Module 5 : Nonparametric Methods and Sampling

Mohit Ravindra Kamble

College of Professional Studies – Northeastern University

ALY6015: Intermediate Analytics

Prof. Roy Wada

February 16, 2024

01. Overview:

In this assignment, I applied nonparametric statistical methods and sampling techniques to address various problems. I focused on hypothesis testing using methods such as the sign test, Wilcoxon rank sum test, signed-rank test, Kruskal-Wallis test, and the runs test. Additionally, I computed the Spearman rank correlation coefficient and demonstrated knowledge of basic sampling methods. The problems covered diverse scenarios, including game attendance, lottery ticket sales, lengths of prison sentences, and baseball game wins. I utilized R programming and MS Excel to perform the analyses, presenting critical values, test values, and decisions. The assignment aimed to enhance my understanding of nonparametric methods, hypothesis testing, and simulation techniques.

02. Analysis:

2.1) Section 13-2

6 - Game Attendance

- ➔ I conducted a binomial test on the attendance data for 20 local football games, where the athletic director claimed a median of 3000.
- ➔ With a p-value of 1 at a 0.05 significance level, I *failed to reject the null hypothesis.*
- ➔ This implies there's no significant difference, and I would use the claimed median of 3000 as a guide for printing programs for the games.

O/P:

```
> result
      Exact binomial test
data:  c(p, n)
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
> result$p.value
[1] 1
> ifelse(result$p.value>alpha,"Failed to reject the null hypothesis","Rejected
the null hypothesis")
[1] "Failed to reject the null hypothesis"
```

10 - Lottery Ticket Sales

- ➔ In this analysis, I investigated the owner's hypothesis that the lottery outlet sells 200 tickets daily.
- ➔ After randomly sampling 40 days and observing sales below 200 tickets on 15 days, I conducted a Wilcoxon rank sum test with a significance level of 0.05.
- ➔ The resulting p-value of 0.00002312556 *led me to reject the null hypothesis*, providing sufficient evidence to conclude that the median daily sales are indeed below 200 tickets.

O/P:

```
> result

      Wilcoxon rank sum test with continuity correction

data:  rep(200, sample) and c(rep(200, sample - db_200), seq(1, db_200))
W = 1100, p-value = 0.00002313
alternative hypothesis: true location shift is not equal to 0

> result$p.value
[1] 0.00002312556
> ifelse(result$p.value>alpha,"Failed to reject the null hypothesis","Rejected
the null hypothesis")
[1] "Rejected the null hypothesis"
```

2.2) Section 13-3

4 - Lengths of Prison Sentences

- ➔ Upon conducting a Wilcoxon rank sum test at a significance level of 0.05, I examined the claim that there is no difference in the sentences received by men and women in prison for a specific crime.
- ➔ The p-value obtained was 0.1425, surpassing the significance threshold.
- ➔ Therefore, I *fail to reject the null hypothesis*, suggesting that there is no compelling evidence of a difference in sentence lengths between genders based on the provided data.

O/P:

```
> result
      Wilcoxon rank sum test with continuity correction
data:  m and f
W = 113, p-value = 0.1425
alternative hypothesis: true location shift is not equal to 0
> result$p.value
[1] 0.1425439
> if (result$p.value < alpha) {
+  cat("\nRejected the null hypothesis")
+ } else {
+  cat("\nFailed to reject the null hypothesis")
+ }
Failed to reject the null hypothesis
```

8 - Winning Baseball Games

- ➔ In analyzing the baseball game wins for the National League (NL) and the American League (AL) Eastern Divisions from 1970 to 1993, I conducted a Wilcoxon rank sum test with a significance level of 0.05.
- ➔ The test aimed to determine if there was sufficient evidence to conclude a difference in the number of wins between the two leagues.The results indicated a p-value of 0.6883, exceeding the chosen alpha level.
- ➔ Therefore, I ***failed to reject the null hypothesis***, suggesting that there is no significant evidence of a difference in the number of wins between the NL and AL Eastern Divisions during the specified period.

O/P:

```
> result
      Wilcoxon rank sum test with continuity correction
data:  nl and al
W = 59, p-value = 0.6883
alternative hypothesis: true location shift is not equal to 0
> result$p.value
[1] 0.6883179
> if (result$p.value > alpha) {
+  cat("\nRejected the null hypothesis")
+ } else {
+  cat("\nFailed to reject the null hypothesis")
+ }
Rejected the null hypothesis
```

2.3) Section 13-4

- ➔ For the first scenario (5th sum), where the Wilcoxon signed-rank test was conducted with a sample size of 15 and a significance level of 0.01 for a two-tailed test, the p-value obtained was 0.3173. As this p-value is greater than the significance level, I *failed to reject the null hypothesis*, suggesting no significant difference in medians between the two groups.
- ➔ Similarly, in the second case (6th sum) with a one-tailed test and a significance level of 0.025, the p-value remained 0.3173, leading to a *failure to reject the null hypothesis.*
- ➔ The third case (7th sum), a one-tailed test with a significance level of 0.05, yielded the same p-value, resulting in a *failure to reject the null hypothesis.*
- ➔ Finally, in the fourth case (8th sum), a two-tailed test with a significance level of 0.10, the p-value once again stayed at 0.3173, *leading to a rejection of the null hypothesis*. Despite the variations in sample size and significance levels, the consistent p-value suggests a lack of evidence to support differences in medians between the groups in each scenario.

O/P:

```
> res5
Kruskal-Wallis rank sum test
data:  list(ws = c(13), n = c(15))
Kruskal-Wallis chi-squared = 1, df = 1, p-value = 0.3173
> res5$p.value
[1] 0.3173105
> if (res5$p.value > alpha) {
+ cat("\nRejected the null hypothesis")
+ } else {
+   cat("\nFailed to reject the null hypothesis")
+ }
Rejected the null hypothesis
```

```
> res6
Kruskal-Wallis rank sum test
data:  list(ws = c(32), n = c(28))
Kruskal-Wallis chi-squared = 1, df = 1, p-value = 0.3173
> res6$p.value
[1] 0.3173105
> if (res6$p.value > alpha) {
+ cat("\nRejected the null hypothesis")
+ } else {
+   cat("\nFailed to reject the null hypothesis")
+ }
Rejected the null hypothesis
```

```
> res7
Kruskal-Wallis rank sum test
data:  list(ws = c(65), n = c(20))
Kruskal-Wallis chi-squared = 1, df = 1, p-value = 0.3173
> res7$p.value
[1] 0.3173105
> if (res7$p.value > alpha) {
+ cat("\nRejected the null hypothesis")
+ } else {
+   cat("\nFailed to reject the null hypothesis")
+ }
Rejected the null hypothesis
```

```
> res8
Kruskal-Wallis rank sum test
data:  list(ws = c(22), n = c(14))
Kruskal-Wallis chi-squared = 1, df = 1, p-value = 0.3173
> res8$p.value
[1] 0.3173105
> if (res8$p.value > alpha) {
+ cat("\nRejected the null hypothesis")
+ } else {
+   cat("\nFailed to reject the null hypothesis")
+ }
Rejected the null hypothesis
```

2.4) Section 13-5

2 - Mathematics Literacy Scores

- ➔ After conducting the Kruskal-Wallis test on randomly selected total mathematics literacy scores from different parts of the world, including the Western Hemisphere, Europe, and Eastern Asia, I found that the p-value is 0.1245, exceeding the significance level of 0.05.
- ➔ Therefore, I *fail to reject the null hypothesis*, indicating no significant difference in means among the selected countries.
- ➔ In simpler terms, the test does not provide enough evidence to conclude that the mathematics literacy scores differ significantly between the regions.

O/P:

```
> result
      Kruskal-Wallis rank sum test
data:  scores by group
Kruskal-Wallis chi-squared = 4.1674, df = 2, p-value = 0.1245
> result$p.value
[1] 0.1244662
> if (result$p.value < alpha) {
+  cat("\nRejected the null hypothesis")
+ } else {
+  cat("\nFailed to reject the null hypothesis")
+ }
Failed to reject the null hypothesis
```

2.5) Section 13-6

Subway and Commuter p Passengers

- ➔ In this analysis, I aimed to determine if there is a correlation between the daily passenger trips for subways and commuter rail services in six randomly selected cities.
- ➔ The Spearman rank correlation coefficient was calculated, and the null hypothesis (H0) stated that there is no correlation, while the alternative hypothesis (H1) posited the existence of a correlation.
- ➔ With a significance level set at 0.05, the decision was made based on the p-value from the Spearman rank correlation test.
- ➔ The obtained p-value was greater than 0.05, leading to the conclusion of *failing to reject the null hypothesis.* Consequently, there is no evidence of a significant correlation between subway and commuter rail usage in the selected cities.

O/P:

```
> result
      Spearman's rank correlation rho
data:  data$s and data$r
S = 14, p-value = 0.2417
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.6
> if (result$p.value < alpha) {
+  cat("\nRejected the null hypothesis")
+ } else {
+  cat("\nFailed to reject the null hypothesis")
+ }
Failed to reject the null hypothesis
```

2.6) Section 14-3

16 - Prizes in Caramel Corn Boxes

- ➔ In this simulation, I designed a function to replicate the experience of buying boxes from a caramel corn company until obtaining all four different prizes: Prize A, Prize B, Prize C, and Prize D.
- ➔ The function randomly selects prizes and keeps track of the boxes purchased until all four unique prizes are obtained.
- ➔ By repeating this experiment 40 times, I calculated the average number of boxes a person needs to buy to collect all four prizes.
- ➔ The result revealed that, on average, *a person would need to purchase approximately 9 boxes to obtain the complete set of prizes.*

O/P:

```
> results
 [1]  5 14  5 19  5  9  6  4  6  5  4 11 12  6  8 15  6 17  6 12  5 16  7  4
4  7  5 12  5  5 14  6 16 11  8
[36] 16 10  7  5  4
> avg_b
[1] 8.55
```

18 - Lottery Winner

- ➔ In simulating the lottery winner scenario, I created a function to replicate the experience of buying tickets until the word "big" is spelled.
- ➔ The lotto tickets contain the letters 'b,' 'i,' and 'g' with probabilities of 60%, 30%, and 10%, respectively. After repeating the experiment 30 times, I calculated *the average number of tickets a person would need to purchase to win the prize, resulting in an average of approximately 48.3 tickets.*

O/P:

```
> results
 [1]  63  19   9  10 104  80 137  50   5  38  24  21  26  16 116  15  28 172
3  21   9  70   6  92  16  96
[27]  24  84  39  62
> avg_t
[1] 48.5
```

03. Conclusion:

In conclusion, this assignment on nonparametric methods and sampling techniques has provided valuable insights into diverse statistical scenarios. By applying hypothesis testing methods such as the sign test, Wilcoxon rank sum test, and Kruskal-Wallis test, I addressed questions related to game attendance, lottery ticket sales, lengths of prison sentences, and baseball game wins. The analyses involved rigorous statistical procedures, including simulations and Spearman rank correlation coefficient calculations. I presented results with p-values, critical values, and decisions based on significance levels. This assignment significantly enhanced my understanding of nonparametric methods, hypothesis testing, and simulation techniques, contributing to a broader skill set in the field of analytics.

04. Citations:

- Kruskal-Wallis Test: source.
- Wilcoxon Tests: source.

05. Appendix:

```
cat("\014") # clears console
rm(list = ls()) # clears global environment
try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots
try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) # clears packages
options(scipen = 100) # disables scientific notation for entire R session
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>Week - 5<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
# Section 13-2
# 6 - Game Attendance
#Stating the Hypothesis
#H0: median = 3000
#H1: median =! 3000
# median
median = 3000
#Significance level
alpha = 0.05
fig <- c(6210,3150,2700,3012,4875,3540,6127,2581,2642,2573,2792,2800,2500,3700,6030,
         5437,2758,3490,2851,2720)
guide <- fig-median
#positive
p <- length(guide[guide>0])
#negative
n <- length(guide[guide<0])
result<- binom.test(x= c(p,n), alternative = "two.sided")
result$p.value
ifelse(result$p.value>alpha,"Failed to reject the null hypothesis","Rejected the null hypothesis")
# 10 - Lottery Ticket Sales
```

```
#Stating the hypothesis

#H0: Sells 200 lottery t/day

#H1: sells < 200 lottery t/day

#Significance level

alpha=0.05

sample <- 40

db_200 <- 15

result <- wilcox.test(x=sample, y=db_200, alternative = "two.sided", correct = FALSE)

test_value <- sum(1:db_200)

# Performing Wilcoxon rank sum test

result <- wilcox.test(x = rep(200, sample), y = c(rep(200, sample - db_200), seq(1, db_200)), exact = FALSE, correct = TRUE)

result$p.value

ifelse(result$p.value>alpha,"Failed to reject the null hypothesis","Rejected the null hypothesis")
```

 # Section 13-3

 # 4 - Lengths of Prison Sentences

```
#Stating the hypothesis

#H0: No difference in sentence received

#H1: Difference in sentence received

alpha=0.05

m <- c(8,12,6,14,22,27,32,24,26,19,15,13)

f <- c(7,5,2,3,21,26,30,9,4,17,23,12,11,16)

result <- wilcox.test(x=m, y=f)

result$p.value

if (result$p.value < alpha) {

  cat("\nRejected the null hypothesis")

} else {  cat("\nFailed to reject the null hypothesis")}
```

```
# 8 - Winning Baseball Games
#Stating the Hypothesis
#H0: Difference between wins
#H1: No difference between wins
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)
result <- wilcox.test(x= nl, y= al)
result$p.value
if (result$p.value > alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
# Section 13-4
# 5)
#Stating the Hypothesis
#H0: The two groups have the same median.
#H1: The two groups do not have the same median.
alpha = 0.01
res5 <- kruskal.test(list(ws = c(13), n = c(15)))
res5$p.value
if (res5$p.value > alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
```

```
# 6)
#Stating the Hypothesis
#H0: Second group = median as or Median < the first group median.
#H1: Second group median > first group median.
alpha = 0.025
res6 <- kruskal.test(list(ws = c(32), n = c(28)))
res6$p.value
if (res6$p.value > alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
# 7)
#Stating the Hypothesis
#H0: Second group = median as or median < first group median.
#H1: Second group median > first group median.
alpha = 0.05
res7 <- kruskal.test(list(ws = c(65), n = c(20)))
res7$p.value
if (res7$p.value > alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
# 8)
#Stating the Hypothesis
```

```
#H0: Both the groups have = median.
#H1: Both the groups different median.
alpha=0.10
res8 <- kruskal.test(list(ws = c(22), n = c(14)))
res8$p.value
if (res8$p.value > alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
```

 # Section 13-5

 # 2 - Mathematics Literacy Scores

```
#Stating the Hypothesis
#H0: No difference in means
#H1: Difference in means
alpha = 0.05
wh <- data.frame( scores =c(527, 406, 474, 381, 411), group = rep("wh",5))
eu <- data.frame( scores = c(520, 510, 513, 548, 496), group = rep("eu",5))
ea <- data.frame( scores = c(523, 547, 547, 391, 549) , group = rep("ea",5))
data <- rbind(wh,eu,ea)
result <- kruskal.test(scores ~ group, data = data)
result$p.value
if (result$p.value < alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
```

```
# Section 13-6
# Subway and Commuter p Passengers
#Stating the Hypothesis
#H0: No correlation
#H1: There is a correlation
alpha = 0.05
c <- c(1,2,3,4,5,6)
s <- c(845, 494, 425, 313, 108, 41)
r <- c(39, 291, 142, 103, 33, 38)
data <-data.frame(c=c, s=s, r=r)
result <- cor.test(x= data$s, y= data$r, method="spearman")
if (result$p.value < alpha) {
  cat("\nRejected the null hypothesis")
} else {
  cat("\nFailed to reject the null hypothesis")
}
# Section 14-3
# 16 - Prizes in Caramel Corn Boxes
sim_exp <- function() {
  prizes <- c("Prize A", "Prize B", "Prize C", "Prize D")
  boxes <- c()
  num_boxes <- 0
  while(length(unique(boxes)) < length(prizes)) {
    prize <- sample(prizes, 1)
    boxes <- c(boxes, prize)
    num_boxes <- num_boxes + 1
  }
```

```
 return(num_boxes)
}
numtr <- 40
results <- replicate(numtr, sim_exp())
avg_b <- mean(results)
```

 # 18 - Lottery Winner

```
sim_exp <- function() {
 l <- c("b", "i", "g")
 t <- c()
 while(TRUE) {
  letter <- sample(l, 1, prob = c(0.6, 0.3, 0.1), replace = TRUE)
  t <- c(t, letter)
  if (length(t) >= 3 && paste(t[(length(t)-2):length(t)], collapse = "") == "big") {
   break
  }
 }
 return(length(t))
}
numtr <- 30
results <- replicate(numtr, sim_exp())
avg_t <- mean(results)
```