Northeastern University

Toronto

Nonparametric Methods and Sampling

Soni Manan[a]

[a] *College of Professional Studies, Master of Professional Studies in Analytics.*

***Subject**: ALY6015 **NUID**: 002982645*

Under the guidance of

Dr. Prof. Alex Maizlish

Introduction

This assignment is to resolve problem-statements with parametric and non-parametric models and sampling. The purpose of this assignment is to learn the hypothesis testing methods such as Sign Test, Wilcoxon test, Kruskal Willis test and more. Through testing we need to accept or reject as per the conclusions whether accept or reject hypothesis.

Previously we learned about Chi-squared test and ANOVA like where to use and when. This are methods are based on sampling.

Sign Test or Non-Parametric Test

Statement 1: From the data, it can be said that median attendance of 20 football games is 3000?

Null Hypothesis: Yes median is 3000.

Alternative: No its not.

Result: We tried sign test for this problem and as outcome, p value is 1 which is higher than the significance value (0.05). Because of that we failed to reject the Null Hypothesis.

We don't have enough evidence to reject. Another fact is the probability is 0.5 which is 50-50% for success and failure.

```
        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
```

Statement 2: Can it be conclude that median sales of lottery ticket is 200?

Null Hypothesis: Yes median is 200.

Alternative: No median sales of tickets is not 200.

Result: As result, P-value is 0.1539 which is greater than level of significance which 0.05. Therefore, we cant reject Null hypothesis. In addition, success rate is 37.5% that median could be sales could be 200.

```
        Exact binomial test

data:  c(pos, neg)
number of successes = 15, number of trials = 40, p-value = 0.1539
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2272627 0.5419852
sample estimates:
probability of success
                 0.375
```

Wilcoxon Rank Sum Test

Statement 1: There is a claim that statement received by each gender is same.

Null Hypothesis: There is no difference.

Alternative: Some difference exist in the sentence received.

Result: As per the instruction, used Wilcoxon rank sum test using the males and females entries. And at end of the test we got p-value as 0.9 that is greater than the alpha 0.05. Because of that we fail to reject the null hypothesis. In nutshell, we have some statistical evidence to prove that men and women have served equal sentences.

```
        Wilcoxon rank sum test

data:  males and female
W = 113, p-value = 0.1357
alternative hypothesis: true location shift is not equal to 0
```

Statement 2: From the data, can it be concluded that number of wins of National League and the American league is different in 1970 to 1933.

Null Hypothesis: Yes number of wins is different.

Alternative: It is not true.

Result: As per a result, used Wilcoxon rank sum test using the NL and AL data. And at end of the test we got p-value which is 0.6657 that is greater than the alpha 0.05. Because of that we fail to reject the null hypothesis.

```
        Wilcoxon rank sum test

data:  Nl and Al
W = 59, p-value = 0.6657
alternative hypothesis: true location shift is not equal to 0
```

K Table Test

Statement 1: To check whether the null hypothesis should be rejected based on data.

Null Hypothesis: Yes.

Alternative: No.

Result: Value of combination is 25 and Ws is 22, so 22 is less than 25, so we reject null hypothesis.

Kruskal Willis Test

Statement 1: To check whether there is difference in the means of literacy of Mathematics in different region.

Null Hypothesis: There is no difference in the means.

Alternative: There is some difference.

Result: We learned this test using this problem statement. And at last p value is 0.12 which is greater than value of alpha 0.05.

So we don't have enough evidence to reject the null hypothesis that mean literacy is different in regions of world.

```
        Kruskal-Wallis rank sum test

data:  score by group
Kruskal-Wallis chi-squared = 4.3478, df = 2, p-value = 0.1137
```

Spearman Rank Correlation Coefficient Test

Statement 1: Check whether there is any relation between subway and rail passenger.

Null Hypothesis: There is relationship between these two variables.

Alternative: There is no relationship.

Result: From this test, we got the p-value is 0.2417. That is good enough greater than alpha which is 0.05 and value of rho is 0.6. If the value of rho is near to 1 means the variables have positive correlation and less than 0 close to -1 will have negative correlation.

```
        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
```

Simulations

Statement 1: Find the average number of boxes a person needs to buy to get all four prizes.

Answer: If someone pickups 1 out of 4, probability will be 25% of choosing a box. So it will be equally distributes. For average 10 boxes should be bought by person for all the prizes.

```
> x<-c("1",
+      "2",
+      "3",
+      "4")
> prizes <- sample(x, 40, replace = TRUE, prob = c(0.25, 0.25, 0.25,0.25))
> mean(table(prizes))
[1] 10
```

Statement 2: Find the average number of tickets a person must buy to win the prize.

Answer: If consider word is 'Big' and for I and g probabilities have given which are 30% and 10% so for big it will be 60%. So from the calculations in R, we found a average person should buy 10 tickets.

```
x<-c("b",
     "i",
     "g")
ticket<-sample(x, 30, replace = TRUE, prob = c(0.6, 0.3, 0.1))
mean(table(ticket))
1] 10
```

Conclusion

Thus, we've got with success understood and created a conclusion on each individual problem-statements given to us. we've got made sensible use of assorted non parameterized hypothesis testing techniques to unravel the things and create a selected decision concerning the behaviour of the problem-sets given.

References

Zach, May 22 2020, How to Perform a Binomial Test in R *Source*: https://www.statology.org/binomial-test-r/

Paired Sample T-Test. (n.d.). Statistics Solutions. *Source:* https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired - sample-t-test/

Appendix

```
alpha <- 0.05
median <- 3000

attendance <- c(6210,3150,2700,3012,4875,3540,6127,2581,2642,2573,2792,2800,2500,3700,6030,5437,2758,3490,2851,2720)

difference <- attendance - median

pos <- length(difference[difference>0])

neg <- length(difference[difference<0])
result <- binom.test(x=c(pos,neg),alternative = "two.sided")
result

result$p.value
ifelse(result$p.value>alpha ,"fail to reject the null","reject the null")

#13.

alpha <- 0.05
n <- 40
pos <- 15
neg <- n - pos

result2 <- binom.test(x=c(pos,neg),alternative = )
result2

result$p.value
ifelse(result$p.value>alpha ,"fail to reject the null","reject the null")


#13.3 (4)

alpha = 0.05

males <- c(8,12,6,14,22,27,32,24,26,19,15,13)
female <- c(7,5,2,3,21,26,30,9,4,17,23,12,11,16)

result3 <- wilcox.test(x = males,y= female,alternative = "two.sided",correct = FALSE)
result3
result3$p.value
ifelse(result3$p.value>alpha ,"fail to reject the null","reject the null")

#13.3(8)

alpha <- 0.05
Nl <- c(89,96,88,101,90,91,92,96,108,100,95)
Al <- c(108,86,91,97,100,102,95,104,95,89,88,101)
```

```
result4 <- wilcox.test(x = Nl,y= Al,alternative = "two.sided",correct = FALSE)
result4
result4$p.value
ifelse(result4$p.value>alpha ,"fail to reject the null","reject the null")

#13.5

rm(list=ls())
# performing kruskal wallis test on diffrent regions for the maths scores>
western<-data.frame(score = c(527,406,474,381,411),group=rep("Western",5))
europe<-data.frame(score = c(520,510,513,548,596),group=rep("Europe",5))
E.asia<-data.frame(score = c(523,547,547,391,549),group=rep("Eastern Asia",5))

data <- rbind(western,europe,E.asia)

result6 <- kruskal.test(score~group,data=data)
result6

#13.6

alpha <- 0.05
city <- 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=city,subway = subway,rail = rail)

result7 <- cor.test(data$subway,data$rail,method="spearman")
result7

result7$p.value
result7$estimate

#14

x<-c("1",
    "2",
    "3",
    "4")
prizes <- sample(x, 40, replace = TRUE, prob = c(0.25, 0.25, 0.25,0.25))
mean(table(prizes))

x<-c("b",
    "i",
    "g")
ticket<-sample(x, 30, replace = TRUE, prob = c(0.6, 0.3, 0.1))
ticket
mean(table(ticket))
```