# CORE S119 FA24 Prog 5

Grusha Prasad

2024-10-22

## Introduction

In order to compute bootstrapped confidence intervals for the cards data, you sampled cards with replacement multiple times. In this notebook, you will learn how to use **loops** in R to do the same task over and over again (just like sampling with replacement mover and over again). You will then use loops to compute bootstrapped confidence intervals and p-values.

## For loops

The following code uses a for loop to print “hello” 10 times.

```
for(i in c(1:10)){
  print('hello')
}
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
## [1] "hello"
```

**Coding Q1** Write code to print “hello” 15 times

```
## WRITE YOUR CODE BELOW THIS LINE
```

**Coding Q2** Write code to print “hi” 5 times

```
## WRITE YOUR CODE BELOW THIS LINE
```

**Written Q1** What do you think the following code does?

**Answer**

```
for(i in c(1:13)){
  num = rbinom(n= 1, size = 100, prob = 0.5)
  print(num)
}
## [1] 44
## [1] 48
## [1] 56
## [1] 50
## [1] 44
## [1] 53
## [1] 51
## [1] 58
## [1] 44
## [1] 48
## [1] 54
## [1] 49
## [1] 50
```

**Written Q2** What do you think the following code does?

**Answer**

```
total = 0
num_iters = 13
exp_size = 1000


for(i in c(1:num_iters)){
  num = rbinom(n = 1, size = exp_size, prob = 0.5)
  total = total + num
}

print(paste('total is:', total))
## [1] "total is: 6399"
print(paste('% success is:', (total*100)/(num_iters*exp_size)))
## [1] "% success is: 49.2230769230769"
```

**Coding Q3** Write code **that uses a for loop** to print 17 different samples from a random distribution with mean 7 and sd 0.5

```
## WRITE YOUR CODE BELOW THIS LINE
```

**Coding Q4** Write code **that uses a for loop** to print the sum of 17 different samples from a random distribution with mean 7 and sd 0.5

```
## WRITE YOUR CODE BELOW THIS LINE
```

## Bootstrapped confidence intervals

Let us plot bootstrapped confidence intervals for the `titanic` dataset.

```
titanic = read.csv('titanic.csv')
```

The following code takes the titanic data, create a new dataframe many times by sampling with replacement, and combines all of the dataframes together using a function called `rbind`

```
combined = data.frame()

for(i in c(1:100)){
  curr_data = titanic %>%
    select(Survived) %>%
    mutate(Survived = sample(Survived, size = n(), replace=TRUE)) %>%
    summarize(prop_survived = mean(Survived)) %>%
    mutate(exp = i)

  combined = rbind(combined, curr_data)
}
```

**Written Q3** How many rows will `combined` have after you run the code? Why?

**Answer**

**Written Q4** In the code above what does `replace=TRUE` do? Why do we need this?

**Answer**

**Coding Q5** Use the `combined` dataframe to compute the 95% bootstrapped confidence interval.

*Hint: Review the `quantile()` function from confidence interval demo*

```
## WRITE YOUR CODE BELOW THIS LINE
```

We can now look at how the proportion of survived differed by gender.

```
combined2 = data.frame()

for(i in c(1:100)){
  curr_data = titanic %>%
    select(Survived, Sex) %>%
    group_by(Sex) %>%
    mutate(Survived = sample(Survived, size = n(), replace=TRUE)) %>%
    summarize(prop_survived = mean(Survived)) %>%
    mutate(exp = i)

  combined2 = rbind(combined2, curr_data)
}
```

**Written Q4** How many rows will `combined2` have after you run the code? Why?

**Answer**

**Written Q5** Why does the code group_by `Sex` before sampling?

**Answer**

**Coding Q6** Use the dataframe `combined2` to create a **summary** dataframe called `combined2_summ` which has two rows (one for male, one for female), and three columns: `mean` (the mean of `prop_survived` across all exp), `lower` (the lower end of the 95% confidence interval for `prop_survived` across all exp) and `upper` (the upper end of `prop_survived` across all exp).

*Hint: if you want just one end of the quantile, you can pass in just one number into the function*

```
## WRITE YOUR CODE BELOW THIS LINE
```

**Coding Q7** Generate a plot from `combined2_summ` with `prop_survived` on the y-axis, `Sex` on the x-axis. You should use the `upper` and `lower` columns for the error bars.

```
## WRITE YOUR CODE BELOW THIS LINE
```

## Bootstrapped p-values

If we are interested in testing whether the proportion of people who survived significantly differs by Sex, we can adopt the following null hypothesis: there is no difference in proportion survived.

The following code can be used to generate the null distribution that is consistent with the null hypothesis.

```
null_dist = data.frame()

for(i in c(1:100)){
  curr_data = titanic %>%
    select(Survived, Sex) %>%
    mutate(Survived = sample(Survived, size = n(), replace=TRUE)) %>%
    group_by(Sex) %>%
    summarize(prop_survived = mean(Survived)) %>%
    mutate(exp = i)

  null_dist = rbind(null_dist, curr_data)
}
```

**Written Q6** The code above is nearly identical to the code we used to generate `combined2` but differs in one crucial way. What is the difference? Why is this difference important if we want the null distribution?

### Answer

For each bootstrapped experiment (i.e., `exp`) under the null hypothesis, the code below computes the difference in proportion survived. It also computes true difference in proportion of male and female passengers who survived.

```
null_dist_wide = null_dist %>%
  spread(key=Sex, value=prop_survived) %>%
  mutate(diff = female-male)

titanic_summ_wide = titanic %>%
  select(Survived, Sex) %>%
  group_by(Sex) %>%
  summarize(prop_survived = mean(Survived, na.rm=TRUE)) %>%
  ungroup() %>%
  spread(key=Sex, value=prop_survived) %>%
  mutate(diff = female-male)

true_prop_diff = titanic_summ_wide$diff
```

The code below now adds a column that checks whether the difference in each experiment was "more or as extreme" as the true prop difference.

```
null_dist_wide = null_dist_wide %>%
  mutate(extreme = ifelse(abs(diff) > abs(true_prop_diff), 1, 0))
```

**Written Q7** What does the `abs()` function do? (Look it up on Google if you are uncertain). Why are we using it here?

**Answer**

**Coding Q8** Use `null_dist_wide` to calculate the p-value that indicates whether the difference in proportion survived by sex is statistically significant.

```
## WRITE YOUR CODE BELOW THIS LINE
```

**Written Q8** Based on the p-value, do you think the difference is significant?

**Answer**

## Homework

For homework, you will explore if the age of female passengers was statistically different from the age of male passengers.

*Hint: carefully look through what we did earlier in this notebook and see which parts you would need to modify for these question*

1. Generate bootstrapped samples, and use these samples to generate a plot where`Age` is on the y-axis and `Sex` is on the x-axis. The plot should include one point each, for male and female, which is the mean age across the bootstrapped samples. It should also include the 95% bootstrapped confidence interval for each of these points.

2. Compute the p-value that indicates whether the mean difference in age observed in the data is statistically significant.

3. Based on the p-value, do you think there is a statistically significant difference in the age of male and female passengers? Why or why not?