t and permutation homework

Wenjia Xie

April 17, 2019

Snodgrass Problem

In 1861, 10 essays appeared in the New Orleans Daily Crescent. They were signed ”Quintus Curtuis Snodgrass” and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words found in an author’s work.

From eight Twain essays we have:

.225 .262 .217 .240 .230 .229 .235 .217

From 10 Snodgrass essays we have:

.209 .205 .196 .210 .202 .207 .224 .223 .220 .201

Use a permutation test the equality of the means. What is your conclusion?

```
# 1. Test the equality of the means with a parametric test
x <- c(.225,.262,.217,.240,.230,.229,.235,.217)
y <- c(.209,.205,.196,.210, .202,.207,.224,.223,.220,.201)
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 3.7036, df = 11.671, p-value = 0.003156
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.009088475 0.035261525
## sample estimates:
## mean of x mean of y
##  0.231875  0.209700
```

```
# 2. Test the equality of the means with a permutation test

## calculation of means
x_mean <- mean(x)
y_mean<- mean(y)
diff <- abs(x_mean-y_mean)

## Assemble sample for permutation distribution
sample <- c(x, y)
```

```r
## simulation for permutation distribution
N <- 1000
T <- NULL
set.seed(10)
for(i in 1:N){
  index = sample(x = 1:length(sample), size = length(sample),replace = TRUE)
  s = sample[index]
  t = abs(mean(s[1:length(x)]) - mean(s[(length(x)+1):length(sample)]))
  T = c(T,t)
}

## calculate number of differences in the permutation distribution where
## x mean >= y difference
D <- T[(T >= diff )]

num_D <- length(D)

## calculate p-value
p_value <- num_D/N

print(p_value)

## [1] 0.002
```

From both t-test and permutation test, we found the the p-values are so small that we reject the null hupothesis. This means that these essays might not written by Mark Twain.

Hot dog problem

Do the problem using the t distribution. Then calculate a 90% confidence interval using the permutation distribution.

Consider the the numbers of calories in 20 different hot dog brands: 186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132. [Source: Consumer Reports, June 1986] Assume that these numbers are the observed values from a random sample of twenty independent normal random variables

```r
hotdog <- c(186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141,
153, 190, 157, 131, 149, 135, 132)
low <- round(mean(hotdog)-qt(0.95,19)*sd(hotdog)/sqrt(20),2)
high <- round(mean(hotdog)+qt(0.95,19)*sd(hotdog)/sqrt(20),2)
print(paste("The 90% confidence interveral is [",low,",",high,"]" ))

## [1] "The 90% confidence interveral is [ 148.1 , 165.6 ]"
```

Reading Score Problem

Do the t-test specified in the problem. Then test the hypothesis with a permutation test.

An exam is given to two classes of students with the following results: Class 1: 1.23, 1.42, 1.41, 1.62, 1.55, 1.51, 1.60, and 1.76. Class 2: 1.76, 1.41, 1.87, 1.49, 1.67, and 1.81 Assuming that all the observations have a normal distribution with a common unknown variance, test the following hypotheses: H0 : μ1 ≥ μ2 H1 : μ1 < μ2

```
# t-test
class1 <- c(1.23, 1.42, 1.41, 1.62, 1.55, 1.51, 1.60, 1.76)
class2 <- c(1.76, 1.41, 1.87, 1.49, 1.67, 1.81)
t.test(x = class1,y = class2,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  class1 and class2
## t = -1.6596, df = 10.048, p-value = 0.9361
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3259393        Inf
## sample estimates:
## mean of x mean of y
##  1.512500  1.668333
```

From the test, we reject the null hypothesis that the mean score of first class is higher than that of the second class.

```
# Permutation test
set.seed(2019)
n1 <- length(class1)
n2 <- length(class2)
n <- n1+n2
x <- c(class1, class2)
index <- c(rep(2,n2),rep(1,n1))
mean.diff <- function(x, index){
mean(x[index==2]) - mean(x[index==1])
}
base <-  mean.diff(x,index)
results <- replicate(999, mean.diff(sample(x), index))
p.value <- length(results[results>base])/1000
print(p.value)
```

```
## [1] 0.986
```