Data Science 1
Statistical Inference for Two Population Variances

Ann Maharaj

Statistical Inference for Two Population Variances

1. Introduction
2. Sampling Distribution
3. Confidence Interval Estimation
4. Hypothesis Testing
5. Examples
   - Example 1
   - Example 2

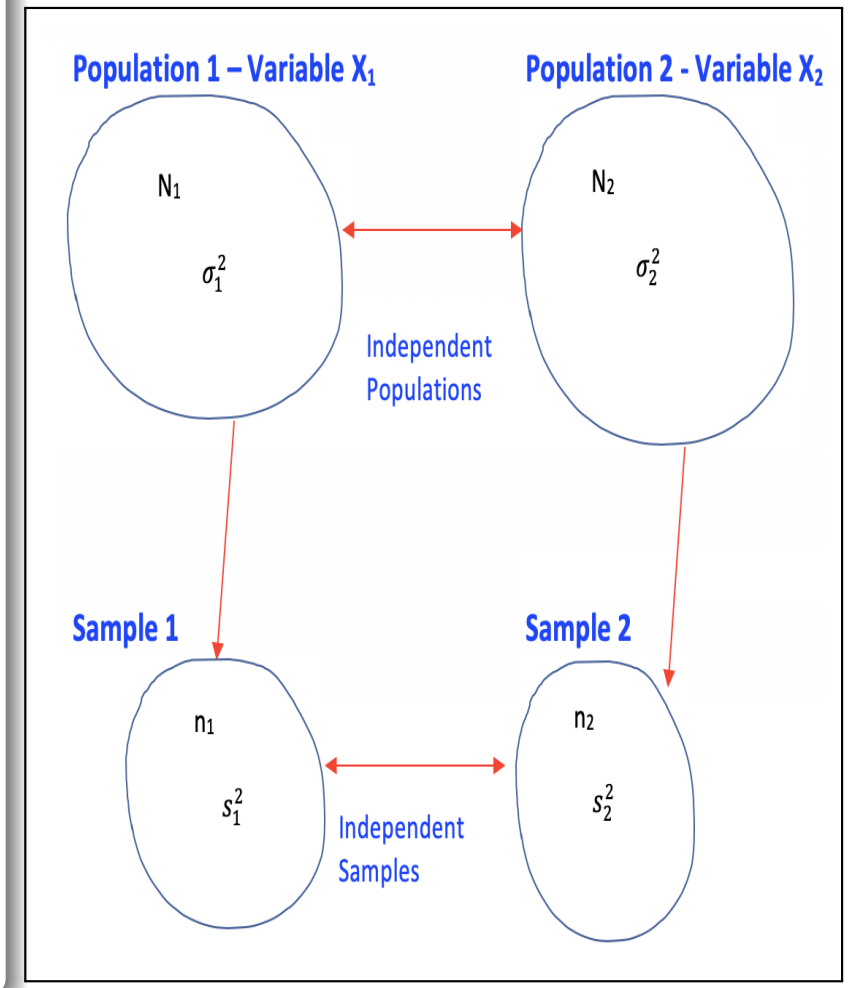
Introduction

There are circumstances where we might be interested in obtaining a confidence interval estimate or testing a hypothesis for the equality of two population variances. For example:

- When obtaining conference interval estimates or testing a hypothesis for the difference in means of two independent for two populations, we can test the assumption that the variances of the two populations are equal.
- In quality control:
  - Suppose a manufacturing plant makes two batches of an item, produces the items on different machines or produces the items on two different shifts.
  - Management may be interested in comparing the variances from two batches or two machines in order to determine whether there is more variability in one than another.

Introduction

- We are interested in some variable of interest for each of two independent populations.
- X_1 and X_2 which are random variables associated with the first and second populations, respectively, are assumed to be normally distributed.
- We take random samples of size n_1 and n_2 from the populations to estimate the ratio of the population variances.
- $\frac{s_1^2}{s_2^2}$ is an estimator of $\frac{\sigma_1^2}{\sigma_2^2}$, where s_1^2 and s_2^2 are estimators of the population variances σ_1^2 and σ_2^2, respectively.
- It can be shown that $\frac{s_1^2}{s_2^2}$ is an unbiased estimator of $\frac{\sigma_1^2}{\sigma_2^2}$.

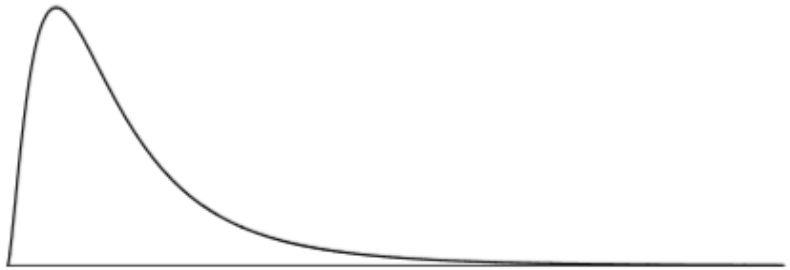


The random variables of interest X_1 and X_2 are normally distributed.

- It can be shown that

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

  follows an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.
- Because the F distribution always describes a ratio, there are two degrees of freedom parameters, one for the numerator and one for the denominator.
- The numerator degrees of freedom is always quoted first.
- The F distribution is a non-symmetric distribution that is skewed to the right. All its values are positive.



F

The random variables of interest X_1 and X_2 are normally distributed.

- Since $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ follow an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

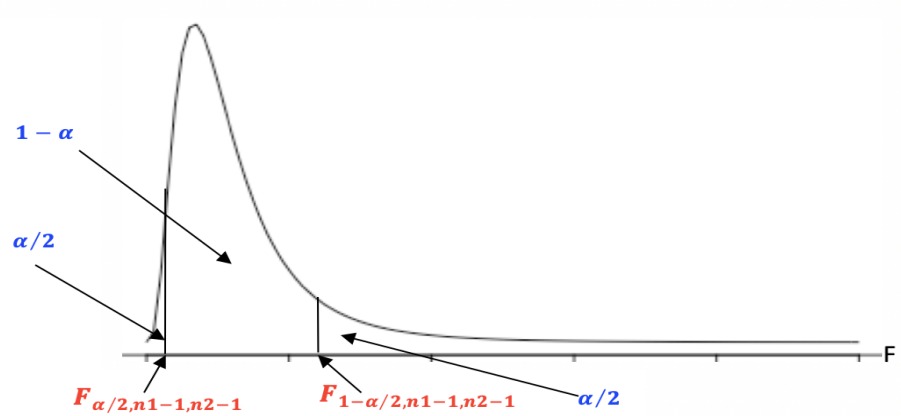
$$P\left(F_{\alpha/2,n_1-1,n_2-1} < F < F_{1-\alpha/2,n_1-1,n_2-1}\right) = 1 - \alpha$$

$$\Rightarrow P\left(F_{\alpha/2,n_1-1,n_2-1} < \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} < F_{1-\alpha/2,n_1-1,n_2-1}\right) = 1 - \alpha.$$

It can be shown that on making $\frac{\sigma_1^2}{\sigma_2^2}$ the subject of the formula:

- A $100(1-\alpha)\%$ Confidence Interval estimator of $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2,n_1-1,n_2-1}}, \frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}\right]$$



The random variables of interest X_1 and X_2 are normally distributed.

Table: 4

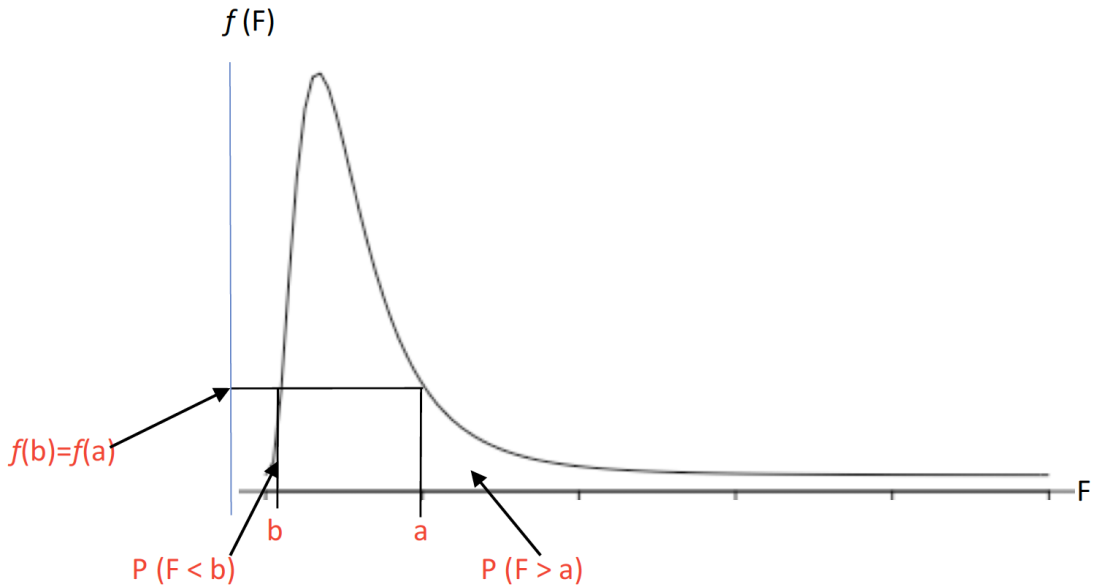
Null Hypothesis	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq 1$	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \geq 1$
Alternative Hypothesis	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$
Level of Significance	α	α	α
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
P-Value Approach Reject H_0 if	$p - value < \alpha$	$p - value < \alpha$	$p - value < \alpha$

P-value approach

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$

$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$



a = numerical value of the test statistic

$f(F)$ is the density function of the F-distribution and b is the quantile of the F-distribution such that $f(b) = f(a)$. In practice b is determined by trial and error.

p-value = $P(F > a) + P(F < b)$

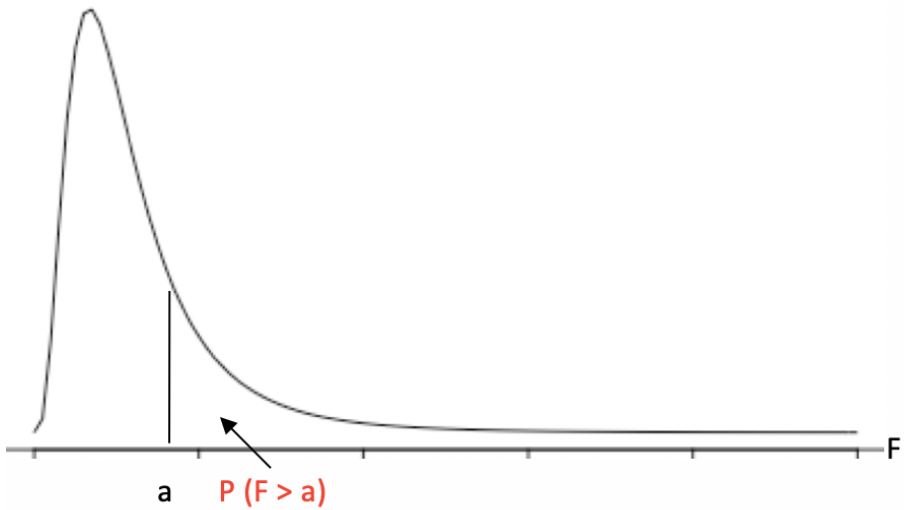
Reject H_0 if $p - value < \alpha$

P-value approach

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$

$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$



a = numerical value of the test statistic

p-value = P($F > a$)

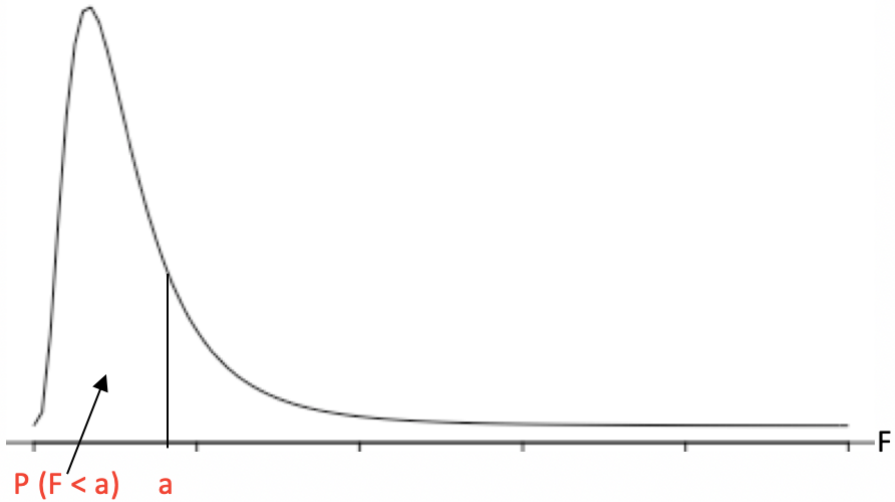
Reject H_0 if $p - value < \alpha$

P-value approach

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $\frac{\sigma_1^2}{\sigma_2^2} \geq 1$

$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$



a = numerical value of the test statistic

p-value = $P(F < a)$

Reject H_0 if $p - value < \alpha$

Example 1

Suppose two machines produce metal sheets that are specified to be 22mm. thick. Because of the machines, the operator, the raw material, the manufacturing environment and other factors, there is variability in the thickness. Management is concerned about the consistency of the two machines. To test consistency, they randomly sample 12 sheets produced by Machine 1 and 10 sheets by Machine 2. The thickness measurements of the sheets sampled from each machine is given in the file *sheets.csv*.

1. Can we conclude that the consistency in production of the two machines is significantly different? Use the 5% level of significance.
2. Obtain and report 95% and 99% confidence intervals estimates of ratio of the population variances of the thickness measures produced by the two machines.
3. Perform diagnostic checks for outliers, and to assess the validity of the assumption of normality of the thickness measures from the two machine.

Example 1- Part 1

σ_1^2 and σ_2^2: population variances associated with Machines 1 and 2, respectively.

1. Hypotheses:
   $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$
   $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$
2. Level of significance: $\alpha = 0.05$
3. Test statistic : F $= 4.133$
4. P-value: $p - value = 0.0420$
5. Decision rule: Reject H_0 if $p - value < 0.05$
6. Conclusion: Since the $p - value < 0.05$, reject H_0 at the 5% level of significance.
   We conclude that the consistency in production of the two machines is significantly different.

```
> #sample variances
> cbind(var(x), var(y))
          [,1]  [,2]
[1,] 0.1033333 0.025
>
> # test for equality of variances
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 4.1333, num df = 11, denom df = 9, p-value = 0.04204
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.056558 14.829981
sample estimates:
ratio of variances
          4.133333
```

However, if we test the hypothesis at the 1% level of significance, we will not reject H_0 and conclude that the consistency in production of the two machines is not significantly different.

Example 1- Part 2

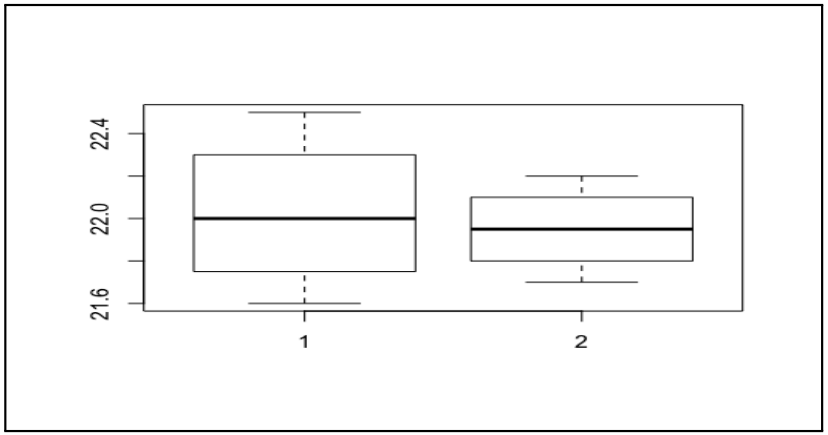
- A 95% confidence interval estimate of the $\frac{\sigma_1^2}{\sigma_2^2}$ is [1.057, 14,830].
- We are 95% confident that the variance in thickness of sheets produced by Machine 1 may be at least 1.057 larger than, but no more than 14.83 larger than that of Machine 2. The interval does not contain 1, implying that H_0 is rejected at the 5% of significance.
- A 99% confidence interval estimate of the $\frac{\sigma_1^2}{\sigma_2^2}$ is [0.655 22.885].
- We are 99% confident that the variance in thickness of sheets produced by Machine 1 may be at most 0.655 smaller than, but no more 22.885 larger than that of Machine 2. The interval contains 1, implying that H_0 is not rejected at the 1% of significance.

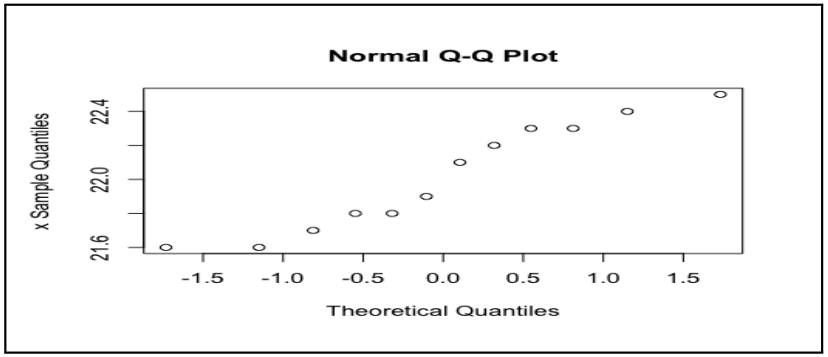
```
> #sample variances
> cbind(var(x), var(y))
          [,1]  [,2]
[1,] 0.1033333 0.025
>
> # test for equality of variances
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 4.1333, num df = 11, denom df = 9, p-value = 0.04204
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.056558 14.829981
sample estimates:
ratio of variances
          4.133333
```

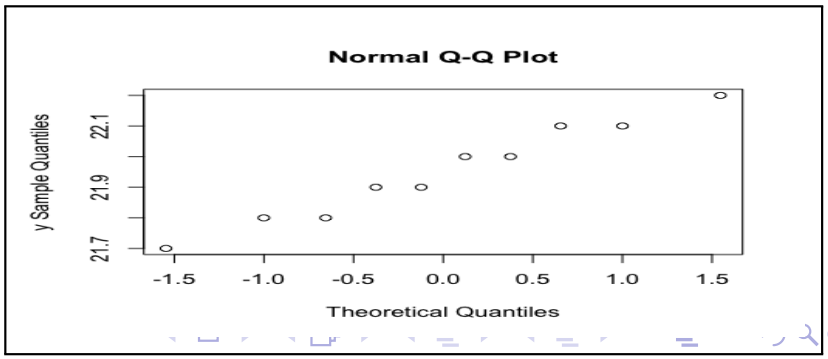
```
> # test for equality of variances with 99% C.I.
> var.test(x,y,conf.level = 0.99)

        F test to compare two variances

data:  x and y
F = 4.1333, num df = 11, denom df = 9, p-value = 0.04204
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
  0.6546051 22.8854079
sample estimates:
ratio of variances
          4.133333
```

Example 1 - Part 3

```
> #check for outliers
> boxplot(x,y)
>
> #assess normality assumptions of x and y
> qqnorm(x, ylab = "x Sample Quantiles")
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9188, p-value = 0.2762

> qqnorm(y, ylab = "y Sample Quantiles")
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.96572, p-value = 0.8486
```







- From the boxplots, we observe that the thickness measures of each of the two machines do not display any outliers.
- From the Normal QQ-plot as well as from the results of the Shapiro-Wilk normality test, we conclude that the assumption of normality of the thickness measures of sheets from each of the two machines is valid.

Example 2

A primary school has anticipated that a new instructional method will more effectively improve the reading ability of Grade 2 students than the standard method currently in use. To test this hypothesis, 20 children were randomly divided into two groups of 10 each. One group was instructed using the standard method while the other group was instructed using the new method. The students scores on reading tests are in the file reading.csv.

1. Test the hypothesis that the variances of the new and standard instructional method reading scores are equal. Use the 5% level of significance.
2. Based on the outcome of the test in Part (1), test the hypothesis that the new instructional method has improved reading ability. Use the 5% level of significance.
3. Perform diagnostic checks for outliers and to assess the validity of the normality assumptions.

Example 2- Part 1

σ_1^2 and σ_2^2: population variances associated with the standard and new reading instructional methods, respectively.

1. Hypotheses:
   $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$
2. Level of significance: $\alpha = 0.05$
3. Test statistic : F $= 0.766$
4. P-value: $p - value = 0.697$
5. Decision rule: Reject H_0 if $p - value < 0.05$
6. Conclusion: Since the $p - value > 0.05$ do not reject H_0 at the 5% level of significance and indeed any reasonable level of significance.

```
> #sample variances with 95% C.I.
> cbind(var(x), var(y))
         [,1]     [,2]
[1,] 13.06667 17.06667
>
> # test for equality of variances
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 0.76562, num df = 9, denom df = 9, p-value = 0.6972
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1901704 3.0824018
sample estimates:
ratio of variances
          0.765625
```

We conclude that the variances of the new and standard instructional methods reading scores are not significantly different.

Example 2: Part 2

μ_1 and μ_2: population means associated with the standard and new reading instructional methods, respectively. Based on the outcome of the test in Part (1), we assume equal population variances.

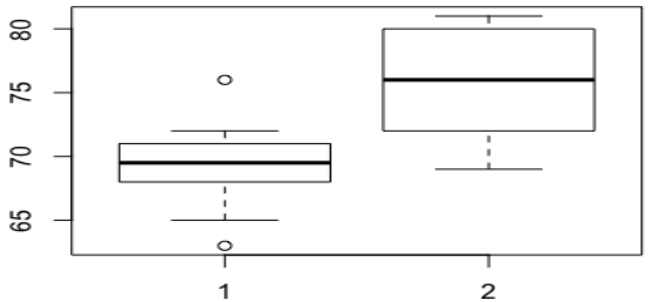
1. Hypotheses:
   $H_0 : \mu_1 - \mu_2 = 0$
   $H_1 : \mu_1 - \mu_2 < 0$
2. Level of significance: $\alpha = 0.05$
3. $t = -3.8021$ with $df = 18$
4. P-value: $p - value = 0.00065$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05, reject H_0 at the 5% level or indeed any reasonable level of significance.
   There is sufficient evidence to conclude that the new instructional method has improved reading performance.

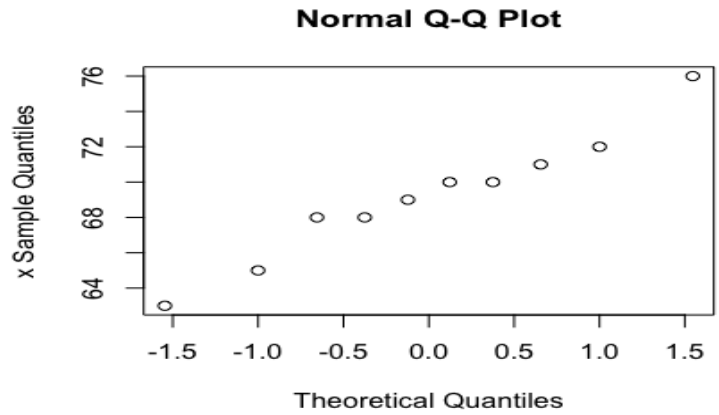
```
> #H0: mu1 -mu2 = 0 , H1: mu1 - mu2 < 0
> #t-test
> t.test(x,y, alternative="less", var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = -3.8021, df = 18, p-value = 0.0006527
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.589847
sample estimates:
mean of x mean of y
     69.2      75.8
```

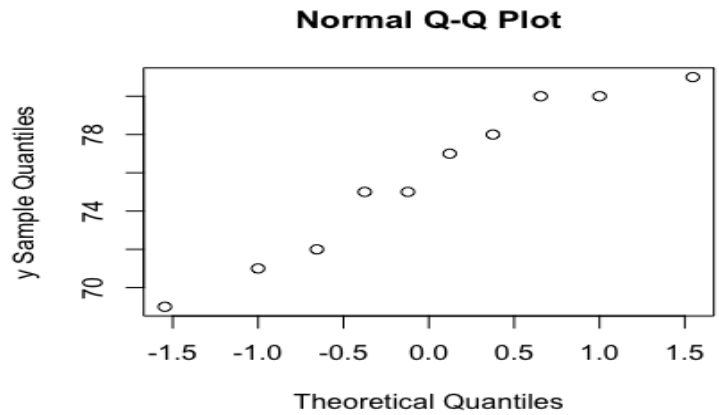
```
> # this will enable us to get a 95% two-sided confidence iterval
> #H0: mu1-mu2 = 0 , H1: mu1-mu2 not= 0
> t.test(x,y, alternative="two.sided", var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = -3.8021, df = 18, p-value = 0.001305
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.246981  -2.953019
sample estimates:
mean of x mean of y
     69.2      75.8
```

Example 2 - Part 3

```
> #check for outliers
> boxplot(x,y)
>
> #assess normality assumptions of x and y
> qqnorm(x, ylab = "x Sample Quantiles")
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9717, p-value = 0.9061

> qqnorm(y, ylab = "y Sample Quantiles")
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.93885, p-value = 0.5403
```





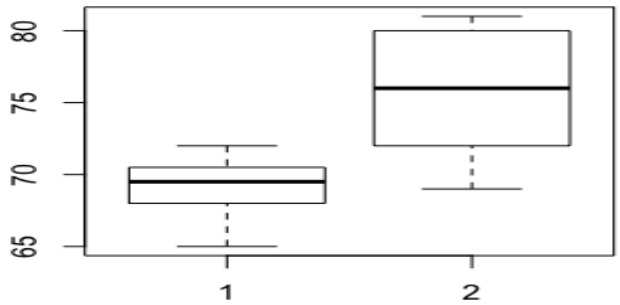


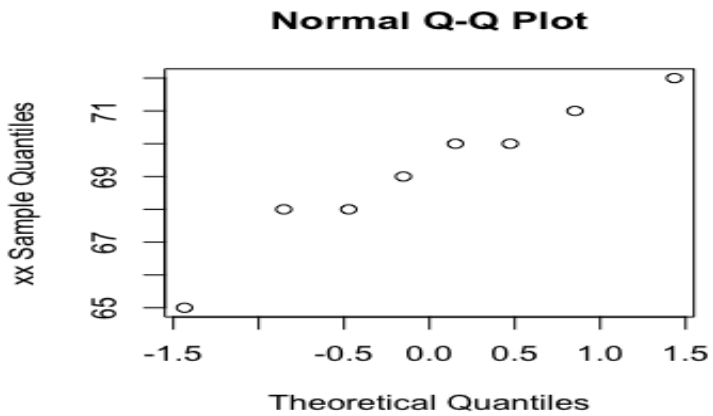
Example 2 - Part 3

- From the boxplot, we observe that the reading scores from the standard instructional method display two outliers, while those of the new instructional method do not display any outliers.
- From the Normal QQ-plot as well as from the results of the Shapiro-Wilk normality test for each variable, we conclude that the assumption of normality of the reading scores from both methods is valid.
- We remove the two outliers from the standard instructional method reading scores, and compute the results for the test in Part 1 and the confidence intervals in Part 2 again.

Example 2 - After removal of outliers

```
> boxplot(xx,y)
> qqnorm(xx, ylab = "xx Sample Quantiles")
> shapiro.test(xx)

        Shapiro-Wilk normality test

data:  xx
W = 0.95087, p-value = 0.72
```





- xx refers to the reading scores for the standard instructional method after the two outliers have been removed.
- The assumption of normality of reading scores for the standard instructional method is still valid.

Example 2 - After removal of outliers

1. Hypotheses:
   $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$
2. Level of significance: $\alpha = 0.05$
3. Test statistic : F $= 0.275$
4. P-value: $p - value = 0.103$
5. Decision rule: Reject H_0 if $p - value < 0.05$
6. Conclusion: Since the $p - value > 0.05$ do not reject H_0 at the 5% level of significance.
   We conclude variances of the new and standard instructional methods reading scores are not significantly different.

```
> #sample variances with 95% C.I.
> cbind(var(xx), var(y))
         [,1]     [,2]
[1,] 4.696429 17.06667
>
> # test for equality of variances
> var.test(xx,y)

        F test to compare two variances

data:  xx and y
F = 0.27518, num df = 7, denom df = 9, p-value = 0.1029
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.06556548 1.32725944
sample estimates:
ratio of variances
         0.2751814
```

Example 2: After removal of outliers

1. Hypotheses:
   $H_0 : \mu_1 - \mu_2 = 0$
   $H_1 : \mu_1 - \mu_2 < 0$
2. Level of significance: $\alpha = 0.05$
3. $t = -4.122$ with $df = 16$
4. P-value: $p - value = 0.0004$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05, reject H_0 at the 5% level or indeed any reasonable level of significance.
   There is sufficient evidence to conclude that the new instructional method has improved reading performance.

```
> #H0: mu1 -mu2 = 0 , H1: mu1 - mu2 < 0
> #t-test
> t.test(xx,y, alternative="less", var.equal = TRUE)

        Two Sample t-test

data:  xx and y
t = -4.122, df = 16, p-value = 0.0003994
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.847798
sample estimates:
mean of x mean of y
   69.125    75.800
```

```
> # this will enable us to get a 95% two-sided confidence iterval
> #H0: mu1-mu2 = 0 , H1: mu1-mu2 not= 0
> t.test(xx,y, alternative="two.sided", var.equal = TRUE)

        Two Sample t-test

data:  xx and y
t = -4.122, df = 16, p-value = 0.0007988
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.107875  -3.242125
sample estimates:
mean of x mean of y
   69.125    75.800
```