STAT 3480 Lab1

Shaoran Sun

February 3, 2016

The binomial test

Step 1

H_0 : Smokers start smoking before the age of 18.

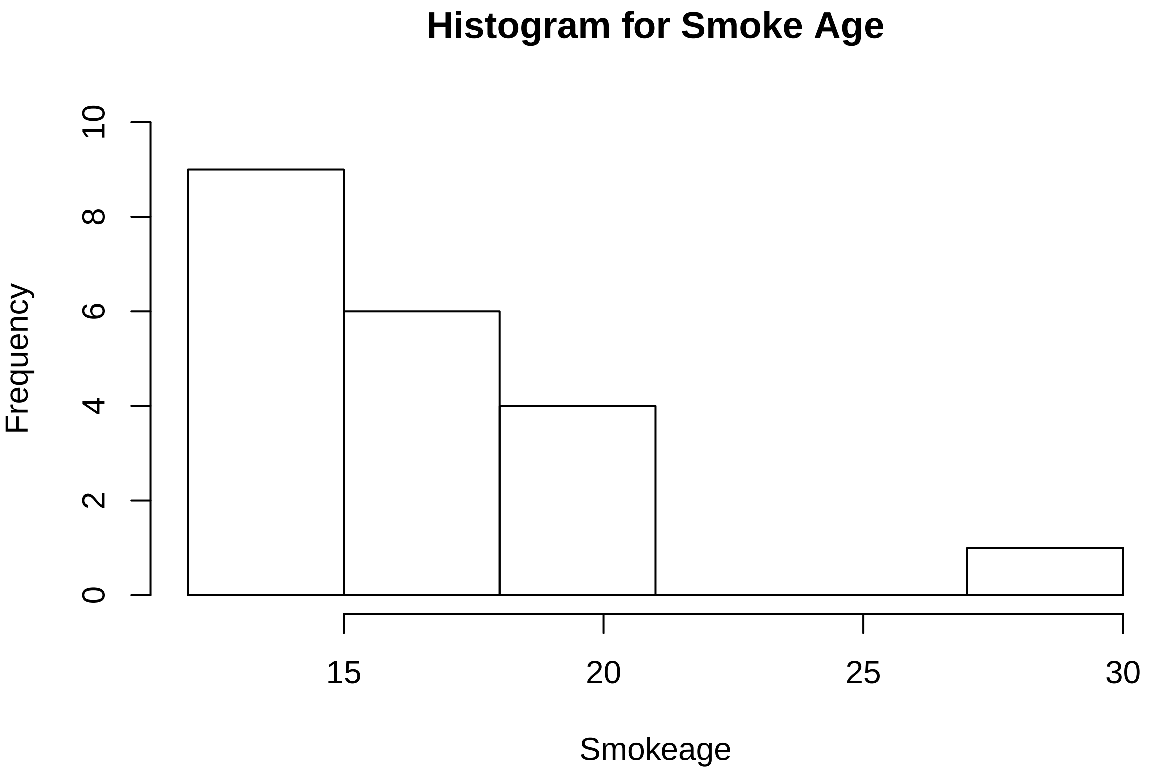
H_a : Smokers start smoking after the age of 18.

Data has mean 16.55, standard deviation 4.006245. Applying one-sample t-test, $t = \frac{\bar{X}-\mu}{S}\sqrt{n}$. We get $t = -1.6186$, $p - value = 0.122$. The $p - value$ is 0.122, which is greater than 0.05.

Therefore, we fail to reject our null hepothesis and conclude that there is not enough evidence of smokers start smoking before the age of 18.

Step 2

```r
Smokeage = c(18,19,30,16,17,15,14,14,17,12,14,13,19,19,17,13,20,12,17,15)
hist(Smokeage, main="Histogram for Smoke Age",  breaks=c(12, seq(15,30,3)),ylim=c(0,10))
```



The majority of people start to smoke before age 18. The assumption may not met because the graph is not normally distributed and the sample size is not large enough.

Step 3

$Z_B = \frac{B-n(.5)}{\sqrt{n(.5)(1-.5)}}$
$= \frac{B-n(.5)}{\sqrt{n(.25)}}$
$= 1.7889$

$P_Z = 0.03681914$

H_0 : Smokers start smoking before the age of 18. H_a : Smokers start smoking after the age of 18.

14 out of 20 data values are less than 18. Afrer applying binomial test, we obtain $Z_B = 1.7889$ and $P_Z = 0.03682$. Thus, we can reject null hypothesis with 95% confidence interval.

Step 4

Binomial is more accurate summarizing the data because as we can see from the data histogram, the data is very skewed to right, not normally distributed. Whereas binomial test does not need any assumption from the data, so that the data is more accurately described.

The binomial test in R

Step 5

After putting in command provided, we get p-value of 0.03178

Step 6

If we take infinite many samples from the poplulation, 95% of the time the true median is upper-bounded by 17.21.

Step 7

With 99% confidence interval two-sided test, we get lower-bound at 13.8033, upper-bound at 19. Again, if we take infinite many samples within the population, 99% of the chance that the real median is within this interval.

Step 8

With different alternative each time, the lower-bound is negative infinite with "less", and the upper-bound is positive infinite with "greater".

H_0 : The median of smokers start smoking is 18. H_a : The median of smokers start smoking is less then 18. The $p-value$ is **0.03178**. We fail to reject the null hypothesis at a .01 significance level, and conclude that there is not enough evidence to show that the mean of the smoke age is not less than 18.

H_0 : The median of smokers start smoking is 18. H_a : The median of smokers start smoking is greater then 18. The $p-value$ is **0.9904**. We reject the null hypothesis at a .01 significance level, and conclude that the mean of the smoke age is greater than 18.

H_0 : The median of smokers start smoking is 18. H_a : The median of smokers start smoking is not 18. The $p-value$ is **0.06357**. We fail to reject the null hypothesis at a .01 significance level, and conclude that there is not enough evidence to show that the mean of the smoke age is not equal to 18. ##Step 9 ###Part 1

```
library(BSDA)
```

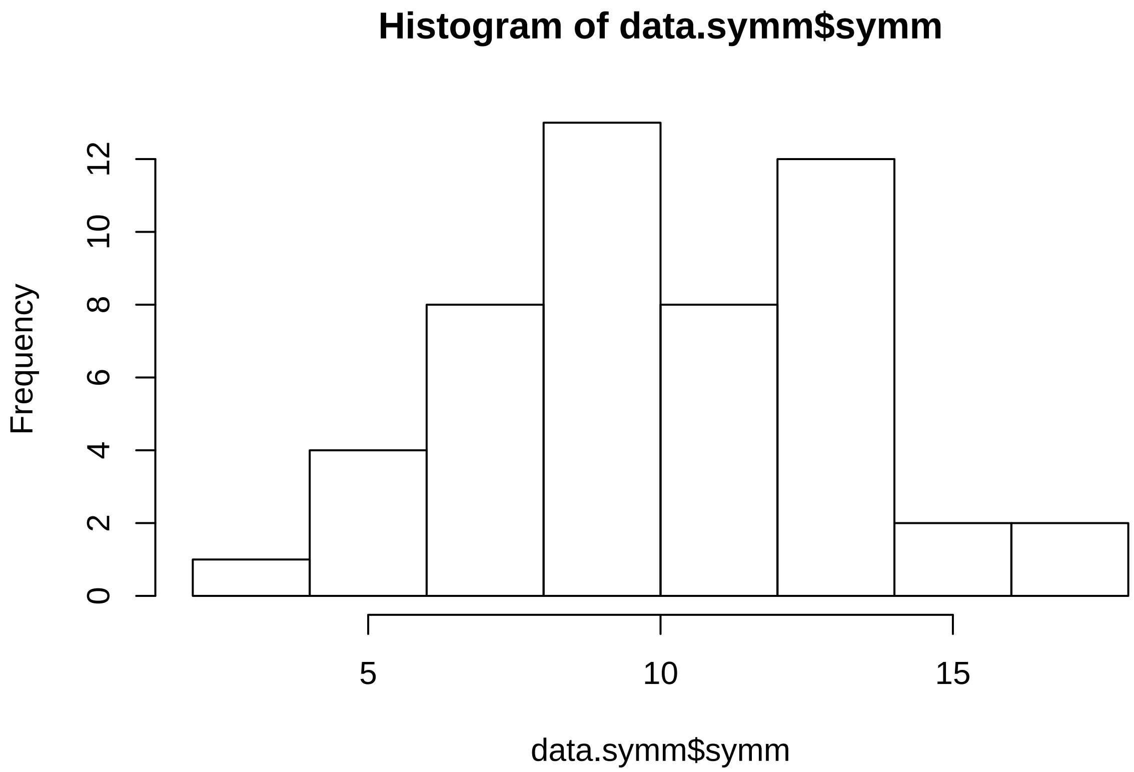
```
## Loading required package: e1071

## Loading required package: lattice

##
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':
##
##     Orange
```

```
data.symm <- read.csv("~/Desktop/data.symm.txt", sep="")
hist(data.symm$symm)
```



The shape of the histogram is nearly symmetric, therefore the median and mean are approximately the same.

Part 2

With One Sample t-test, we have $p-value = 7.571e-06$. H_0 : The true mean of data is 8. H_a : The true mean of data is greater than 8. With 95% confidence interval, we have $(9.3509, \infty)$. We can reject the null hypothesis at 0.05 significant level, and conclude that the true mean of data is greater than 8.

Part 3

With One Sample binomial-test, we have $p-value = 0.0004681$. H_0 : The true median of data is 8. H_a : The true median of data is greater than 8. With 95% confidence interval, we have $(9.2781, \infty)$. We can reject the null hypothesis at 0.05 significant level, and conclude that the true median of data is greater than 8.

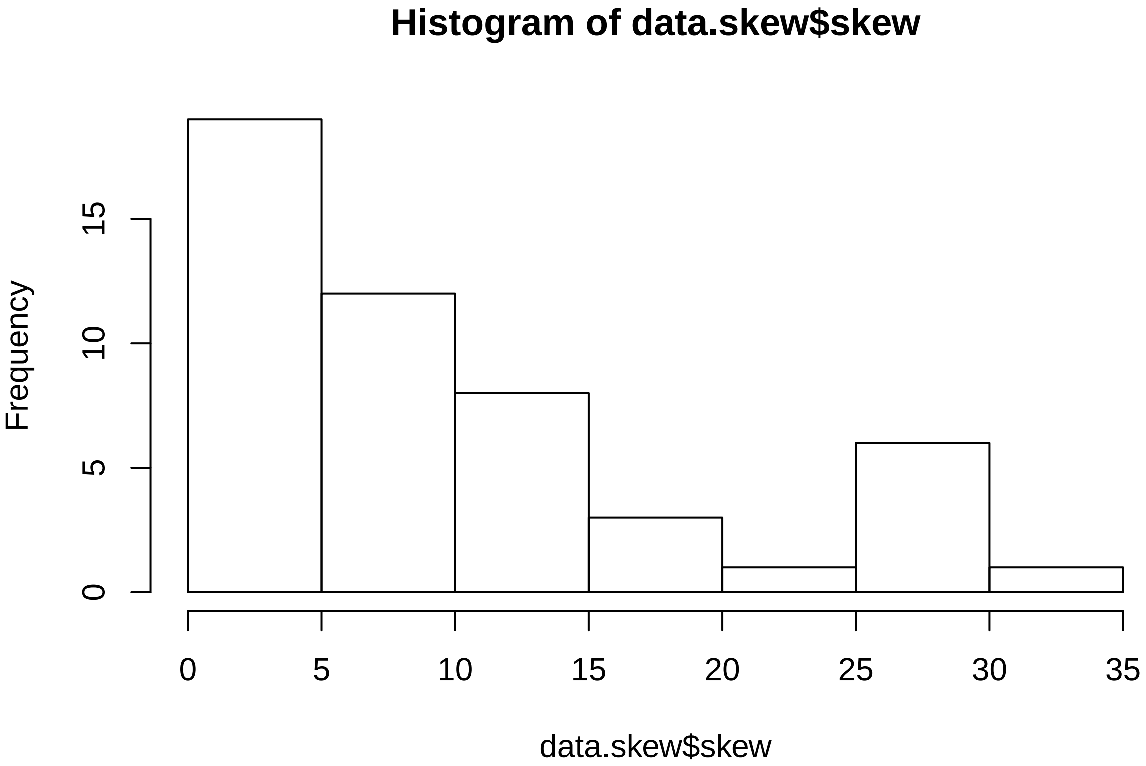
Part 4

Both tests give the same result that the true mean/median is greater than 8. Since the graph is symmetric, the true mean and median fall in the same point.

Step 10

Part 1

```
data.skew <- read.csv("~/Desktop/data.skew.txt", sep="")
hist(data.skew$skew)
```



The shape of the histogram is skewed to the right, therefore the median should be to the left of mean.

Part 2

With One Sample t-test, we have $p - value = 0.0703$. H_0 : The true mean of data is 8. H_a : The true mean of data is greater than 8. With 95% confidence interval, we have $(7.7773, \infty)$. We fail to reject the null hypothesis at 0.05 significant level, and conclude that there is not enough evidence to show that the mean of the data is greater than 8.

Part 3

With One Sample binomial-test, we have $p - value = 0.4439$. H_0 : The true median of data is 8. H_a : The true median of data is greater than 8. With 95% confidence interval, we have $(5.0228, \infty)$. We fail to reject the null hypothesis at 0.05 significant level, and conclude that there is not enough evidence to show that the mean of the data is greater than 8.

Part 4

Both tests give the same result there is not enough evidence to show that the mean of the data is greater than 8. However, binomial test shows a stronger rejection with $p - value$ at 0.4439, comparing to t-test's result is almost at he edge of rejecting the null hypothesis.

Lab summary

Based on the two test I applied to the data sample of 20 smokers and the age they started smoking. There are 14 out of 20 samples started smoking before the age of 18. The two tests are t-test and binomial test. T-test assumes a relatively large sample size, and data being normally distributed. Binomial test does not require any assumption of the sample size or data's distribution.

With t-test, we get $p - value$ of 0.122. $p - value$ is a statist that describe how likely the event happens. We make hypothesis that smokers start smoking before the age of 18. We usually consider 95% confident interval, which means that 95% of the chance, the population mean/median is within our desired range. In this case, 0.122 is larger than 0.05. Therefore, we do not have enough evidence to show that smokers start smoking before the age of 18. However, with binomial test, we get $p - value$ of 0.03178. This is smaller than our siginificant level of 0.05, and conclude that smokers start smoking before the age of 18.

Binomial test is a better choice for this dataset, because we do not require any assumption to the data. So that it is more accurate to use such test, and conclude that, indeed, smokers sart smoking before the age of 18.

Appendix

Step 1 Code

```
smokeage = c(18,19,30,16,17,15,14,14,17,12,14,13,19,19,17,13,20,12,17,15)
xbar <- mean(smokeage)
s <- sd(smokeage)
a <- 18
n <- 20
t <- (xbar-a)/(s/sqrt(n))
t <- (16.55-18)/4.006254*sqrt(20)
t #t value
```

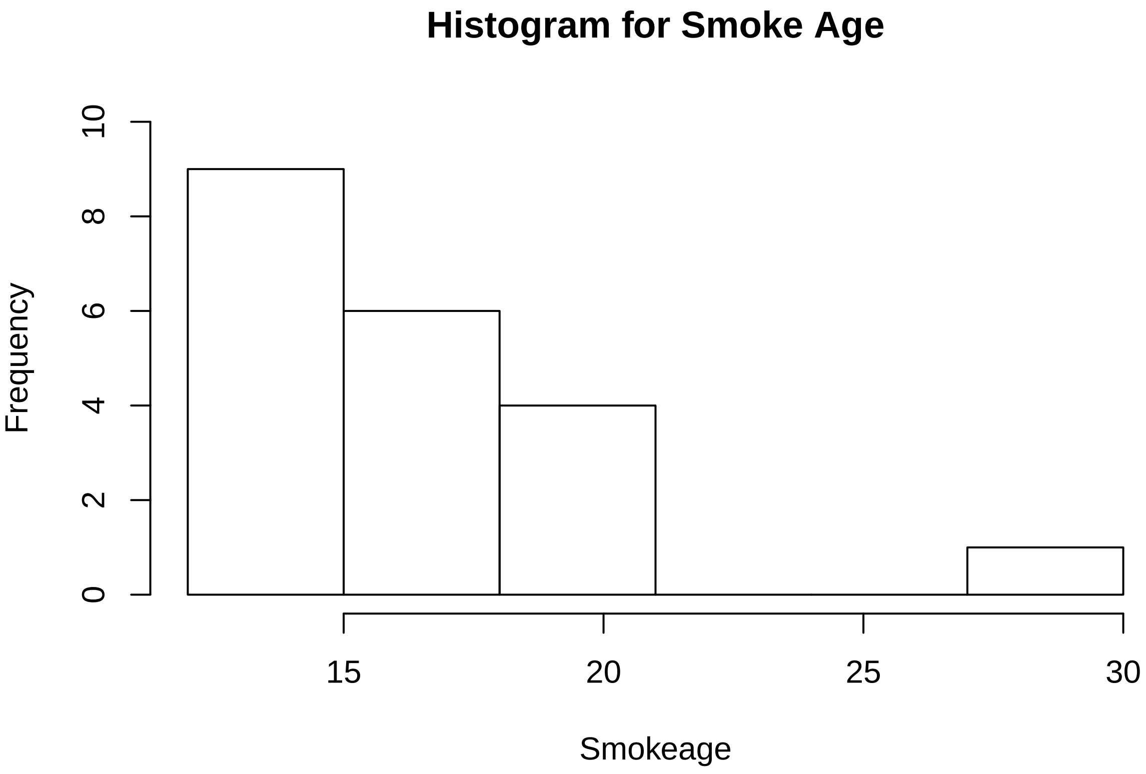
```
## [1] -1.618619
```

```
#t.test(smokeage,mu=18)
2*pt(-abs(t),df=n-1) #p-value
```

```
## [1] 0.1220099
```

Step 2 Code

```
Smokeage = c(18,19,30,16,17,15,14,14,17,12,14,13,19,19,17,13,20,12,17,15)
hist(Smokeage, main="Histogram for Smoke Age",  breaks=c(12, seq(15,30,3)),ylim=c(0,10))
```



Step 3 Code

```
sum(smokeage<18)
```

```
## [1] 14
```

```
pnorm(-abs(4/sqrt(5)))
```

```
## [1] 0.03681914
```

Step 5 Code

```
library(BSDA)
SIGN.test(smokeage, md=18, alternative="less")
```

```
##
##  One-sample Sign-Test
##
## data:  smokeage
## s = 5, p-value = 0.03178
## alternative hypothesis: true median is less than 18
## 95 percent confidence interval:
##     -Inf 17.2072
## sample estimates:
## median of x
##        16.5

##                     Conf.Level L.E.pt  U.E.pt
## Lower Achieved CI     0.9423   -Inf 17.0000
## Interpolated CI       0.9500   -Inf 17.2072
## Upper Achieved CI     0.9793   -Inf 18.0000
```

Step 7 Code

```
library(BSDA)
SIGN.test(smokeage, md=18, alternative="two.sided", conf.level = .99)
```

```
##
##  One-sample Sign-Test
##
## data:  smokeage
## s = 5, p-value = 0.06357
## alternative hypothesis: true median is not equal to 18
## 99 percent confidence interval:
##  13.80328 19.00000
## sample estimates:
## median of x
##        16.5
##                      Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.9882 14.0000     19
## Interpolated CI       0.9900 13.8033     19
## Upper Achieved CI     0.9974 13.0000     19
```

Step 8 Code

```
SIGN.test(smokeage, md=18, alternative="greater", conf.level = .99)
```

```
##
##  One-sample Sign-Test
##
## data:  smokeage
## s = 5, p-value = 0.9904
## alternative hypothesis: true median is greater than 18
## 99 percent confidence interval:
##   14 Inf
## sample estimates:
## median of x
##        16.5
```

```
##                      Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9793     14    Inf
## Interpolated CI       0.9900     14    Inf
## Upper Achieved CI     0.9941     14    Inf
```

```
SIGN.test(smokeage, md=18, alternative="less", conf.level = .99)
```

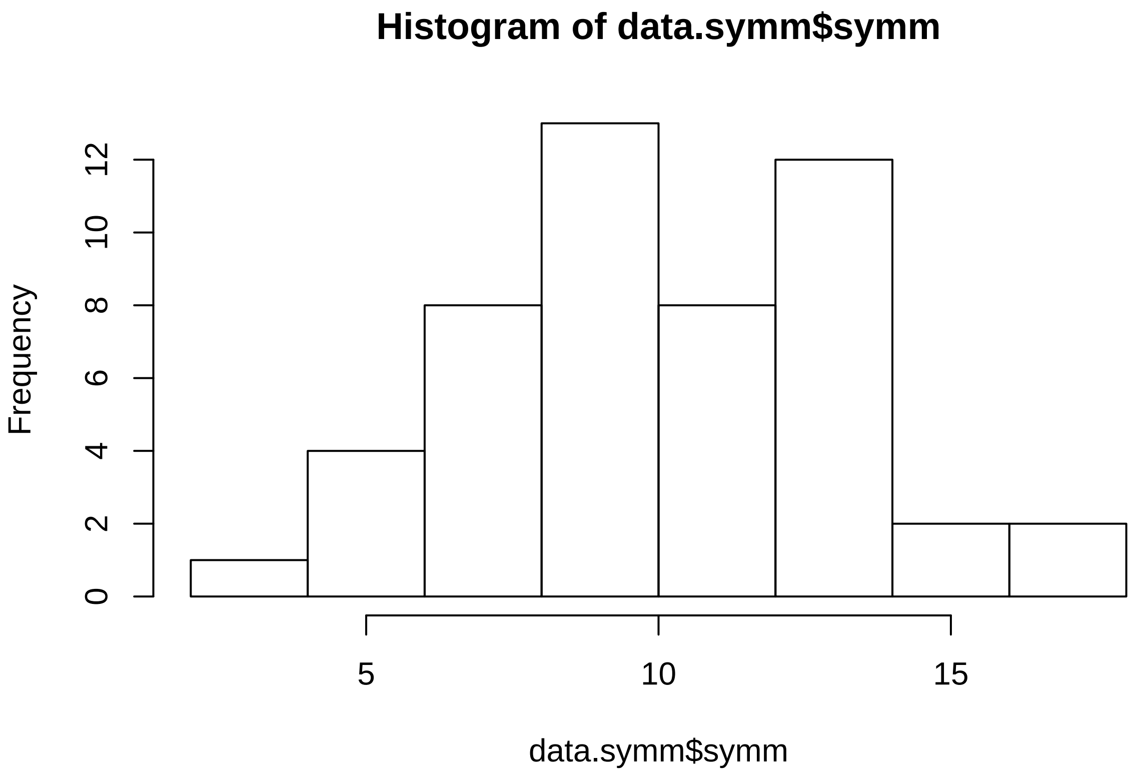
```
##
##  One-sample Sign-Test
##
## data:  smokeage
## s = 5, p-value = 0.03178
## alternative hypothesis: true median is less than 18
## 99 percent confidence interval:
##      -Inf 18.72331
## sample estimates:
## median of x
##        16.5
```

```
##                      Conf.Level L.E.pt  U.E.pt
## Lower Achieved CI     0.9793   -Inf 18.0000
## Interpolated CI       0.9900   -Inf 18.7233
## Upper Achieved CI     0.9941   -Inf 19.0000
```

Step 9 Code

Part 1

```
library(BSDA)
data.symm <- read.csv("~/Desktop/data.symm.txt", sep="")
hist(data.symm$symm)
```



###Part 2

```
library(BSDA)
t.test(data.symm$symm, mu = 8, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  data.symm$symm
## t = 4.8032, df = 49, p-value = 7.571e-06
## alternative hypothesis: true mean is greater than 8
## 95 percent confidence interval:
##  9.350882      Inf
## sample estimates:
## mean of x
##  10.07525
```

Part 3

```
SIGN.test(data.symm$symm,md=8,alternative="greater",conf.level=.95)
```

```
##
```

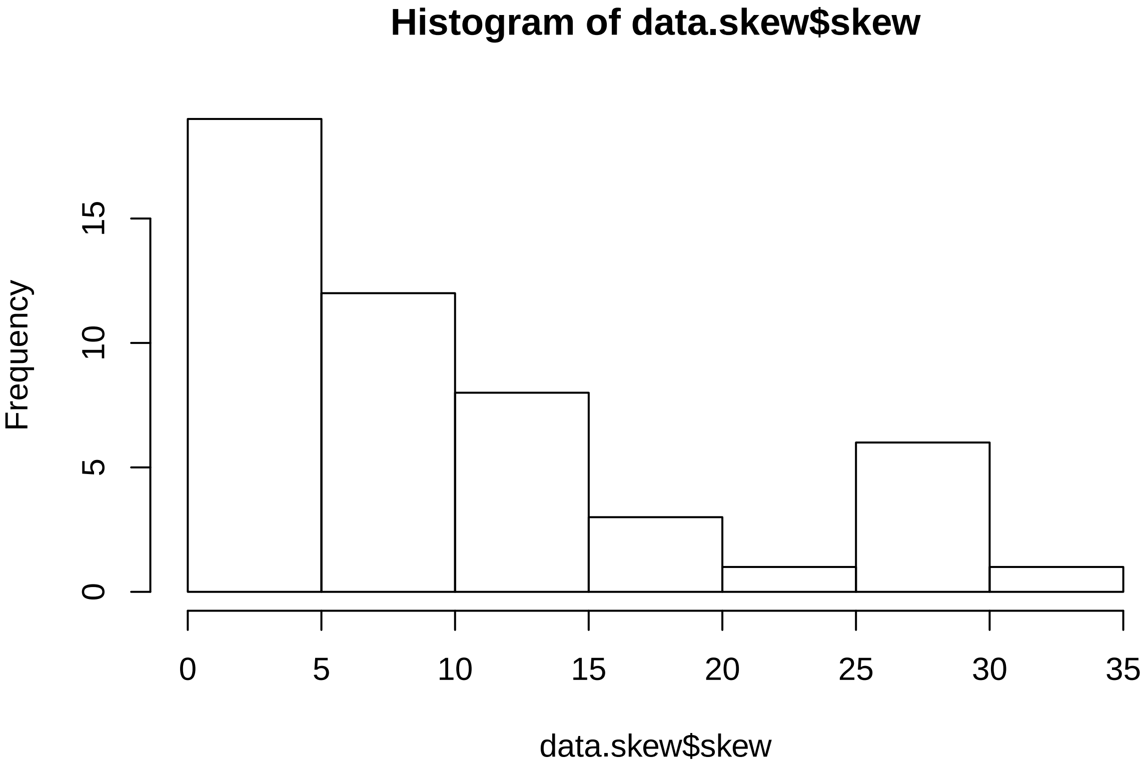
```
##  One-sample Sign-Test
##
## data:  data.symm$symm
## s = 37, p-value = 0.0004681
## alternative hypothesis: true median is greater than 8
## 95 percent confidence interval:
##  9.278101      Inf
## sample estimates:
## median of x
##    9.950267
```

```
##                     Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9405 9.2892    Inf
## Interpolated CI       0.9500 9.2781    Inf
## Upper Achieved CI     0.9675 9.2576    Inf
```

Step 10 Code

Part 1

```
data.skew <- read.csv("~/Desktop/data.skew.txt", sep="")
hist(data.skew$skew)
```



###Part

2

```
t.test(data.skew$skew, mu = 8, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  data.skew$skew
## t = 1.4978, df = 49, p-value = 0.0703
## alternative hypothesis: true mean is greater than 8
## 95 percent confidence interval:
##  7.777324      Inf
## sample estimates:
## mean of x
##  9.865746
```

Part 3

```
SIGN.test(data.skew$skew,md=8,alternative="greater",conf.level=.95)
```

```
##
##  One-sample Sign-Test
##
## data:  data.skew$skew
## s = 26, p-value = 0.4439
## alternative hypothesis: true median is greater than 8
## 95 percent confidence interval:
##  5.022828      Inf
## sample estimates:
## median of x
##    8.261484
```

```
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9405 5.0364    Inf
## Interpolated CI       0.9500 5.0228    Inf
## Upper Achieved CI     0.9675 4.9976    Inf
```