Bootstrap Confidence Intervals

Key ideas from last class

Algorithm:

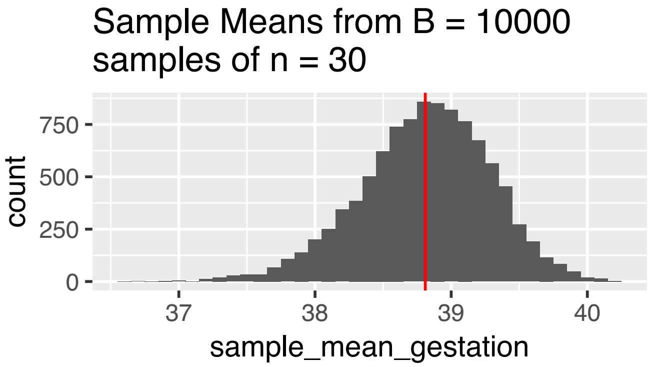
1. For $b = 1, \dots, B$:
   a. Draw a bootstrap sample of size n **with replacement** from the observed data
   b. Calculate the estimate $\hat{\theta}_b$ based on that bootstrap sample (a number)
2. The distribution of estimates $\{\hat{\theta}_1, \dots, \hat{\theta}_B\}$ from different bootstrap samples approximates the sampling distribution of the estimator $\hat{\Theta}$ (the random variable).

Notation:

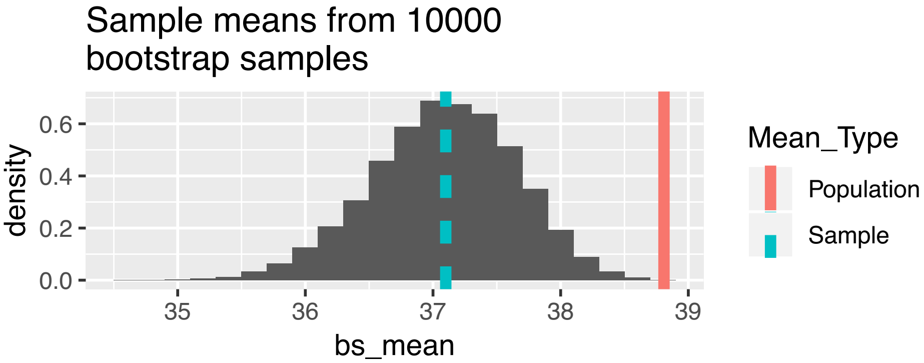
- θ is the unknown parameter to estimate
- $\hat{\theta}$ is the estimate from our sample
- $\hat{\theta}_b$ is the estimate from bootstrap resample number b

Compare the approximations from sampling directly from the population and from bootstrap resampling:

Many means, based on samples from the population:



Many means, based on bootstrap resamples with replacement from the sample:



- The relationship of $\hat{\theta}$ to θ is like the relationship of $\hat{\theta}_b$ to $\hat{\theta}$:
  - Bootstrap distribution **reproduces shape and bias** of actual sampling distribution
  - Bootstrap distribution **does not reproduce mean** of actual sampling distribution
    - E.g., centered at sample mean instead of population mean

Bootstrap applied to pivotal quantity

- We could also do all of the above with a pivotal quantity instead:
    - T is the pivotal quantity (a random variable) → for example $T = \frac{\hat{\Theta} - \Theta}{\widehat{SE}(\hat{\Theta})}$

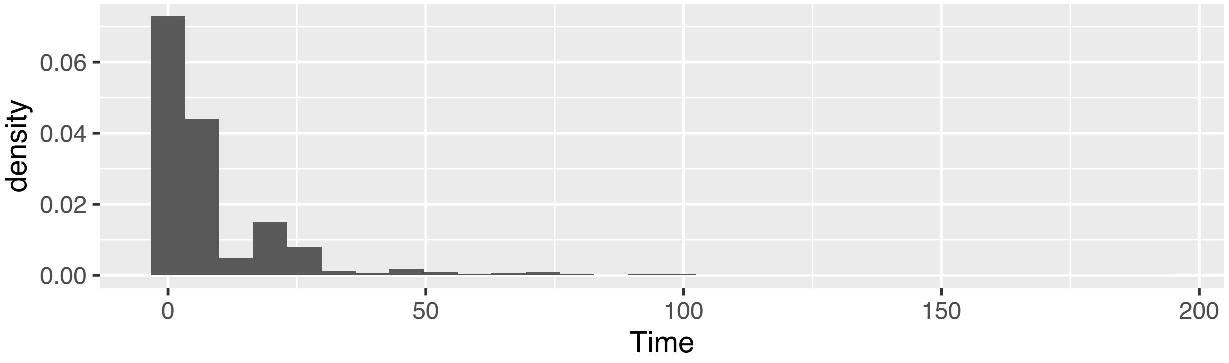
      $\hat{\Theta} = \bar{X}$, $\widehat{SE}(\hat{\Theta}) = \left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right]^{1/2}$
    - t is the value of the pivotal quantity based on our observed sample
    - t_b is the value of the pivotal quantity based on bootstrap sample number b
    - The distribution of $\{t_1, \dots, t_B\}$ approximates the distribution of the pivotal quantity.
        * Since the distribution of a pivotal quantity does not depend on any unknown parameters (like θ), the bootstrap approximation to the disribution of T will generally be located in the correct place.
        * For example, the mean of the t pivotal quantity $T = \frac{\bar{X}-\mu}{SE(\bar{X})}$ is always 0 since $\bar{X}$ is an unbiased estimator of μ. The distribution of T does not depend on μ.

↳ actual sampling dist'n and the bootstrap approx. to sampling dist'n will both be centered at Θ.

Running Example: Verizon Repair Times

This example is taken from Hesterberg (2014). We have data on the amount of time it took Verizon to repair problems in their telephone lines in the state of New York. They were brought to court, accused of taking longer than their contractual obligation to do repairs. For legal reasons, there was interest in estimating the mean repair time.

```r
library(readr)
library(dplyr)
library(ggplot2)

verizon <- read_csv("http://www.evanlray.com/data/chihara_hesterberg/Verizon.csv")
verizon_ilec <- verizon %>%
  filter(Group == "ILEC")

ggplot(data = verizon_ilec, mapping = aes(x = Time)) +
  geom_histogram(mapping = aes(y = ..density..))
```



Based on sample data, the estimate is $\hat{\theta} = 8.412$

```r
mean(verizon_ilec$Time)
```

```
## [1] 8.411611
```

Two approaches to bootstrap confidence intervals (there are many others)

1. bootstrap percentile
2. bootstrap t
    a. with standard error from formulas
    b. using nested bootstrap to estimate standard error

$T = \frac{\hat{\Theta} - \Theta}{\widehat{SE}(\hat{\Theta})}$

Confidence Interval Idea #1: Bootstrap Percentile CI

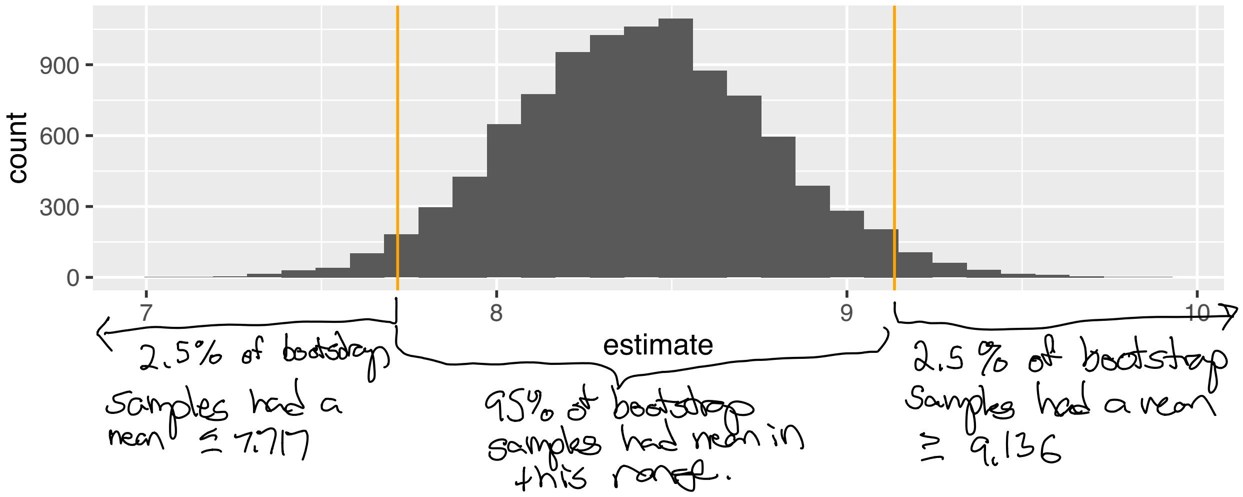
Take percentiles of the bootstrap approximation to the sampling distribution of the estimator $\hat{\Theta}$

- If you have seen bootstrap confidence intervals before, this is probably what you have seen.
- Essentially unjustifiable unless sampling distribution is symmetric
- Regardless of no great formal justification, tends to work well for moderate to large sample sizes
- For small n, actual coverage rate $\neq$ nominal coverage rate

```
# sample size
n <- nrow(verizon_ilec)

# how many bootstrap samples to take, and storage space for the results
num_samples <- 10^4
bs_percentile_results <- data.frame(
  estimate = rep(NA, num_samples)
)

# draw many samples from the observed data and calculate mean of each simulated sample
for(i in seq_len(num_samples)) {
  ## Draw a bootstrap sample of size n with replacement from the observed data
  bs_sample <- verizon_ilec %>%
    sample_n(size = n, replace = TRUE)

  ## Calculate mean of bootstrap sample
  bs_percentile_results$estimate[i] <- mean(bs_sample$Time)
}

# 95% Bootstrap Percentile Interval
bs_percentile_interval <- quantile(bs_percentile_results$estimate, prob = c(0.025, 0.975))
bs_percentile_interval
```

```
##     2.5%    97.5%
## 7.717432 9.135773
```

```
# Plot
ggplot(data = bs_percentile_results, mapping = aes(x = estimate)) +
  geom_histogram() +
  geom_vline(xintercept = bs_percentile_interval, color = "orange")
```

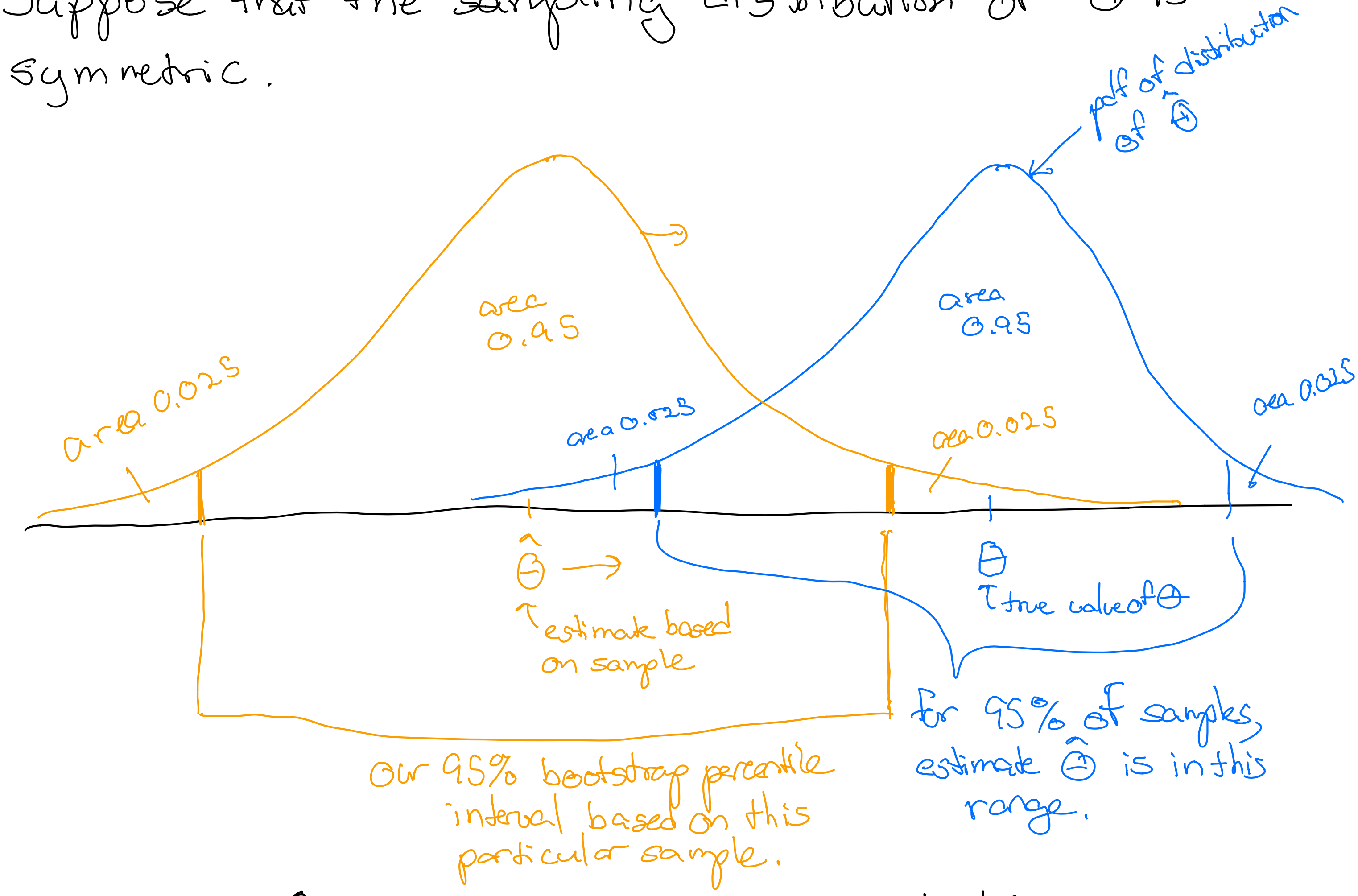


2.5% of bootstrap samples had a mean $\leq$ 7.717

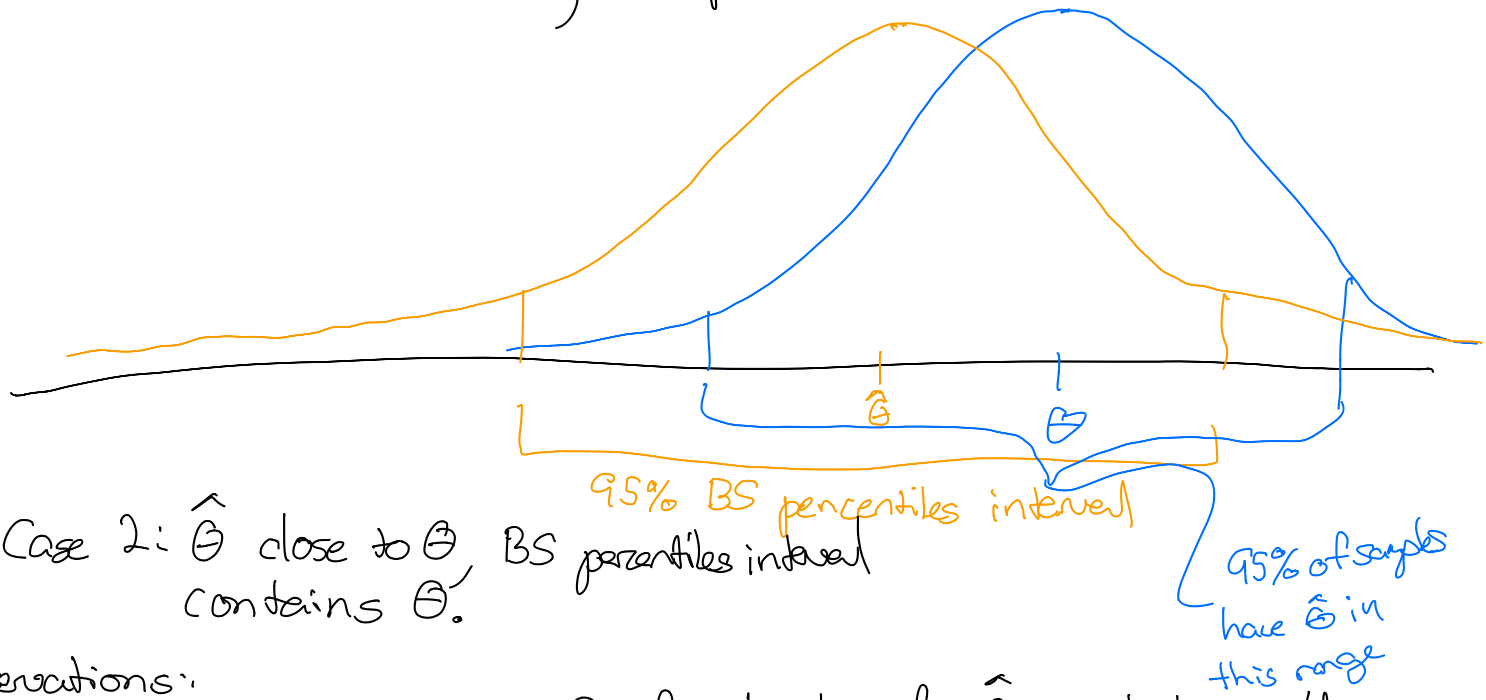
95% of bootstrap samples had mean in this range.

2.5 % of bootstrap samples had a mean $\geq$ 9.136

Suppose that the sampling distribution of $\hat{\Theta}$ is symmetric.


pdf of distribution of $\hat{\Theta}$
area 0.025
area 0.95
area 0.025
area 0.025
area 0.95
area 0.025
$\hat{\Theta}$
estimate based on sample
Θ
true value of Θ
for 95% of samples, estimate $\hat{\Theta}$ is in this range.
Our 95% bootstrap percentile interval based on this particular sample.

Case 1: $\hat{\Theta}$ far from Θ, BS percentile interval did not contain Θ.


$\hat{\Theta}$
Θ
95% BS percentiles interval
95% of samples have $\hat{\Theta}$ in this range

Case 2: $\hat{\Theta}$ close to Θ, BS percentiles interval contains Θ.

Observations:

- percentiles CI contains Θ if and only if $\hat{\Theta}$ is between the 2.5th and 97.5th percentiles of the actual sampling distribution.
- by definition, this happens for 95% of samples.
- for 95% of samples, BS percentiles interval contains Θ (if sampling dist'n symmetric)

Confidence Interval Idea #2: Bootstrap t

- Based on the pivotal quantity $T = \frac{\hat{\Theta} - \theta}{\widehat{SE}(\hat{\Theta})}$.
- If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\theta, \sigma^2)$ then the exact distribution of T is $T \sim t_{n-1}$.
- If n is large and regularity conditions are satisfied, the approximate distribution of T is $T \sim \text{Normal}(0, 1)$
- Otherwise, approximate the sampling distribution of the t statistic using the bootstrap
- How to translate T to bootstrap?

$$T = \frac{\bar{X} - \Theta}{\underline{\widehat{SE}(\bar{X})}} = \frac{\bar{X} - \Theta}{S/\sqrt{n}}$$ where $S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$ is the estimated std. dev. of X_i

and $\boxed{S/\sqrt{n}}$ is estimated std. error of $\hat{\Theta}$.

(in this procedure we were able to calculate $\widehat{SE}(\bar{X}) = S/\sqrt{n}$

Translation to bootstrap:

$\Theta \leftrightarrow \hat{\Theta} \ (= \bar{x})$, $\bar{x} \longleftrightarrow \bar{x}_b$, $S \longleftrightarrow S_b$

Replace Θ (pop. mean) with its sample-based equivalent $\hat{\Theta} = \bar{x}$

Replace $\bar{x}$ (sample mean) with its bootstrap-sample based equivalent $\bar{x}_b$

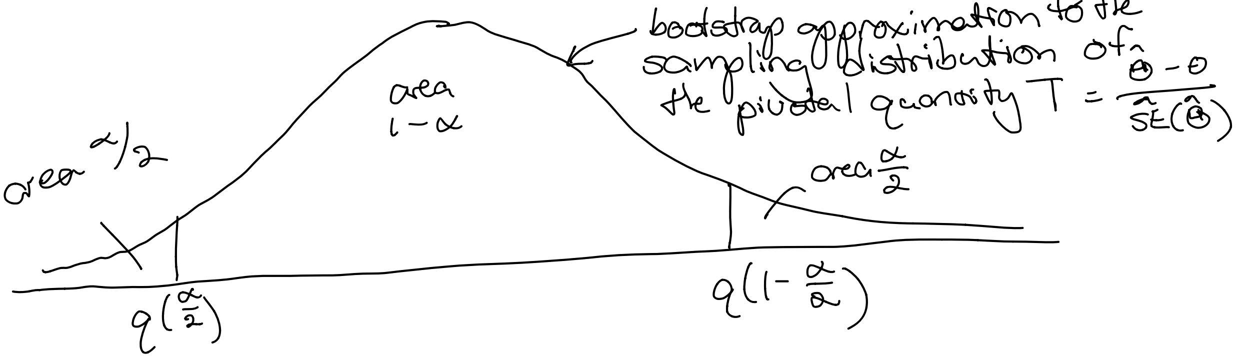
Replace s (sample std. dev.) " " " " " " S_b

For bootstrap sample b, we calculate

$$t_b = \frac{\bar{x}_b - \bar{x}}{\widehat{SE}_b(\bar{X})} = \frac{\bar{x}_b - \bar{x}}{S_b/\sqrt{n}}$$

- Algorithm:

1. For $b = 1, \dots, B$:
   a. Draw a bootstrap sample of size n **with replacement** from the observed data
   b. Calculate $t_b = \frac{\hat{\theta}_b - \hat{\theta}}{\widehat{SE}_b(\hat{\Theta})}$ based on that bootstrap sample
2. The distribution of $\{t_1, \dots, t_B\}$ from different bootstrap samples approximates the sampling distribution of T.
3. Form a confidence interval as $[\hat{\theta} - q(1 - \frac{\alpha}{2})\widehat{SE}(\hat{\Theta}), \hat{\theta} - q(\frac{\alpha}{2})\widehat{SE}(\hat{\Theta})]$
   - $\hat{\theta}$ and $\widehat{SE}(\hat{\Theta})$ are numbers calculated based on original sample data (example: $\hat{\theta} = \bar{x}$, $\widehat{SE}(\hat{\theta}) = s/\sqrt{n}$)
   - $q(a)$ is the a'th quantile of the bootstrap approximation to the distribution of T.


bootstrap approximation to the sampling distribution of the pivotal quantity $T = \frac{\hat{\Theta} - \theta}{\widehat{SE}(\hat{\Theta})}$
area $1-\alpha$
area $\alpha/2$
area $\frac{\alpha}{2}$
$q(\frac{\alpha}{2})$
$q(1-\frac{\alpha}{2})$

Based on bootstrap approx. to distribution of T,

$$P\left(q\left(\frac{\alpha}{2}\right) \le T \le q\left(1-\frac{\alpha}{2}\right)\right) \approx 1-\alpha$$

approximately equal because based on bootstrap approximation to sampling distribution of T not the exact sampling dist'n of T.

$$\Rightarrow P\left(\boxed{q\left(\frac{\alpha}{2}\right)} \le \frac{\boxed{\hat{\Theta}} - \boxed{\theta}}{\boxed{\widehat{SE}(\hat{\Theta})}} \le \boxed{q\left(1-\frac{\alpha}{2}\right)}\right) \approx 1-\alpha$$

random variables

known numbers: quantiles from bootstrap

unknown parameter

$$\Rightarrow P\left(q\left(\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta}) \le \hat{\Theta} - \theta \le q\left(1-\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta})\right) \approx 1-\alpha$$

$$\Rightarrow P\left(-\hat{\Theta} + q\left(\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta}) \le -\theta \le -\hat{\Theta} + q\left(1-\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta})\right) \approx 1-\alpha$$

$$\Rightarrow P\left(\hat{\Theta} - q\left(\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta}) \ge \theta \ge \hat{\Theta} - q\left(1-\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta})\right) \approx 1-\alpha$$

$$\Rightarrow P\left(\underbrace{\hat{\Theta} - q\left(1-\frac{\alpha}{2}\right)\cdot\widehat{SE}(\hat{\Theta})}_{A} \le \theta \le \underbrace{\hat{\Theta} - q\left(\frac{\alpha}{2}\right)\widehat{SE}(\hat{\Theta})}_{B}\right) \approx 1-\alpha$$

An approximate $(1-\alpha)*100\%$ CI for θ is

$$\left[\boxed{\hat{\Theta}} - \underline{q\left(1-\frac{\alpha}{2}\right)}\boxed{\widehat{SE}(\hat{\Theta})}, \boxed{\hat{\Theta}} - \underline{q\left(\frac{\alpha}{2}\right)}\boxed{\widehat{SE}(\hat{\Theta})}\right]$$

random variables

quantiles from bootstrap.

Based on observed data, observed CI is

$$\left[\boxed{\hat{\theta}} - q\left(1-\frac{\alpha}{2}\right)\boxed{\widehat{SE}(\hat{\Theta})}, \boxed{\hat{\theta}} - q\left(\frac{\alpha}{2}\right)\boxed{\widehat{SE}(\hat{\Theta})}\right]$$

number calculated based on observed data.

```r
# sample size
n <- nrow(verizon_ilec)

# how many bootstrap samples to take, and storage space for the results
num_samples <- 10^4
bs_t_results <- data.frame(
  t = rep(NA, num_samples)
)

# draw many samples from the observed data and calculate ~~mean~~ t of each simulated sample
for(i in seq_len(num_samples)) {
  ## Draw a bootstrap sample of size n with replacement from the observed data
  bs_sample <- verizon_ilec %>%
    sample_n(size = n, replace = TRUE)
```

$$t = \frac{\bar{x}_b - \bar{x}}{(s_b/\sqrt{n})}$$

```r
  ## Calculate t statistic based on bootstrap sample
  bs_t_results$t[i] <- (mean(bs_sample$Time) - mean(verizon_ilec$Time)) /
    (sd(bs_sample$Time) / sqrt(n))
}
```

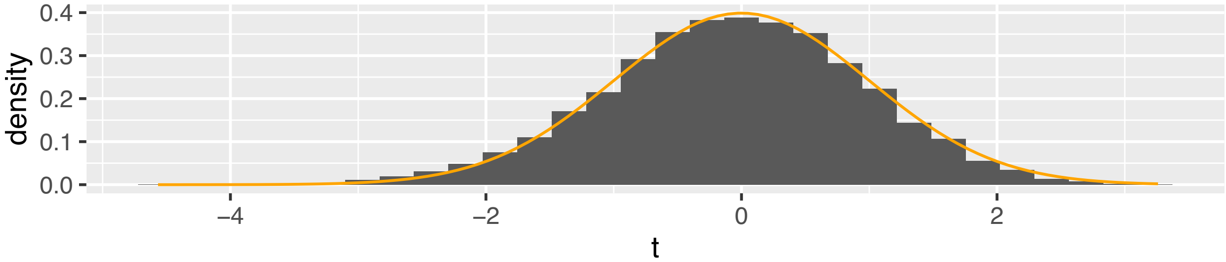
$$\left[\hat{\theta} - q(1-\tfrac{\alpha}{2})\hat{SE}(\hat{\theta}), \ \hat{\theta} - q(\tfrac{\alpha}{2})\hat{SE}(\hat{\theta})\right]$$

$$\left[\bar{x} - q(1-\tfrac{\alpha}{2}) \cdot s/\sqrt{n}, \ \bar{x} - q(\tfrac{\alpha}{2})\, s/\sqrt{n}\right]$$

```r
# 95% Bootstrap t Interval
bs_t_interval <- c(
  mean(verizon_ilec$Time) -
    quantile(bs_t_results$t, prob = 0.975) * sd(verizon_ilec$Time) / sqrt(n),
  mean(verizon_ilec$Time) -
    quantile(bs_t_results$t, prob = 0.025) * sd(verizon_ilec$Time) / sqrt(n)
)

bs_t_interval
```

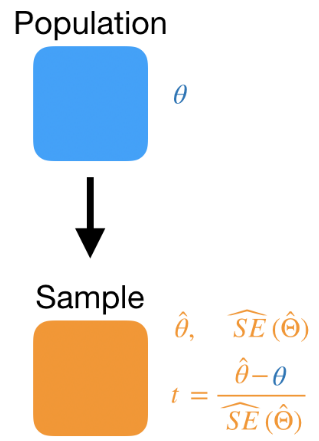
```
##    97.5%     2.5%
## 7.750175 9.193114
```

```r
# Plot to compare bootstrap estimate of distribution of t statistic to theoretical result
ggplot(data = bs_t_results, mapping = aes(x = t)) +
  geom_histogram(mapping = aes(y = ..density..)) +
  stat_function(fun = dt, args = list(df = n - 1), color = "orange")
```

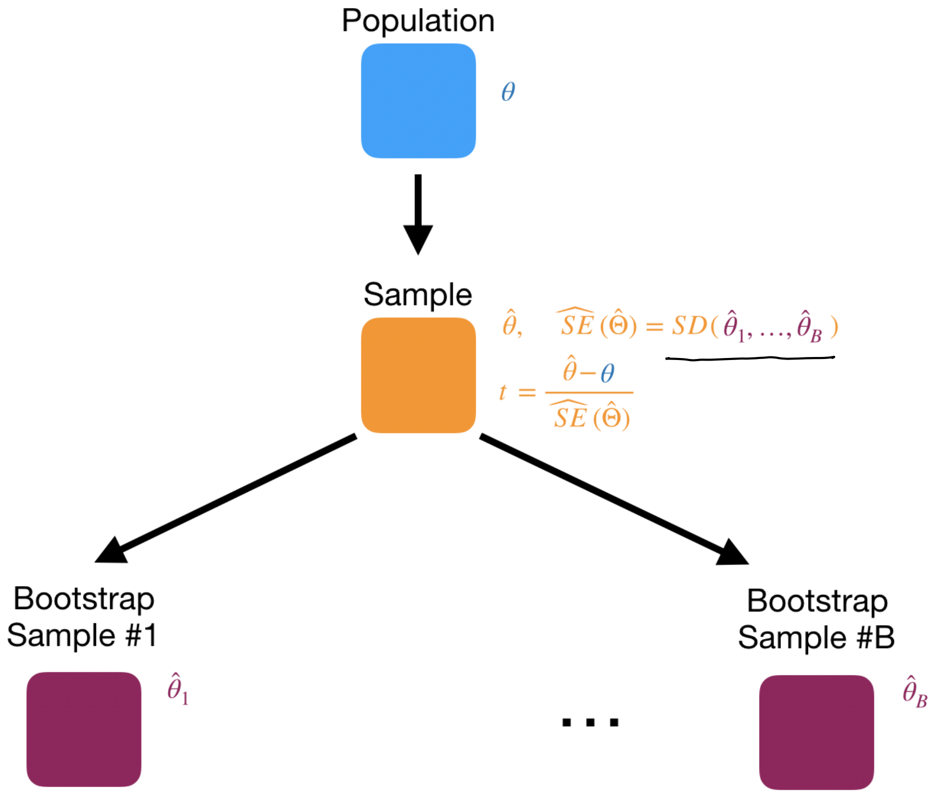


Q: What if we didn't have a formula for $\widehat{SE}(\hat{\Theta})$?

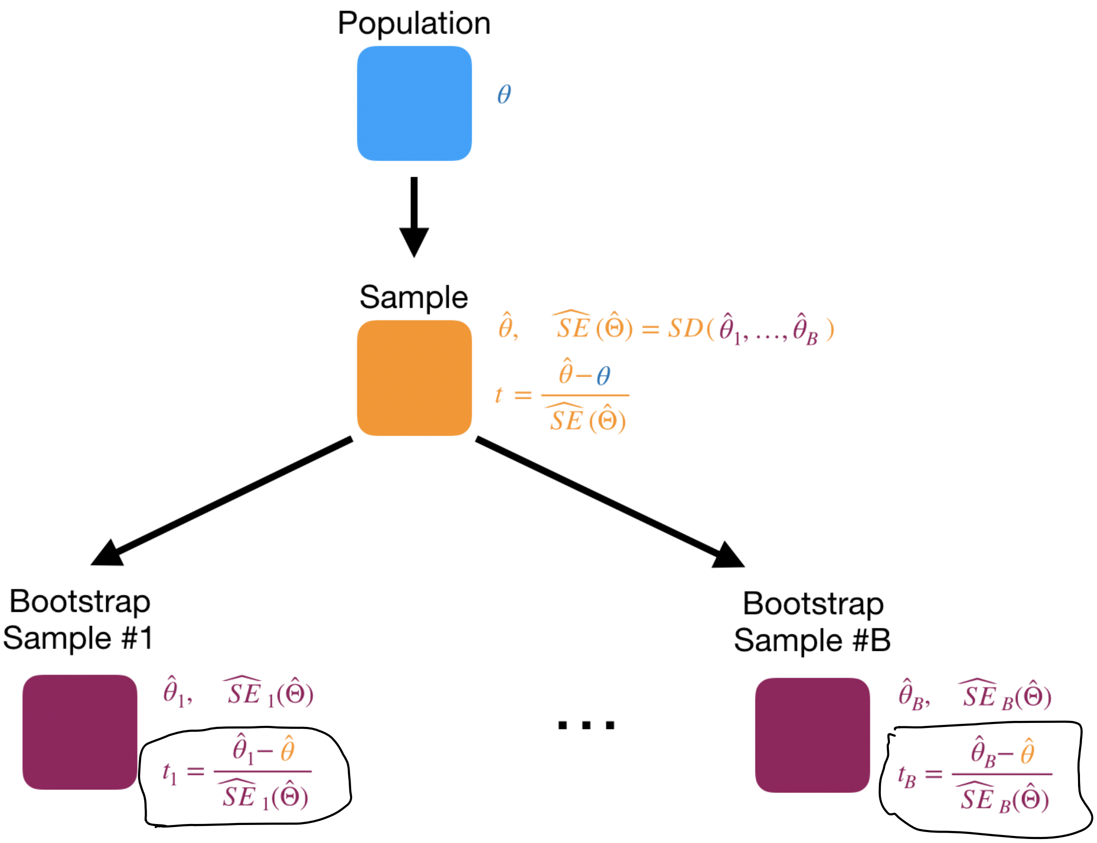
- Suppose we wanted to calculate $t = \frac{\hat{\theta}-\theta}{\widehat{SE}(\hat{\Theta})}$ based on our original sample.


Population
θ
Sample
$\hat{\theta}, \quad \widehat{SE}(\hat{\Theta})$
$t = \frac{\hat{\theta}-\theta}{\widehat{SE}(\hat{\Theta})}$

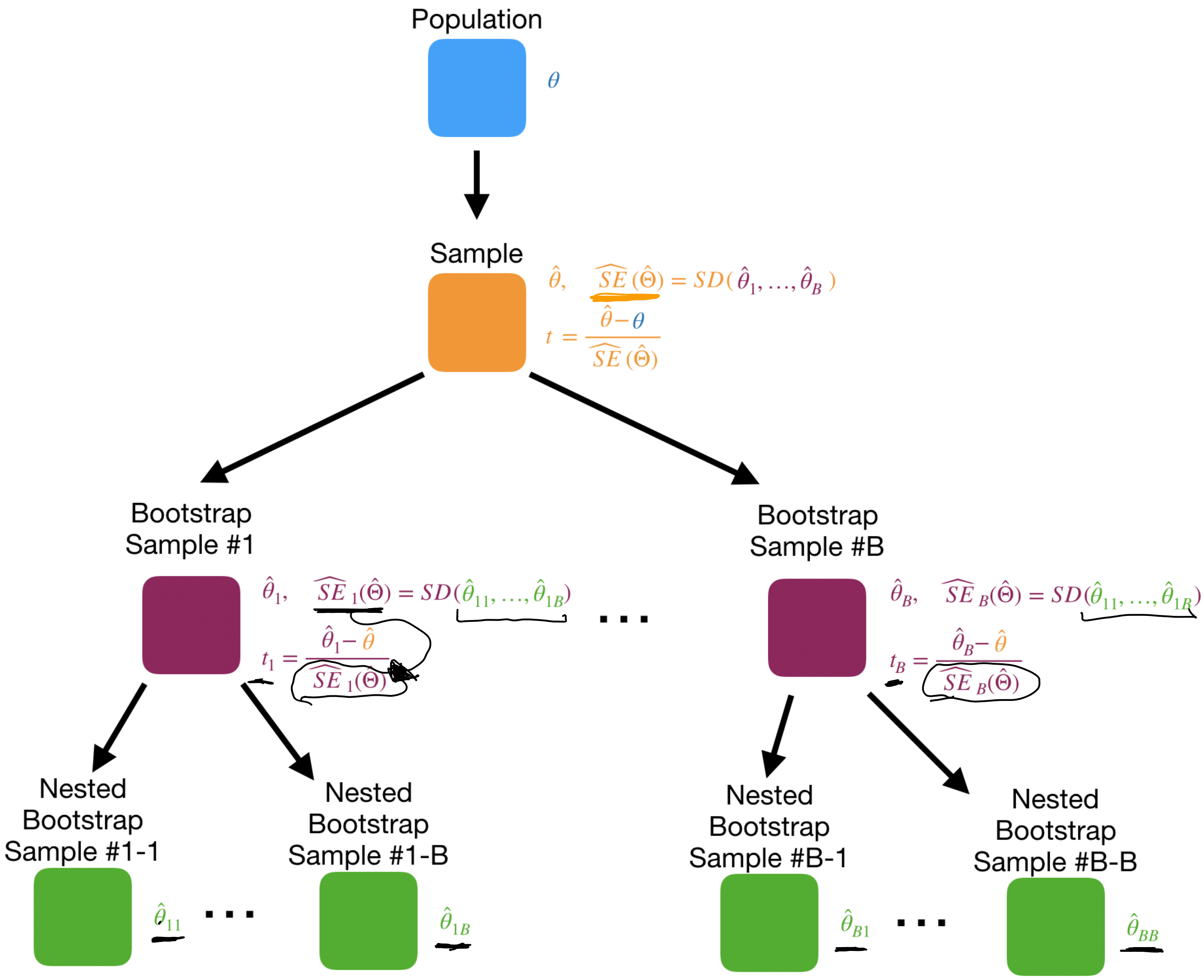
- Here's how we could do that, using a bootstrap to obtain $\widehat{SE}(\hat{\Theta})$:


Population
θ
Sample
$\hat{\theta}, \quad \widehat{SE}(\hat{\Theta}) = \underline{SD(\hat{\theta}_1, \ldots, \hat{\theta}_B)}$
$t = \frac{\hat{\theta}-\theta}{\widehat{SE}(\hat{\Theta})}$
Bootstrap Sample #1
$\hat{\theta}_1$
...
Bootstrap Sample #B
$\hat{\theta}_B$

- Now suppose we want to calculate bootstrap estimate of the sampling distribution of $T = \frac{\hat{\Theta}-\theta}{\widehat{SE}(\hat{\Theta})}$
- We need to calculate $t_1, \ldots, t_B$ based on B bootstrap samples


Population
θ
Sample
$\hat{\theta}, \quad \widehat{SE}(\hat{\Theta}) = SD(\hat{\theta}_1, \ldots, \hat{\theta}_B)$
$t = \frac{\hat{\theta}-\theta}{\widehat{SE}(\hat{\Theta})}$
Bootstrap Sample #1
$\hat{\theta}_1, \quad \widehat{SE}_1(\hat{\Theta})$
$t_1 = \frac{\hat{\theta}_1-\hat{\theta}}{\widehat{SE}_1(\hat{\Theta})}$
...
Bootstrap Sample #B
$\hat{\theta}_B, \quad \widehat{SE}_B(\hat{\Theta})$
$t_B = \frac{\hat{\theta}_B-\hat{\theta}}{\widehat{SE}_B(\hat{\Theta})}$

- But to calculate t_b, we must replicate the full process that would have been used to calculate t for the original sample.
- This means we need to obtain a bootstrap estimate $\widehat{SE}_b(\hat{\Theta})$ **within the processing of bootstrap sample** b!


Population
θ
Sample
$\hat{\theta}$, $\widehat{SE}(\hat{\Theta}) = SD(\hat{\theta}_1, \dots, \hat{\theta}_B)$
$t = \frac{\hat{\theta} - \theta}{\widehat{SE}(\hat{\Theta})}$
Bootstrap Sample #1
$\hat{\theta}_1$, $\widehat{SE}_1(\hat{\Theta}) = SD(\hat{\theta}_{11}, \dots, \hat{\theta}_{1B})$
$t_1 = \frac{\hat{\theta}_1 - \hat{\theta}}{\widehat{SE}_1(\hat{\Theta})}$
...
Bootstrap Sample #B
$\hat{\theta}_B$, $\widehat{SE}_B(\hat{\Theta}) = SD(\hat{\theta}_{11}, \dots, \hat{\theta}_{1B})$
$t_B = \frac{\hat{\theta}_B - \hat{\theta}}{\widehat{SE}_B(\hat{\Theta})}$
Nested Bootstrap Sample #1-1
$\hat{\theta}_{11}$
...
Nested Bootstrap Sample #1-B
$\hat{\theta}_{1B}$
Nested Bootstrap Sample #B-1
$\hat{\theta}_{B1}$
...
Nested Bootstrap Sample #B-B
$\hat{\theta}_{BB}$

Here's a statement of our full algorithm (omitting pre-allocation of storage space for clarity, though that is an important implementation detail to keep it from being too slow in R):

```
1. For $b = 1, ..., B$   (main goal: $t_1, \dots, t_b$, tell us about distribution of $T$ to get a CI for $\theta$.
    i. Draw a bootstrap sample from the original data, with replacement
    ii. For $j = 1, ..., B$        only new part
        a. Draw a bootstrap sample from the bootstrap sample obtained in step 1 i, with replacement
        b. Calculate $\hat{\theta}_{bj}$ based on the bootstrap sample from step 1 ii a
    iii. Calculate $\widehat{SE}_b(\hat{\Theta}) = \sqrt{\frac{1}{B-1}\sum_{j=1}^{B}(\hat{\theta}_{bj} - \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}_{bk})^2}$
    iv. Calculate $\hat{\theta}_b$ based on the bootstrap sample from step 3 i
    v. Calculate $t_b = \frac{\hat{\theta}_b - \hat{\theta}}{SE(\hat{\theta}_b)}$
2. Calculate $\widehat{SE}(\hat{\Theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}_k)^2}$
3. Calculate $\hat{\theta}$ based on the original observed data
4. Calculate the confidence interval as $[\hat{\theta} - q(1 - \frac{\alpha}{2})SE(\hat{\theta}), \hat{\theta} - q(\frac{\alpha}{2})SE(\hat{\theta})]$, where $q(1 - \frac{\alpha}{2})$ and $q(\frac{\alpha}{2})$ are quantiles of
   the bootstrap estimate of the sampling distribution of the $t$ statistic obtained in step 1, $\hat{\theta}$ was computed in step 2,
   and $SE(\hat{\theta})$ was computed in step 4.
```

Same bootstrap t CI from previous video.

This is really really slow.

```r
# sample size
n <- nrow(verizon_ilec)

# how many bootstrap samples to take, and storage space for the results
num_samples <- 10^3
bs_results <- data.frame(
  t = rep(NA, num_samples),
  theta_hat = rep(NA, num_samples)
)

num_inner_samples <- 10^3 # fewer to make this take an achievable amount of time
inner_bs_se_results <- data.frame(
  theta = rep(NA, num_inner_samples)
)

# Step 1
for(b in seq_len(num_samples)) {
  # Step 1 i: Draw a bootstrap sample of size n with replacement from the observed data
  bs_sample <- verizon_ilec %>%
    sample_n(size = n, replace = TRUE)

  # Step 1 ii: Use a nested bootstrap to estimate SE(\hat{\theta}_b), based on this bootstrap sample
  for(j in seq_len(num_inner_samples)) {
    # Step 1 ii a
    inner_bs_sample <- bs_sample %>%
      sample_n(size = n, replace = TRUE)

    # Step 1 ii b
    inner_bs_se_results$theta[j] <- mean(inner_bs_sample$Time)
  }
  # Step 1 iii
  bs_se <- sd(inner_bs_se_results$theta)

  # step 1 iv: Calculate theta hat based on bootstrap sample
  bs_results$theta_hat[b] <- mean(bs_sample$Time)

  # Step 1 v: Calculate t statistic based on bootstrap sample
  bs_results$t[b] <- (mean(bs_sample$Time) - mean(verizon_ilec$Time)) / bs_se
}

# Step 2: Calculate bootstrap standard error based on "outer" bootstrap samples
bs_se <- sd(bs_results$theta_hat)

# Step 3: Calculate theta hat based on original sample
theta_hat <- mean(verizon_ilec$Time)

# Step 4: Calculate 95% Bootstrap t Interval
bs_t_interval <- c(
  theta_hat - quantile(bs_t_results$t, prob = 0.975) * bs_se,
  theta_hat - quantile(bs_t_results$t, prob = 0.025) * bs_se
)

bs_t_interval
```

```
##    97.5%     2.5% 
## 7.753299 9.189423
```