Resampling: Nonparametric Bootstrapping

Mindy Lee

11/27/2017

Introduction

In class, we learned all about the fundamental concepts of R computation, mostly on how to clean and tidy our data for further analysis. So far, we have only looked at datasets that provide all the information we need for the analysis, but this is not always the case. You may face situations such as you only have one sample's data with no information on the population, yet you need to draw conclusions for the entire population. How would you perform analysis on your data then? As simple as it sounds: you estimate. In this post, I will talk about nonparametric bootstrapping, a resampling method that is commonly used in estimation. You can read this post as a preview of what you will be learning in future statistic courses (e.g. Stat 135) that requires R computation.

Background

Nonparametric bootstrapping is a useful tool when the distribution of a population is unknown. It is often implicated on really small samples when normal distribution is not applicable. We can resample using nonparametric bootstrapping method to find estimated standard error, form confidence interval, and perform hypothesis testing.

Nonparametric boostrapping generally follows the following steps:

1. resample from the original sample with replacement
2. find the desired statistics (e.g. mean) from the new sample
3. repeat the first two steps for as many times to get your bootstrapped estimated samples

Examples

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

For this example, let's reuse the fictional Stat 133 grades we have already cleaned up in HW4, since we all spent so much time working on it. Simply load the CLEAN data 'cleanscores.csv' from your HW4 data folder.

```
dat <- read.csv("../data/cleanscores.csv")
```

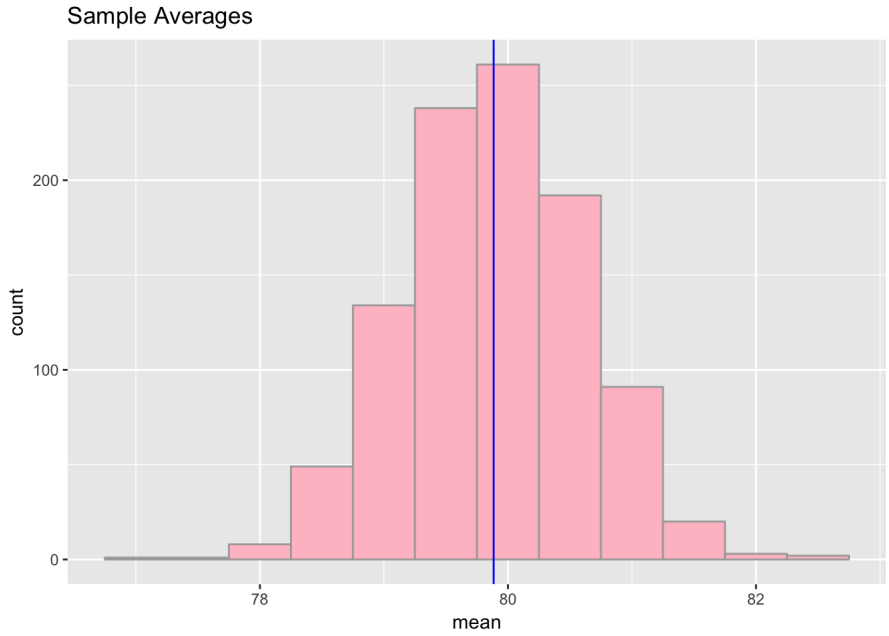
Let's assume that Stat 133 is a class that has been around for a long long time. I want to know the average overall grades of all the past semesters (= population). How can we estimate that using just the data we have?

```
set.seed(34)
# write a function that finds the mean af the sample you draw
find_mean <- function(){
             resample <- dat$Overall %>% sample(334, replace=TRUE)
             mean(resample)
}
# repeat the sampling and finding mean process 1000 times
means <- replicate(1000, find_mean())

# the mean of the sample averages is the estimated population mean
mean(means)
```

```
## [1] 79.88441
```

```
# we can plot the sample averages (aka sampling distribution of averages)
# the blue line indicates the mean of the sample averages
means.df <- data.frame(means)
means.df %>% ggplot(aes(x = means)) +
  geom_histogram(binwidth = 0.5, col = "dark grey", fill ="pink") +
  geom_vline(xintercept = mean(means), color = "blue") +
  labs(title = "Sample Averages", x = "mean", y ="count")
```



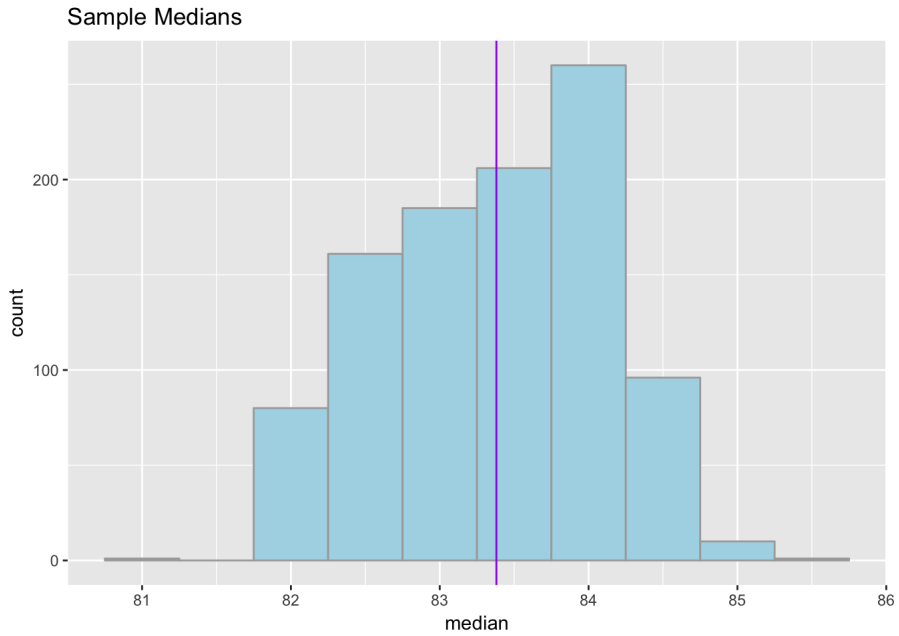
Not limited to mean, you can use nonparametric bootstrapping to estimate other statistics as well. Let's now estimate the median of overall grades over all the past semesters using similar codes as estimating means.

```
set.seed(97)
# write a function that finds the mean af the sample you draw
find_median <- function(){
            resample <- dat$Overall %>% sample(334, replace=TRUE)
            median(resample)
}
# repeat the sampling and finding median process 1000 times
medians <- replicate(1000, find_median())

# the mean of the sample medians is the estimated population mean
mean(medians)
```

```
## [1] 83.38067
```

```
# plot the medians, and the line indicates the estimated median
medians.df <- data.frame(medians)
medians.df %>% ggplot(aes(x = medians)) +
  geom_histogram(binwidth = 0.5, col = "dark grey", fill ="light blue") +
  geom_vline(xintercept = mean(medians), color = "purple") +
  labs(title = "Sample Medians", x = "median", y ="count")
```



An interesting thing we can also do with nonparametric bootstrapping is resample for different confidence interval. If we get 1000 different 95% CI, how many of them cover the true average? Note that in practice you would be unable to do this since you only get one sample.

```r
set.seed(18)

# function that finds CI from the sample
find_CI <- function(){
           CI_resample <- sample(dat$Overall, 100)
           g_mean <- mean(CI_resample)
           g_sd <- sd(CI_resample)/sqrt(100)
           z <- abs(qnorm(.05/2))
           CI.wt <- c(g_mean - z * g_sd, g_mean + z * g_sd)
}

# repeat to get 1000 different CI
CIs <-  replicate(1000, find_CI())

# the actual mean of the overall grades
true_mean <- mean(dat$Overall)

# how many of the CIs cover the true mean
cover <- sum(cbind(CIs[1, ] <= true_mean & CIs[2, ] >= true_mean))
cover
```

```
## [1] 964
```

Discussion

In this post, I introduced a resampling method, nonparametric boostrapping. The examples demonstrate how nonparametric bootstrapping is carried out. After getting your reampled samples, a more in depth statistical analysis can be performed, which you will learn in future statistic courses. For now, just know the following: There will be times when you will have little information on the population or have a sample that is too small to perform analysis on. This is when nonparametric bootstrappings come in handy. You can resample from your original sample, estimate statistics from the resamples, and perform hypothesis testings.

Again, read this post as a preview of Stat 135 if you haven't taken that course. You will be learning a more extensive version of bootstrapping, not limited to nonparametric bootstrapping but also parametric bootstrapping. If you have already taken Stat 135, this is a great post to refresh the main concept of nonparametric boostrapping.

I hope you learned something from this post! Thank you for reading!

References

1. https://www.zoology.ubc.ca/~schluter/R/resample/
2. http://faculty.washington.edu/kenrice/rintro/sess06.pdf
3. http://www.stat.wisc.edu/~larget/stat302/chap3.pdf
4. http://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf
5. https://github.com/ucb-stat133/stat133-fall-2017/blob/master/data/rawscores.csv
6. https://www.zoology.ubc.ca/~schluter/R/resample/
7. https://www.r-bloggers.com/im-all-about-that-bootstrap-bout-that-bootstrap/