IST772– Problem Set 4

Sathish Kumar Rajendiran

Attribution statement: 1. I did this homework by myself, with help from the book and the professor.

Chapter 4, Exercise 7

_The built-in warpbreaks data set contains data for the number of warp breaks per loom with different amounts of tension (we will not consider the variable for the type of wool). The tensions are labelled "L", "M" or "H" for low, medium and high tension. Run the summary() command on warpbreaks and explain the output. Create a histogram of the breaks for low tension (1 pt). As a reminder about R syntax, here is one way that you can access the low tension data:Also create histograms of the breaks for medium and high tensions. What can you say about the differences in the effects of tension by looking at the histograms? (1 pt)___

1) Summary(): warpbreaks

- Total number of observations: 54
- Number of variables : 3
  - [,1] breaks numeric The number of breaks
  - [,2] wool factor The type of wool (A or B)
  - [,3] tension factor The level of tension (L, M, H)
- Summary() breaks wool tension Min. :10.00 A:27 L:18
  1st Qu.:18.25 B:27 M:18
  Median :26.00 H:18
  Mean :28.15
  3rd Qu.:34.00
  Max. :70.00
  From the summary above,
  - There two types of "Wool" A and B equally distributed with 27 observations each. Dataype is "Factor"
  - There three types of "tension" L = low; H = High; M= medium; with equal distribution of 18 observations each. Dataype is "Factor"
  - breaks - numeric value with a mean value of 28.15; median of 26; Min value 10, max value 70 , 1st quartile = 18.25 and 3rd quartile as 34

```
# ?warpbreaks
# This data set gives the number of warp breaks per loom, where a loom
# corresponds to a fixed length of yarn.
# dim(warpbreaks)
# 1. Summary() of warpbreaks dataset
summary(warpbreaks)
```

```r
str(warpbreaks)
View(warpbreaks)
summary(warpbreaks$breaks[warpbreaks$tension=="L"])
```

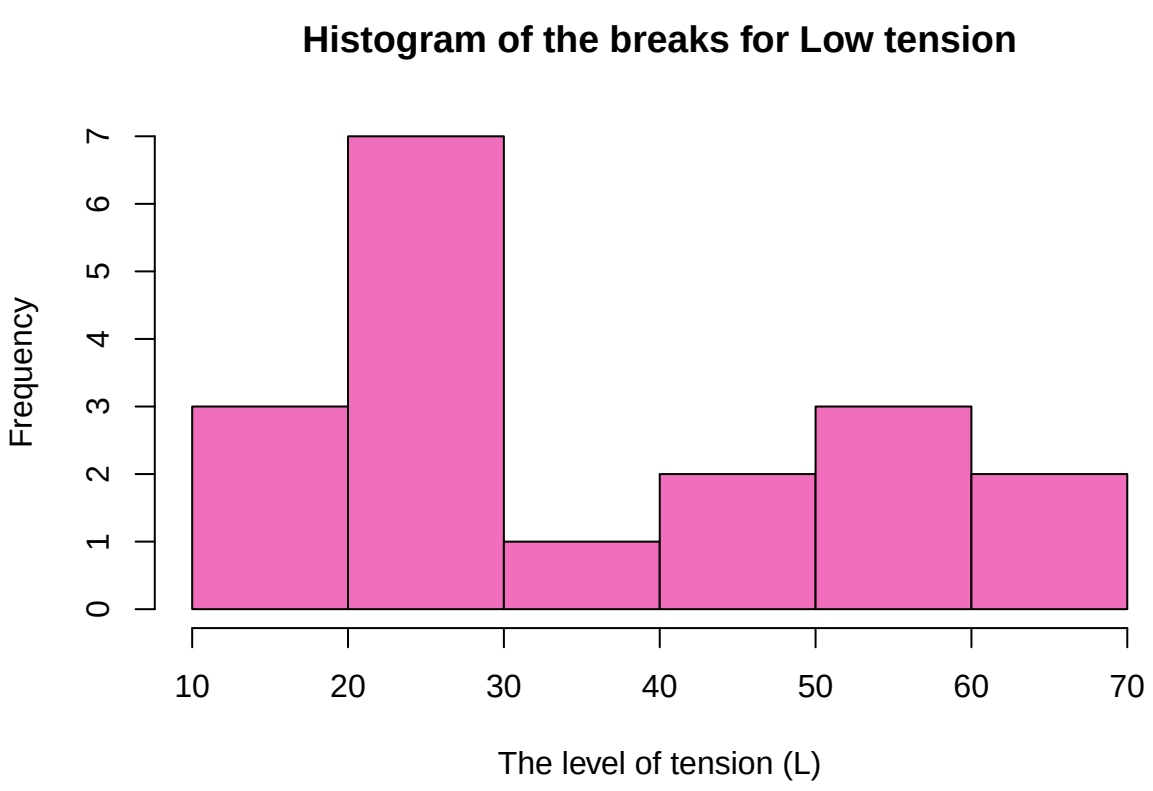
```r
#1. histogram of the breaks for low tension

cat("Summary of breaks for Low tension:\n")
```

```
## Summary of breaks for Low tension:
```

```r
summary(warpbreaks$breaks[warpbreaks$tension=="L"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   26.00   29.50   36.39   49.25   70.00
```

```r
hist(warpbreaks$breaks[warpbreaks$tension=="L"]
     ,main="Histogram of the breaks for Low tension"
     ,xlab="The level of tension (L)"
     ,col="#F06EBB")
```



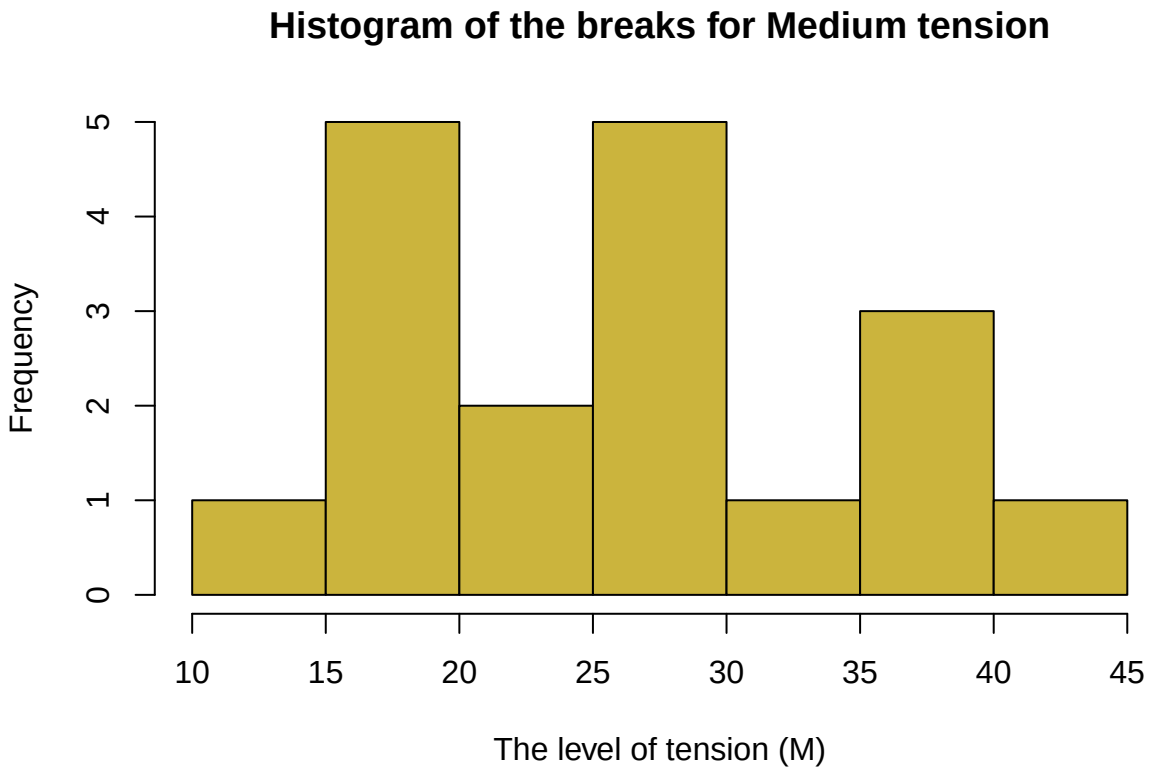

```r
cat("summary of breaks for Medium tension:\n")
```

```
## summary of breaks for Medium tension:
```

```r
summary(warpbreaks$breaks[warpbreaks$tension=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   18.25   27.00   26.39   33.75   42.00
```

```r
#histogram of the breaks for low tension
hist(warpbreaks$breaks[warpbreaks$tension=="M"]
     ,main="Histogram of the breaks for Medium tension"
     ,xlab="The level of tension (M)"
     ,col="#CBB43D")
```



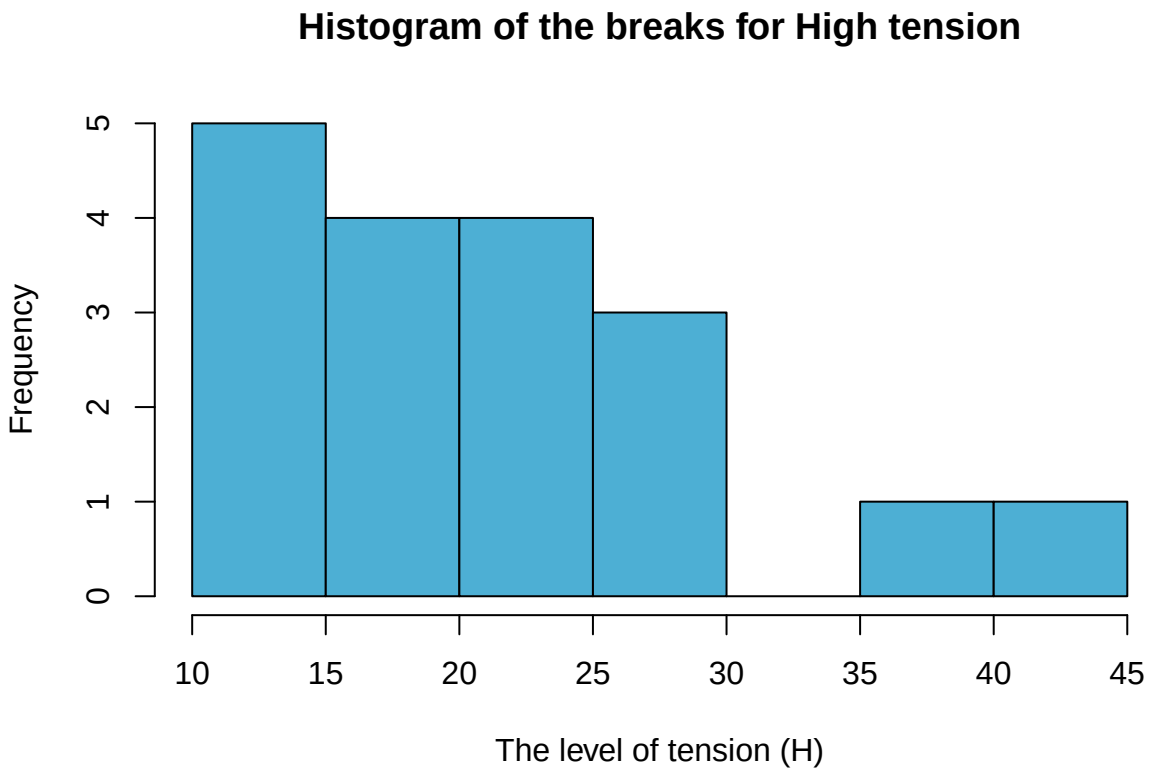

```r
#histogram of the breaks for low tension
cat("summary of breaks for High tension:\n")
```

```
## summary of breaks for High tension:
```

```r
summary(warpbreaks$breaks[warpbreaks$tension=="H"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   15.25   20.50   21.67   25.50   43.00
```

```r
hist(warpbreaks$breaks[warpbreaks$tension=="H"]
     ,main="Histogram of the breaks for High tension"
     ,xlab="The level of tension (H)"
     ,col="#4DAFD4")
```



Effects of tension by looking at the histograms? Differnces among groups:

- Summary of breaks for Low tension: We can see that "low" tension doesnt really have an impact as the "breaks" values are spread across 14 to 70; Average value is 36.9 there is one spike between the ranges 20 to 30; Otherwise, looks like a uniform distribution across.
  - Min. 1st Qu. Median Mean 3rd Qu. Max.
  - 14.00 26.00 29.50 36.39 49.25 70.00
- summary of breaks for Medium tension: We can see that "medium" tension does have an impact as the "breaks" values are spread across 14 to 42 with Average value is 26.39 However, we still see inconsistent spikes at 3 places.Median and Mean almost merge having possible normal distribution having both low and high tails symmetrically placed on both ends.
  - Min. 1st Qu. Median Mean 3rd Qu. Max.
  - 12.00 18.25 27.00 26.39 33.75 42.00
- summary of breaks for High tension: We can see that "high" tension does have an impact heavily densed pouplation between Min to 3rd quartile.However beyond that there isnt any between 30 to 35 and low values between 35 to 45. Since the mean is reduced, its possible that some of the observations beyond 3rd quartile may have outliers.Shape may look like right skewed curve.
  - Min. 1st Qu. Median Mean 3rd Qu. Max.
  - 10.00 15.25 20.50 21.67 25.50 43.00

Using the dplyr package, you can instead write: (Note that a select function is defined in multiple packages, so if you want to be sure you're using the one from the dplyr library, call dplyr::select.)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
warpbreaks %>% filter(tension == "L") %>% dplyr::select(breaks)
```

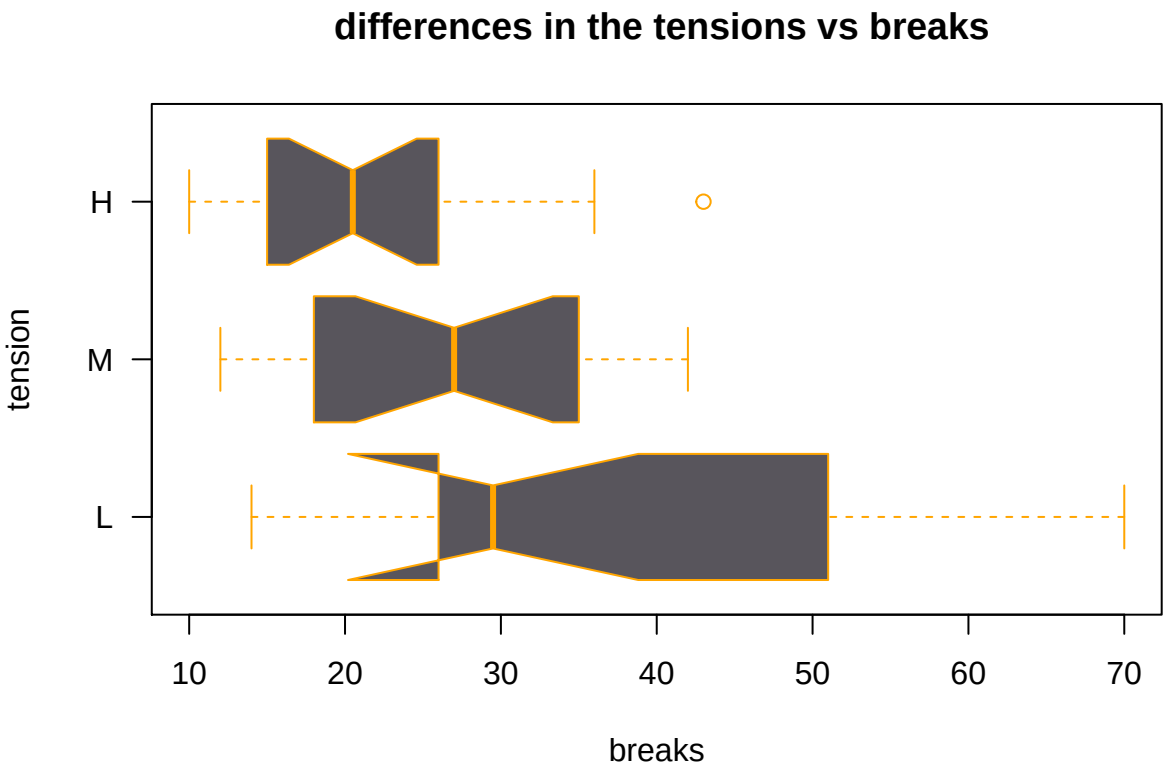
```
##    breaks
## 1      26
## 2      30
## 3      54
## 4      25
## 5      70
## 6      52
## 7      51
## 8      26
## 9      67
## 10     27
## 11     14
## 12     29
## 13     19
## 14     29
## 15     31
## 16     41
## 17     20
## 18     44
```

Chapter 4, Exercise 8

Create a boxplot (or violin plot) of the breaks data, using the model "breaks ~ tension". (1 pt) What can you say about the differences in the tensions by looking at the boxplots for the different tensions? (1 pt)

```
#box plot to compare the batteries
# with(batterydata_r, boxplot(Time~Battery))
boxplot(breaks~tension, data=warpbreaks,
       border="orange",
       col="#58555C",
       freq=FALSE,
       las=1,
       breaks=5,
       notch = TRUE,
       horizontal = TRUE ,main=" differences in the tensions vs breaks")
```

```
## Warning in bxp(list(stats = structure(c(14, 26, 29.5, 51, 70, 12, 18, 27, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```



Effects of tension by looking at the histograms? Differnces among groups:

- From the box plot above,
  - Low group doesnt really have an impact as the values are spread across. Minimum value starts at ~14 and max value is at ~70. Median or (50th percentile) is at ~30. 3rd Quartile is recorded at ~50. There are more observations between 50th percentile and 3rd quartile.Values are inconsistent across.
  - Medium group does show an impact as the values are spread evenly across median value. Minimum value starts at ~12 and max value is at ~42. Median or (50th percentile) is at ~27. 3rd Quartile is recorded at ~34. Mean and Median are almost converging on the same value suggesting a normal distribution. No outliers seen.
  - Similarly, High group does have an impact as the values are spread evenly across median value. Minimum value starts ~10 (lowest) and max value is at ~43. Median or (50th percentile) is at ~20.50. 3rd Quartile is recorded at ~25.5 Mean and Median are almost converging on the same. However, there is an outlier recorded almost at ~45. Otherwise, this plot clearly shows the impact of having High tension on the breaks

Chapter 4, Exercise 9

Run a t-test to compare the means of high and medium tension in the warpbreaks data. (1 pt) Report and interpret the confidence interval. (1 pt) Make sure to include a carefully worded statement about what

the confidence interval implies with respect to the population mean difference between the high and medium tensions. (1 pt)

```
#t-test
t.test(warpbreaks$breaks[warpbreaks$tension=="H"],warpbreaks$breaks[warpbreaks$tension=="M"])
```

```
##
##  Welch Two Sample t-test
##
## data:  warpbreaks$breaks[warpbreaks$tension == "H"] and warpbreaks$breaks[warpbreaks$tension == "M"]
## t = -1.6199, df = 33.74, p-value = 0.1146
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.648042   1.203597
## sample estimates:
## mean of x mean of y
##  21.66667  26.38889
```

t-test to compare the means of high and medium tension in the warpbreaks data

- From the t-test above,
  - critical value t* is at -1.6199 with ~34 degrees of freedom and p-value 0.1146
  - Sample mean of "High" tension is at ~21.67
  - Sample mean of "Medium" tension is at ~26.39
  - 95 percent confidence interval is between -10.6480 (lower band) and 1.2035 (higher band).

implication of population mean

- 95 percent confidence interval is between -10.6480 (lower band) and 1.2035 (higher band). It includes the possibility of 0 in between In addition,The size of the space between the lower and upper bounds represents our uncertainity considering the differences of the means of x and y. Mean difference -4.72222 (i.e. 21.66667-26.38889)

Chapter 4, Exercise 10

Run a t-test to compare the means of low and medium tension in the warpbreaks data. (1 pt) Report and interpret the confidence interval. (1 pt + 1 pt for statement about means)

```
#t-test
t.test(warpbreaks$breaks[warpbreaks$tension=="L"],warpbreaks$breaks[warpbreaks$tension=="M"])
```

```
##
##  Welch Two Sample t-test
##
## data:  warpbreaks$breaks[warpbreaks$tension == "L"] and warpbreaks$breaks[warpbreaks$tension == "M"]
## t = 2.256, df = 26.554, p-value = 0.03252
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.8976796 19.1023204
## sample estimates:
## mean of x mean of y
##  36.38889  26.38889
```

t-test to compare the means of low and medium tension in the warpbreaks data

- From the t-test above,
  - critical value t* is at 2.256 with ~27 degrees of freedom and p-value 0.03252
  - Sample mean of "low" tension is at ~36.39
  - Sample mean of "medium" tension is at ~26.39
  - 95 percent confidence interval is between 0.89767 (lower band) and 19.1023 (higher band).

implication of population mean

- 95 percent confidence interval is between 0.89767 (lower band) and 19.1023 (higher band). It excludes the possibility of 0 in between In addition,The size of the space between the lower and upper bounds represents our uncertainity considering the differences of the means of x and y. Mean difference 10 (i.e. 36.38889-26.38889) is at the center of that region of uncertainity.