# STAT537: Statistics for Research I: Midterm

Due on November 3, 2016

*Dr. Schmidhammer TR 11:10am – 12:25pm*

**Wenqiang Feng**

# Contents

**Problem 1** 3

**Problem 2** 4

**Problem 3** 6

**Problem 4** 10

**Appendix** 11

R code for Midterm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

# Problem 1

Potencies dataset

*Solution.* (a) **Create a stem-and-leaf plot for these data.**

```
The decimal point is at the |

 22 | 79
 23 | 0234
 23 | 68
 24 | 013
 24 | 589
 25 | 0244
 25 | 89
 26 | 144
 26 | 7799
 27 | 123
```

(b) **Assess the normality of these data.** The following Shapiro-Wilk normality indicates that the p-value is $0.08275 > 0.05$. Hence we do not have enough information to reject the $H_0$. Therefore, we may consider this data obeys normal distribution. And the corresponding Normal QQ plot can be found in Figure.1 which confirms our conclusion.

```
> shapiro.test(potencies);

Shapiro-Wilk normality test

data:  potencies
W = 0.93847, p-value = 0.08275
```

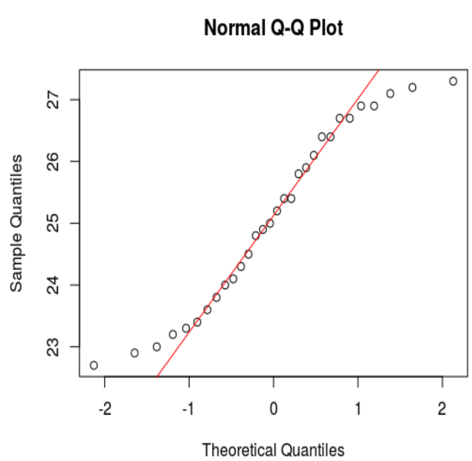


Figure 1: Normal QQ plot.

(c) **Provide a 99% Confidence Interval for the average potency.** Since we do not know the variance of the population, hence the formula of the 99% ($\alpha = 0.01$) Confidence Interval is

$$\left[\mu - t_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \mu + t_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

We can get the result directly from the One Sample t-test package. And the 99% Confidence Interval for the average potency is [24.35735, 25.83599] .

```
One Sample t-test

data:  potencies
t = 0.3604, df = 29, p-value = 0.7212
alternative hypothesis: true mean is not equal to 25
99 percent confidence interval:
 24.35735 25.83599
sample estimates:
mean of x
 25.09667
```

(d) **Based on your results for part (c), can we conclude that the average potency is 25 mg as advertised?** Based on the results from part (c), we have that the p-value $= 0.7212 > 0.01$, hence we do not have enough information to reject $H_0$. Therefore we may conclude that the average potency is 25 mg as advertised. Moreover, $25 \in [24.35735, 25.83599]$ which confirms our conclusion.

□

## Problem 2

WLabor dataset

*Solution.* (a) **Compute the difference scores between percentages for each year and create a stem-and-leaf plot for these difference scores.**

- **The difference scores difference=Year_68-Year_72:**

```
> data
             City Year_68 Year_72 diffence
1            N.Y.    0.42    0.45    -0.03
2            L.A.    0.50    0.50     0.00
3         Chicago    0.52    0.52     0.00
4    Philadelphia    0.45    0.45     0.00
5         Detroit    0.43    0.46    -0.03
6   San Francisco    0.55    0.55     0.00
7          Boston    0.45    0.60    -0.15
8           Pitt.    0.34    0.49    -0.15
9       St. Louis    0.45    0.35     0.10
10    Connecticut    0.54    0.55    -0.01
11    Wash., D.C.    0.42    0.52    -0.10
12          Cinn.    0.51    0.53    -0.02
13      Baltimore    0.49    0.57    -0.08
```

```
14          Newark   0.54    0.53     0.01
15 Minn/St. Paul   0.50    0.59    -0.09
16         Buffalo   0.58    0.64    -0.06
17         Houston   0.49    0.50    -0.01
18       Patterson   0.56    0.57    -0.01
19          Dallas   0.63    0.64    -0.01
```

- **Stem-and-leaf plot for these difference scores:**

```
The decimal point is 1 digit(s) to the left of the |

-1 | 550
-0 | 9863321111
 0 | 00001
 1 | 0
```

(b) **Assess the normality of these difference scores.** The following Shapiro-Wilk normality indicates that the p-value is $0.04503 < 0.05$, then reject the $H_0$. Therefore, we may conclude that this data does not obey normal distribution. And the corresponding Normal QQ plot can be found in Figure.2 which confirms our conclusion.

```
Shapiro-Wilk normality test

data:  diffence
W = 0.89814, p-value = 0.04503
```

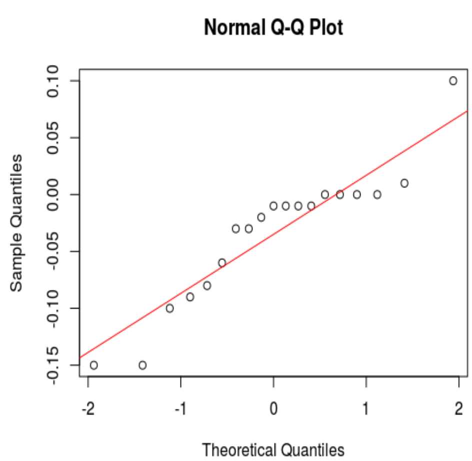


Figure 2: Normal QQ plot.

(c) **Based upon these results, use either the Wilcoxon Signed Ranks test or the t-test to determine whether there is a difference between the average percentages in 1968 and the average percentages in 1972. Use** $\alpha = 0.05$.

- **Wilcoxon Signed Ranks test:** The Wilcoxon signed rank test indicates that the p-value is $0.01324 < 0.05$, hence reject $H_0$. Therefore we may conclude that alternative hypothesis is valid, i.e. there is a difference between the average percentages in 1968 and the average percentages in 1972.

```
Wilcoxon signed rank test with continuity correction

data:  diffence
V = 16, p-value = 0.01324
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 -0.08004133 -0.01002685
sample estimates:
(pseudo)median
   -0.04498224
```

- **t-test:** Similarly, the Paired t-test indicates that the p-value is $0.02435 < 0.05$, hence reject $H_0$. Therefore we may conclude that alternative hypothesis is valid, i.e. there is a difference between the average percentages in 1968 and the average percentages in 1972.

```
Paired t-test

data:  Year_68 and Year_72
t = -2.4577, df = 18, p-value = 0.02435
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.062478527 -0.004889895
sample estimates:
mean of the differences
            -0.03368421
```

(d) **If you decide that the mean percentages differ, estimate this difference with a 95% Confidence Interval.** Based on the results from part (c), the 99% Confidence Interval of the difference is [-0.08004133, -0.01002685] for Wilcoxon signed rank test, and the 99% Confidence Interval of the difference is [-0.062478527 -0.004889895] for t-test .

□

# Problem 3

Weight dataset

*Solution.* (a) **Create a stacked histogram of the data from each of the two therapies, as well as side-by-side box-and-whisker plots.** The histogram and boxplot of the data can be found in Figure.3.

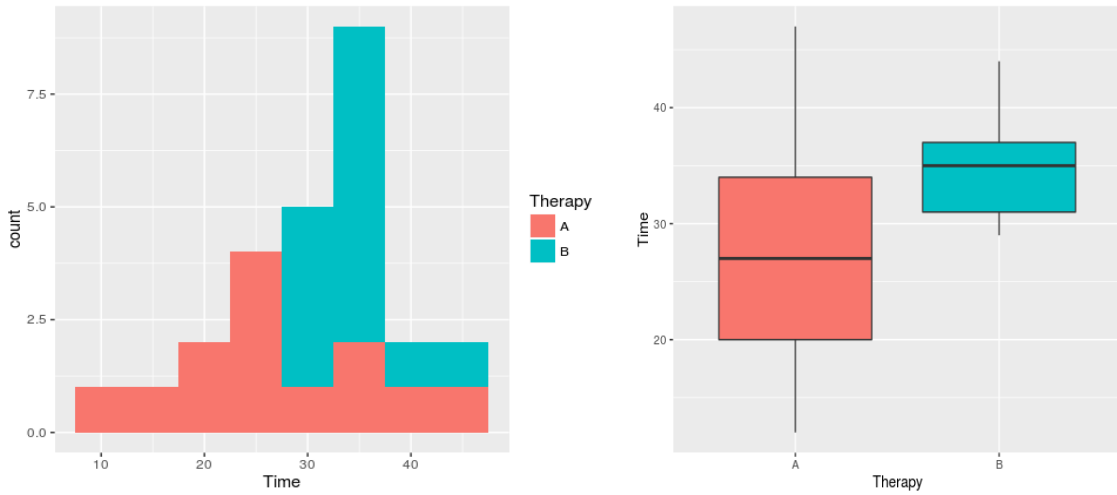


Figure 3: Histogram and boxplot of the data from each of the two therapies.

(b) **Assess the normality of the data from each of the two therapies.**

- **For therapy A:** The following Shapiro-Wilk normality indicates that the p-value is $0.7462 > 0.05$. Hence we do not have enough information to reject the $H_0$. Therefore, we may conclude that this data obeys normal distribution. And the corresponding Normal QQ plot can be found in Figure.4 which confirms our conclusion.

```
Shapiro-Wilk normality test

data:  group_a
W = 0.95952, p-value = 0.7462
```

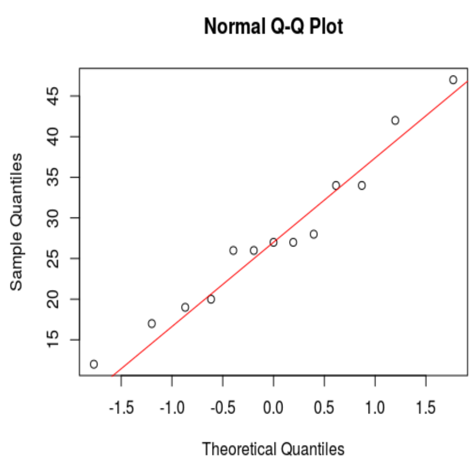


Figure 4: Normal QQ plot.

- **For therapy B:** The following Shapiro-Wilk normality indicates that the p-value is $0.4024 > 0.05$. Hence we do not have enough information to reject the $H_0$. Therefore, we may conclude that this data obeys normal distribution. And the corresponding Normal QQ plot can be found in Figure.4 which confirms our conclusion.

```
Shapiro-Wilk normality test

data:  group_b
W = 0.93559, p-value = 0.4024
```

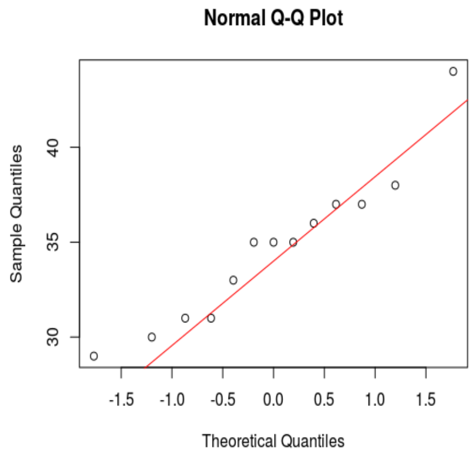


Figure 5: Normal QQ plot.

(c) **Provide a 95% Confidence Interval for the mean times for each of the two therapies.**

- **For therapy A:** From the following One Sample t-test, we get the 95% Confidence Interval for the mean times for each of the therapy A is $[21.67638, 33.55439]$

```
One Sample t-test

data:  group_a
t = 10.131, df = 12, p-value = 3.111e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 21.67638 33.55439
sample estimates:
mean of x
 27.61538
```

- **For therapy B:**From the following One Sample t-test, we get the 95% Confidence Interval for the mean times for each of the therapy B is $[32.25776, 37.12685]$

```
One Sample t-test

data:  group_b
t = 31.048, df = 12, p-value = 7.835e-13
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 32.25776 37.12685
sample estimates:
mean of x
 34.69231
```

(d) **Determine if the variances of the times from the two therapies are equal.** According to the following F test, the p-value = $0.00426 < 0.05$. Hence reject $H_0$. Therefore, we may conclude that the variances of the times from the two therapies are not equal to each other.

```
F test to compare two variances

data:  group_a and group_b
F = 5.951, num df = 12, denom df = 12, p-value = 0.00426
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.815845 19.503164
sample estimates:
ratio of variances
          5.951027
```

(e) **Test whether the means of the times of the two therapies are equal, using either the t-test or the Wilcoxon Rank Sum test, based on the information you obtained above.**

- **Wilcoxon Signed Ranks test:** The Wilcoxon signed rank test indicates that the p-value is $0.01185 < 0.05$, hence reject $H_0$. Therefore we may conclude that alternative hypothesis is valid, i.e. the means of the times of the two therapies are not equal to each other.

```
Wilcoxon rank sum test with continuity correction

data:  group_a and group_b
W = 35, p-value = 0.01185
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -13.999970  -2.000003
sample estimates:
difference in location
             -8.000047
```

- **t-test:** Similarly, the Welch Two Sample t-test indicates that the p-value is $0.02885 < 0.05$, hence reject $H_0$. Therefore we may conclude that alternative hypothesis is valid, i.e.the means of the times of the two therapies are not equal to each other.

```
Welch Two Sample t-test

data:  group_a and group_b
t = -2.4023, df = 15.922, p-value = 0.02885
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -13.3244968  -0.8293494
sample estimates:
mean of x mean of y
 27.61538  34.69231
```

(f) **If you decide that the means of these two therapies are not equal, estimate this difference with a 95% Confidence Interval.** Based on the results from part (d), the 99% Confidence Interval of the difference is [ -13.999970 , -2.000003] for Wilcoxon signed rank test, and the 99% Confidence Interval of the difference is [-13.3244968 , -0.8293494] for t-test .

□

# Problem 4

Weight dataset

*Solution.* (a) **Determine whether clearance to return to work in independent of employee type.** Since the values of each cell are larger than 5, hence we can use $\chi^2$ test to test it. The following Pearson's Chi-squared test indicates that the p-value $= 6.997e-06 < 0.01$, hence reject $H_0$ : Variable A and Variable B are independent. Therefore, we may claim that the clearance to return to work in is dependent on the employee type.

```
Pearson's Chi-squared test with Yates' continuity correction

data:  ctbl
X-squared = 20.194, df = 1, p-value = 6.997e-06
```

(b) **Estimate the proportion of salaried workers granted clearance.** The estimated proportion of salaried workers granted clearance is 0.597561 with the 99% Confidence Interval:[ 0.4499513 , 0.7299512].

```
1-sample proportions test with continuity correction

data:  49 out of 49 + 33, null probability 0.5
X-squared = 2.7439, df = 1, p-value = 0.09763
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.4499513 0.7299512
sample estimates:
       p
0.597561
```

(c) **Estimate the proportion of wage-earning workers granted clearance.** The estimated proportion of wage-earning workers granted clearance is 0.8731343 with the 99% Confidence Interval:[0.7767396 , 0.9326389].

```
1-sample proportions test with continuity correction

data:  117 out of 117 + 17, null probability 0.5
X-squared = 73.142, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.7767396 0.9326389
sample estimates:
        p
0.8731343
```

(d) **Estimate the difference in these two proportions.** The estimated difference is 0.5975610 − 0.8731343 = -0.2755733 with the 99% Confidence Interval:[-0.4433355, -0.1078112]

```
2-sample test for equality of proportions with continuity correction

data:  ctbl
X-squared = 20.194, df = 1, p-value = 6.997e-06
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.4433355 -0.1078112
sample estimates:
   prop 1    prop 2
0.5975610 0.8731343
```

□

# Appendix

## R code for Midterm

Listing 1: Source code for problem 1

```
rm(list = ls())
# set the path or enverionment
#setwd("/Users/wenqiangfeng/Dropbox/UTK_Course/Stat537/Midterm/data")
setwd("/home/feng/Dropbox/UTK_Course/Stat537/Midterm/data")
###############  problem 1 #########################
rawdata = read.table("Potencies.txt")  # read text file
potencies = unlist(rawdata)

# (a)
stem(potencies)

# (b)
## Perform the test
shapiro.test(potencies);

## Plot using a qqplot
```

```
qqnorm(potencies);qqline(potencies, col = 2)

# (c)
t.test(potencies, alternative = c("two.sided"),
                    mu = 25,conf.level = 0.99)

################  problem 2 ##########################
#install.packages("readxl") # CRAN version
library(readxl)
#install.packages("moments")
rawdata = read_excel("WLabor.xlsx",sheet = 1)
attach(rawdata)

# (a)
diffence = Year_68-Year_72
data = cbind(rawdata,diffence)
data
stem(diffence)

# (b)
## Perform the test
shapiro.test(diffence);

## Plot using a qqplot
qqnorm(diffence);qqline(diffence, col = 2)

# (c)
wilcox.test(diffence, conf.int = T,
            alternative="two.sided", conf.level = 0.95)

t.test(Year_68, Year_72,alternative="two.sided",
               paired = TRUE, conf.level = 0.95)

################  problem 3 ##########################
weight = read.table("Weight.dat", header = TRUE)  # read text file
attach(weight)
library(ggplot2)

# (a)
# Overlaid histograms with means
ggplot(weight, aes(x=Time, fill=Therapy)) +
  geom_histogram(binwidth = 5)

# A basic box plot
#ggplot(weight, aes(x=Therapy, y=Time)) + geom_boxplot()
# The above adds a redundant legend. With the legend removed:
ggplot(weight, aes(x=Therapy, y=Time, fill=Therapy)) + geom_boxplot() +
  guides(fill=FALSE)

group_a =weight[c(Therapy=='A'),2]
group_b =weight[c(Therapy=='B'),2]

# (b)
```

```
## Perform the test
shapiro.test(group_a);

## Plot using a qqplot
qqnorm(group_a);qqline(group_a, col = 2)

## Perform the test
shapiro.test(group_b);

## Plot using a qqplot
qqnorm(group_b);qqline(group_b, col = 2)

# (c)
t.test(group_a,alternative = c("two.sided"),conf.level = 0.95)
t.test(group_b,alternative = c("two.sided"),conf.level = 0.95)

# (d)
var.test(group_a,group_b,
         alternative = c("two.sided"),conf.level = 0.95)

# (e)
wilcox.test(group_a,group_b, conf.int = T)

t.test(group_a,group_b,alternative="two.sided",conf.level = 0.95)


################  problem 4 ##########################
# extra
group = c("Salaried", "Wearning")
granted = c(49,117)
Ngranted = c(33,17)
data = data.frame(granted,Ngranted)
data

ctbl=cbind(data$granted,data$Ngranted)
ctbl

# (a)
chisq.test(ctbl)

prop.table(ctbl)

# (b)
prop.test(49, 49+33)

# (c)
prop.test(117, 117+17)

# (d)
prop.test(ctbl, alternative="two.sided",conf.level = 0.99)
```