Unit 3 – Introduction to Nonparametrics
Practice Problems
Solutions – *R*

1. (*Source: Moore, D and McCabe, GP. Introduction to the Practice of Statistics, Third Edition)*
Food products are often enriched with vitamins and other supplements. Does the level of a supplement decline over time so that the user receives less than the manufacturer intended? The following are n=9 observations of vitamin C levels (milligrams per 100 grams) in a wheat soy blend, a flour-like product supplied by international aid programs mainly for feeding children. Each bag was measured twice, first at the factory and a second time five months later in Haiti. Researchers suspect that vitamin C levels are generally higher at the factory than they were five months later. We would like to test the hypotheses:

Null, H_O: vitamin C has the same distribution at both times

Alternative, H_A: vitamin C is systematically higher at the factory (one sided)

Bag	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Factory	45	32	47	40	38	41	37	52	37
Haiti	38	40	35	38	34	35	38	38	40

For this exercise, a simple way to get the data into R is to first create a table and then convert the table to a dataframe

```
table1 = read.table(text="
bag factory haiti
1.00    45 38
2.00    32 40
3.00    47 35
4.00    40 38
5.00    38 34
6.00    41 35
7.00    37 38
8.00    52 38
9.00    37 40", header=TRUE)
df1 <- as.data.frame.matrix(table1)
df1$diff <- df1$factory - df1$haiti
```

Tip. In the above, the green is a cut and paste from excel. Specifically, I started with the following R code exactly as you see it here. Then, I replaced the highlighted yellow and green with my edit/copy/paste from Excel. Convenient!

```
table1 = read.table(text="
edit/copy/paste from excel goes here", header=TRUE)
```

(a) What is the correct nonparametric test here?

Wilcoxon Signed Rank Test for single sample of paired data.

(b) Produce a copy of the table above that shows the differences and the ranks of the *absolute* differences.

```
# abs( ) produce absolute values
# rank( ) produces ranks

df1$rank_absdiff <- rank(abs(df1$diff), ties.method = c("average"))
df1

##   bag factory haiti diff rank_absdiff
## 1   1      45    38    7            6
## 2   2      32    40   -8            7
## 3   3      47    35   12            8
## 4   4      40    38    2            2
## 5   5      38    34    4            4
## 6   6      41    35    6            5
## 7   7      37    38   -1            1
## 8   8      52    38   14            9
## 9   9      37    40   -3            3
```

(c) Consider your answer to "a". What is the corresponding normal theory test that would have been performed if the assumptions were met?

Paired t-test for single sample of paired data.

(d) By any means you like, perform the nonparametric test you gave in "a". In 1-2 sentences at most, report. What do you conclude?

```
# wilcox.test( ) calculates difference = 1stvar - 2ndvar
# HA says factory scores are greater than haiti scores --> (difference) is POSITVE -->
# p-value = Pr [ sum positive ranks is observed or greater]
wilcox.test(df1$factory, df1$haiti, paired=TRUE,alternative="greater")

##
##  Wilcoxon signed rank test
##
## data:  df1$factory and df1$haiti
## V = 34, p-value = 0.1016
## alternative hypothesis: true location shift is greater than 0
```

The null hypothesis is ***not*** rejected (p-value = .10). Conclude this sample does ***not*** provide statistically significant evidence that the level of vitamin C supplement declines with time after it leaves the factory.

(e) By any means you like (and just for comparison), perform the normal theory test you gave in "c", even knowing that it is not appropriate here.

```
# Assumes data are in WIDE format (the data for each group is in its own variable (here, factory and Haiti)
t.test(df1$factory, df1$haiti, paired=TRUE,alternative="greater")

##
##  Paired t-test
##
## data:  df1$factory and df1$haiti
## t = 1.5595, df = 8, p-value = 0.07874
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.7053589        Inf
## sample estimates:
## mean of the differences
##                3.666667
```

The p-value from the (*admittedly not appropriate*) normal theory paired t-test is not so different (p-value = .08) from what was obtained by the Wilcoxon Signed Rank test (p-value = .10).

2. (*Source: Moore, D and McCabe, GP. Introduction to the Practice of Statistics, Third Edition).*

The most commonly used measure of economic growth is the rate of growth in a country's total output of goods and services gauged by the gross domestic product (GDP) adjusted for inflation. The level of a country's GDP growth reflects the growth of businesses, jobs, and personal income. Here are World Bank data on the average growth of GDP (percent per year) for the period 2010 to 2013 in developing countries of Europe:

Developing Countries: Europe

Country	Growth	Country	Growth
Albania	2.3	Macedonia, FYR	2.1
Armenia	4.4	Moldova	5.5
Azerbaijan	3.2	Montenegro	1.7
Belarus	4.0	Romania	1.3
Bosnia and Herzegovina	0.4	Serbia	0.9
Bulgaria	0.9	Turkey	6.0
Georgia	5.6	Ukraine	2.9
Kosovo	3.4		

Developing Countries: Central Asia

Country	Growth	Country	Growth
Uzbekistan	8.2	Kyrgyz Republic	4.0
Turkmenistan	11.3	Kazakhstan	6.5
Tajikistan	7.2		

Here, too, I find it simplest to do an EDIT/COPY/PASTE from Excel into my R code

```
table2 = read.table(text="
country growth
europe  2.3
europe  4.4
europe  3.2
europe  4.0
europe  0.4
europe  0.9
europe  5.6
europe  3.4
europe  2.1
europe  5.5
europe  1.7
europe  1.3
europe  0.9
europe  6.0
europe  2.9
asia    8.2
asia    11.3
asia    7.2
asia    4.0
asia    6.5", header=TRUE)
df2 <- as.data.frame.matrix(table2)
```

(a). Suppose we are interested in researching the similarity of average growth of GDP in the two groups of developing countries: Europe versus Central Asia. State the null and alternative hypotheses.

H_O: Distribution of growth in Europe = distribution of growth in Central Asia.
H_A: Distribution of growth in Europe ≠ distribution of growth in Central Asia (two-sided).

(b). Produce a copy of the table above that shows the ranks of the 20 observations. Take care in your ranking to handle ties.

```
df2$rank_growth <- rank(df2$growth, ties.method = c("average"))
df2
```

```
##    country growth rank_growth
## 1   europe    2.3         7.0
## 2   europe    4.4        13.0
## 3   europe    3.2         9.0
## 4   europe    4.0        11.5
## 5   europe    0.4         1.0
## 6   europe    0.9         2.5
## 7   europe    5.6        15.0
## 8   europe    3.4        10.0
## 9   europe    2.1         6.0
## 10  europe    5.5        14.0
## 11  europe    1.7         5.0
## 12  europe    1.3         4.0
## 13  europe    0.9         2.5
## 14  europe    6.0        16.0
## 15  europe    2.9         8.0
## 16    asia    8.2        19.0
## 17    asia   11.3        20.0
## 18    asia    7.2        18.0
## 19    asia    4.0        11.5
## 20    asia    6.5        17.0
```

(c). What is the correct nonparametric test here?

Wilcoxon Rank Sum Test for Two Independent Groups

(d) Consider your answer to "c". What is the corresponding normal theory test that would have been performed if the assumptions were met?

Two Sample t-Test for Two Independent Groups

(e) By any means you like, perform the nonparametric test you gave in "c". In 1-2 sentences at most, report. What do you conclude?

```
# wilcox.test( ) wants data in wide format
europe <- subset(df2, country=="europe")
asia <- subset(df2, country=="asia")

# wilcox.test(x,y,alternative= ) calculates sum of ranks in smaller sample size group
# HA is that growth in europe ≠ growth in asia
wilcox.test(europe$growth,asia$growth, alternative="two.sided")

## Warning in wilcox.test.default(europe$growth, asia$growth, alternative =
## "two.sided"): cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  europe$growth and asia$growth
## W = 4.5, p-value = 0.004526              Assumption of the null has led to an Unlikely result.  Reject the null
## alternative hypothesis: true location shift is not equal to 0
```

The null hypothesis *is* rejected (p-value = .005). Assumption of the null hypothesis and its application to the data has led to a highly unlikely result. Conclude these data provide statistically significant evidence that the average growth of GDP (percent per year) for the period 2010 – 2013 was different in developing countries of Central Asia than it was in developing countries of Europe.

NOTE: The R output here doesn't come with much by way of clues regarding the nature of this two-sided difference, however. A cursory look at the data, however, suggests that the average growth of GDP was higher in the developing countries of Central Asia.

(f) By any means you like (and just for comparison), perform the normal theory test you gave in "d", even knowing that it is not appropriate here.

```
t.test(europe$growth,asia$growth, alternative="two.sided")

##
##  Welch Two Sample t-test
##
## data:  europe$growth and asia$growth
## t = -3.4936, df = 5.3188, p-value = 0.01573
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.694863 -1.238470
## sample estimates:
## mean of x mean of y
##  2.973333  7.440000
```

Interesting. Here, the p-value from the (*admittedly not appropriate*) normal theory two independent samples t-test (p-value = .02) is quite a bit different from what was obtained by the Wilcoxon Rank Sum test (p-value = .005). Note – the discrepancy could go either way.

Supplement - Learn R

Practice 1 - Vectors.

__1a) Create a vector of 100 elements such that the first 20 elements are 1, 2, …, 20, the next 10 elements are 10, 20, 30, …. 100, and the last 70 elements are 31, 32, …., 100.

__1b) Display

```
v1 <- c( seq(from=1,to=20,by=1), seq(from=10,to=100,by=10),seq(from=31,to=100,by=1))

v1
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  10  20  30  40  50  60  70  80
 [29]  90 100  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
 [57]  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
 [85]  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
```

Practice 2 - Matrices.

__2a) Create the following two matrices

m1

1	2	3	4
5	6	7	8

m2

1	3	5	7
2	4	6	8

__2b) Display

```
m1 <- matrix(data=c(1,2,3,4,5,6,7,8),ncol=4,byrow=TRUE)
m2 <- matrix(data=c(1,2,3,4,5,6,7,8),nrow=2,byrow=FALSE)

m1
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8


m2
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
```

Practice 3 - Probability Distribution Calculations.

__3a) Write the code to calculate Pr [Normal (mean=75, sd=4) $\leq$ 76.1].
Show your result.

__3b) Write the code to calculate Pr [Normal (mean=75, sd=4) $\geq$ 76.1].
Show your result.

__3c) Write the code to calculate Pr [Student t-distribution (degrees of freedom=8) $\geq$ 1.645].
Show your result.

__3d) Write the code to obtain
the value of the 90th percentile of a Chi Square distribution with degrees of freedom = 3.
Show your result

```
# a)
# Calculate Pr [ Normal (mean=75, sd=4)  < 76.1 ].  Show result
pnorm(76.1,mean=75,sd=4)                # display, do not save
[1] 0.6083419
pa <- pnorm(76.1,mean=75,sd=4)          # save
pa <- round(pa,digits=4)                # user chooses number of digits to show
pa
[1] 0.6083

# b)
# Calculate Pr [ Normal (mean=75, sd=4)  > 76.1 ].  Show result
pb <- pnorm(76.1,mean=75,sd=4,lower.tail=FALSE)
pb <- round(pb,digits=4)
pb
[1] 0.3917

# c)
# Calculate Pr [ Student T (df = 8)  > 1.645 ].  Show result
pc <- pt(1.645, df=8, lower.tail=FALSE)
pc <- round(pc,digits=4)
pc
[1] 0.0693

# d)
# Calculate the 90th percentile of the Chi Square Distribution with df = 3
pd <- qchisq(.90,df=3)
pd <- round(pd,digits=4)
pd
[1] 6.2514
```