Analysis of Variance (ANOVA)

2021

Analysis of Variance (ANOVA)

In ANOVA we assume that the considered random variables have finite variance.

From practical point of view the **Analysis of Variance (ANOVA) / дисперсионен анализ/** is used for checking if the influence of one or more independent variables on one dependent variable Y is statistically significant. The dependent variable is obligatory numeric. As far as regression analysis is a better approach to model dependencies between numerical variables the Analysis of Variance is usually applied when at least one of the independent variables is categorical.

When we have one independent variable we speak about **One-way Analysis of Variance (ANOVA)**. When we have two independent variables we speak about **Two-way Analysis of Variance (ANOVA)**. When we have more independent variables we speak about **Multiple Analysis of Variance ANOVA**. Its methodology is developed by **Ronald Fisher** and **Jerzy Neyman**.

One-way Analysis of Variance (ANOVA)

The terminology of ANOVA is largely from the **statistical Design of experiments**. The experimenter adjusts factors (different values of independent variables) and measures responses in an attempt to determine an **effect**.

Suppose the **independent variable (factor)** X (categorical) has k **groups and denote by** Y the dependent variable and by

$Y_i := (Y|X = Category), i = 1, 2, \ldots, k.$

We assume that $Y_i \in N(\mu_i; \sigma^2)$, $i = 1, 2, \ldots, k$ are independent.

Suppose that we have $n = n_1 + n_2 + \ldots + n_k$ independent observations on random variable Y and the results are grouped according to values of the independent categorical variable (let's call it X). n_1 of them fall in the Category 1 of X, n_2 fall in the Category 2 of X, and so on n_k fall in the Category k of X.

X	Y	Average (Sample mean)
Category 1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\overline{Y}_1$
Category 2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\overline{Y}_2$
...	...	...
Category k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$	$\overline{Y}_k$

Therefore, the model is

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \varepsilon_{ij} \in N(0,\sigma_\varepsilon^2) \text{ i.i.d}$$

The last means that by assumption the data are normal, independent and **homoscedastic**.

The requirement for normality could be replaced with a requirement in each group to have large sample size and the observed random variables to have finite variance. If this assumption is violated we can use **non-parametric** approaches and to use medians instead of the means (**Kruskal-Wallis**).

Let us summarize the assumptions:

- Independence of observations.
- Normality - the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called **homoscedasticity** - the variance of data in groups should be the same.

The component

$$\mu_i = \mathbb{E}Y_i = \mathbb{E}(Y \mid X = Category\ i), i = 1, 2, \ldots, k.$$

is called **explained from the factor X.**

The component ε_{ij} is called unexplained from the model, or this is the random residual.

As we already see `t-test` was used to test hypotheses about the mean of two independent samples.

Analysis of variance (ANOVA) is used to compare the means for 2 or more independent samples. It is a generalization of the two-sample `t-test`. It is a hypothesis test to see if the means of the variables in different categories are all equal. Therefore, it is a generalization to corresponding `t-test`.

In the typical application of ANOVA, the null hypothesis is that all groups are random samples from the same population. More precisely, $H_0 : \mathbb{E}Y_1 = \ldots = \mathbb{E}Y_k$ The last means that the **differences** between the averages of the observed data in different groups **are not statistically significant** which is the same the categorical variable according to which we make differentiation between different groups does not have statistically significant influence of the dependent variable Y.

For example if the milelife of a tire depends of its manufacturer, then in one and the same of these subpopulations we would have tires produced by one and the same manufacturer and k would be the number of manufacturers. Y_i would be milelife of the considered tires produced by i-th manufacturer, $i = 1, 2, \ldots, k$.

The alternative hypothesis would be

H_A : At least one of $\mathbb{E}Y_1, \ \ldots, \mathbb{E}Y_k$ is different from the others. The **differences** between the sample means in different groups **are statistically significant** which is the same as the categorical variable according to which we have made differentiation between the groups have statistically significant influence of the dependent variable Y.

Let us denote **overall (grand) mean /общо средно/** by

$$\overline{Y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}}{n},$$

Total Sum of Squares (SS_{Total}) /обща сума от квадратите/

$$SS_{Total} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2$$

Total Sample Variance (S^2_{Total}, S^2_Y) or Mean Square Total (MS_{Total})

$$MS_{Total} \equiv S^2_{Total} \equiv S^2_Y := \frac{SS_{Total}}{n-1}.$$

The denominator is the degrees of freedom of the numerator.

ANOVA separates this Total Sample Variance into a variance explained from the factor X and unexplained variance.

First we divide the **Total Sum of squares** to a **Sum of Square explained from the factor X (The Sum of squares Between different Groups)** (therefore, we denote it by SS_B) and **Unexplained Sum of Squares (The Sum of squares Within the Groups)** (therefore, we denote it by SS_W**).**

$$SS_{Total} = SS_B + SS_W,\ SS_B := \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2,$$

$$SS_W := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

Between groups is frequently call also **treatment effect**.

The statistical significance of the experiment is determined by a ratio of two variances.

The sample variance explained from the factor X (The Variance Between different Groups, Mean Square Between these groups) (therefore, we denote it by $S^2_B = MS_B$) and **Unexplained**

Variance (The Variance Within the Groups, Mean Square Within these groups) (therefore, we denote it by $S_W^2 = MS_B$)**. Their sum is the Total variance (**S_{Total}^2**)**

$$S_B^2 \equiv MS_B = \frac{SS_B}{k-1}, S_W^2 \equiv MS_W = \frac{SS_W}{n-k}, S_{Total}^2 \equiv MS_{Total} = \frac{SS_{Total}}{n-1}.$$

Here the denominators are the degrees of freedom of the corresponding numerators. Note that

$$(n-1) = (k-1) + (n-k)$$

Now we can reformulate the above hypothesis in the following way

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

$H_A : \sigma_B^2 \neq \sigma_W^2$

and we can use the **Fisher test**. The ratio that we are going to use is

$$(F_{stat} | H_0) = \left(\frac{MS_B}{MS_W} \middle| H_0 \right) \in F(k-1, n-k)$$

The first parameter is the degrees of freedom of MS_B. The second parameter is the degrees of freedom of MS_W.

This ratio is independent of several possible alterations to the experimental observations:

- Adding a constant to all observations does not alter significance.
- Multiplying all observations by a constant does not alter significance.

If H_0 is true we would have $F_{stat} \approx 1$.

If H_A is true we would have $F_{stat} \neq 1$.

More precisely

$$W_\alpha = \left\{ \frac{MS_B}{MS_W} \geq x_{1-\alpha, F(k-1, n-k)} \right\}$$

When we replace the values from the sample in $F_{stat} = \dfrac{MS_B}{MS_W}$ we obtain F_{emp} and $p-value = \mathbb{P}(F_{stat} > F_{emp} \,|\, H_0) = \mathbb{P}(\eta > F_{emp})$ where $\eta \in F(k-1, n-k)$.

Example 1

In order to check if the milelife (in 1000 km) of a tire depends of its manufacturer (A, B, C, D) let us suppose that we have observations on 490 tires. In order to obtain the data simulate

- 100 observations on $(Y \,|\, X = A) \in N(30, 5^2)$ to be milelife of tires that are produced by the manufacturer A,
- 120 observations on $(Y \,|\, X = B) \in N(32, 3^2)$ to be milelife of tires that are produced by the manufacturer B,
- 130 observations on $(Y \,|\, X = C) \in N(27, 4^2)$ to be milelife of tires that are produced by the manufacturer C,
- 140 observations on $(Y \,|\, X = D) \in N(30, 3^2)$ to be milelife of tires that are produced by the manufacturer D.

a. Check if the manufacturer has statistically significant influence on the milelife of the tires

b. Extract the residuals and check if they are normal

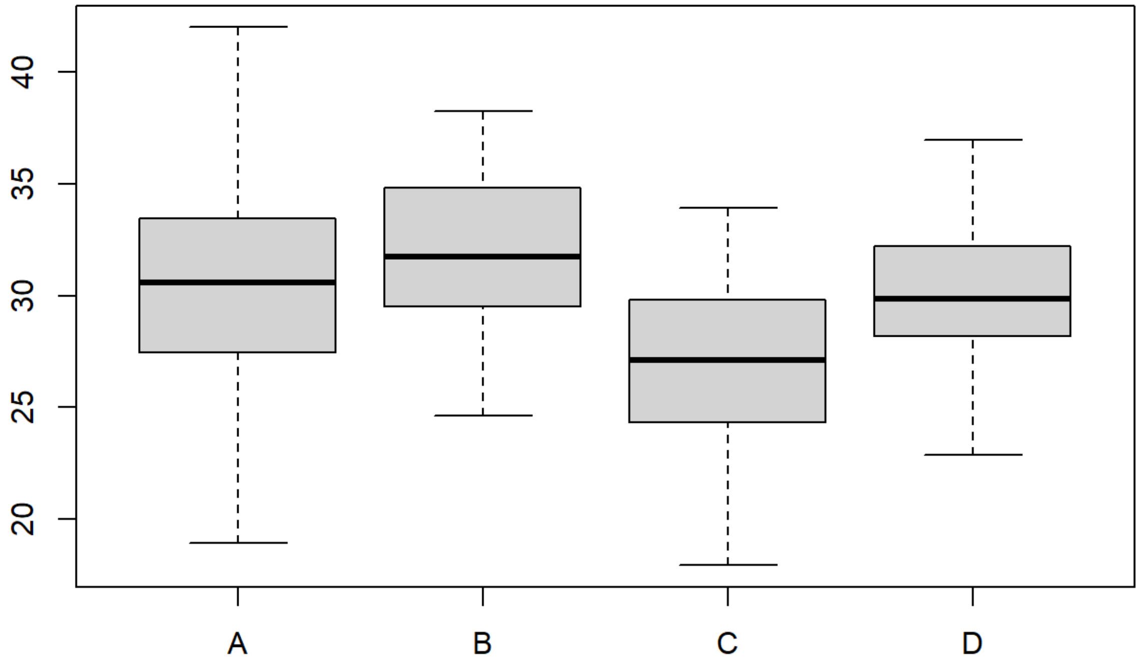
c. If the answer in a) is “Yes” which is the most influential group?

d.

```
> set.seed(1)
> Y_A <- rnorm(100, 30, 5)
> set.seed(2)
> Y_B <- rnorm(120, 32, 3)
> set.seed(3)
> Y_C <- rnorm(130, 27, 4)
> set.seed(4)
> Y_D <- rnorm(140, 30, 3)
> all <- c(Y_A, Y_B, Y_C, Y_D)
> meanAll <- mean(all)
> meanGroups <- c(mean(Y_A), mean(Y_B), mean(Y_C),
mean(Y_D)); meanGroups
[1] 30.54444 32.09595 27.01025 30.16771
> var <- c(var(Y_A), var(Y_B), var(Y_C), var(Y_D)); var
[1] 20.169052 11.529016 11.539114  8.167186
```

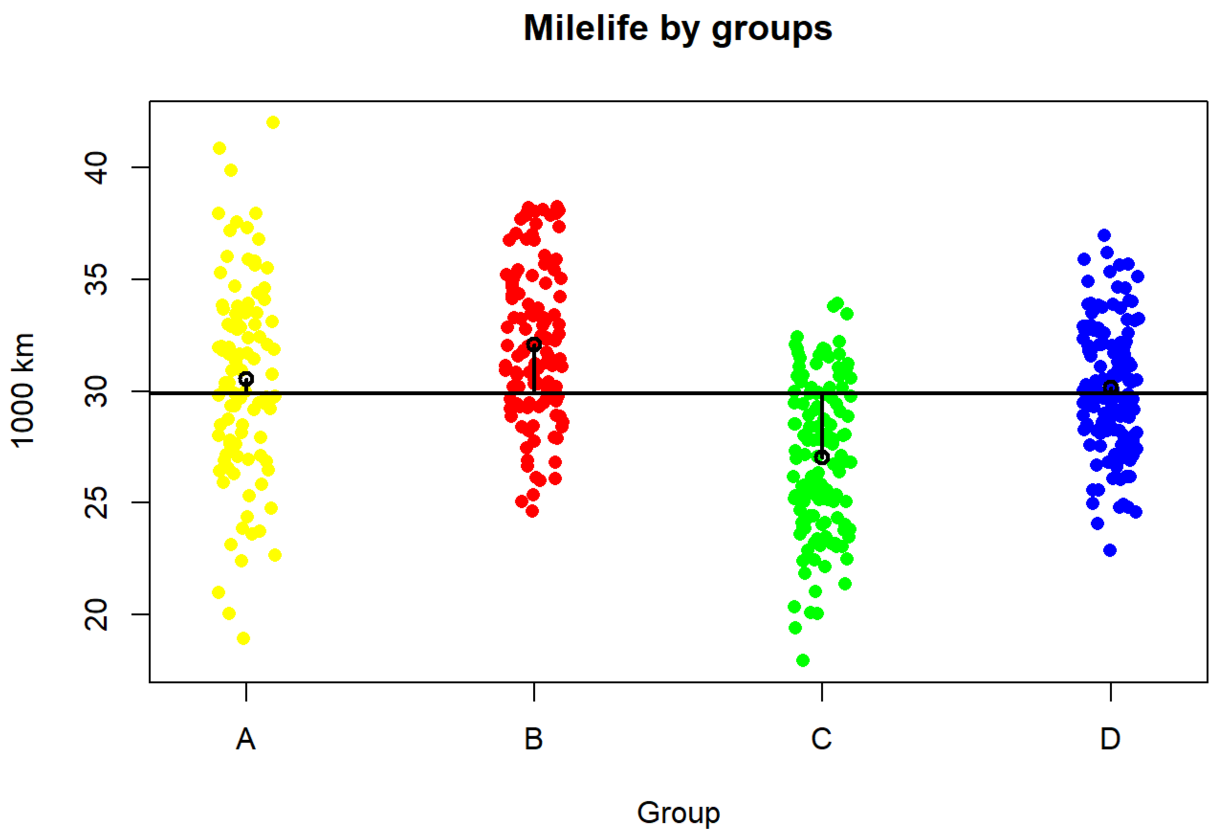
First let’s compare the distributions of the samples in different groups.

```
> boxplot(Y_A, Y_B, Y_C, Y_D, names = c("A", "B", "C",
"D"))
```



or

```
> milelife <- list("A" = Y_A, "B" = Y_B, "C" = Y_C, "D" =
Y_D)
> stripchart(milelife,
+            main = "Milelife by groups",
+            ylab = "1000 km",
+            xlab = "Group",
+            method = "jitter",
+            col = c("yellow", "red", "green", "blue"),
+            pch = 16,
+            vertical = TRUE)
> abline(h = meanAll, lwd = 2)
> segments(c(1, 2, 3, 4), meanGroups, c(1, 2, 3, 4),
rep(meanAll, 4), lwd = 2)
> points(meanGroups ~ c(1, 2, 3, 4), lwd = 2)
```



Are the differences between these means statistically significant?

$$H_0 : \mathbb{E}(Y|X = A) = \mathbb{E}(Y|X = B) = \mathbb{E}(Y|X = C) = \mathbb{E}(Y|X = D)$$

H_A : At least two of $\mathbb{E}(Y|X = A)$, $\mathbb{E}(Y|X = B)$,
$\mathbb{E}(Y|X) = C$ and $\mathbb{E}(Y|X = D)$ are different.

or which is the same as

$H_0 : \sigma_B^2 = \sigma_W^2$
$H_A : \sigma_B^2 \neq \sigma_W^2$

First we organize the data as a data frame

```
> data <- data.frame(Y = c(Y_A, Y_B, Y_C, Y_D),
+                     Manufacturer = factor(rep(c("A",
"B", "C", "D"),
+                                                  times =
c(length(Y_A), length(Y_B),
+
length(Y_C), length(Y_D)))))
> head(data)
         Y Manufacturer
1 26.86773            A
2 30.91822            A
3 25.82186            A
4 37.97640            A
5 31.64754            A
6 25.89766            A
```

Then we perform ANOVA via the functions `aov` and `anova`

```
> myanova <- aov(Y ~ Manufacturer, data = data)
> anova(myanova)
Analysis of Variance Table

Response: Y
              Df Sum Sq Mean Sq F value    Pr(>F)    
Manufacturer   3 1715.6  571.87  46.379 < 2.2e-16 ***
Residuals    486 5992.5   12.33                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

The first three numbers in the row `Manufacturer` are for the degrees of freedom, SS_B and MS_B between the groups.

The first three numbers in the row `Residuals` are for the degrees of freedom, SS_W and MS_W within the groups.

The number `F value` is for

$$F_{emp} = \frac{MS_B}{MS_W} = 46.379$$

The corresponding $p-value = \mathbb{P}(F_{stat} > F_{emp} \,|\, H_0) = \mathbb{P}(\eta > F_{emp}) < 2.2e-16 < 0.05 = \alpha$, so we reject H_0. According to the data the manufacturer is statistically significant for the milelife of the tires.

We can use also

```
> summary(myanova)
              Df Sum Sq Mean Sq F value Pr(>F)    
Manufacturer   3   1716   571.9   46.38 <2e-16 ***
Residuals    486   5992    12.3                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

or

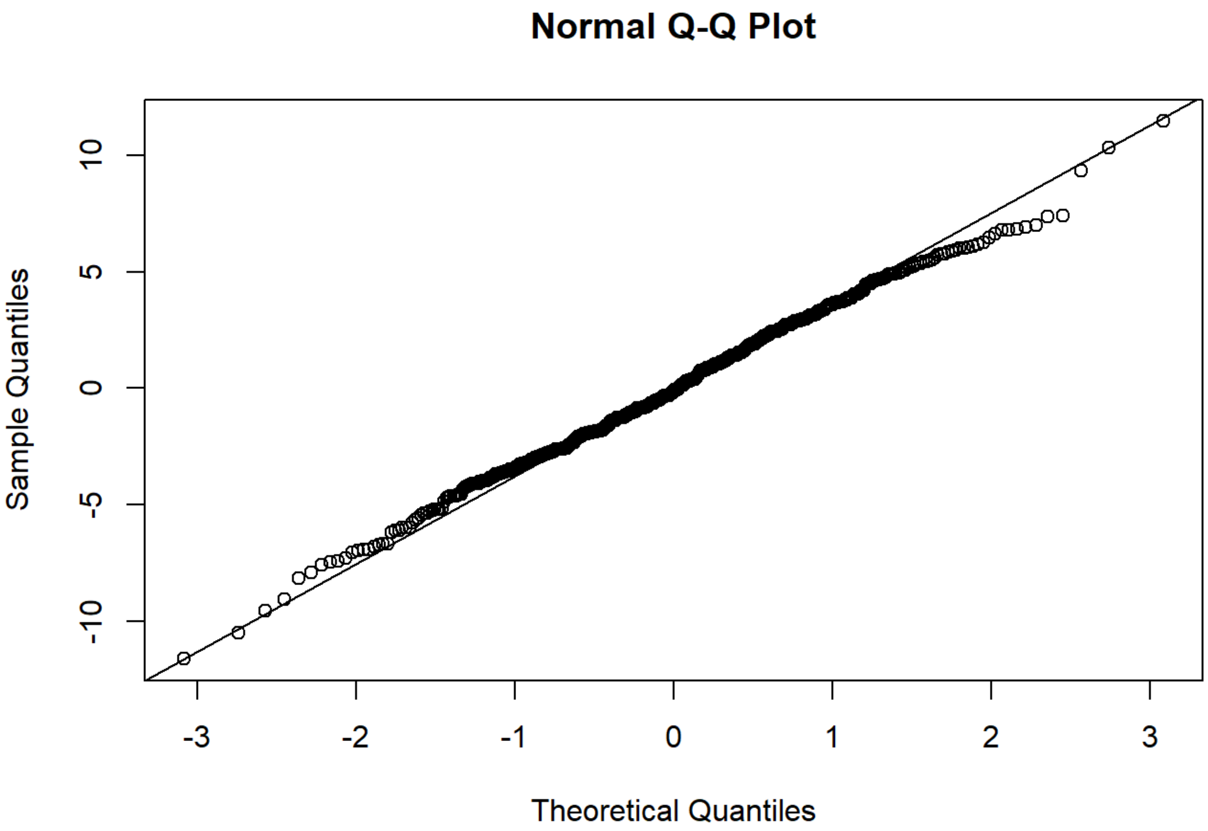
```
> oneway.test(Y ~ Manufacturer, data = data, var.equal =
TRUE)

    One-way analysis of means

data:  Y and Manufacturer
F = 46.379, num df = 3, denom df = 486, p-value < 2.2e-16
```

b. we can see the attributes of `myanova`

```
> attributes(myanova)
$names
 [1] "coefficients"  "residuals"     "effects"      
"rank"         
 [5] "fitted.values" "assign"        "qr"           
"df.residual"  
 [9] "contrasts"     "xlevels"       "call"         
"terms"        
[13] "model"        

$class
[1] "aov" "lm" 
```

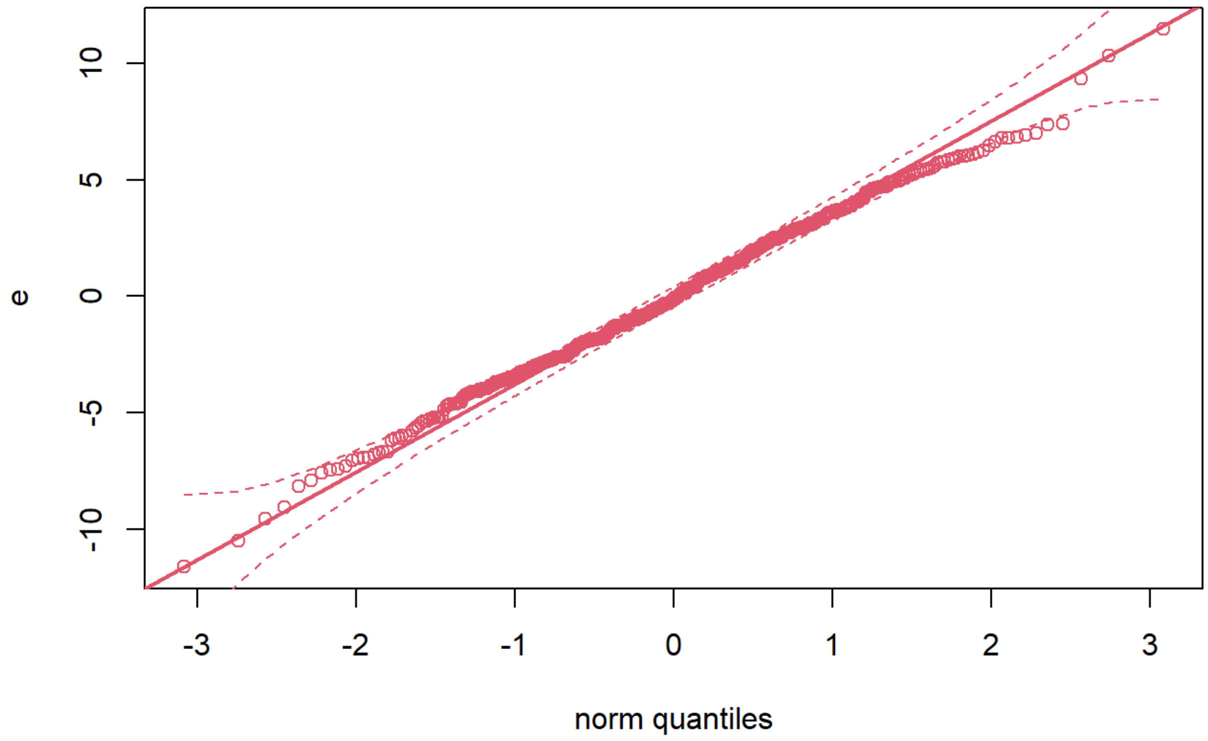
First we make the normal `qq-plot`

```
> e <- myanova$residuals
> qqnorm(e)
> qqline(e)
```



```
> library(StatDA)
Warning: package 'StatDA' was built under R version 4.0.3
Loading required package: sgeostat
Warning: package 'sgeostat' was built under R version
4.0.3
Registered S3 method overwritten by 'geoR':
  method         from
  plot.variogram sgeostat
```

```
> qqplot.das(e)
```



We can perform also Shapiro test

H_0 : ε is normally distributed
H_A : ε is not normally distributed

Now we use the function `shapiro.test` in R

```
> shapiro.test(e)
```

```
    Shapiro-Wilk normality test

data:  e
W = 0.99721, p-value = 0.5765
```

The $p-value = 0.5765 > 0.05 = \alpha$ therefore we have no evidence to reject H_0. ε is normally distributed.

Now we have to answer the question. **How to determine which one is different?**

Example 2

Now let’s determine which manufacturer causes rejection of H_0.

We will assume that these comparisons are independent and perform all possible pairwise comparisons via `TukeyHSD` function.

We compare A and B; A and C; A and D; B and C; B and D; C and D.

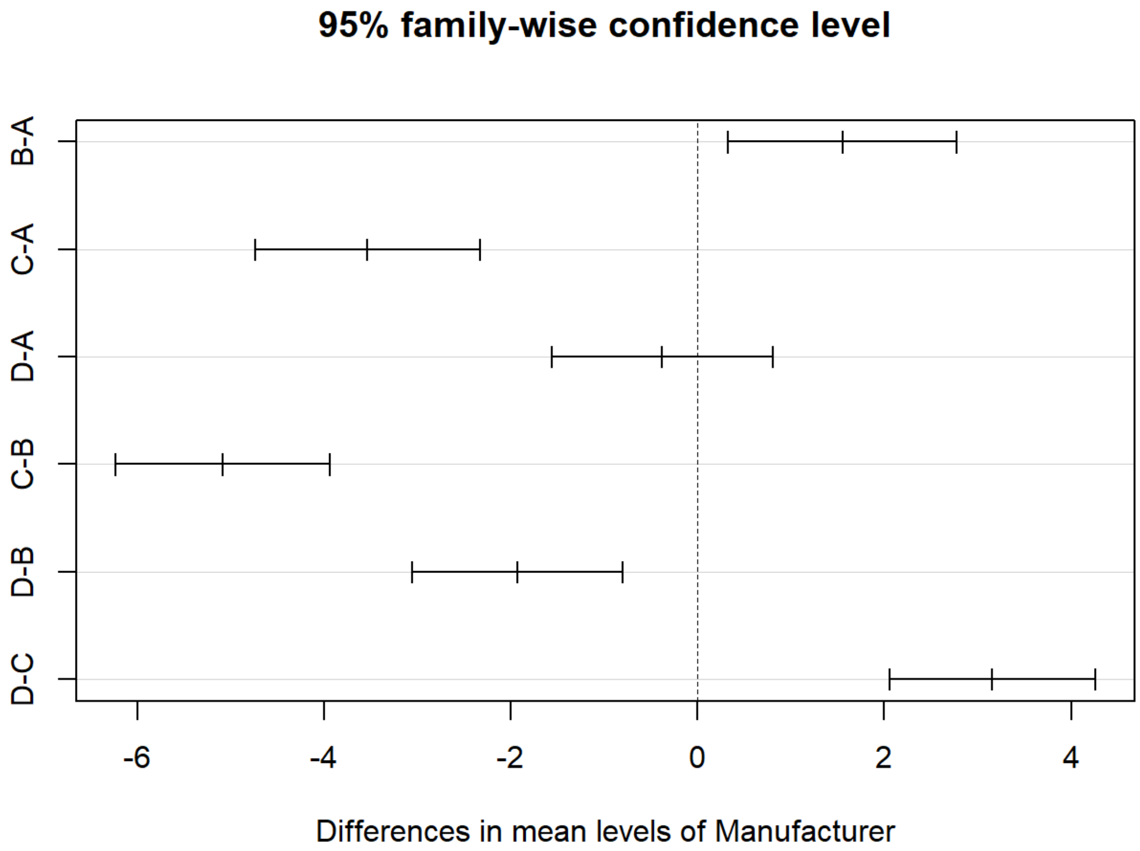
For any one of these couples we use independent two-sample t-test or the corresponding two confidence intervals.

```
> TukeyHSD(myanova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ Manufacturer, data = data)

$Manufacturer
          diff        lwr        upr     p adj
B-A  1.5515132  0.3258287  2.7771977 0.0064566
C-A -3.5341849 -4.7382507 -2.3301191 0.0000000
D-A -0.3767255 -1.5619472  0.8084962 0.8452953
C-B -5.0856981 -6.2316481 -3.9397481 0.0000000
D-B -1.9282387 -3.0543725 -0.8021048 0.0000736
D-C  3.1574594  2.0548945  4.2600243 0.0000000
```

The overall confidence is $95\ \%$ and the **adjusted p-values** show that only the difference between A and D is not statistically significant.

We can plot the corresponding confidence intervals by

```
> plot(TukeyHSD(myanova))
```



We observe that only the confidence interval between the means for A and D contains zero. The last means again that this difference is not statistically significant.

Example 3

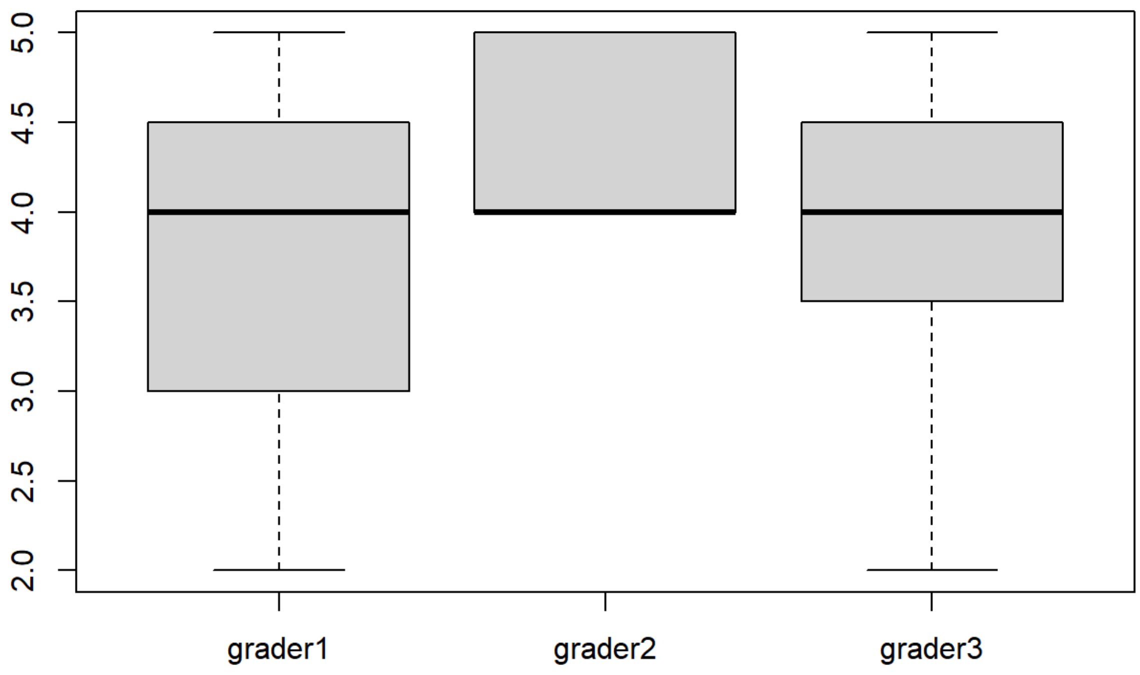
Suppose a school is trying to grade 300 different scholarship applications. As the job is too much work for one grader, suppose 6 are used. The scholarship committee would like to ensure that each grader is using the same grading scale, as otherwise the students aren't being treated equally. One approach to checking if the graders are using the same scale is to randomly assign each grader 50 exams and have them grade. Then compare the grades for the 6 graders knowing that the differences should be due to chance errors if the graders all grade equally.

To illustrate, suppose we have just 24 tests and 3 graders. Furthermore, suppose the grading scale is on the range $1-5$ with 5 being the best and the scores are reported as

```
> grader1 <- c(4, 3, 4, 5, 2, 3, 4, 5)
> grader2 <- c(4, 4, 5, 5, 4, 5, 4, 4)
> grader3 <- c(3, 4, 2, 4, 5, 5, 4, 4)
> scores <- data.frame(grader1, grader2, grader3)
```

First let's compare the three distributions.

```
> boxplot(scores)
```

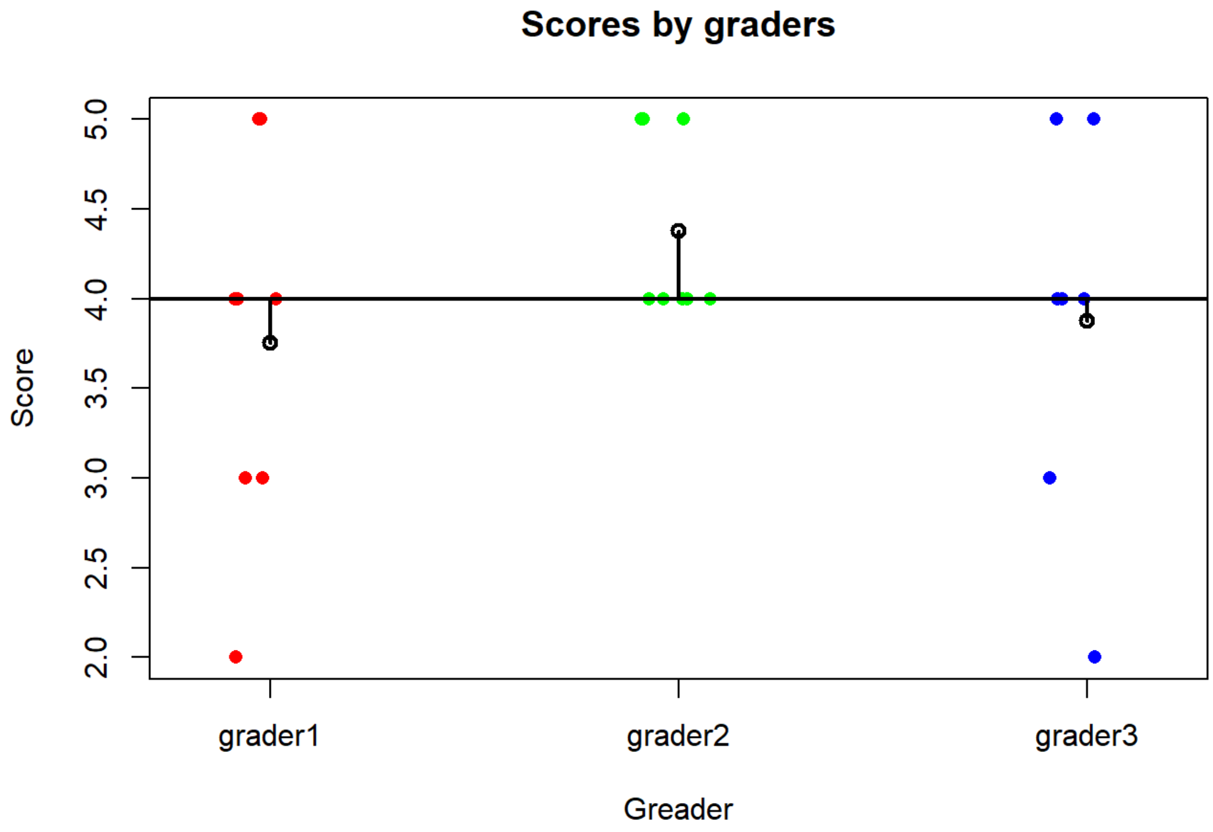


From this graph it appears that grader 2 is different from graders 1 and 3.

Or

```
> scoresStack <- stack(scores)
> head(scoresStack)
  values     ind
1      4 grader1
2      3 grader1
3      4 grader1
4      5 grader1
5      2 grader1
6      3 grader1
```

```
> meanAll <- mean(scoresStack$values); meanAll
[1] 4
> meanGroups <- c(mean(grader1), mean(grader2),
mean(grader3))
> stripchart(scores,
+            main = "Scores by graders",
+            ylab = "Score",
+            xlab = "Greader",
+            method = "jitter",
+            col = c("red", "green", "blue"),
+            pch = 16,
+            vertical = TRUE)
> abline(h = meanAll, lwd = 2)
> segments(c(1, 2, 3), meanGroups, c(1, 2, 3),
rep(meanAll, 3), lwd = 2)
> points(meanGroups ~ c(1, 2, 3), lwd = 2)
```



Are the differences between these means statistically significant?

$H_0 : \mathbb{E}(Y|X = Grader\ 1) = \mathbb{E}(Y|X = Grader\ 2) = \mathbb{E}(Y|X = grader\ 3)$
H_A : At least two of $\mathbb{E}(Y|X = grader\ 1)$,
$\mathbb{E}(Y|X = Grader\ 2)$ and $\mathbb{E}(Y|X = Grader\ 3)$ are different.

or which is the same as

$$H_0 : \sigma_B^2 = \sigma_W^2$$
$$H_A : \sigma_B^2 \neq \sigma_W^2$$

Analysis of variance allows us to investigate if all the graders have the same mean. We can run this using the `oneway.test` on the data organized as a data frame containing single variable holding the scores, and a factor variable describing the grader or category. The formula notation of `oneway.test` is `dependent variable ~ independent variables`

```
> oneway.test(values ~ ind, data = scoresStack, var.equal
= TRUE)

	One-way analysis of means

data:  values and ind
F = 1.1308, num df = 2, denom df = 21, p-value = 0.3417
```

The $F_{emp} = 1.1308$ the critical value is the quantile of $F(2{,}21)$ distribution. The $p-value = 0.3417 > 0.05 = \alpha$, so we have no evidence to reject the H_0, so we accept the H_0 that the variables have equal means.

Another way is by using the `anova` function

```
> anova(lm(values ~ ind, data = scoresStack))
Analysis of Variance Table

Response: values
          Df Sum Sq Mean Sq F value Pr(>F)
ind        2   1.75 0.87500  1.1308 0.3417
Residuals 21  16.25 0.77381
```

The row `ind` gives df of between sum of squares. They are $k-1 = 3-1 = 2$. The $Sum\ Sq = SS_B$. Let us remind that

$$SS_{Total} = SS_B + SS_W, SS_B := \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2 = 1.75,$$

$$SS_W := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 = 16.25$$

Here the denominators are the degrees of freedom of the corresponding numerators.

Note that

$$n = 24, k = 3, (n-1) = (k-1) + (n-k)$$

When we replace the values from the sample in $F_{stat} = \dfrac{MS_B}{MS_W}$ we obtain $F_{emp} = 1.1308$.

The final $p - value = \mathbb{P}(F_{stat} > F_{emp} \,|\, H_0) = \mathbb{P}(\eta > F_{emp}) = 0.3417$, where $\eta \in F(k-1, n-k) \equiv F(3-1, 24-3)$.

We can also use

```
> summary(aov(values ~ ind, data = scoresStack))
            Df Sum Sq Mean Sq F value Pr(>F)
ind          2   1.75  0.8750   1.131  0.342
Residuals   21  16.25  0.7738
```

Or we can use the formulas to calculate them and we can check the above results.

- The means within the groups $Y, i = 1, 2, 3$

```
> grader1mean <- mean(grader1); grader1mean
[1] 3.75
> grader2mean <- mean(grader2); grader2mean
[1] 4.375
> grader3mean <- mean(grader3); grader3mean
[1] 3.875
```

- The overall mean $\overline{Y}$

```
> allmean <- mean(scoresStack$values); allmean
[1] 4
```

- The sum of squares SS_B - sum of squares between the diferent groups (treatment effect)

```
> n1 <- length(grader1)
> n2 <- length(grader2)
> n3 <- length(grader3)
> betweenSS <- n1 * (grader1mean - allmean)^2 +
+              n2 * (grader2mean - allmean)^2 +
+              n3 * (grader3mean - allmean)^2
> betweenSS
[1] 1.75
```

SS_W - sum of squares within the groups

```
> withinSS <- sum((grader1 - grader1mean)^2) +
+             sum((grader2 - grader2mean)^2) +
+             sum((grader3 - grader3mean)^2)
> withinSS
[1] 16.25
```

SS_{Total} - total sum of squares

```
> totalSS <- sum((grader1 - allmean)^2) +
+            sum((grader2 - allmean)^2) +
+            sum((grader3 - allmean)^2)
> totalSS
[1] 18
```

Now we observe that $SS_{Total} = SS_B + SS_W$

- The degrees of freedom

```
> n <- n1 + n2 + n3
> k <- 3
> n-1
[1] 23
```

```
> k-1
[1] 2
> n-k
[1] 21
```

- The mean squares

$MS_W = S_W^2$ - mean sum of squares between the different groups

```
> betweenMS <- betweenSS / (k - 1)
> betweenMS
[1] 0.875
```

$MS_{Total} = S_{Total}^2 = S_Y^2$ - mean sum of squares within the group

```
> withinMS <- withinSS / (n - k)
> withinMS
[1] 0.7738095
```

$MS_{Total} = S_{Total}^2 = S_Y^2$ - Total variance

```
> totalMS <- totalSS / (n - 1)
> totalMS
[1] 0.7826087
```

or

```
> totalMS <- var(c(grader1, grader2, grader3))
> totalMS
[1] 0.7826087
```

- The $F-static$

```
> Fstatistic <- betweenMS / withinMS
> Fstatistic
[1] 1.130769
```

- Corresponding p-value

```
> pf(Fstatistic, k - 1, n - k, lower.tail = FALSE)
[1] 0.3416639
```

Kruskal-Wallis test

If we are not sure if the observed random variables are normal and if we cannot say that all of them have finite variance instead of means we use medians. The corresponding test is called **Kruskal-Wallis test**.

It is a non-parametric ranks test

$H_0 : Me(Y|X = Grader\ 1) = Me(Y|X = Grader\ 2) = Me(Y|X = Grader\ 3)$
H_A : At least two of

$Me(Y|X = Grader\ 1)$,
$Me(Y|X = Grader\ 2)$ and $Me(Y|X = Grader\ 3)$ are different.

```
> kruskal.test(values ~ ind, data = scoresStack)

    Kruskal-Wallis rank sum test

data:  values by ind
Kruskal-Wallis chi-squared = 1.9387, df = 2, p-value =
0.3793
```

The $p-value = 0.3793 > 0.05 = \alpha$, so we have no evidence to reject H_0. We accept that the medians are equal.

Example 4

In order to check if the wages (in $) of citizens in four different cities depend on the location (A, B, C, D) let us suppose that we have observations on 560 citizens. In order to obtain the data simulate observations on Pareto distributed random variable

$$Y \in Pareto(X_{\min}, \alpha), F_X(x) = \begin{cases} 0, & x < x_{\min} \\ 1 - \left(\frac{x_{\min}}{x}\right)^{\alpha}, & x \geq x_{\min} \end{cases},$$

$$F_{\overleftarrow{X}}(p) = \inf\{x \in R : F_X(x) \geq p\} = x_{\min}\left(\frac{1}{1-p}\right)^{\frac{1}{\alpha}}, p \in (0,1]$$

110 observations on $Y_A \in Pareto(650, 1.75)$ to be wage of the citizens of the city A

130 observations on $Y_B \in Pareto(690, 1.85)$ to be wage of the citizens of the city B

150 observations on $Y_C \in Pareto(750, 1.55)$ to be wage of the citizens of the city C

170 observations on $Y_D \in Pareto(800, 1.65)$ to be wage of the citizens of the city D.

Use **ANOVA** and check if the city has statistically significant influence on the wages of the citizens.

Extract the residuals and check if they are normal.

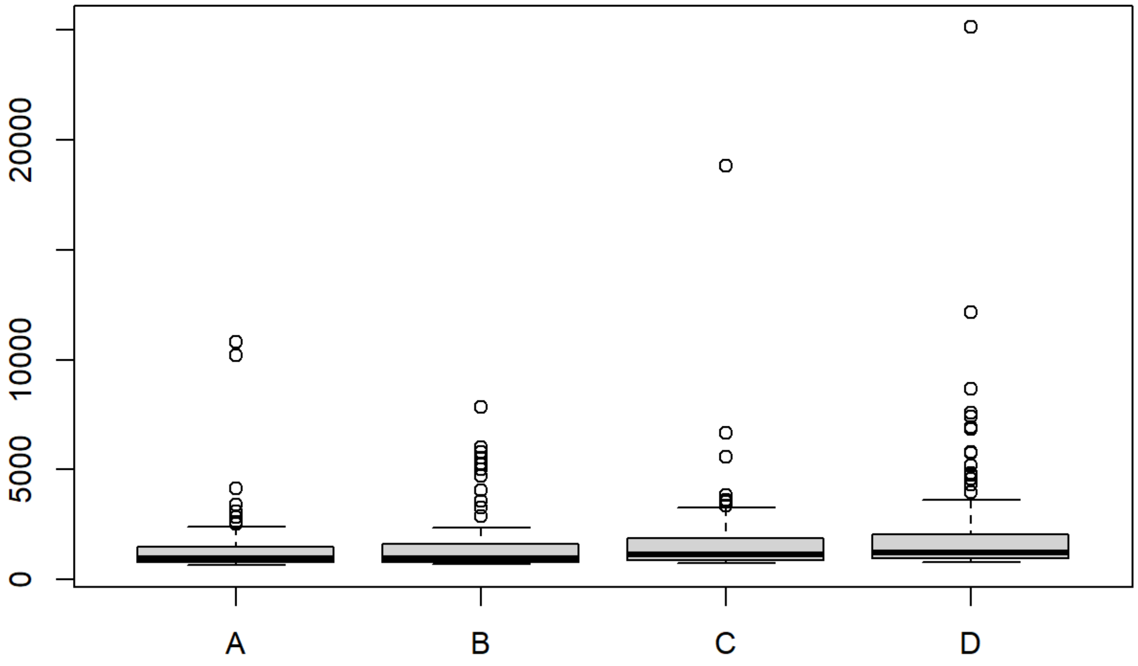
If they are not normal perform **Kruskal-Wallis test**.

Compare the results.

```
> set.seed(1)
> n_A <- 110; xmin <- 650; alpha <- 1.75
> Y_A <- xmin / (1 - runif(n_A, 0, 1))^(1/alpha)
> set.seed(2)
> n_B <- 130; xmin <- 690; alpha <- 1.85
> Y_B <- xmin / (1 - runif(n_B, 0, 1))^(1/alpha)
> set.seed(3)
> n_C <- 150; xmin <- 750; alpha <- 1.55
> Y_C <- xmin / (1 - runif(n_C, 0, 1))^(1/alpha)
> set.seed(4)
> n_D <- 170; xmin <- 800; alpha <- 1.65
> Y_D <- xmin / (1 - runif(n_D, 0, 1))^(1/alpha)
> all <- c(Y_A, Y_B, Y_C, Y_D)
> means <- c(mean(Y_A), mean(Y_B), mean(Y_C), mean(Y_D)); 
means
[1] 1379.038 1489.232 1596.143 2031.014
> var <- c(var(Y_A), var(Y_B), var(Y_C), var(Y_D)); var
[1] 1940695 1616750 2806385 5976501
```

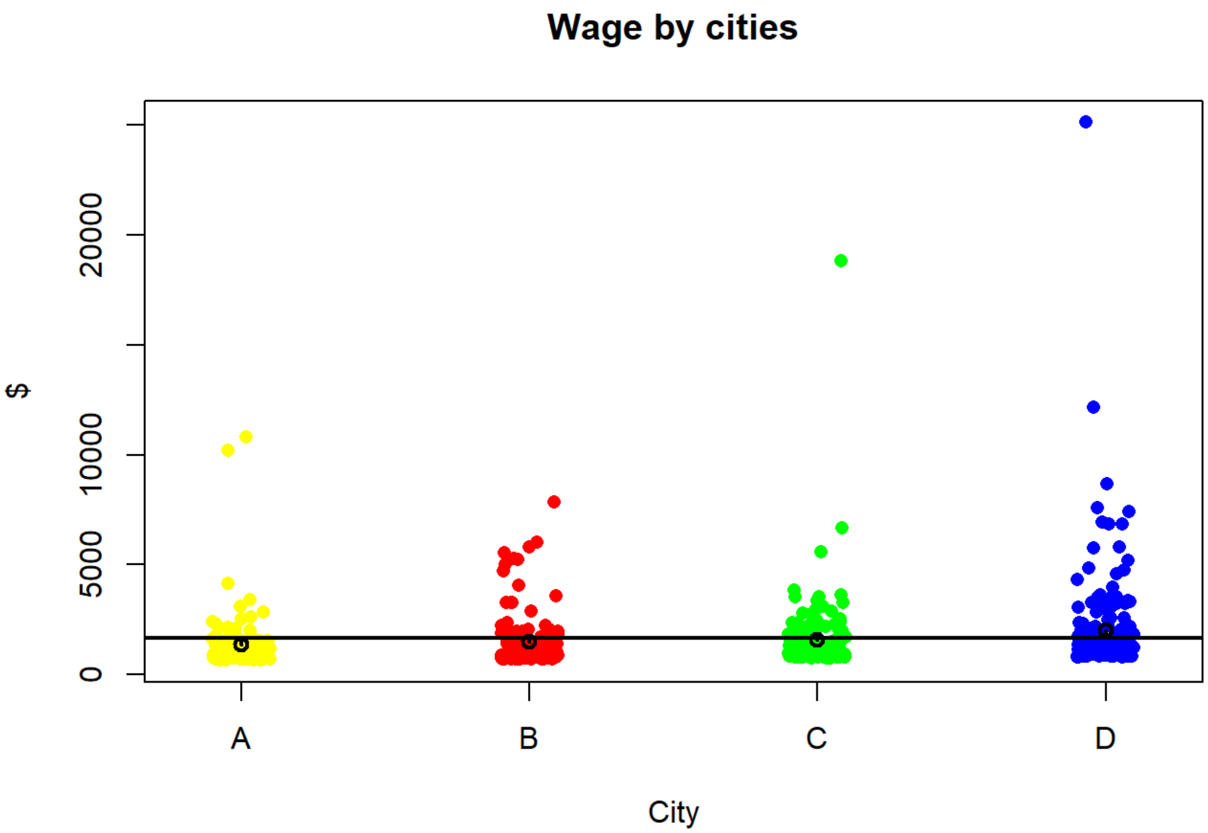
First let’s compare the distributions of the samples in different groups.

```
> boxplot(Y_A, Y_B, Y_C, Y_D, names = c("A", "B", "C",
"D"))
```



or

```
> salary <- list("A" = Y_A, "B" = Y_B, "C" = Y_C, "D" =
Y_D)
> meanAll <- mean(all)
> meanGroups <- c(mean(Y_A), mean(Y_B), mean(Y_C),
mean(Y_D))
> stripchart(salary,
+            main = "Wage by cities",
+            ylab = "$",
+            xlab = "City",
+            method ="jitter",
+            col = c("yellow", "red", "green", "blue"),
+            pch = 16,
+            vertical = TRUE)
> abline(h = meanAll, lwd = 2)
> segments(c(1, 2, 3, 4), meanGroups, c(1, 2, 3, 4),
rep(meanAll, 4), lwd = 2)
> points(meanGroups ~ c(1, 2, 3, 4), lwd = 2)
```



Are the differences between these means statistically significant?

$H_0 : \mathbb{E}(Y\,|\,X = A) = \mathbb{E}(Y\,|\,X = B) = \mathbb{E}(Y\,|\,X = C) = \mathbb{E}(Y\,|\,X = D)$
H_A : At least two of $\mathbb{E}(Y\,|\,X = A)$, $\mathbb{E}(Y\,|\,X = B)$,
$\mathbb{E}(Y\,|\,X = C)$ and $\mathbb{E}(Y\,|\,X = D)$ are different.

or which is the same as

$H_0 : \sigma_B^2 = \sigma_W^2$
$H_A : \sigma_B^2 \neq \sigma_W^2$

First let's organize the data as a data frame

```
> data <- data.frame(Y = c(Y_A, Y_B, Y_C, Y_D),
+                    City = factor(rep(c("A", "B", "C",
"D"),
+                                      times =
c(length(Y_A), length(Y_B),
+
length(Y_C), length(Y_D)))))
> head(data)
          Y City
```

```
1  775.3404    A
2  848.0353    A
3 1056.8472    A
4 2544.4607    A
5  739.2866    A
6 2400.9188    A
```

Then we perform **ANOVA** via the functions `aov` and `anova`

```
> myanova <- aov(Y ~ City, data = data)
> anova(myanova)
Analysis of Variance Table

Response: Y
           Df     Sum Sq  Mean Sq F value Pr(>F)
City        3   36486571 12162190  3.6586 0.0124 *
Residuals 556 1848276578  3324238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

The first three numbers in the row `City` are for the degrees of freedom, SS_B and MS_B.

The first three numbers in the row `Residuals` are for the degrees of freedom, SS_W and MS_W.

The number `F value` is for

$F_{emp} = \dfrac{MS_B}{MS_W} = 3.6586$The corresponding $p{-}value = \mathbb{P}(F_{stat} > F_{emp} | H_0) = \mathbb{P}(\eta > F_{emp}) = 0.0124 < 0.05 = \alpha,$ so we reject H_0. According to the data the city is statistically significant for the wage of the citizens.

We can use also

```
> summary(myanova)
             Df    Sum Sq  Mean Sq F value Pr(>F)
City          3 3.649e+07 12162190   3.659 0.0124 *
Residuals   556 1.848e+09  3324238
---
```

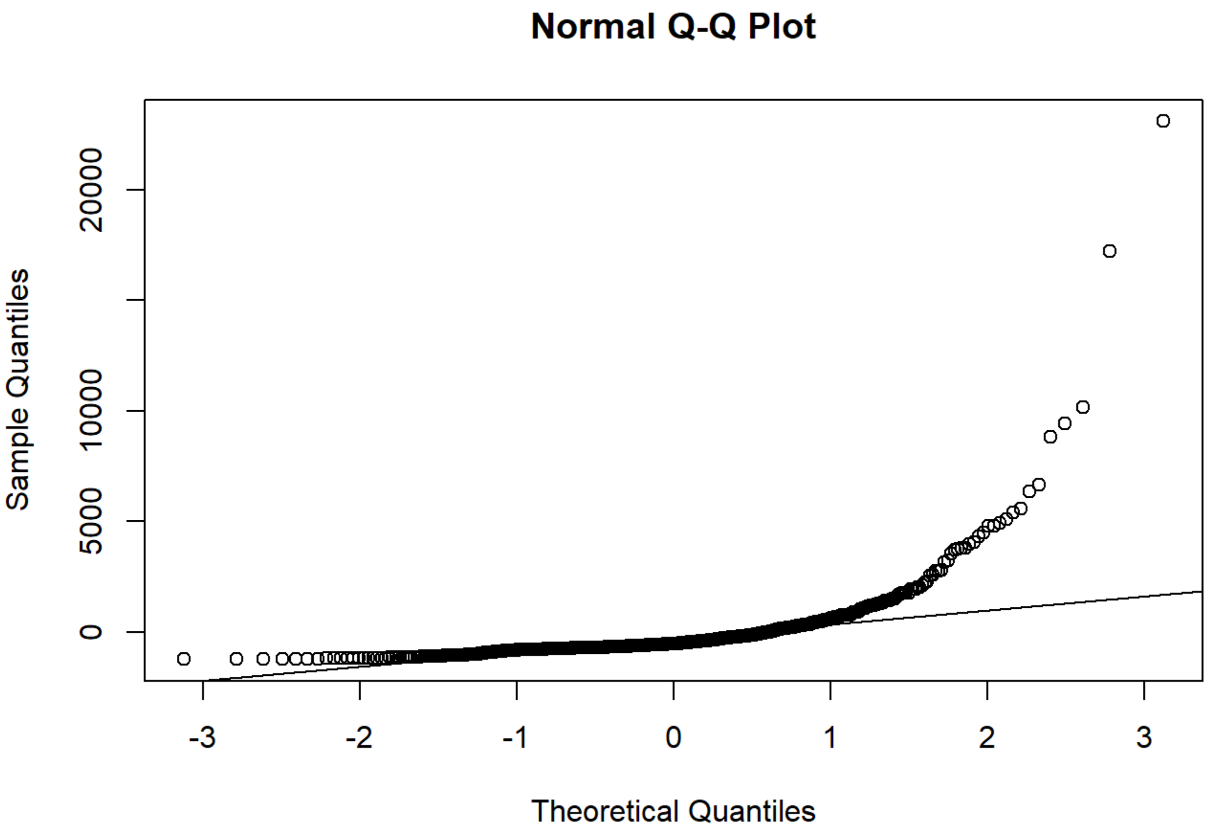
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

Or

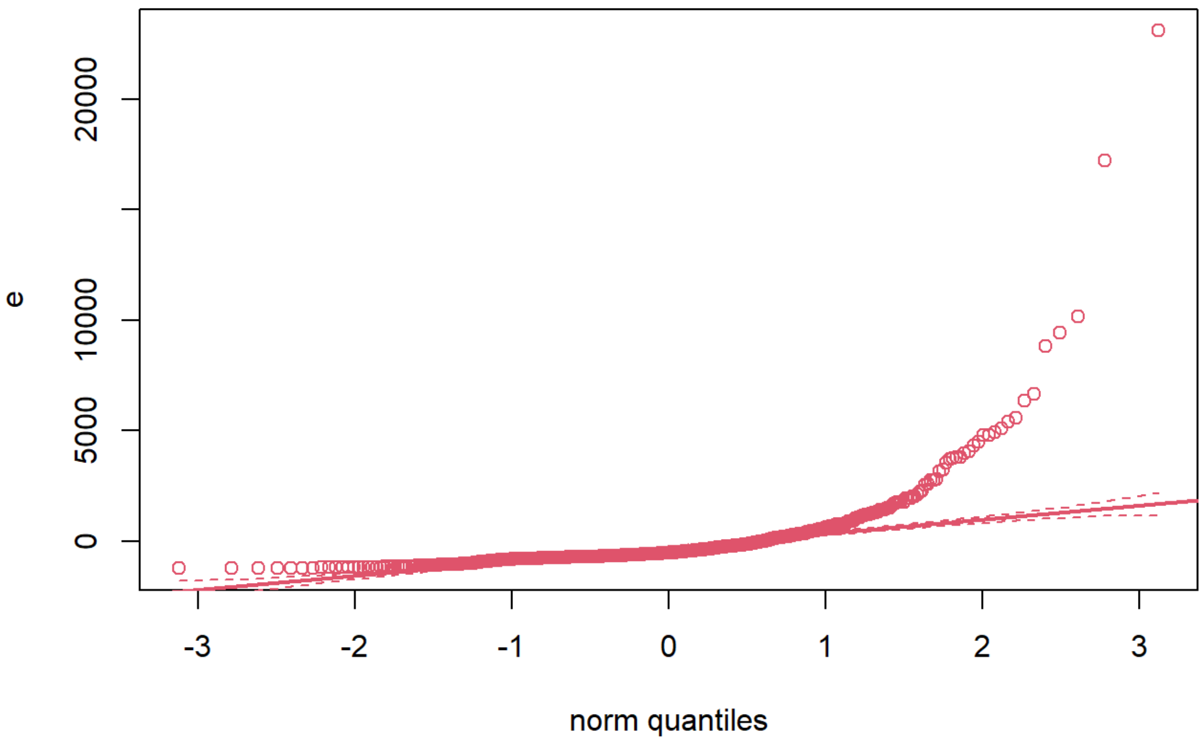
```
> oneway.test(Y ~ City, data = data, var.equal = TRUE)

    One-way analysis of means

data:  Y and City
F = 3.6586, num df = 3, denom df = 556, p-value = 0.0124
```

We can extract the residuals and see if hey are normal.

```
> e <- myanova$residuals
> qqnorm(e)
> qqline(e)
```



```
> qqplot.das(e)
```



From the plot we observe that the residuals are not normal. Therefore, ANOVA is not appropriate.

We can perform also Shapiro test to prove this

H_0 : ε is normally distributed
H_A : ε is not normally distributed

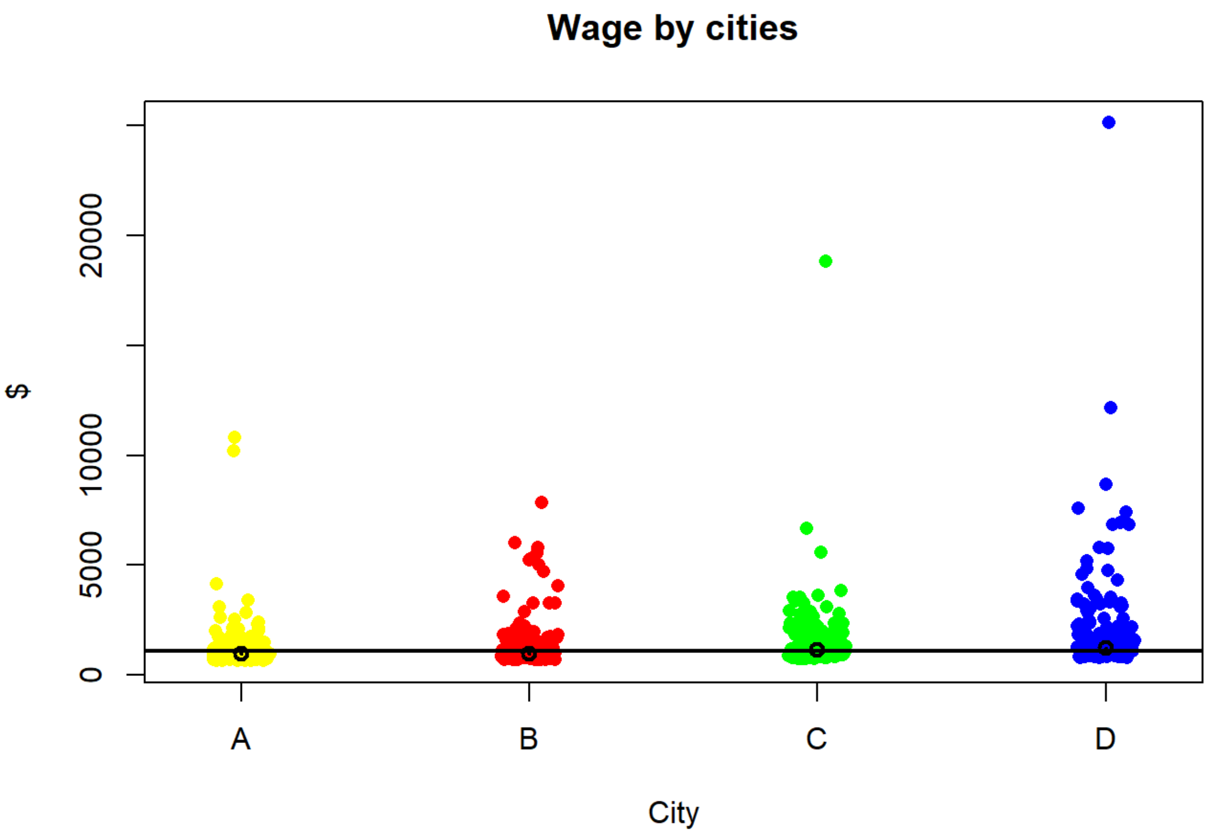
```
> shapiro.test(e)

    Shapiro-Wilk normality test

data:  e
W = 0.47838, p-value < 2.2e-16
```

The $p-value < 2.2e-16 < 0.05 = \alpha$, so we reject H_0. The residuals are not Normal.

Therefore, we perform **Kruskal-Wallis test** for equality of the medians

```
> medianAll <- median(all); medianAll
[1] 1106.174
```

```
> medianGroups <- c(median(Y_A), median(Y_B),
median(Y_C), median(Y_D)); medianGroups
[1]  952.7468  960.2859 1158.8570 1236.0503
> stripchart(salary,
+            main = "Wage by cities",
+            ylab = "$",
+            xlab = "City",
+            method = "jitter",
+            col = c("yellow", "red", "green", "blue"),
+            pch=16,
+            vertical=TRUE)
> abline(h = medianAll, lwd = 2)
> segments(c(1, 2, 3, 4), medianGroups, c(1, 2, 3, 4),
rep(medianAll, 4), lwd = 2)
> points(medianGroups ~ c(1, 2, 3, 4), lwd = 2)
```



$$H_0 : Me(Y\,|\,X = City\ A) = Me(Y\,|\,X = City\ B) =$$
$$= Ne(Y\,|\,X = City\ C) = Me(Y\,|\,X = City\ D)$$

H_A : At least two of $Me(Y\,|\,X = City\ A)$, $Me(Y\,|\,X = City\ B)$, $Me(Y\,|\,X = City\ C)$ and $Me(Y\,|\,X = City\ D)$ are different.

```
> kruskal.test(Y ~ City, data = data)
```

```
    Kruskal-Wallis rank sum test

data:  Y by City
Kruskal-Wallis chi-squared = 35.323, df = 3, p-value =
1.041e-07
```

The $p-value = 1.041e-07 < 0.05 = \alpha$, so we reject H_0 of equal medians.

In this case as far as the values are not normal the results are very different from the ANOVA results.