Homework 3: boostrap methods

Alyssa Vanderbeek

4/17/2020

```
cores = detectCores()
clust <- makeCluster(cores[1] - 1)
registerDoParallel(clust)
```

Problem 1

```
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)
```

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

The question of interest is whether there is a meaningful difference in the reading accuracies between the treatment groups. Since we have both paired and unpaired data, we need to test differences on these two sets with a paired t-test and two-sample t-test, respectively. For both datasets, we boostrap the adjusted observations (observed - expected + pulled mean).

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

```
# Total sample size is 40
paired = 1:20
indep = 21:30

boottest = function(paired.index, indep.index, alpha = 0.05, nboot = 200) {

  ####### Paired observations
  paired.diff = blue[paired.index] - red[paired.index] # observed paired differences
  diff.adj = paired.diff - mean(paired.diff) # adjusted observations

  teststatvec.paired = rep(NA, nboot)
  teststatvec.paired <- foreach(i = 1:nboot, .combine = rbind) %dopar% {
    diff.boot <- sample(diff.adj, replace = TRUE) # boot sample
    diff.bar = mean(diff.boot) # mean
    diff.se = sd(diff.boot) / length(paired.index) # SE

    z = qnorm(1 - (alpha/2))
    lower.ci = diff.bar - z*diff.se
    upper.ci = diff.bar + z*diff.se
    #stat = diff.bar / diff.se # t-statistic
```

```
    out <- c(diff.bar, lower.ci, upper.ci) # vector of the sampled mean diff, lower CI, upper CI
    out
  }

  ####### Unpaired observations
  x = blue[indep.index] # observed blue
  y = red[indep.index] # observed red
  mu = mean(c(x, y)) # pulled mean

  # adjusted observations
  x.adj = x - mean(x) + mu
  y.adj = y - mean(x) + mu

  teststatvec.indep = rep(NA, nboot)
  teststatvec.indep <- foreach(i = 1:nboot, .combine = rbind) %dopar% {
    xsamp <- sample(x.adj, replace = TRUE)
    ysamp <- sample(y.adj, replace = TRUE)
    d = xsamp - ysamp

    bar = mean(d)
    se = sd(d) / length(indep.index)

    z = qnorm(1 - (alpha/2))
    lower.ci = bar - z*se
    upper.ci = bar + z*se

    out <- c(bar, lower.ci, upper.ci) # # vector of the sampled mean diff, and test statistic
    out
  }

  return(list(paired = teststatvec.paired,
              indep = teststatvec.indep))

}

set.seed(1)
boot1 = boottest(paired, indep, nboot = 200)

apply(boot1[["paired"]], MARGIN = 2, FUN = mean) # average and CI for paired data
```

```
## [1]  0.769000 -1.783619  3.321619
```

```
apply(boot1[["indep"]], MARGIN = 2, FUN = mean) # average and CI for independent data
```

```
## [1]  2.878000 -1.837701  7.593701
```

Results suggest that for both the paired and independent samples, there is little evidence of a difference in reading accuracies between treatment groups (Paired: 0.42, 95% CI [-2.03, 2.88]; Independent: 2.38, 95% CI -2.36, 7.12]).

Problem 2

The Galaxy data consist 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 seen as evidence for the existence of voids and superclusters in the far universe.

Statistically, the question of multimodality can be formulated as a test problem $H_0 : n_{\text{modes}} = 1$ vs. $H_1 : n_{\text{modes}} \geq 1$.

Considered nonparametric kernel density estimates $\hat{f}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$

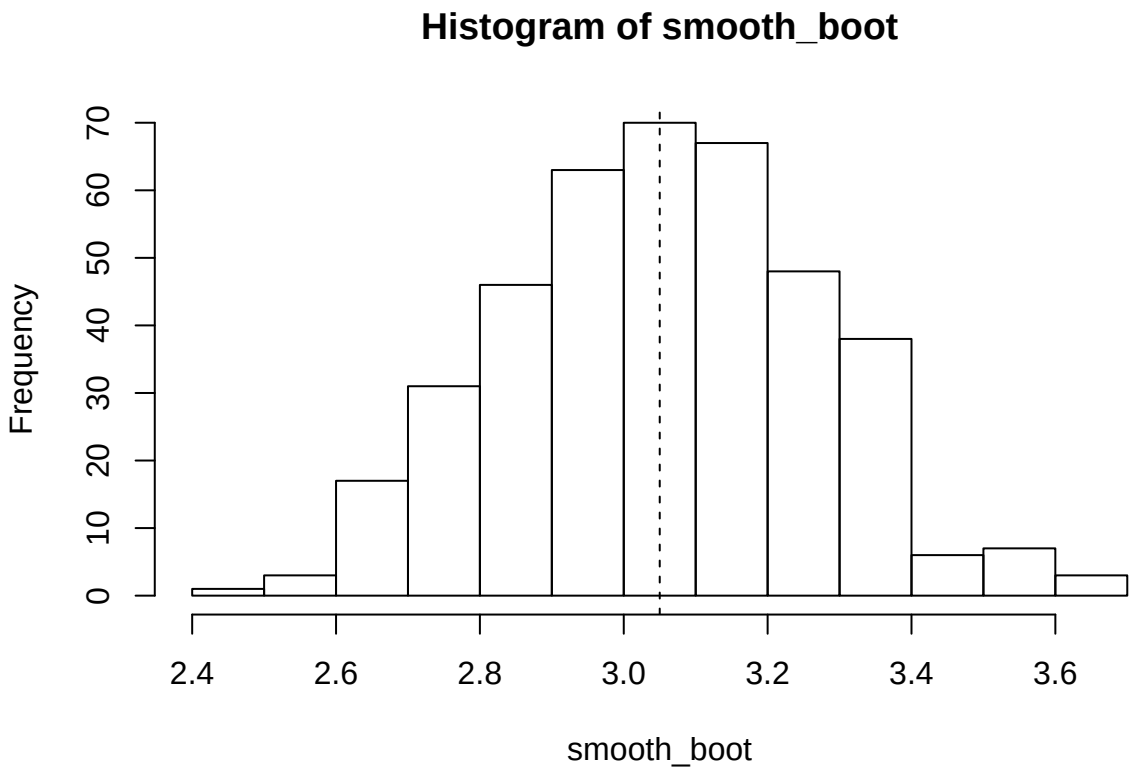
It can be shown that the number of modes in $\hat{f}_{K,h}(x)$ decreases as h increase. Let H_1 be the minimal bandwidth for which $\hat{f}_{K,H_1}(x)$ is unimodal. In the galaxy data, $h_1 = 3.05$.

Since multimodal densities need more smoothing to become unimodal, the minimal bandwidth H_1 can be used as a test statistic, and one reject the null hypothesis if $Pr(H_1 > h_1) \leq \alpha$.

We need to know the distribution of H_1 under the null.

```
data(galaxies)

#calculate the number of modes in the density
n.modes = function(data, bw) {
  den <- density(data/1000, 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)
}

# Calculate h*
get.h = function(data){
  samp = sample(data, replace = TRUE)
  bw = 1
  n = n.vec = n.modes(samp, bw)

  while (n != 1) {
    n = n.modes(samp, bw)
    bw = bw + 0.01

    n.vec = append(n.vec, n)
  }

  bw = floor(bw * 100) / 100
  return(bw)

}


#############
B = 400
h1 = 3.05

set.seed(1)
smooth_boot = foreach(i = 1:B, .combine = c) %dopar% {
  get.h(galaxies)
}
```

```
hist(smooth_boot)
abline(v = h1, lty = 2)
```




```
p.val = sum(smooth_boot > h1) / B
p.val
```

```
## [1] 0.5175
```

The p-value of this test is 0.5175, suggesting that at a bandwidth of 3.059075, taken as the mean of the bootstrap estimates, we are comfortable saying that the density is unimodal.