***

FINAL ASSIGNMENT

TITLE: NONPARAMETRIC METHOD

NAME: NAM HO PHAN

COURSE: 6015

PROFESSOR: SHAPIRO VLADIMIR

***

INTRODUCTION:

Statistician shave developed a branch of statistics known as nonparametric statistics or distribution-free statistics to use when the population from which the samples are selected is not normally distributed. Nonparametric statistics can also be used to test hypotheses that do not involve specific population parameters, such as m, s, or p.

The nonparametric tests explained in this chapter are the sign test, the Wilcoxon rank sum test, the Wilcoxon signed-rank test, the Kruskal-Wallis test, and the runs test. In addition, the Spearman rank correlation coefficient, a statistic for determining the relationship between ranks, is explained.

There are five advantages that nonparametric methods have over parametric methods:

1. They can be used to test population parameters when the variable is not normally distributed.

2. They can be used when the data are nominal or ordinal.

3. They can be used to test hypotheses that do not involve population parameters.

4. In some cases, the computations are easier than those for the parametric counterparts.

5. They are easy to understand.

6. It is useful to understand if there are many outliers.

There are three disadvantages of nonparametric methods:

1. They are less sensitive than their parametric counterparts when the assumptions of the parametric methods are met. Therefore, larger differences are needed before the null hypothesis can be rejected.

2. They tend to use less information than the parametric tests. For example, the sign test requires the researcher to determine only whether the data values are above or below the median, not how much above or below the median each value is.

3. They are less efficient than their parametric counterparts when the assumptions of the parametric methods are met. That is, larger sample sizes are needed to overcome the loss of information. For example, the nonparametric sign test is about 60% as efficient as its parametric counterpart, the z test. Thus, a sample size of 100 is needed for use of the sign test, compared with a sample size of 60 for use of the z test to obtain the same results.

ANALYSIS:

***

The Sign Test

The simplest nonparametric test, the sign test for single samples, is used to test the value of a median for a specific sample.

```
#State the hypotheses and identify the claim.
#Find the critical value(s).
#Compute the test value.
#Make the decision.
#Summarize the results.
```

Section 13-2

```
#Game Attendance
cat("H0: median = 3000(claim)\nH1: median != 3000")

## H0: median = 3000(claim)
## H1: median != 3000

n = 20
cat(paste("\ncritical_value:", 5))

##
## critical_value: 5

attendance = c(6210,3150,2700,3012, 4875, 3540, 6127, 2581, 2642, 2573, 2792,
2800, 2500, 3700, 6030, 5437, 2758, 3490, 2851, 2720)
greater_30 = 0
lower_30 = 0

for (i in attendance) {
  if (i >= 3000) {
  greater_30 = greater_30 + 1 }
  else { lower_30 = lower_30 + 1
  }
  if (greater_30 > lower_30){
    test_value = lower_30}
  else if (greater_30 < lower_30) {
    test_value = greater_30}
  else {test_value = lower_30}}
cat(paste("\ntest_value:",test_value))
```

```
##
## test_value: 10
```

```
cat(paste("\nBecause test value is greater than the critical value. We cannot
reject the null hypothesis. Therefore, there is not enough evidence to reject
the claim that the median number of the paid attendance is 3000."))
```

```
##
## Because test value is greater than the critical value. We cannot reject th
e null hypothesis. Therefore, there is not enough evidence to reject the clai
m that the median number of the paid attendance is 3000.
```

```
#Lottery Ticket Sales
cat("H0: median = 200(claim)\nH1: median < 200")
```

```
## H0: median = 200(claim)
## H1: median < 200
```

```
cat(paste("\ncritical_value:", -1.65))
```

```
##
## critical_value: -1.65
```

```
X= 15
n= 40
z = ((X + 0.5) - (n/2))/ sqrt(n)/2
cat(paste("\ntest_value:",z))
```

```
##
## test_value: -0.355756236768943
```

```
cat(paste("\nBecause test value is less than the critical value. We reject th
e null hypothesis. Therefore, there is enough evidence to conclude that the m
edian is below 200."))
```

```
##
## Because test value is less than the critical value. We reject the null hyp
othesis. Therefore, there is enough evidence to conclude that the median is b
elow 200.
```

When the sample size is 26 or more, the normal approximation can be used to find the test value.

```
#Lengths of Prison Sentences
cat("H0: There is no difference in the length of sentence by each gender (cla
im)\nH1: There is difference in the length of sentence by each gender")
```

```
## H0: There is no difference in the length of sentence by each gender (claim
)
## H1: There is difference in the length of sentence by each gender
```

```
alpha= 0.05
cat(paste("\ncritical_value:", -1.96, "< z <", 1.96))
```

```
##
## critical_value: -1.96 < z < 1.96

frame = c(2,3,4,5,6,7,8,9,11,12,12,13,14,15,16,17,19,21,22,23,24,26,26,27,30,
32)
M_F = c("F","F","F","F","M","F","M","F","F","M","F","M","M","M","F","F","M","
F","M","F","M","M","F","M","F","M")
M_F = data.frame(M_F)
M_F['time'] = frame
M_F['rank'] = c(1,2,3,4,5,6,7,8,9,10.5,10.5,12,13,14,15,17,17,18,19,20,21,22.
5,22.5,24,25,26)
M_F

##    M_F time rank
## 1    F    2  1.0
## 2    F    3  2.0
## 3    F    4  3.0
## 4    F    5  4.0
## 5    M    6  5.0
## 6    F    7  6.0
## 7    M    8  7.0
## 8    F    9  8.0
## 9    F   11  9.0
## 10   M   12 10.5
## 11   F   12 10.5
## 12   M   13 12.0
## 13   M   14 13.0
## 14   M   15 14.0
## 15   F   16 15.0
## 16   F   17 17.0
## 17   M   19 17.0
## 18   F   21 18.0
## 19   M   22 19.0
## 20   F   23 20.0
## 21   M   24 21.0
## 22   M   26 22.5
## 23   F   26 22.5
## 24   M   27 24.0
## 25   F   30 25.0
## 26   M   32 26.0

M = M_F[M_F['M_F'] == 'M',]
R = sum(M['rank'])

#Because the sample size of M < L so we sum the ranks of the group M

uR = (12*(12+14+1)) / 2
std = sqrt((12*14*(12+14+1)) / 12)
z = (R-uR) / std
cat(paste("\ntest_value:",z))
```

```
##
## test_value: 1.49159904963356
```

```
cat(paste("\nBecause test value is in the critical value. I don't reject the
null hypothesis. Therefore, there is enough evidence to conclude that there i
s no difference in the length of sentence received by each gender"))
```

```
##
## Because test value is in the critical value. I don't reject the null hypot
hesis. Therefore, there is enough evidence to conclude that there is no diffe
rence in the length of sentence received by each gender
```

The Wilcoxon Rank Sum Test

The Wilcoxon rank sum test is used for independent samples, and the Wilcoxon signed-rank test is used for dependent samples. Both tests are used to compare distributions.

The Wilcoxon rank sum test

```
#Winning Baseball Games
cat("H0: There is no difference in the number of wins by each league\nH1: The
re is difference in the number of wins by each league(claim)")
```

```
## H0: There is no difference in the number of wins by each league
## H1: There is difference in the number of wins by each league(claim)
```

```
alpha= 0.05
cat(paste("\ncritical_value:", -1.96, "< z <", 1.96))
```

```
##
## critical_value: -1.96 < z < 1.96
```

```
frame = c(86,88,88,89,89,90,91,91,92,95,95,95,96,96,97,100,100,101,101,102,10
4,108,108)
game = c("A","A","N","A","N","N","N","A","N","N","A","A","N","N","A","N","A",
"A","N","A","A","N","A")
base = data.frame(frame)
base['game'] = game
base['rank'] = c(1,2.5,2.5,4.5,4.5,6,7.5,7.5,9,10,11.5,11.5,13.5,13.5,15,16.5
,16.5,18.5,18.5,20,21,22.5,22.5)
base
```

```
##    frame game rank
## 1     86    A  1.0
## 2     88    A  2.5
```

```
## 3     88    N  2.5
## 4     89    A  4.5
## 5     89    N  4.5
## 6     90    N  6.0
## 7     91    N  7.5
## 8     91    A  7.5
## 9     92    N  9.0
## 10    95    N 10.0
## 11    95    A 11.5
## 12    95    A 11.5
## 13    96    N 13.5
## 14    96    N 13.5
## 15    97    A 15.0
## 16   100    N 16.5
## 17   100    A 16.5
## 18   101    A 18.5
## 19   101    N 18.5
## 20   102    A 20.0
## 21   104    A 21.0
## 22   108    N 22.5
## 23   108    A 22.5
```

```
NL = 11
AL = 12

#Because the sample size of NL < AL so we sum the ranks of the group NL
base_NL = base[base['game'] == 'N',]
R = sum(base_NL['rank'])

uR = (NL*(NL+AL+1)) / 2
std = sqrt((NL*AL*(NL+AL+1)) / 12)
z = (R-uR) / std
cat(paste("\ntest_value:",z))
```

```
##
## test_value: -0.492365963917331
```

```
cat(paste("\nBecause test value is in the critical value. I don't reject the
null hypothesis. Therefore, there is not enough evidence to conclude that the
re is difference in the number of wins by each league."))
```

```
##
## Because test value is in the critical value. I don't reject the null hypot
hesis. Therefore, there is not enough evidence to conclude that there is diff
erence in the number of wins by each league.
```

The Wilcoxon Signed-Rank Test

13-4.A No Title

```r
#ws = 13, n = 15, α = 0.01, two-tailed
ws = 13
critical_value = 16
cat(paste("Because test value",ws, "< critical value",critical_value, "so the decision is to reject the null hypothesis."))
```

```
## Because test value 13 < critical value 16 so the decision is to reject the null hypothesis.
```

```r
#ws = 32, n = 28, α = 0.025, one-tailed
ws = 32
critical_value = 117
cat(paste("\n\nBecause test value",ws, "< critical value",critical_value, "so the decision is to reject the null hypothesis."))
```

```
##
##
## Because test value 32 < critical value 117 so the decision is to reject th
e null hypothesis.
```

```r
#ws = 65, n = 20, α = 0.05, one-tailed
ws = 65
critical_value = 60
cat(paste("\n\nBecause test value",ws, "> critical value",critical_value, "so the decision is not to reject the null hypothesis."))
```

```
##
##
## Because test value 65 > critical value 60 so the decision is not to reject
the null hypothesis.
```

```r
#ws = 22, n = 14, α = 0.10, two-tailed
ws = 22
critical_value = 26
cat(paste("\n\nBecause test value",ws, "< critical value",critical_value, "so the decision is to reject the null hypothesis."))
```

```
##
##
## Because test value 22 < critical value 26 so the decision is to reject the
null hypothesis.
```

The Kruskal-Wallis Test

The analysis of variance uses the F test to compare the means of three or more populations. The assumptions for the ANOVA test are that the populations are normally distributed and that the population variances are equal. When these

assumptions cannot be met, the nonparametric Kruskal-Wallis test, sometimes called the H test, can be used to compare three or more means.

13-5.2

```
#Mathematics Literacy Scores
cat("H0: There is no difference in the scores of three different parts of the world(claim)\nH1: There is no difference in the scores of three different parts of the world\n")
```

```
## H0: There is no difference in the scores of three different parts of the world(claim)
## H1: There is no difference in the scores of three different parts of the world
```

```
order = c(527,406,474,381,411,520,510,513,548,496,523,547,547,391,549)
score = sort(order)
score
```

```
##  [1] 381 391 406 411 474 496 510 513 520 523 527 547 547 548 549
```

```
region = c("Western","Eastern","Western","Western","Western","Europe","Europe","Europe","Eastern","Western","Eastern","Eastern","Eastern","Europe","Eastern")
rank= c(1,2,3,4,5,6,7,8,9,10,11,12.5,12.5,14,15)
OECD = data.frame(score=score,region=region,rank=rank)
N = 15
alpha = 0.05
df = 2
critical_value = 5.991

cat(paste("\ncritical_value:", critical_value))
```

```
##
## critical_value: 5.991
```

```
#Western
Western = OECD[OECD[,'region'] == 'Western',]
R_W = sum(Western['rank'])
#Europe
Europe = OECD[OECD[,'region'] == 'Europe',]
R_EU = sum(Europe['rank'])
#Eastern
Eastern = OECD[OECD[,'region'] == 'Eastern',]
R_E = sum(Eastern['rank'])

H = (12/(N*(N+1))) * (((R_W)**2)/5+((R_EU)**2)/5+((R_E)**2)/5) - 3*(N+1)

cat(paste("\ntest_value:",H))
```

```
##
## test_value: 7.98
```

```
cat(paste("\nBecause test value is greatern than the critical value. I reject the null hypothesis. Therefore, there is not enough evidence to conclude that there is no difference in the scores of three different parts of the world.")
)
```

```
##
## Because test value is greatern than the critical value. I reject the null hypothesis. Therefore, there is not enough evidence to conclude that there is no difference in the scores of three different parts of the world.
```

The Spearman Rank Correlation Coefficient and the Runs Test

One assumption for testing the hypothesis that r0 for the Pearson coefficient is that the populations from which the samples are obtained are normally distributed. If this requirement cannot be met, the nonparametric equivalent, called the Spearman rank correlation coefficient (denoted by rs), can be used when the data are ranked.

Section 13-6

```
cat("H0: p=0 and H1: p != 0")
```

```
## H0: p=0 and H1: p != 0
```

```
city = c(1,2,3,4,5,6)
subway = c(845,494,425,313,108,41)
rail = c(39,291,142,103,33,38)
rank1 = c(1,2,3,4,5,6)
rank2 = c(4,1,2,3,6,5)
frame = data.frame(city,subway,rank1,rail,rank2)
frame
```

```
##   city subway rank1 rail rank2
## 1    1    845     1   39     4
## 2    2    494     2  291     1
## 3    3    425     3  142     2
## 4    4    313     4  103     3
## 5    5    108     5   33     6
## 6    6     41     6   38     5
```

```
n=6
alpha=0.05
critical_value = 0.886
cat(paste("\ncritical_value:", critical_value))
```

```
##
## critical_value: 0.886
```

```
#subtract
d = frame['rank1'] - frame['rank2']
#squared the difference
d = d**2
E = sum(d)
Rs = 1 - ((6*E)/(6*((6**2)-1)))
cat(paste("\ntest_value:",Rs))
```

```
##
## test_value: 0.6
```

```
cat("\n\nI don't reject the null hypothesis since test value is less than cri
tical value. Therefore, there is not enough evidence to say there is a relati
onship between the daily trips between two services")
```

```
##
##
## I don't reject the null hypothesis since test value is less than critical
value. Therefore, there is not enough evidence to say there is a relationship
between the daily trips between two services
```

Section 14-3

For this assignment, to know the average number of boxes a person buys to get 4 prizes, I just sample for the random boxes and each time 5 persons guess. After that, I have the sum of getting 4 prizes, then divide it by the times of 5 persons to get the average number of boxes a person should buy.

```
set.seed(10)
```

14-3.16

```
o = sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))
```

```r
a= sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))
b = sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))
c = sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))
d = sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))
e = sample(c(1,2,3,4),size = 40,replace=TRUE,p=c(0.25,0.25,0.25,0.25))

total = sum(a == o) + sum(b == o) +sum(c  == o) +sum(d  == o)+sum(e  == o)
average = total/5

average
```

```
## [1] 10.2
```

14-3.18

```r
o = sample(c("b","i","g"),size = 30,replace=TRUE,p=c(0.6,0.3,0.1))


a = sample(c("b","i","g"),size = 30,replace=TRUE)
b = sample(c("b","i","g"),size = 30,replace=TRUE)
c = sample(c("b","i","g"),size = 30,replace=TRUE)
d = sample(c("b","i","g"),size = 30,replace=TRUE)
e = sample(c("b","i","g"),size = 30,replace=TRUE)

total = sum(a == o) + sum(b == o) +sum(c  == o) +sum(d  == o)+sum(e  == o)
average = total/5
average
```

```
## [1] 8.4
```

CONCLUSION:

Non-parametric method although is not the best method to find the result accurately, it is helpful in some situations, for example, we do not have the large sample as population to examine, there are many outliers, we want to understand simply and quickly about small research we are interested, or we have the un normal distribution sample. Since there are many methods for using, we need to conduct its feature to use appropriately. For example, Spearman Rank Correlation is used when you want to find a correlation between two sets of data, or Kruskal-Wallis test is used to find out if two or more medians are different. While 1-sample sign test is used to estimate the median conveniently for 1 sample. Paired-Sample Sign Test is used to test sample means in a comparison of two dependent samples, such as a before-and-after test.

REFERENCES:

Bluman, A. G. (2009). Elementary statistics: A step by step approach. New York, NY: McGraw-Hill Higher Education.