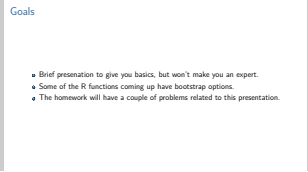
Bootstrap Hypothesis Testing

Goals

- Brief presenation to give you basics, but won't make you an expert.
- Some of the R functions coming up have bootstrap options.
- The homework will have a couple of problems related to this presentation.

Bootstrap Hypothesis Testing

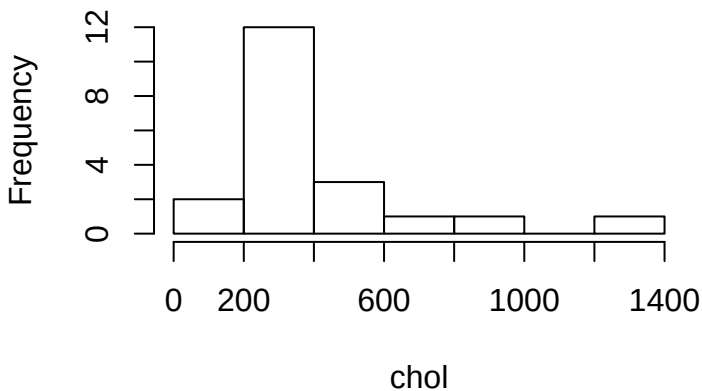
Goals

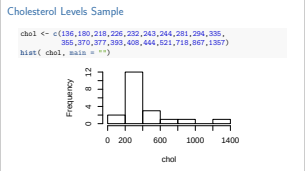


- audio01.mp3
- We just want you to get some idea about how bootstrap hypothesis testing works.
- Some of the R functions for hypothesis testing we'll use this week and next have bootstrap options available.
- You'll also see a bit more about bootstrap hypothesis testing in this weeks other presenation.

Cholesterol Levels Sample

```
chol <- c(136,180,218,226,232,243,244,281,294,335,
          355,370,377,393,408,444,521,718,867,1357)
hist( chol, main = "")
```





- audio02.mp3
- this is the same sample of cholesterol levels we used to introduce bootstrap confidence intervals in Lesson 2.
- in this case we want to test a hypothesis about the population mean cholesterol level, but the sample clearly isn't from a normally distributed population and has at least one pretty large outlier.

The t.test() in R

$$H_0 : \mu = 310, H_a : \mu > 310$$

```
t.test( chol, mu = 310, alternative = 'greater')$p.value
```

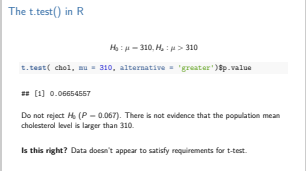
```
## [1] 0.06654557
```

Do not reject H_0 ($P = 0.067$). There is not evidence that the population mean cholesterol level is larger than 310.

Is this right? Data doesn't appear to satisfy requirements for t-test.

Bootstrap Hypothesis Testing

The t.test() in R



- audio03.mp3
- what if we go ahead with the t-test anyway even though the sample data appears quite skewed with possibly a large outlier.
- We don't expect the t-test to work well for the small sample of size 20 we have here.
- the underlying theory that repeated samples from this population yield a t-distribution requires a normally distributed population,
- but we can try using bootstrapping to approximate the sampling distribution by a boot distribution
- one way to do this is by bootstrapping a confidence interval and using it to draw

Hypothesis Testing using Confidence Intervals

Confidence interval

$\Updownarrow$

Plausible values for population parameter

$\Updownarrow$

Retain H_0 for plausible parameter values

Bootstrap Hypothesis Testing

Hypothesis Testing using Confidence Intervals

Hypothesis Testing using Confidence Intervals

Confidence interval
⇕
Plausible values for population parameter
⇕
Retain H_0 for plausible parameter values

- no audio

Equivalence between Tests and Intervals

Significance level α.
$H_0 : \theta = \theta_0$.

Alternative Hypothesis H_a	Confidence Level	Reject H_0 if
$\theta \neq \theta_0$	$100(1-\alpha)\%$	if θ_0 not in CI
$\theta > \theta_0$	$100(1-2\alpha)\%$	if θ_0 is below CI
$\theta < \theta_0$	$100(1-2\alpha)\%$	if θ_0 is above CI

Equivalence between Tests and Intervals

Significance level α.
$H_0 : \theta = \theta_0$.

Alternative Hypothesis H_a	Confidence Level	Reject H_0 if
$\theta \neq \theta_0$	$100(1-\alpha)\%$	if θ_0 not in CI
$\theta > \theta_0$	$100(1-2\alpha)\%$	if θ_0 is below CI
$\theta < \theta_0$	$100(1-2\alpha)\%$	if θ_0 is above CI

- audio04.mp3
- we've written this slide using a generic parameter which we call theta, but this could be a populatoin mean or proportion, or any other paramter.
- a confidence interval is equivalent to a two-sided test. A 95% confidence interval for a mean gives you all the values of the mean for which we'd retain the null hypothesis.
- for one-sided tests we have to adjust the confidence level if we want to use a regular two-sided interval to test the hypothesis
- some hypothesis test in R give you one-sided confidence bounds when you do one-tailed hypothesis tests, in those cases there is no need to adjust the confidence level. We will see an example of this below

Example 1

Test $H_0 : \mu = 100$ vs $H_a : \mu \neq 100$ at $\alpha = 0.05$.

```
x = rnorm( 20, mean = 100, sd = 3)
t.test( x, conf.level=.95, alternative='two.sided')$conf.int
```

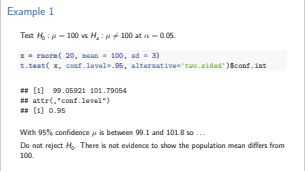
```
## [1]  99.05921 101.79054
## attr(,"conf.level")
## [1] 0.95
```

With 95% confidence μ is between 99.1 and 101.8 so ...

Do not reject H_0. There is not evidence to show the population mean differs from 100.

Bootstrap Hypothesis Testing

Example 1



Example 1

Test $H_0 : \mu = 100$ vs $H_a : \mu \neq 100$ at $\alpha = 0.05$.

```
x = rnorm( 20, mean = 100, sd = 3)
t.test( x, conf.level=.95, alternative='two.sided')$conf.int
```

```
## [1]  99.05921 101.79054
## attr(,"conf.level")
## [1] 0.95
```

With 95% confidence μ is between 99.1 and 101.8 so ...

Do not reject H_0. There is not evidence to show the population mean differs from 100.

- no audio

Example 2

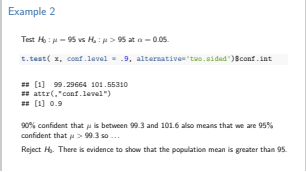
Test $H_0 : \mu = 95$ vs $H_a : \mu > 95$ at $\alpha = 0.05$.

```
t.test( x, conf.level = .9, alternative='two.sided')$conf.int
```

```
## [1]  99.29664 101.55310
## attr(,"conf.level")
## [1] 0.9
```

90% confident that μ is between 99.3 and 101.6 also means that we are 95% confident that $\mu > 99.3$ so ...

Reject H_0. There is evidence to show that the population mean is greater than 95.


Example 2

Test $H_0 : \mu = 95$ vs $H_1 : \mu > 95$ at $\alpha = 0.05$.

```
t.test( x, conf.level = .9, alternative='two.sided')$conf.int
```

```
## [1]  99.29664 101.55310
## attr(,"conf.level")
## [1] 0.9
```

90% confident that μ is between 99.3 and 101.6 also means that we are 95% confident that $\mu > 99.3$ so ...

Reject H_0. There is evidence to show that the population mean is greater than 95.

- audio05.mp3
- our 90% confidence interval tells us the mean is between 99.3 and 101.6 so we can be 5% confident that the mean is below 99.3 and also 5% confident the mean is above 101.6. Combininng the interval and the piece above 101.6 we can be 95% confident the mean is above 99.3.
- if we're 95% confident the mean is bigger than 99.3 then we can certainly accept our alternative hypothesis, at the 5% significance level, that the mean is larger than 95.

Example 2 again

Test $H_0 : \mu = 95$ vs $H_a : \mu > 95$ at $\alpha = 0.05$.

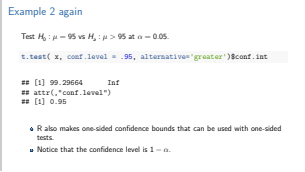
```
t.test( x, conf.level = .95, alternative='greater')$conf.int
```

```
## [1] 99.29664      Inf
## attr(,"conf.level")
## [1] 0.95
```

- R also makes one-sided confidence bounds that can be used with one-sided tests.
- Notice that the confidence level is $1 - \alpha$.

Bootstrap Hypothesis Testing

Example 2 again

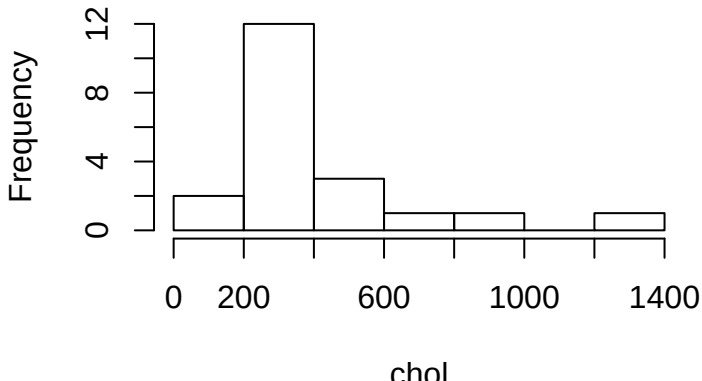


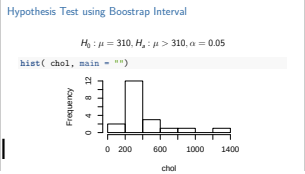
- no audio

Hypothesis Test using Boostrap Interval

$$H_0 : \mu = 310, H_a : \mu > 310, \alpha = 0.05$$

```
hist( chol, main = "")
```





- audio06.mp3
- returning to the original hypothesis test with cholesterol levels
- we didn't expect the t-test to be accurate because we have a small sample from a clearly non-normal population, so we'll bootstrap a confidence interval and use it to make a decision in our hypothesis test.

The BCa interval

```
bootMean <- function( x, i, trim = 0){ mean(x[i]) }
require(boot)
boot.object <- boot( chol, bootMean, R=5000)
CI <- boot.ci( boot.object, type='bca', conf=0.90)
CI
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.object, conf = 0.9, type = "bca")
##
## Intervals :
## Level       BCa
## 90%   (332.7, 556.9 )
```

Bootstrap Hypothesis Testing

└The BCa interval

The BCa interval

```
bootMean <- function( x, i, trim = 0){ mean(x[i]) }
require(boot)
boot.object <- boot( chol, bootMean, R=5000)
CI <- boot.ci( boot.object, type='bca', conf=0.90)
CI
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.object, conf = 0.9, type = "bca")
##
## Intervals :
## Level      BCa
## 90%   (332.7, 556.9 )
## Calculations and Intervals on Original Scale
```

- no audio

BCa interval conclusion

- We are 90% confidence that the population mean cholesterol level is between 332.7 and 556.9.
- We are also 95% confident that the population mean cholesterol level is greater than 332.7.
- At the 5% significance level we reject H_0. There is significant evidence to show the population mean cholesterol level is greater than 310.
- Different result than the analytic t-interval!

Bootstrap Hypothesis Testing

└BCa interval conclusion

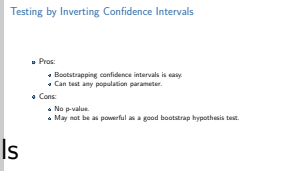
BCa interval conclusion

- We are 90% confidence that the population mean cholesterol level is between 332.7 and 556.9.
- We are also 95% confident that the population mean cholesterol level is greater than 332.7.
- At the 5% significance level we reject H_0. There is significant evidence to show the population mean cholesterol level is greater than 310.
- Different result than the analytic t-interval!

- no audio

Testing by Inverting Confidence Intervals

- Pros:
  - Bootstrapping confidence intervals is easy.
  - Can test any population parameter.
- Cons:
  - No p-value.
  - May not be as powerful as a good bootstrap hypothesis test.



- audio07.mp3
- for many applications a confidence interval is better than a hypothesis test. Using a CI we can make hypothesis test decisions, but even better we get parameter estimates

Bootstrap Hypothesis Testing

- Complicated.
- IDEA
  1. Create null population from sample data.
  2. Collect resamples.
  3. Compute test statistic for each resample $\rightarrow$ null boot distribution.
  4. Approximate p-value from boot distribution.

Bootstrap Hypothesis Testing

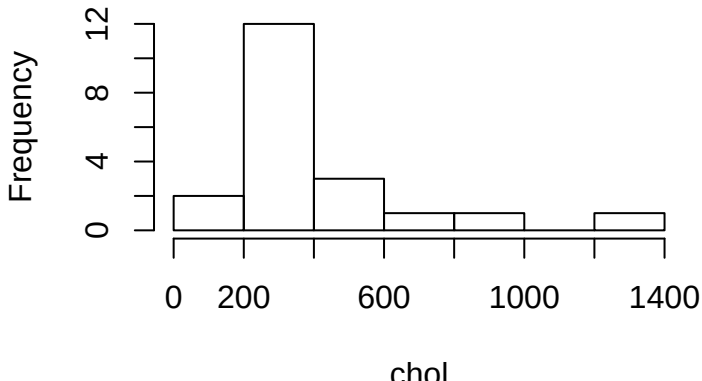
- Complicated.
- IDEA
  1. Create null population from sample data.
  2. Collect resamples.
  3. Compute test statistic for each resample → null boot distribution.
  4. Approximate p-value from boot distribution.

- audio08.mp3
- all hypothesis tests work from the idea that the null is assumed to be true and that the sampling distribution represents all the possible values of the test statistic that can occur when the null is true.
- so while bootstrap hypothesis testing and confidence intervals are similar, a bootstrap hypothesis tests starts by shifting the original sample to create a pseudo population in which the null hypothesis is true.
- the next several slides take you through these steps for our cholesterol hypothesis test.

Cholesterol Revisited

$$H_0 : \mu = 310, H_a : \mu > 310, \alpha = 0.05$$

```
hist( chol, main = "")
```



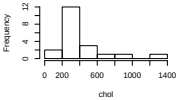
Bootstrap Hypothesis Testing

Cholesterol Revisited

Cholesterol Revisited

$H_0 : \mu = 310, H_a : \mu > 310, \alpha = 0.05$

```
hist( chol, main = "")
```



-no audio

Step 1 - Create null pseudo-population

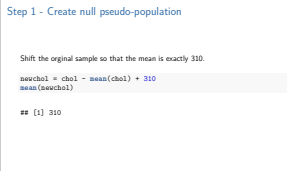
Shift the orginal sample so that the mean is exactly 310.

```
newchol = chol - mean(chol) + 310
mean(newchol)
```

```
## [1] 310
```

Bootstrap Hypothesis Testing

Step 1 - Create null pseudo-population



- no audio

Step 2 - Collect resamples

- This could be done with a `for` loop, but using `replicate` is faster.
- Each column is a resample.

```
resamples <- replicate( 5000, sample( newchol, replace = T) )
```

Bootstrap Hypothesis Testing

Step 2 - Collect resamples

- no audio

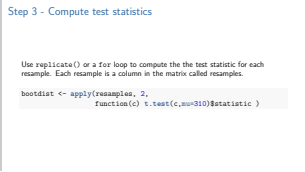
Step 3 - Compute test statistics

Use `replicate()` or a `for` loop to compute the the test statistic for each resample. Each resample is a column in the matrix called resamples.

```
bootdist <- apply(resamples, 2,
                  function(c) t.test(c,mu=310)$statistic )
```

Bootstrap Hypothesis Testing

Step 3 - Compute test statistics



- no audio

Step 4 - Approximate p-value

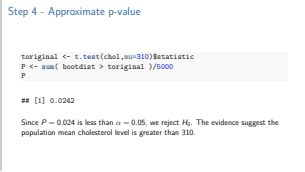
```
toriginal <- t.test(chol,mu=310)$statistic
P <- sum( bootdist > toriginal )/5000
P
```

```
## [1] 0.0242
```

Since $P = 0.024$ is less than $\alpha = 0.05$, we reject H_0. The evidence suggest the population mean cholesterol level is greater than 310.

Bootstrap Hypothesis Testing

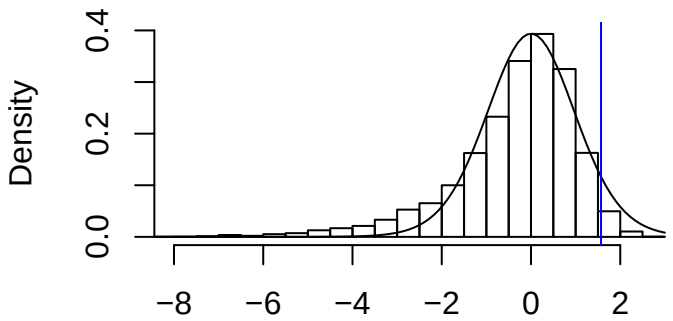
Step 4 - Approximate p-value



- audio09.mp3
- Using a bootstrap distribution or a mathematical sampling distribution the p-value is still the probability of observing a test statistic at least as extreme as the original test statistic computed from the observed data.

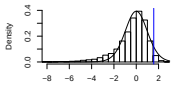
Why did the bootstrap do better?

```
hist( bootdist, probability = TRUE, breaks = 40, main = "",
      xlab="",xlim=c(-8,3), ylim = c(0,.4))
curve( dt( x, df = 19), add = TRUE)
abline( v = toriginal, col = 'blue')
```



Why did the bootstrap do better?

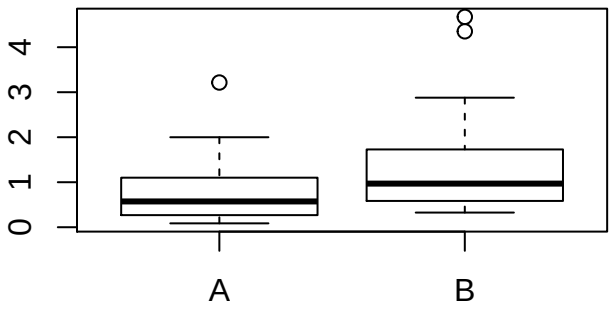
```
hist( bootdist, probability = TRUE, breaks = 40, main = "",
      xlab="",xlim=c(-8,3), ylim = c(0,.4))
curve( dt( x, df = 19), add = TRUE)
abline( v = toriginal, col = 'blue')
```



- audio10.mp3
- the histogram shows our bootstrap distribution of t test statistics
- the curve shows the mathematical t distribution which assumes the population was normally distributed
- the blue vertical line shows the location of t computed from the original sample so the P-value is the area to the right.
- we can see that mathematical t distibrution over estimates the p-value

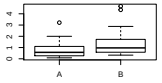
A Two Means Example

```
set.seed(321)
x <- c( rexp(20), rexp(20)+.25)
g <- factor( rep(c('A','B'),each=20) )
boxplot(x~g)
```



A Two Means Example

```
set.seed(321)
x <- c( rexp(20), rexp(20)+.25)
g <- factor( rep(c('A','B'),each=20) )
boxplot(x~g)
```



- audio11.mp3
- we've created an artificial example here to demonstrate how bootstrapping works for a two sample test.
- the B population is shifted upward from the A populatoin so the means definitely aren't the same.
- in this case the Wilcoxon Rank Sum test would also be suitable

The usual t-test

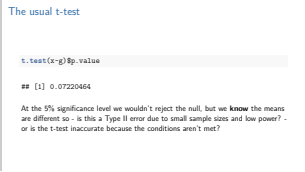
```
t.test(x~g)$p.value
```

```
## [1] 0.07220464
```

At the 5% significance level we wouldn't reject the null, but we **know** the means are different so - is this a Type II error due to small sample sizes and low power? - or is the t-test inaccurate because the conditions aren't met?

Bootstrap Hypothesis Testing

└The usual t-test



- no audio

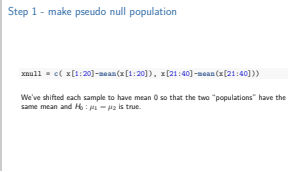
Step 1 - make pseudo null population

```
xnull = c( x[1:20]-mean(x[1:20]), x[21:40]-mean(x[21:40]))
```

We've shifted each sample to have mean 0 so that the two "populations" have the same mean and $H_0 : \mu_1 = \mu_2$ is true.

Bootstrap Hypothesis Testing

Step 1 - make pseudo null population



- no audio

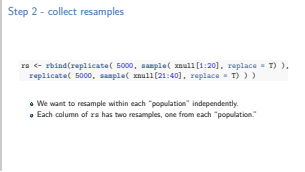
Step 2 - collect resamples

```
rs <- rbind(replicate( 5000, sample( xnull[1:20], replace = T) ),
  replicate( 5000, sample( xnull[21:40], replace = T) ) )
```

- We want to resample within each "population" independently.
- Each column of `rs` has two resamples, one from each "population."

Bootstrap Hypothesis Testing

Step 2 - collect resamples



- no audio

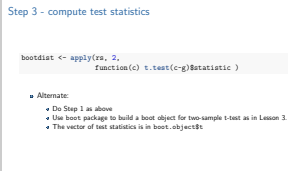
Step 3 - compute test statistics

```
bootdist <- apply(rs, 2,
                  function(c) t.test(c~g)$statistic )
```

- Alternate:
  - Do Step 1 as above
  - Use `boot` package to build a boot object for two-sample t-test as in Lesson 3.
  - The vector of test statistics is in `boot.object$t`

Bootstrap Hypothesis Testing

Step 3 - compute test statistics



- no audio

Step 4 - approximate p-value

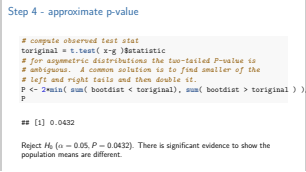
```
# compute observed test stat
toriginal = t.test( x~g )$statistic
# for asymmetric distributions the two-tailed P-value is
# ambiguous.  A common solution is to find smaller of the
# left and right tails and then double it.
P <- 2*min( sum( bootdist < toriginal), sum( bootdist > toriginal ) )
P
```

```
## [1] 0.0432
```

Reject H_0 ($\alpha = 0.05, P = 0.0432$). There is significant evidence to show the population means are different.

Bootstrap Hypothesis Testing

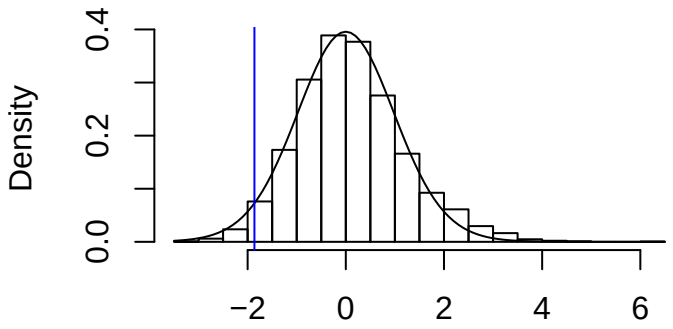
Step 4 - approximate p-value

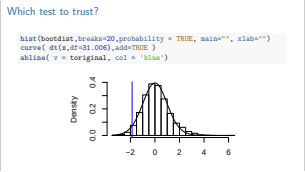


- no audio

Which test to trust?

```
hist(bootdist,breaks=20,probability = TRUE, main="", xlab="")
curve( dt(x,df=31.006),add=TRUE )
abline( v = toriginal, col = 'blue')
```





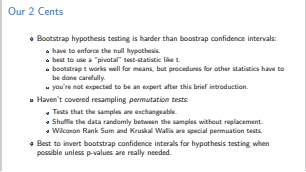
- audio12.mp3
- We can see that the mathematical t-distribution doesn't quite agree with the boostrap distribution of Welch t test statistics.
- With these small samples from skewed populations we don't expect t to be accurate, but the boot distribution accounts for the skewness so we'd go with that one.
- Try doing the wilcoxon rank sum test on this artifical data and see what happens.

Our 2 Cents

- Bootstrap hypothesis testing is harder than boostrap confidence intervals:
  - have to enforce the null hypothesis.
  - best to use a "pivotal" test-statistic like t.
  - bootstrap t works well for means, but procedures for other statistics have to be done carefully.
  - you're not expected to be an expert after this brief introduction.
- Haven't covered resampling *permutation tests*:
  - Tests that the samples are exchangeable.
  - Shuffle the data randomly between the samples without replacement.
  - Wilcoxon Rank Sum and Kruskal Wallis are special permuation tests.
- Best to invert bootstrap confidence interals for hypothesis testing when possible unless p-values are really needed.

Bootstrap Hypothesis Testing

Our 2 Cents



- no audio