Two Sample Tests

- **Two sample z-test (proportion)**
- **Two sample z-test (mean)**
- **Two sample t-test (mean)**
  - **Independent samples**
  - **Paired t-test**
- **Two sample variance test**

Two Sample z-test for proportion

$$H_0 : P_1 - P_2 = d$$

$$H_1 : P_1 - P_2 > d$$

or

$$H_1 : P_1 - P_2 < d$$

or

$$H_1 : P_1 - P_2 \neq d$$

Problem:

A university wants to compare the pass rates of students taught using two different teaching methods.

Method A: Out of 200 students, 148 passed the examination.

Method B: Out of 180 students, 117 passed the examination.

At the 5% level of significance, test whether there is a significant difference in the proportion of students passing under the two methods.

$$H_0 : P_1 - P_2 = 0$$

$$H_1 : P_1 - P_2 \neq 0$$

```r
# Number of successes (passes)
successes <- c(148, 117)

# Sample sizes
samples <- c(200, 180)

# Two-sample proportion test
p=prop.test(successes, samples, correct = FALSE)
if (p[3]<0.05) print("Reject H_0") else print("Fail to reject H_0")
```

```
## [1] "Fail to reject H_0"
```

Two Sample z-test for mean

$$H_0 : \mu_1 - \mu_2 = d$$

$$H_1 : \mu_1 - \mu_2 > d$$

or

$$H_1 : \mu_1 - \mu_2 < d$$

or

$$H_1 : \mu_1 - \mu_2 \neq d$$

Problem: Suppose we want to compare the mean IQ scores of two cities A and B. If we know the sample for both the cities as raw data, we use z.test() function (available from BSDA)

```
x <- c(7.8, 6.6, 6.5, 7.4, 7.3, 7.0, 6.4, 7.1, 6.7,
       7.6, 6.8, 4.5, 5.4, 6.1, 6.1, 5.4, 5.0, 4.1, 5.5, 4.8)
y <- c(10.1, 11.5, 10.0, 10.8, 10.7, 10.6, 11.2, 10.9,
       10.4, 10.9, 11.1, 10.5, 10.8, 11.0, 11.2, 11.1, 10.9, 10.7, 10.6, 10.8)
library(BSDA)
```

```
## Loading required package: lattice

##
## Attaching package: 'BSDA'

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

```
z.test(x, y, sigma.x = 15, sigma.y = 15, alternative = "two.sided", mu = 0)
```

```
##
##  Two-sample z-Test
##
## data:  x and y
## z = -0.9666, p-value = 0.3337
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.881925   4.711925
## sample estimates:
## mean of x mean of y
##     6.205    10.790
```

```
zsum.test(mean.x=100.65, sigma.x=15, n.x=20, mean.y=108.8, sigma.y=15, n.y=20,
          alternative = "two.sided",mu = 0, conf.level = 0.95)
```

```
##
##  Two-sample z-Test
##
## data:  Summarized x and y
## z = -1.7182, p-value = 0.08577
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.446925   1.146925
## sample estimates:
## mean of x mean of y
##    100.65    108.80
```

Two sample t-test (mean) - Independent samples

Problem:

Two different teaching methods were used to teach Statistics to two independent groups of students. After the course, their test scores were recorded.

- Group A (Traditional method): 56, 60, 58, 62, 59, 61, 57, 63
- Group B (Interactive method): 64, 66, 65, 67, 68, 66, 69, 65

Assuming that the population variances are unknown, test at the 5% level of significance whether there is a significant difference in the mean scores of the two groups.

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

If population variances are equal, we use student's t-test

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows t distribution with $n_1 + n_2 - 2$ as degrees of freedom. Here

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

```r
# Marks of Group A
groupA <- c(56, 60, 58, 62, 59, 61, 57, 63)

# Marks of Group B
groupB <- c(64, 66, 65, 67, 68, 66, 69, 65)

# Two-sample t-test
t.test(groupA, groupB, alternative = "two.sided", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  groupA and groupB
## t = -6.4411, df = 14, p-value = 1.545e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.997645 -4.502355
## sample estimates:
## mean of x mean of y
##     59.50     66.25
```

If population variances are unequal, we use Welch t-test:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

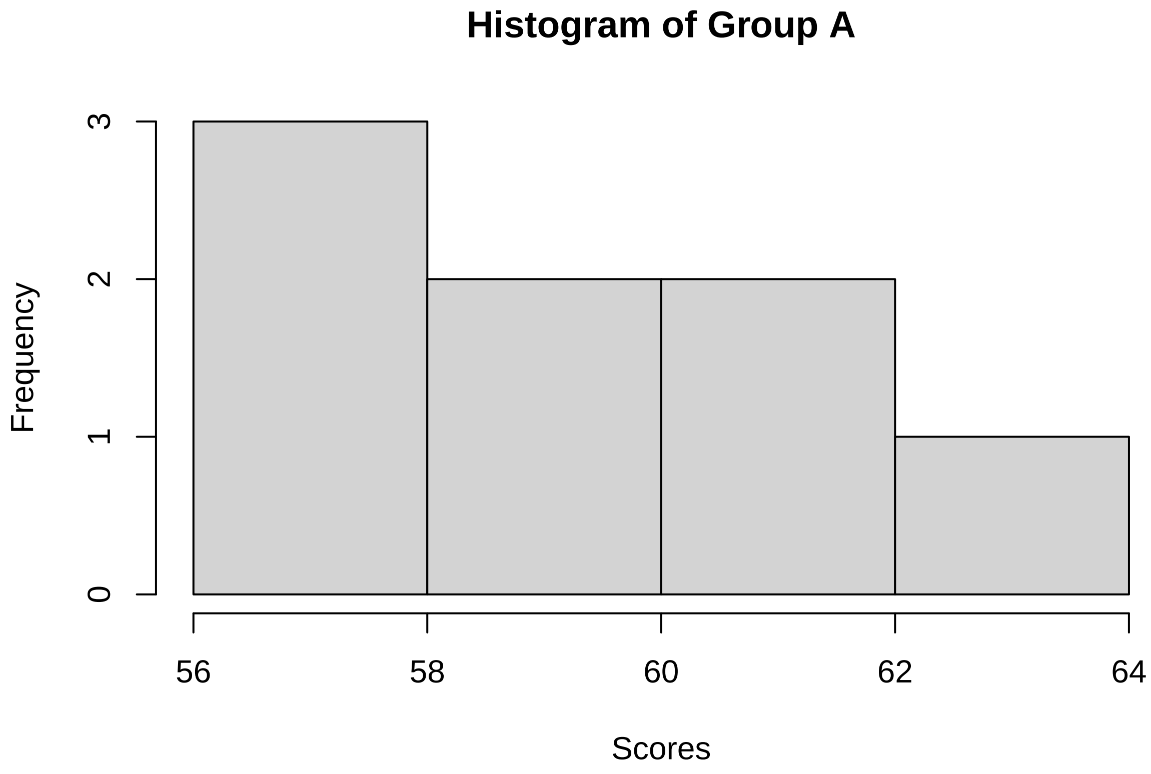
follows a t-distribution with degrees of freedom

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

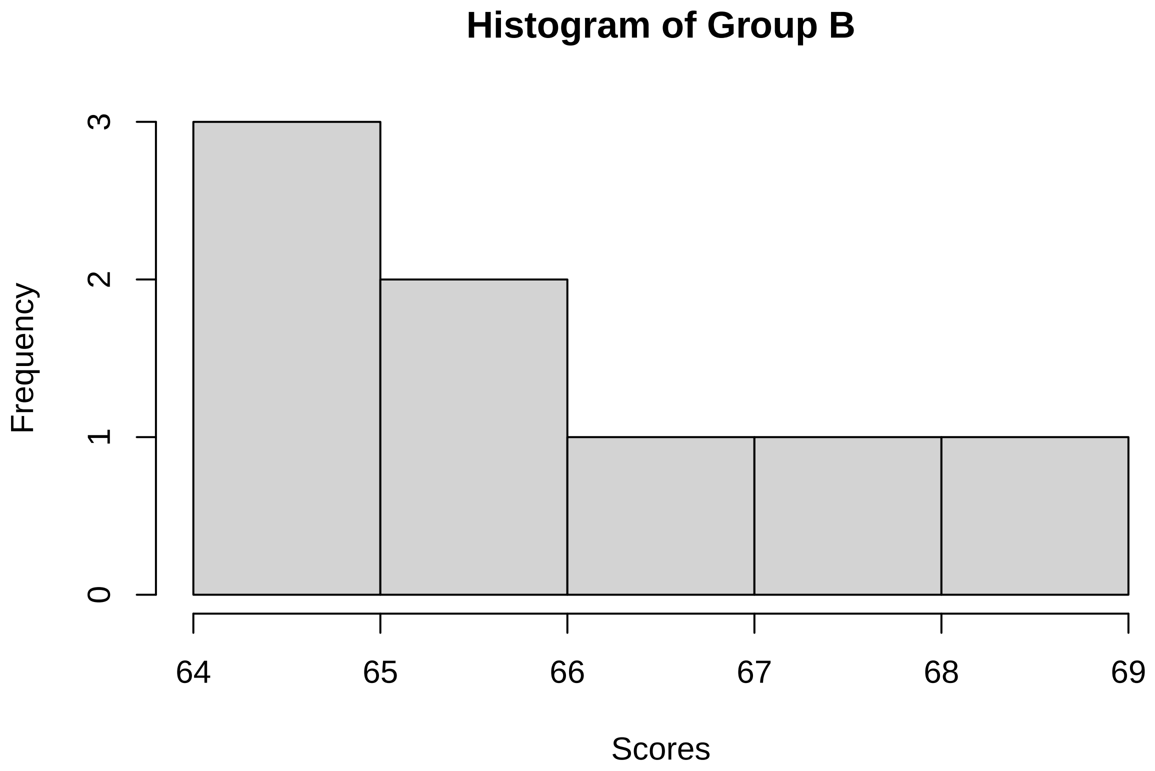
```r
# Marks of Group A
groupA <- c(56, 60, 58, 62, 59, 61, 57, 63)

# Marks of Group B
groupB <- c(64, 66, 65, 67, 68, 66, 69, 65)

# Two-sample t-test
tv=t.test(groupA, groupB, alternative = "two.sided", var.equal=FALSE)
if (tv[3]<0.05) print("Reject H_0") else print("Fail to reject H_0")
```

```
## [1] "Reject H_0"
```

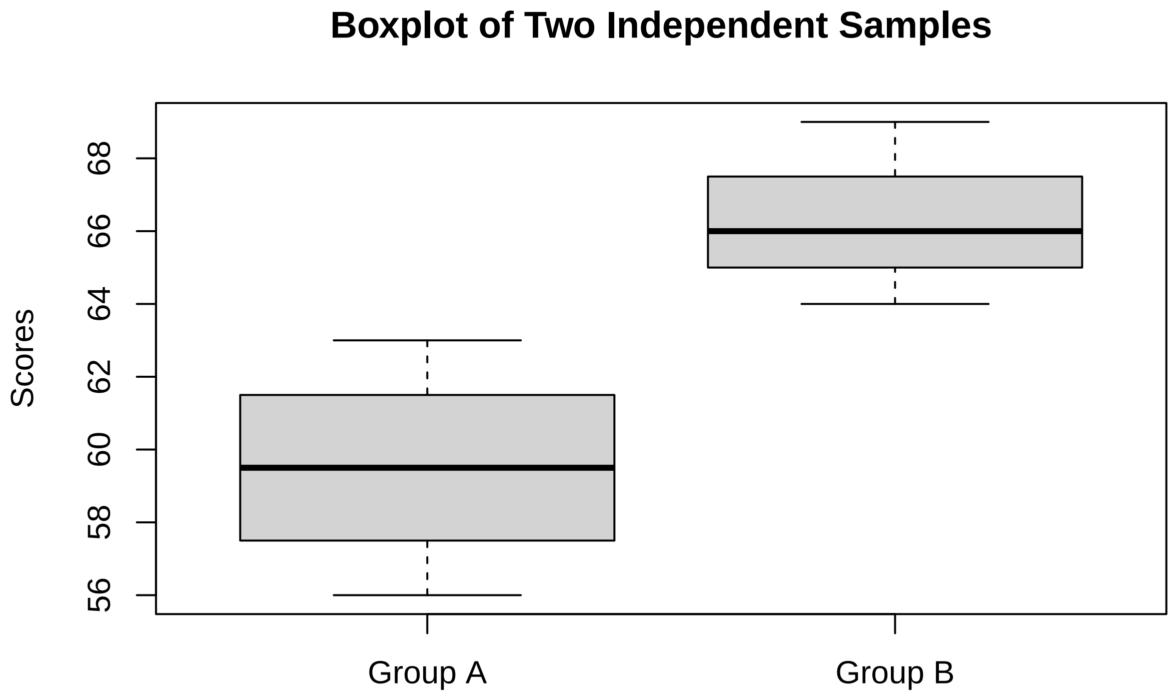
```r
hist(groupA, main = "Histogram of Group A", xlab = "Scores")
```



```r
hist(groupB, main = "Histogram of Group B", xlab = "Scores")
```



```
boxplot(groupA, groupB,
        names = c("Group A", "Group B"),
        main = "Boxplot of Two Independent Samples",
        ylab = "Scores")
```



Paired t-test

$$t = \frac{\mu_d}{\frac{s}{\sqrt{n}}}$$

Problem: The marks obtained by 8 students in a Mathematics test were recorded before and after a special coaching programme.

$$\text{Before: } 52, 55, 50, 48, 60, 57, 53, 56$$

$$\text{After: } 58, 60, 55, 54, 65, 62, 59, 61$$

At the 5% level of significance, test whether the coaching programme has significantly improved the students' performance.

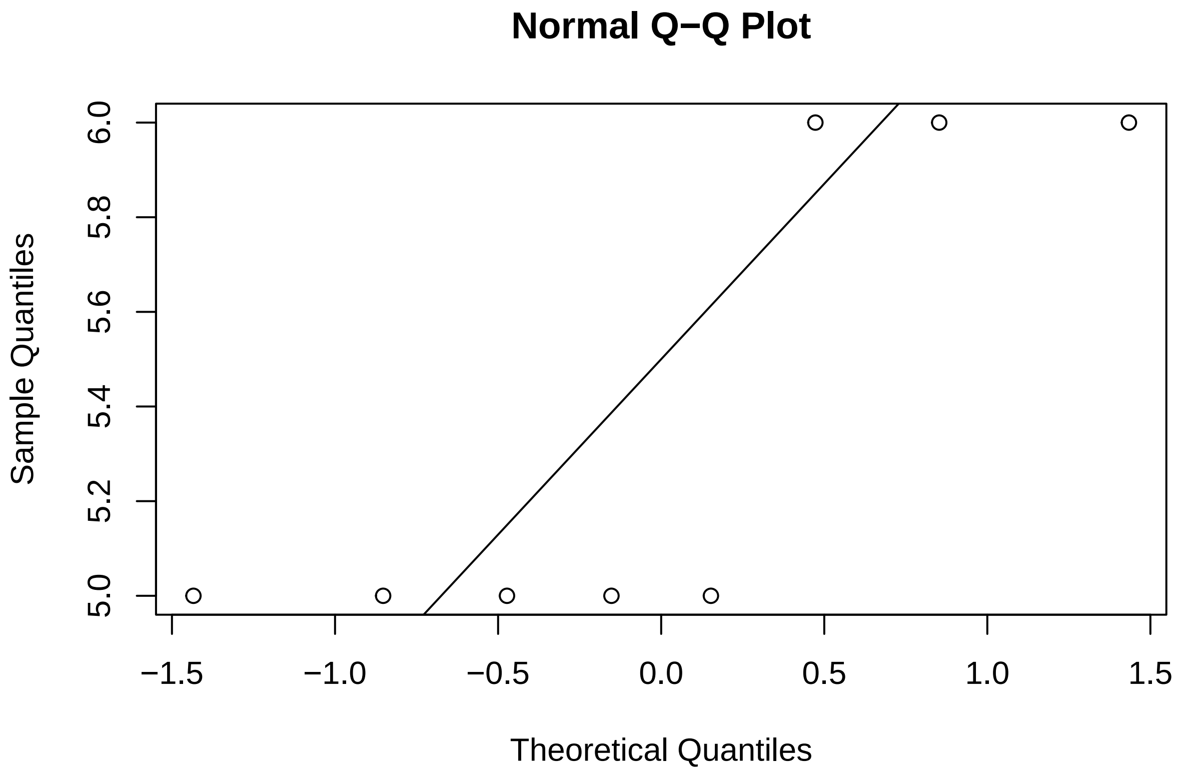
$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 < 0$$

```r
# Marks before coaching
before <- c(52, 55, 50, 48, 60, 57, 53, 56)

# Marks after coaching
after <- c(58, 60, 55, 54, 65, 62, 59, 61)

# Paired t-test
t.test(after, before, paired = TRUE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  after and before
## t = 29.375, df = 7, p-value = 1
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##      -Inf 5.721672
## sample estimates:
## mean difference
##           5.375
```

```
diff=after-before

qqnorm(diff)
qqline(diff)
```



Two sample variance test

A manufacturer produces light bulbs in two different factories: Factory A and Factory B. A quality control engineer wants to check whether the variability in the lifetime of bulbs from the two factories is the same.

The lifetimes (in hours) of randomly selected bulbs from each factory are:

Factory A: 1020, 980, 1005, 1015, 990, 995, 1000

Factory B: 1010, 1030, 1025, 1000, 1015, 1020

Test at 5% significance level whether the variances of the lifetimes from the two factories are equal.

$$H_0 : \sigma_A^2 = \sigma_B^2$$

$$H_0 : \sigma_A^2 \neq \sigma_B^2$$

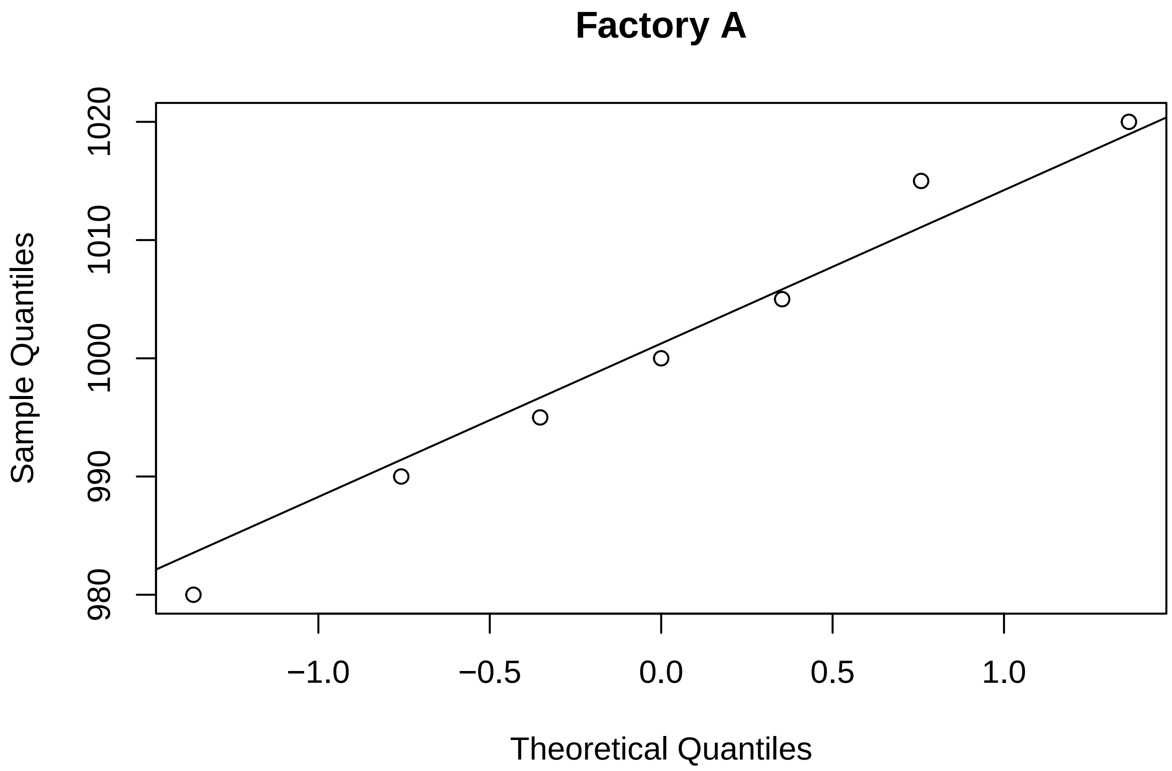
Test statistic:

$$F = \frac{S_1^2}{S_2^2}$$

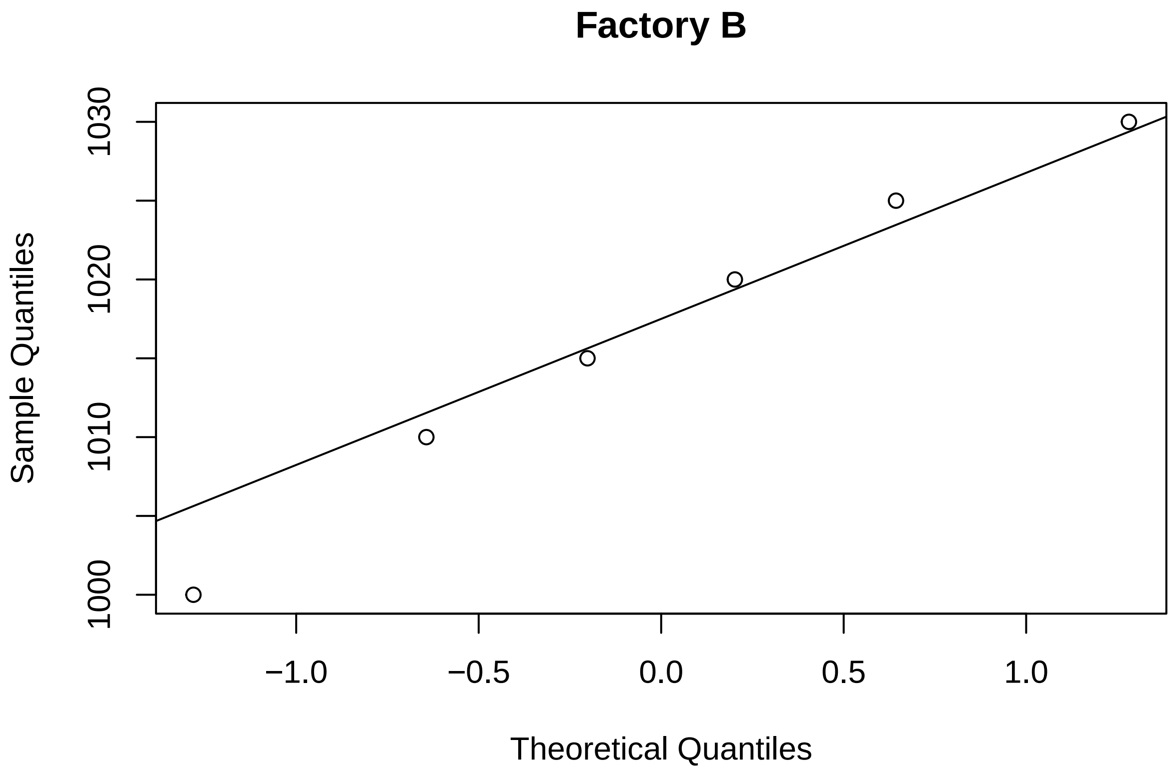
```r
# Data
factoryA <- c(1020, 980, 1005, 1015, 990, 995, 1000)
factoryB <- c(1010, 1030, 1025, 1000, 1015, 1020)

# Two-sample variance test (F-test)
var_test <- var.test(factoryA, factoryB,ratio=1)

# Display the result
print(var_test)
```

```
##
##  F test to compare two variances
##
## data:  factoryA and factoryB
## F = 1.6735, num df = 6, denom df = 5, p-value = 0.5886
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.239831 10.020007
## sample estimates:
## ratio of variances
##           1.673469
```

```r
#Normality checking
qqnorm(factoryA,main="Factory A")
qqline(factoryA)
```



```r
qqnorm(factoryB,main="Factory B")
qqline(factoryB)
```


Factory B
Sample Quantiles
1000
1010
1020
1030
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles