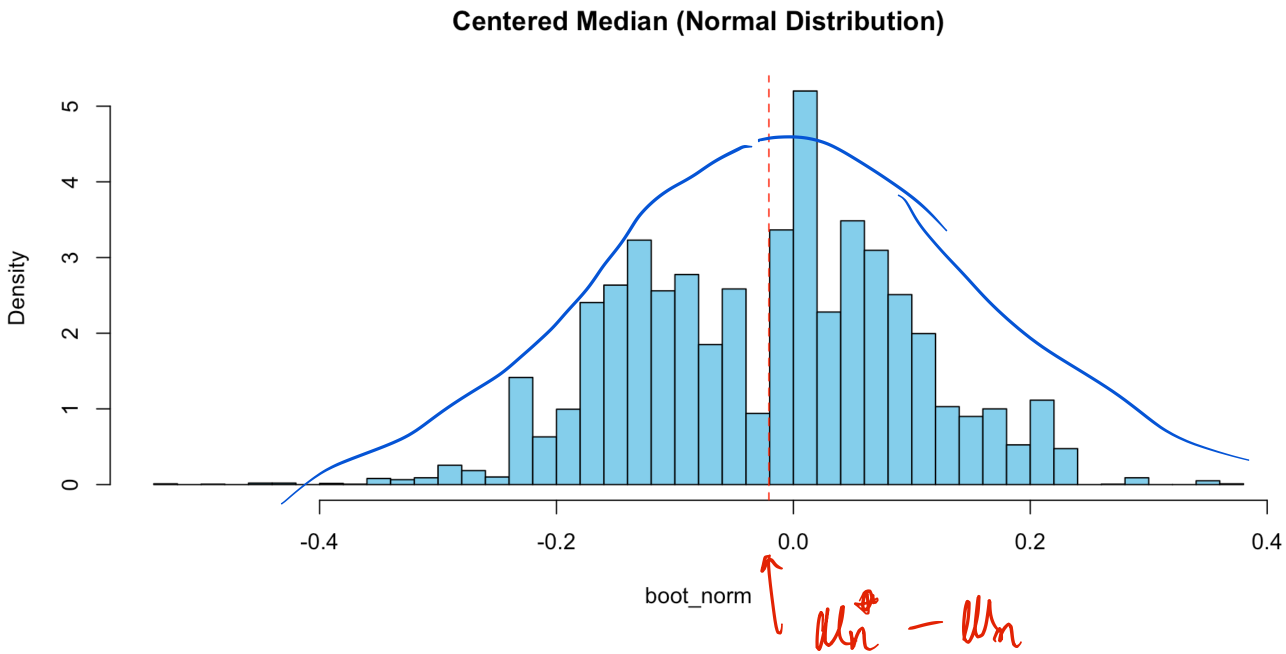
1. Sample n observations from the data with replacement to get x^*. $\longrightarrow \hat{F}_n$
2. Bootstrap Statistic: Calculate $M_n^* = \mathrm{Med}(x^*)$
3. Center: Compute the difference $\delta^* = M_n^* - M_n$.

Repeat: Do this B times to form the empirical distribution of δ^*.

Example: centered median

Simulation showing the bootstrap distribution of $\delta^* = M_n^* - M_n$ when the underlying distribution is standard normal $\mathcal{N}(0, 1)$. The simulation used a sample size $n = 100$ and $B = 10000$ bootstrap repetitions.

$\in \{x_1, \dots, x_n\}$



Bootstrap estimate for the standard error of the sample mean

One of the most important examples of a sample statistic to estimate with the bootstrap is the **standard error**.

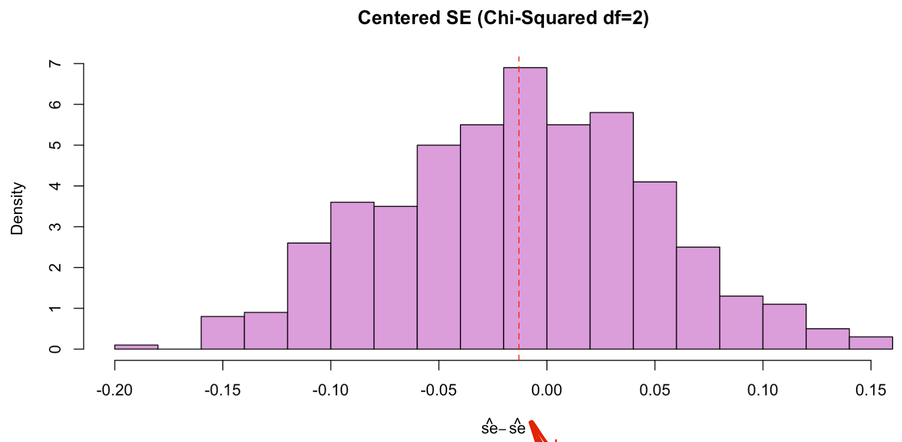
For an estimator $\hat{\theta} = s(X_1, \ldots, X_n)$, we wish to assign a standard error se to $\hat{\theta}$ $\implies$ estimate $\mathrm{sd}(\hat{\theta})$. For the sample mean, if $\sigma = \sqrt{\mathbf{Var}(X_i)}$, then the standard error is given by $\mathrm{se}(\bar{x}) = \sqrt{\sigma^2/n}$. $= \sigma/\sqrt{n}$ σ standard deviation

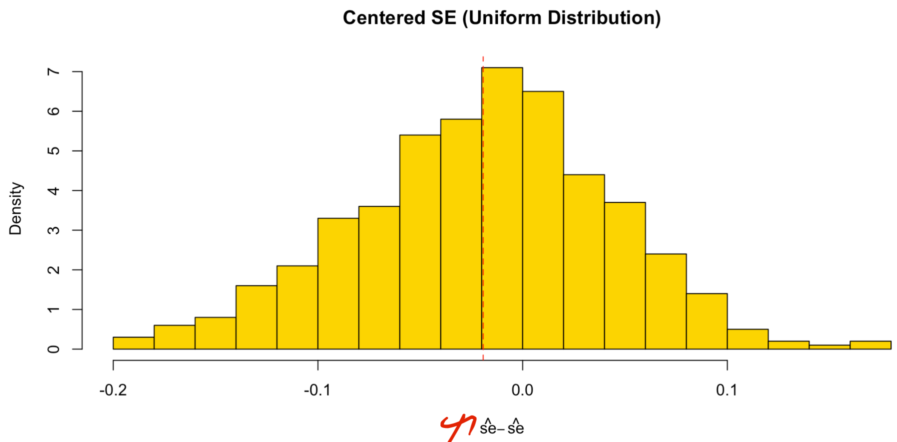
> We don't really need the bootstrap for this statistic in the case of the sample mean. (Why?)

how to estimate $se(\bar{x})$? $\longrightarrow$ $\widehat{se}(\bar{x}) = \hat{\sigma}/\sqrt{n}$, $\hat{\sigma} = \sqrt{s^2}$

Example: Bootstrap for the standard error of the sample mean

Simulation showing the bootstrap distribution of $\hat{se}$ when the distribution of the original samples is chi squared χ^2_2 or uniform $\mathrm{Unif}(0, 10)$. The simulation used a sample size $n = 40$ and $B = 500$ bootstrap repetitions.



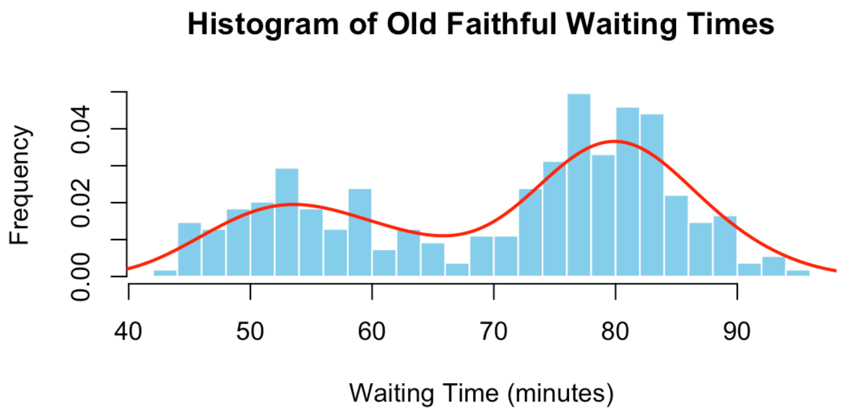


$\hat{se}^* - \hat{se}$

well-centered indicates the sampling distribution of the $\hat{se}^*$ bootstrap has expectation $\hat{se}$.

Example: Old Faithful dataset

Recall the dataset we

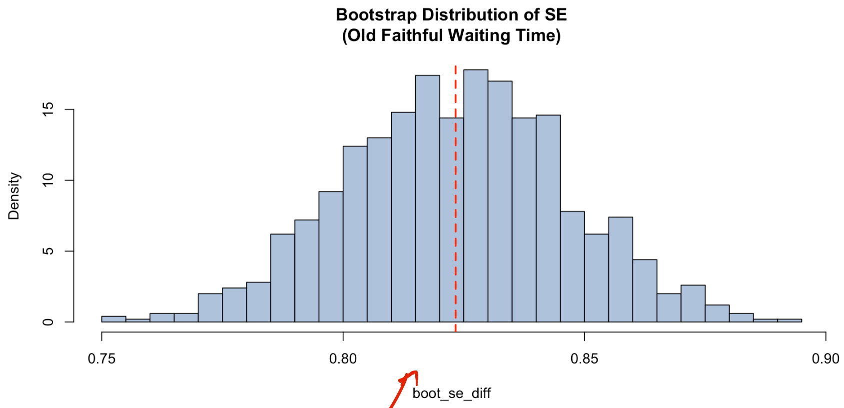


Old Faithful geyser in Yellowstone National Park

Example: Old Faithful dataset

```
data(faithful)
waiting_data <- faithful$waiting
n <- length(waiting_data)

# Define the sample statistic (Standard Error of the mean)
calc_se <- function(x) sd(x) / sqrt(length(x))
se_n <- calc_se(waiting_data)

B <- 1000
boot_se_diff <- replicate(B, {
  x_star <- sample(waiting_data, size = n, replace = TRUE)
  se_star <- calc_se(x_star)
  return(se_star) # Centering: SE* - SE
})

hist(boot_se_diff, breaks = 20, col = "lightsteelblue", prob = TRUE,
     main = "Bootstrap Distribution of SE\n(Old Faithful Waiting Time)")
abline(v = mean(boot se diff), col = "red", lwd = 2, lty = 2)
```

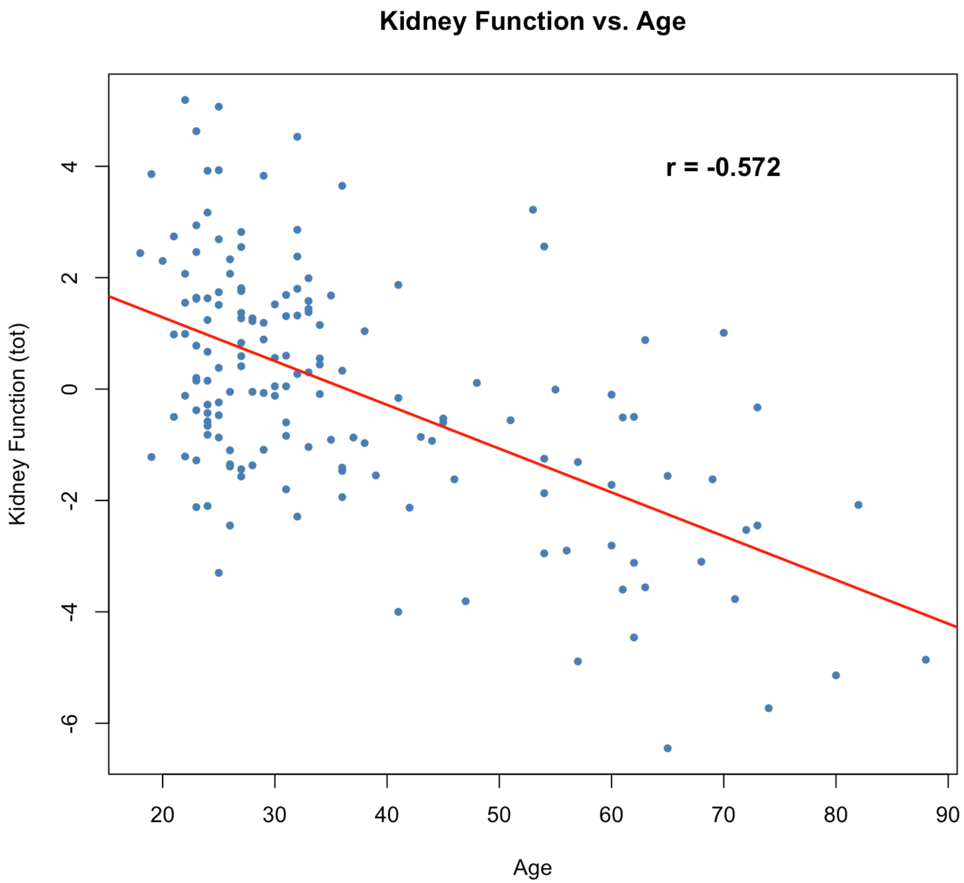
Bootstrap simulation of size $B = 1000$.




```
[1] 0.8232931
```

0.82

Example: Correlation coefficient



The following scatterplot from Efron and Hastie (2014) shows the age against a total measure of kidney function for $n = 157$ healthy individuals.

Can we get a sense for how accurate the value of the correlation coefficient r might be? $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ then

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Bootstrap samples (x_i^*, y_i^*)

$\longrightarrow r(x^*, y^*)$

B many times

Example: Correlation coefficient

```
B <- 1000
set.seed(123) # For reproducibility

boot_corrs <- replicate(B, {
  # Resample the row indices with replacement
  resample_indices <- sample(1:nrow(kidney), replace = TRUE)
  sample_data <- kidney[resample_indices, ]

  # Calculate correlation for this bootstrap sample
  cor(sample_data$age, sample_data$tot)
})

# Calculate the Standard Error (Standard Deviation of the bootstrap
distribution)
se_boot <- sd(boot_corrs)
```

Using the bootstrap we get $\hat{\text{se}}_{\text{boot}} = 0.058$.

The classical Taylor expansion for the standard error, which is given in (Efron and Hastie, 2014) would be

$$\widehat{\text{se}}_{\text{taylor}} = \left\{ \frac{\hat{\theta}^2}{4n} \left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right] \right\}^{1/2}$$

where

$$\hat{\mu}_{hk} = \sum_{i=1}^{n} (x_i - \bar{x})^h (y_i - \bar{y})^k / n.$$

This gives us $\hat{\text{se}}_{\text{taylor}} = 0.057$.

almost the same with $B = 1000$

Parametric variation

$$\hat{F}_n \longrightarrow \boldsymbol{x}^* \longrightarrow \hat{\theta}^*$$

Suppose we assume that F comes from a parametric model $F = F_\mu$. If we assume $x \sim F_\mu$, then we know the form of $\hat{F}_\mu$.

Example:

Suppose that we have random realizations $x_1, \ldots, x_n \sim \mathcal{N}(\mu, 1)$ and we take $\hat{\mu} = \bar{x}$. Then, the bootstrap samples would follow $x_i^* \sim \mathcal{N}(\bar{x}, 1)$.

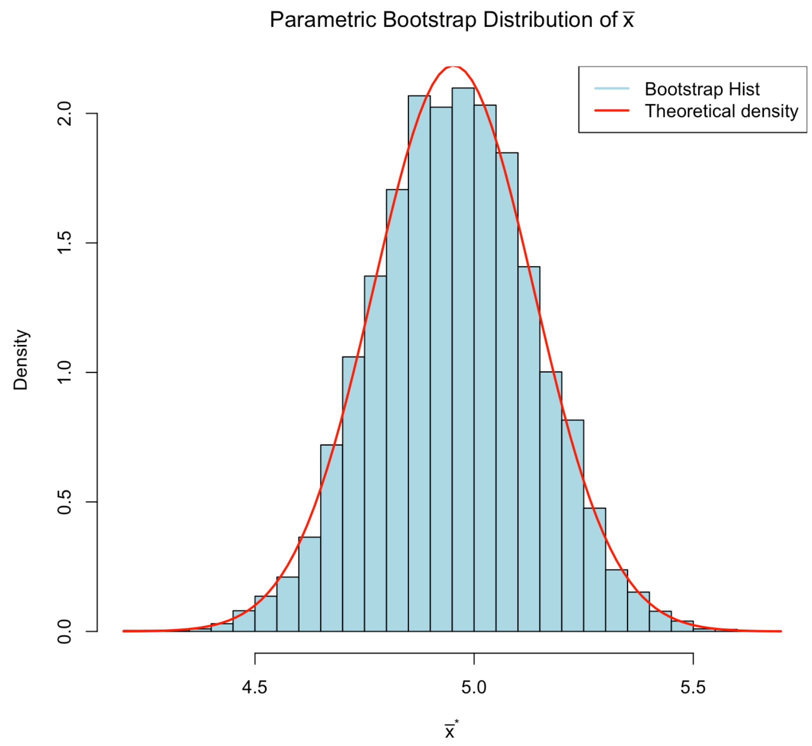
Assume $x_i \sim \text{Pois}(\lambda)$, create bootstrap samples $x_i^* \sim \text{Pois}(\bar{x})$

This allows us to combine the bootstrap with the modelling assumptions that we like.

Parametric variation

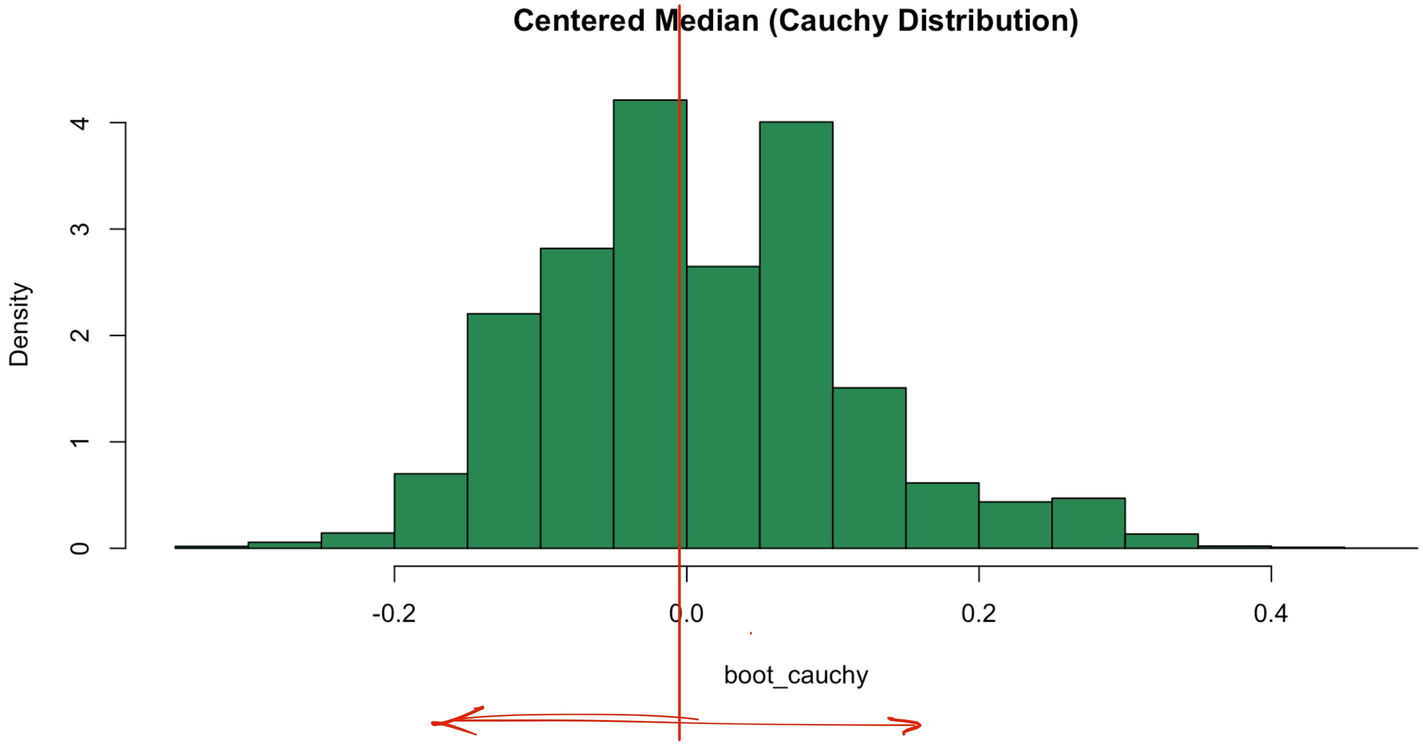
$x_i \sim N(\bar{x}, 1)$

```
n <- 30          # Sample size
mu_true <- 5     # True population mean
B <- 10000       # Number of bootstrap replicates

# Generate the Observed Data
x_obs <- rnorm(n, mean = mu_true, sd = 1)
x_bar_obs <- mean(x_obs)

# The Parametric Bootstrap: we generate new samples from N(x_bar_obs, 1)
boot_means <- replicate(B, {
  x_star <- rnorm(n, mean = x_bar_obs, sd = 1)
  mean(x_star)
})

hist(boot_means, breaks = 50, col = "lightblue", probability = TRUE,
     main = bquote("Parametric Bootstrap Distribution of" ~ bar(x)),
     xlab = expression(bar(x)^"*"))
```



Bootstrap with heavy tails

Another simulation from $n = 20$ samples of a $\mathrm{Cauchy}(0, 1)$ distribution (which has heavy tails). We can use the bootstrap (in this case $B = 1000$) to analyze the median.



Bootstrap of robust statistics

Why is so much of classical statistical theory centered around the sample mean?

easy to analyze mathematically

Consider our discussion around the trimmed mean.

hard to derive CLT-like results

Robust statistics are harder to analyze mathematically.

median, MAD, etc.

CLT only works for sums of iid r.v.s, i.e. $\bar{X}_n$

Remarks

- The bootstrap is completely automatic. A master algorithm can be written that inputs the data x and the function $s(\cdot)$, and outputs $\hat{\text{se}}_{\text{boot}}$.
- The bootstrap is more dependable than the jackknife or the CLT for statistics that are not smooth, e.g. quantiles.
- $B \in [200, 500]$ is usually sufficient for evaluating $\hat{\text{se}}_{\text{boot}}$. Still, larger values such as $B = 1000$ or $B = 2000$, will be required for the bootstrap confidence intervals (to be seen in future lectures).
- There is nothing special about standard errors. We could just as well use the bootstrap replications to estimate the expected absolute error $\mathbb{E}\left[\left|\hat{\theta} - \theta\right|\right]$, or any other accuracy measure.

$$\frac{1}{n}\sum_{b=1}^{B}\left|\hat{\theta}_b^* - \hat{\theta}\right|$$