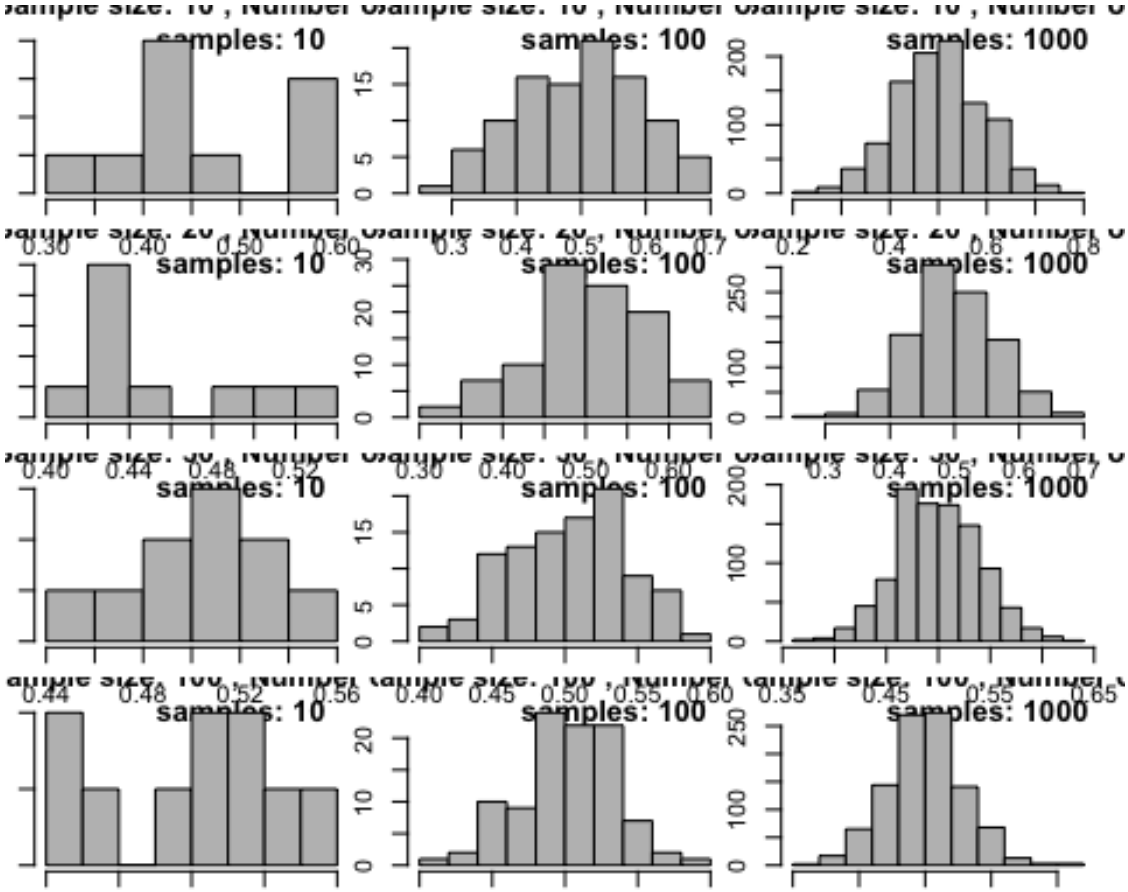
Assignment2_stat359

Koki Itagaki

2023-01-30

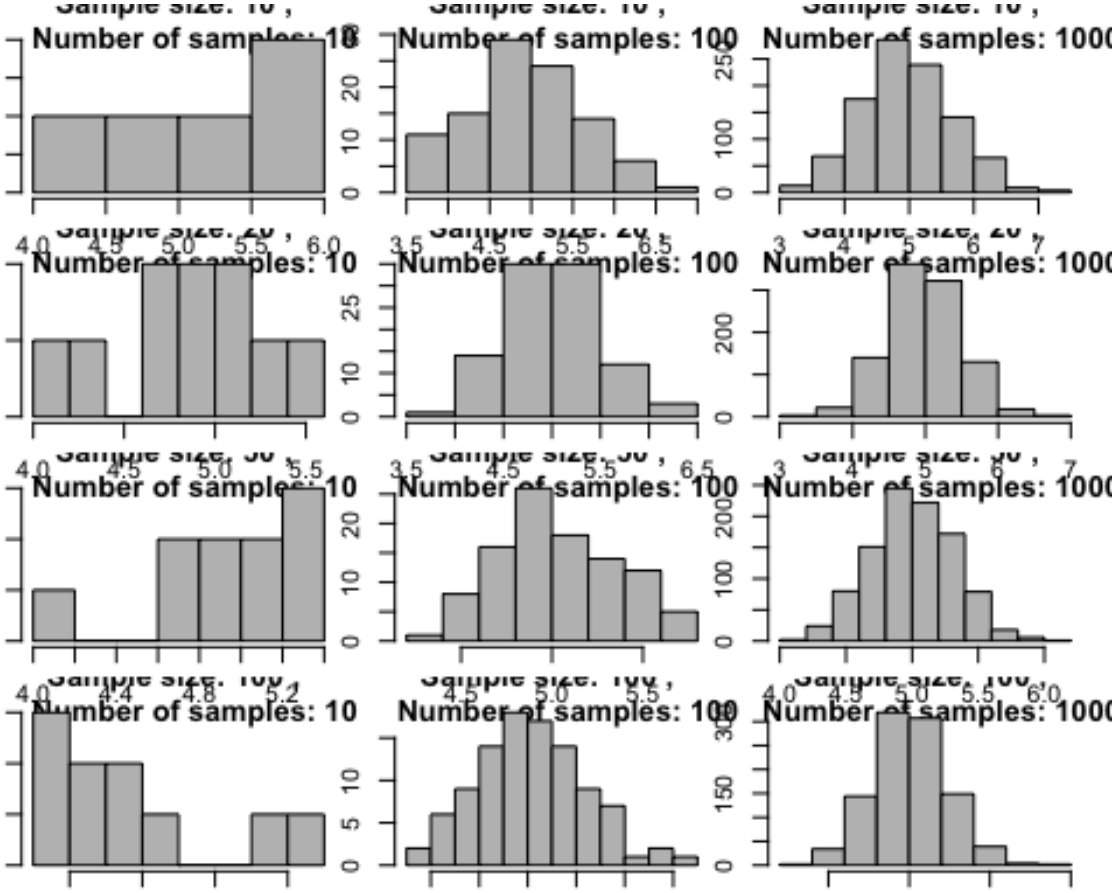
#1. Use R to take 10, 100, 1000 samples of size 10, 20, 50, 100 from 3 #distributions. #i) Uniform(a=0, b=1) ii) Poisson(λ = 5) #iii) Bernoulli(p=0.20) (or Binomial(n=1, p=0.20))

```
#The more the sample size increases, the more the graph close to the normal
#distribution by using central limit theorem
#begin code

#(a) Investigate how the Central Limit Theorem works when
#sampling from all of the above distributions, and for each
#sample size. Plot a histogram illustrating the distributions of
#the sample mean for each sample size, number of samples, and distribution.

#begin code
n.size<-c(10,20,50,100)
n.samples<-c(10,100,1000)
# plot window split as required
par(mar=c(1,1,1,1))

par(mfrow=c(length(n.size),length(n.samples)))
for(i in 1:length(n.size))
{
  for (j in 1:length(n.samples))
  {
    n.total<-n.size[i]*n.samples[j]
    ## generate all of the samples at once
    samples.all<-runif(n.total,0,1) ## uniform case
    #samples.all<-rpois(n.total,5) ## Poisson case
    #samples.all<-rbinom(n.total,size=1,p=0.2) ## rbinom
    ## store the samples in a matrix where columns correspond to
    ##individual replicates then take the mean accross the columns which
    ##yields a vector of length n.samples finally plot the histogram
    #par(mar=c(1,1,1,1))
    #I made a matrix to put sample.all into it
    samples.matrix <- matrix(samples.all, nrow = n.size[i], ncol =
n.samples[j],
                             byrow = TRUE)
    sample.means <- apply(samples.matrix, 2, mean)
    hist(sample.means, main = paste("Sample size:", n.size[i], "; Number of
            samples:", n.samples[j]), xlab = "Sample mean", col = "Grey")
  }
}
```



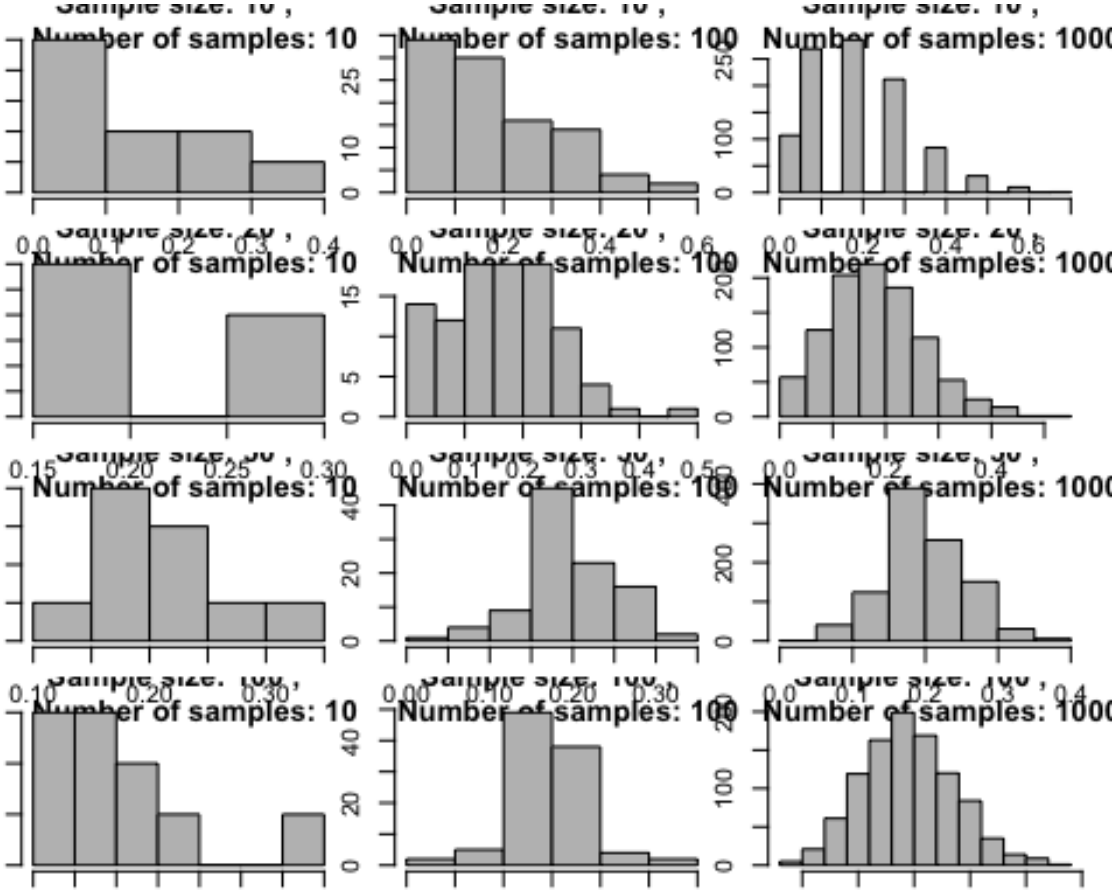

```
par(mfrow=c(length(n.size),length(n.samples)))
for(i in 1:length(n.size))
{
  for (j in 1:length(n.samples))
  {
    n.total<-n.size[i]*n.samples[j]
    ## generate all of the samples at once
    #samples.all<-runif(n.total,0,1) ## uniform case
    samples.all<-rpois(n.total,5) ## Poisson case
    #samples.all<-rbinom(n.total,size=1,p=0.2) ## rbinom
    ## store the samples in a matrix where columns correspond to individual
replicates
    ## then take the mean accross the columns which yields a vector of length
n.samples
    ## finally plot the histogram
    #par(mar=c(1,1,1,1))
     samples.matrix <- matrix(samples.all, nrow = n.size[i], ncol =
n.samples[j],
                              byrow = TRUE)
    sample.means <- apply(samples.matrix, 2, mean)
    hist(sample.means, main = paste("Sample size:", n.size[i], ";
    Number of samples:", n.samples[j]), xlab = "Sample mean", col = "Grey")
```

```
  }
  }
```




```
par(mfrow=c(length(n.size),length(n.samples)))
for(i in 1:length(n.size))
{
  for (j in 1:length(n.samples))
  {
    n.total<-n.size[i]*n.samples[j]
    ## generate all of the samples at once
    #samples.all<-runif(n.total,0,1) ## uniform case
    #samples.all<-rpois(n.total,5) ## Poisson case
    samples.all<-rbinom(n.total,size=1,p=0.2) ## rbinom
    ## store the samples in a matrix where columns correspond to individual
    #replicates then take the mean accross the columns which yields a vector
    #of length n.samples finally plot the histogram
    #par(mar=c(1,1,1,1))
    samples.matrix <- matrix(samples.all, nrow = n.size[i], ncol =
n.samples[j],
                             byrow = TRUE)
    sample.means <- apply(samples.matrix, 2, mean)
    hist(sample.means, main = paste("Sample size:", n.size[i], ";
    Number of samples:", n.samples[j]), xlab = "Sample mean", col = "Grey")
```

```
  }
}
```

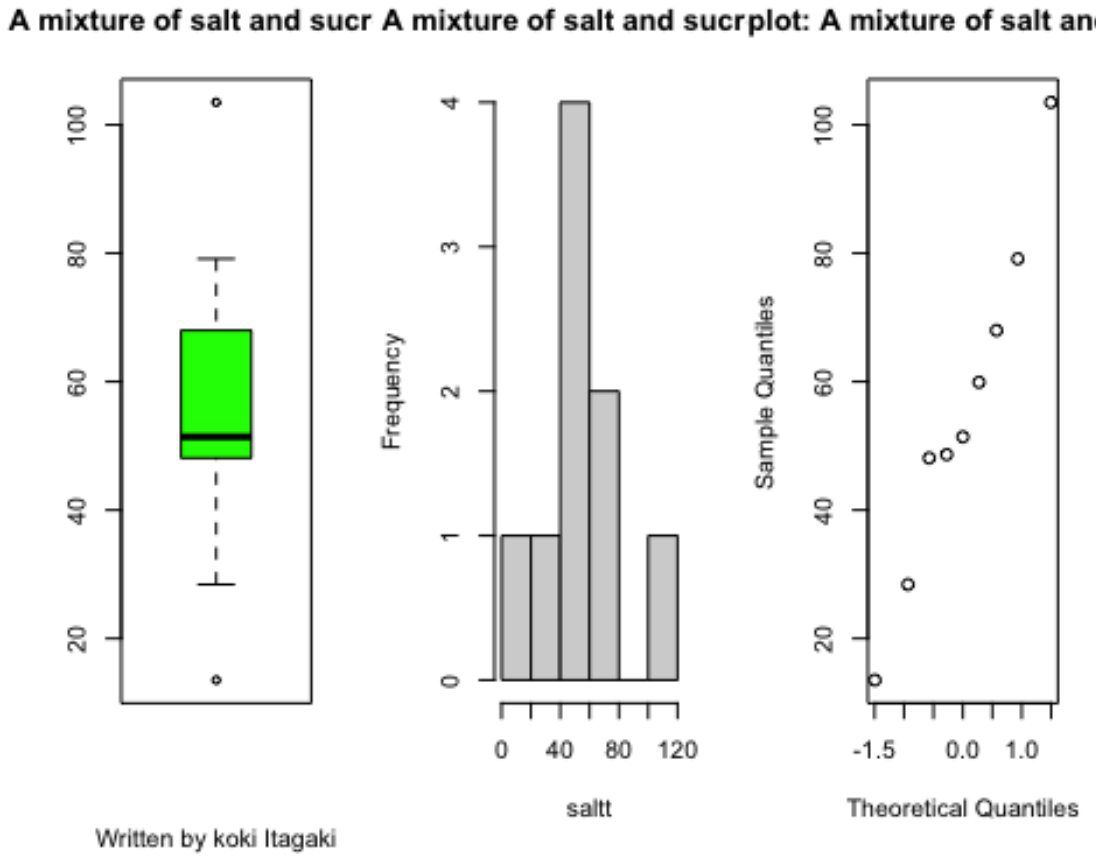


#(b) What do you notice as sample sizes increase? It is clear that the data in the graph is gathered to the middle of the graph and form a bell-shaped as sample sizes increase, Moreover, the distribution resembles the normal distribution more closely. #(c) What do you notice as the number of samples increases? The more the number of samples increases, the more the data are concentrated around the mean of the population. This also means the variability decreases and form a similar shape as the normal distribution since the graph becomes the graph similar to the normal distribution. #(d) How does the distribution affect the outcome?

#Question 2: A mixture of salt and sucrose was tasted to investigate how #saltiness was judged depending on sucrose concentration and the #data are contained in the file: salt.txt

```
#Question 2: A mixture of salt and sucrose was tasted to investigate how
#saltiness was judged depending on sucrose concentration and the
#data are contained in the file: salt.txt

salt<-read.table(file ='~/Desktop/stat359/data/salt.txt', sep="",header=TRUE)
salt

##     salt
## 1  13.53
```

```
## 2  28.42
## 3  48.11
## 4  48.64
## 5  51.40
## 6  59.91
## 7  67.98
## 8  79.13
## 9 103.50

saltt<-salt$salt
summary(saltt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.53   48.11   51.40   55.62   67.98  103.50

#From the summary function, we can see the mean and the median are different.
#This means I cannot say that bthe graph is symmentric.


#(a) Examine several graphical summaries to determine whether
#the data come from a symmetric distribution.
#Use par function to show boxplot and a histgram at the same time
par(mfrow = c(1,3))
boxplot(saltt,main= "A mixture of salt and sucrose",sub =
          "Written by koki Itagaki",col='green',names='salt')

hist(saltt, main = "A mixture of salt and sucrose")
qqnorm(saltt, main = "QQ-plot: A mixture of salt and sucrose")
```




```
#According to the graph above, we cannot say the data is symmentric.
#The histgram would be right-skewed and also the box plot shows the mean is
not
#close to the center of the box and has 2 outliers.
#Therefore, this is not symmentric.


#(b) Estimate the skew (γ1) and  (γ2) of this distribution
#using the data
skew<-function(x){
  m<-sum((x-mean(x))^3)/length(x)
  s<-sqrt(var(x))^3
  m/s
  }
skew.saltt<-skew(saltt)
skew.saltt

## [1] 0.1723753

#Now I will find kurtosis (γ2) of this distribution

kurtosis<-function(x){
  m2<-sum((x-mean(x))^4)/length(x)
```

```
  s2<-(var(x))^2
  m2/s2 -3
}

kurtosis(saltt)
```

```
## [1] -0.9342198
```

```
#The result above shows that the kurtosis is about -0.934.
#This means that the middle of the graph is flatter and the tail is heavier
#than the normal distribution.




#According to the result above, I could say that there is no skewness in this
#graph

#(c) Using the bootstrap construct a 95% confidence interval for the
#population skewness? Does it seem that the population distribution
#generating the data may be skewed?

#Now I use bootstarapping for skewness
x<-saltt ## data for bootstrapping
B<-15000
x.boot<-matrix(data=sample(x=x,size=B*length(x),replace=TRUE),
               nrow=length(x),ncol=B)
skew.boot.sampled<-apply(x.boot,2,skew)
boot.interval<-quantile(skew.boot.sampled,probs=c(0.025,0.975))
boot.interval
```

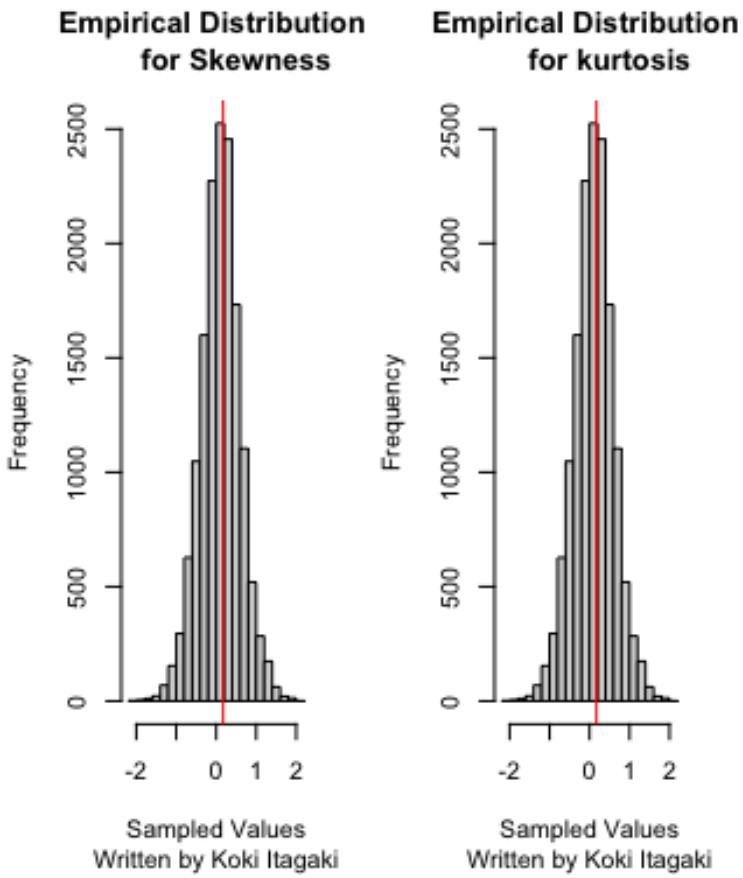
```
##      2.5%     97.5%
## -0.909979  1.093681
```

```
hist(skew.boot.sampled, main='Empirical Distribution
     for Skewness',sub = "Written by Koki Itagaki",xlab='Sampled Values')
abline(v= skew.saltt, col='red')

#According to the hist graph, we cannot clearly see that there is a skewness.
#Also,The 95% confidence interval for the skewness is approximately(-0.90,
1.09)
#This interval includes 0, it means the skewness is not neither positive nor
#negative skewness.




#(d) Using the bootstrap construct a 95% confidence interval for
#the population kurtosis. Does it seem that the population
#distribution generating the data may have non-zero kurtosis?
```

```r
#Now I use bootstarapping for kurtosis.

x<-saltt ## data for bootstrapping
B<-15000
x.boot<-matrix(data=sample(x=x,size=B*length(x),replace=TRUE),
               nrow=length(x),ncol=B)
kurtosis.boot.sampled<-apply(x.boot,2,kurtosis)
boot.interval<-quantile(kurtosis.boot.sampled,probs=c(0.025,0.975))
boot.interval
```

```
##      2.5%     97.5%
## -1.881980  0.565753
```

```r
hist(skew.boot.sampled, main='Empirical Distribution
     for kurtosis',sub = "Written by Koki Itagaki",xlab='Sampled Values')
abline(v= skew.saltt, col='red')

#According to the hist graph, we cannot clearly see that there is a kurtosis.
#Also,The 95% confidence interval for the skewness is
#approximately(-1.871, 0.586)
#This interval includes 0, so we are 95% confident that the data does not
have
#non-zero kurtosis.

#(e) Based on your analysis above, what do you conclude about the
#distribution?

#Since the results show that there is not neither large skewness nor
kurtosis,
#I can conclude that the graph is approximately symmentric that means that
#the data is evenly spread around the center.
```



#(e)Based on your analysis above, what do you conclude about the distribution? Since the results show that there is not neither large skewness nor kurtosis, I can conclude that the graph is approximately symmentric that means that the data is evenly spread around the center.

#3 Data on the per diem fecundity (fecundity.txt) #(number of eggs laidper female per day for the first 14 days of life) for 25 #females on 2genetic lines of the fruit fly Drosophila melanogaster are provided #Resistant (RS) to DDT were selectively bred and non-selected #(NS) was the control. Do RS and NS lines differ in population #variance? Do RS and NS lines differ in population mean #fecundity?

```
#1 = rs, 2 = ns

#Lets assume that:Ho: σ1 = σ2 , Ha: σ1 != σ2  (σ1 is not equal to σ2)

fecundity<-read.table(file ='~/Desktop/stat359/data/fecundity.txt'
                      ,sep="",header=TRUE)
rs<-fecundity$RS
ns<-fecundity$NS


var.test(rs,ns)
```

```
##
##  F test to compare two variances
##
## data:  rs and ns
## F = 0.75551, 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:
##  0.3329286 1.7144557
## sample estimates:
## ratio of variances
##          0.7555074
```

```
#Since (p-value = 0.4974) > a = 0.05, wecannot reject Ho.
#There is a insignificant evidence that the variances are
#different form each other.
#This means that the variance of RS is the same as the variance of NS.
```

```
#Since there are not enough sample sizes m and n, we use t-test insted of z-
test

#Lets assume that:Ho: u1 = u2 , Ha: u1 != u2  (u1 is not equal to u2)
```

```
t.test(rs,ns,alternative = "two.sided", mu = 0, var.equal = TRUE)
```

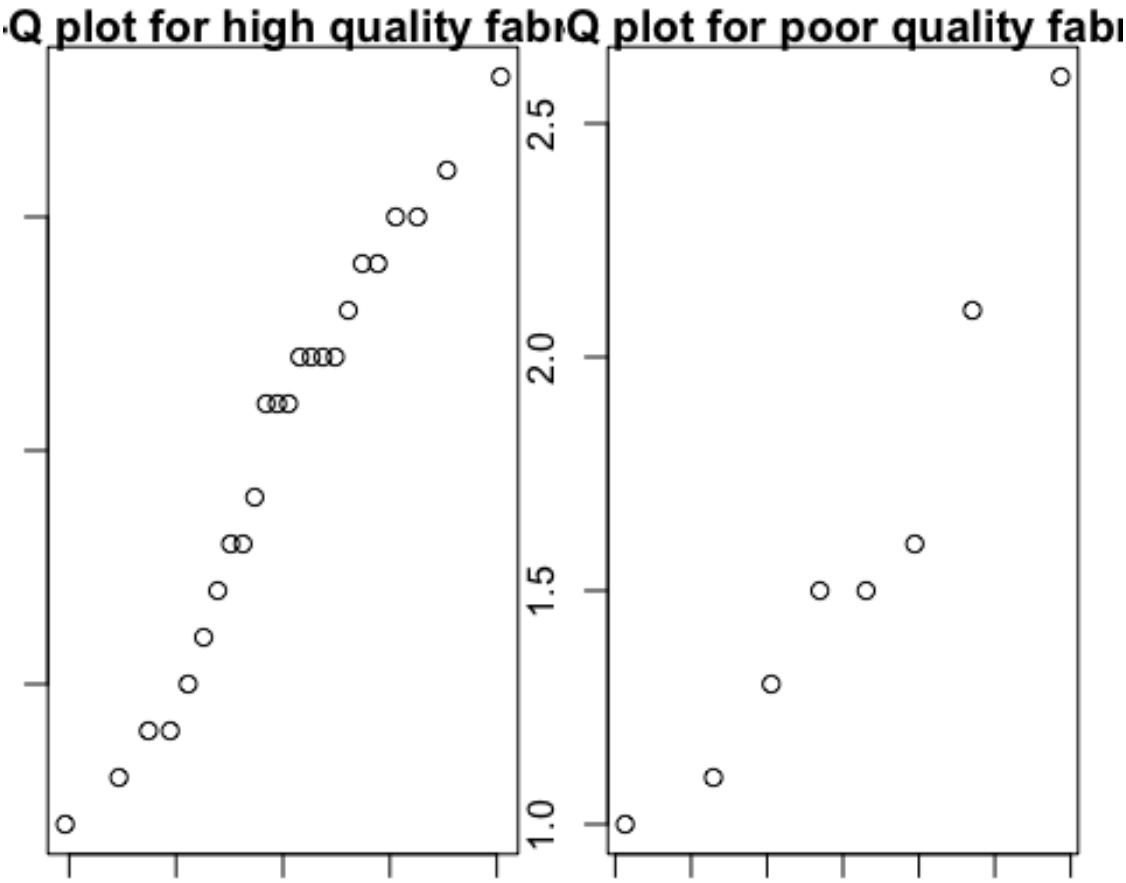
```
##
##  Two Sample t-test
##
## data:  rs and ns
## t = -3.4251, df = 48, p-value = 0.001268
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.880308  -3.351692
## sample estimates:
## mean of x mean of y
##    25.256    33.372
```

```
#Since (p-value = 0.0013) < a = 0.05, we reject Ho.
#There is a significant evidence that the means are different form each
other.
#This means that the mean of RS is different from the mean of NS.
```

```
#Since the sample size isnt enough, we use t test and to decide
#which t test I am going to use, I need to test if the variance of
#two population distributions is the same
```

#Q4. Fusible interlinings are being used with increasing frequency to #support outer fabrics and improve the shape and drape of various #pieces of clothing. The data on extensibility (100%) at 100 gm/cm #for both high quality fabric (H) and poor-quality fabric (P) #specimens is given in fabric.txt.

```
H<-c(1.2, .9, .7, 1.0, 1.7, 1.7, 1.1, .9, 1.7, 1.9, 1.3, 2.1, 1.6, 1.8, 1.4,
     1.3, 1.9, 1.6, .8, 2.0, 1.7, 1.6, 2.3, 2.0)

P<-c(1.6, 1.5, 1.1, 2.1, 1.5, 1.3, 1.0, 2.6)

#(a) Construct normal qq plots to verify the plausibility of both
#samples having been selected from normal population
#distribution
poor<-P
high<-H
par(mar = c(1, 1, 1, 1))
par(mfrow = c(1,2))
qqnorm(high,main = "Q-Q plot for high quality fabric", sub = "Written by
Koki")
qqnorm(poor,main = "Q-Q plot for poor quality fabric", sub = "Written by
Koki")
```



```r
#According to the Q-Q plots, it looks the points make a stright line.
#Therefore, both of the data are normally distributed.



#(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?
boxplot(high,poor, names = c("extensibility for high-quality",
"extensibility for poor-quality"),main = "box plots of 2 different qualities
specimens", sub = "Written by Koki")

#From the boxplots, the distributions are almost the same and canot decide
#if the population means are different.



#(c) Decide whether true average extensibility differs for
#the two types of fabric.

#Since the number of samles are less than 30(n< 30, m < 30),
#I use t-test. First of all, I need find if the variance is the same or not
# to dicide which t-test i can use.
#Assume Ho: σ1 = σ2 , Ha: σ1 != σ2
var.test(high,poor)
```

```
##
##  F test to compare two variances
##
## data:  high and poor
## 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
```

```r
#Since (p-value = 0.4862) > (a = 0.05), we cannot reject Ho.
#There is a insignificant evidence that both variance are different.

#since the variance are almost the same, I use pooled t-test.

#Ho: u1 = u2
#Ha: u1!= u2

t.test(high,poor,alternative = "two.sided", mu = 0, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
```

```
## data:  high and poor
## 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

#Since (p-value = 0.6801) >> a = 0.05, we cannot reject Ho.
#There is an insignificant evidence that the means are not the same.
```