Resample Package

Tim Hesterberg
UseR! 2014
Goals

- Easy to use
  - Introductory statistics (and everyone else)
  - In this poster – compare resample to boot
- Plans
- Accurate bootstrap confidence intervals
One-sample bootstrap

x <- rnorm(100); y <- rnorm(100)
data <- data.frame(x, y)

## Mean of a vector
bootstrap(x, mean)
boot(x, function(x, i) mean(x[i]), R = 1000)

## Mean of a column in a data frame
Bootstrap(data, mean(x))
Boot(data, function(data, i) mean(data[i, "x"]), R = 1000)

## Pass arguments to a statistic
bootstrap(x, mean(x, trim = .25))
bootstrap(x, mean, args.stat = list(trim = .25))
boot(x, function(x, i) mean(x[i], trim = .25), R = 1000) # write new fun
Two-sample bootstrap

x1 <- rnorm(30); x2 <- rnorm(50);
x12 <- c(x1, x2); g <- rep(1:2, c(30, 50))
data12 <- data.frame(x = x12, arm = g)

# Three options for resample:
bootstrap2(x1, data2 = x2, mean)
bootstrap2(x12, mean, treatment = g)
bootstrap2(data12, mean(x), treatment = arm)
# The replicates for each sample are saved, and their differences.

# For boot, need to write a function
DiffMeans <- function(df, i) {
  df <- df[i, ]
  with(df, mean(x[arm == 1]) - mean(x[arm == 2]))
}
boot(data12, DiffMeans, strata = data12$arm, R = 1000)
Two-sample permutation test

```r
permutationTest2(x1, data2 = x2, mean)
permutationTest2(x12, mean, treatment = g)
permutationTest2(data12, mean(x), treatment = arm)

# For boot, need a different version of DiffMeans
DiffMeans2 <- function(df, i) {
  df <- df[i, ]
  with(df, mean(x[data12$arm == 1]) - mean(x[data12$arm == 2]))
}
result <- boot(data12, DiffMeans2, sim = "permutation", R = 999)

# Calculate P-value by hand
# Calculate P-value by hand
pvalueL <- (1 + mean(result$t <= result$t0)) / (1 + result$R)
pvalueG <- (1 + mean(result$t <= result$t0)) / (1 + result$R)
2 * min(pValueL, pValueG)
```
One-sample permutation test

size <- data.frame(height = rnorm(100), weight = rnorm(100))
with(size, cor(height, weight))
permutationTest(size, cor(height, weight), resampleColumns = "height",
                alternative = "less")
result <- boot(size, function(data, i) cor(data$height[i], data$weight),
               R = 999)
(1 + mean(result$t <= result$t0)) / (1 + result$R)
Confidence Intervals

• Current:
  • Bootstrap Percentile Interval
    `limits.percentile(x)`
    `boot.ci(x, type = "perc")`

• Future: More accurate intervals
  • BCa interval
    `boot.ci(x, type = "bca")`
  • Improvements on the Percentile Interval
Accuracy of Percentile Interval

• For mean
  • Like \( \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{(n-1)/n} \)
  • Plus partial skewness correction
  • Plus added variability

• More generally
  • Too narrow, under-covers
  • Doesn’t handle skewness right
  • Does exactly the wrong thing for bias
  • First-order correct, errors \( O(1/n^{1/2}) \)
Accuracy of BCa Interval

- For mean
  - Like \( \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{(n - 1)/n} \)
  - Plus skewness correction
  - Plus bias correction
  - Plus added variability

- More generally
  - Too narrow, under-covers
  - Handles skewness and bias right
  - Second-order correct, errors \( O(1/n) \)
Do-it-yourself vs Use a Package

- Do it yourself
  - Ordinary bootstrap sampling
  - Bootstrap percentile interval

- Package
  - Better sampling methods
  - Better confidence intervals
  - …