STAT805 Portfolio - Mohammed El Essawy

Introduction

In Autumn 2016, I joined a group of hiking enthusiasts who planned a long camping/hiking trip called the *Red Leaves Loop*. The trip started from Auckland and went towards the South East coasts of the North island, all the way thrrough Tauranga, Rotorua, Gisborne, Napier and finally Wellington. The way back was from the west coast, through Whanganui, New Plymouth and Hamilton.

The objective of the trip was to chase the colour transformation happening to many trees and plants, to become slightly more reddish. It's a known fact that the drop in temperature during Autumn nights causes the Chlorophyll to stop functioning properly, and therefore stop absorbing sunlight as usual, which makes the leaves go red. This doesn't work 100% in New Zealand as it does in colder countries, as the temperature doesn't go very low in Autumn.

I enjoyed every moment of that trip, and to document my findings, I took a couple of leaves' photos every day and also noted the temperature of the day. For each of the locations we camped in, I gave a score from 1 to 5 describing the amount of red shades in the leaves around us, where 1 is dark green, 5 is clear red. The scores 2 to 4 described the middle shades between green and red, like yellow, orange .. etc. My objective of noting these details was to review them later when I return and see what facts can be uncovered. I never did that review, and I think that now is the right chance to do so.

Slight transformation into orange/red

More..

Me with red leaves

Dataset

My dataset has 61 observations (61 days in total). Below I am showing the first 15 rows.

```
library(knitr)
library(kableExtra)
data <- read.csv("data/data.csv")
kable(head(data, 15)) %>% kable_styling(bootstrap_options = "striped", full_width = F, position = "left")
```

date	temprature	red_degree
1/04/2016	16	1
2/04/2016	15	1
3/04/2016	16	2
4/04/2016	17	2
5/04/2016	17	1
6/04/2016	17	2
7/04/2016	16	1
8/04/2016	16	1
9/04/2016	13	1
10/04/2016	16	3
11/04/2016	11	2
12/04/2016	10	3
13/04/2016	12	4
14/04/2016	13	1
15/04/2016	12	2

I will assign each column to a variable:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

```
temp <- data %>% pull(temprature)
red_deg <- data %>% pull(red_degree)
```

Let’s calculate some stats:

```
mean(red_deg)
```

```
## [1] 2.278689
```

```
sd(red_deg)
```

```
## [1] 1.226528
```

```
mean(temp)
```

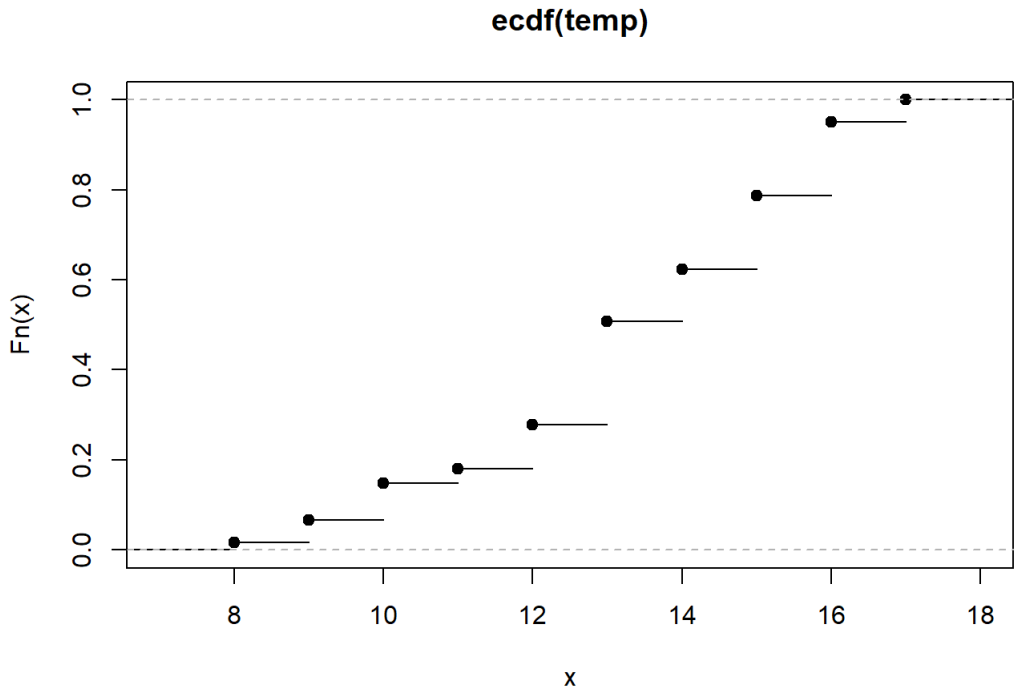
```
## [1] 13.44262
```

```
sd(temp)
```

```
## [1] 2.254807
```

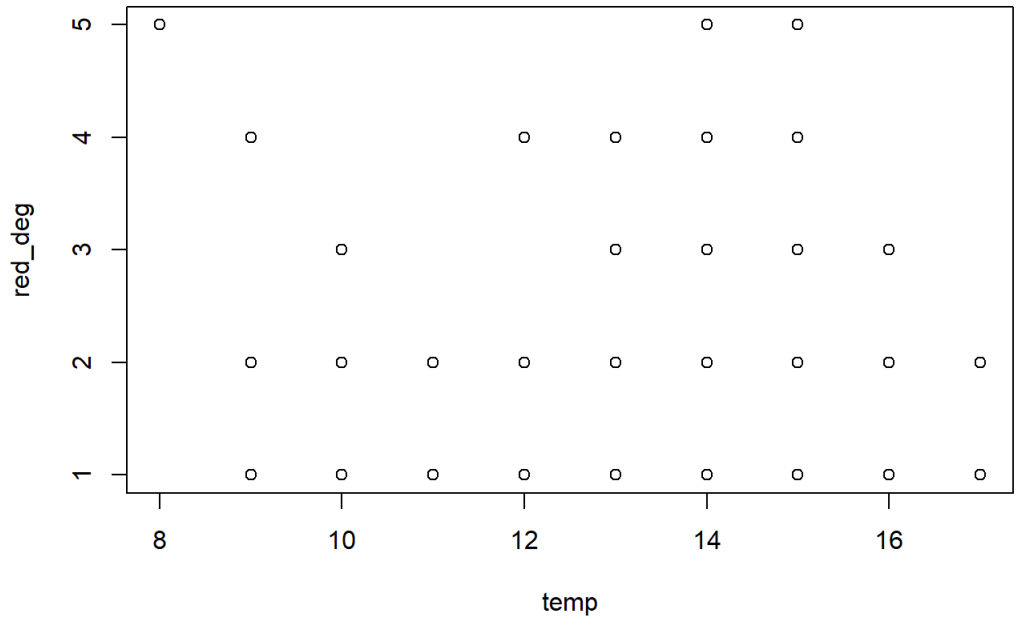
Let’s generate some plots to help understand the dataset. I will start by an ECDF plot of the temprature:

```
plot(ecdf(temp))
```



Around 50% of our dataset observations are representing tempratures between 8 to 14. The other 50% is between 14 and 18. Let's look at the temprature against the red degree:

```
plot(temp, red_deg)
```

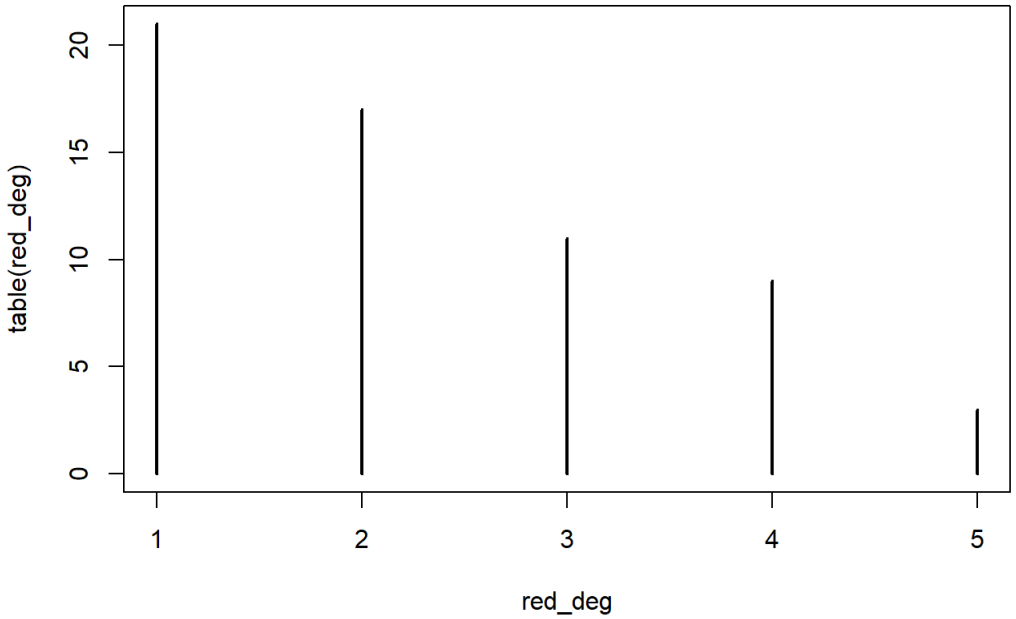


We can see that the data points are scattered and there is no visible relationship between the two variables.
Let's study the frequency of each red degree in the dataset:

```
table(red_deg)
```

```
## red_deg
##  1  2  3  4  5
## 21 17 11  9  3
```

```
plot(table(red_deg))
```



It’s clear that early red degrees are very common, while deep red degrees are very rare (3 observations only out of 61).
Let’s calculate the probability of each red degree:

```
21/61 ##probability of red_degree = 1##
```

```
## [1] 0.3442623
```

```
17/61 ##probability of red_degree = 2##
```

```
## [1] 0.2786885
```

```
11/61 ##probability of red_degree = 3##
```

```
## [1] 0.1803279
```

```
9/61 ##probability of red_degree = 4##
```

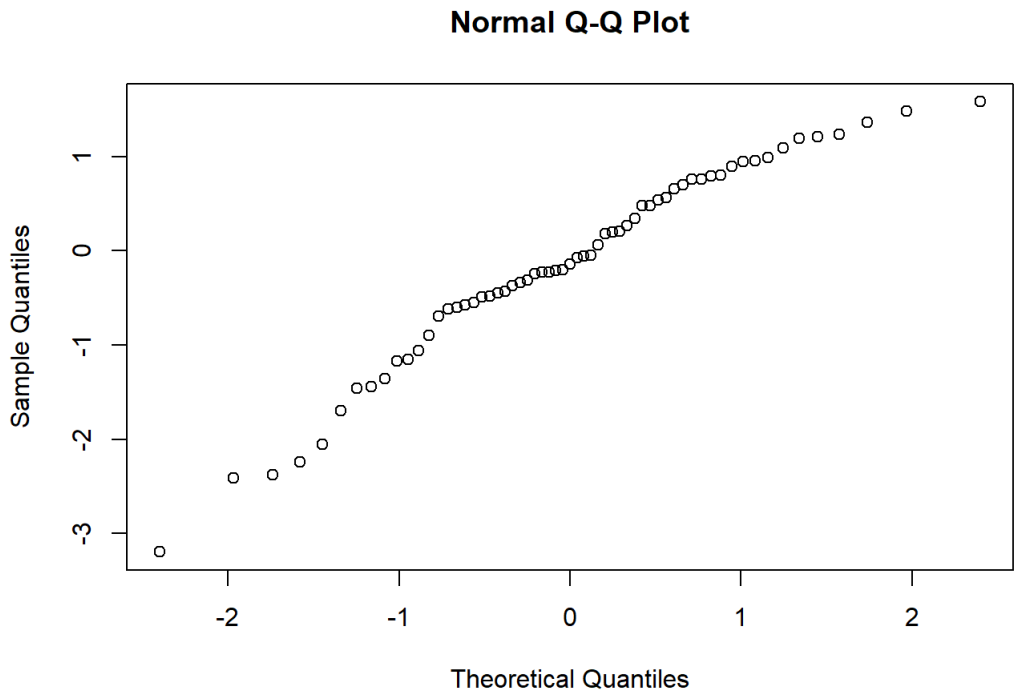
```
## [1] 0.147541
```

```
3/61 ##probability of red_degree = 5##
```

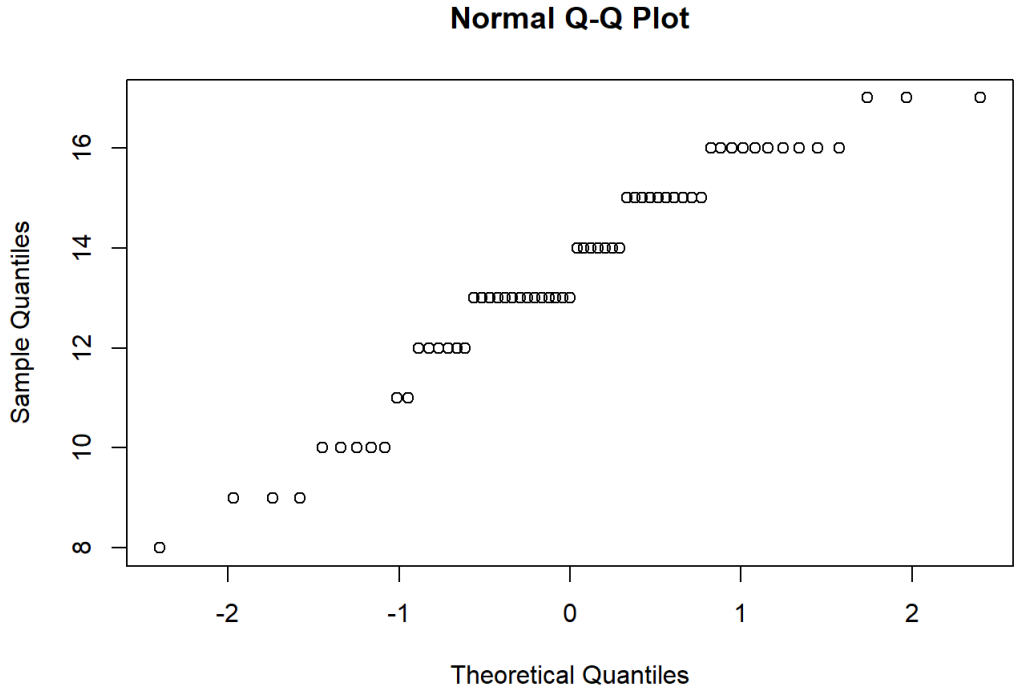
```
## [1] 0.04918033
```

Let’s test the normality of the temprature data against random data from the normal distribution:

```
qqnorm(rnorm(61))
```




```
qqnorm(temp)
```



Although the datapoints look (somehow) like a straight line, it doesn't match how genuinely gaussian data should look.

t-test and p-value

My trip lasted for 61 days. In the first 35 days of the trip, I travelled through forests located on the Eastern cost of the North island. In the last 26 days of the trip, I travelled through the west coast.

North island residents commonly say that the red degree seen in the East coast forests is significantly different than what is seen in the west coast, where the east coast leaves tend to visibly have redder shades.
Here, I will evaluate this common say using Student t-test. My null hypothesis will be that red degree in east and west are similar.

```
red_deg
```

```
##  [1] 1 1 2 2 1 2 1 1 1 3 2 3 4 1 2 1 3 3 2 4 1 1 2 2 4 2 1 1 3 1 3 3 2 5 2
## [36] 2 5 5 4 4 4 3 4 1 1 2 2 4 3 1 2 1 2 1 1 1 3 1 2 3 4
```

```
red_deg_east <- red_deg[1:35]
red_deg_east
```

```
##  [1] 1 1 2 2 1 2 1 1 1 3 2 3 4 1 2 1 3 3 2 4 1 1 2 2 4 2 1 1 3 1 3 3 2 5 2
```

```
red_deg_west <- red_deg[36:61]
red_deg_west
```

```
##  [1] 2 5 5 4 4 4 3 4 1 1 2 2 4 3 1 2 1 2 1 1 1 3 1 2 3 4
```

The p-value is defined as the probability, if the null is true, of obtaining the observation or an observation more extreme.

I will use Student's t-test to find the p-value:

```
t.test(red_deg_east, red_deg_west)
```

```
##
##  Welch Two Sample t-test
##
## data:  red_deg_east and red_deg_west
## t = -1.3925, df = 46.791, p-value = 0.1704
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.106925  0.201430
## sample estimates:
## mean of x mean of y
##  2.085714  2.538462
```

The p-value 0.1704 is not less than 5%. Therfore, I fail to reject the null hypothesis and conclude that the red degree in forests of the East and West coasts don't significantly differ.

Pearson's chi-square test

I will take a small 5-day sample out of my 61-day dataset.

```
o<- red_deg[1:5]
o
```

```
## [1] 1 1 2 2 1
```

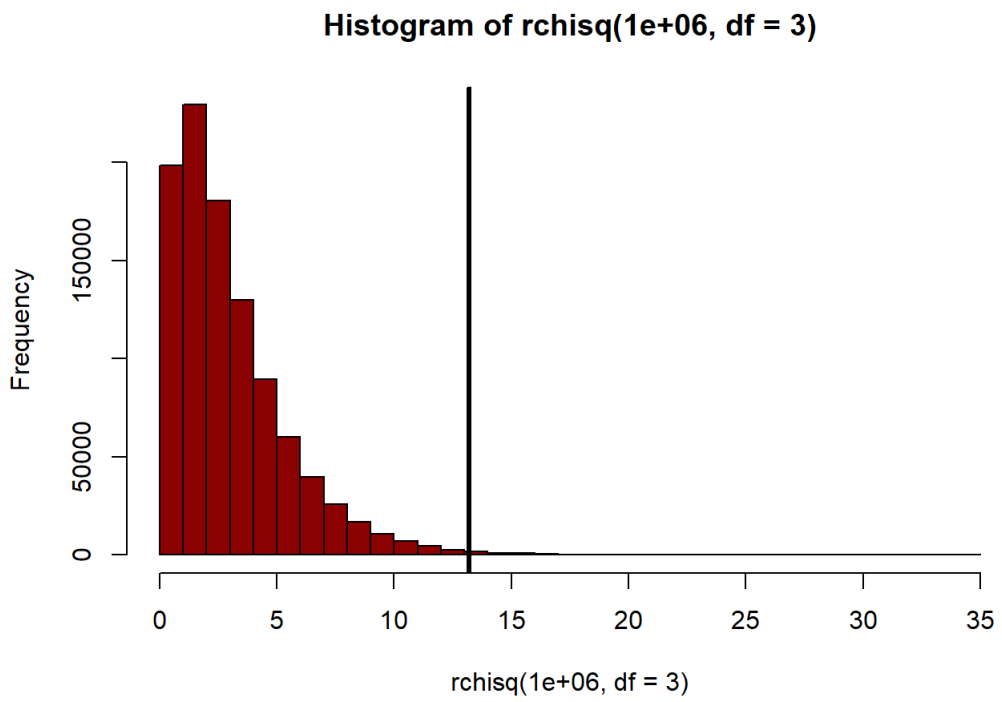
This small sample represents the degree of red colour seen in the first 5 days of the trip. I spent these 5 days in a nice regional park in Tauranga. The leaves were not very red during that time but it was slightly starting to turn into light green and yellow.
Here I will try to utilise Pearson's chi-square test to evaluate the probability of seeing leaves on the maximum red degree (5) during these 5 days.

```
e<- c(5,5,5,5,5)
B <- sum((o-e)^2/e)     #Calculating goodness-of-fit
B
```

```
## [1] 13.2
```

If my null is true, B will be distributed with a chi-square distribution with 3 degrees of freedom. I will plot a diagram to evaluate that.

```
hist(rchisq(1e6,df=3),nclass=25,col='red4')
abline(v=B,lwd=3)
```



The B value of 13.2 is marked by the vertical line.
The p-value would be all the area located to the right side of that vertical line.
Now I will try to calculate the exact pvalue:

```
pchisq(13.2,df=3,lower.tail=FALSE)
```

```
## [1] 0.004223464
```

As we see above, there is a very small probability of seeing clear red leaves during the first 5 days of the trip. This perfectly matches the common knowledge saying that leaves don't go red until after 15-20 days of Autumn start.

The Hypergeometric Distribution

When previewing the trip photos, I see that leaves with red_degree of 1, 2 are not at all red; they are preogressing on shades between the green and the light yellow. Leaves start to be on the early shades of red when they hit the red_degree 3.
It's also a known fact that leaves go red when the temperature goes below 13.

Here I will add a new field `filter` to flag each of the 61 days as below:
1 temperature 13 or above and red_degree 3 or above
2 temperature less than 13 and red_degree 3 or above
3 temperature 13 or above and red_degree less than 3
4 temperature less than 13 and red_degree less than 3

```
x<- 1:61
filter<-0

for (val in x) {
   if (temp[val] >= 13 & red_deg[val] >= 3){
        filter[val]<-1
    }
   if (temp[val] <= 12 & red_deg[val] >= 3){
        filter[val]<-2
    }
   if (temp[val] >= 13 & red_deg[val] <= 2){
        filter[val]<-3
    }
   if (temp[val] <= 12 & red_deg[val] <= 2){
        filter[val]<-4
    }
}

filter
```

```
##  [1] 3 3 3 3 3 3 3 3 3 1 4 2 2 3 4 3 1 1 3 1 3 3 3 3 1 3 3 3 1 3 1 1 4 2 4
## [36] 3 1 1 1 1 2 1 1 3 3 3 3 2 1 3 3 3 3 4 4 4 2 4 4 2 2
```

```
table(filter)
```

```
## filter
##  1  2  3  4
## 15  8 29  9
```

Let's represent this in a matrix style for better previewing:

```
a <- matrix(c(table(filter)[1],table(filter)[2],table(filter)[3],table(filter)[4]),2,2,byrow=TRUE) # define matrix
dimnames(a) <- list(Red=c("TRUE","FALSE"),temp_13=c("MORE","LESS"))
a
```

```
##        temp_13
## Red     MORE LESS
##   TRUE    15    8
##   FALSE   29    9
```

What is the probability of seeing leaves of high red_degree (3 or more) in a high temperature day (13 degrees or more)? As per the matrix above, this happened in 15 days out of the total 61 days.
I will utilise the R idiom `dhyper()` to calculate the probability of seeing this observation.

```
dhyper(15,23,38,44)
```

```
## [1] 0.1488868
```

The definition of p-value: "the probability, if the null is true, of obtaining the observation or an observation more extreme".
In this case (as I use a one-sided test), "more extreme" means greater than the red_degree observation of 15. I also know that the number of days with high temperature (above 13) can not be more than 23. So the p-value is:

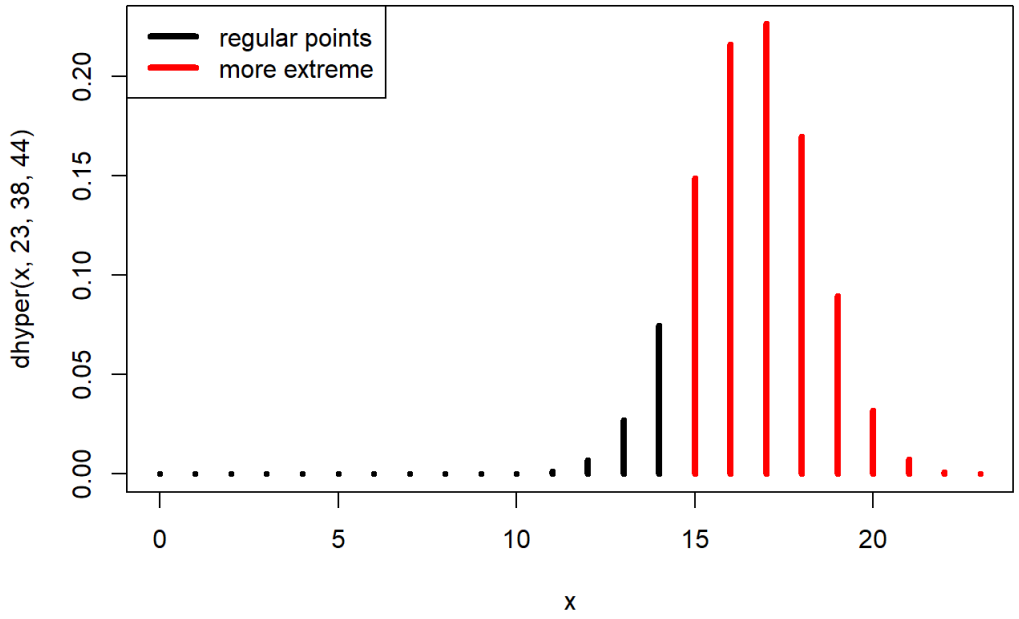
```
sum(dhyper(15:23,23,38,44))
```

```
## [1] 0.8902728
```

As we see here, the p-value is way larger than 5% and it is not significant.

I will try to visualise this:

```
x <- 0:23
plot(x,dhyper(x,23,38,44),type='h',lwd=4,col=c(rep("black",15),rep("red",20)))
legend("topleft",lwd=4, col=c("black","red"),legend=c("regular points","more extreme"))
```



Fisher's Exact Test

The `fisher.test()` function can save time and find out the p-value easily out of the matrix I used above:

```
a
```

```
##        temp_13
## Red     MORE LESS
##   TRUE    15    8
##   FALSE   29    9
```

```
fisher.test(a, alternative="greater")
```

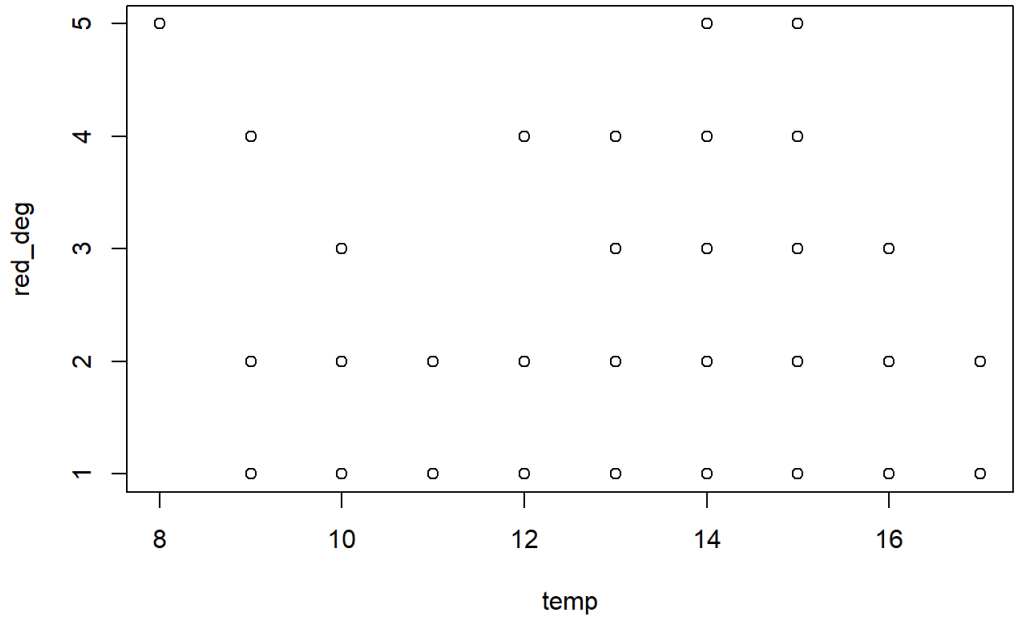
```
##
##  Fisher's Exact Test for Count Data
##
## data:  a
## p-value = 0.8903
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.194232      Inf
## sample estimates:
## odds ratio
##  0.5872479
```

Regression

In this section, I will try to apply regression methods to examine the relationship between my two variables representing the leaves' red degree and the temperature.

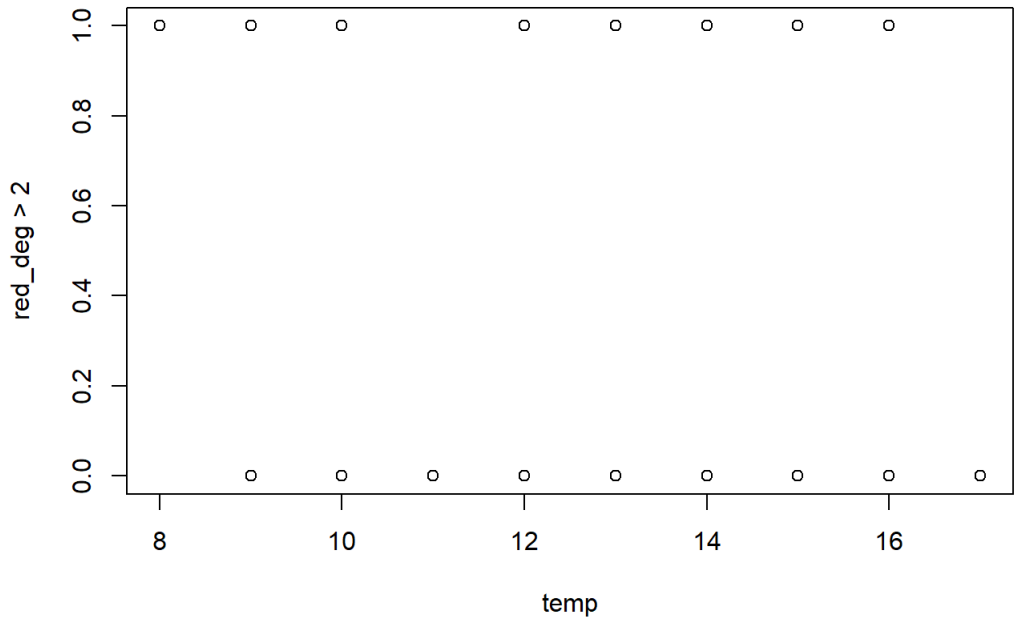
I will start by plotting the two variables to decide the right method of regression suitable for this relationship:

```
plot(red_deg~temp, data=data)
```



Based on how the scatterplot look like, linear regression doesn't seem to be an applicable solution.
I will utilise the same concept used in the Hypergeometric Distribution section (distributing red_degree to 3 and above, below 3) to try the Logistic Regression instead.
I will first plot the relationship between temperature and a binary variable red_deg>2.

```
plot(red_deg>2~temp, data=data)
```



Now I will try to apply The logistic regression.

```
fit <- glm(red_deg>2~temp,family='binomial')
summary(fit)
```

```
##
## Call:
## glm(formula = red_deg > 2 ~ temp, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0998  -0.9841  -0.9020   1.3519   1.4805
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.46248    1.59774   0.289    0.772
## temp        -0.07198    0.11793  -0.610    0.542
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80.837  on 60  degrees of freedom
## Residual deviance: 80.464  on 59  degrees of freedom
## AIC: 84.464
##
## Number of Fisher Scoring iterations: 4
```

Let's interpret the output of the `summary()` function above:

First, We see the deviance residuals, which are a measure of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model. The numbers here look good, as they are closely centred around zero, and are roughly symmetrical.

The second part of the summary shows the estimate of the regression beta coefficients, their standard errors, the z-statistic, and the associated p-values.

The intercept (b0) is 0.46248 and the coefficient of temperature variable is -0.07198. Both regression correficients correspond to the following model:

$$Red\_degree = 0.46248 - 0.07198 * temperature$$

Both p-values are well above the 0.05 value, which means that they are insignificant.

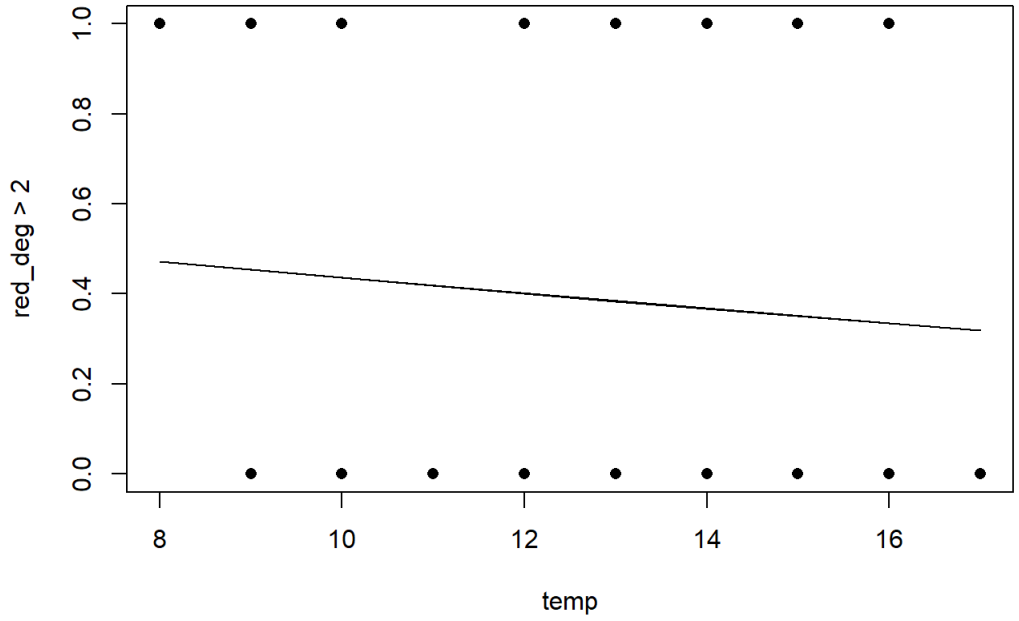
Next we see the Dispersion parameter used. Because we are doing logistic regression here, the variance is derived from the mean and not from the data. Therefore, it is possible that the variance is underestimated.

Below that is the table of Null deviance and Residuals deviance, which can be used to compare models. I also see the AIC value, which is basically the Residuals deviance adjusted for the number of the parameters in this model. The AIC is also useful if we need to compare this model to another.

And finally we see the number of Fisher Scoring iterations, which tells us how quickly the `glm()` function was able to converge on the maximum likelihood estimates for the coefficients.

Now, let's plot the regression line:

```
plot(red_deg>2~temp,pch=16)
points(temp,plogis(0.46248 + (-0.07198*temp)),type='l')
```



Conclusion

My dataset of red leaves and temperatures was analysed using statistical methods and a number of results were obtained. Based on the data, the probability of seeing red leaves is inversely proportional to the degree of the red colour, where the higher the degree of the red colour, the lower the probability. I also evaluated the assumption that the red degree seen in the east coast leaves is significantly higher than what is seen in the west coast ones, and concluded that, based on the data, the red degree in east and west coasts are not significantly different. On another hand, I utilised Pearson's chi-square test to study the probability of seeing maximum red degree (5) in the very early days of autumn (first five days), and concluded that it's very rare to happen (probability 0.004). I also applied Fisher's Exact Test to asses the hypothesis stating that leaves go red when the temperature goes below 13, and failed to reject it based on the collected data. And finally, I applied logistic regression methods to find an appropriate regression model describing the relationship between the high red degrees (3, 4, 5) and the temprature, and reached an insignificant result ($p>0.05$).