Assignment2

Zimeng Ming V00844078

2019/2/22

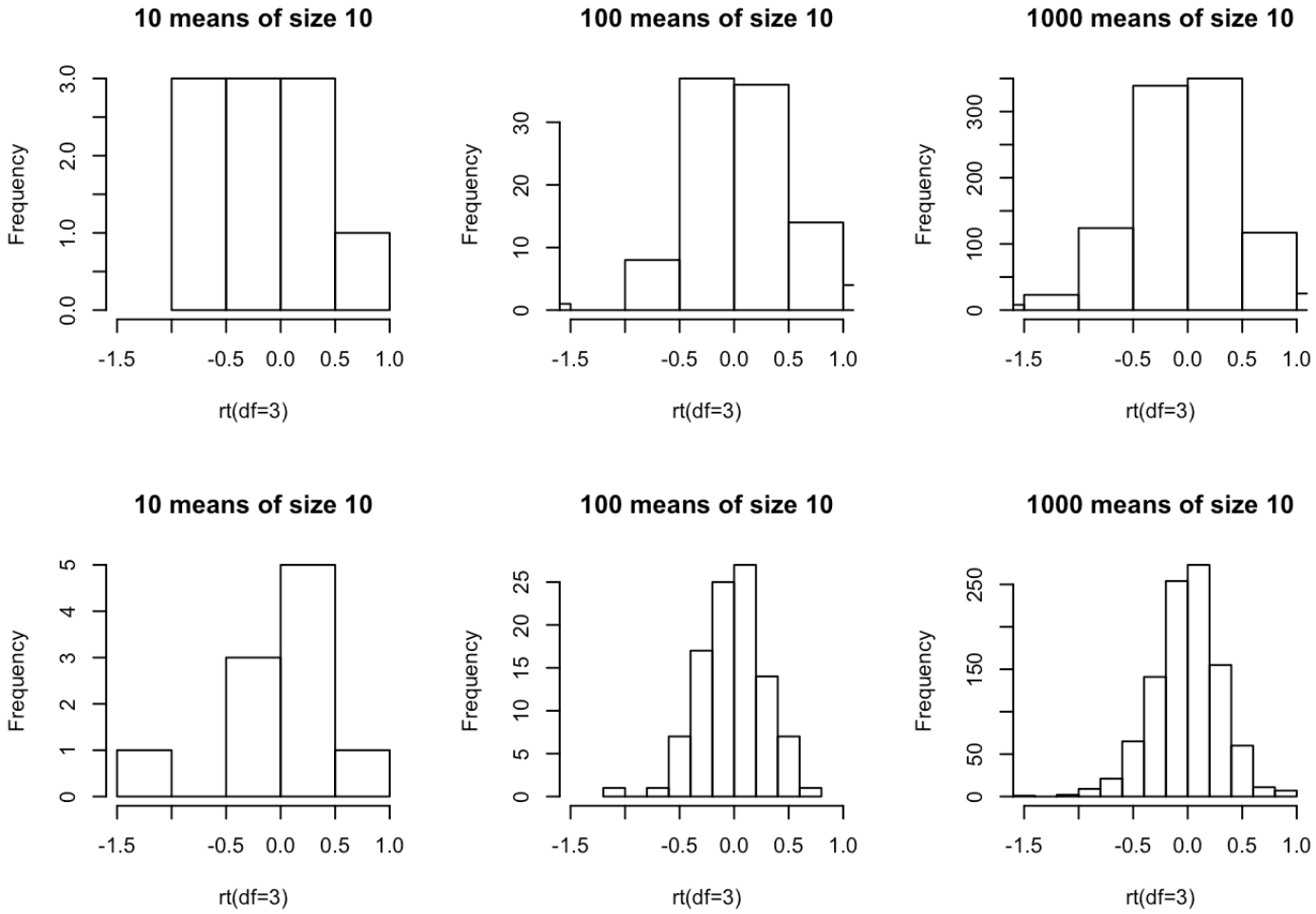
Question1:

```
#first we initialize the sample and size of sample.
sample_Q1<-c(10,100,1000)
size<-c(10,30,100)

#bulid the matrix for the two distribution
#for t-distribution
t10<-matrix(rep(0,times=length(size)*10),nrow = length(size),ncol = 10)
t100<-matrix(rep(0,times=length(size)*100),nrow = length(size),ncol = 100)
t1000<-matrix(rep(0,times=length(size)*1000),nrow = length(size),ncol = 1000)

#for binomial distribution
b10<-matrix(rep(0,times=length(size)*10),nrow = length(size),ncol = 10)
b100<-matrix(rep(0,times=length(size)*100),nrow = length(size),ncol = 100)
b1000<-matrix(rep(0,times=length(size)*1000),nrow = length(size),ncol = 1000)


#create a function for t-distribution
t_mean_function<-function(n){
  #n is the sample size
   return(mean(rt(n,df=3)))
}

#create a function for binomial distribution
bi_mean_function<-function(n){
  #n is the sample size
  return(mean(rbinom(n,1,0.85)))
}



#using the for loop to get the values in the matrix, m is the sample
for(m in sample_Q1){
  #set the martrix index to zero
  i<-0
  for (n in size) {
    i<-i+1
    for(k in 1:m){
      #when sample is 10
```

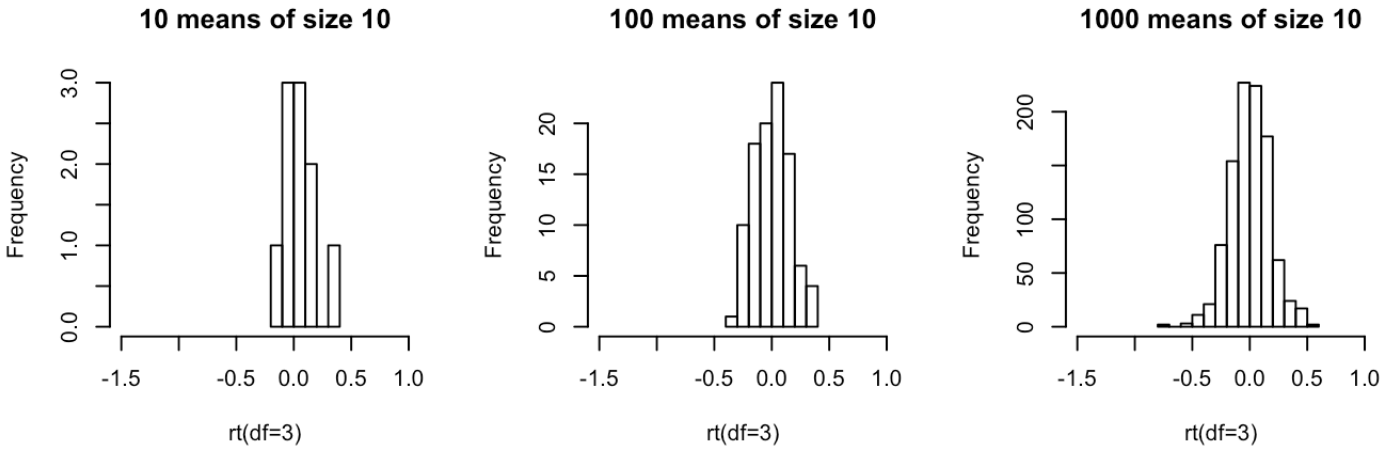
```
      if(m==10){
      t10[i,k]<-t_mean_function(n)
      b10[i,k]<-bi_mean_function(n)
      }
       #when sample is 100
      if(m==100){
      t100[i,k]<-t_mean_function(n)
      b100[i,k]<-bi_mean_function(n)
      }
       #when sample is 1000
      if(m==1000){
      t1000[i,k]<-t_mean_function(n)
      b1000[i,k]<-bi_mean_function(n)
     }
    }
  }
}
```

for the t-distribution:

```
par(mfrow=c(2,3))
#here we using the range -1.5 - 1 in order to show all the data in the distribution

hist(t10[1,],main="10 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t100[1,],main="100 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t1000[1,],main="1000 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t10[2,],main="10 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t100[2,],main="100 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t1000[2,],main="1000 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
```

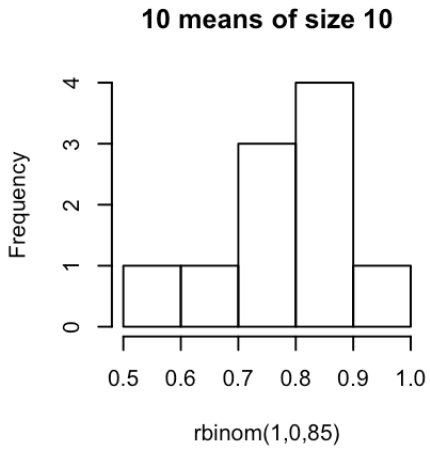


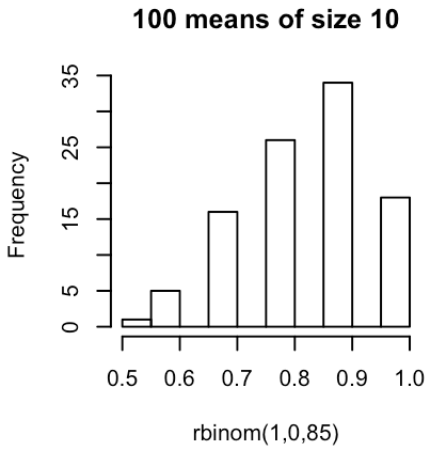

```
hist(t10[3,],main="10 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t100[3,],main="100 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
hist(t1000[3,],main="1000 means of size 10",xlab = "rt(df=3)",xlim = c(-1.5,1))
```

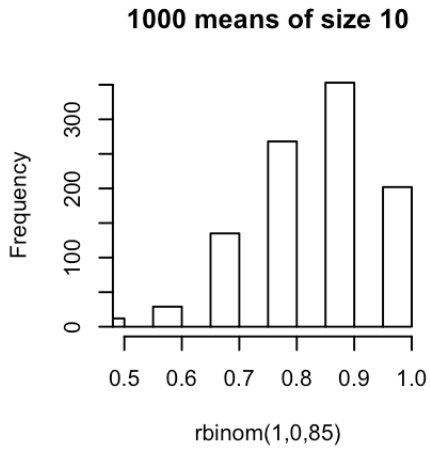


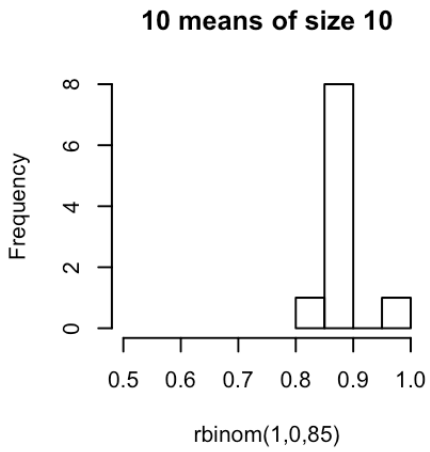
for the binomial distribution:

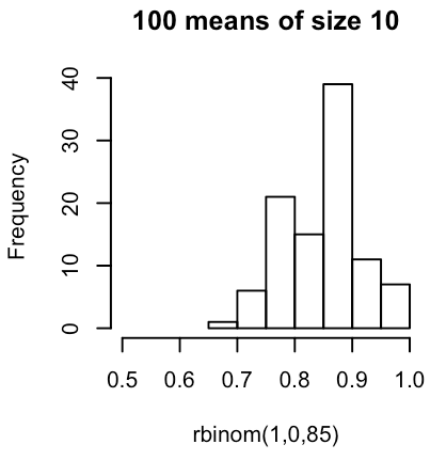
```
par(mfrow=c(2,3))
#here I try the range lots of times and decided using xlim 0.5-1 to show correctly

hist(b10[1,],main="10 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b100[1,],main="100 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b1000[1,],main="1000 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b10[2,],main="10 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b100[2,],main="100 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b1000[2,],main="1000 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
```

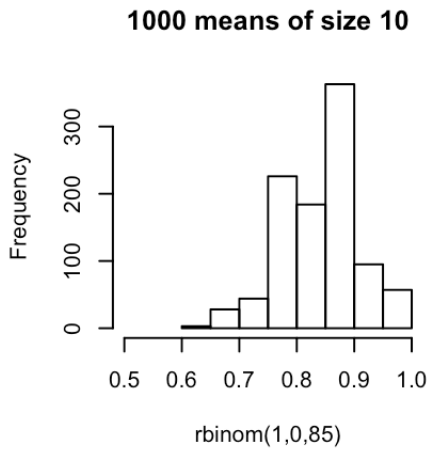




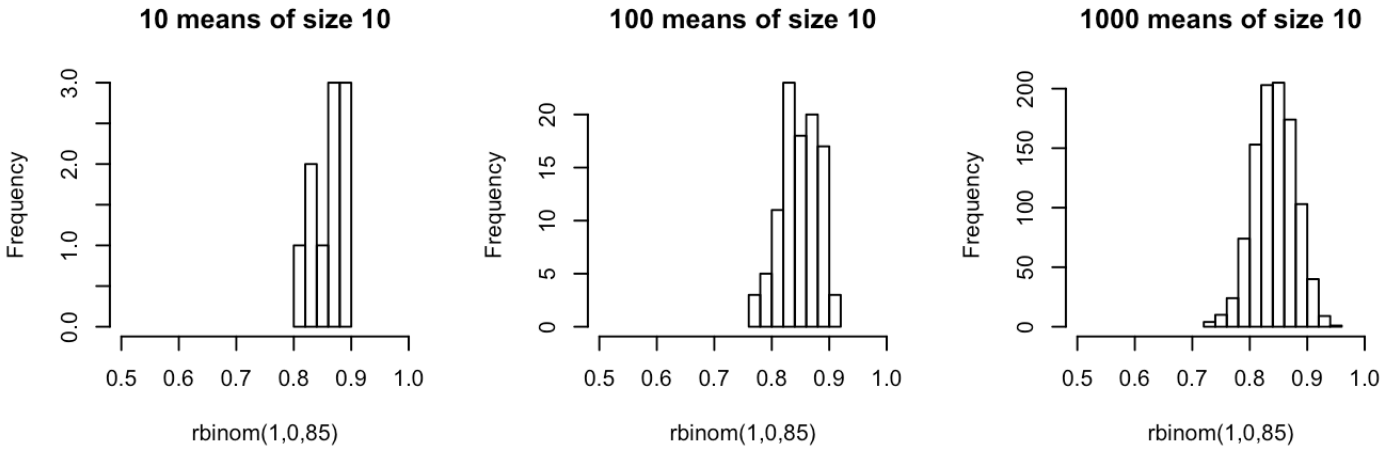









```
hist(b10[3,],main="10 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b100[3,],main="100 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
hist(b1000[3,],main="1000 means of size 10",xlab = "rbinom(1,0,85)",xlim = c(0.5,1))
```



b)when the sample size is increase, the distribution will be more central and get close finally will form a central peak. It not look symmetric when the size is small, but will look symmetric when the size is large

C)when the number of sample increase, the distribution are more likly to get close and form a peak since it become large. Similarly with the question b, the distribution will looks symmetric when the size is large

D)for the t-distribution, it will first very large range and not like symmetric. but in the end it will be symmetric. Also for the binormial distribution, the situation is same.

Question2:

```
#import the data
salt<-read.table("salt.txt",header = T)
#show the data
salt
```

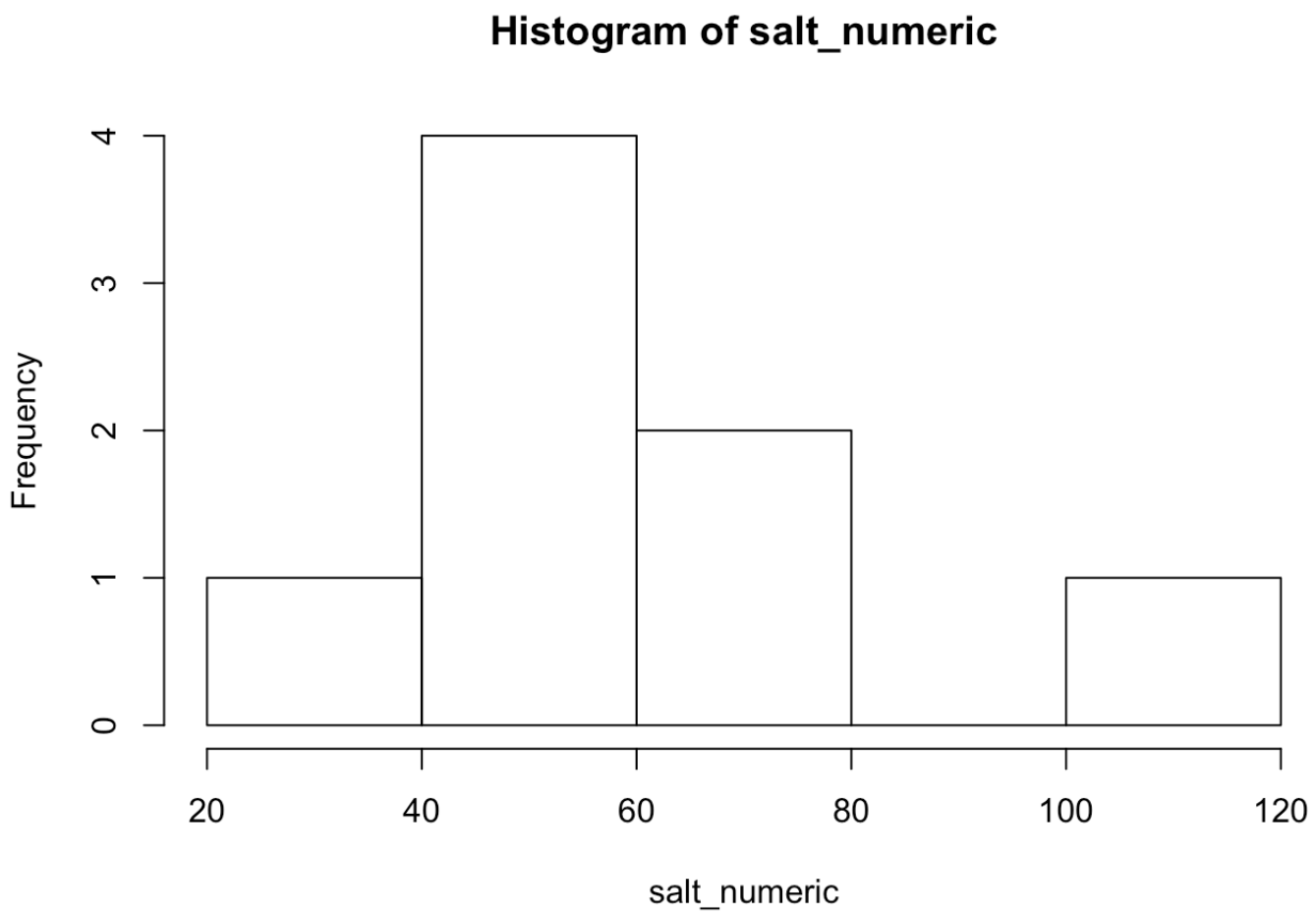
```
##   X13.53
## 1  28.42
## 2  48.11
## 3  48.64
## 4  51.40
## 5  59.91
## 6  67.98
## 7  79.13
## 8 103.50
```

a)Determine whether the data come from a symmetric distribution. Comment.

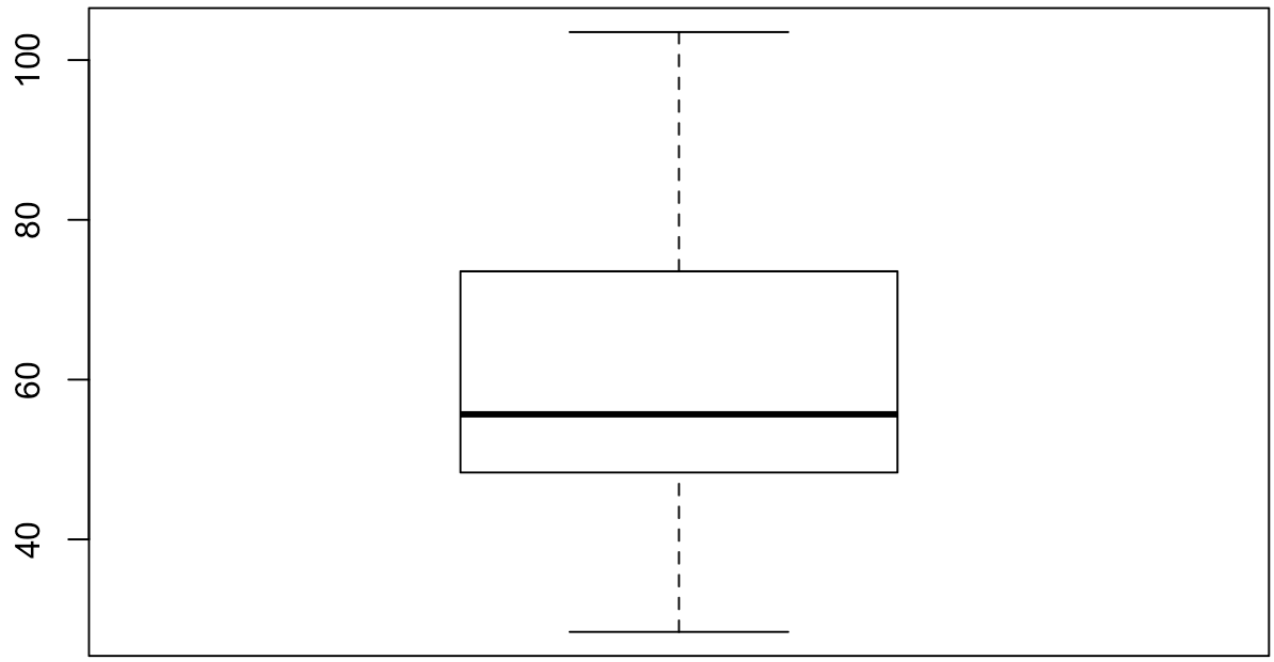
```
#summary the data
summary(salt)
```

```
##      X13.53
##  Min.   : 28.42
##  1st Qu.: 48.51
##  Median : 55.66
##  Mean   : 60.89
##  3rd Qu.: 70.77
##  Max.   :103.50
```

```
salt_numeric<-salt[,1]
#get the historgram
hist(salt_numeric)
```




```
#get the boxplot
boxplot(salt_numeric)
```



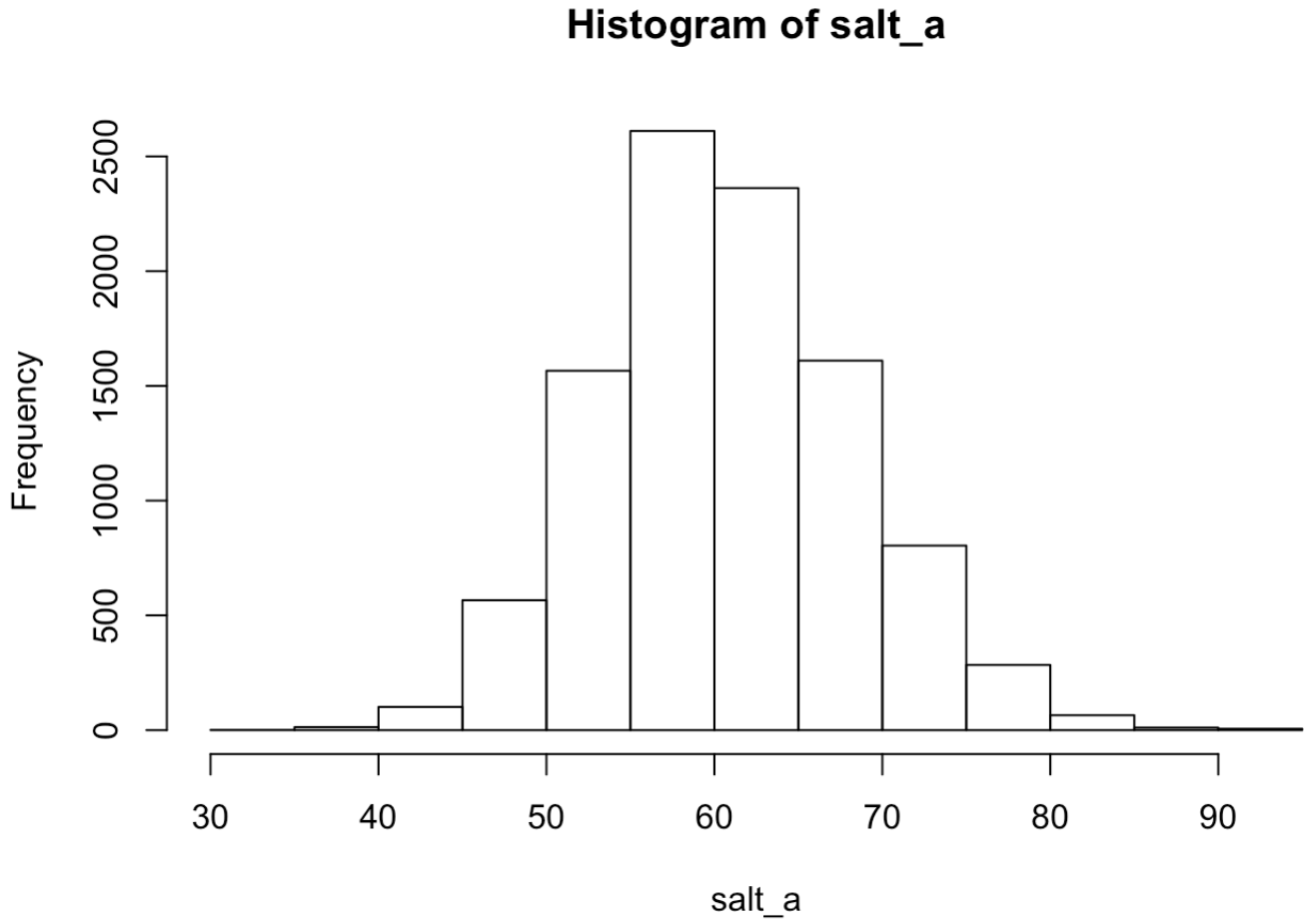
So as we can see from the summary and Histogram of the data. We can see clearly from the histogram that the data are not looks like symmetry and the boxplot also shows there are not outliers and the median is not on the middle. so the data are not looks like symmetric.

B)Develop a bootstrap test to determine if the mean and median are equal. What do you conclude?

```
#take sample size 5 with replacement from data
sample(salt_numeric,5,replace = T)
```

```
## [1] 48.11 48.64 48.11 67.98 67.98
```

```
#get the initializing vector to 0 which has a length of 10000
salt_a<-numeric(10000)

#calculate the sample mean
for(i in 1:10000){
  salt_a[i]<- mean(sample(salt_numeric,length(salt_numeric),replace = T))
}

#get the histogram of sample.
hist(salt_a)
```




```
#get the quantile
quantile(salt_a,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 46.68125 76.36097
```

The true Median : 55.66 and Mean : 60.89 are in the range of 95% confidence interval. and the histogram graph is symmetric. Thus we can say the median and mean are likely equal.

C)Estimate the skew (y1) and kurtosis (y2) of this distribution using the data.

For Skew:

```
#for the skew
skew<-(sum((salt_numeric-mean(salt_numeric))^3)/length(salt_numeric))/(sqrt(var(salt_
numeric))^3)

skew
```

```
## [1] 0.4690721
```

For Kurtosis:

```
((sum((salt_numeric-mean(salt_numeric))^4)/length(salt_numeric))/(var(salt_numeric)^2
))-3
```

```
## [1] -0.9031081
```

D)Based on your analysis above, what do you conclude about the distribution?

```
#doing the skew and Kurtosis test:
0.4690721>2*sqrt(6/length(salt_numeric))
```

```
## [1] FALSE
```

```
-0.9031081>2*sqrt(24/length(salt_numeric))
```

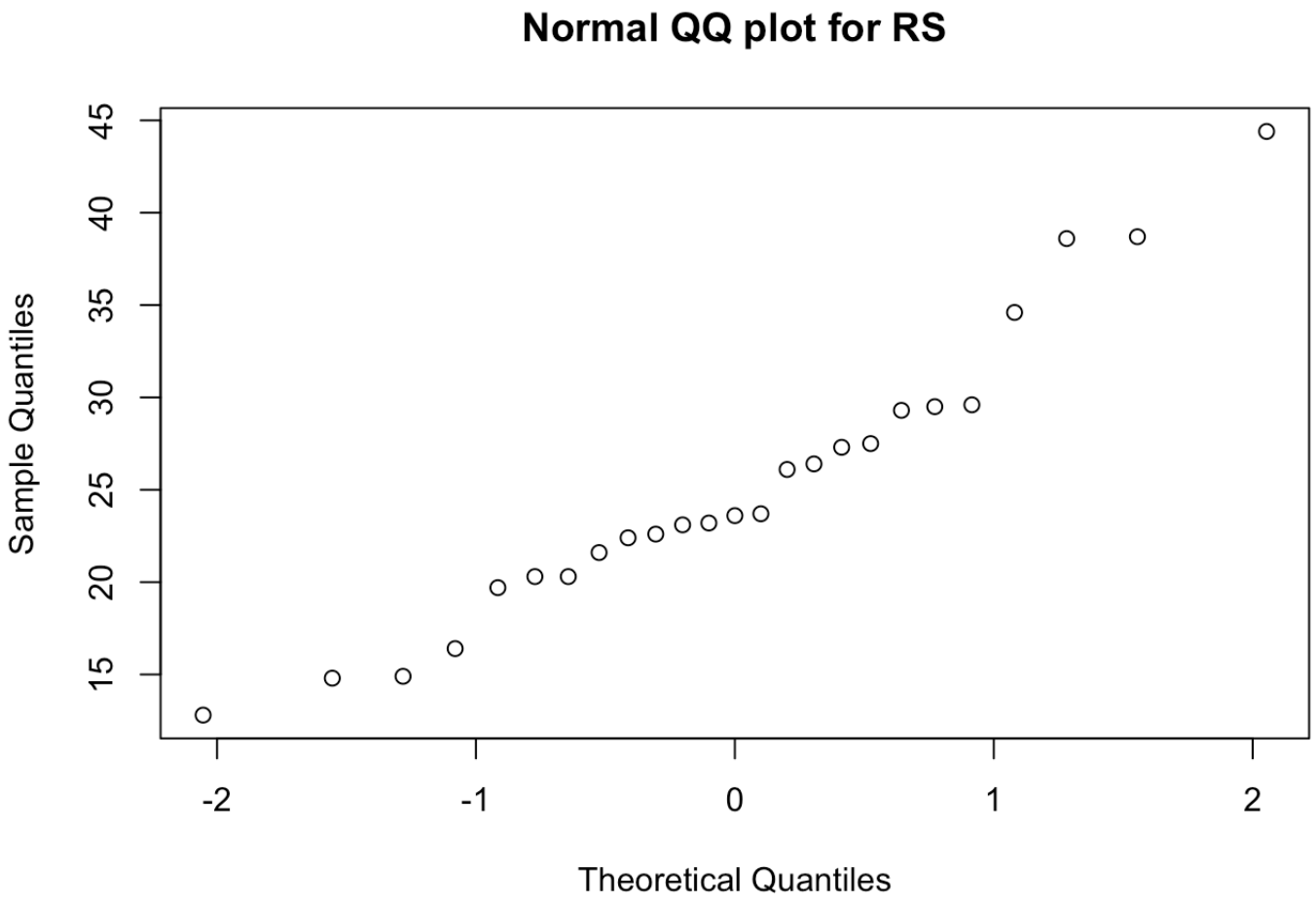
```
## [1] FALSE
```

Conclusion: So it seems not skewed and is symmetric based on the two rules above. so we can concluded that the distriution is symmetric

---

Question3: a)Construct a normal QQplot for the RS line and comment on the distribution of RS.

```
#load the data
fecundity<-read.table("fecundity.txt",header = T)
attach(fecundity)

#get the qqplot

qqnorm(RS,main = "Normal QQ plot for RS")
```



The qqplot for the RS shows that it is nearly linear, so we can say that it is symmetric and related.

B):Compute the variances of the RS and NS lines. Construct side-by-side boxplots for RS and NS lines. Perform a hypothesis test to determine if RS and DS lines dier in population variance and provide the output from this test. Comment on the equal variance assumption for RS and NS lines.

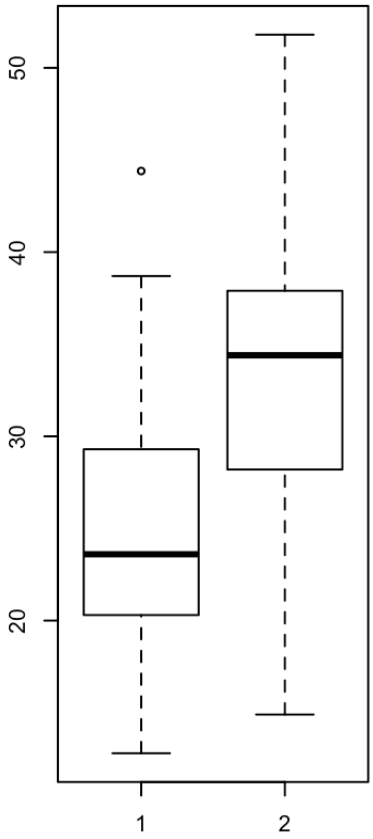
```
#first compute two variances.
var(RS)
```

```
## [1] 60.41007
```

```
var(NS)
```

```
## [1] 79.9596
```

```
#construct side-by-side boxplots
par(mfrow=c(1,3))
boxplot(RS,NS)
```



Hypothesis test: H0:the population variance are equal

```
var.test(NS,RS)
```

```
##
##  F test to compare two variances
##
## data:  NS and RS
## F = 1.3236, num df = 24, denom df = 24, p-value = 0.4974
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5832755 3.0036468
## sample estimates:
## ratio of variances
##           1.323614
```

AS the p-value is large, there is no strong evidenece againts H0, so we can conclude that the population variance are equal

C):Do RS and NS lines differ in population mean fecundity? Comment on your results.

H0: The population mean are equal

```
t.test(NS,RS,var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  NS and RS
## t = 3.4251, df = 48, p-value = 0.001268
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.351692 12.880308
## sample estimates:
## mean of x mean of y
##    33.372    25.256
```

As for the p-value is very small. There is strong evidence against H0, so we can conclude that the population mean are not equal.

D)Perform a nonparametric Wilcoxon test to see whether RS and NS lines dier in population mean fecundity.

```
wilcox.test(NS, RS)
```

```
## Warning in wilcox.test.default(NS, RS): cannot compute exact p-value with
## ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  NS and RS
## W = 468.5, p-value = 0.002547
## alternative hypothesis: true location shift is not equal to 0
```

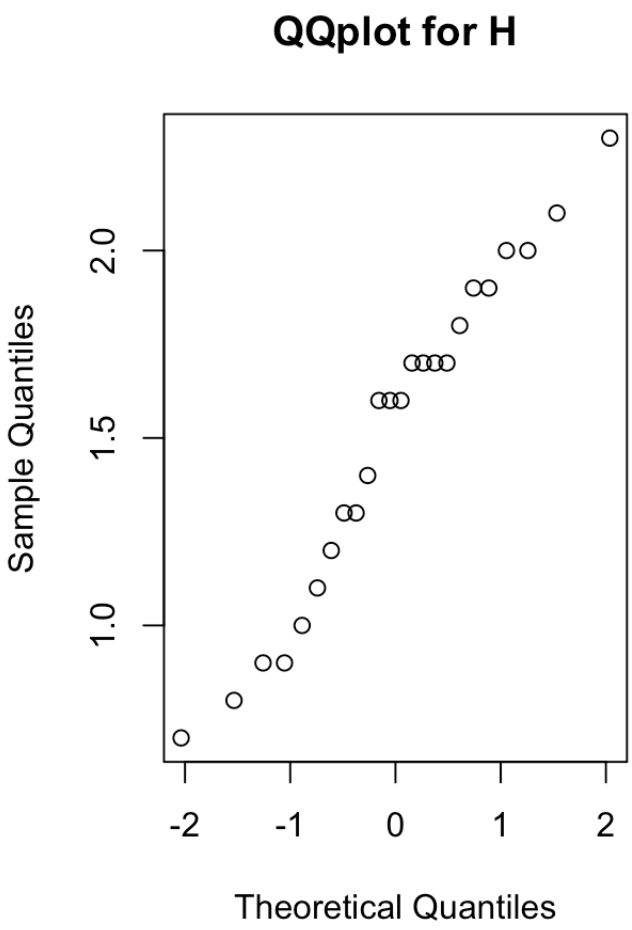
Also we can see the p-value are very small, So there is strong evidence against H0, so we can conclude that the population mean are not equal. 0 q io

---

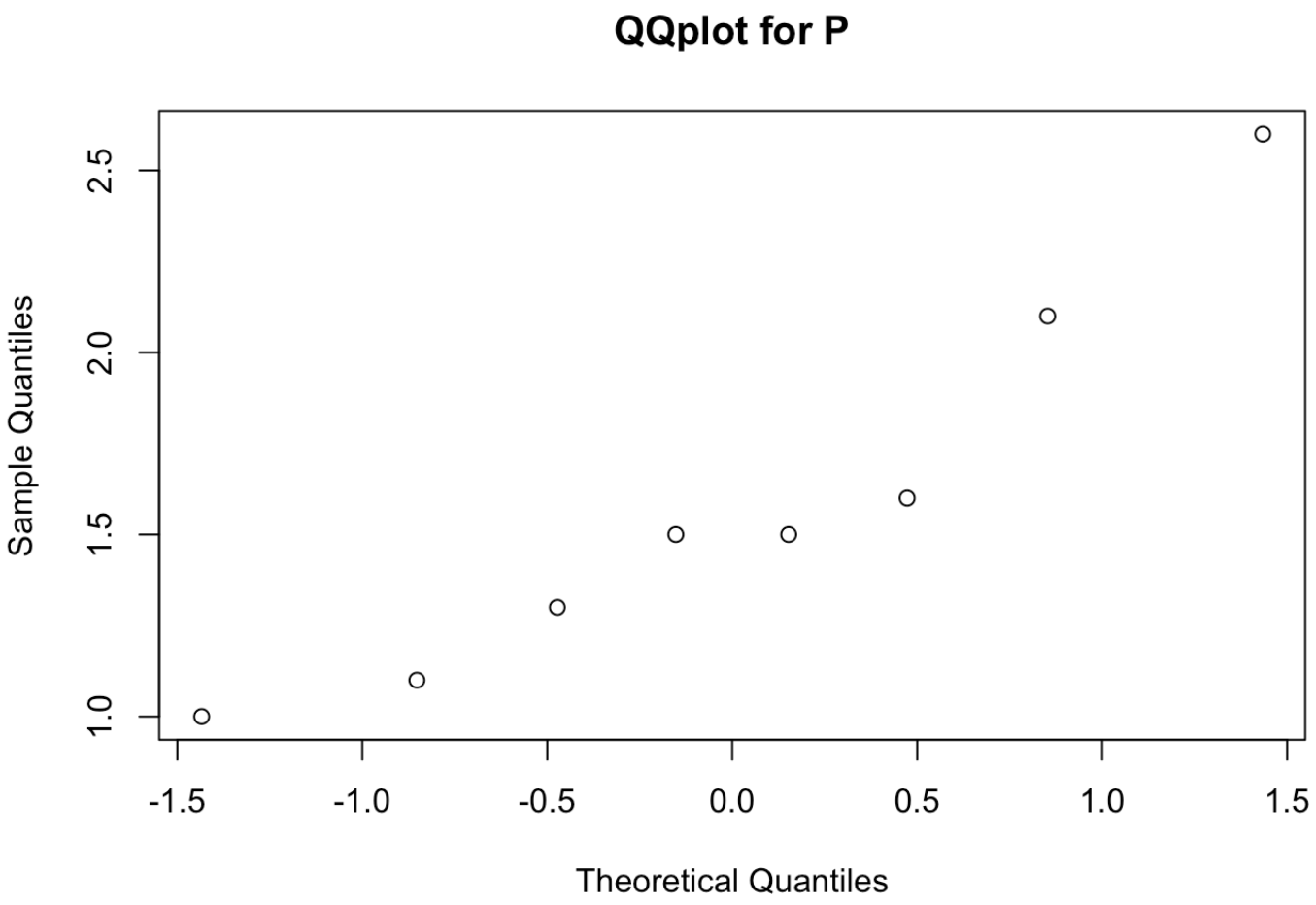
Question4:

a)Construct normal qq plots to verify the plausibility of both samples having been selected from normal population distributions. Comment.

```
#I could not load the data in the system so I just create a numerical sequence for th
at
H<-c(1.2, 0.9, 0.7, 1.0, 1.7, 1.7, 1.1, 0.9, 1.7, 1.9, 1.3 ,2.1, 1.6, 1.8, 1.4, 1.3,
1.9, 1.6, 0.8 ,2.0, 1.7, 1.6, 2.3, 2.0)
par(mfrow=c(1,2))
qqnorm(H,main="QQplot for H")
```

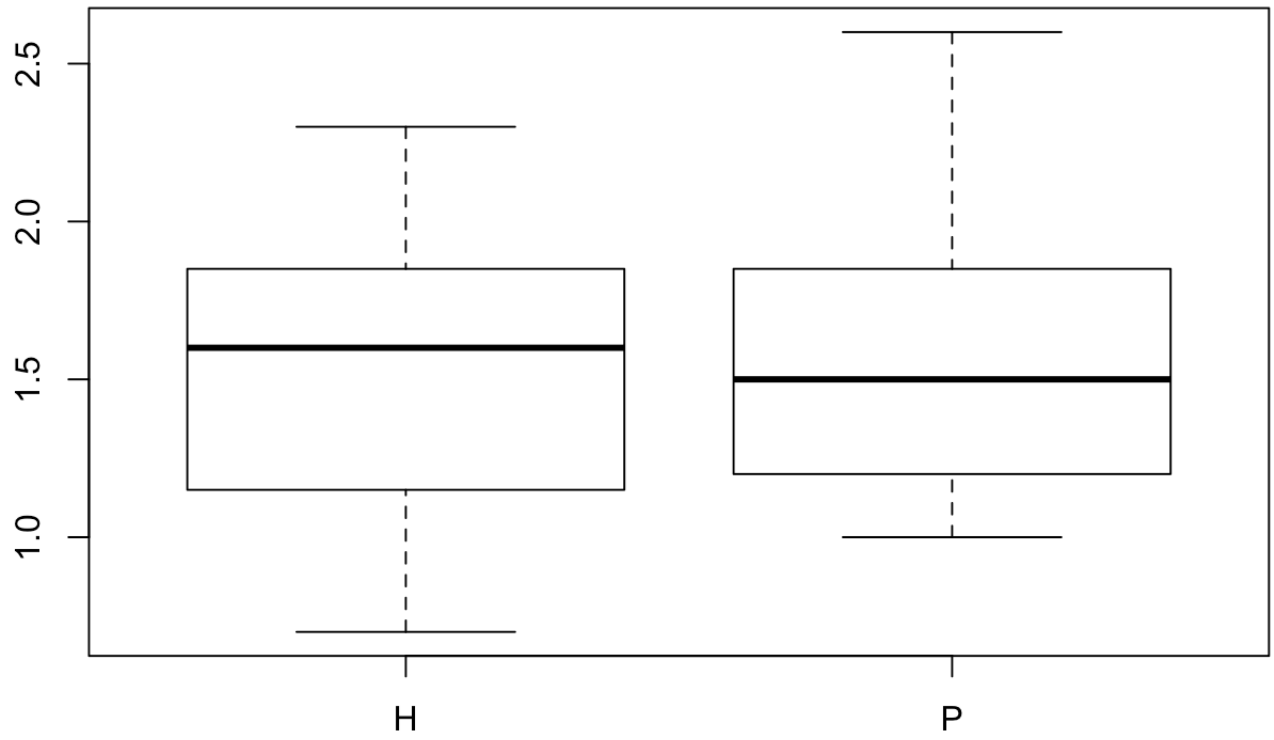



```
#get the values of P
P<-c(1.6, 1.5, 1.1 ,2.1, 1.5, 1.3, 1.0 ,2.6)
qqnorm(P,main="QQplot for P")
```



B)Construct a comparative boxplot. Does it suggest that there is a difference between true average extensibility for high-quality fabric specimens and that for poor-quality specimens?

```
#get the comparative boxplot
par(mfrow=c(1,1))
boxplot(H,P,names=c("H","P"))
```



As we can see from above. because the majority of the range of two plots are same, thus the mean of two types do not have a lot difference. Therefore, the true average extensibility of two population means are similar.

C)Determine whether true average extensibility differs for the two types of fabric. Comment.

we need to assume the variance is equal or not H0: no difference in variance of H and P

```
var.test(H,P)
```

```
##
##  F test to compare two variances
##
## data:  H and P
## F = 0.70158, num df = 23, denom df = 7, p-value = 0.4862
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1585015 2.0362234
## sample estimates:
## ratio of variances
##          0.7015781
```

As 95 percent confidence interval is not include 0,and the p-value is big so we can conclude that there is no eveidence against the hypothesis. so we can using var.equal= T in t.test

we do t.test for H0:the mean is equal

```
t.test(H,P,var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  H and P
## t = -0.41638, df = 30, p-value = 0.6801
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4674695  0.3091362
## sample estimates:
## mean of x mean of y
##  1.508333  1.587500
```

The p-value is big so there is no evidence against H0, the true average extensibilities is not diffrent with H and P.

Also for more accurate we can do nonparametic wilcox test:

```
wilcox.test(H,P)
```

```
## Warning in wilcox.test.default(H, P): cannot compute exact p-value with
## ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  H and P
## W = 96, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is 1, which there is no eveidence against H0. So according to the two tests above, we can conclude there is no difference between two types.