Analysis of variance (ANOVA)

Input file contains two columns: the first column `y` is the value, and the second column `group` indicates the group number.

```
> bvtv=read.csv("bvtv.csv")
> bvtv=data.frame(y=bvtv[,1],group=factor(bvtv[,2]))
> head(bvtv)
```

```
      y group
1  6.25     1
2  8.13     1
3 25.27     1
4 14.36     1
5  6.36     1
6  7.62     1
```

As you can see, we only have one column called `group`, which is a factor, to indicate the classification. That's why we called it ***one-way ANOVA***. If there are 2 factor variables to be involved, we called it ***two-way ANOVA***.

```
> result=aov(y~group,data=bvtv)
> summary(result)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
group         5   3457   691.4   17.01 5.28e-10 ***
Residuals    53   2154    40.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a significant p-value of 5.28e-10<0.05, which means we reject the null hypothesis that $H_0: \mu_1 = \mu_2 = \cdots = \mu_6$, i.e. there are significant differences among the 6 population means.

But *ANOVA* will not give any information about which 2 are different. Therefore, when we reject H_0 in *ANOVA*, we need to perform multiple comparisons, i.e. 6 *pair-wise t-test*, which should be adjusted for multiple testing.

```
> pairwise.t.test(bvtv$y,bvtv$group,p.adjust="holm")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  bvtv$y and bvtv$group

  1       2       3       4      5
2 0.8062  -       -       -      -
3 4.0e-06 4.1e-09 -       -      -
4 1.0000  1.0000  1.7e-07 -      -
5 1.0000  1.0000  2.1e-08 1.0000 -
6 0.1156  0.0013  0.0131  0.0131 0.0041

P value adjustment method: holm
```

Based on the results, for example, we know that group1 are significantly different from group3 with an adjusted p-value of 4.0e-06<0.05.

You can choose any p-value adjustment by change `p.adjust="holm"` to anyone of these: `"holm", "hochberg","hommel","bonferroni","BH","BY","fdr"`. The `pooled SD` means we calculate a common standard deviation for all 6 groups and use it for all *t-test*.

Instead of *pair-wise t-test*, you may see some people use *Tukey's method*, or they may say that they use *ANOVA with post-hoc Tukey HSD test*. It is just another way to find which two are different when rejecting H_0 in *ANOVA*.

```
> TukeyHSD(result,conf.level=0.95)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ group, data = bvtv)

$group
           diff        lwr        upr     p adj
2-1  -4.2741667 -12.585847   4.037514 0.6529630
3-1  17.5355556   8.649996  26.421115 0.0000048
4-1  -2.6222222 -11.507782   6.263338 0.9514253
5-1  -3.3769697 -11.849019   5.095079 0.8450439
6-1   7.4411111  -1.444449  16.326671 0.1499314
3-2  21.8097222  13.498042  30.121402 0.0000000
4-2   1.6519444  -6.659736   9.963625 0.9914496
5-2   0.8971970  -6.970869   8.765263 0.9993900
6-2  11.7152778   3.403598  20.026958 0.0015257
4-3 -20.1577778 -29.043338 -11.272218 0.0000002
5-3 -20.9125253 -29.384574 -12.440476 0.0000000
6-3 -10.0944444 -18.980004  -1.208885 0.0173083
5-4  -0.7547475  -9.226796   7.717301 0.9998177
6-4  10.0633333   1.177773  18.948893 0.0178123
6-5  10.8180808   2.346032  19.290130 0.0051693
```

Similar to *pair-wise t-test*, we can find that group1 are significantly different from group3 with an adjusted p-value of 4.8e-06<0.05. The p-value between *pair-wise t-test* and *Tukey's method* are different, but they pretty much the same thing, and you can choose one of them as you like.