IST772 Problem Set 4

Abhijith Anil Vamadev

The homework for week 4 is based is based on exercises 7-10 on page 66, but with changes as noted in the text in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 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:

```
warpbreaks <- warpbreaks #assing warbreaks variable
summary(warpbreaks) #summary of the warpbreaks variable
hist(warpbreaks$breaks[warpbreaks$tension=="L"]) #histogram of the L tension
```

- The summary for warpbreaks shows the 3 variables within warpbreaks: breaks, wool and tension. Wool and Tension are categorical variables and breaks is a numerical variable.

Using the dplyr package, you can instead write:

```
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") %>% select(breaks)
```

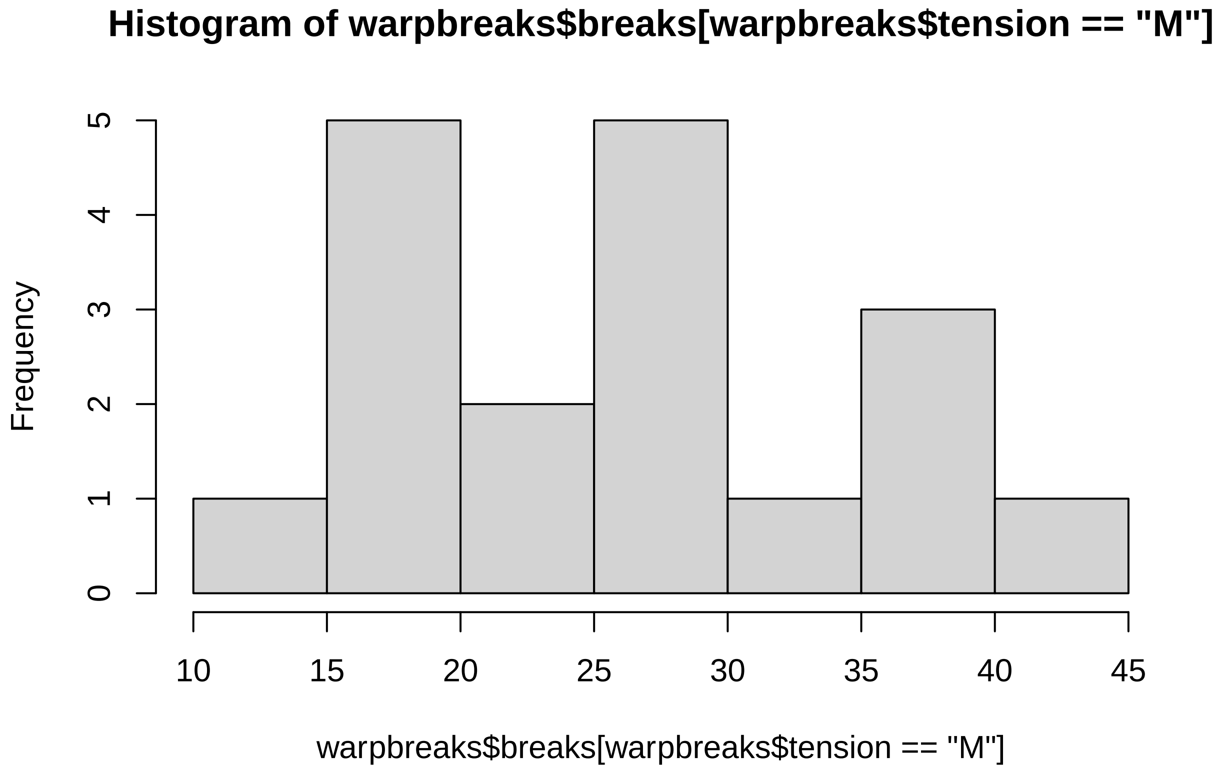
```
##    breaks
## 1      26
## 2       30
## 3       54
## 4       25
## 5       70
## 6       52
## 7       51
## 8       26
## 9       67
## 10      27
## 11      14
## 12      29
## 13      19
## 14      29
## 15      31
## 16      41
## 17      20
## 18      44
```

```
#filter then select using dplyr to find L tension
```

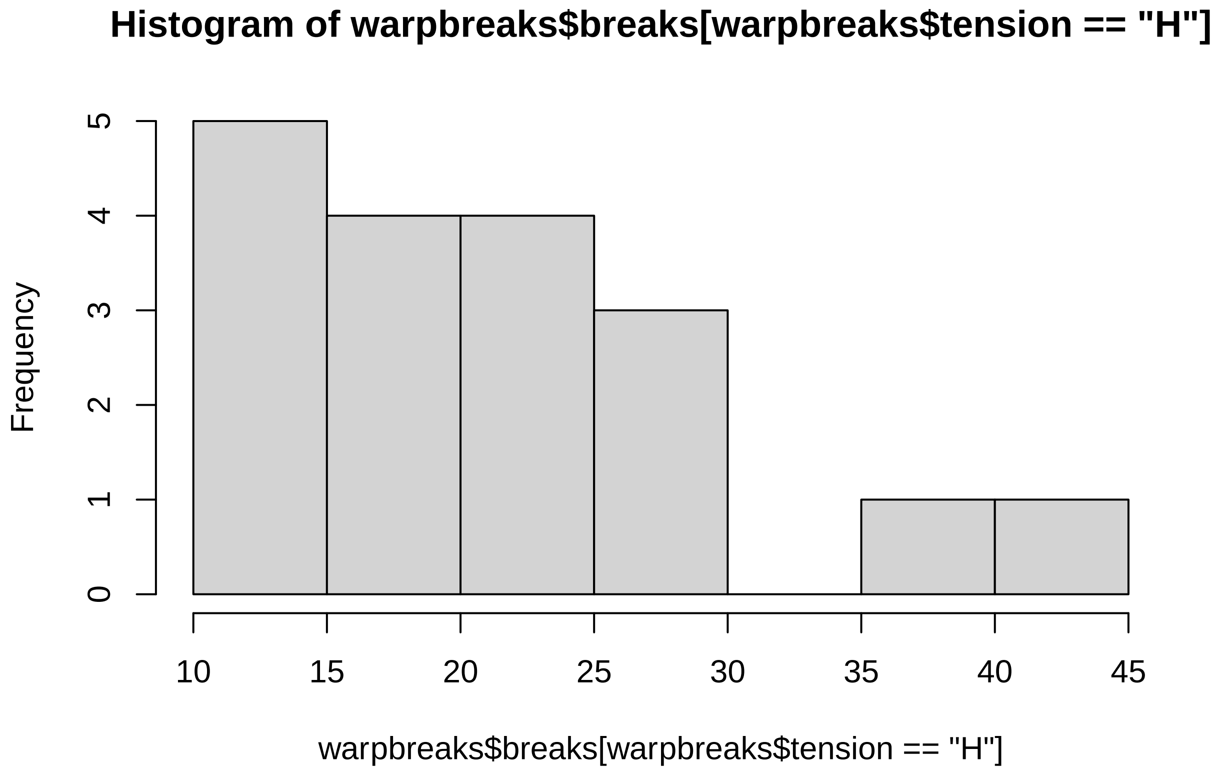
(Note that a select function is defined in multiple packages, so if you want to be sure you're using the one from the dplyr library, call dplyr::select.)

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)

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



```
#histogram of M tension breaks
hist(warpbreaks$breaks[warpbreaks$tension=="H"])
```



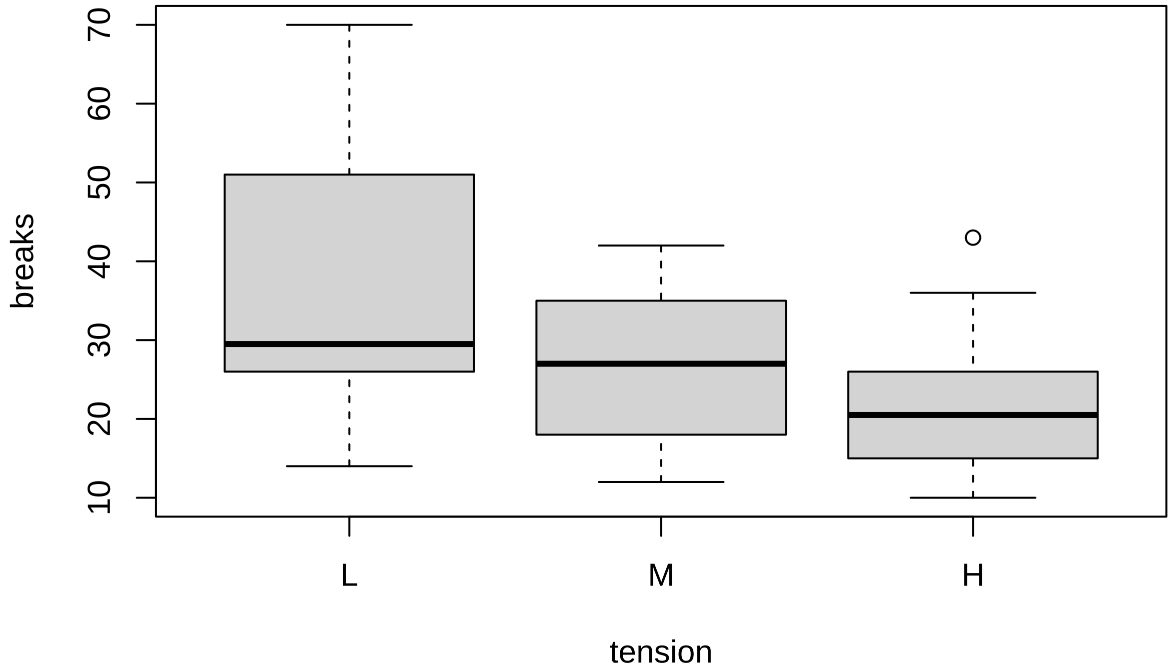
```
#histogram ofo H tension breaks
```

- For the medium tension, it ranges from 10 to 45, with high frequency between 15-20 and between 25 to 30. Whereas for the high tension, the highest frequency mostly ranges from 10 to 30. And 35 to 45 having very low frequency.

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)

```
with(warpbreaks, boxplot(breaks~tension))
```



```
#boxplot of the different tensions
```

- On average L tension has higher breaks, followed by M and H. The max value for L-tension is higher and is followed by M and then H tension. The range of L is the highest compared to M and H. H-tension has an outlier above the maximum value.

Chapter 4, Exercise 9

Run a t-test to compare the means of high and low 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 low tensions. (1 pt)

```
high <- warpbreaks$breaks[warpbreaks$tension=="H"] #assign values to H
medium <- warpbreaks$breaks[warpbreaks$tension=="M"] #assing values to M
low <- warpbreaks$breaks[warpbreaks$tension=="L"] #assign values to L
t.test(high, low) #t-test for H and L
```

```
##
##  Welch Two Sample t-test
##
## data:  high and low
## t = -3.3862, df = 25.222, p-value = 0.002327
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.67256  -5.77188
## sample estimates:
## mean of x mean of y
##  21.66667  36.38889
```

```
mean(high - low) #mean diffeernce of H and L
```

```
## [1] -14.72222
```

- The confidence interval of 95% is within the -23.67 and -5.77. The 95% confidence interval indicates that if the process was to be run 100 times, then 95 times the mean difference would be contained within this confidence interval. The low tension has higher breaks than the high tension by 14.72 breaks per loom.

Chapter 4, Exercise 10

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 + 1 pt for statement about means)

```
t.test(high, medium) #t-test for high and medium
```

```
##
##  Welch Two Sample t-test
##
## data:  high and medium
## 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
```

```
mean(high - medium) #mean difference for H and M.
```

```
## [1] -4.722222
```

- The confidence interval for the t-test between high and medium are -10.68 and 1.203. That means for a process that is conducted 100 times,the confidence interval would contain the true population 95 out of 100 times. 95% interval indicates that, if the process was to repeat 100 times 95 of the test, would contain the mean difference. The medium tension has higher breaks per loom by 4.722 compared to the High tension.