```
## 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
```

very similar
not biased.

```
## 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 $\hat{\theta}$.

```
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?

All very similar.

Maybe Percentile if needing to explain to a judge/jury.

Maybe BCA b/c has shown better coverage.

1.7 Bootstrapping for the difference of two means

Given iid draws of size n and m from two populations, to compare the means of the two groups using the bootstrap,

1. For replicates $b = 1, \dots, B$
   a) Draw a sample of size n w/ replacement from sample 1 and separately a sample of size m w/ replacement from sample 2.
   b) Compute a statistic that compares the groups (i.e. $\hat\theta = \bar y_1 - \bar y_2$).
2. Have a bootstrap dsn of $\hat\theta^{(1)}, \dots, \hat\theta^{(B)}$ — shape, bias, se, etc.
3. Compute an appropriate CI.

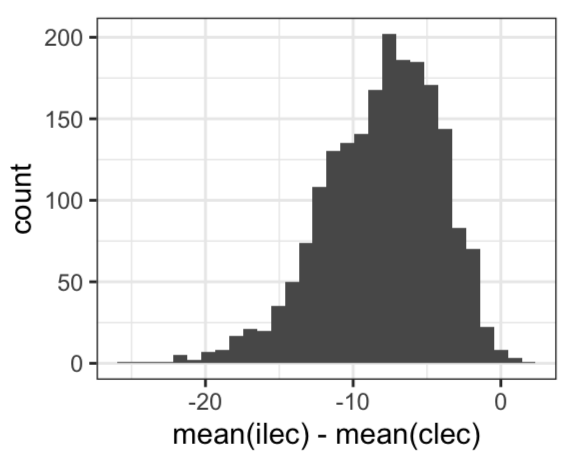
again feel free to write your own.

The function `two.boot` in the `simpleboot` package is used to bootstrap the difference between univariate statistics. Use the bootstrap to compute the shape, bias, and bootstrap sample error for the samples from the `Verizon` data set of CLEC and ILEC customers.

```
library(simpleboot)

clec_times <- Verizon[Verizon$Group == "CLEC",]$Time
diff_means.boot <- two.boot(ilec_times, clec_times, "mean", R = 2000)

ggplot() +
  geom_histogram(aes(diff_means.boot$t)) +
  xlab("mean(ilec) - mean(clec)")
```



looks skewed.

```r
# Your turn: estimate the bias and se of the sampling distribution
```

Which confidence intervals should we use?

```r
# Your turn: get the chosen CI using boot.ci
```

Is there evidence that

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 < 0$$

is rejected?

2 Parametric Bootstrap

In a **nonparametric bootstrap**, we resample the observed data

Create a bootstrapped sample $y_1^*, \dots, y_n^*$ iid from empirical dsn $\hat{F}_n$.

$\uparrow$

iid case, equivalent to resampling original data w/ replacement.

In a **parametric bootstrap**, Assume a parametric model.

Key idea: use a fitted parametric model $\hat{F}(y) = F(y \mid \hat{\psi})$ to estimate F where $\hat{\psi}$ estimate using MLE (or another method) from data.

Create a bootstrapped sample $y_1^*, \dots, y_n^*$ iid from $F(y \mid \hat{\psi})$.

i.e. resample from a model w/ parameters estimated using original data.

For both methods,

1. Compute $\hat{\theta}^{*(b)}$ for each bootstrapped sample $y_1^{*(b)}, \dots, y_n^{*(b)}$

2. repeat procedure B times to get

$$\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(B)}$$

and make inferences using the result.

2.1 Bootstrapping for linear regression

Consider the regression model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i, i = 1, \dots, n$ with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

$Y_1, \dots, Y_n$ not iid! They have different conditional means.

Resampling in the bootstrap must be completed on iid quantities.

Two approaches for bootstrapping linear regression models –

1. Bootstrapping the residuals (model based, parametric).

2. Paired bootstrapping (case resampling, nonparametric)

2.1.1 Bootstrapping the residuals (model-based).

1. Fit the regression model using the original data to get $\hat{\beta}$
2. Compute the residuals from the regression model, errors ε_i are assumed iid.

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \quad i = 1, \dots, n$$

3. Sample $\hat{\epsilon}_1^*, \dots, \hat{\epsilon}_n^*$ with replacement from $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$.
4. Create the bootstrap sample

fitted values based on model/data

$$y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \epsilon_i^*, \quad i = 1, \dots, n$$

→ get

bootstrapped data.

$\{y_1^*, x_1\}$
$\vdots$
$\{y_n^*, x_n\}$

5. Estimate $\hat{\boldsymbol{\beta}}^*$

fit regression model on bootstrapped data to get $\hat{\beta}^*$

6. Repeat steps 3-5 B times to create B bootstrap estimates of $\hat{\beta}$.

Assumptions:

- ε_i are iid
  ↳ i.e. we have fit a "good" model.

- design matrix $\underline{X}$ is fixed.

2.1.2 Paired bootstrapping (case resampling).

Resample $z_i^* = (y_i, \boldsymbol{x}_i)^*$ from the empirical distribution of the pairs $(y_i, \boldsymbol{x}_i)$.

Fit regression model w/ n bootstrapped pairs $(y_i, \underline{x}_i)^*$.

$$y_i^* = (\underline{x}_i^*)^T \underline{\beta} + \varepsilon_i \quad i = 1, \ldots, n$$

Assumptions:

Assume $(y_i, \underline{x}_i)$ are iid from a population.

Can have varying design matrix $\underline{X}$.

2.1.3 Which to use?

1. Standard inferences - i.e. earlier part of this class, likelihood approaches.

   Most of the time.

2. Bootstrapping the residuals -

   - most appropriate for designed experiments where $\underline{X}$ is fixed in advance.
   - model based, model must be reasonable fit for the data.
   - useful if complex sampling dsn for $\hat{\beta}_k$ maybe some weird nonlinear function maybe.

3. Paired bootstrapping -

   - robust to model misspecification.
   - useful for observational studies where values of predictors aren't fixed in advance

     $\Rightarrow$ bootstrap mirrors data generating process.