Swetha Adike

Venu Goud Raparti

Problems from Chapter 4

MSIS 545

18.

In this problem, a study is done on six patients where 3 of them are randomly selected and given treatment while the other 3 serve as a control group. Ranks are assigned to these patients on the basis of severity of the symptoms with 1 being the rank of the patient with severe symptoms and 6 being less symptoms.

The patients in the treatment group got the ranks 3,5,6 and 1,2,4 for control group patients. Below is the solution to find evidence that treatment group has an effect on severity of the symptoms by randomization distribution by calculating sum of ranks.

Below is the combination of ranks to find the sum of ranks in treatment group

treatment group	sum of ranks
1 2 3	6
1 2 4	7
1 2 5	8
1 2 6	9
1 3 4	8
1 3 5	9
1 3 6	10
1 4 5	10
1 4 6	11
1 5 6	12
2 3 4	9
2 3 5	10
2 3 6	11
2 4 5	11
2 4 6	12
2 5 6	13
3 4 5	12
3 4 6	13
3 5 6	14
4 5 6	15

Sum of ranks of treatment group = 3+5+6 = 14

Sum of ranks of control group = 1+2+4 = 7

From the above 20 possible combinations, only last two combinations give the sum of ranks greater or equal to treatment group ranks sum, 14.

Therefore, p value = 2/20 = 0.1

From the p value, we can say that there is no strong evidence that the patients who were offered treatment has effect in showing less symptoms and hence, the treatment does not affect the severity of the symptoms.

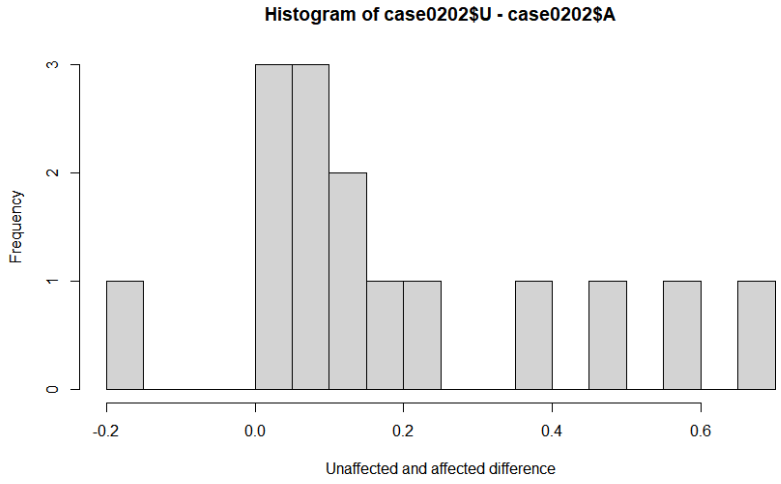
---

26. a

This is a problem for studying the data before evaluating if there is any relation to hippocampus volume in brain size in ideal twins, of which one twin is schizophrenic. The data is a record of 15 hippocampus volumes with respect to the twins affected and unaffected by schizophrenia.

Let’s draw a histogram of the differences in hippocampus volumes for affected and unaffected twins. The code for this is written below

```
> hist(case0202$U - case0202$A, xlab = "Unaffected and affected difference", breaks = 15)
```



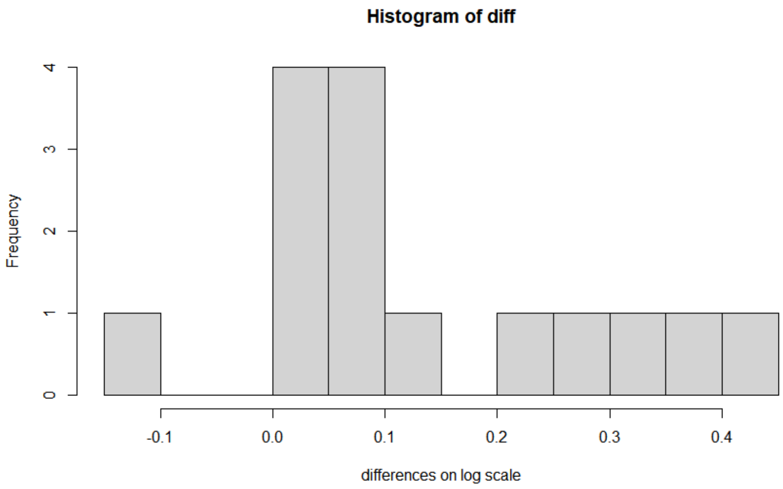
From the histogram, we can see that there is a bit skewness in the data and is not completely symmetric.

b.

Let’s draw a histogram of differences after applying logarithms to hippocampus volumes

```
> logu <- log(case0202$U)

> loga <- log(case0202$A)

> diff <- logu - loga

> hist(diff, xlab = "differences on log scale", breaks = 15)
```

And below is the histogram for the above code, where we can clearly see that the skewness has decreased when previous histogram in 26 a.



c.

Now let’s perform paired t-tests for log transformed data and untransformed data to see if there is any relation between the two

Below is the t-test result performed after changing the data to log scale

```
> t.test(logu, loga, paired = "True")


     Paired t-test
```

```
data:  logu and loga

t = 3.1967, df = 14, p-value = 0.006463

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

 0.04228227 0.21470597

sample estimates:

mean difference

      0.1284941
```

Before going to analyze the t-test results, let’s go for t-test on untransformed data

```
> t.test(case0202$U, case0202$A, paired="True")


     Paired t-test


data:  case0202$U and case0202$A

t = 3.2289, df = 14, p-value = 0.006062

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

 0.0667041 0.3306292

sample estimates:

mean difference

      0.1986667
```

Comparing both the p values are more or less in same zone (<0.005) giving strong evidence that the hippocampus volumes are different for schizophrenia unaffected and affected twins. However, since the data sample is very small, the implication of this evidence to population may or may be considered.

d.

In continuation with the above results, let’s find 95% confidence interval for the mean difference

```
> t.test(logu, loga, paired = "True")$conf
[1] 0.04228227 0.21470597
attr(,"conf.level")
[1] 0.95
> ci <- t.test(logu, loga, paired = "True")$conf
> exp(ci)
[1] 1.043189 1.239497
attr(,"conf.level")
[1] 0.95
```

The above code shows that the confidence interval of p value for hippocampus volumes to be different for schizophrenia affected and unaffected twins is between 4% and 23% with mean of 13.6%. The confidence interval from t test of untransformed data also gives more or less the same interval with similar mean difference.

27.

Let’s find two-sided p-value using the signed-rank test after taking log transformation of the hippocampus volumes and compare with the test on untransformed data to see if there is similarity.

The signed rank test on untransformed data is find by the below code with the results

```
> wilcox.test(case0202$U - case0202$A)


	Wilcoxon signed rank exact test
```

```
data:  case0202$U - case0202$A

V = 111, p-value = 0.002014

alternative hypothesis: true location is not equal to 0
```

And below is the code and results for the signed rank test on the log transformed data (in continuation to the variables assigned in the previous problem, 26)

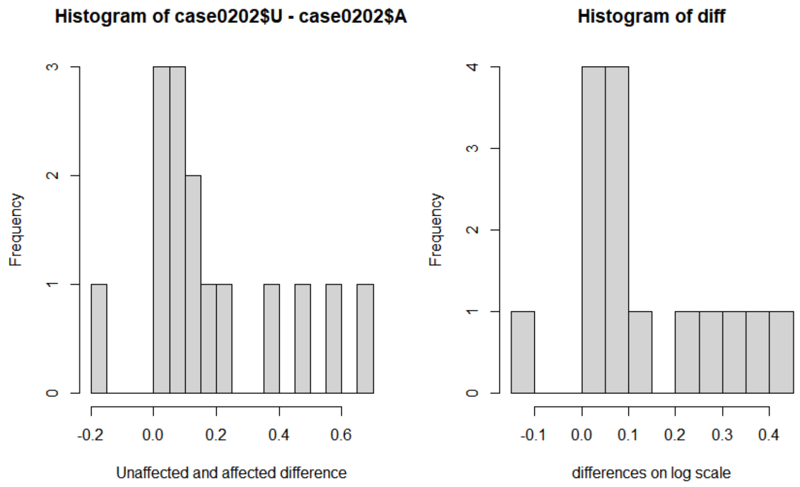
```
> wilcox.test(logu - loga)


     Wilcoxon signed rank exact test


data:  logu - loga

V = 111, p-value = 0.002014

alternative hypothesis: true location is not equal to 0
```

From the above two tests, the two-sided p-value using the signed-rank test on log-transformed data and normal data is 0.002 which is strong evidence (in accordance to the t-tests find in previous problem, 26c). Therefore, we can see that the two-sided p-values of both log-transformed and untransformed data are the same.

Let’s compare the log transformed and untransformed histograms to see if the two are appropriate

Below is the code for drawing histograms of both the two

```
> par(mfrow = c(1,2))

> hist(case0202$U - case0202$A, xlab = "Unaffected and affected difference", breaks = 15)

> hist(diff, xlab = "differences on log scale", breaks = 15)

> par(mfrow = c(1,1))
```



As the two histograms are nearly same, we can say that the assumptions behind the signed rank test are more appropriate on any of these scales.

31.

In this exercise, researchers randomly assigned breast cancer patients to two groups – treatment and control group. The treatment group received weekly 90-minute session of group therapy and self-hypnosis to see if the treatment can improve patient's quality of life and the control group was the observation group which did not receive any.

The data records showed that the patients who was served with treatment lived longer than the control group and this is study to find the statistical evidence whether there is any effect of group therapy treatment on survival time.

Let's head to the statistical summary

```
> summary(ex0431)

    Survival          Group         Censor

 Min.    :  2.00   Control:24   Min.    :0.00000

 1st Qu.: 12.00   Therapy:34   1st Qu.:0.00000

 Median : 18.00                Median :0.00000

 Mean    : 31.41                Mean    :0.05172

 3rd Qu.: 44.50                3rd Qu.:0.00000

 Max.    :122.00                Max.    :1.00000
```

The data shows that the control group has 24 sample patients and therapy 34. The sample is not so bad to draw the statistical conclusions. However, there are 3 patients with life expected months greater than 122 months who were still alive at the time of data collection as given in the problem.

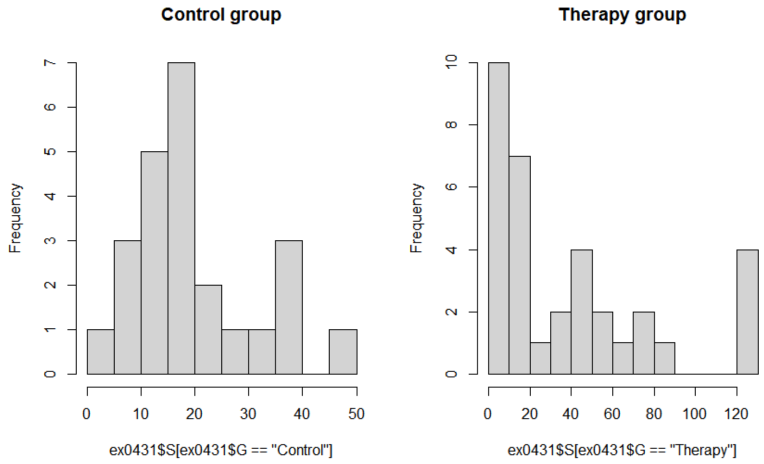
```
> tapply(ex0431$S, ex0431$G, summary)

$Control

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

    2.0    13.5    17.0    20.0    23.0    48.0



$Therapy

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

   2.00   10.00   21.00   39.47   58.00  122.00
```

The above summary tells that both the therapy and control group differ a lot. However, the medians seem to be a bit closer. Looking at the data distribution and spread, it tells that logarithm transformations has to be made on the data since the minimum and maximum values vary a lot.

Let’s see if there is any skewness in the data by histogram. Below is the code for it

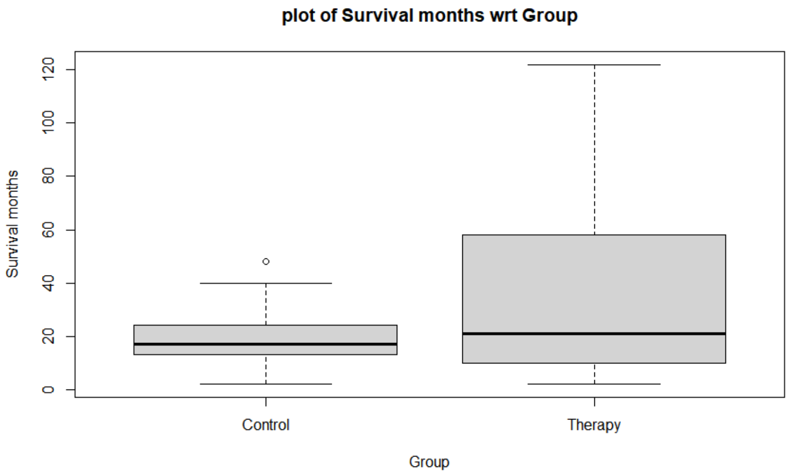
```
> par(mfrow = c(1,2))

> hist(ex0431$S[ex0431$G == "Control"], main = "Control group", breaks = 15)

> hist(ex0431$S[ex0431$G == "Therapy"], main = "Therapy group", breaks = 15)

> par(mfrow = c(1,1))
```

And below is histogram for the code, which shows a little skewness in the data.



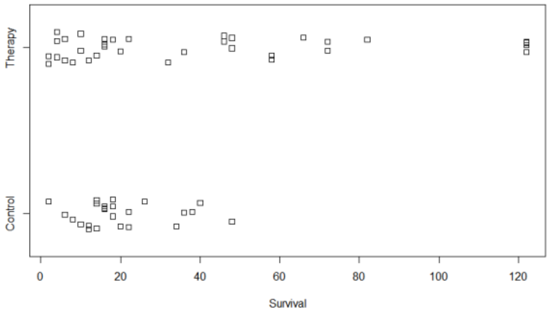
Let's get to boxplot to see how the data is distributed

```
> boxplot(ex0431$S ~ ex0431$G, xlab = "Group", ylab = "Survival months",
+         main = "plot of Survival months wrt Group")
```



The below boxplot shows that the tail of therapy group is long while control group is quite contended which tells the need for log transformations but let's take a look at strip chart

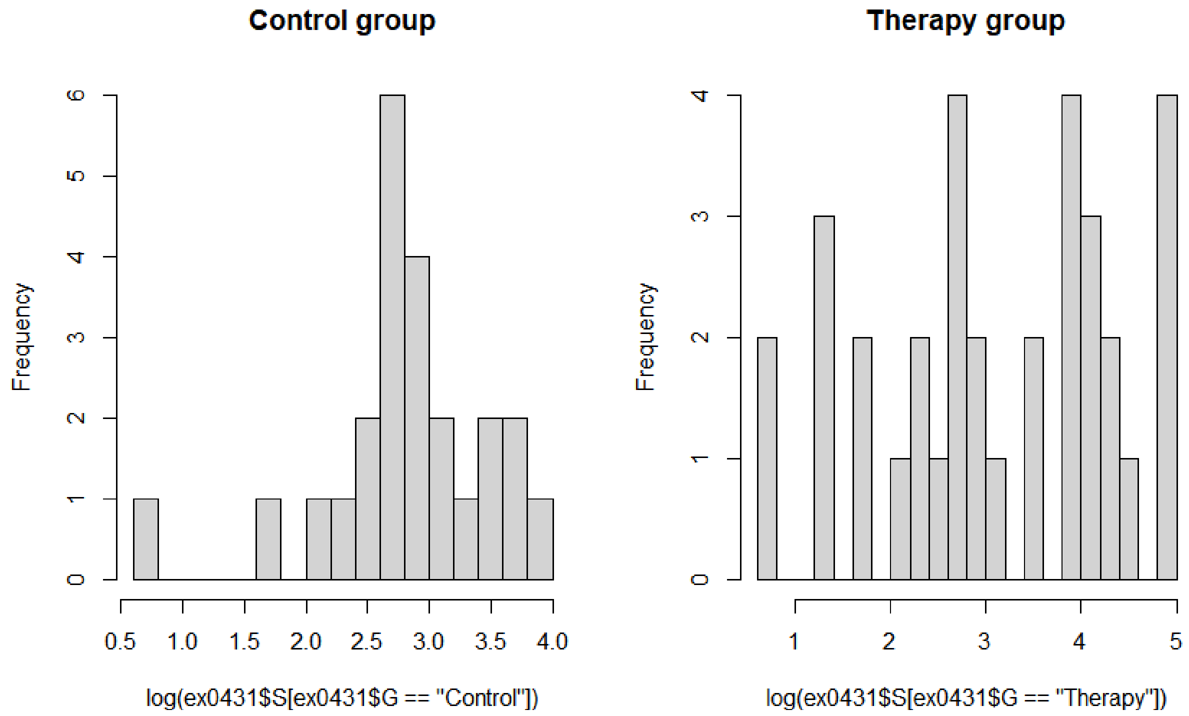
```
> stripchart(Survival ~ Group, data = ex0431, method = "jitter")
```



The strip chart clearly says to go for log transformations.

Let's see how the histogram and boxplot changes after applying log scale on the survival group

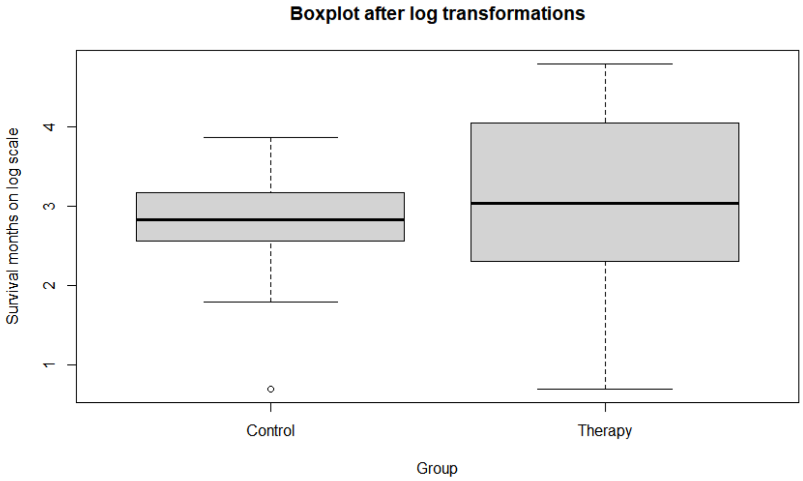
```
> par(mfrow = c(1,2))
> hist(log(ex0431$S[ex0431$G == "Control"]), main = "Control
group", breaks = 15)

> hist(log(ex0431$S[ex0431$G == "Therapy"]), main = "Therapy
group", breaks = 15)

> par(mfrow = c(1,1))
```



The above histogram shows more good spread of the data than compared to data on normal survival scale.

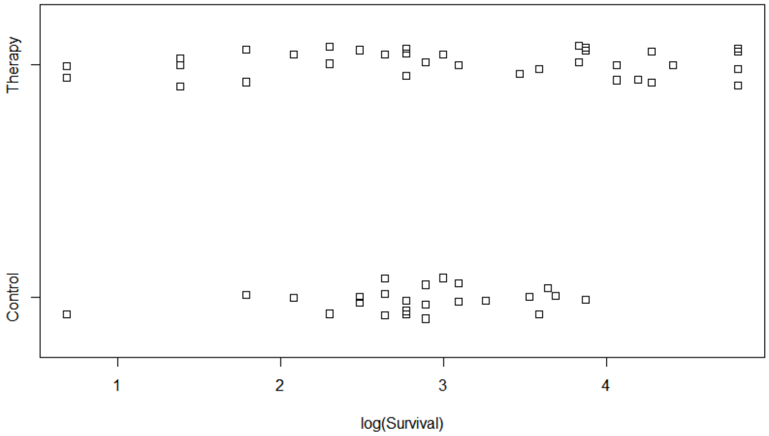
Let’s take a look at boxplot, the code is as below

```
> boxplot(log(ex0431$S) ~ ex0431$G, xlab = "Group", ylab =
"Survival months on log scale",

+          main = "Boxplot after log transformations")
```



The boxplot also seems to be good with no long tails. Below is the code for strip chart to see the overall spread and check for skewness

```
> stripchart(log(Survival) ~ Group, data = ex0431, method =
"jitter")
```



The strip chart also gives good picture of the data now.

To infer the statistical evidences to see if there is any effect in patient’s life when therapy is offered, let’s perform t-test

```
>                                   t.test(log(ex0431$S)[ex0431$G
=="Therapy"],log(ex0431$S)[ex0431$G =="Control"])
```

```
     Welch Two Sample t-test


data: log(ex0431$S)[ex0431$G == "Therapy"] and
log(ex0431$S)[ex0431$G == "Control"]

t = 1.1209, df = 53.683, p-value = 0.2673

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:

 -0.2218461  0.7843323

sample estimates:

mean of x mean of y

 3.093171  2.811927
```

By the t-test, a statistical interpretation can be made that the evidence that there is indeed no effect of therapy group on the patient’s survival time. The results show that there is a mean difference is 0.28 on log scale. Converting this mean to normal scale by exponentiation we get e^0.28 = 1.32 which tells that median of therapy group is 37% higher than control group.

Now let’s perform Wilcox Rank Sum test infer further

```
> wilcox.test(log(Survival) ~ Group, data = ex0431, conf.int =
TRUE)


     Wilcoxon rank sum test with continuity correction


data:  log(Survival) by Group

W = 337, p-value = 0.265
```

```
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.9807764  0.2513536
sample estimates:
difference in location
            -0.3184537


Warning messages:
1: In wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...) :
  cannot compute exact confidence intervals with ties
```

The p-values in the both tests are equal to 0.266 approximately, which shows that there is no evidence of treatment effect. And so, there is no way we estimate the expectant life time when women receive the therapy.