Northeastern University

R Practice 5 ALY6015

R practice week 5 - Module 5 Assignment — Non-parametric Methods and Sampling

by Mohammad Hossein Movahedi

movahed.m@northeastern.edu

18 May 2022

Table of contents

Introduction

This R practice mainly focuses on Non-parametric Methods and Sampling. During this assignment, I will answer multiple questions using different methods such as the sign test, Wilcox Nonparametric Test, Kruskal-Wallis test, and Spearman rank correlation coefficients.

Analytics

Section 13-2

Question 6: Game Attendance

An athletic director suggests the median number for the paid attendance at 20 local football games is 3000. The data for a random sample are shown. At α = 0.05, is there enough evidence to reject the claim? If you were printing the programs for the games, would you use this figure as a guide?

6210	3150	2700	3012	4875
3540	6127	2581	2642	2573
2792	2800	2500	3700	6030
5437	2758	3490	2851	2720

the first step is to state the hypotheses and identify the claim.

$$H_0 : \text{the median is } 3000$$
$$H_1 : \text{the median is not } 3000$$
$$\begin{cases} H_0 : median = 3000 \\ H_1 : \text{otherwise} \end{cases}$$

then I ran the code

```
#significant value
alpha <- 0.05

#sign test
#claim
median <- 3000
#data
game <- c(16210,  3150, 2700, 3012, 4875,
          3540, 6127, 2581, 2642, 2573,
          2792, 2800, 2500, 3700, 6030,
          5437, 2758, 3490, 2851, 2720)
diff <- game - median
```

```
#determin positives
pos <- length(diff[diff>0])

#determin negatives
neg <- length(diff[diff<0])

#run the test
results <- binom.test(x =  c(pos,neg),alternative = "two.sided")
results
#view p-value
p<-results$p.value

#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the code results are shown below

```
> results

  Exact binomial test

data:  c(pos, neg)
number of successes = 10, number of trials = 20, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2719578 0.7280422
sample estimates:
probability of success
                   0.5

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

The resulting p-value is the strongest; therefore, with great confidence, we fail to reject the null hypothesis that the median is 3000.

Question 10: Lottery Ticket Sales

A lottery outlet owner hypothesizes that she sells 200 lottery tickets a day. She randomly sampled 40 days and found that on 15 days, she sold fewer than 200 tickets. At α = 0.05, is there sufficient evidence to conclude that the median is below 200 tokens?

The first step is stating the hypothesis.

$$H_0 : \text{the median is 200 or more}$$
$$H_1 : \text{the median is less than 200}$$
$$\begin{cases} H_0 : median => 200 \\ H_1 : \text{otherwise} \end{cases}$$

Then I ran the code

```
#question 10
#claim
median <- 200
#data
#determin positives
pos <- 40-15

#determin negatives
neg <- 15

#run the test
```

```
results <- binom.test(x =  c(pos,neg),alternative = "less")
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the result is shown below

```
> results

  Exact binomial test

data:  c(pos, neg)
number of successes = 25, number of trials = 40, p-value = 0.9597
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7527053
sample estimates:
probability of success
                 0.625

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

As can be seen in the result, the p-value of the test is 0.95, which is enormously significant, suggesting that we can't reject the null hypothesis that the median is more than 200.

Section 13-3

Question 4: Lengths of Prison Sentences

A random sample of men and women in prison was asked to give the length of sentence each received for a certain type of crime. At α = 0.05, test the claim that there is no difference in the sentence received by each gender. The data (in months) are shown here.

Males	8 12 6 14 22 27 3 2 2 2 4 6 19 15 13
Females	7 5 2 3 21 26 3 9 4 17 23 12 11 16

The first step is stating the hypothesis.

$$H_0 : \text{there is no difference in the sentence received by each gender}$$
$$H_1 : \text{there is a difference in the sentence received by each gender}$$
$$\begin{cases} H_0 : \text{length of sentence is independent from gender} \\ H_1 : \text{otherwise} \end{cases}$$

then I ran the code

```
#section 13-3
#question 4
males <- c(8, 12, 6 ,14 ,22,27, 3,2 ,2, 2,4,6 , 19, 15, 13)
females <-c(  7,  5 ,2, 3,  21, 26, 3,  9,  4 , 17, 23, 12, 11, 16)
results <- wilcox.test(x = males,y = females ,
                       alternative = 'two.sided',correct = F,exact = F)
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the result is shown below

```
> results

  Wilcoxon rank sum test

data:  males and females
W = 95.5, p-value = 0.6778
alternative hypothesis: true location shift is not equal to 0

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

As can be seen in the results, the p-value is `0.67`, which is greater than alpha; therefore, we fail to reject the null hypothesis.

Question 8: Winning Baseball Games

In the years 1970–to 1993, the National League (NL) and the American League (AL) (major league baseball) were each divided into two divisions: East and West. Below are random samples of the number of games won by each league's Eastern Division. At α = 0.05, is there sufficient evidence to conclude a difference in the number of wins?

NL	89 9 8 101 90 91 9 96 108 100 9 6 8 2 5
AL	08 8 9 97 100 102 9 104 95 89 8 101 6 1 5 8

First, I state the hypothesis.

$$H_0 : \text{there is no difference in the wins}$$
$$H_1 : \text{there is a difference in wins}$$
$$\begin{cases} H_0 : \text{the wins are random} \\ H_1 : \text{otherwise} \end{cases}$$

then I ran the code

```
#question 8
nl <- c(  89, 9,  8 ,101, 90, 91, 9 ,96,  108,  100,  9 ,6, 8 , 2 , 5)
al <- c(108,  8 ,9, 97  ,100, 102,  9 ,104, 95, 89, 8,  101 , 6,  1 ,5,8)
results <- wilcox.test(x = nl,y = al ,
                       alternative = 'two.sided',correct = F,exact = F)
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the results are shown below

```
> results

  Wilcoxon rank sum test

data:  nl and al
W = 111, p-value = 0.7208
alternative hypothesis: true location shift is not equal to 0

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

As can be seen, the p-value is `0.72`, which is greater than alpha; therefore, we fail to reject the null hypothesis.

Section 13-4

Use Table K to determine whether the null hypothesis should be rejected.

1. w = 13, n = 15, α = 0.01, two-tailed

the critical value for these parameters is 16, which is more than w; therefore, the null hypothesis should be rejected

1. w = 32, n = 28, α = 0.025, one-tailed

the critical value for these parameters is 117, which is more than w; therefore, the null hypothesis should be rejected

1. w = 65, n = 20, α = 0.05, one-tailed

the critical value for these parameters is 60, which is less than w; therefore, the null hypothesis can't be rejected

1. w = 22, n = 14, α = 0.10, two-tailed

the critical value for these parameters is 26, which is more than w; therefore, the null hypothesis should be rejected

Section 13-5

Question 2: Mathematics Literacy Scores

Through the Organization for Economic Cooperation and Development (OECD), 15-year-olds are tested in member countries in mathematics, reading, and science literacy. Listed are randomly selected total mathematics literacy scores (i.e., both genders) for selected countries in different parts of the world. Using the Kruskal-Wallis test to see if there is a difference in means at α = 0.05.

Western Hemisphere	Europe	Eastern Asia
527	520	523
406	510	547
474	513	547
381	548	391
411	496	549

First of all, I state the hypothesis.

$$H_0 : \text{here is a difference in means of scores}$$
$$H_1 : \text{there is no difference in means of scores}$$
$$\begin{cases} H_0 : \overline{x}_{WesternHemisphere} = \overline{x}_{Europe} = \overline{x}_{EasternAsia} \\ H_1 : \text{otherwise} \end{cases}$$

then I ran the code

```
#13.5
wh <-data.frame(ndata = c(527, 406, 474, 381 ,411),countries = 'wh')
eu <-data.frame(ndata = c(520 ,510, 513 ,548, 496),countries = 'eu')
ea <-data.frame(ndata = c(523, 547 ,547, 391, 549),countries = 'ea')
data <- rbind(wh,eu,ea)
results <- kruskal.test(ndata ~ countries,data = data )
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the results are shown below

```
> results

  Kruskal-Wallis rank sum test
```

```
data:  ndata by countries
Kruskal-Wallis chi-squared = 4.1674, df = 2, p-value = 0.1245

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

As can be seen, the p-value is `0.12`, which is greater than alpha; therefore, we fail to reject the null hypothesis.

Section 13-6

Question 6: Subway and Commuter Rail Passengers

Six cities are randomly selected, and the number of daily passenger trips (in thousands) for subways and commuter rail services is obtained. At α = 0.05, is there a relationship between the variables? Suggest one reason why the transportation authority might use the results of this study.

City	1	2	3	4	5	6
Subway	845	494	425	313	108	41
Rail	39	291	142	103	33	38

First of all, I state the hypothesis.

$$
\begin{aligned}
&H_0 : \text{here is no relationship between the number of passengers} \\
&H_1 : \text{here is a relationship between the number of passengers} \\
&\quad \begin{cases} H_0 : \text{there is no relationship between variables} \\ H_1 : \text{otherwise} \end{cases}
\end{aligned}
$$

then I ran the code

```
#13.6
citys <- c(1, 2 ,3, 4,  5 ,6)
Subway <- c(845,  494,  425,  313,  108 ,41)
Rail <- c(39, 291,  142 ,103, 33, 38)

data<- data.frame(city = citys,Subway = Subway , Rail = Rail )
results <- cor.test(data$Subway,data$Rail,method = 'spearman')
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
```

the results are shown below

```
> results

  Spearman's rank correlation rho

data:  data$Subway and data$Rail
S = 14, p-value = 0.2417
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.6

> #view p-value
> p<-results$p.value
> #determine reject of accept
> ifelse(p >alpha , 'fail to reject','reject null' )
[1] "fail to reject"
```

As can be seen, the p-value is `0.24`, which is greater than alpha; therefore, we fail to reject the null hypothesis.

Section 14-3

Question 16: Prizes in Caramel Corn Boxes

A caramel corn company gives four different prizes, one in each box. They are placed in the boxes at random. Find the average number of boxes a person needs to buy to get all four prizes. (40)

This is a random test run and the loop breaks when four different results are seen.

then I ran the code

```
#section 14-3
set.seed(96)
y <-c()
x <- c()
z <- c()
z<-replicate(40,{
  repeat{
    y <- sample(1:4,1)

    if (any(x==4)&any(x==3)&any(x==2)&any(x==1)){
      z <- append(z,length(x),1)
      print(x)
      x <- c()
      break
    }
    x <- append(x,y,1)

  }

  ;z})

mean(z)
```

the result is shown below

```
> mean(z)
[1] 8.125
```

As can be seen, on average after buying 8 products all 4 products will be purchased.

Question 18: Lottery Winner

To win a certain lotto, a person must spell the word big. Sixty percent of the tickets contain the letter b, 30% contain the letter i, and 10% contain the letter g. Find the average number of tickets a person must buy to win the prize. (30)

This is a random test run with probability and the loop breaks when three different results are seen.

then I ran the code

```
#question 18
set.seed(96)
y <-c()
x <- c()
z <- c()
z<-replicate(30,{
  repeat{
    y <- sample(c('b','i','g'),1, prob = c(0.6,0.3,0.1))

    if (any(x=='b')&any(x=='i')&any(x=='g')){
      z <- append(z,length(x),1)
      print(x)
      x <- c()
      break
    }
    x <- append(x,y,1)

  }

  ;z})
```

```
mean(z)
```

the the mean of z which is the number of trials turned out to be 10.2 which means after buying ten ticket there is a good chance of wining this lottary.

Bibliography

Statistics Globe. (2022). *sample Function in R (6 Examples) | How to Apply size, replace & prob*. [online] Available at: https://statisticsglobe.com/sample-function-in-r/ [Accessed 19 May 2022].

Kurt, W. (2015). *Count Bayesie*. [online] Count Bayesie. Available at: https://www.countbayesie.com/blog/2015/3/3/6-amazing-trick-with-monte-carlo-simulations [Accessed 19 May 2022].

Appendix

```
#significant value
alpha <- 0.05

#sign test
#section 13-2
#question 6
#claim
median <- 3000
#data
game <- c(16210,  3150, 2700, 3012, 4875,
          3540, 6127, 2581, 2642, 2573,
          2792, 2800, 2500, 3700, 6030,
          5437, 2758, 3490, 2851, 2720)
diff <- game - median

#determin positives
pos <- length(diff[diff>0])

#determin negatives
neg <- length(diff[diff<0])

#run the test
results <- binom.test(x =  c(pos,neg),alternative = "two.sided")
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )

#question 10
#claim
median <- 200
#data
#determin positives
pos <- 40-15

#determin negatives
neg <- 15

#run the test
results <- binom.test(x =  c(pos,neg),alternative = "less")
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )

#section 13-3
#question 4
males <- c(8, 12, 6 ,14 ,22,27, 3,2 ,2, 2,4,6 , 19, 15, 13)
females <-c(  7,  5 ,2, 3,  21, 26, 3,  9,  4 , 17, 23, 12, 11, 16)
results <- wilcox.test(x = males,y = females ,
```

```
                      alternative = 'two.sided',correct = F,exact = F)
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
#question 8
nl <- c(  89, 9,  8 ,101, 90, 91, 9 ,96,  108,  100,  9 ,6, 8 , 2 , 5)
al <- c(108,  8 ,9, 97  ,100, 102,  9 ,104, 95, 89, 8,  101 , 6,  1 ,5,8)
results <- wilcox.test(x = nl,y = al ,
                      alternative = 'two.sided',correct = F,exact = F)
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )

#13.5
wh <-data.frame(ndata = c(527, 406, 474, 381 ,411),countries = 'wh')
eu <-data.frame(ndata = c(520 ,510, 513 ,548, 496),countries = 'eu')
ea <-data.frame(ndata = c(523, 547 ,547, 391, 549),countries = 'ea')
data <- rbind(wh,eu,ea)
results <- kruskal.test(ndata ~ countries,data = data )
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
#13.6
citys <- c(1, 2 ,3, 4,  5 ,6)
Subway <- c(845,  494,  425,  313,  108 ,41)
Rail <- c(39, 291,  142 ,103, 33, 38)

data<- data.frame(city = citys,Subway = Subway , Rail = Rail )
results <- cor.test(data$Subway,data$Rail,method = 'spearman')
results
#view p-value
p<-results$p.value
#determine reject of accept
ifelse(p >alpha , 'fail to reject','reject null' )
#section 14-3
#question 16
set.seed(96)
y <-c()
x <- c()
z <- c()
z<-replicate(40,{
  repeat{
    y <- sample(1:4,1)

    if (any(x==4)&any(x==3)&any(x==2)&any(x==1)){
      z <- append(z,length(x),1)
      print(x)
      x <- c()
      break
    }
    x <- append(x,y,1)

  }

  ;z})

mean(z)
#question 18
set.seed(96)
y <-c()
x <- c()
z <- c()
z<-replicate(30,{
  repeat{
    y <- sample(c('b','i','g'),1, prob = c(0.6,0.3,0.1))

    if (any(x=='b')&any(x=='i')&any(x=='g')){
      z <- append(z,length(x),1)
      print(x)
      x <- c()
      break
    }
    x <- append(x,y,1)

  }
```

```
  ;z})

mean(z)
```