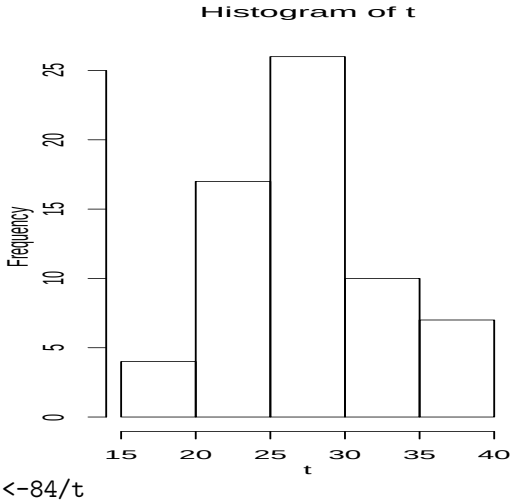
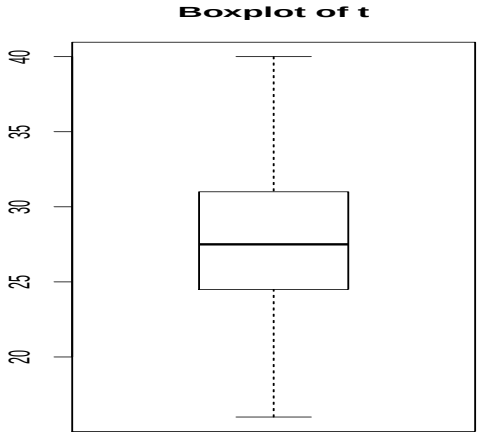
Illustration of Elementary Statistical Methods

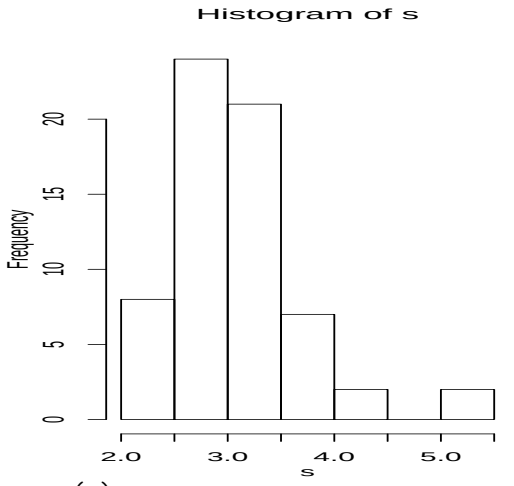
1. Simon Newcomb measured the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 8400 meters. The experiment was repeated 64 times and the data file named `speed_of_light.data` contains these 64 time measurements in micro-seconds (1 micro-second=10^{-6} second). Based on this experimental data, we shall like to get a confidence interval for the mean speed of light and verify whether there is sufficient evidence against the now accepted mean speed of light of 3×10^5 km/sec.

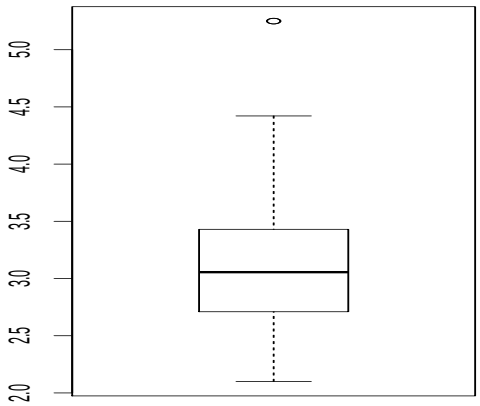
```
> d1<-read.table("speed_of_light.data")
> t<-d1$$V1
t
 [1] 28 22 36 26 28 28 26 24 32 30 27 24 33 21 36 32 31 25 24 25 28 36 27 32 34
[26] 30 25 26 26 25 23 21 30 33 29 27 29 28 22 26 27 16 31 29 36 32 28 40 19 37
[51] 23 32 29 24 25 27 24 16 29 20 28 27 39 23
> hist(t)
> boxplot(t)
```



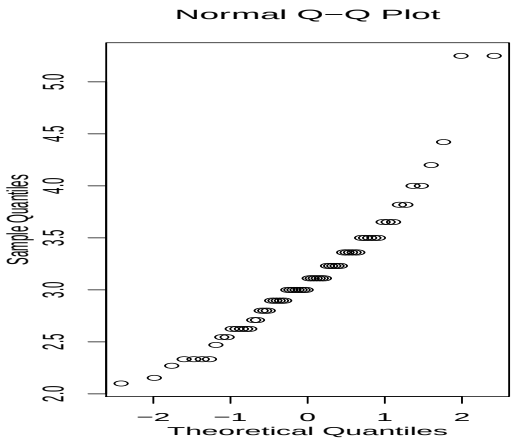


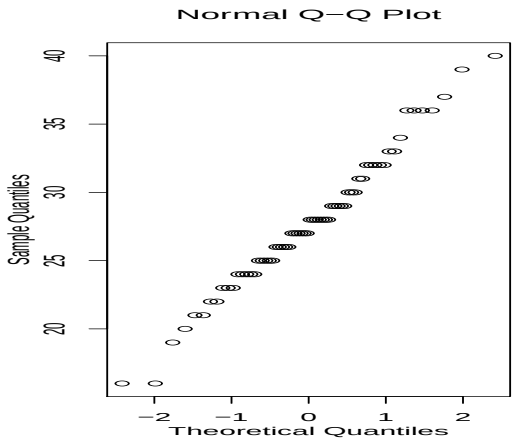

```
> s<-84/t
> hist(t)
> boxplot(t)
```



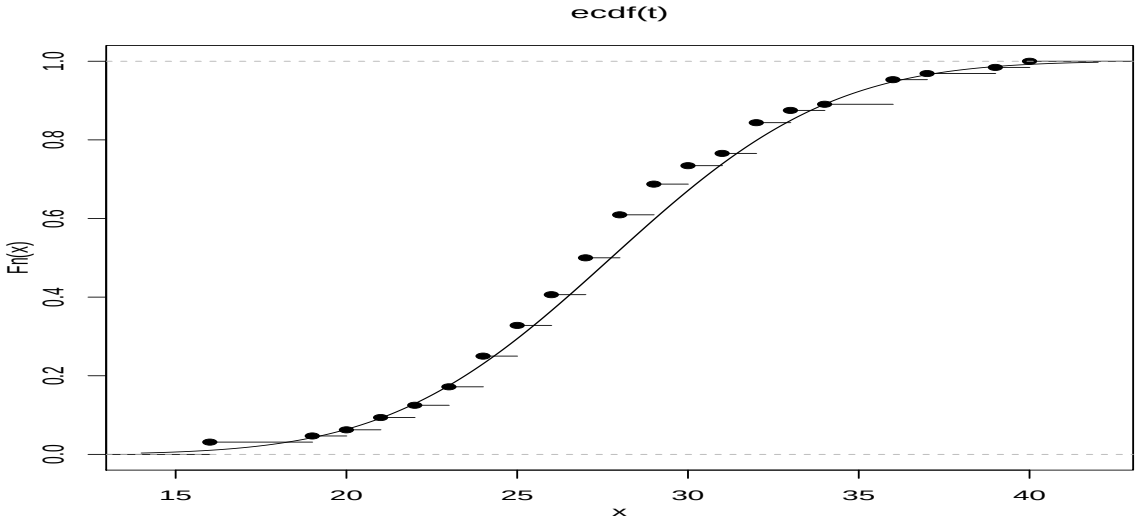



```
> qqnorm(s)
> qqnorm(t)
```






```
> range(t)
[1] 16 40
> mean(t)
[1] 27.75
> sd(t)
[1] 5.083431
> plot(ecdf(t))
> lines(seq(14,42,0.05),pnorm(seq(14,42,0.05),27.75,5.083431))
```

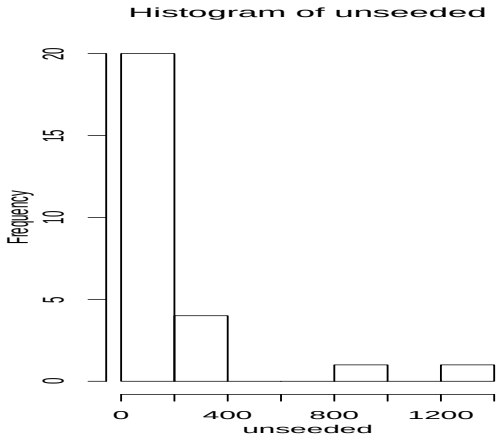


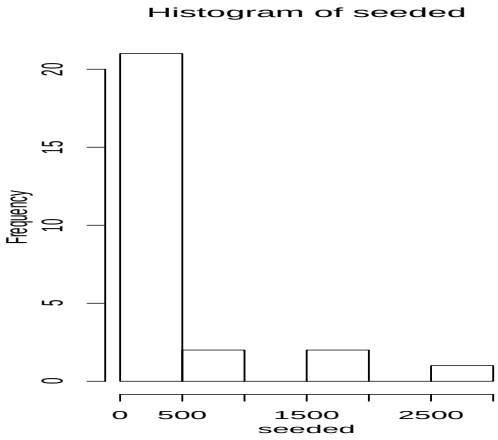

```
> normtest<-function(x)
+ {
+  library(nortest)
+  s<-shapiro.test(x)
+  ad<-ad.test(x)
+  cvm<-cvm.test(x)
+  ll<-lillie.test(x)
+  sf<-sf.test(x)
+  df<-data.frame(Method=c(s$method, ad$method, cvm$method, ll$method, sf$method),
+  P.Value=c(s$p.value, ad$p.value, cvm$p.value, ll$p.value, sf$p.value))
+  df
+ }
```

```
> normtest(s)
                                        Method      P.Value
1                    Shapiro-Wilk normality test 0.0004692483
2                Anderson-Darling normality test 0.0075964850
3                Cramer-von Mises normality test 0.0185316038
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.0460139178
5                 Shapiro-Francia normality test 0.0006860844

> normtest(t)
                                        Method   P.Value
1                    Shapiro-Wilk normality test 0.6082205
2                Anderson-Darling normality test 0.3908665
3                Cramer-von Mises normality test 0.3321172
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.2178599
5                 Shapiro-Francia normality test 0.5533636

> mu0<-(8.4/300000)*10^6
> mu0
[1] 28

> t.test(t,mu=mu0)
        One Sample t-test
data:  t
t = -0.3934, df = 63, p-value = 0.6953
alternative hypothesis: true mean is not equal to 28
95 percent confidence interval:
 26.48020 29.01980
sample estimates:
mean of x
    27.75

> wilcox.test(t,mu=mu0)
        Wilcoxon signed rank test with continuity correction
data:  t
V = 759.5, p-value = 0.5964
alternative hypothesis: true location is not equal to 28
```

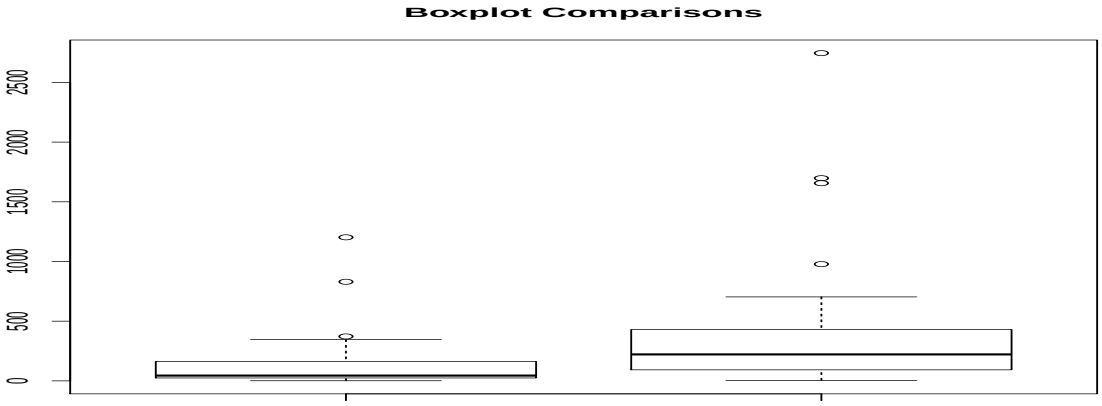
2. Some clouds were randomly seeded with silver nitrate while some were not. File `cloud_seeding.data` contains the rainfall volumes (in acre-feet) from both these clouds. Based on this data we are to determine whether cloud seeding by silver nitrate increases rainfall or not.

```
> d2<-read.table("cloud_seeding.data",header=T)
> d2
   Unseeded_Clouds Seeded_Clouds
1           1202.6        2745.6
2            830.1        1697.8
.................................
.................................
25             4.9           7.7
26             1.0           4.1
> unseeded<-d2$Unseeded_Clouds
> seeded<-d2$Seeded_Clouds
> hist(unseeded)
> hist(seeded)
```

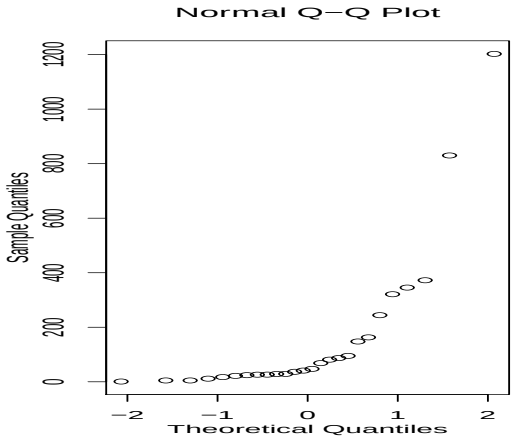


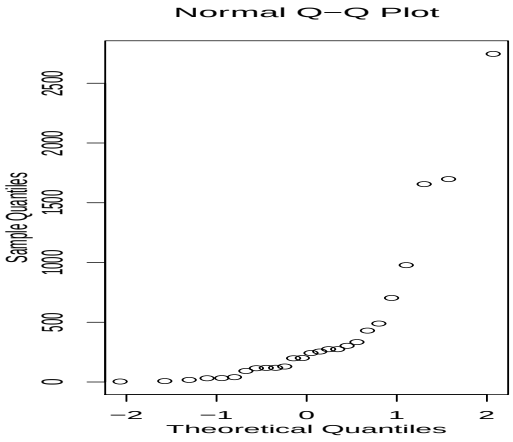



```
> boxplot(unseeded,seeded)
```

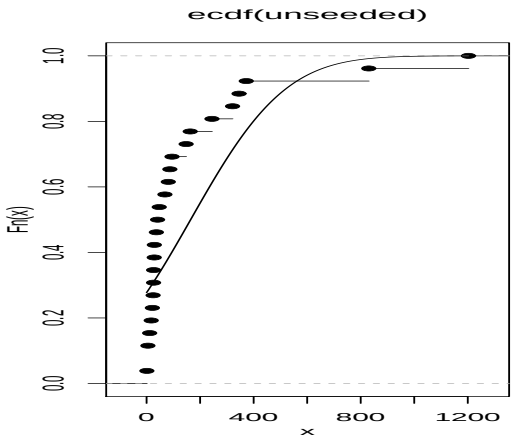


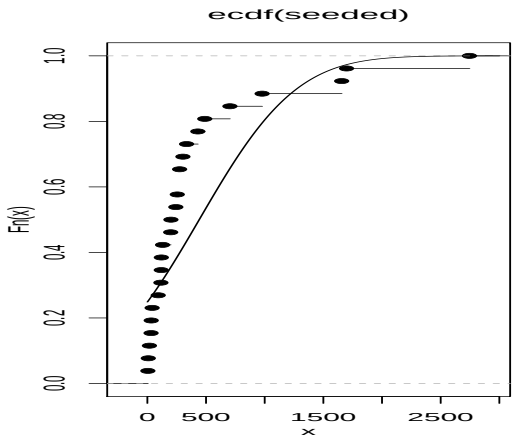

```
> qqnorm(unseeded)
> qqnorm(seeded)
```



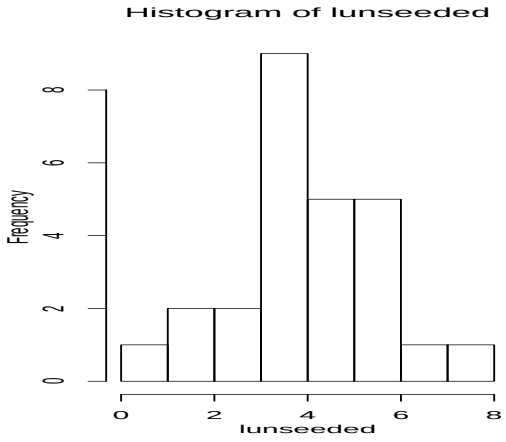


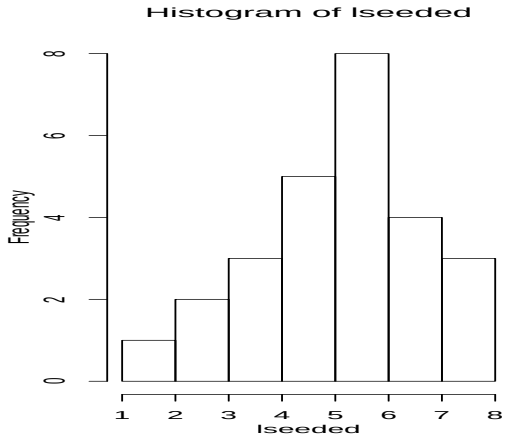

```
> plot(ecdf(unseeded))
> lines(0:1205,pnorm(0:1205,mean(unseeded),sd(unseeded)))
> plot(ecdf(seeded))
> lines(0:3000,pnorm(0:3000,mean(seeded),sd(seeded)))
```



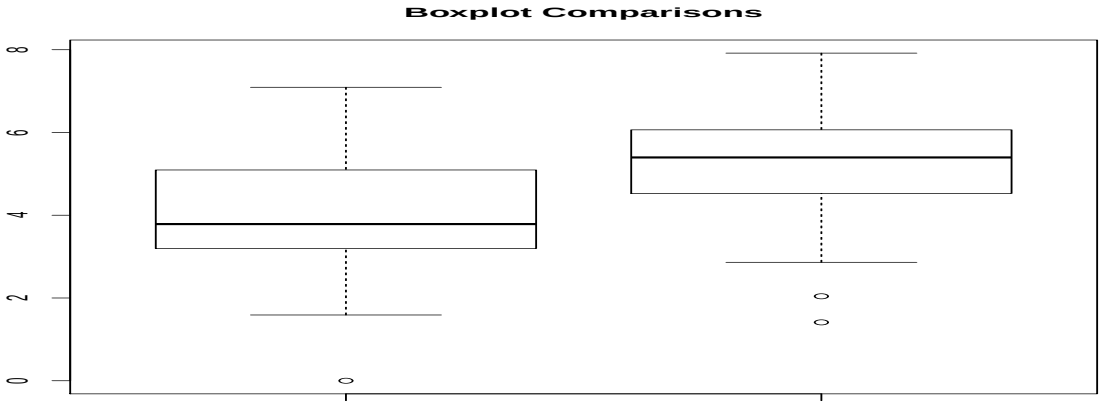



```
> normtest(unseeded)
                                      Method      P.Value
1                   Shapiro-Wilk normality test 3.131400e-07
2               Anderson-Darling normality test 9.048556e-10
3               Cramer-von Mises normality test 4.692917e-08
4 Lilliefors (Kolmogorov-Smirnov) normality test 4.649066e-06
5                Shapiro-Francia normality test 1.467887e-06

> normtest(seeded)
                                      Method      P.Value
1                   Shapiro-Wilk normality test 1.411025e-06
2               Anderson-Darling normality test 9.605483e-09
3               Cramer-von Mises normality test 1.276676e-07
4 Lilliefors (Kolmogorov-Smirnov) normality test 2.614762e-06
5                Shapiro-Francia normality test 5.359383e-06

> fligner.test(seeded,unseeded)
        Fligner-Killeen test of homogeneity of variances
data:  seeded and unseeded
Fligner-Killeen:med chi-squared = 24.6561, df = 24, p-value = 0.4246

> wilcox.test(seeded,unseeded,alt="g")
        Wilcoxon rank sum test with continuity correction
data:  seeded and unseeded
W = 473, p-value = 0.006916
alternative hypothesis: true location shift is greater than 0

> lseeded<-log(seeded)
> lunseeded<-log(unseeded)

> hist(lunseeded)
> hist(lseeded)
```

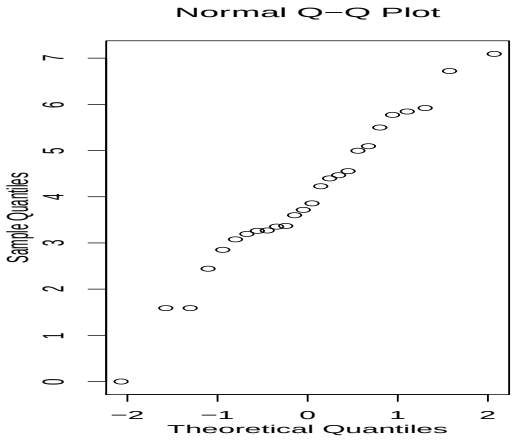


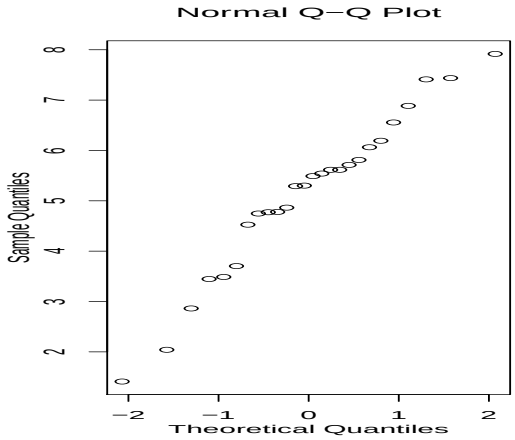



```
> boxplot(lunseeded,lseeded)
```

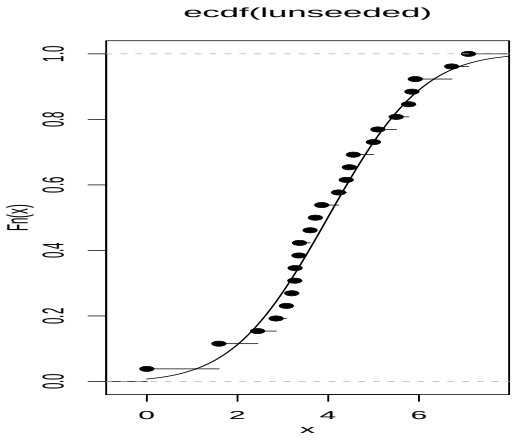


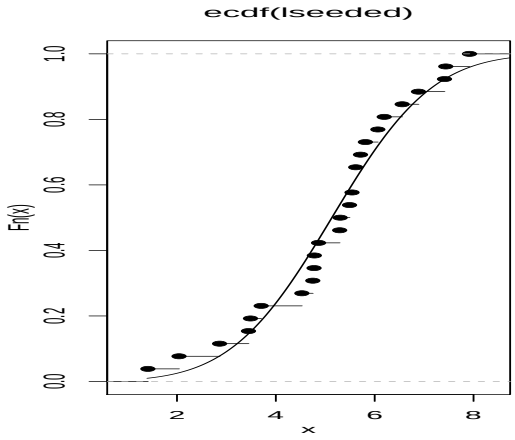

```
> qqnorm(lunseeded)
> qqnorm(lseeded)
```






```
> plot(ecdf(lunseeded))
> lines(seq(0,8,0.05),pnorm(seq(0,8,0.05),mean(lunseeded),sd(lunseeded)))
> plot(ecdf(lseeded))
> lines(seq(1.4,9,0.05),pnorm(seq(1.4,9,0.05),mean(lseeded),sd(lseeded)))
```



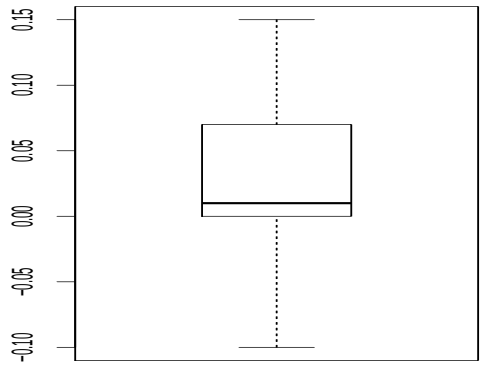


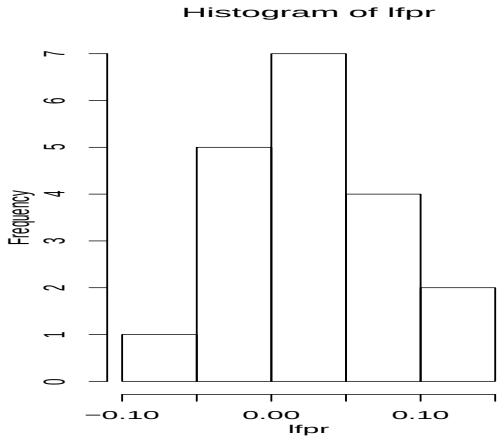

```
> normtest(lseeded)
                                      Method   P.Value
1                   Shapiro-Wilk normality test 0.5207568
2               Anderson-Darling normality test 0.3986896
3               Cramer-von Mises normality test 0.2895569
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.2549802
5                Shapiro-Francia normality test 0.4662411

> normtest(lunseeded)
                                      Method   P.Value
1                   Shapiro-Wilk normality test 0.8726644
2               Anderson-Darling normality test 0.7717395
3               Cramer-von Mises normality test 0.7284970
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.7676689
5                Shapiro-Francia normality test 0.7280068

> var.test(lseeded,lunseeded)
        F test to compare two variances
data:  lseeded and lunseeded
F = 0.9491, num df = 25, denom df = 25, p-value = 0.8971
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4255461 2.1167714
sample estimates:
ratio of variances
         0.9490963

> t.test(lseeded,lunseeded,var.equal=T,alt="g")
        Two Sample t-test
data:  lseeded and lunseeded
t = 2.5444, df = 50, p-value = 0.007041
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3904045       Inf
sample estimates:
mean of x mean of y
 5.134187  3.990406
```

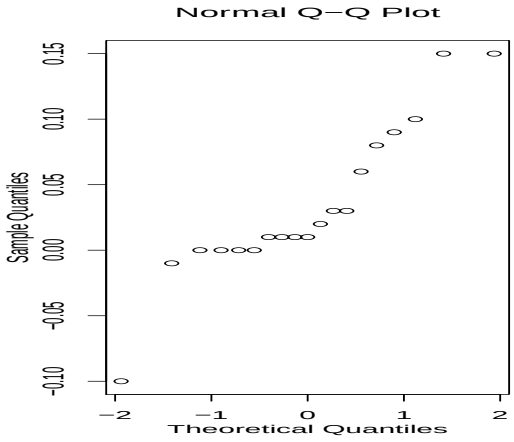
```
> ks.test(seeded,unseeded,alternative="l")
        Two-sample Kolmogorov-Smirnov test
data:  seeded and unseeded
D^- = 0.4231, p-value = 0.009525
alternative hypothesis: the CDF of x lies below that of y

> ks.test(lseeded,lunseeded,alternative="l")
        Two-sample Kolmogorov-Smirnov test
data:  lseeded and lunseeded
D^- = 0.4231, p-value = 0.009525
alternative hypothesis: the CDF of x lies below that of y
```

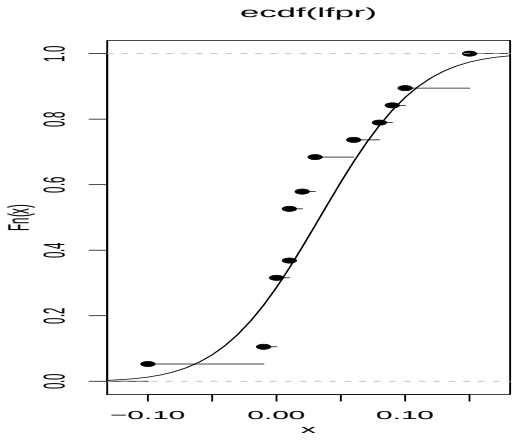
3. File `lfpr.data` contains labor force participation rate (LFPR) of women in 19 cities in the United States in the years 1968 and 1972. Based on this we are to determine whether LFPR of women has increased.over these 4 years.

```
> d3<-read.table("lfpr.data",header=T)
> d3
            City X1972 X1968
1           N.Y.  0.45  0.42y
2           L.A.  0.50  0.50
.............................
.............................
18     Patterson  0.57  0.56
19        Dallas  0.64  0.63

> lfpr<-d3$X1972-d3$X1968
> boxplot(lfpr)
> hist(lfpr)
> qqnorm(lfpr)
> plot(ecdf(lfpr))
> lines(seq(-0.15,0.20,0.01),pnorm(seq(-0.15,0.20,0.01),mean(lfpr),sd(lfpr)))
```










```
> normtest(lfpr)
                                        Method     P.Value
1                    Shapiro-Wilk normality test 0.045034838
2                Anderson-Darling normality test 0.012726926
3                Cramer-von Mises normality test 0.009219975
4 Lilliefors (Kolmogorov-Smirnov) normality test 0.028738967
5                 Shapiro-Francia normality test 0.030783310

> wilcox.test(lfpr,alt="g")
        Wilcoxon signed rank test with continuity correction
data:  lfpr
V = 104, p-value = 0.006622
alternative hypothesis: true location is greater than 0

> wilcox.test(d3$X1972,d3$X1968,paired=T,alt="g")
Wilcoxon signed rank test with continuity correction
data:  d3$X1972 and d3$X1968
V = 104, p-value = 0.006622
alternative hypothesis: true location shift is greater than 0

> wilcox.test(d3$X1972,d3$X1968,alt="g")
        Wilcoxon rank sum test with continuity correction
data:  d3$X1972 and d3$X1968
W = 237, p-value = 0.05059
alternative hypothesis: true location shift is greater than 0
```

4. A group of dolphin was observed off the coast of Iceland near Keflavik in 1998. The time of the day, namely Morning, Noon, Afternoon and Evening; and the main activity of the group at that time, namely Feeding, Travelling or Socializing, were observed on different occassions. Data file `dolphin.data` contains observations on these two variables for each of these occassions. Based on these data we are to determine whether the dolphin activities are significantly different from one time of the day to another.

```
> d4<-read.table("dolphin.data")
Travel  Morning
Feed    Noon
.................
.................
Social  Afternoon
Feed    Evening
```

```
> table(d4)
        V2
V1       Afternoon Evening Morning Noon
  Feed           0      56      28    4
  Social         9      10      38    5
  Travel        14      13       6    6

> chisq.test(table(d4))
        Pearson's Chi-squared test
data:  table(d4)
X-squared = 68.4646, df = 6, p-value = 8.44e-13

> chisq.test(table(d4),simulate.p.value=T)
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)
data:  table(d4)
X-squared = 68.4646, df = NA, p-value = 0.0004998
```

5. Since the 1978/79 India in Pakistan series till the DLF Cup 2006, leaving aside the abandoned matches, India has played 89 One Day International (ODI) Cricket matches against Pakistan, in which India has batted first 41 times. Among these 41 matches, India has won against Pakistan 16 times, while in the remaining 48 matches, in which India has tried to chase Pakistan's Total, it has won 18 times. (Source: `http://thatscricket.oneindia.in`) Thus on the surface it appears that batting first might have a slight advantage of winning for India against Pakistan (16/41=0.3902) compared to a chase (18/48=0.375). The question is, is this difference in probability of winning while batting first, statistically significant?

```
> cricket<-matrix(c(16,25,18,30),nrow=2,ncol=2,byrow=T,dimnames=list(c("Batted First",
  "Chased"), c("Won","Lost")))

> cricket
             Won Lost
Batted First  16   25
Chased        18   30

> fisher.test(cricket,alt="g")
        Fisher's Exact Test for Count Data
data:  cricket
p-value = 0.5278
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.4752376       Inf
sample estimates:
odds ratio
  1.065883
```