# StatR 101: Fall 2012

Homework 9

Rod Doe

Wedneday, November 28, 2012

# Point Estimates and Confidence Intervals

## GetCI function

I read about bootstrap confidence intervals and how they yield more realistic values with skewed data. It sounds like a clever idea, so I included the option of using the bootstrap method in this function. It returns a list object with various goodies of interest in it.

GetCI <- function(X, sd, confidencePercent, bootstrap=FALSE, bootstrap.sample.size = 30, bootstrap.iterations = 10000) {

n = length(X)

stdErr = sd / sqrt(n) # standard error of the mean.

meanX = mean(X)

upper = meanX

lower = meanX

ci = 0

alpha = (1 - (confidencePercent/100)) / 2

if (bootstrap == FALSE) {

# Use theoretical distributions.

# Determine which theoretical distribution to use.

if (n < 30) {

# Use Student's t distribution for small samples.

dof = n - 1

tValue = qt(1 - alpha, dof)

ci = tvalue \* stdErr

} else {

# Use the normal distribution for large samples.

zValue = qnorm(1 - alpha, meanX, sd, lower.tail=TRUE)

ci = zValue \* stdErr

}

upper = as.numeric(meanX + ci)

lower = as.numeric(meanX - ci)

}

else {

# Get a bootstrap confidence interval.

a <- numeric(bootstrap.iterations)

for (i in 1:bootstrap.iterations) {

a[i] = mean(sample(X, bootstrap.sample.size, replace=TRUE))

}

lower = as.numeric(quantile(a, alpha))

upper = as.numeric(quantile(a, 1 - alpha))

ci = NA

}

list(mean=meanX, ci=ci, percent=confidencePercent, n=n, alpha=alpha, lower=lower, upper=upper)

}

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Command** | **CI** | **Percent** | **Alpha** | **N** | **Lower** | **Upper** |
| GetCI(W, 38, 95, bootstrap=TRUE) | NA | 95 | 0.025 | 6202 | 17.34889 | 44.14196 |
| GetCI(W, 38, 95) | 49.85679 | 95 | 0.025 | 6202 | -21.01014 | 78.70345 |
| GetCI(W.5, 38, 95) | 47.18323 | 95 | 0.025 | 5 | -37.66323 | 56.70323 |
| GetCI(W.10, 38, 95) | 27.18356 | 95 | 0.025 | 10 | -4.946896 | 49.42023 |
| GetCI(W.30, 38, 95) | 743.6669 | 95 | 0.025 | 30 | -710.9552 | 776.3786 |
|  |  |  |  |  |  |  |
| GetCI(W, 38, 50) | 26.28652 | 50 | 0.25 | 6202 | 2.560141 | 55.13317 |
| GetCI(W, 38, 50, bootstrap=TRUE) | NA | 50 | 0.25 | 6202 | 23.88931 | 32.90319 |
| GetCI(W.5, 38, 50) | 12.58749 | 50 | 0.25 | 5 | -3.067493 | 22.10749 |
| GetCI(W.10, 38, 50) | 8.44437 | 50 | 0.25 | 10 | 13.7923 | 30.68104 |
| GetCI(W.30, 38, 50) | 404.7682 | 50 | 0.25 | 30 | -372.0565 | 437.4798 |

## Count of confidence intervals that do not span the true mean

trueSigma = 38

trueMean = 29

spanTrueMean = numeric(10000)

for (i in 1:length(spanTrueMean))

{

W.X <- sample(W,5)

r = GetCI(W.X, trueSigma, 95)

if ((r$lower < trueMean) && (r$upper > trueMean)) {spanTrueMean[i] = TRUE}

else {spanTrueMean[i] = FALSE}

}

> length(which(spanTrueMean == TRUE))

[1] 9779

Here are the results. I repeated results 5X to get a quick-and-dirty [\*1] reality check.

|  |  |
| --- | --- |
| **Sample Size** | **Confidence intervals that span mean (out of 10000)** |
| 5 | 9779 |
| 5 | 9795 |
| 5 | 9797 |
| 5 | 9802 |
| 5 | 9824 |
| 10 | 9698 |
| 10 | 9712 |
| 10 | 9714 |
| 10 | 9724 |
| 10 | 9727 |
| 30 | 10000 |
| 30 | 10000 |
| 30 | 10000 |
| 30 | 10000 |
| 30 | 10000 |

The smaller sample size exhibited more variation. (This exercise convinced me of the superiority of the bootstrapped confidence interval.)

[\*1] – The Rule of Five occasionally provides a nice reality check. This rule states that there is a 93.75% probability that the median of any population is between the largest and smallest values in a random sample. The way this rule works is quite simple – there is a ½ probability that the true population median is larger than the smallest sample. It then follows that there is a 1/32 probability that the true population median is larger than all five sorted random samples. The logic mirrors on the left, with a 1/32 probability that the true population median is smaller than all five random samples.

> 1 - 1/32 - 1/32

[1] 0.9375

Source: [www.HowToMeasureAnything.com](http://www.HowToMeasureAnything.com)

# Hypothesis Tests

We observed that the mean waiting time between 10 earthquakes in a row was 15 minutes, compared to a known mean and standard deviation of 29 and 39 minutes respectively.

## Null Hypothesis

The mean earthquake period of the earth is 29 minutes with a standard deviation of 39 minutes.

H0: mean earthquake period = 29 minutes

HA: mean earthquake period < 29

## Alternative Hy**pothesis**

The mean earthquake period is becoming more frequent.

## Z statistic (Normal distribution)

> xbar = 15

> mu = 29

> sigma = 39

> n = 10

> z <- (xbar - mu)/sigma/sqrt(n)

> z

[1] -0.1135177

The Z statistic must be outside of +/- 1.96 to state with 95% confidence that the mean earthquake is less than 29 minutes. This observation is -1.96 <= -0.114 <= 1.96. As such, this observation can be attributed to chance alone, *so I reject the alternative hypothesis*. It’s not the end of the world as we know it, and I feel fine.

The p-value of this experiment can be expressed thus:

> pnorm(xbar, mu, sigma/sqrt(n))

[1] 0.1281507

This value is within the alpha values of 0.025 and 0.975. Were it outside of the alpha values, it would be statistically significant.