1.5.2 Basic Bootstrap CI (corrects for bias).

based on residuals.

The $100(1-\alpha)\%$ Basic Bootstrap CI for θ is

$$\Big(\underbrace{\hat\theta}_{\text{estimate from original sample}} - [\underbrace{\hat\theta_{1-\alpha/2}}_{1-\alpha/2 \text{ quantile based on bootstrap distribution}} - \hat\theta],\ \hat\theta - [\hat\theta_{\alpha/2} - \hat\theta]\Big) = \left(2\hat\theta - \hat\theta_{1-\alpha/2},\ 2\hat\theta - \hat\theta_{\alpha/2}\right).$$

Why?

Let $\varepsilon = \hat\theta - \theta$ and $\varepsilon_{1-\alpha/2}$, $\varepsilon_{\alpha/2}$ quantiles of dsn of ε's.

$$P(\varepsilon_{\alpha/2} \le \hat\theta - \theta \le \varepsilon_{1-\alpha/2}) = 1-\alpha \Rightarrow (1-\alpha)\%\text{ CI is } \left(\hat\theta - \varepsilon_{1-\alpha/2},\ \hat\theta - \varepsilon_{\alpha/2}\right).$$

$\varepsilon_{\alpha/2} \approx \hat\theta_{\alpha/2} - \hat\theta$ bootstrap version.

Assumptions/usage

- corrects for bias, but slightly harder to explain than percentile.
- Not transformation invariant.

1.5.3 Bootstrap t CI (Studentized Bootstrap)

Consider $z = \dfrac{\hat{\theta} - E(\hat{\theta})}{se(\hat{\theta})}$.

Even if the distribution of $\hat{\theta}$ is Normal and $\hat{\theta}$ is unbiased for θ, the Normal distribution is not exactly correct for z.

because we need to estimate $se(\hat{\theta})$. $\Rightarrow t^* = \dfrac{\hat{\theta} - E(\hat{\theta})}{\hat{se}(\hat{\theta})} \sim t_{n-1}$? No

where $\hat{se}(\hat{\theta}) = sd(\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(B)})$.

Additionally, the distribution of $\hat{se}(\hat{\theta})$ is unknown.

So we cannot just claim $t^* \sim t_{n-1}$.

$\Rightarrow$ The bootstrap t interval does not use a Student t distribution as the reference distribuion, instead we estimate the distribution of a "t type" statistic by resampling.

The $100(1-\alpha)\%$ Boostrap t CI is

based on bootstrap samples $\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{(B)}$

$$\left(\hat{\theta} - t^*_{1-\alpha/2} \cdot \hat{se}(\hat{\theta}),\ \hat{\theta} + t^*_{\alpha/2}\, \hat{se}(\hat{\theta})\right).$$

estimate from original data sample

quantile of the bootstrap dsn of our "t-type" statistic.

Overview

t-type statistic : $t^{(1)} = \dfrac{\hat{\theta}^{*(1)} - \hat{\theta}}{\underbrace{\hat{se}(\hat{\theta}^{*(1)})}}, \dots, t^{(B)} = \dfrac{\hat{\theta}^{*(B)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{*(B)})}$

= bootstrap estimate of $se(\hat{\theta}^{*(1)})$ based on the first bootstrap sample?? Double bootstrap!

To estimate the "t style distribution" for $\hat{\theta}$,

1. Compute $\hat{\theta}$
2. For $b = 1, \dots, B$
   (a) Sample w/ replacement
   $\underline{y}^{(b)} = (y_1^{(b)}, \dots, y_n^{(b)})$
   (b) Compute $\hat{\theta}^{*(b)}$
   (c) For each $r = 1, \dots, R$
   (i) Sample w/ replacement from $\underline{y}^{(b)}$
   $\underline{y}^{(b)(r)} = (y_1^{(b)(r)}, \dots, y_n^{(b)(r)})$
   (ii) Compute $\hat{\theta}^{*(b)(r)}$
   (d) Compute $\hat{se}(\hat{\theta}^{*(b)}) = sd(\hat{\theta}^{*(b)(1)}, \dots, \hat{\theta}^{*(b)(R)})$
   (e) Compute $t^{(b)} = \dfrac{\overset{2(b)}{\hat{\theta}^{*(b)}} - \overset{1.}{\hat{\theta}}}{\hat{se}(\hat{\theta}^{*(b)})} \leftarrow 2(d)$
3. get quantiles $t^*_{\alpha/2}, t^*_{1-\alpha/2}$
4. Compute CI.

Assumptions/usage

- Computationally intensive.
- Not doing anything for bias/skewness.
- Need $\hat{\theta}$ independent of $\hat{se}(\hat{\theta})$.

This idea is based on "pivot quantities" = a function of observations and parameters whose probability dsn doesn't depend on the parameter.

E.g. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ ← no μ in here!

This can help us to obtain a bootstrap CI for μ.

You could create other pivot quantities to extend via bootstrap, e.g.

If $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ $\quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\Rightarrow \left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} , \frac{(n-1)S^2}{\chi^2_{\alpha/2}} \right)$$ would be a 95% CI for σ^2.

Bootstrap version does not assume $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Now need to estimate χ^2-type quantile using the bootstrap.

1.5.4 BCa CIs

correct for skew.

" bias-corrected accelerated"

Modified version of percentile intervals that adjusts for bias of estimator and skewness of the sampling distribution.

This method automatically selects a transformation so that the normality assumption holds.

Idea:

Assume there exists a monotonically increasing function g and constants a and b s.t.

$$U = \frac{g(\hat\theta) - g(\theta)}{1 + a g(\theta)} + b \sim N(0,1) \quad \text{where } 1 + a g(\theta) > 0.$$

By the bootstrap principle

$$U^* = \frac{g(\hat\theta^*) - g(\hat\theta)}{1 + a g(\hat\theta)} + b \overset{\text{approx}}{\sim} N(0,1).$$

$\Rightarrow$ For any quantile of a standard Normal dsn,

$$\begin{aligned}\alpha &\approx P^*\left[U^* \le z_\alpha\right] \\ &= P^*\left[\frac{g(\hat\theta^*) - g(\hat\theta)}{1 + a g(\hat\theta)} + b \le z_\alpha\right] \\ &= P^*\Big[\hat\theta^* \le \underbrace{g^{-1}\big(g(\hat\theta) + (z_\alpha - b)(1 + a g(\hat\theta))\big)}_{\hat\theta_\alpha}\Big]\end{aligned}$$

The α quantile from the bootstrap dsn of $\hat\theta^*$, denoted $\hat\theta_\alpha$, is observable from BS dsn.

$$\Rightarrow g^{-1}\big(g(\hat\theta) + (z_\alpha - b)(1 + a g(\hat\theta))\big) \approx \hat\theta_\alpha$$

To use this, consider U: $1-\alpha = P(U > z_\alpha)$

$$= P\Big(\theta < \underbrace{g^{-1}\Big(g(\hat\theta) + \frac{b - z_\alpha}{1 - a(b - z_\alpha)}\left[1 + a g(\hat\theta)\right]\Big)}\Big)$$

Notice similarity to above.

$\Rightarrow$ If we could find β such that $\frac{b - z_\alpha}{1 - a(b - z_\alpha)} = z_\beta - b$ then the bootstrap principle can be applied to conclude $\theta < \hat\theta_\beta$ will be an appropriate $1-\alpha$ upper CI limit.

2-sided argument is similar.

$$\Rightarrow z_\beta = \frac{b - z_\alpha}{1 - a(b - z_\alpha)} + b \quad \Rightarrow \quad \beta = \Phi\left(\frac{b + z_{1-\alpha}}{1 - a(b + z_{1-\alpha})} + b\right).$$

β^{th} quantile of $N(0,1)$

$\Rightarrow$ ① find a, b, ② Compute β, ③ Find β^{th} quantile of empirical dsn of $\hat\theta^*$

b: Let p_0 denote the fraction of obs. from bootstrap dsn s.t. $\hat\theta^* < \hat\theta$. Since g is monotone, this is the same fractions of $\hat\theta^*$'s s.t. $g(\hat\theta^*) < g(\hat\theta)$.

$\Rightarrow P(Z < b) = p_0$ where $Z \sim N(0,1)$ gives us a way to estimate b (If the bootstrap dsn has $\hat\theta$ as its median, then $b = 0$).

$\Rightarrow b$ corrects for bias.

a: Let $S_{-i} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$ and let $\hat\theta_{-i}$ denote the estimate of θ based on S_{-i}: $a = \dfrac{\sum_{i=1}^{n}(\hat\theta_{(\cdot)} - \hat\theta_{-i})^3}{6\left[\sum_{i=1}^{n}(\hat\theta_{(\cdot)} - \hat\theta_{-i})^2\right]^{3/2}}$ based on skewness estimator where $\hat\theta_{(\cdot)}$ is the mean of $\hat\theta_{-i}$'s.

If $a=0$, don't adjust for skewness $\Rightarrow$ BC interval.

The BCa method uses bootstrapping to estimate the bias and skewness then modifies which percentiles are chosen to get the appropriate confidence limits for a given data set.

In summary,

BCa is like the percentile bootstrap CI, but instead of $(\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2})$, choose better quantiles to account for bias and skewness.

Has better coverage than percentile method in empirical studies, but harder to explain.

Your Turn

We will consider a telephone repair example from Hesterberg (2014). `Verizon` has repair times, with two groups, CLEC and ILEC, customers of the "Competitive" and "Incumbent" local exchange carrier.

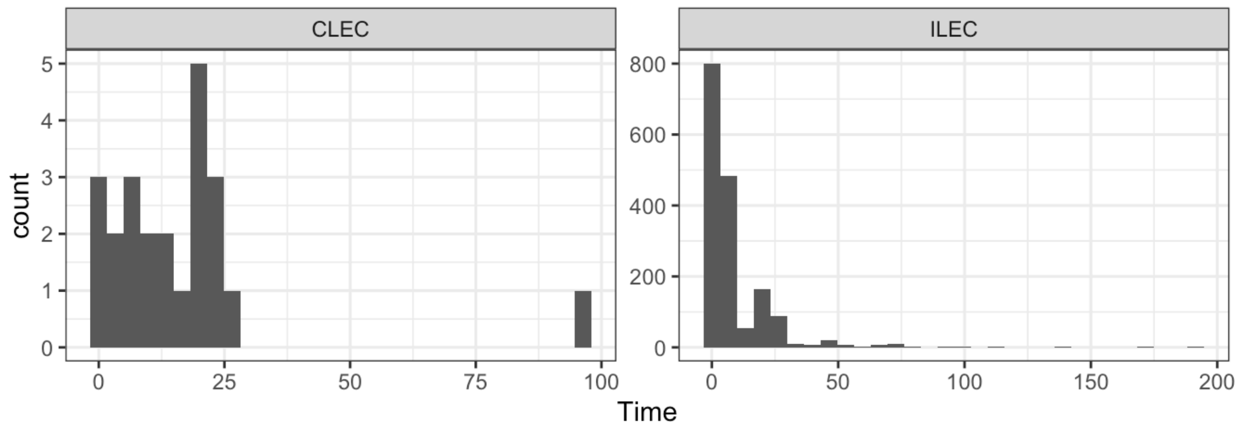
```
library(resample) # package containing the data

data(Verizon)
head(Verizon)
```

```
##    Time Group
## 1 17.50  ILEC
## 2  2.40  ILEC
## 3  0.00  ILEC
## 4  0.65  ILEC
## 5 22.23  ILEC
## 6  1.20  ILEC
```

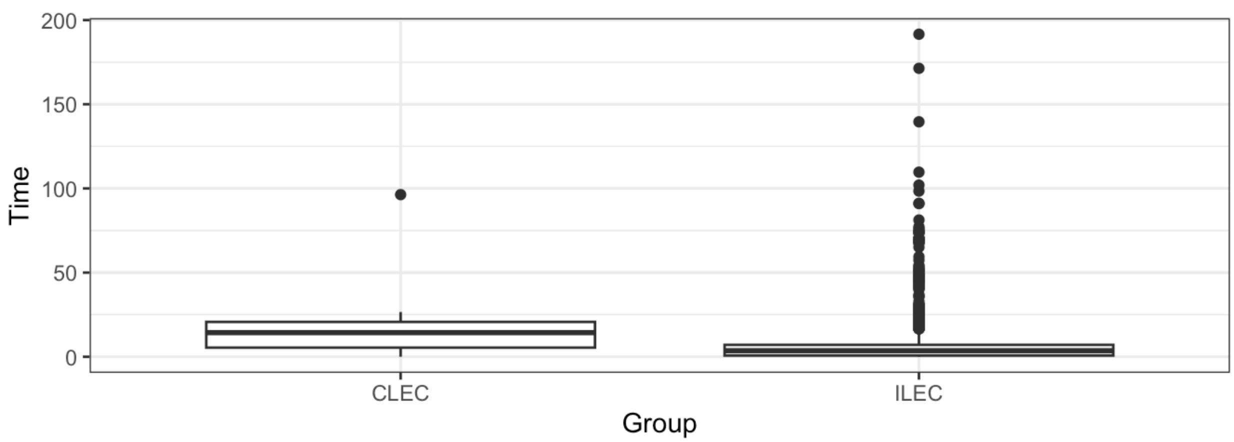
```
Verizon |>
  group_by(Group) |>
  summarize(mean = mean(Time), sd = sd(Time), min = min(Time), max =
        max(Time)) |>
  kable()
```

Group	mean	sd	min	max	n
CLEC	16.509130	19.50358	0	96.32	23
ILEC	8.411611	14.69004	0	191.60	1664

```
ggplot(Verizon) +
  geom_histogram(aes(Time)) +
  facet_wrap(.~Group, scales = "free")
```




```
ggplot(Verizon) +
  geom_boxplot(aes(Group, Time))
```



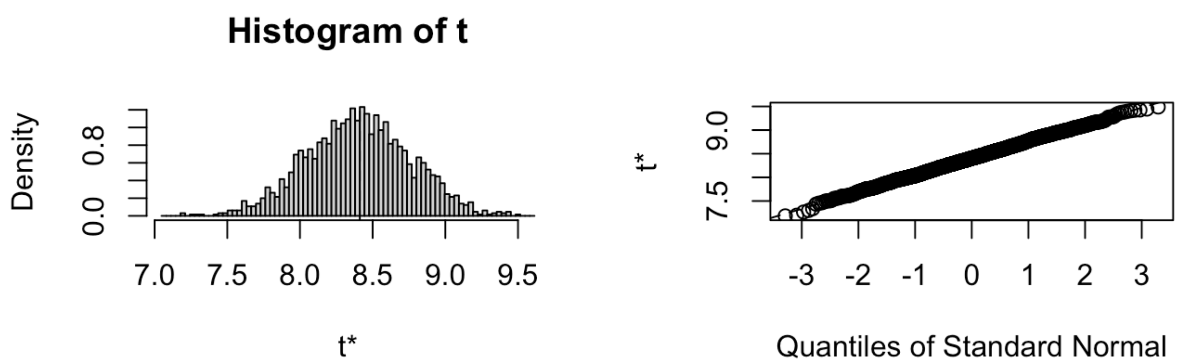
1.6 Bootstrapping CIs

you could write your own.

There are many bootstrapping packages in `R`, we will use the `boot` package. The function `boot` generates R resamples of the data and computes the desired statistic(s) for each sample. This function requires 3 arguments:

1. data = the data from the original sample (data.frame or matrix).
2. statistic = a function to compute the statistic from the data where the first argument is the data and the second argument is the indices of the obervations in the boostrap sample.
3. R = the number of bootstrap replicates.

```r
library(boot) # package containing the bootstrap function

mean_func <- function(x, idx) {
  mean(x[idx])
}

ilec_times <- Verizon[Verizon$Group == "ILEC",]$Time
boot.ilec <- boot(ilec_times, mean_func, 2000)

plot(boot.ilec)
```



If we want to get Bootstrap CIs, we can use the `boot.ci` function to generate the different nonparametric bootstrap confidence intervals.

(*) →

```r
boot.ci(boot.ilec, conf = .95, type = c("perc", "basic", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.ilec, conf = 0.95, type = c("perc",
"basic",
##     "bca"))
##
## Intervals :
## Level      Basic              Percentile            BCa
## 95%   ( 7.733,  9.110 )   ( 7.714,  9.091 )   ( 7.755,  9.125 )
## Calculations and Intervals on Original Scale
```

```r
## we can do some of these on our own
## percentile
quantile(boot.ilec$t, c(.025, .975))
```

```
##     2.5%    97.5%
## 7.714075 9.084725
```

```r
## basic
2*mean(ilec_times) - quantile(boot.ilec$t, c(.975, .025))
```

```
##    97.5%     2.5%
## 7.738496 9.109147
```

To get the studentized bootstrap CI, we need our statistic function to also return the variance of $\hat{\theta}$.

```r
mean_var_func <- function(x, idx) {
  c(mean(x[idx]), var(x[idx])/length(idx))
}

boot.ilec_2 <- boot(ilec_times, mean_var_func, 2000)
boot.ci(boot.ilec_2, conf = .95, type = "stud")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.ilec_2, conf = 0.95, type = "stud")
##
## Intervals :
## Level    Studentized
## 95%   ( 7.728,  9.183 )
## Calculations and Intervals on Original Scale
```

Which CI should we use?