Homework 3

Mengyu Zhang / mz2777

4/17/2020

Your homework for bootstrap methods is to complete the two exercises on the slides Lecture 7.pdf, page 2 and 4; When you implement them in R, please use the parallel computing codes for bootstrapping replicates.

Problem 1

A randomized trial on eye treatment. Two laser treatments were randomized to eyes on patients. The response is visual acuity, measured by the number of letters correctly identified ina standard eye test. Some patients had only one suitable eye, and they received one treatment allocated at random. There are 20 patientswith paired data and 20 patients for whom just one observation is available, so we have a mixture of paired comparison and two-sampledata.

(1) How would you analyze the data to investigate whether the expected accuracies between the two treatments are different.

Answer:

Using Bootstrap to generate the null distributionsfor hypothesis testing. Data is divided into three parts which are paired data, 10 patients only treated by red laser and 10 patients only treated by blue laser, and then bootstrapping is applied on three parts respectively in order to maintian the empirical distribution.

Also, we need to adjust the data to generate samples under null hypothesis which is equal mean.

Weighted t statistic is applied to the algorithm. The first term is for two sample t test and second is paired t test.

$$T_{\text{wgt}} = \sqrt{\gamma}\frac{\bar{X}_{0,U} - \bar{X}_{1,U}}{\sqrt{S_{0,U}^2/n_0 + S_{1,U}^2/n_1}} + \sqrt{1-\gamma}\frac{\bar{D}}{S_D/\sqrt{n}}$$

```
blue <- c(4,69,87,35,39,79,31,79,65,95,68,62,70,80,84,79,66,75,59,77,36,86,39,85,74,72,69,85,85,72)
red <-c(62,80,82,83,0,81,28,69,48,90,63,77,0,55,83,85,54,72,58,68,88,83,78,30,58,45,78,64,87,65)
acui<-data.frame(str=c(rep(0,20),rep(1,10)),red,blue)
```

```
# test statistic computation
teststat = function(x,y,d){
  x = as.matrix(x)
  y = as.matrix(y)
  d = as.matrix(d)
return(sqrt(20/40) * (mean(x) - mean(y))/(sqrt(var(x)/10 + var(y)/10)) + sqrt(1-20/40)*mean(d)/(sqrt(va
}

hypo_test <- function(data, nboot=10000){
  x = data %>% filter(str == 1) %>% dplyr::select(red)
  y = data %>% filter(str == 1) %>% dplyr::select(blue)
  d = data %>% filter(str == 0) %>%  mutate(d = red - blue) %>%  dplyr::select(d)
```

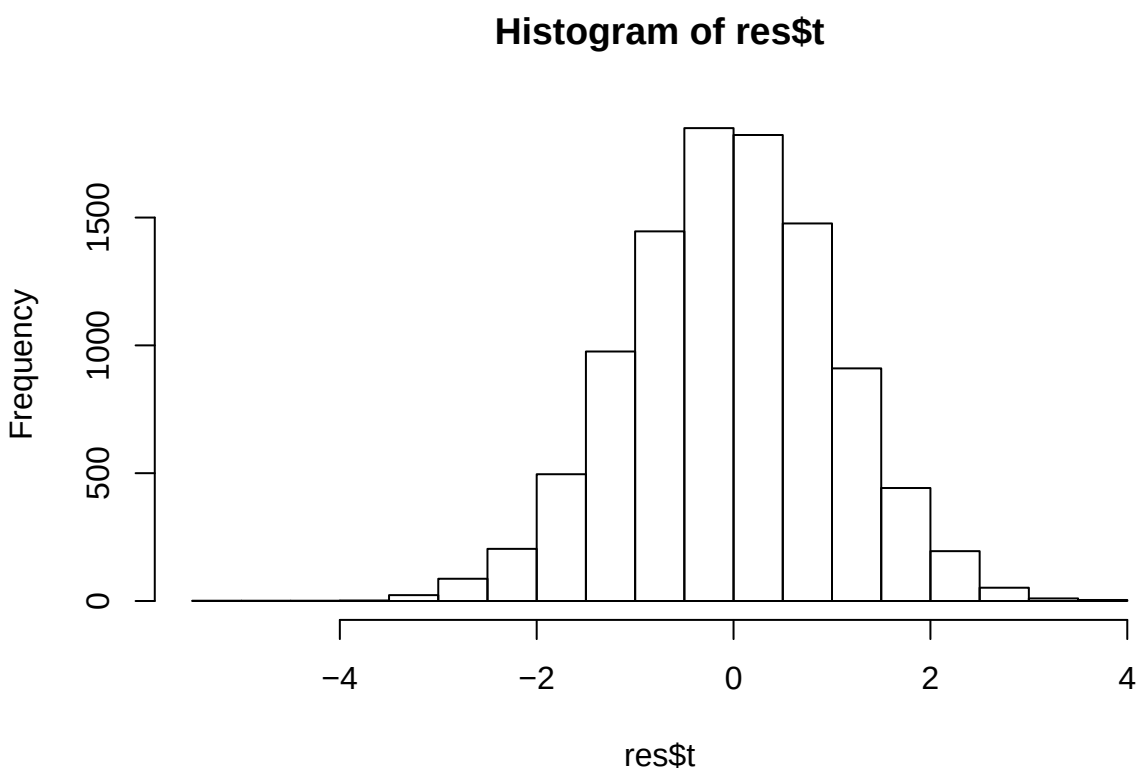
```r
  # The mean of the combined
  combmean <- mean(c(data[,2],data[,3]))

  # split data
  pair_pat = data %>%
    filter(str == 0) %>%
    mutate(adj_red = red - mean(red) + combmean,
           adj_blue = blue - mean(blue) + combmean,
           d = adj_red -adj_blue) %>%
    dplyr::select(d)


  red_pat = data %>%
    filter(str == 1) %>%
    mutate(adj = red - mean(red) + combmean) %>%
    dplyr::select(adj)

  blue_pat = data %>%
    filter(str == 1) %>%
    mutate(adj = blue - mean(blue) + combmean) %>%
    dplyr::select(adj)


  nCores <- 10  # to set manually
  registerDoParallel(nCores)
  teststatvec <- vector()
  out <- foreach(i = 1:nboot, .combine = c) %dopar% {

    new_pair = sample(as.matrix(pair_pat), replace = T)
    new_red = sample(as.matrix(red_pat), replace = T)
    new_blue = sample(as.matrix(blue_pat), replace = T)

    teststatvec <- as.matrix(teststat(x=new_red, y=new_blue, d=new_pair))
    teststatvec
  }

  return(list(bootpval = sum(rep(teststat(x,y,d),nboot) < out)/nboot, t = out, obs_t = as.matrix(teststa
}
data = acui

set.seed(123)
res = hypo_test(acui) #### results are different
p_value = res$bootpval
p_value
```

```
## [1] 0.7149
```

```r
res$obs_t
```

```
##            red
## red -0.6113838
```

```r
hist(res$t)
```



(2) Use bootstrap to construct confidence interval of the treatment effect. What is your conclusion?

Answer:

Get estimate of standard error by bootstrap first, then get standard confidence interval.

```r
CI <- function(data, nboot=10000){
  # The mean of the combined
  combmean <- mean(c(data[,2],data[,3]))

  # split data
  pair_pat = data %>%
    filter(str == 0) %>%
    mutate(adj_red = red - mean(red) + combmean,
           adj_blue = blue - mean(blue) + combmean) %>%
    dplyr::select(adj_red, adj_blue)


  red_pat = data %>%
    filter(str == 1) %>%
    mutate(adj = red - mean(red) + combmean) %>%
    dplyr::select(adj)

  blue_pat = data %>%
```

```
    filter(str == 1) %>%
    mutate(adj = blue - mean(blue) + combmean) %>%
    dplyr::select(adj)


  nCores <- 10  # to set manually
  registerDoParallel(nCores)
  meandiffvec <- NULL
  out <- foreach(i = 1:nboot, .combine = c) %dopar% {
    #sampling
    new_pair = pair_pat[sample(c(1:20), replace = T),]
    new_red = sample(red_pat, replace = T)
    new_blue = sample(blue_pat, replace = T)

    # frame new data
    red_ = c(new_pair[,1], as.matrix(new_red))
    blue_ = c(new_pair[,2], as.matrix(new_blue))

    meandiffvec = mean(blue_) - mean(red_)
    meandiffvec
  }
  bootse = sqrt(var(out))
  interval = data.frame(t(c(mean(out) + c(0,qnorm(0.025),-qnorm(0.025)) * bootse)))
  colnames(interval) = c("point_estimate", "low_bound", "higher_bound")
  return(list(CI = interval, SE = bootse))
}

set.seed(123)

CI(acui)
```

```
## $CI
##   point_estimate low_bound higher_bound
## 1    -0.07690333 -7.700214     7.546408
##
## $SE
## [1] 3.889516
```

0 is included in the confidence interval, which means true difference may be 0. Therefore, based on the p-value $= 0.7149 > 0.05$ we get from part 1, we fail to reject the null and can conclude that at 0.05 significant level, the mean for blue lazer treatment is equal to the mean for red lazer treatment.

Problem 2

The Galaxy dataconsist of the velocities (in km/sec) of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. The structure in the distribution of velocities corresponds to the spatial distribution of galaxies in the far universe. In particular, a multimodal distribution of velocities indicates a strong heterogeneity in the spatial distribution of the galaxies and thus is seenas evidence for the existence of voids and superclusters in the far universe.

Anwser

bootstrap algorithm

1. draw B bootstrap samples if size n from $\hat{f}_{K,h_1}(x)$
2. for each bootstrap, find $h_1^{*(b)}$, the smallest h for which this bootstrap sample has just 1 mode
3. approximate p-value of test is $\frac{\#h_1^{*(b)} > h_1}{B}$

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

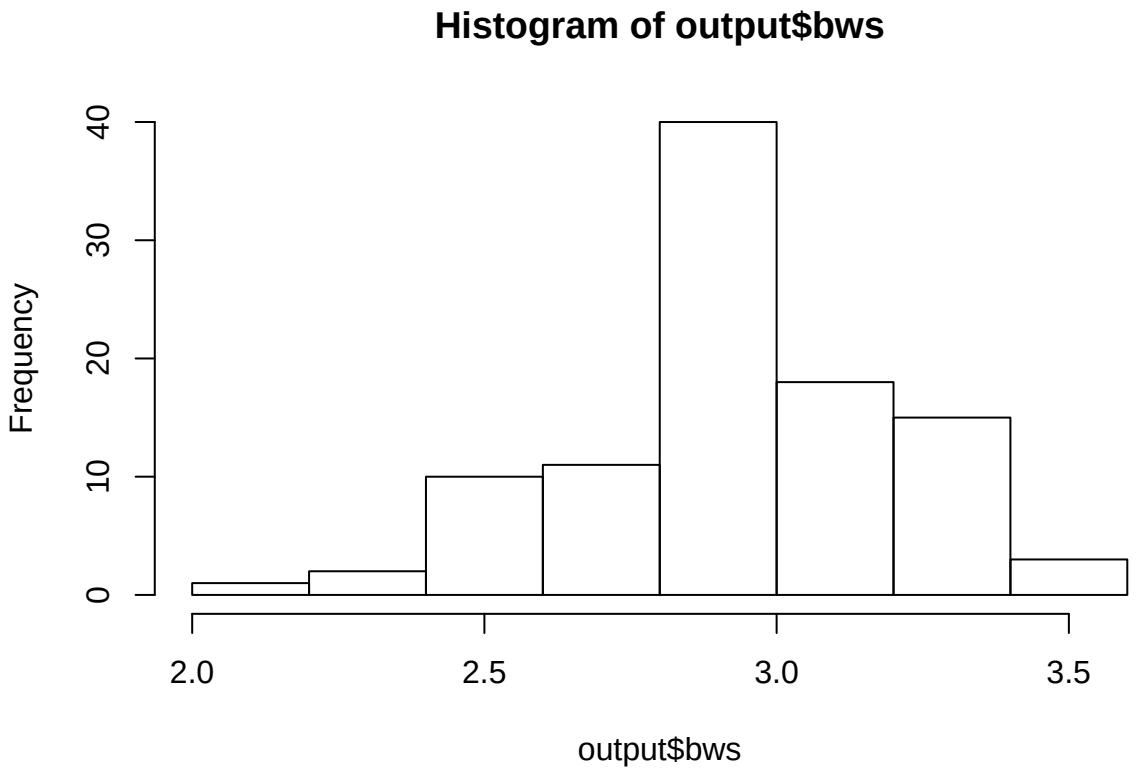
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
data(galaxies)


#calculate the number of modes in the density
num_modes <- function(data, bw){
  den = density(data, bw=bw)
  den.s = smooth.spline(den$x, den$y, all.knots=TRUE, spar=0.8)
  s.1 = predict(den.s, den.s$x, deriv=1)
  nmodes = length(rle(den.sign <- sign(s.1$y))$values)/2
  return(nmodes)
}

hypo_gala <- function(data, B, N){

  # h1
  i = 0
  n_mode = 2
  while (n_mode > 1) {
    i = i + 0.01
    n_mode = num_modes(data/1000, bw = i)
  }
  bw = i

  # generate bootstrap samples from estimated kernal density with bw
  dens = density(data, bw=bw)
  res = rerun(B, sample(data/1000, size = N, replace = TRUE) + rnorm(N, dens$bw))

  # get bws
  nCores <- 10  # to set manually
  registerDoParallel(nCores)

  out <- foreach(j = 1:B, .combine = c) %dopar% {
    i = 0
```

```
    n_mode = 2
    while (n_mode > 1) {
      i = i + 0.01
      n_mode = num_modes(res[[j]], bw = i)
    }
    i
  }

  return(list(pval = sum(out > bw)/B, bws = out))
}


output = hypo_gala(data = galaxies, B = 100, N = 82)
output
```

```
## $pval
## [1] 0.33
##
## $bws
##   [1] 2.59 3.28 3.51 2.95 2.49 2.62 3.12 3.39 2.74 3.06 2.96 2.83 2.75 3.22
##  [15] 2.88 3.15 2.92 2.81 3.41 3.33 3.08 3.38 3.04 2.83 2.83 2.80 2.68 2.93
##  [29] 2.90 3.26 3.32 2.96 3.23 3.12 2.83 2.23 3.10 2.64 2.60 2.74 2.95 3.34
##  [43] 2.43 2.90 3.27 3.20 3.01 2.96 2.72 2.89 2.46 2.47 2.81 2.93 3.44 2.73
##  [57] 3.18 3.09 2.96 3.31 3.04 3.12 2.56 3.19 2.45 2.51 2.71 2.92 2.91 2.83
##  [71] 2.97 3.16 2.85 2.82 3.33 2.92 2.18 2.85 3.27 2.90 3.12 3.08 2.92 2.97
##  [85] 2.81 2.86 2.33 2.59 2.91 2.86 2.86 2.94 2.96 3.29 3.26 2.91 2.80 2.92
##  [99] 3.16 2.86
```

```
hist(output$bws)
```



P-value is $0.33 > 0.05$, so we can not reject the null and conclude that at significant level 0.05, the number of modes of density of velocity for galaxies is 1.