lab8.R

Feipeng Huang

2022-11-02

```
veg =
read.csv("/Users/stonehuang/Documents/environmental_data/data/vegdata.csv")
dat_bird =
read.csv("/Users/stonehuang/Documents/environmental_data/data/bird.sub.csv")
dat_habitat =
read.csv("/Users/stonehuang/Documents/environmental_data/data/hab.sub.csv")
require(palmerpenguins)
```

```
## Loading required package: palmerpenguins
```

```
penguin_dat = droplevels(subset(penguins, species != "Gentoo"))
t.test(flipper_length_mm ~ species, data = penguin_dat, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm by species
## t = -5.7804, df = 119.68, p-value = 3.025e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.186534
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                189.9536                195.8235
```

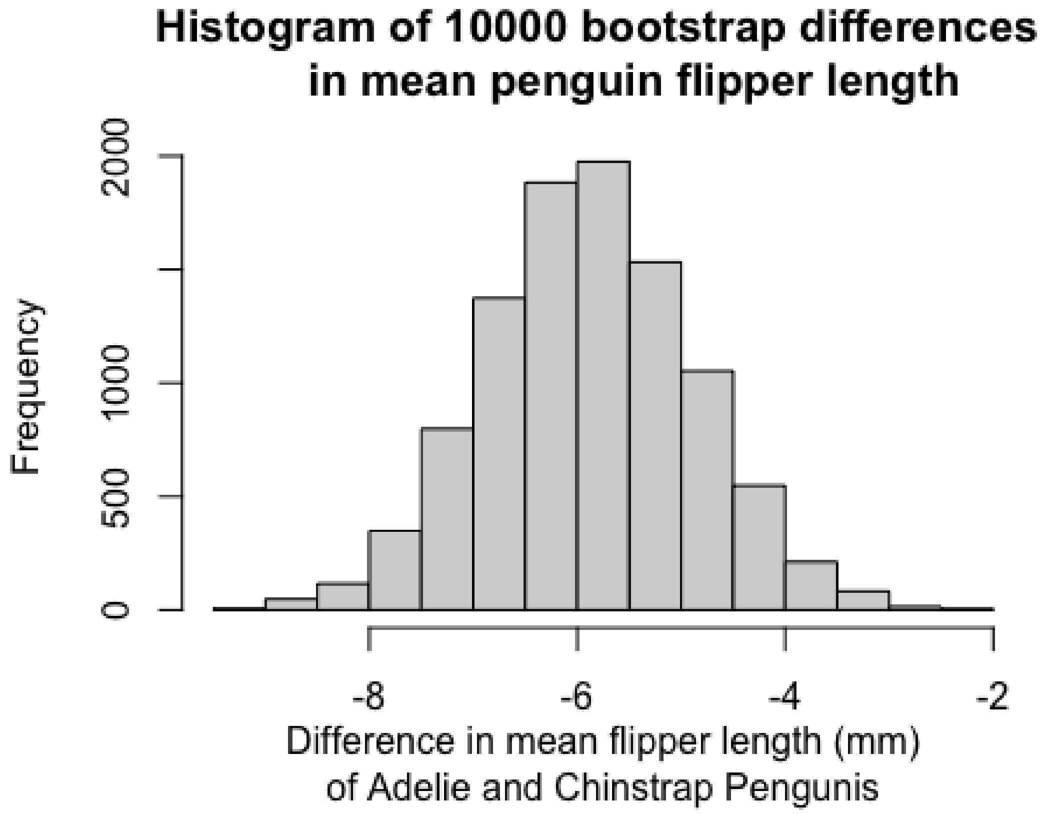
```
#install.packages("simpleboot")
require(simpleboot)
```

```
## Loading required package: simpleboot
```

```
## Simple Bootstrap Routines (1.1-7)
```

```
dat_adelie = subset(penguins, species == "Adelie")
dat_chinstrap = subset(penguins, species == "Chinstrap")
boot_mean = function(x, i)
{
  return(mean(x[i], na.rm = TRUE))
}
pen_boot = two.boot(dat_adelie$flipper_length_mm,
dat_chinstrap$flipper_length_mm, boot_mean, 10000, student = FALSE, weight =
NULL)
str(pen_boot)
```

```
## List of 12
##  $ t0       : num -5.87
##  $ t        : num [1:10000, 1] -6.43 -5.81 -6.1 -5.18 -4.8 ...
##  $ R        : num 10000
##  $ data     : int [1:220] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ seed     : int [1:626] 10403 1 1878503375 -718147370 -1323315072 -
1509060925 -1822959519 487631056 -1559371218 1432003545 ...
##  $ statistic:function (x, idx)
##  $ sim      : chr "ordinary"
##  $ call     : language boot(data = c(sample1, sample2), statistic =
boot.func, R = R, strata = ind,      weights = weights)
##  $ stype    : chr "i"
##  $ strata   : num [1:220] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weights  : num [1:220] 0.00658 0.00658 0.00658 0.00658 0.00658 ...
##  $ student  : logi FALSE
##  - attr(*, "class")= chr "simpleboot"
##  - attr(*, "boot_type")= chr "boot"

t = pen_boot[["t"]]
##########Q1##########
sd = sd(t)
#sd = 1.021529
##########Q2##########
hist(t, xlab = "Difference in mean flipper length (mm)
of Adelie and Chinstrap Pengunis", main = "Histogram of 10000 bootstrap
differences
     in mean penguin flipper length")
```



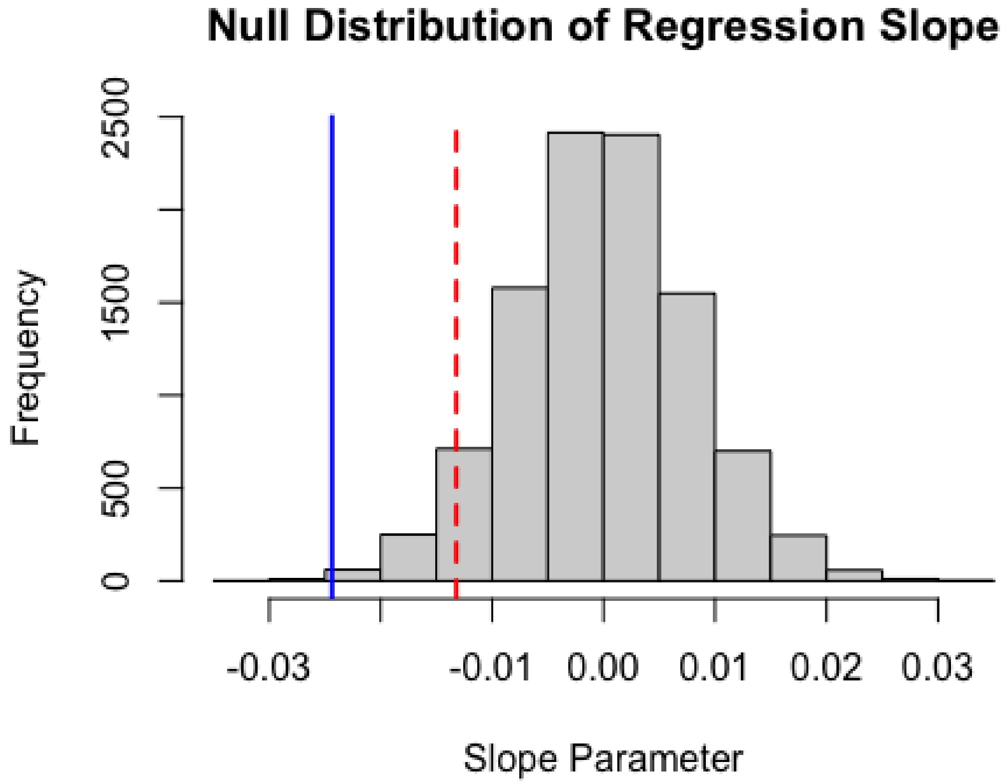
```
mean(t)

## [1] -5.883202

median(t)

## [1] -5.891703

##########Q3##########
quantile(
  pen_boot$t,
  c(0.025, 0.975))

##      2.5%     97.5% 
## -7.843715 -3.895663

#95% bootstrap CI = -7.897064, -3.855186
##########Q4##########
#I think the resampled differences in means do not follow a skewed distribution. The mean is similar to the median and the peak of the histogram centers around the mean/median.
##########Q5##########
pen_ecdf = ecdf(t)
```

```
##########Q6##########
1 - pen_ecdf(-4.5)

## [1] 0.0864

#0.088
##########Q7##########
pen_ecdf(-8)

## [1] 0.0171

#0.02
##########Q8##########
#null: There is no difference in mean flipper lengths between Adelie and
Chinstrap Pengunis.
#alternative: There is difference in mean flipper lengths between Adelie and
Chinstrap Pengunis.
head(veg)

##   block plot date treatment birch pine fern
## 1     A   A3 1995   control     0    4  260
## 2     A   A7 1995   control     0    0  186
## 3     A   A4 1995   removed     8    8   46
## 4     A   A6 1995   removed     6   28    1
## 5     A   A5 1995     mixed     0    1  309
## 6     A   A8 1995     mixed     0    0  258

boxplot(pine ~ treatment, dat = veg)

dat_tree = droplevels(subset(veg, treatment %in% c("control", "clipped")))
boxplot(pine ~ treatment, dat = dat_tree)

table(dat_tree$treatment)

## 
## clipped control 
##       8       8

##########Q9##########
wilcox.test(pine ~ treatment, data = dat_tree, alternative = "two.sided")

## Warning in wilcox.test.default(x = c(11L, 0L, 16L, 3L, 49L, 17L, 0L, 47L:
cannot
## compute exact p-value with ties

## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  pine by treatment
## W = 48, p-value = 0.1005
## alternative hypothesis: true location shift is not equal to 0
```

```
#p-value = 0.1005
#Bootstrap
dat_clipped = subset(dat_tree, treatment == "clipped")
dat_control = subset(dat_tree, treatment == "control")
tree_boot = two.boot(dat_clipped$pine, dat_control$pine, boot_mean, 10000,
student = FALSE, weight = NULL)
##########Q10##########
quantile(
  tree_boot$t,
  c(0.025, 0.975))

##   2.5%  97.5%
##  4.125 29.750

#4.25000 29.50312
##########Q11##########
observed_difference = mean(dat_clipped$pine) - mean(dat_control$pine)
#The observed difference in mean tree counts is 16 and it falls within the
95% bootstrap CI.

dat_all = merge(
  dat_bird,
  dat_habitat,
  by = c("basin", "sub"))

head(dat_all[, c("b.sidi", "s.sidi")])

##       b.sidi s.sidi
## 1 0.06678912   0.12
## 2 0.06509689   0.34
## 3 0.06092608   0.78
## 4 0.06012721   0.57
## 5 0.04112905   0.84
## 6 0.06086158   0.73

#Simpson’s diversity index for breeding birds: b.sidi
#Simpson’s diversity index for vegetation cover types: s.sidi

# Calculate the sample mean and sd:
b_sidi_mean = mean(dat_all$b.sidi, na.rm = TRUE)
b_sidi_sd   = sd(dat_all$b.sidi, na.rm = TRUE)
# Use the subset-by-name symbol ($) to create a
# new column of z-standardized values.
dat_all$b.sidi.standardized = (dat_all$b.sidi - b_sidi_mean)/b_sidi_sd
mean(dat_all$b.sidi.standardized)

## [1] 7.166938e-17

sd(dat_all$b.sidi.standardized)

## [1] 1
```

```
##########Q12##########
#Simpson diversity index measures diversity. It quantifies the number of
species and the relative abundance of each species.
##########Q13##########
s_sidi_mean = mean(dat_all$s.sidi, na.rm = TRUE)
s_sidi_sd   = sd(dat_all$s.sidi, na.rm = TRUE)
dat_all$s.sidi.standardized = (dat_all$s.sidi - s_sidi_mean)/s_sidi_sd
mean(dat_all$s.sidi.standardized)

## [1] 2.984718e-17

sd(dat_all$s.sidi.standardized)

## [1] 1

fit_1 = lm(b.sidi ~ s.sidi, data = dat_all)
coef(fit_1)

## (Intercept)      s.sidi
##  0.07116980 -0.02437131

slope_observed = coef(fit_1)[2]
plot(
  b.sidi ~ s.sidi, data = dat_all,
  main = "Simpson's diversity indices",
  xlab = "Vegetation cover diversity",
  ylab = "Bird diversity")
abline(fit_1)

dat_1 =
  subset(
    dat_all,
    select = c(b.sidi, s.sidi))

set.seed(123)
index_1 = sample(nrow(dat_1), replace = TRUE)
index_2 = sample(nrow(dat_1), replace = TRUE)

dat_resampled_i =
  data.frame(
    b.sidi = dat_1$b.sidi[index_1],
    s.sidi = dat_1$s.sidi[index_2]
  )

fit_resampled_i = lm(b.sidi ~ s.sidi, data = dat_resampled_i)
slope_resampled_i = coef(fit_resampled_i)[2]

print(slope_resampled_i)

##      s.sidi
## 0.006235381
```

```r
plot(
  b.sidi ~ s.sidi, data = dat_resampled_i,
  main = "Simpson's diversity indices (MC resampled data)",
  xlab = "Vegetation cover diversity",
  ylab = "Bird diversity")
abline(fit_resampled_i)

##########Q14##########
m = 10000
result_mc = numeric(m)
for(i in 1:m)
{
  index_1 = sample(nrow(dat_1), replace = TRUE)
  index_2 = sample(nrow(dat_1), replace = TRUE)

  dat_resampled_i =
    data.frame(
      b.sidi = dat_1$b.sidi[index_1],
      s.sidi = dat_1$s.sidi[index_2]
    )
  fit_resampled_i = lm(b.sidi ~ s.sidi, data = dat_resampled_i)
  result_mc[i] = coef(fit_resampled_i)[2]
}
##########Q15##########
hist(
  result_mc,
  main = "Null Distribution of Regression Slope",
  xlab = "Slope Parameter")
abline(v = slope_observed, lty = 1, col = "blue", lwd = 2)
abline(v = quantile(result_mc, c(.05)), lty = 2, col = "red", lwd = 2)
```

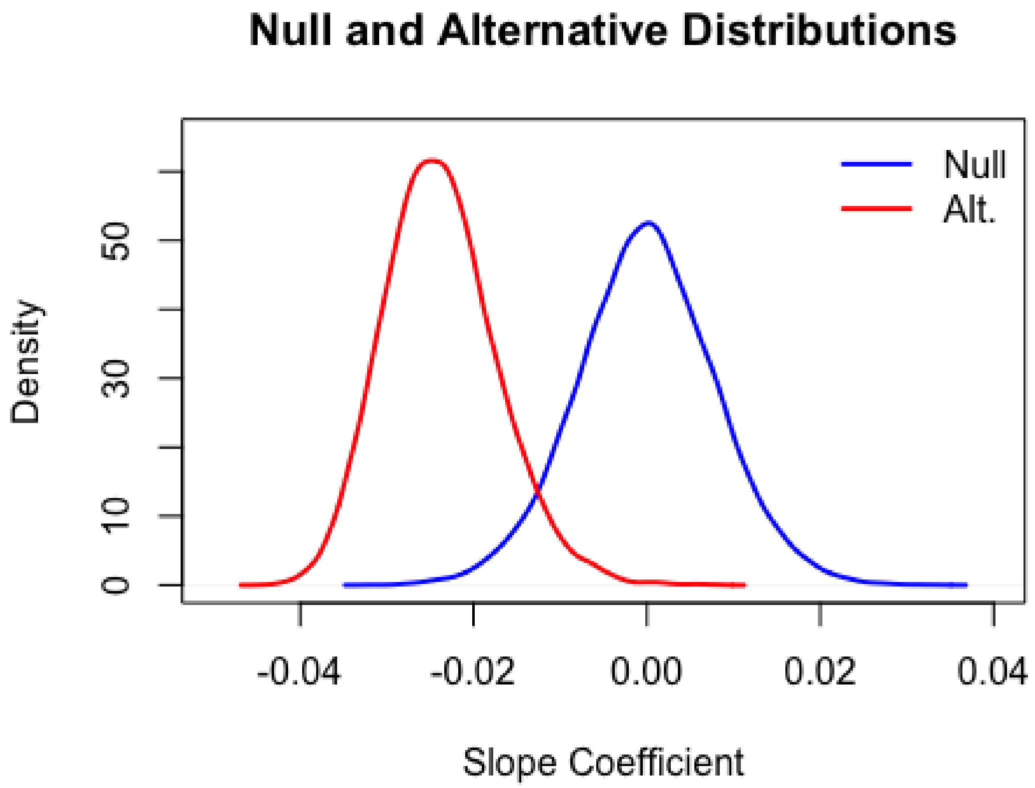


```
quantile(result_mc, c(.05))

##          5%
## -0.01320388

##########Q16##########

#-0.01320388. The observed slope is less than the critical value.

##########Q17##########
#The chance of getting the observed value only by chance is very low (<5%),
which provides evidence to reject the null hypothesis. It is likely that a
negative relationship between vegetation cover diversity and bird diversity
exists.

set.seed(345)
index_1 = sample(nrow(dat_1), replace = TRUE)

dat_boot = dat_1[index_1, ]
head(dat_boot)

##        b.sidi s.sidi
## 29 0.08263485   0.00
## 23 0.05705873   0.62
## 19 0.05820778   0.54
## 21 0.07254766   0.41
## 18 0.06365076   0.49
## 28 0.06284046   0.36
```

```
fit_bs1 = lm(b.sidi ~ s.sidi, data = dat_boot)

coef(fit_bs1)
```

```
## (Intercept)      s.sidi
##  0.07489893 -0.03146039
```

```
##########Q18##########
b = 10000
result_boot = numeric(b)
for(i in 1:b)
{
  index = sample(nrow(dat_1), replace = TRUE)

  dat_boot_i =
    data.frame(
      b.sidi = dat_1$b.sidi[index],
      s.sidi = dat_1$s.sidi[index]
    )
  fit_bs1 = lm(b.sidi ~ s.sidi, data = dat_boot_i)
  result_boot[i] = coef(fit_bs1)[2]
}
hist(
  result_boot,
  main = "Alternative Distribution of Regression Slope",
  xlab = "Slope Parameter")
abline(v = slope_observed, lty = 2, col = "red", lwd = 2)
abline(v = 0, lty = 2, col = 1, lwd = 2)

##########Q19##########
plot(
  density(result_mc),
  main = "Null and Alternative Distributions",
  xlab = "Slope Coefficient",
  xlim = c(-0.05, 0.04),
  ylim = c(0, 65),
  col = "blue",
  lwd = 2
  )
lines(density(result_boot), col = "red", lwd = 2)
legend(c("Null", "Alt."), lty = 1, lwd = 2, col = c("blue", "red"), x =
"topright", bty = "n")
```



```
##########Q20##########
#The region that falls under both curves is a region of uncertainty. If we
observe a slope there, we are not sure if we can reject the null hypothesis.
```