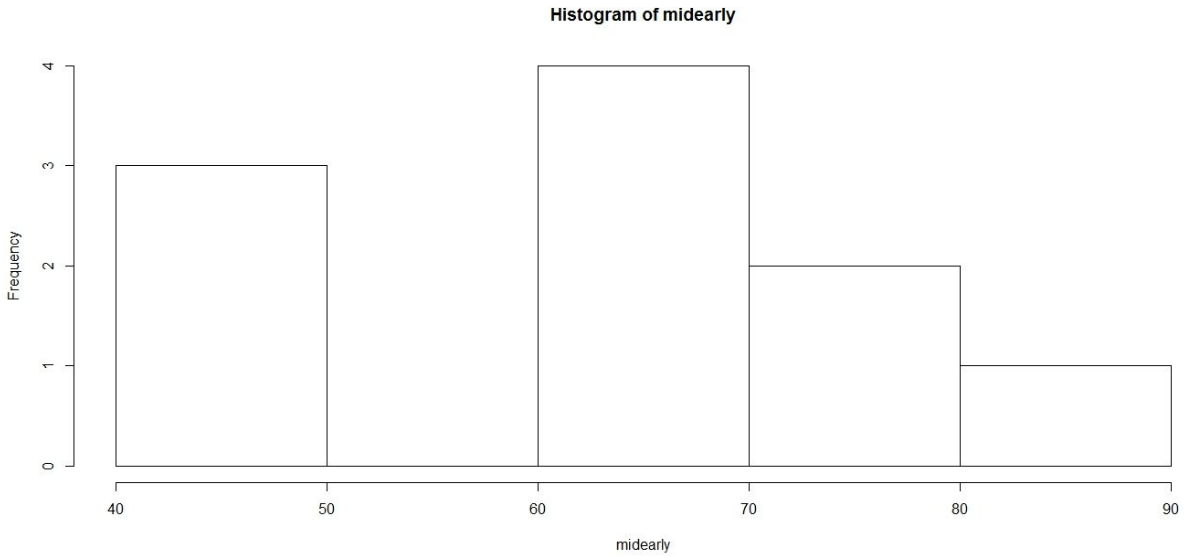
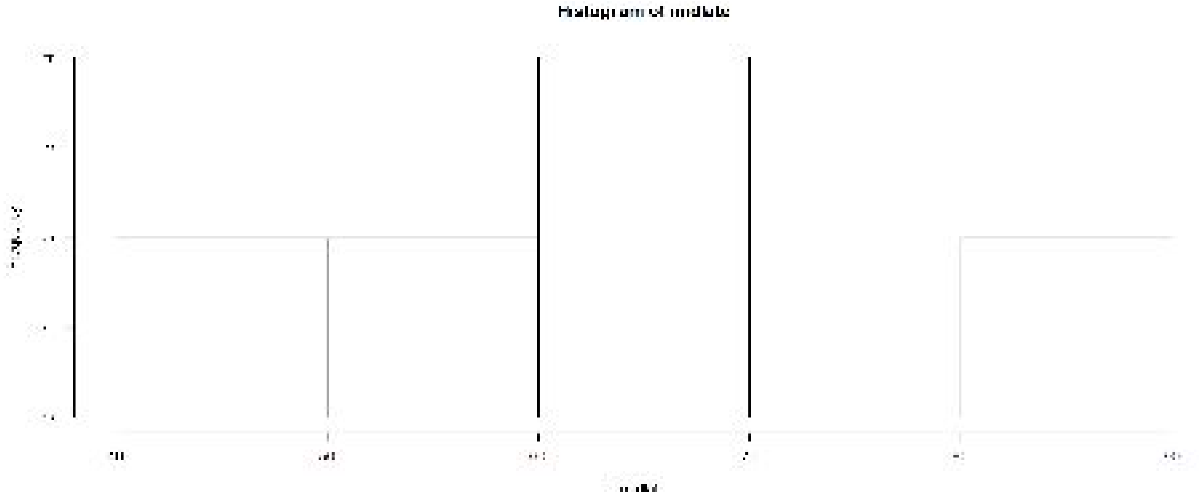
Question 1

(ii)
```
hist(midearly)
hist(midlate)
```

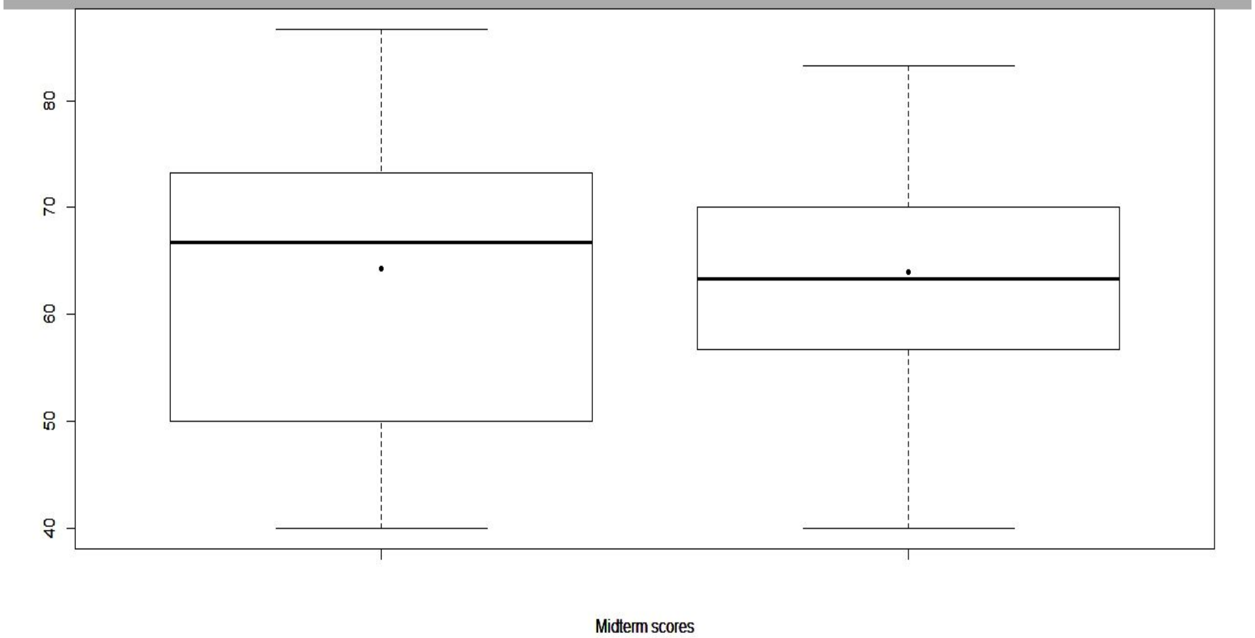


This is the histogram representing the sample of the midterms which were submitted early. The scores here seem to be random and do not conform to a standard distribution as the histogram does not display the required bell-curve. No scores were in the 50-60 range, this is likely an anomaly due to small sample size.



The histogram representing the sample of the midterms which were submitted late skews somewhat to the right as there are a few values greater than the main body of data (2 scores in the 80-90 range with none in the 70-80 range).

```
boxplot(midearly, midlate, xlab="Midterm scores")
points(x=1,y=mean(midearly),pch=20)
points(x=2,y=mean(midlate),pch=20)
```



It can be seen from this boxplot that the first quartile of the midearly values(on the left) is lower than the first quartile of the midlate values but also that the third quartile is greater. The mean is about even while the median (second quartile) is slightly higher for the midearly values.

```
dataTable <- matrix(c(mean(midearly),sd(midearly),quantile(midearly),IQR(midearly),mean(midlate),sd(midlate),quantile(midlate),IQR(midlate)),ncol=2,nrow=8,byrow=FALSE)
colnames(dataTable) <- c("Midearly","Midlate")
rownames(dataTable) <- c("Mean","SD","Min","1st Quantile","2nd Quantile","3rd Quantile","Max","IQR")
dataTable <- as.table(dataTable)
dataTable
```

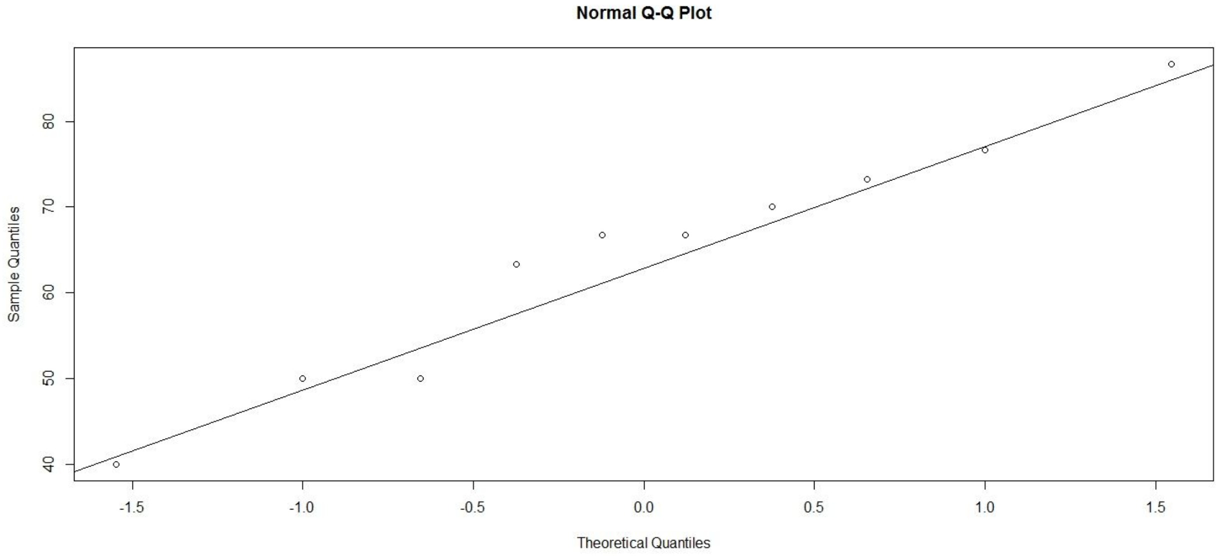
	Midearly	Midlate
Mean	64.34000	63.99000
SD	14.06724	13.57796
Min	40.00000	40.00000
1st Quantile	53.32500	57.52500
2nd Quantile	66.70000	63.30000
3rd Quantile	72.47500	70.00000
Min	86.70000	83.30000
IQR	19.15000	12.47500

This table confirms what the box-plot demonstrated earlier: that the dispersion of the Midearly values is greater than that of the Midlate values. This is shown by the fact that the values of the sample of midterms submitted early have a higher standard distribution and a higher InterQuartile range.

(iii)

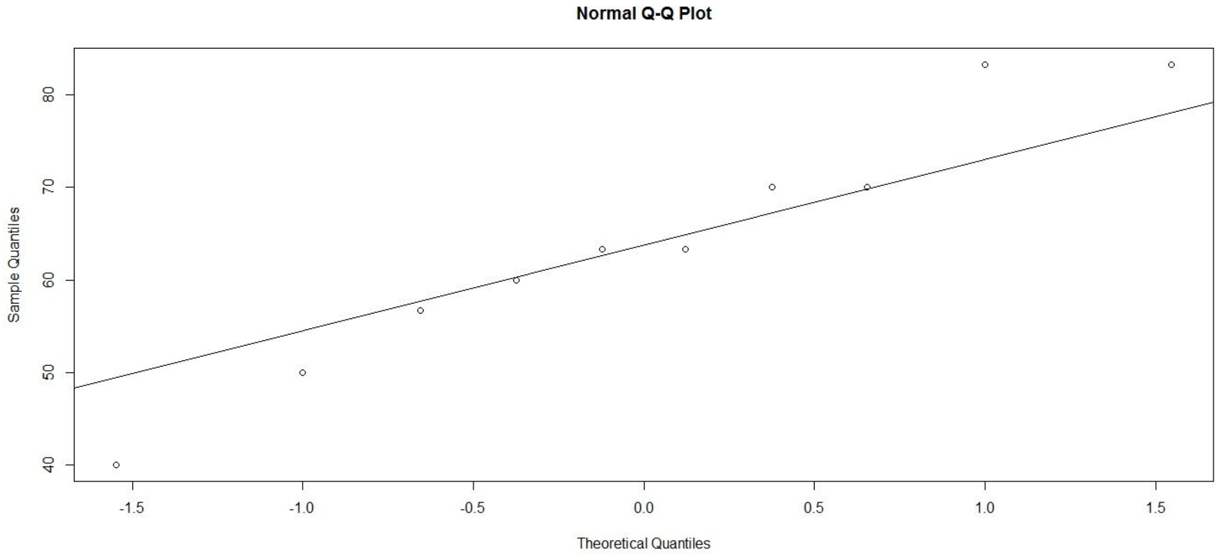
```
qqnorm(midearly)
```

```
qqline(midearly)
```



In this qqplot the values start on the line, go below, then above, then onto the line again. This shows that the midearly values are too wide to be considered normally distributed.

```
qqnorm(midlate)
qqline(midlate)
```



In this qqplot the values start below the line, go slightly above, below again, above, back onto the line before finishing above the line. This pattern indicates that the midlate data vector is also too wide for normality..

We perform the Shapiro-Wilks test to check the hypothesis that the midearly values represent a population which is normally distributed. The Shapiro-Wilks test, and more generally any significance test, answers the question: is there enough evidence for non-normality to overthrow the null hypothesis.

```
shapiro.test(midearly)
```

Null Hypothesis = The sample midearly represents a population which is normally distributed.
Alternative Hypothesis = The sample midearly does not represent a population which is normally distributed.

```
Shapiro-Wilk normality test

data:  midearly
W = 0.96204, p-value = 0.8089
```

If we assume the significance level at 0.05 then the p-value is greater than alpha (0.8089>0.05) and so we cannot reject the hypothesis that the sample (midearly) comes from a population that is normally distributed.

```
shapiro.test(midlate)
```

Null Hypothesis = The sample midlate represents a population which is normally distributed.
Alternative Hypothesis = The sample midlate does not represent a population which is normally distributed.

```
 Shapiro-Wilk normality test

data:  midlate
W = 0.95829, p-value = 0.7662
```

If we assume the significance level at 0.05 then the p-value is greater than alpha (0.7662>0.05) and so we cannot reject that hypothesis that the sample (midlate) comes from a population (all of the midterms submitted late) that is normally distributed.

iv) **> t.test(rows, mu=50)**

```
 One Sample t-test

data:  rows
t = -0.013998, df = 19, p-value = 0.989
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 34.9475 64.8525
sample estimates:
mean of x
     49.9
```

Null hypothesis = 50, **Alternative hypothesis** ≠ 50;

By the results you can clearly see that 50 is between the confidence intervals and the p-value is high so there is little evidence to reject the **Null Hypothesis**.

iv) **> var.test(midlate, midearly)**

```
F test to compare two variances

data:  midlate and midearly
F = 0.93165, num df = 9, denom df = 9, p-value = 0.9177
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.231408 3.750807
sample estimates:
ratio of variances
         0.9316473
```

Null hypothesis = 1, **Alternative hypothesis** ≠ 1;

The p-value is much greater than 0.05 and therefore the hypothesis that the variances of midlate and midearly are equal is accepted.

 iv) **> t.test(midearly, midlate, var.equal=TRUE)**

```
 Two Sample t-test

data:  midearly and midlate
t = 0.05661, df = 18, p-value = 0.9555
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.63921  13.33921
sample estimates:
mean of x mean of y
    64.34     63.99
```

Null hypothesis = 1, **Alternative hypothesis** ≠ 1;

I decided to use the data I collected from part (iii) in this question, as from part (iii) I came to a conclusion that the variances were equal in both groups from the output. Once again the p-value is nowhere close to 0.05 and the difference between the two means(0.35) is also found within the 95 confidence intervals. So it satisfies the **null hypothesis**.

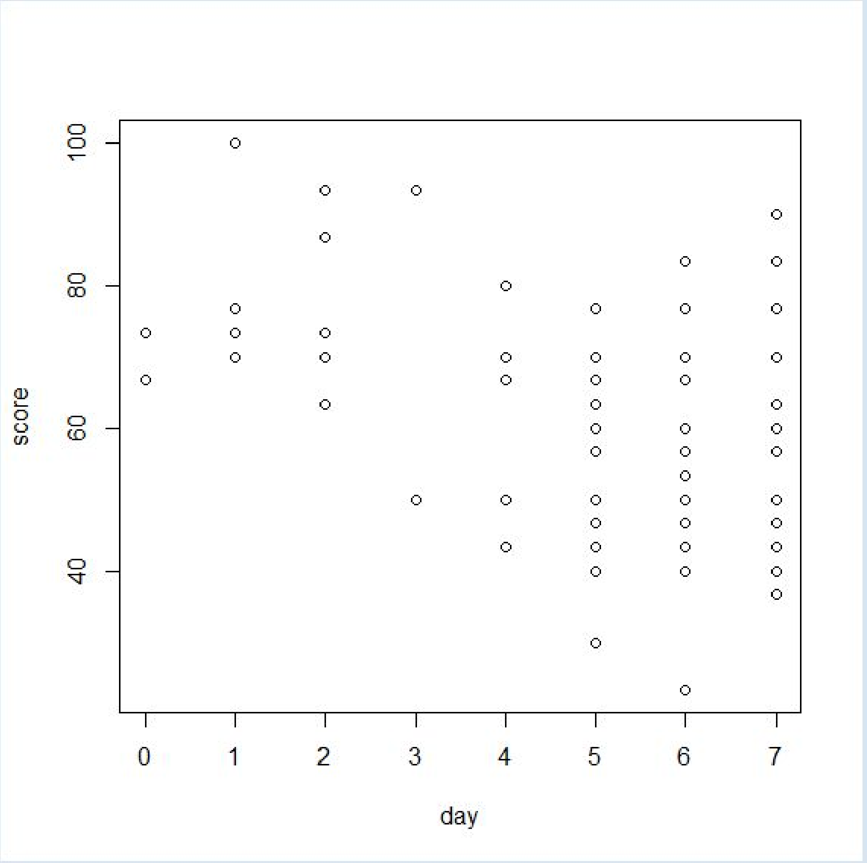
iv) **> wilcox.test(midearly, midlate, mu=0, conf.int=T, conf.level=0.95, correct=T,paired=F, exact=F)**

```
 Wilcoxon rank sum test with continuity correction

data:  midearly and midlate
W = 53.5, p-value = 0.8196
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -13.30005  13.39994
sample estimates:
difference in location
              1.48858
```

Again the p-value is not close to the value of 0.05 and the difference in location is between the two CI values and so the null hypothesis is accepted.

v) **> with(midscores, plot(score~day)**



From the graph you can see the points are quite scattered. Most points are skewed to the right.

ii) There is a negative correlation between the two values [-0.4382445]. This means that it is an inverse correlation, as one value goes decreases the other increases. As the days go on through 1-7 the scores seem to drop as between the first 3 days the average score was quite high but dropped significantly from 4-7.

v)

```
> model=lm(score~day, data=midscores)
> summary(model)
```

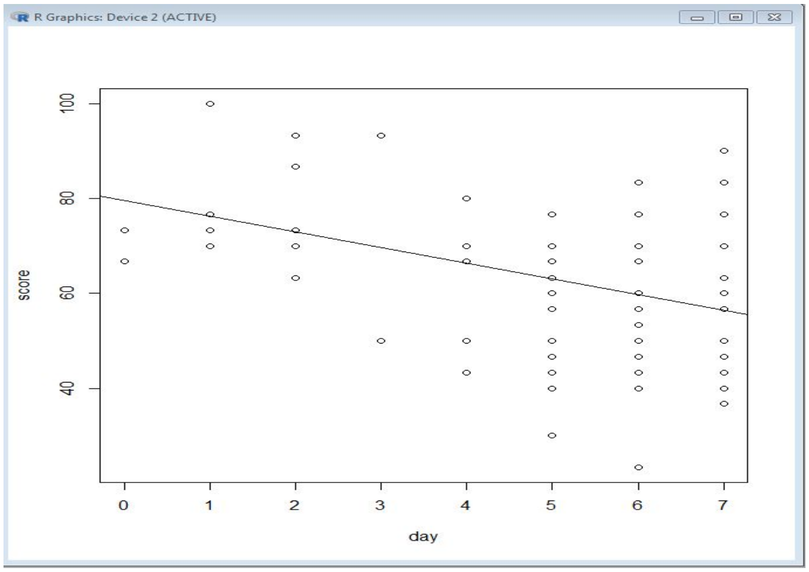
>abline(model)

```
Call:
lm(formula = score ~ day, data = midscores)

Residuals:
    Min      1Q  Median      3Q     Max 
-36.463  -9.763  -1.345  10.237  33.528 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  79.5092     3.4466  23.069  < 2e-16 ***
day          -3.2910     0.6819  -4.827 5.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.01 on 98 degrees of freedom
Multiple R-squared:  0.1921,    Adjusted R-squared:  0.1838 
F-statistic:  23.3 on 1 and 98 DF,  p-value: 5.11e-06
```



(Regression Line)

The regression line above is quite far away from a lot of the points on the graph.

R-squared in summary(model) is very low, the closer this value is to 0 means how accurate the line is which means how close are the points to the line. Here its nearly at 0 which means it is very inaccurate.

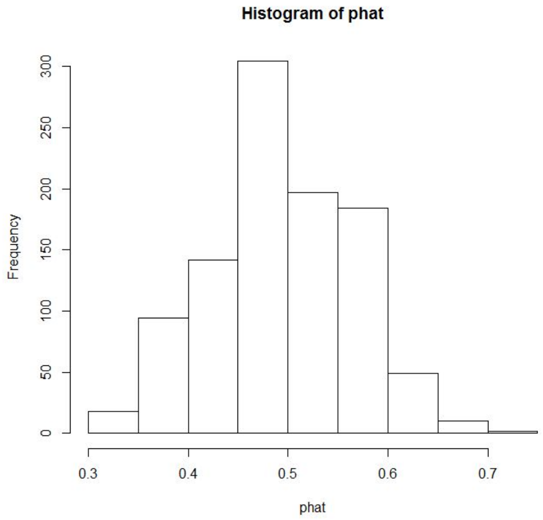
2) Simulation Study
The code and graphs below represent my results for a simulation of Bernoulli data. I set the loop to go for 1000 iterations each time for a fairer and more accurate result.
As seen below we began with getting specific p-hat when our sample size is 50 and the probability of a 1 is .5 .

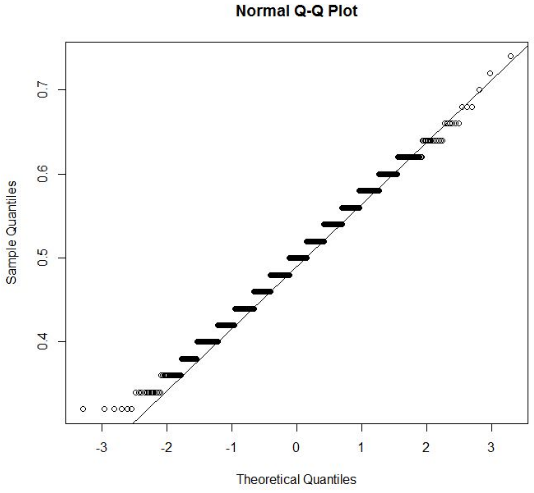
vi)

From part (ii) we can see that the dispersion of the midearly values is greater than the dispersion of the midlate values (this is shown by the greater interquartile range and standard distribution) while the mean of the midearly values is slightly higher. In part (iii) we used qqplots to determine that the midearly and midlate data vectors were too widely distributed to be considered to conform to normal distribution. The Shapiro-Wilks test showed that there was not enough evidence to overthrow the null hypothesis (that the sample represented a population that is normally distributed) for either the midearly or midlate values. We can also see that the correlation between the days to scores is negative and all of the null hypothesis are true.

Question 2

```
> set.seed(15144127)
> simreps=1000
> phat=rep(0,simreps)
> for(i in 1:simreps){phat[i] =mean(rbinom(n=50, size=1, prob=.5)) }
> hist(phat)
> qqplot(phat);qqline(phat)
```

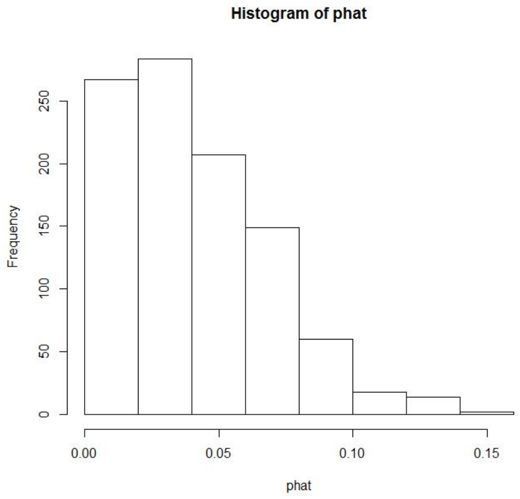


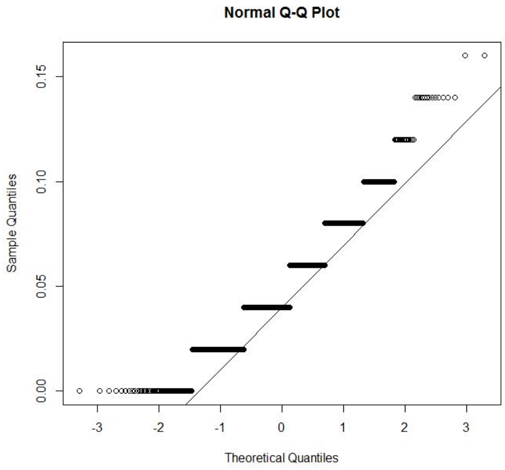


As you can see the results are approximately normally distributed creating a bell shaped graph in the histogram. We can also see how the mean of .5 is very dominant and has a relatively small standard error.

We then repeated the test but this time I changed the probability to .05 .

```
> set.seed(15144127)
> for(i in 1:simreps){phat[i] =mean(rbinom(n=50, size=1, prob=.05))}
> hist(phat)
> qqnorm(phat);qqline(phat)
```

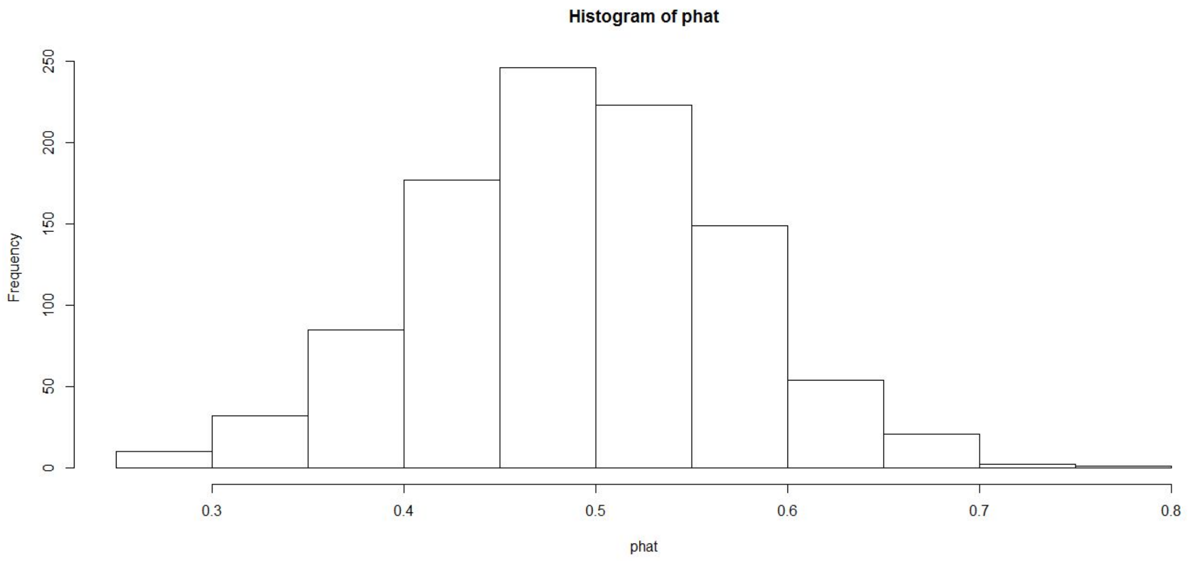


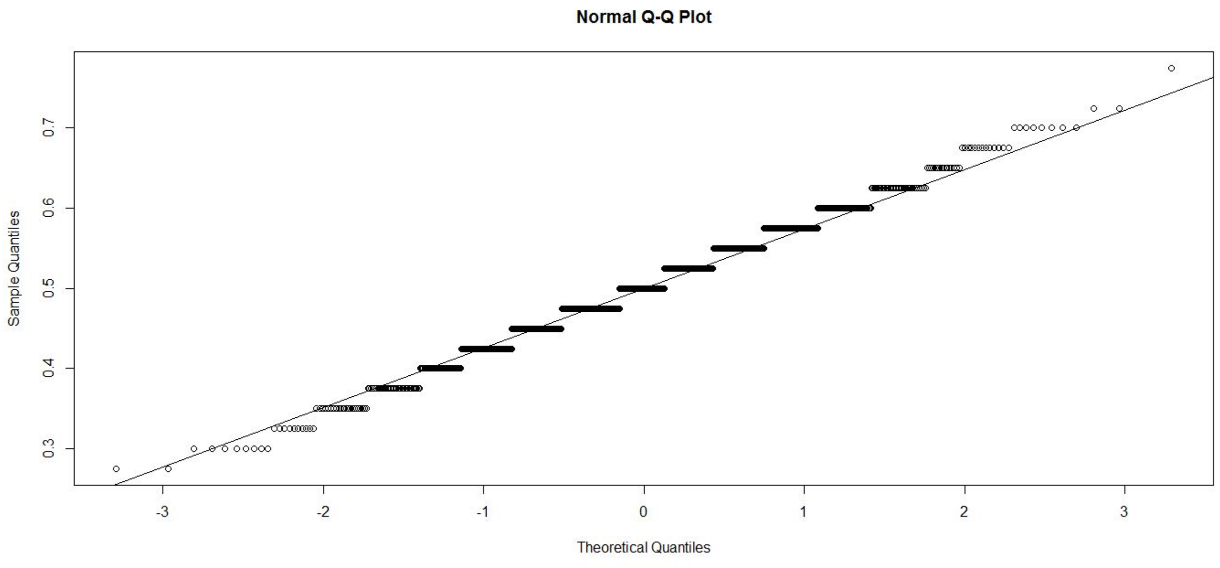


The shape of the graphs for this is different due to the much lower probability and also not normally distributed due to the mean value being just below .05 for the graph.

We then continued to repeat the first simulation test but this time change n to 40.

```
> set.seed(15144127)
> for(i in 1:simreps){phat[i] = mean(rbinom(n=40, size=1, prob=.5))}
> hist(phat)
> qqnorm(phat);qqline(phat)
```

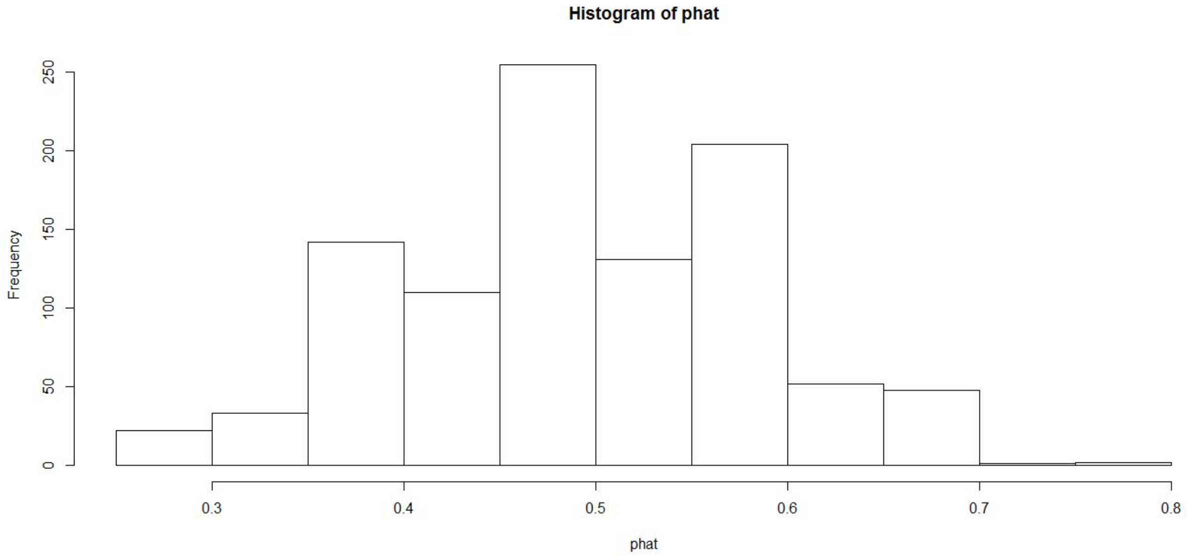


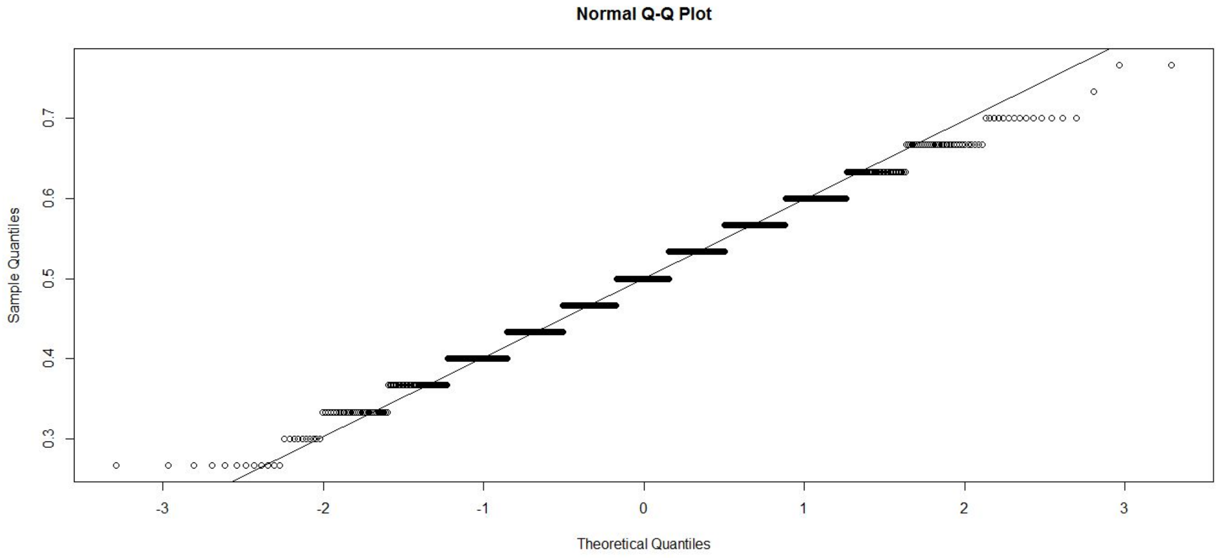


As we can see the graphs are still showing normal distribution around the mean of .5 due to the sufficiently high n value which gives us a lower standard error.

Again we ran the simulation but this time we lowered n again to 30, the value at which the central limit theorem typically works well above.

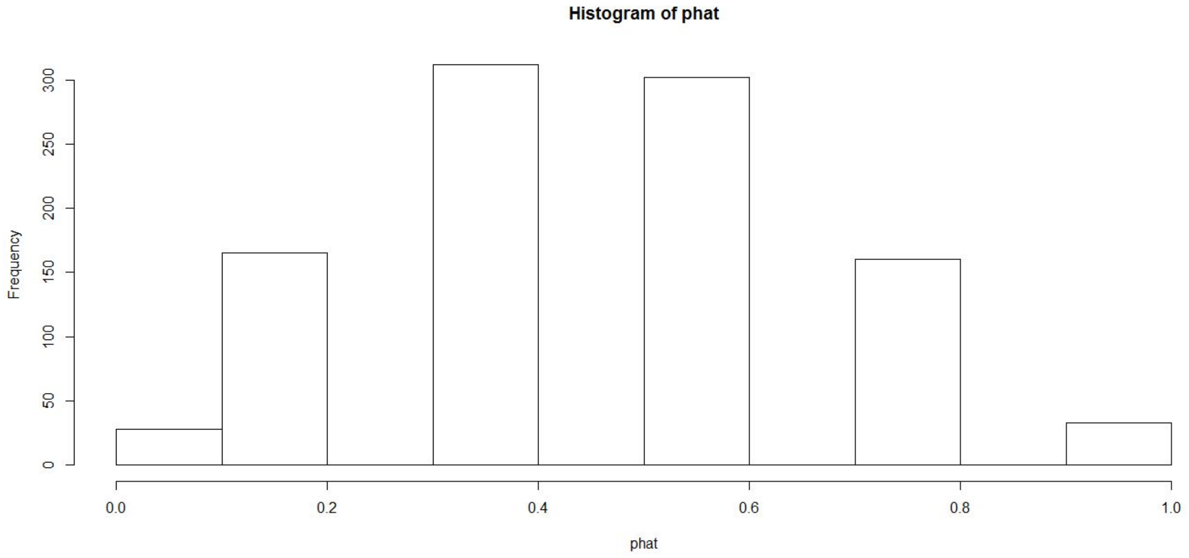
```
> set.seed(15144127)
> for(i in 1:simreps){phat[i] =mean(rbinom(n=30, size=1, prob=.5)) }
> hist(phat)
> qqnorm(phat);qqline(phat)
```

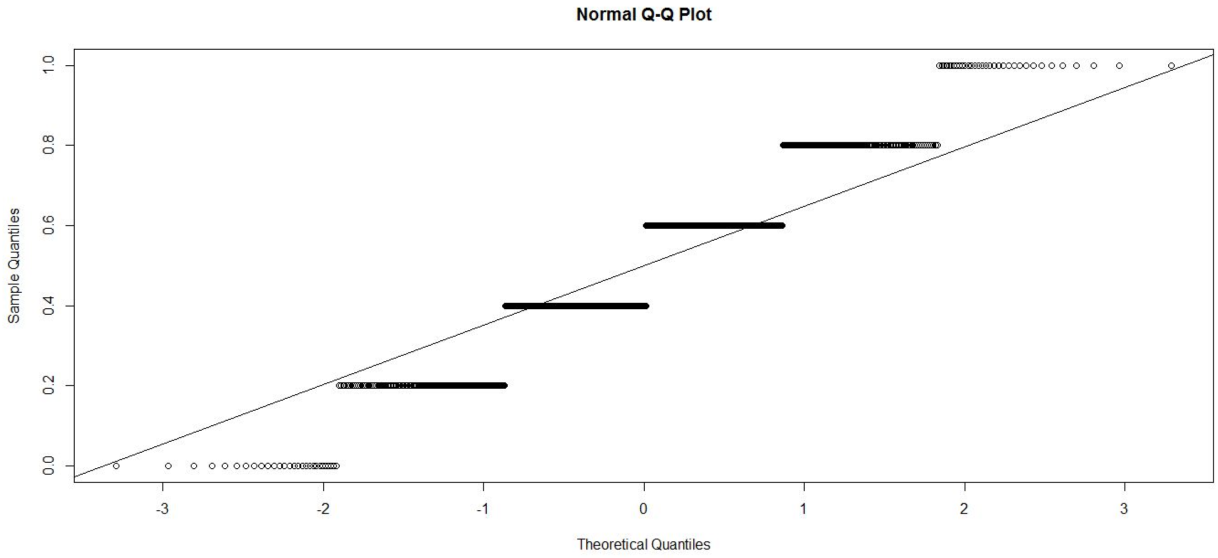




Here we can start to see how the standard error is getting higher as the graphs are starting to lose shape and normal distribution is not as clear.
To finish the test we decided to drop the n value to 5 in the hopes we would lose normal distribution and sufficiently convince ourselves on the central limit theorem.

```
> set.seed(15144127)
> for(i in 1:simreps){phat[i] = mean(rbinom(n=5, size=1, prob=.5)) }
> hist(phat)
> qqnorm(phat);qqline(phat)
```





As expected the graphs are very different now, no normal distribution can be seen and it is obvious that there is a high standard error which was to be expected due to the formula σ(X) = σ/√n which shows how the higher the n value the lower the Standard Error.

This helps show the central limit theorem as it proves to us how regardless of the distribution, the sample mean, p-hat, has a normal distribution when the sample size, n, is large.