Swetha Adike
Venu Goud Raparti
Problems for Chapter 2
MSIS 545

14. This study was a random experiment where male volunteers with high blood pressure were divided into two groups by randomly giving them black and red playing cards. One group was specific to fish oil diet while another with regular oil diet. From this the effect of the diet can be inferred on the blood pressure.

Heading to the fish oil and blood pressure details, we get

```
> #14
> head(ex0112)
  BP    Diet
1  8 FishOil
2 12 FishOil
3 10 FishOil
4 14 FishOil
5  2 FishOil
6  0 FishOil
> table(ex0112$Diet)

   FishOil RegularOil
         7          7
```

It is seen that there are 7 samples (men) are present in each of the oil groups to check the.

The below are the summary statistics of the data which shows that there is a huge difference in means and medians on fish oil diet and regular oil diet.

```
> tapply(ex0112$BP, ex0112$Diet, summary)
$FishOil
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   8.000   6.571  11.000  14.000

$RegularOil
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -6.000  -3.500   0.000  -1.143   1.500   2.000

> tapply(ex0112$BP, ex0112$Diet, sd)
   FishOil RegularOil
  5.855400   3.184785
```

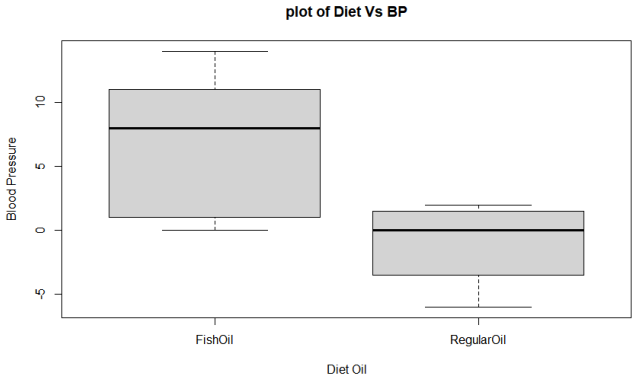
We could see that there is a low change in blood pressure for regular oil with a minimum value of -6.00 to a maximum change of 2.00 while fish oil diet gave a change in blood pressure (lowered BP) varying from minimum of 0.00 to 14.00. The difference between the two groups are varied to large

extent. The means of both the diet groups are very apart where there is 65.7% of change in BP when fish oil diet is given to men while there is a decrease of overall 11.4% when regular oil is used with a standard deviation of (5.86 to 3.18)

By these basic summary statistics an inference can be begin that fish oil diet gave good results on men rather than regular oil.

On this note, lets make a boxplot picturing all the summary statistics

```
> boxplot(ex0112$BP ~ ex0112$Diet, xlab = "Diet Oil", ylab =
"Blood Pressure",
+         main = "plot of Diet Vs BP")
```



The box plot shows that there is a lot of difference between the change in blood pressures with fish oil diet and regular oil diet, as seen from summary statistics. Most of the BP change by regular oil data falls within the first quartile of fish oil BP change. While the fish oil diet data has high variations in blood pressure, from which it is clearly evident that the fish oil diet lowered BP rather than regular oil diet.

The below code gives the 95% confidence interval on t test for making a statistical conclusion

```
>           t.test(ex0112$BP[ex0112$Diet           =="FishOil"],
ex0112$BP[ex0112$Diet == "RegularOil"],
+         var.equal = TRUE)

     Two Sample t-test

data:          ex0112$BP[ex0112$Diet     ==     "FishOil"]     and
ex0112$BP[ex0112$Diet == "RegularOil"]
t = 3.0621, df = 12, p-value = 0.009861
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
  2.225174 13.203398
sample estimates:
mean of x mean of y
 6.571429 -1.142857
```

The above gives the standard deviation for the fish oil and regular oil diet change in BP with a strong evidence (p value = 0.009861 from a two-sample t-test). The mean reduction in BP of men on fish oil diet is 7.71 mmHg more than the mean reduction in BP of men on regular oil diet with a confidence interval of 2.26 mmHg to 13.20 mmHg.

The below code gives the p value on the study when one-sided t test is performed.

```
>           t.test(ex0112$BP[ex0112$Diet            =="FishOil"],
ex0112$BP[ex0112$Diet == "RegularOil"],
+         var.equal = TRUE, alt = "great")

     Two Sample t-test

data:          ex0112$BP[ex0112$Diet     ==     "FishOil"]     and
ex0112$BP[ex0112$Diet == "RegularOil"]
t = 3.0621, df = 12, p-value = 0.004931
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 3.224145      Inf
sample estimates:
mean of x mean of y
 6.571429 -1.142857
```

While in first t test, the alternative is taken as default with “two-sided” while an alternative “great” is used in this test to find a one-sided p value. Hence, the one-sided p value will be half of the p value in default case. The p value, here also gives a strong evidence supporting the above made statistical conclusion.

16.

The motivation and creativity case in chapter one is a study to infer whether or not the rewards, praises promote a professional (or academical) improvement. The statistical conclusion in 1.1.1 says that there is a strong evidence that “intrinsic” group has received high scores than “extrinsic” group.

Lets begin the study by reading the data and summarizing the statistics in it.

```
> #16
> head(case0101)
  Score Treatment
1   5.0 Extrinsic
2   5.4 Extrinsic
3   6.1 Extrinsic
4  10.9 Extrinsic
5  11.8 Extrinsic
```

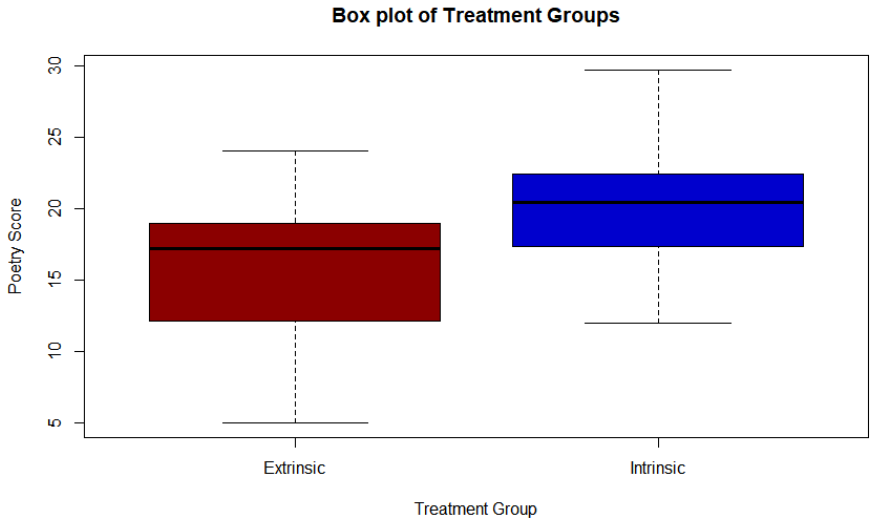
```
6  12.0 Extrinsic
```

The below are the summary of the data where the means of the two groups are fairly close to make a statistical conclusion. We have 23 extrinsic samples and 24 intrinsic samples of which the spread of both the groups is fairly equal.

```
> tapply(case0101$Score, case0101$Treatment, summary)
$Extrinsic
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   12.15   17.20   15.74   18.95   24.00

$Intrinsic
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   17.43   20.40   19.88   22.30   29.70
> tapply(case0101$Score, case0101$Treatment, sd)
Extrinsic Intrinsic
 5.252596  4.439513
```

The same are depicted in the boxplot by below code

```
>   boxplot(case0101$Score   ~   case0101$Treatment,      col   =
c("red4","blue3"),
+         main = "Box plot of Treatment Groups", xlab = "Treatment
Group",
+          ylab = "Poetry Score")
```



Like in summary statistics, box plot also show that both the groups are very competitive in making any statistical conclusion. So, there is a need for t test to find an evidence for this study. Hence imposing t-test on the data

Doing much more than asked for.

```
> t.test(Score ~ Treatment, data = case0101, var.equal = TRUE)

     Two Sample t-test

data:  Score by Treatment
t = -2.9259, df = 45, p-value = 0.005366
alternative hypothesis: true difference in means between group
Extrinsic and group Intrinsic is not equal to 0
95 percent confidence interval:
 -6.996973 -1.291432
sample estimates:
mean in group Extrinsic mean in group Intrinsic
               15.73913                 19.88333
```

As said in the statements of statistical conclusion in 1.1.1, the t test gives a strong statistical evidence that intrinsic group has scored better than extrinsic in the experiment with a p value of 0.005 from a two-sample t test. There is a mean difference of 4.1 points in score between the two groups; though the experiment was done randomly.

As the case started with extrinsic motivation is strong than intrinsic motivation we got a negative confidence interval, which shows that the test results fall under the 95% confidence interval of 1.29 and 6.99

22.

This problem is to find if there is any intelligence difference on gender. A test is made on a sample with respect to different quotients on mind to know the capability. The samples were tested for arithmetic, word knowledge, comprehension and mathematical knowledge. AFQT gives an overall score depicting all these four terms.

Let's head to the data ex0222

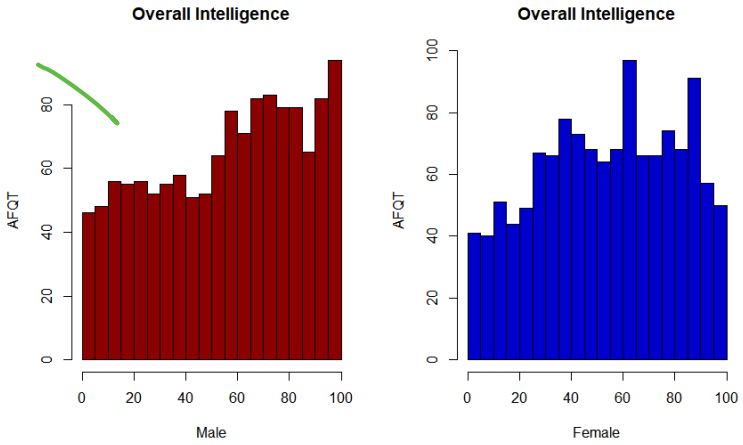
```
> #22
> head(ex0222)
  Gender Arith Word Parag Math AFQT
1   male    19   27    14   14 70.3
2 female    23   34    11   20 60.4
3   male    30   35    14   25 98.3
4 female    30   35    13   21 84.7
5 female    13   30    11   12 44.5
6 female     8   15     6    4  4.0
> table(ex0222$Gender)

female   male
```

```
  1278   1306
```

From the table, there are a total of 2584 test samples of which the test is made on 1278 female and 1306 male. These large sample size supports the conclusions strongly.

Let's go by analyzing the data in each component with respect to gender

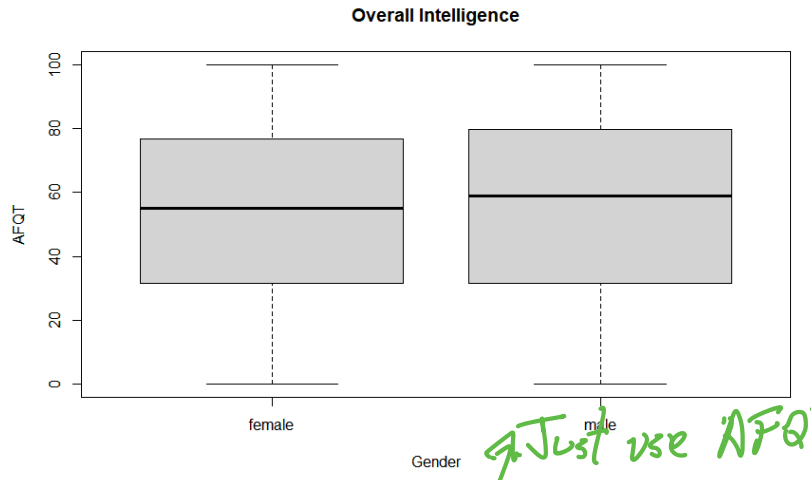
```
> #AFQT Analysis
> par(mfrow = c(1,2))
> hist(ex0222$AFQT[ex0222$Gender == "male"], xlab = "Male", ylab
= "AFQT",
+      main = "Overall Intelligence", col = "red4", breaks = 20)
> hist(ex0222$AFQT[ex0222$Gender == "female"], xlab = "Female",
ylab = "AFQT",
+      main = "Overall Intelligence", col = "blue3", breaks = 20)
> par(mfrow = c(1,1))
```



The above picture gives the histogram of different gender groups with respect to AFQT.

```
> boxplot(ex0222$AFQT ~ ex0222$Gender, xlab = "Gender", ylab =
"AFQT", main = "Overall Intelligence")
> tapply(ex0222$AFQT, ex0222$Gender, summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   31.55   54.90   53.41   76.60  100.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   31.43   59.00   55.45   79.78  100.00
```

The above summary statistics shows that both the groups are very close to make any conclusion. Both the groups spread symmetrically with a close mean. The data till first quartile is very close while a little variation has started to a mean difference of 2.04.



The box plot supports the summaries with symmetrical data distribution and close means and quartiles. The data is okay to use t test.

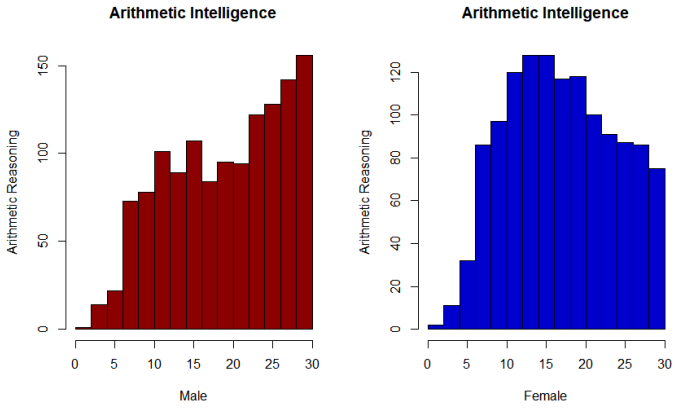
```
>           t.test(ex0222$AFQT[ex0222$Gender          =="male"],
ex0222$AFQT[ex0222$Gender ==="female"],
+           var.equal = TRUE)
```

```
        Two Sample t-test

data:          ex0222$AFQT[ex0222$Gender     ==     "male"]     and
ex0222$AFQT[ex0222$Gender == "female"]
t = 1.8689, df = 2582, p-value = 0.06175
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.1004044  4.1813200
sample estimates:
mean of x mean of y
 55.44625  53.40579
```

The data provide suggestive but inclusive (fair but not strong) evidence that the distribution of AFQT score for the two groups males and females differ with a two-sided p value of 0.062. There is a difference of 2.04% between male and female groups with a 95% confidence interval of - 0.10 and 4.18 points.

When the same analysis is repeated to Arithmetic reasoning component,

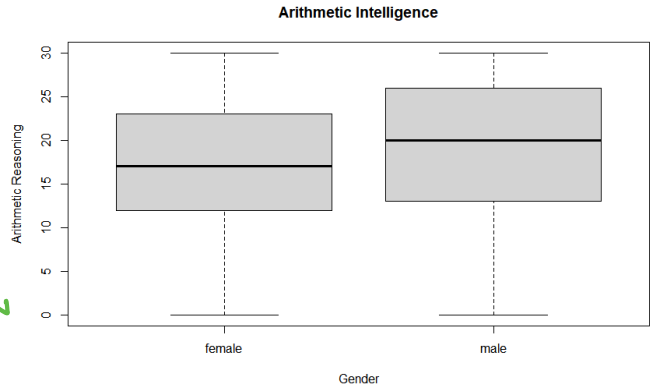
```
> #Arithmetic Reasoning analysis
> par(mfrow = c(1,2))
> hist(ex0222$Arith[ex0222$Gender == "male"], xlab = "Male", ylab
= "Arithmetic Reasoning",
+       main = "Arithmetic Intelligence", col = "red4", breaks =
15)
> hist(ex0222$Arith[ex0222$Gender == "female"], xlab = "Female",
ylab = "Arithmetic Reasoning",
+       main = "Arithmetic Intelligence", col = "blue3", breaks =
15)
> par(mfrow = c(1,1))
```



The above picture gives the histogram of different gender groups with respect to Arithmetic reasoning component. How similar?

```
> boxplot(ex0222$Arith ~ ex0222$Gender, xlab = "Gender", ylab =
"Arithmetic Reasoning",
+         main = "Arithmetic Intelligence")
> tapply(ex0222$Arith, ex0222$Gender, summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.00   17.00   17.49   23.00   30.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   13.00   20.00   19.52   26.00   30.00

> tapply(ex0222$Arith, ex0222$Gender, sd)
  female     male
6.808906 7.346686
```

The above summary statistics shows that both the groups are fairly close. Both the groups spread symmetrically with a close mean. The data till first quartile is very close while a little variation has started to a mean difference of 0.54.



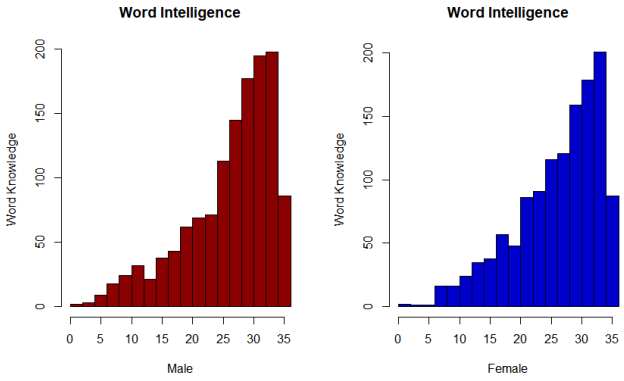
The box plot supports the summaries with symmetrical data distribution and close means and quartiles. The data is okay to use t test.

```
>          t.test(ex0222$Arith[ex0222$Gender          =="male"],
ex0222$Arith[ex0222$Gender ==="female"],
+         var.equal = TRUE)

      Two Sample t-test

data:          ex0222$Arith[ex0222$Gender     ==     "male"]     and
ex0222$Arith[ex0222$Gender == "female"]
t = 7.3064, df = 2582, p-value = 3.639e-13
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.490353 2.583758
sample estimates:
mean of x mean of y
 19.52297  17.48592
```

The data provide strong convincing evidence that the distribution of arithmetic reasoning score for the two groups males and females differ with a two-sided p value of 3.64*e^-13. There is a difference of 2.04% between male and female groups with a 95% confidence interval of 1.49 to 2.58 points.

Proceeding the same analysis process with word knowledge component,

```
> #Word knowledge analysis
> par(mfrow = c(1,2))
> hist(ex0222$Word[ex0222$Gender == "male"], xlab = "Male", ylab
= "Word Knowledge",
+      main = "Word Intelligence", col = "red4", breaks = 15)
> hist(ex0222$Word[ex0222$Gender == "female"], xlab = "Female",
ylab = "Word Knowledge",
+      main = "Word Intelligence", col = "blue3", breaks = 15)
> par(mfrow = c(1,1))
```

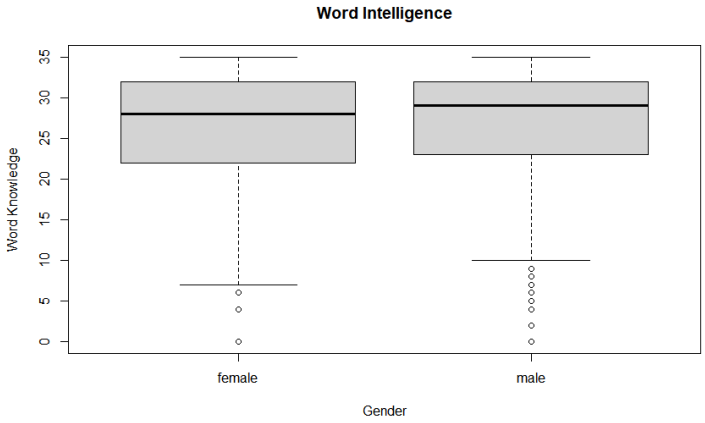


Similar? Diff.

The above picture gives the histogram of different gender groups with respect to word knowledge.

```
> boxplot(ex0222$Word ~ ex0222$Gender, xlab = "Gender", ylab =
"Word Knowledge",
+         main = "Word Intelligence")
> tapply(ex0222$Word, ex0222$Gender, summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   22.00   28.00   26.57   32.00   35.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   23.00   29.00   26.55   32.00   35.00

> tapply(ex0222$Word, ex0222$Gender, sd)
  female     male
6.915509 7.176577
```

The above summary statistics shows that both the groups are fairly close to make any conclusion. Both the groups spread symmetrically with a close mean. The data till first quartile is very close while a very little variation in mean with 0.02.



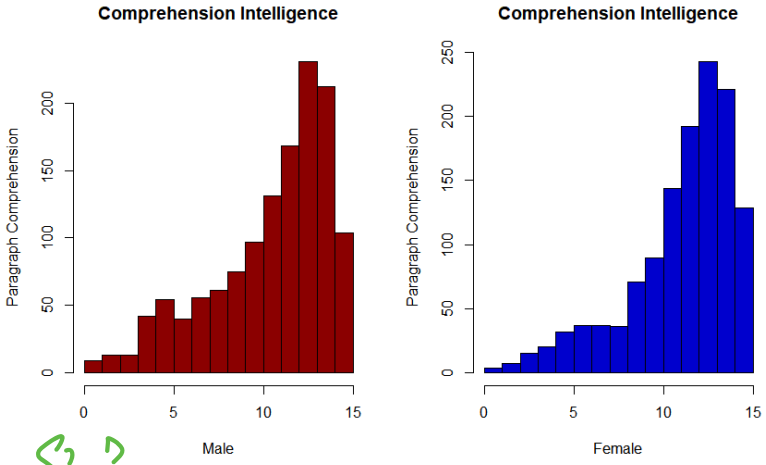
The box plot supports the summaries with symmetrical data distribution and close means and quartiles. The data is okay to use t test.

```
> t.test(ex0222$W[ex0222$Gender =="male"], ex0222$W[ex0222$Gender
=="female"],
+         var.equal = TRUE)

      Two Sample t-test

data:           ex0222$W[ex0222$Gender      ==      "male"]      and
ex0222$W[ex0222$Gender == "female"]
t = -0.079805, df = 2582, p-value = 0.9364
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.5659693  0.5217027
sample estimates:
mean of x mean of y
 26.54594  26.56808
```

The data provides no evidence of gender difference. The distribution of word knowledge scores for the two groups males and females differ with a two-sided p value of 0.94. There is a difference of 0.02% between male and female groups with a 95% confidence interval of – 0.57 to 0.52 points.

Paragraph comprehension analysis,

```
> #Paragraph comprehension analysis
> par(mfrow = c(1,2))
> hist(ex0222$Parag[ex0222$Gender == "male"], xlab = "Male", ylab
= "Paragraph Comprehension",
+      main = "Comprehension Intelligence", col = "red4", breaks =
15)
> hist(ex0222$P[ex0222$Gender == "female"], xlab = "Female", ylab
= "Paragraph Comprehension",
+      main = "Comprehension Intelligence", col = "blue3", breaks
= 15)
> par(mfrow = c(1,1))
```

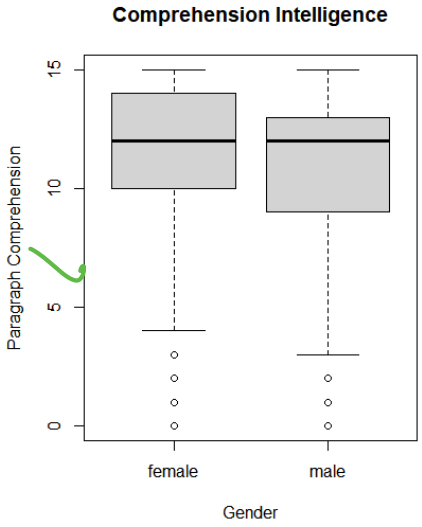


The above picture gives the histogram of different gender groups with respect to word knowledge.

```
> boxplot(ex0222$P ~ ex0222$Gender, xlab = "Gender", ylab =
"Paragraph Comprehension",
+         main = "Comprehension Intelligence")
> tapply(ex0222$P, ex0222$Gender, summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   10.00   12.00   11.49   14.00   15.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   12.00   10.92   13.00   15.00

> tapply(ex0222$P, ex0222$Gender, sd)
  female     male
```

```
2.967768 3.315783
```

The above summary statistics shows that both the groups are fairly close to make any conclusion. Both of the groups have few outliers with equal median. Both the groups spread more or less symmetrically. The data till first quartile is very close while a very little variation in mean with 0.05.



The data is okay to use t test.

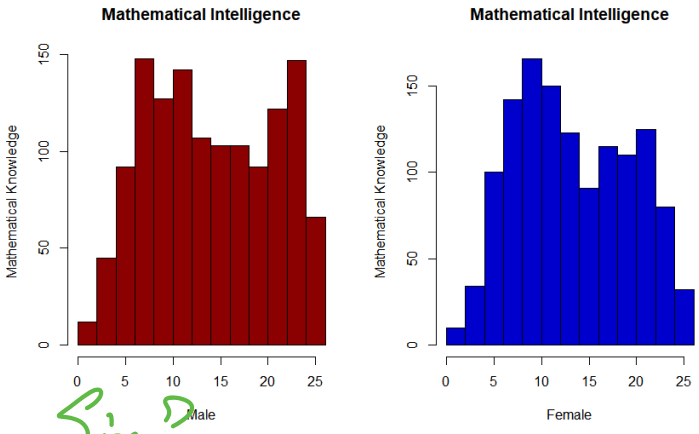
```
> t.test(ex0222$P[ex0222$Gender =="male"], ex0222$P[ex0222$Gender
=="female"],
+         var.equal = TRUE)

     Two Sample t-test

data:           ex0222$P[ex0222$Gender      ==      "male"]      and
ex0222$P[ex0222$Gender == "female"]
t = -4.5968, df = 2582, p-value = 4.497e-06
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.8123791 -0.3265415
sample estimates:
mean of x mean of y
 10.92037  11.48983
```

The data provide strong convincing evidence that the distribution of paragraph comprehension score for the two groups males and females differ with a two-sided p value of 0.0000045. There is a difference of 0.57% between male and female groups. The female mean exceeds the male mean by 0.57 percent points with a confidence interval of 0.32 to 0.81 points.

Some minor points, but overall very well done!

Mathematical knowledge analysis,

```
> #Mathematical knowledge
> par(mfrow = c(1,2))
> hist(ex0222$M[ex0222$Gender == "male"], xlab = "Male", ylab =
"Mathematical Knowledge",
+      main = "Mathematical Intelligence", col = "red4", breaks =
15)
> hist(ex0222$M[ex0222$Gender == "female"], xlab = "Female", ylab
= "Mathematical Knowledge",
+      main = "Mathematical Intelligence", col = "blue3", breaks =
15)
> par(mfrow = c(1,1))
```

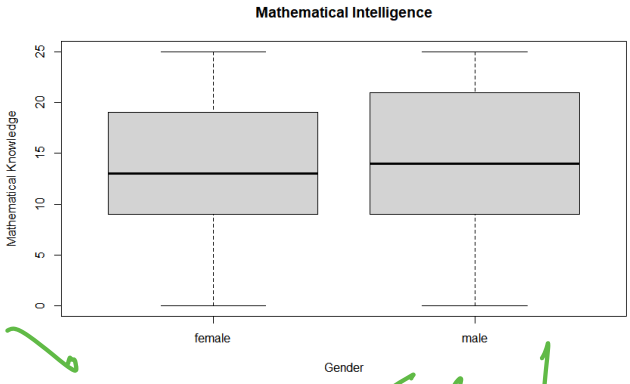


The above picture gives the histogram of different gender groups with respect to mathematical knowledge component

```
> boxplot(ex0222$M ~ ex0222$Gender, xlab = "Gender", ylab =
"Mathematical Knowledge",
+         main = "Mathematical Intelligence")
> tapply(ex0222$M, ex0222$Gender, summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   13.00   13.82   19.00   25.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   14.00   14.57   21.00   25.00

> tapply(ex0222$M, ex0222$Gender, sd)
```

```
  female     male
6.013065 6.519207
```

The above summary statistics shows that both the groups are fairly close to make any conclusion. Both of the groups have no outliers with almost equal mean. Both the groups spread more symmetrically. The data till first quartile is very close while a very little variation in mean with 0.51.



The data is okay to use t test.

```
> t.test(ex0222$M[ex0222$Gender =="male"], ex0222$M[ex0222$Gender
=="female"],
+         var.equal = TRUE)

     Two Sample t-test

data:           ex0222$M[ex0222$Gender      ==      "male"]      and
ex0222$M[ex0222$Gender == "female"]
t = 3.0464, df = 2582, p-value = 0.002339
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.267979 1.236111
sample estimates:
mean of x mean of y
 14.56738  13.81534
```

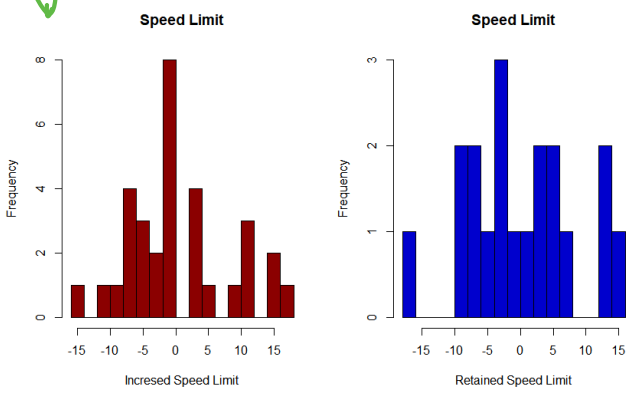
The data provide strong convincing evidence that the distribution of paragraph comprehension score for the two groups males and females differ with a two-sided p value of 0.002. There is a difference of 0.75% between male and female groups. The female mean exceeds the male mean by 0.75 percent points with a confidence interval of 0.27 to 1.24 points.

23. The exercise shows the data conducted to see the percentage change in speed limits in the states when a federal act was proposed giving states to decide the state speed limit for roads.
This study is to find a statistical evidence if there is any greater percent change in states that increases the speed limits from 1995 to 1996.
Heading to the exercise data,

```
> #23
> head(ex0223)
       State Fatalities1995 Fatalities1996 PctChange SpeedLimit
1    Alabama           1114           1146      2.87        Inc
2     Alaska             87             81     -6.90        Ret
3    Arizona           1035            994     -3.96        Inc
4   Arkansas            631            615     -2.54        Inc
5 California           4192           3989     -4.84        Inc
6   Colorado            645            617     -4.34        Inc
> table(ex0223$SpeedLimit)

Inc Ret
 32  19
```

The data shows the percent change, whether increased or retained the speed limit for respective states.
It can be seen that out of 51 states, 32 has increased the speed limit while 19 states retained it.
The histogram can be drawn for the data by below code

```
> par(mfrow = c(1,2))
>  hist(ex0223$PctChange[ex0223$SpeedLimit  ==  "Inc"],  xlab  =
"Incresed Speed Limit",
+        main = "Speed Limit", col = "red4", breaks = 15)
>  hist(ex0223$PctChange[ex0223$SpeedLimit  ==  "Ret"],  xlab  =
"Retained Speed Limit",
+      main = "Speed Limit", col = "blue3", breaks = 15)
> par(mfrow = c(1,1))
```

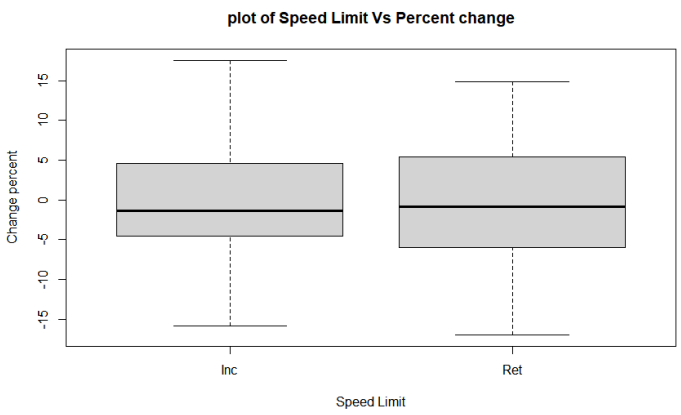


The box plot gives a big picture of the data points, it can be drawn by below code

```
> boxplot(ex0223$PctChange ~ ex0223$SpeedLimit, xlab = "Speed
Limit", ylab = "Change percent",
+         main = "plot of Speed Limit Vs Percent change")
> tapply(ex0223$PctChange, ex0223$SpeedLimit, summary)
$Inc
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-15.8800  -4.4650  -1.3800   0.4938   4.0675  17.5600

$Ret
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-16.98000  -6.01000  -0.82000   0.09947   5.36500  14.86000

> tapply(ex0223$PctChange, ex0223$SpeedLimit, sd)
     Inc      Ret
8.095122 8.557060
```

The summary statistics give that there is nearly a fair spread of speed limit data. The means have little difference with standard deviation intervals of 8.095 and 8.557

Lets have a look at the box plot to compare the percentage change in states which increased their speed limits with respect to those which retained it.



The box plot shows that there is no much percent change in states which increased their speed limits and those retained their speed limits. Lets do a hypothesis test to see if there any evidence that percentage change is greater tin states which increased the speed limits.

Performing a one-sided t-test,

```
> t.test(PctChange ~ SpeedLimit, data = ex0223, var.equal = TRUE,
alt= "great")

     Two Sample t-test

data:  PctChange by SpeedLimit
t = 0.16466, df = 49, p-value = 0.4349
alternative hypothesis: true difference in means between group
Inc and group Ret is greater than 0
95 percent confidence interval:
 -3.620309       Inf
sample estimates:
mean in group Inc mean in group Ret
       0.49375000        0.09947368
```

From the above we can say that there is no evidence that mean percent increase is high in states that increased the speed with the ones which retained it. This no evidence can be supported with a p value of 0.44 with mean percent increase in the states which increased the speed limit to exceed by 0.39 (mean difference) with the states which retained the speed limits.

Lets perform a two sided t-test to find the confidence interval of the difference.

```
> t.test(PctChange ~ SpeedLimit, data = ex0223, var.equal = TRUE)

     Two Sample t-test

data:  PctChange by SpeedLimit
t = 0.16466, df = 49, p-value = 0.8699
```

```
alternative hypothesis: true difference in means between group
Inc and group Ret is not equal to 0
95 percent confidence interval:
 -4.417753  5.206305
sample estimates:
mean in group Inc mean in group Ret
       0.49375000        0.09947368
```

Supporting one-sided t-test, the above test results also give that there is no evidence that the mean percent increase is higher in states that increased their speed with those of retained ones with a p value of 0.87. There is 0.39 percent difference of means in both the groups with confidence interval of -4.4 to 5.2 percent.

Scope of inference

Since the samples of this study were real and not random, the inference to population cannot be done. Inference may be applied to actual states only. If any statistical statements are made based on this study, the population could get biased, producing in wrong interpretations and this study is strongly not applicable to any population.

Good. Wording on tests could be cleaner though.

Please keep a { code, output, comment, code, output, comment, ... } pattern to your responses to data problems.