Swetha Adike

Venu Goud Raparti

Problems from Chapter 3

MSIS 545

24.a

This is a problem regarding the variation or gender discrimination in terms of salaries for men and women. The data consists of a sample of the salaries paid for 32 men and 61 women, skilled entry level clerical employees hired by a bank. Since this group of gender was not decided by previously, this is not a randomized experiment but an observational one.

Heading to summary of the data, we get

```
> summary(case0102)
     Salary         Sex
 Min.   :3900   Female:61
 1st Qu.:4980   Male  :32
 Median :5400
 Mean   :5420
 3rd Qu.:6000
 Max.   :8100
> max(case0102$Salary)/min(case0102$Salary)
[1] 2.076923
```

From the above data, the largest to smallest measurement is 8100/3900 = 2.08 < 10 and hence there is no need of using log scale to make an inference but let's just see if there is any difference in the statistical findings if a log scale is used to express salaries. This can be done by the below code

```
> salary.log <- log(case0102$Salary)
```

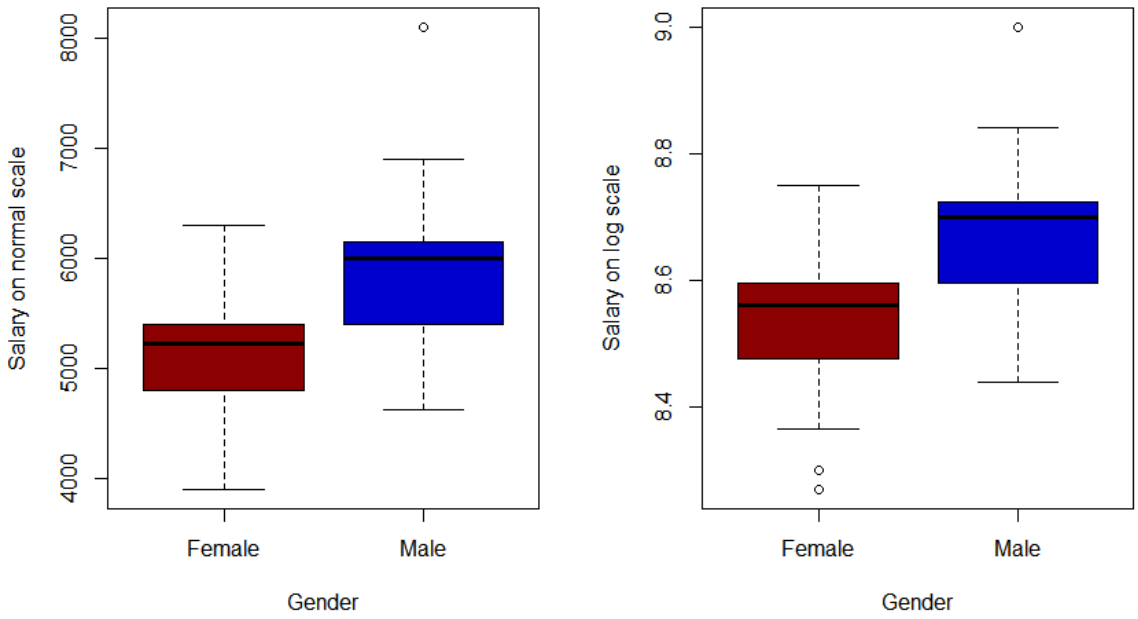
So by the above code, the salaries are turned on to log scale which can be fetched by the variable salary.log. Let's now compare the summary statistics of the data with and without log scale

```
> tapply(case0102$Salary, case0102$Sex, summary)
$Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
   3900    4800    5220    5139    5400    6300


$Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4620    5400    6000    5957    6075    8100


> tapply(salary.log, case0102$Sex, summary)
$Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.269   8.476   8.560   8.539   8.594   8.748


$Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.438   8.594   8.700   8.686   8.712   9.000
```

So, both the statistical summary shows that there is a bit difference in means, for which male mean and median is higher than female ones. One difference that can be observed when both the summaries are compared is that the normal data is widely spread while the log data is spread got decreased (especially for female, where the spread is 0.479). Let's see if this spread can be clearly depicted in a box plot or both of them look similar.

```
> par(mfrow = c(1,2))
> boxplot(case0102$Salary ~ case0102$Sex, col = c("red4","blue3"),
xlab = "Gender",
+         ylab = "Salary on normal scale")
> boxplot(salary.log ~ case0102$Sex, col = c("red4","blue3"), xlab
= "Gender",
+         ylab = "Salary on log scale")
> par(mfrow = c(1,1))
```



From the above box plot a little variation in female plot can be observed where two outliers has appeared on the log scale where as no outliers in the normal salaries. However, there is no much difference in the male group. Both the plots for male on normal scale and log scale look much similar.

b.

The t-test results on the log salary data is given by the below code

```
> t.test(salary.log[case0102$Sex =="Male"],
salary.log[case0102$Sex =="Female"],var.equal = TRUE, alt =
"great")


     Two Sample t-test


data:  salary.log[case0102$Sex == "Male"] and
salary.log[case0102$Sex == "Female"]

t = 6.1715, df = 91, p-value = 9.245e-09

alternative hypothesis: true difference in means is greater than
0

95 percent confidence interval:
```

```
 0.1073767       Inf
sample estimates:
mean of x mean of y
 8.685992  8.539048
```

The t-test results show a strong convincing evidence that the means of the salary groups for male and female differ with one sided p value < 0.0001. There is a difference of 0.147 between male and female groups.

c.

The 95% confidence interval for the data on log scale of salaries is found by the below code for two-sided t test

```
> t.test(salary.log[case0102$Sex =="Male"],
salary.log[case0102$Sex =="Female"],var.equal = TRUE)


     Two Sample t-test


data:  salary.log[case0102$Sex == "Male"] and
salary.log[case0102$Sex == "Female"]
t = 6.1715, df = 91, p-value = 1.849e-08
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.09964777 0.19423950
sample estimates:
mean of x mean of y
 8.685992  8.539048
```

Extracting the confidence interval, we get

```
> ci <- t.test(salary.log[case0102$Sex =="Male"],
salary.log[case0102$Sex =="Female"],
+              var.equal = TRUE)$conf
> ci
[1] 0.09964777 0.19423950
```

```
attr(,"conf.level")
[1] 0.95
```

Converting the confidence interval of salaries from logarithmic scale to normal scale by exponentiation,

```
> exp(ci)
[1] 1.104782 1.214387
attr(,"conf.level")
[1] 0.95
```

Inference: The results in 24 b show that there is a mean difference of 0.147 on log scale between male and female groups, male salary being higher than female group. Converting this mean to normal scale by exponentiation we get e^0.147 = 1.16 which tells that median salary of male is 16% higher than female. Going to the results we got in 24 c, the 95% confidence interval after inverting from the log scale is between 1.11 to 1.21 which mean that the median of male salaries is greater than female salaries with 95% confidence. Since 16% is in between this interval, it is more accurate to impose log scale on salaries when applied to another population, to get better results.

28.

This problem is on a study of house sparrows's humerus length and its effect on them after a winter whether they have survived or perished. The data is a numerical collection of humerus length of house sparrows after a severe winter and it is grouped according to whether they have survived or perished. The study is to analyze and summarize the evidence that if the humerus length distribution differed in the two groups.

Heading to the data,

```
> #28
> head(ex0221)
  Humerus   Status
1   0.687 Survived
2   0.703 Survived
3   0.709 Survived
4   0.715 Survived
5   0.728 Survived
6   0.721 Survived
```

```
> table(ex0221$S)


Perished Survived
      24       35
```

We have a sample of 59 humerus lengths and the status of the sparrows with two groups – either perished or survived. The sample contained 24 humerus lengths which have perished and 35 survived.

The below are the summary statistics

```
> summary(ex0221)
    Humerus            Status
 Min.   :0.6590   Perished:24
 1st Qu.:0.7245   Survived:35
 Median :0.7360
 Mean   :0.7339
 3rd Qu.:0.7470
 Max.   :0.7800
```

The summary show that minimum humerus length is 0.659 inches. Observing the first quartile, median, mean and 3rd quartile, they all seem to be very close, which gives a necessity for a log scale spread.

Before having a deeper look into the summary statistics, lets check whether we can impose log scale on the humerus length in this data

```
> max(ex0221$Humerus)/min(ex0221$Humerus)
[1] 1.183612
```

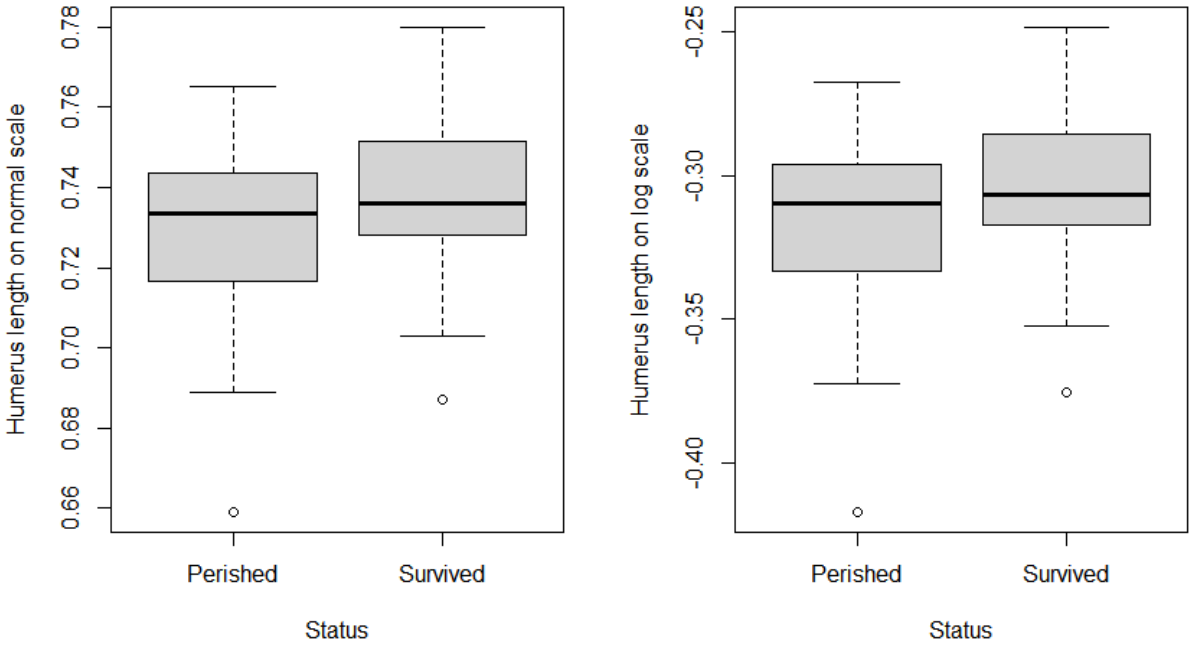
Since 1.18 < 10, there is logically no need for log scale. On this note, moving on to summary again

```
> tapply(ex0221$H, ex0221$S, summary)
$Perished
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6590  0.7183  0.7335  0.7279  0.7432  0.7650


$Survived
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6870  0.7280  0.7360  0.7380  0.7515  0.7800
> tapply(ex0221$H, ex0221$S, sd)
  Perished   Survived
0.02354259 0.01983906
```

The spread of the two groups is more or less the same with almost close numbers. Let's have a look at the box plot to see if there are any outliers

```
> par(mfrow = c(1,1))
> par(mfrow = c(1,2))
> boxplot(ex0221$H ~ ex0221$S, xlab = "Status", ylab = "Humerus length on normal scale")
> boxplot(humerus.log ~ ex0221$S, xlab = "Status", ylab = "Humerus length on log scale")
> par(mfrow = c(1,1))
```



The above boxplots show that there are two common outliers.

More or less there is significantly no difference between the two box plots on normal and log scale.

The one-sided t test results for this data on normal and log scale are as follows

```
>   t.test(ex0221$H[ex0221$S   =="Survived"],   ex0221$H[ex0221$S
=="Perished"], alt = "great")


     Welch Two Sample t-test


data:  ex0221$H[ex0221$S == "Survived"] and ex0221$H[ex0221$S ==
"Perished"]

t = 1.7207, df = 43.824, p-value = 0.04618

alternative hypothesis: true difference in means is greater than
0

95 percent confidence interval:

 0.0002363793          Inf

sample estimates:

mean of x mean of y

0.7380000 0.7279167


> t.test(humerus.log[ex0221$S =="Survived"],
humerus.log[ex0221$S =="Perished"], alt = "great")


     Welch Two Sample t-test


data:  humerus.log[ex0221$S == "Survived"] and
humerus.log[ex0221$S == "Perished"]

t = 1.7135, df = 42.834, p-value = 0.04693

alternative hypothesis: true difference in means is greater than
0

95 percent confidence interval:

 0.0002619841          Inf

sample estimates:

 mean of x  mean of y
```

```
-0.3041635 -0.3180825
```

Inverting the means of t test from log to normal scale by exponentiation, we get the results similar to normal t test. Hence there is necessarily no need of log scale transformations required in this study. Being said that let's see if we get any change in result if one of the outliers (0.659") is removed.

The humerus length 0.659 is an outlier and a minimum value with "Perished" status. The commands for removing this record from the data is

```
> ex0221.2 <- ex0221[-c(36),]
```

Since the outlier 0.659" is the 36th observation, it has been removed and the data after this is as below

```
> ex0221.2
   Humerus   Status
1    0.687 Survived
2    0.703 Survived
3    0.709 Survived
4    0.715 Survived
5    0.728 Survived
6    0.721 Survived
7    0.729 Survived
8    0.723 Survived
9    0.728 Survived
10   0.723 Survived
11   0.726 Survived
12   0.728 Survived
13   0.736 Survived
14   0.733 Survived
15   0.730 Survived
16   0.733 Survived
17   0.730 Survived
18   0.739 Survived
```

19	0.735	Survived
20	0.741	Survived
21	0.741	Survived
22	0.749	Survived
23	0.741	Survived
24	0.743	Survived
25	0.741	Survived
26	0.752	Survived
27	0.752	Survived
28	0.751	Survived
29	0.756	Survived
30	0.755	Survived
31	0.766	Survived
32	0.767	Survived
33	0.769	Survived
34	0.770	Survived
35	0.780	Survived
37	0.689	Perished
38	0.703	Perished
39	0.702	Perished
40	0.709	Perished
41	0.713	Perished
42	0.720	Perished
43	0.729	Perished
44	0.726	Perished
45	0.726	Perished
46	0.720	Perished
47	0.737	Perished

```
48   0.739 Perished
49   0.731 Perished
50   0.738 Perished
51   0.736 Perished
52   0.738 Perished
53   0.744 Perished
54   0.745 Perished
55   0.743 Perished
56   0.754 Perished
57   0.752 Perished
58   0.752 Perished
59   0.765 Perished
```

Though we could see 59 samples, the data record of 36 will be deleted and is not seen in the above sample data.

Performing one-sided t test on this new data, we get

```
> t.test(ex0221.2$H[ex0221.2$S =="Survived"],
ex0221.2$H[ex0221.2$S =="Perished"], alt = "great")


     Welch Two Sample t-test


data:  ex0221.2$H[ex0221.2$S == "Survived"] and
ex0221.2$H[ex0221.2$S == "Perished"]
t = 1.373, df = 48.967, p-value = 0.08801
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 -0.001567186          Inf
sample estimates:
mean of x mean of y
 0.738000  0.730913
```

Hence comparing the first and the above t test, it is clear that the conclusion depends on the value 0.659" as the statistical evidence shifted from strong convincing (p value 0.046) evidence to suggestive but inconclusive evidence (p value 0.089). P value depended on the removal of the lower value in the perished group which tells that there is no significant difference in status of the sparrows with the change in humerus length. Also, since this was not a randomized experiment, inferences cannot be made directly and so care should be taken in interpreting the results when inferring to large population, which is not suggestive.

29. The cloud seeding case study is collection of data to test a hypothesis that injecting silver iodide into the clouds can increase the rainfall. On 52 days of observation, a random mechanism was used to decide whether to seed the cloud or leave it unseeded. The precipitation was measured and the data is collected for both seeded and unseeded days.

This exercise is a study of how the data is varied for additive seeding days, by increasing the rainfall by 100 mm each by four times and multiplicative seeding days, by multiplying the rainfall by 2,3,4, and 5. Lets head to the data

```
> #29
> head(case0301)
  Rainfall Treatment
1   1202.6  Unseeded
2    830.1  Unseeded
3    372.4  Unseeded
4    345.5  Unseeded
5    321.2  Unseeded
6    244.3  Unseeded
```

The summary of the data is shown as below

```
> summary(case0301)
    Rainfall         Treatment
 Min.   :   1.0   Seeded  :26
 1st Qu.:  28.9   Unseeded:26
 Median : 116.8
 Mean   : 303.3
 3rd Qu.: 307.4
```

```
 Max.    :2745.6
```

It is clearly seen that the log scale transformations have to be used to get the evidence of the effect of seeding and unseeding on the rainfall. But here, since we are just examining the data and not performing any t test, let us consider the data as it is.

Creating a variable for seeding and unseeded days, we get

```
> unseeded <- case0301$Rainfall[case0301$T == "Unseeded"]
> seeded <- case0301$R[case0301$T == "Seeded"]
```

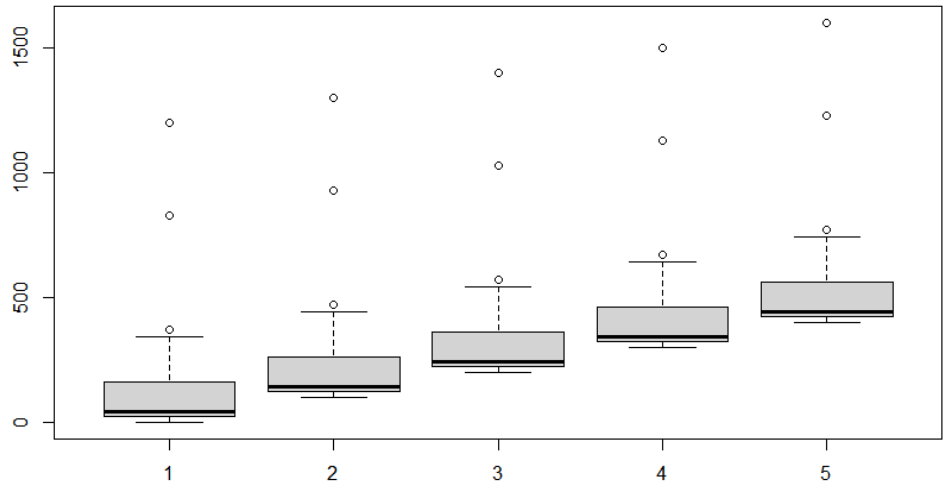
a.

Creating four new variables by adding 100, 200, 300 and 400 to each of the unseeded day rainfall amounts,

```
> unseeded100 <- unseeded + 100
> unseeded200 <- unseeded + 200
> unseeded300 <- unseeded + 300
> unseeded400 <- unseeded + 400
```

The boxplot of the five seeded rainfall data is as follows and let us examine what happens if the seeding goes additive

> boxplot(unseeded, unseeded100, unseeded200, unseeded300, unseeded400)

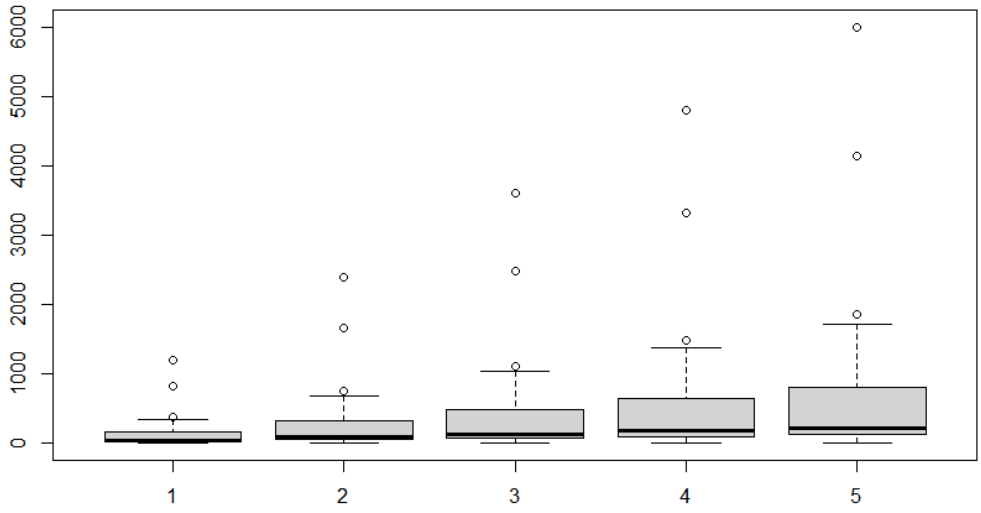


b. creating four additional variables by multiplying the data by 2, 3, 4, and 5 to see the change in the data if the seeding goes multiplicative

```
> unsmul2 <- unseeded*2
> unsmul3 <- unseeded*3
> unsmul4 <- unseeded*4
> unsmul5 <- unseeded*5
```

Boxplot of the above gives

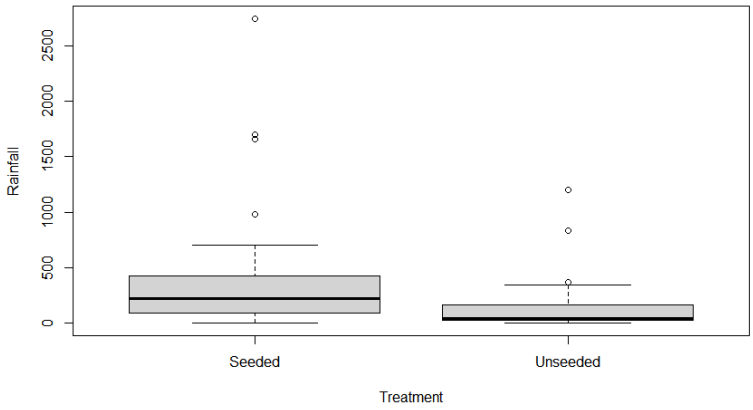
```
> boxplot(unseeded, unsmul2, unsmul3, unsmul4, unsmul5)
```



c.

The above two The original boxplot of the seeded and unseeded rainfall data is

```
 > boxplot(Rainfall ~ Treatment, data = case0301)
```



Comparing the unseeded rainfall data with additive and multiplicative data, it can be seen that with respect to the placement of the outliers or the first quartile, closeness of mean to first quartile, the original data is more similar to multiplicative data boxplot in b than the additive data.

33. This is an observational data which shows relative brain weights for the mammals with average litter size less than 2 as one group (Small) and greater than 2 as another group (Large). The study is to make an evidence if there is any connection between brain sizes to be different for two groups. Getting into the data details,

```
> #33
> head(ex0333)
  BrainSize LitterSize
1      0.42      Small
2      0.86      Small
3      0.88      Small
4      1.11      Small
5      1.34      Small
6      1.38      Small
> summary(ex0333)
   BrainSize      LitterSize
 Min.   : 0.420   Large:45
 1st Qu.: 2.740   Small:51
 Median : 6.635
 Mean   : 8.800
 3rd Qu.:12.357
 Max.   :36.350
```

The data was collected on a sample size of 96 species of mammals with 45 species of Large samples and 51 small samples. There is a great variation in the brain size where the minimum size is found to be 0.420 and maximum size is 36.350.

Let's check if there is a necessity for log scale transformations

```
> max(ex0333$B)/min(ex0333$B)
[1] 86.54762
```

Since the ration of larger value to smaller value is greater than 10, it is obvious to us logarithmic scale on brain size to interpret the results. The transformation can be done by the command

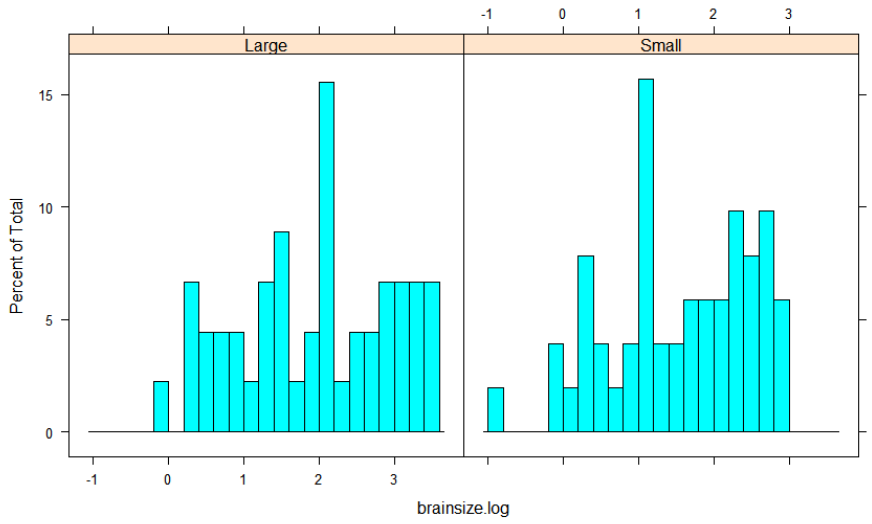
```
> brainsize.log <- log(ex0333$BrainSize)
```

Going into the summary statistics,

```
> tapply(brainsize.log, ex0333$L, summary)
$Large
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.06187  1.22083  2.07568  1.94943  2.92370  3.59319


$Small
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.8675  0.9581  1.6094  1.5525  2.3461  2.9957


> tapply(brainsize.log, ex0333$L, sd)
    Large     Small
1.0162933 0.9522342
```

We can see that there is a good spread of the data in both large and small groups where it is difficult to interpret whether there is a similarity in the data with respect to the brain size and litter size. Let's analyze the data with a histogram

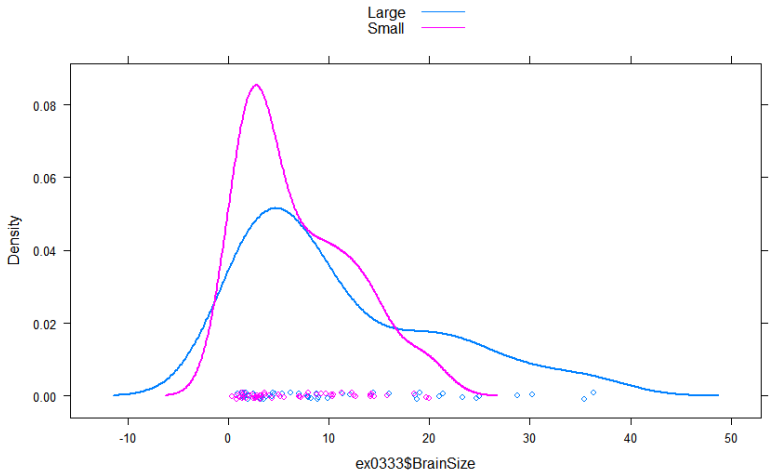
```
> histogram( ~ brainsize.log | ex0333$L, breaks = 15)
```



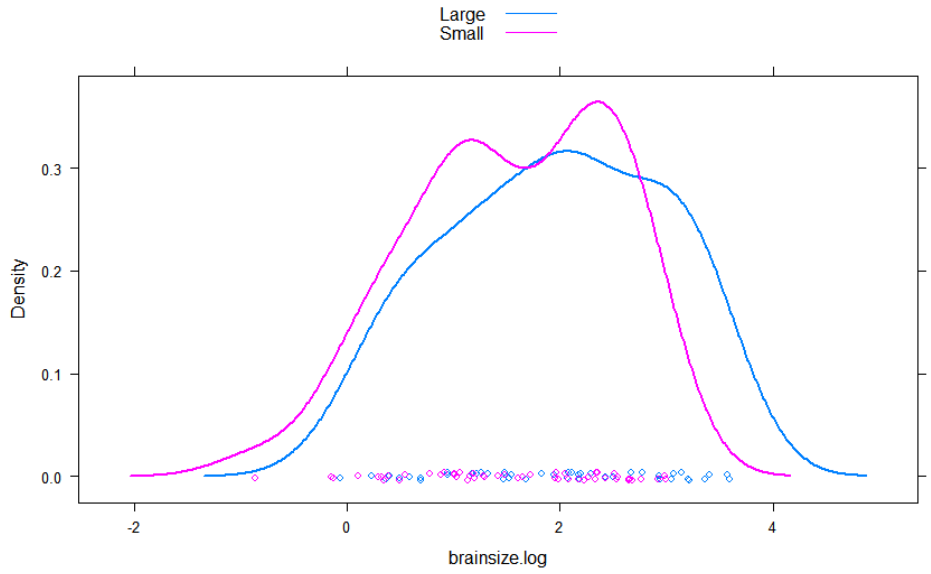
The histogram shows that there is some similarity in the data in its spread and few data points and connections. Lets have clear picture by box plot.

Proceeding further, let's have a look at the density plot of the data before and after the log transformations. The code for it is as below

```
> densityplot(~ex0333$BrainSize, groups = ex0333$LitterSize, auto.key = T, lwd = 2)
```



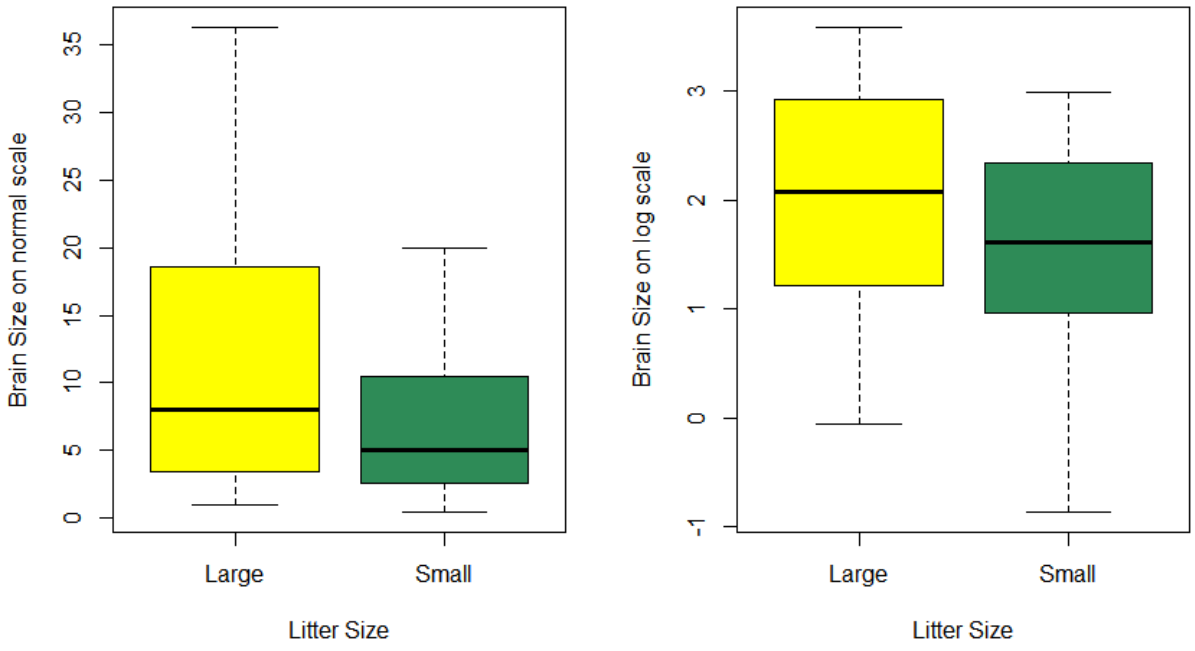

```
> densityplot(~brainsize.log, groups = ex0333$LitterSize, auto.key = T, lwd = 2)
```



It can be seen clearly that the original data is skewed rightly. Since the log transformations reduce the skewness, it changed after log transformations.

Looking at the boxplots before and after the log transformations

```
> par(mfrow = c(1,2))
> boxplot(ex0333$B ~ ex0333$L, col = c("yellow","seagreen"),
+           xlab = "Litter Size", ylab = "Brain Size on normal
scale")
> boxplot(brainsize.log ~ ex0333$L, col = c("yellow","seagreen"),
+         xlab = "Litter Size", ylab = "Brain Size on log scale")
> par(mfrow = c(1,1))
```



Looking at both the box plots, it is clear that the spread and data distribution has changed after imposing log on brain size. By the above density plot and boxplot, we can clearly conclude that using the logarithmic scale is main crux of finding the statistical evidence.

From the log scale box plot on the right, we can see that there is no outlier in the data where the mean in the two groups divide the data into two (almost) equal parts. The tail of the boxplot of small litter sized mammals is long, starting from a very low value. The two box plots in the right show that there is no much difference between the large and small litter sized mammals with the brain size but the 3rd quartile of small litter size plot ends just above the mean of the large litter size.

Though the large litter sized mammals plot seemed to have higher brain sizes by the boxplot, let infer the result by a t-test

```
> t.test(brainsize.log ~ ex0333$L, var.equal = TRUE)


     Two Sample t-test


data:  brainsize.log by ex0333$L

t = 1.975, df = 94, p-value = 0.0512

alternative hypothesis: true difference in means between group
Large and group Small is not equal to 0

95 percent confidence interval:

 -0.00210957  0.79604609

sample estimates:

mean in group Large mean in group Small

           1.949426            1.552458
```

By the t-test, a statistical interpretation can be made that the evidence that the brain size is different for different groups of litter sizes is strong but not convincing with two-sided p value 0.512. The results show that there is a mean difference is 0.397 on log scale between large and small litter size group. Converting this mean to normal scale by exponentiation we get e^0.397 = 1.487 which tells that median of brain size (Relative brain weight with no units) in large size litter mammals is 48.7% higher than in small litter sized mammals.

Considering the confidence intervals,

```
> conf <- t.test(brainsize.log ~ ex0333$L, var.equal = TRUE)$conf

> conf

[1] -0.00210957  0.79604609

attr(,"conf.level")

[1] 0.95

> exp(conf)

[1] 0.9978927 2.2167587

attr(,"conf.level")

[1] 0.95
```

The 95% confidence interval after inverting from the log scale is between 0.997 to 2.217 which mean that the median of large litter size brain size is greater than small litter sized brain size with

95% confidence interval of 0.2% (1-0.997) smaller and 121.7% (1-2.217) larger. Since this is not a randomized experiment but the results are bit swinging with no strong evidence, applying this to larger population can be done accordingly, keeping in view of p value.