Q1

1. The sample mean height of the cross-fertilized seedlings is: **20.19333**

> mean(seedlings$Crossfertilised)

[1] 20.19333

The sample mean height of the self-fertilized seedlings is:**17.58667**

> mean(seedlings$Selffertilised)

[1] 17.58667

The standard error of the cross-fertilized seedlings is: **0.9337**

> sd(seedlings$Crossfertilised)/sqrt(length(seedlings$Crossfertilised))

[1] 0.9336802

The standard error of the self-fertilized seedling is: **0.52625**

> sd(seedlings$Selffertilised)/sqrt(length(seedlings$Selffertilised))

[1] 0.5262506

1. The 95% confidence interval for the mean height of the cross-fertilized plants is:

**(18.191, 22.196)**

> mean(seedlings$Crossfertilised) - qt(0.975,14)\*sd(seedlings$Crossfertilised)/sqrt(length(seedlings$Crossfertilised))

[1] 18.19079

> mean(seedlings$Crossfertilised) + qt(0.975,14)\*sd(seedlings$Crossfertilised)/sqrt(length(seedlings$Crossfertilised))

[1] 22.19588

1. If the experiment was carried out 500 times and I calculated the 95% confidence interval each time, I would expect 500\*0.95 = **475 experiments** to contain the population mean.
2. H0 : mu = 19 inches. H1 : mu ≠ 19 inches. Since our 95% confidence interval is from (18.191,22.196), and 19 inches **lies within the confidence interval**, **there is no evidence to reject the null hypothesis** which means that there is no evidence that the mean is different to 19 inches.

Q2

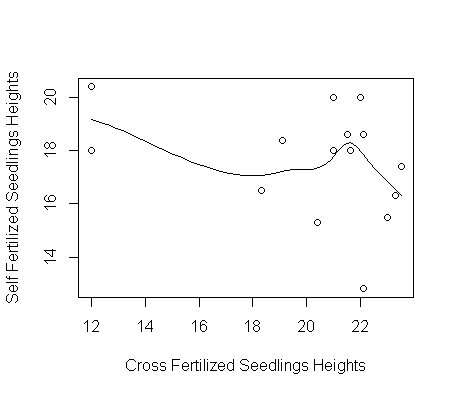
1. The scatter plots are presented below (the line shows the closest line of fit to the scatter plot). There doesn’t seem to be a clear correlation between the two sets of heights when looking at the graphs, however, finding the correlation between the two heights there is a **slight negative correlation** between them. **The pearson’s correlation is -0.3379 while spearman’s correlation is -0.3348**

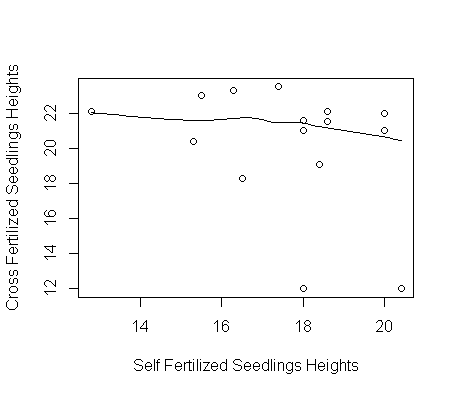
> cor(seedlings$Crossfertilised, seedlings$Selffertilised)

[1] -0.3378578

> cor(seedlings$Selffertilised, seedlings$Crossfertilised, method="spearman")

[1] -0.3348347





1. **H0: µ(cross fertilized plants height) = µ(self fertilized plants height)**

**H1: µ(cross fertilized plants height) > µ(self fertilized plants height)**

1. Carrying out a parametric test:

> t.test(seedlings$Crossfertilised, seedlings$Selffertilised, alt="greater", paired = T)

Paired t-test

data: seedlings$Crossfertilised and seedlings$Selffertilised

t = 2.1422, df = 14, p-value = 0.02512

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.4634257 Inf

sample estimates:

mean of the differences

2.606667

As you can see, **the p-value is 0.02512**. So with a **test at the 5% level, since 0.02512 < 0.05, we would reject the null hypothesis** and conclude that the heights of the Cross fertilized plants are indeed greater than the Self Fertilized plants. However, if we **did test at the 1% level, we would get a different conclusion as 0.02512> 0.01, we would not be able to reject the null hypothesis.**

1. If we wanted to carry out a non-parametric test for part c, we would use the **Wilcoxon signed ranks test** as we are using **paired data**.

Q3

1. E[Y] =

Ῡ =

λ \* Ῡ = 1

λ=

**λ=**

1. The mean for the rainfall data is: **19.93**

> mean(rainfall)

[1] 19.93333

So to calculate λ:

λ= = 1/19.93 = 0.05

> 1/mean(rainfall)

[1] 0.05016722

**So λ=0.05**

1. The Pr[Y > 35] = 0.173

> lambda <- 1/mean(rainfall)

> 1- pexp(35, rate=lambda)

[1] 0.1727598

1. The graph is kind of hard to see so I have shown each of them individually and then shown the overlayed graph. The exponential distribution doesn’t seem appropriate because it doesn’t match up with the shape of the histogram as an exponential distribution would have greater values on the left side while gradually decreasing towards a right tail. I picked a range from 10 to 30 as the minimum value in rainfall was 12 and the maximum was 29.

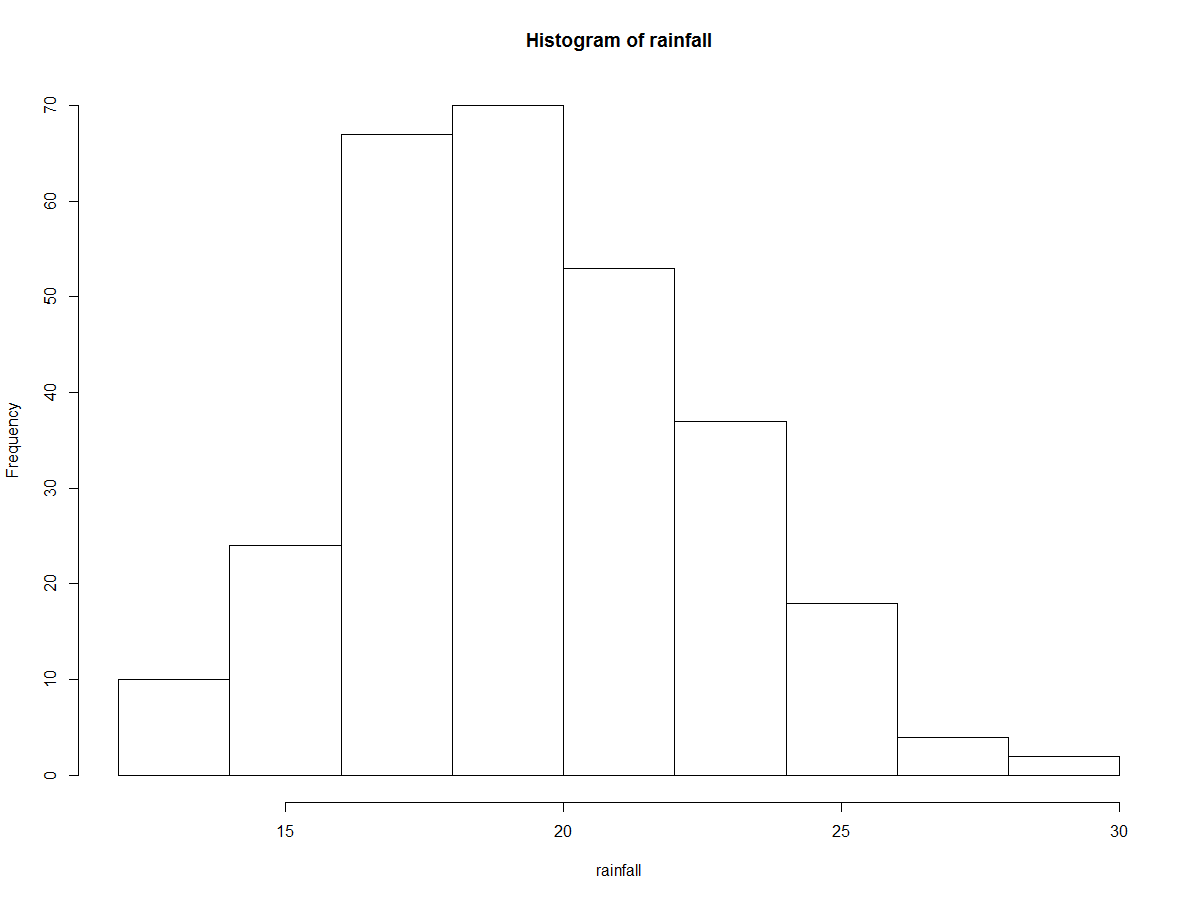
> hist(rainfall)

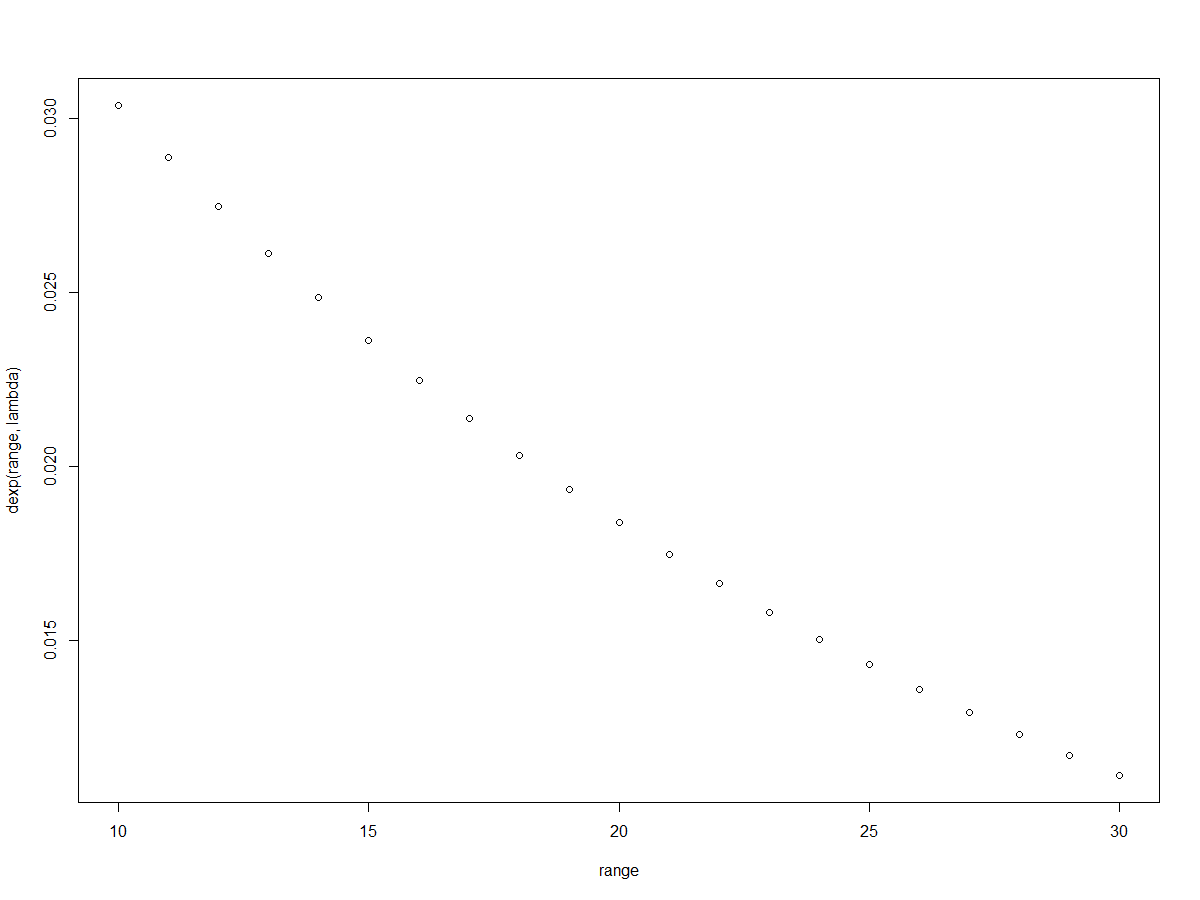
> range <- 10:30

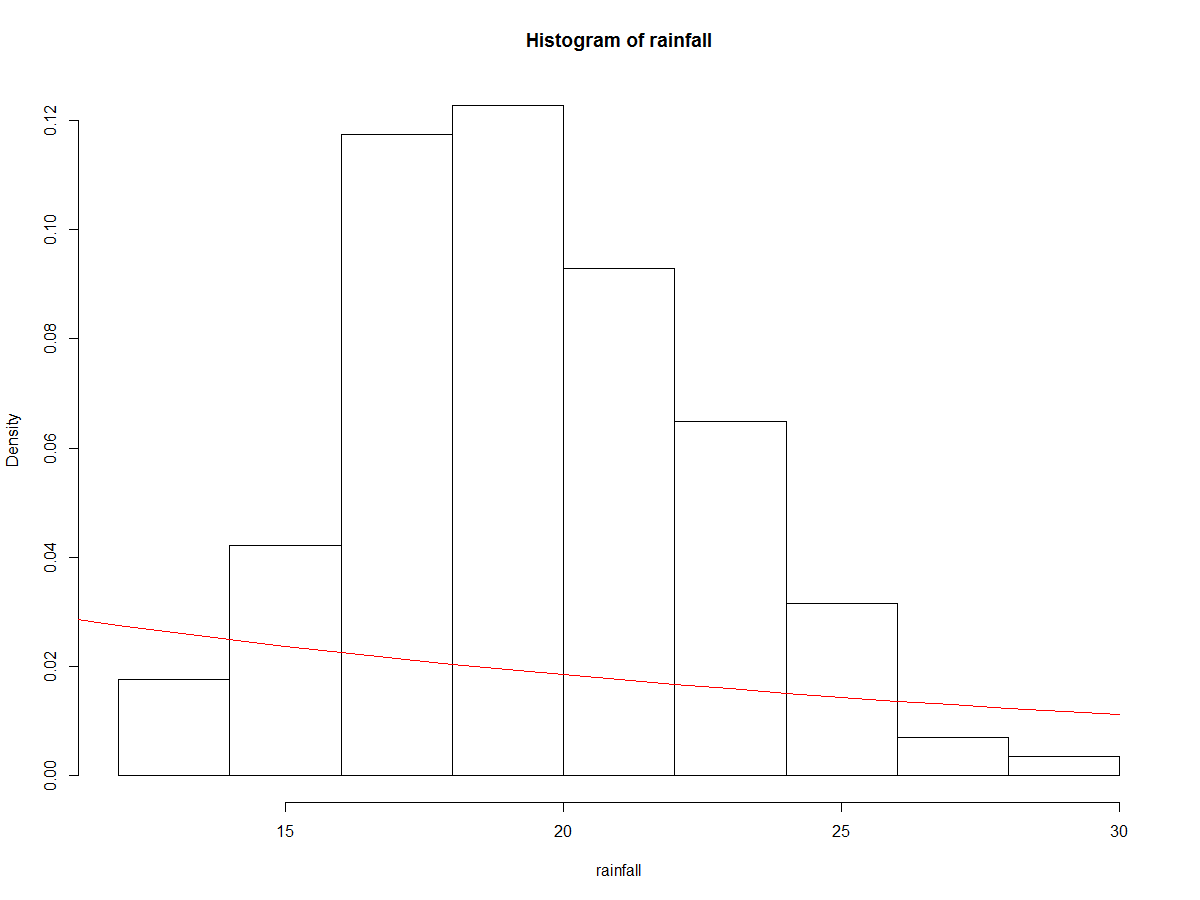
> plot(range, dexp(range, lambda))

> hist(rainfall, freq = FALSE)

> lines(range, y = dexp(range, lambda),col=2)







Q4

1. E[Y] =

Ῡ =

β\* Ῡ = α

α= β \*

Var(Y) =

Var(Y) = E(Y2) - ((E(Y))2

Var(Y) = 2 – ()2

= 2 – ()2

α = β2 \* [ 2 – ()2]

β\* Ῡ = α

β\* = α

β\* = β2 \* [ 2 – ()2]

= β \* [ 2 – ()2]

β= / [ 2 – ()2]

**β= / [ 2 – ()2]**

α= β \*

**α= 2 / [ 2 – ()2]**

or written differently

**β= / [** E(Y2) **– ()2]**

**α=** Ῡ2 **/ [** E(Y2) **– ()2]**

1. Using R and the rainfall data, we get **Beta = 2.018 and Alpha = 40.217**:

> Ybar <- sum(rainfall) / length(rainfall)

> Ybar

[1] 19.93333

> secondexpectation <- sum(rainfall^2) / length(rainfall)

> secondexpectation

[1] 407.2175

> beta <- (Ybar) / (secondexpectation - (Ybar^2))

> beta

[1] 2.017592

> alpha <- (Ybar^2) / (secondexpectation - (Ybar^2))

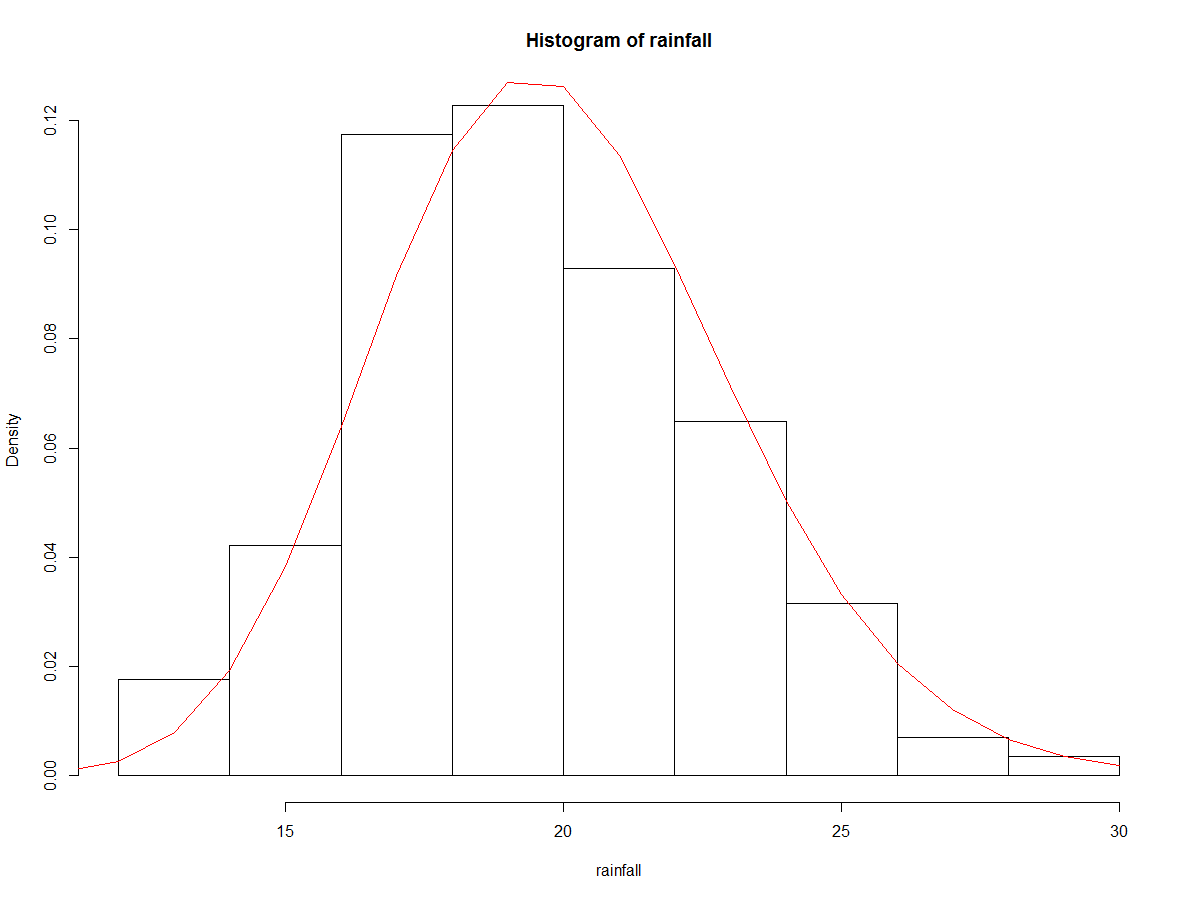
> alpha

[1] 40.21733

1. Here is the plot of the density of the fitted model using dgamma, as you can see with the graph, the fitted model of dgamma (the red line) matches up with the shape of the histogram a lot better then the exponential model and is thus a more appropriate model to use:

> hist(rainfall, freq=FALSE)

> lines(range, dgamma(range, shape=alpha, rate=beta), col=2)



1. **The Pr[Y>35] = 3.319e-05** under this model so it is extremely small which matches up with the graph as well.

> 1 - pgamma(35, shape = alpha, rate=beta)

[1] 3.31941e-05

1. Here is the code to generate the bootstrap sampling distribution for Pr[Y>35]. The 95% confidence interval is (0.9999030, 0.9999928) that the Pr[Y>35] for 1000 bootstrapped samples. Ofcoarse the histogram shape and the confidence interval will vary because of RNG.

> probBS = c()

> dataYbar <- c()

> datasecondexp <- c()

> dataalpha <- c()

> databeta <- c()

> for( j in 1:1000){

+ data = sample(rainfall, replace=T)

+ dataYbar[j] <- sum(data)/length(data)

+ datasecondexp[j] <- sum(data^2)/length(data)

+ dataalpha[j] <- (dataYbar[j]^2) / (datasecondexp[j] - (dataYbar[j])^2)

+ databeta[j] <- (dataYbar[j]) / (datasecondexp[j] - (dataYbar[j])^2)

+ probBS[j] = 1-dgamma(35, shape = dataalpha[j], rate = databeta[j])

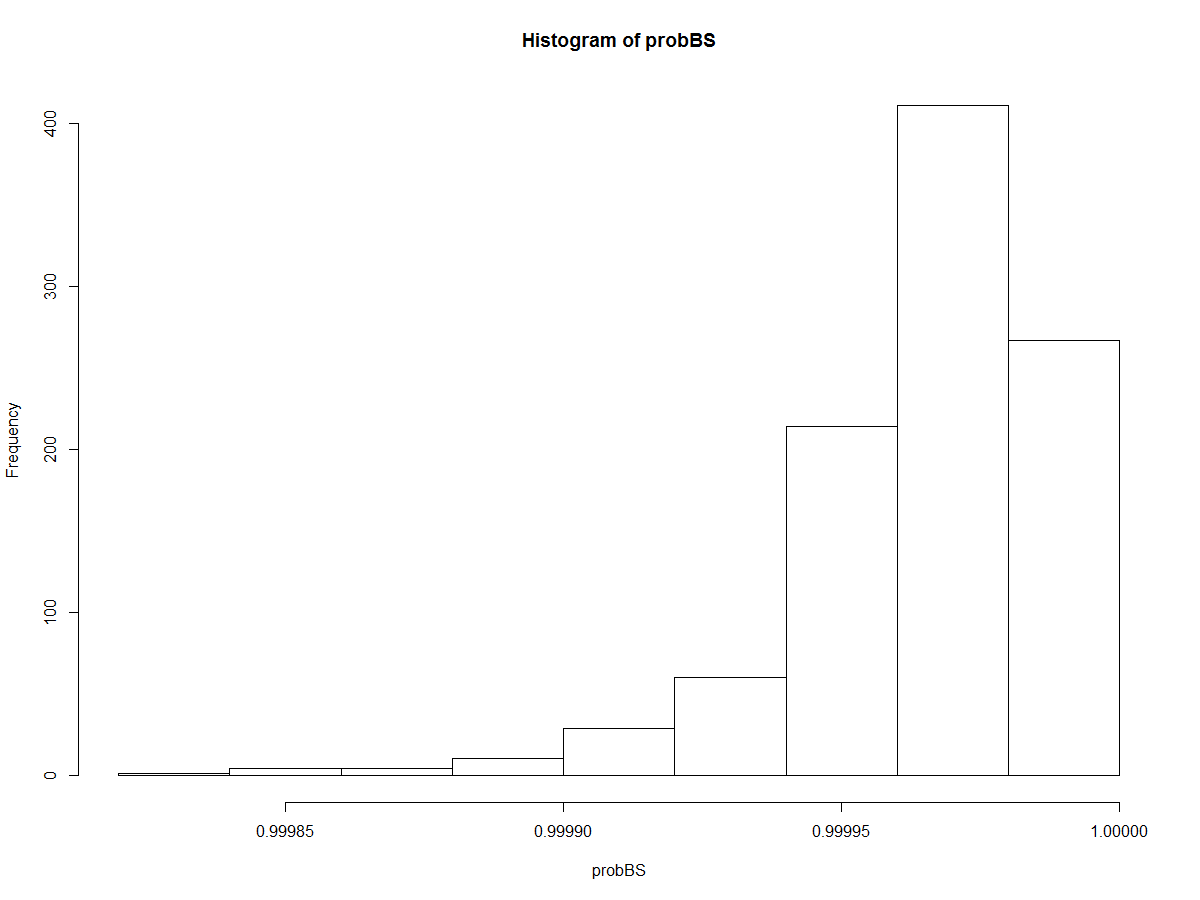
+ }

> quantile(probBS, c(0.025, 0.975))

2.5% 97.5%

0.9999030 0.9999928

> hist(probBS)



Q5

1. Here are the row, column, and overall totals

|  |  |  |  |
| --- | --- | --- | --- |
|  | Invaded | Not Invaded | Row Totals |
| Species A | 2 | 13 | Species A Total: 15 |
| Species B | 10 | 3 | Species B Total: 13 |
| Column Totals | Invaded Total: 12 | Not Invaded Total: 16 | Overall Total: 28 |

1. I have performed both the chi-squared test manually and automatically below both resulting in the same conclusion.

**Doing the chi-squared test manually**:

This is the expected values table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Invaded | Not Invaded | Row Totals |
| Species A | 15\*12/28 = 6.429 | 15\*16/28 = 8.571 | Species A Total: 15 |
| Species B | 12\*13/28 = 5.571 | 16\*13/28 = 7.429 | Species B Total: 13 |
| Column Totals | Invaded Total: 12 | Not Invaded Total: 16 | Overall Total: 28 |

H0 = Ants do not prefer one species of tree over the other

H1 = Ants do prefer one species of tree over the other

X2 test = (2-6.429)^2 / 6.429 + (13-8.571)^2/8.571 + (10-5.571)^2 / 5.571 + (3-7.429)^2 / 7.429 = 11.499

> 1-pchisq(11.49915, 1, 1)

[1] 0.008406009

This will lead us to **reject H0 when using a confidence level of 0.05 as 0.008 < 0.05 which will lead us to conclude that ants do prefer one species of tree over the other.**

**Doing the Chi-squared test automatically in R:**

> chisq.test(ants)

Pearson's Chi-squared test with Yates' continuity correction

data: ants

X-squared = 9.0491, df = 1, p-value = 0.002628

Leads us to the same result with a slightly different p-value of 0.0026. **However, we will still reject H0 and conclude that ants do prefer one species of tree over the other as 0.0026 < 0.05.**

1. **Fisher’s exact test is an exact test of association for a 2x2 data set, however, unlike the X2  test, it doesn’t rely on asymptotic results and it is useful when we don’t have a large data set and when at least one of the expected responses is less then 5**.
2. Doing the fisher’s test automatically in R:

> fisher.test(ants)

Fisher's Exact Test for Count Data

data: ants

p-value = 0.001624

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.003786123 0.425475250

sample estimates:

odds ratio

0.05401494

**We obtain a p-value of 0.001 so our conclusion doesn’t change as we still reject H0 and conclude that ants do prefer one species of tree over another as 0.001 < 0.05.**

Q6. The total number of ants in the sample is: 8+14+8+6+4 = 40 ants

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species of Ants | Small | Semi-Small | Medium | Semi-Large | Large |
| Frequency | 8 | 14 | 8 | 6 | 4 |
| Expected | 40\*20% = 8 | 40\*30% = 12 | 40\*20% = 8 | 40\*25% = 10 | 40\*5% = 2 |

1. I am going to use the X2 test to test whether the ant colony in the tree colonization experiment is consistent with the findings of the original study. So it barely fulfils the criteria as all expected frequencies are greater than 1 and at most 20% of expected frequencies is less then 5 (only 1 in this case out of 5 which is 20%).

H0 = Sample probabilities equal population properties

H1 = Sample probabilities do not equal population properties.

Test Statistic: (((8-8)^2)/8) + (((14-12)^2)/14) + (((8-8)^2)/8) + (((6-10)^2)/10) + (((4-2)^2)/2) = 3.886

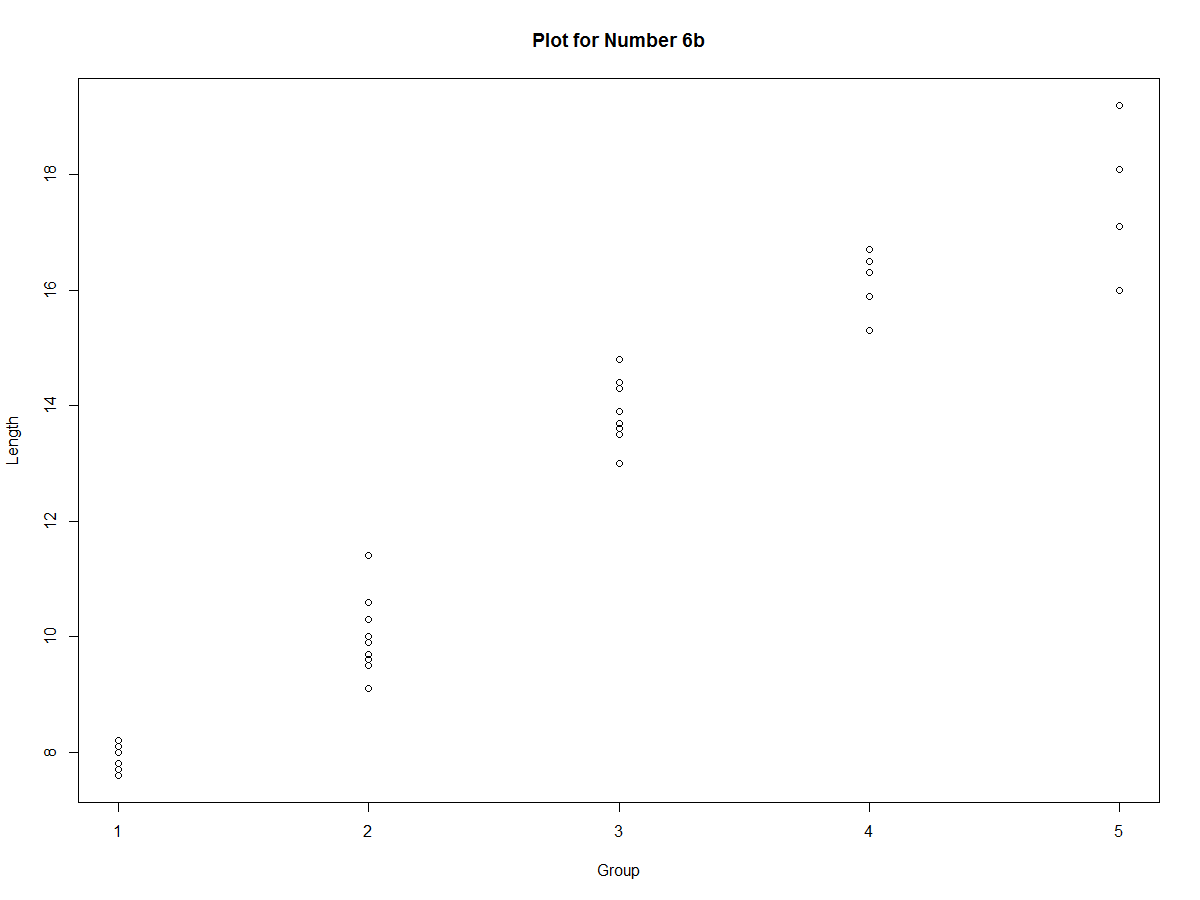
The degrees of freedom is 4 because there are 5 groups and you subtract 1.

> 1-pchisq(3.885714, 4)

[1] 0.4216937

**Since we get a p-value of 0.42 and 0.42 > 0.05 or 0.01, we can not reject H0 and therefore conclude that the ant colony experiment is consistent with the finding of the original study in terms of the proportions of workers in each of the size categories.**

1. Here is the plot which shows group on the x-axis while showing the ant lengths on the y axis. I used a simple plot to show the different in ant lengths heights in different groups.



> onegroup <- antLengths$Length[antLengths$Group == 1]

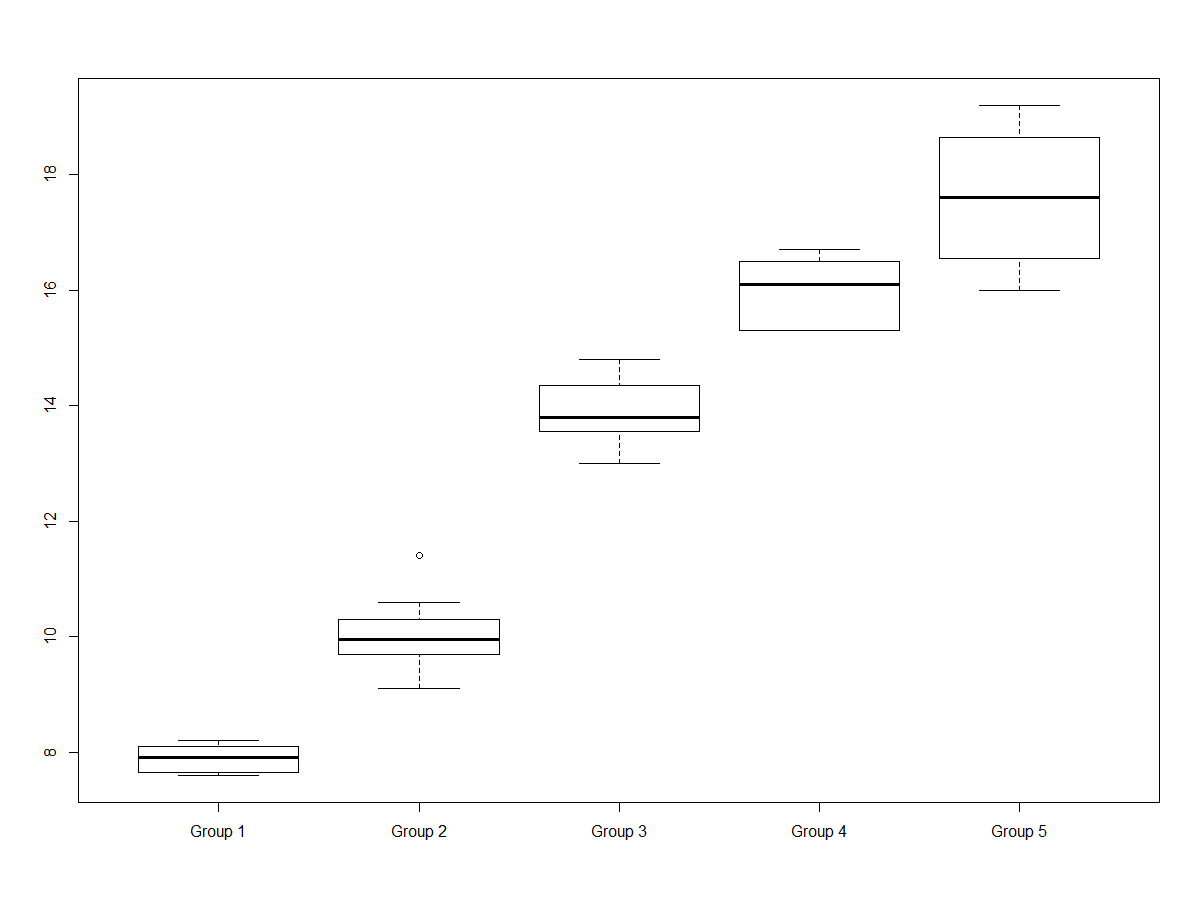
> twogroup <- antLengths$Length[antLengths$Group == 2]

> threegroup <- antLengths$Length[antLengths$Group == 3]

> fourgroup <- antLengths$Length[antLengths$Group == 4]

> fivegroup <- antLengths$Length[antLengths$Group == 5]

> boxplot(list(onegroup, twogroup, threegroup, fourgroup, fivegroup), names=c("Group 1","Group 2","Group 3","Group 4","Group 5"))



I have also depicted a boxplot separating each group and showing the lengths distribution on the y-axis, **it clearly shows that the mean lengths of ants are different depending on which group they are in with an increasing length when going from Group 1 to Group 5.**

1. I have used the non-parametric Kruskal-Wallis test because our data set is small and it doesn’t require any normality assumption, is less influenced by outliers, and is generally useful for small samples.

H0 : µ1 = µ2 = µ3 = µ4 = µ5

H1 : µ1 ≠ µ2 ≠µ3 ≠µ4 ≠ µ5

> kruskal.test(unlist(antLengths$Length) ~ antLengths$Group)

Kruskal-Wallis rank sum test

data: unlist(antLengths$Length) by antLengths$Group

Kruskal-Wallis chi-squared = 36.405, df = 4, p-value = 2.388e-07

**As you can see from the results of this test, we obtain a p-value of 2.388e-07 < 0.05. Which is less then our 5% confidence level so we can reject H0 and conclude that there is significant evidence that the mean ant lengths differ between the five groups.**