Module 8. Analysis of Variance (ANOVA).

Overview:

This module introduces Analysis of Variance which compares multiple population means. This is an extension of the one population and two population statistical inference methods.

We start with ANOVA with one factor, and use it to introduce the terminologies. The linear model setup for ANOVA is introduced. Also, the group differences are determined through pairwise group means comparisons.

We then introduce ANOVA with two factors. Finally, we introduce diagnostic tests for checking the ANOVA model assumptions. Some robust tests are discussed for when the assumptions are violated.

By the end of this module, you should be able to conduct the one-way and two-way ANOVA in R. You need to be able to read the R outputs, finding the variance components and parameter estimates. You should be able to locate the group differences through pairwise group means comparison. You should know how to check the model assumptions. Finally, you should know how to do the robust tests in R, including programming permutation tests.

Learning Objectives

- Write out **one-way ANOVA** model correctly. Decompose the total sum of squares into two components: **within group** and **between group variations**.
- Carry out the ANOVA in R, use **F-test** to test the null hypothesis.
- Identify which groups differ using a **pairwise t-test**, with **adjustment for multiple testing** by false discovery rate.
- Create a **two-way ANOVA** model and conduct the analysis in R
- **Check the model assumptions** and perform **robust test** for non-normal or heteroscedastic data

Readings:

Seefeld & Linder’s book pages 263-274.
Krijnen’s book pages 73-90.

Lesson 1: One-way Analysis of Variance

Objectives

By the end of this lesson you will have had opportunity to:

- Separate the variances from two sources: between-group and within-group
- Carry out one-way ANOVA in R

Overview

We have learned about statistical inferences for one population and two populations. For comparing means of more than two populations, we will use the analysis of variance (ANOVA). We will cover the basic concepts and setup of the one-way ANOVA here.

Analysis of Variance: Two Variance Components

In previous modules, we have compared two population means using the two-sample tests. When there are more than two populations, we will need the ANOVA (Analysis Of Variance) procedure to test the differences in their means.

Mathematically, we have g random samples $Y_{1,1},...,Y_{n,1} \sim N(mean = \mu_1, sd = \sigma)$, …, $Y_{1,g},...,Y_{n,g} \sim N(mean = \mu_g, sd = \sigma)$. We generally use the small case letters to denote the actual observations: $y_{1,1},...,y_{n,1}$, …, $y_{1,g},...,y_{n,g}$ from the g groups. We want to know if the group means are different. That is, to test
$H_0 : \mu_1 = ... = \mu_g$ versus $H_A :$ at least two of the group means are different.

For the hypothesis, we will separate the variance in data into two components: **between group variance** and **within group variance**. Then compare these two components, hence the name **analysis of variance**.

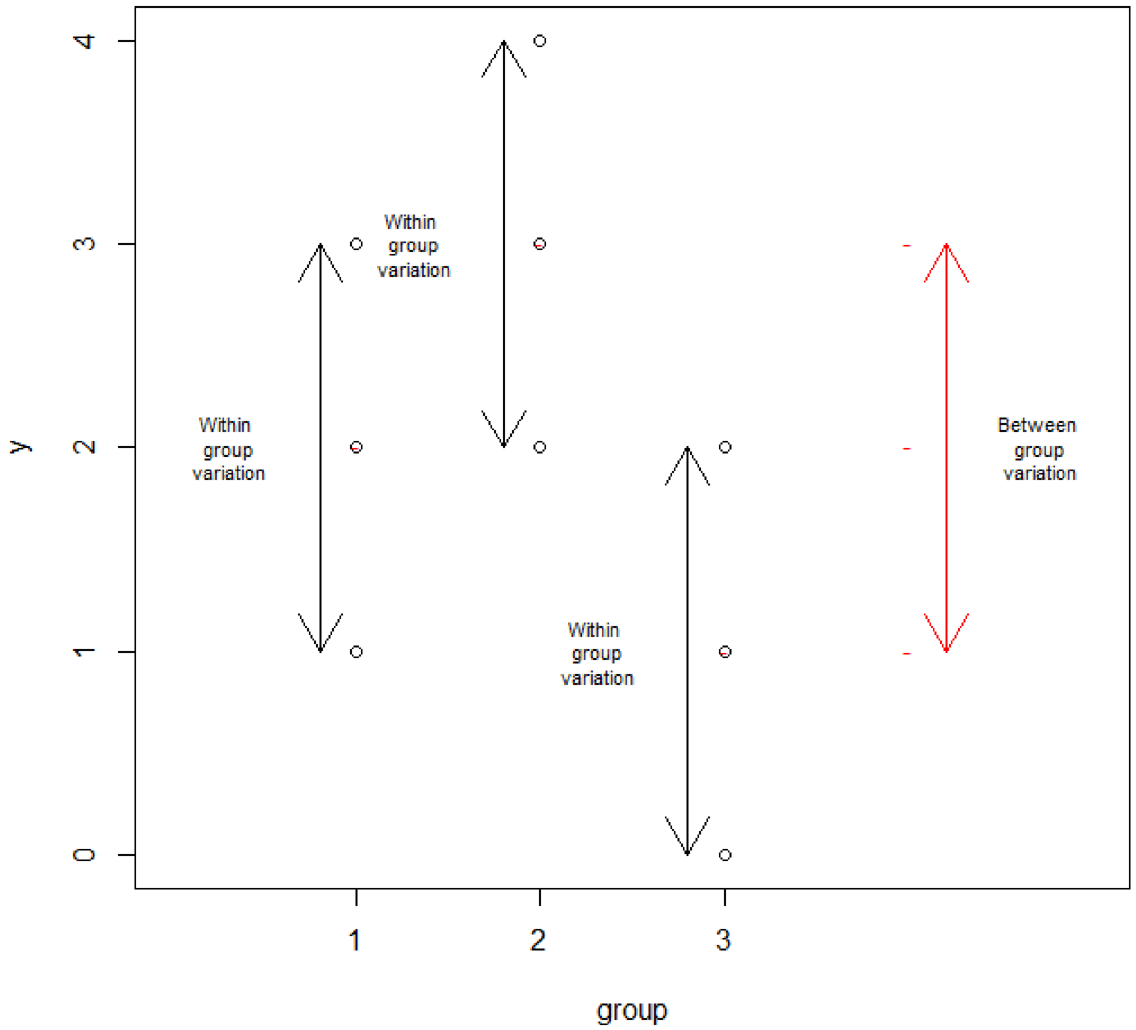
We illustrate these two components graphically on a small artificial data set next.

Graphical Display of The Two Variance Components

Suppose we have the following artificial gene expressing values 2,3,1 of Group 1, and 3,2,4 of Group 2, and 2,0,1 of Group 3. We may store the data and the group indicator as

```
y<-c(2,3,1,3,2,4,2,0,1)
group<-c(1,1,1,2,2,2,3,3,3)
```

The following plot displays the data by groups. The within group variation is reflected by the black arrows. The group means are plotted on the right with red colored "-" symbols. The variation among the group means are the between group variation reflected by the red arrow. These are the two components we need to compare.



ANOVA: The Variance Decomposition

Quantitatively, the variation in data is summarized by the sum of squares of the deviations from mean. Mathematically, the within group variation and between group variation exactly sums up to total variation in data.

Let $\bar{y} = \frac{1}{N}\sum_{j=1}^{g}\sum_{i=1}^{n} y_{i,j}$ denote the overall mean, with $N = gn$ being the total number of observations. And let $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}$ denote the j-th group mean. The total variation in data is represented by the Sum of Squares Total (SST)

$$SST = \sum_{j=1}^{g}\sum_{i=1}^{n}(y_{i,j} - \bar{y})^2.$$

The within group variation is represented by the Sum of Squares Within (SSW)

$$SSW = \sum_{j=1}^{g}\sum_{i=1}^{n}(y_{i,j} - \bar{y}_j)^2.$$

The between group variation is represented by the Sum of Squares Between (SSB)

$$SSB = \sum_{j=1}^{g}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2 = n\sum_{j=1}^{g}(\bar{y}_j - \bar{y})^2.$$

Then with some algebra, we get the basic equality of ANOVA:

SST = SSW + SSB.

For the derivation of this equation, click on

http://www.livescribe.com/cgi-bin/WebObjects/LDApp.woa/wa/MLSOverviewPage?sid=F44444444444

ANOVA: The Variances Comparison

We compare the two components of variation to test the null hypothesis

$$H_0 : \mu_1 = ... = \mu_g .$$

An equivalent representation for the model $Y_{1,j}, ..., Y_{n,j} \sim N(mean = \mu_j, sd = \sigma)$ is that

$$Y_{i,j} = \mu_j + \varepsilon_{i,j},$$

where $\varepsilon_{i,j} \sim N(mean = 0, sd = \sigma)$ is random noise.

The within group variation SSW only depends on the noise $\varepsilon_{i,j}$'s. The SSB, under the null hypothesis, also depends only on noise $\varepsilon_{i,j}$'s. However, under alternative hypothesis, SSB contains both variations in μ_j's and the noise variation. Hence, we will reject the null hypothesis if SSB is bigger than purely noise variation (which can be measured by SSW). That is, we reject the null hypothesis when the ratio SSB/SSW is too big.

To decide the cutoff value for rejection, we need the distribution of SSB/SSW under null hypothesis.

ANOVA: The F-Test

Under the null hypothesis $H_0 : \mu_1 = ... = \mu_g$,

$$SSB = \sum_{j=1}^{g}\sum_{i=1}^{n}(\bar{y}_j - \bar{y})^2 = n\sum_{j=1}^{g}(\bar{y}_j - \bar{y})^2 \sim \sigma^2 \chi^2_{df=g-1}.$$

$$SSW = \sum_{j=1}^{g}\sum_{i=1}^{n}(y_{i,j} - \bar{y}_j)^2 \sim \sigma^2 \chi^2_{df=N-g}.$$

Therefore, the expected values of SSB and SSW are respectively $\sigma^2(g-1)$ and $\sigma^2(N-g)$. Hence we should compare their values scaled by their degrees of freedom $df_B = g-1$ and $df_W = N-g$ respectively.

That is, we compare the Mean of Squares $MSB = SSB/(g-1)$ and $MSW = SSW/(N-g)$. Then the ratio $F_{obs} = \frac{MSB}{MSW}$ should be around 1 under null hypothesis.

Theorem: Under the null hypothesis $H_0 : \mu_1 = ... = \mu_g$, $F_{obs} = \frac{MSB}{MSW}$ follows an F-distribution with degrees of freedom (g-1) and (N-g).

Therefore, an α level test would reject the null hypothesis if $F_{obs} > F_{1-\alpha,g-1,N-g}$.

Note that the Theorem still holds for unequal group sizes: the j-th group has n_j observations $Y_{1,j},...,Y_{n_j,j} \sim N(mean = \mu_j, sd = \sigma)$, $N = n_1 + ... + n_g$.

ANOVA: The F-Test For Group Means

When we have g random samples $Y_{1,1},...,Y_{n_1,1} \sim N(mean = \mu_1, sd = \sigma)$, …, $Y_{1,g},...,Y_{n_g,g} \sim N(mean = \mu_g, sd = \sigma)$. We test $H_0 : \mu_1 = ... = \mu_g$ versus H_A : at least two of the group means are different.

Then we reject the null hypothesis when $F_{obs} = \frac{MSB}{MSW} > F_{1-\alpha,g-1,N-g}$.

In R, this can be done with anova(). The results are generally presented in a table, often referred to as an *ANOVA table*.

Source of variation	Degrees of freedom	Sum of Squares	Mean Squares	F-statistic	p-value
Treatment (between group)	g-1	SSB	MSB	$F_{obs} = \frac{MSB}{MSW}$	$P(F_{g-1,N-g} \geq F_{obs})$
Residuals (within group)	N-g	SSW	MSW		
Total	N-1	SST			

We can see all the quantities mentioned before are listed. Particularly the first two rows add up to the third row for the decomposition of two sources of variation. The test statistic and p-values are summarized in the last two columns.

We next illustrated this with examples of the gene expression data sets.

ANOVA Example for a SKI-Like Oncogene

We illustrate ANOVA on the “ALL” data from the “ALL” package. (We have used this data set before. If somehow you did not have “ALL” package installed, review page 2 of Krijnen’s textbook.). Specifically, we consider the gene with name “1866_g_at”, which refers to an SKI-like oncogene related to oncoproteins. For illustration purposes, we focus on the patients in three disease stages: B1, B2, and B3. We are interested in this if the expression values of “1866_g_at” gene differ within the disease stages.

This is exactly the setting of ANOVA procedure for testing if three groups mean all equal. We use R to conduct the test.

```
> library(ALL);data(ALL) #load the package and the data set.
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in three stages
> y <- exprs(ALLB123)["1866_g_at",] #exprs function gives gene expression
values. We only take gene 1866_g_at values.
> anova(lm(y ~ ALLB123$BT)) #anova. Group indicator in ALLB123$BT
Analysis of Variance Table
Response: y
               Df  Sum Sq Mean Sq F value    Pr(>F)
ALLB123$BT  2   5.4563  2.72813  19.848 1.207e-07 ***
Residuals      75  10.3091 0.13745
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the ANOVA table, p-value=1.207×10^{-7} is very small and we reject the null hypothesis. Hence we conclude that the 1866_g_at gene expression is related to the disease stages for B-cells: B1, B2 and B3.

ANOVA Example for An Ets2 Repressor Gene.

We now test if the mean expression values for probe 1242_at differ for patients in stages B1, B2, and B3 from the ALL data set. This probe corresponds to the Ets2 repressor factor which plays a role in telomerase regulation in human cancer cells. Running ANOVA with R, we get the following:

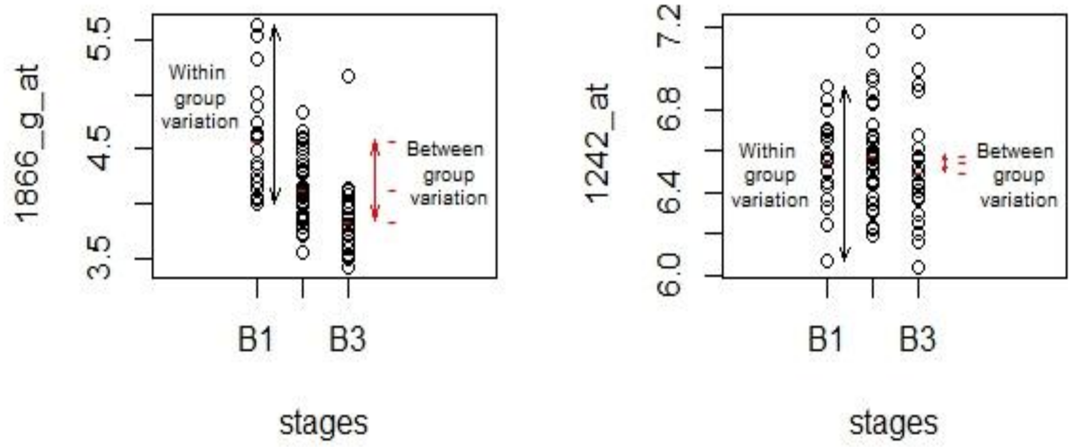
```
> library(ALL);data(ALL)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in three stages
> y <- exprs(ALLB123)["1242_at",] #exprs function gives gene expression values.
We only take gene 1241_at values.
> anova(lm(y ~ ALLB123$BT)) #anova. Group indicator in ALLB123$BT
Analysis of Variance Table

Response: y
             Df Sum Sq  Mean Sq F value Pr(>F)
ALLB123$BT  2 0.0900 0.045012  0.7362 0.4823
Residuals      75 4.5854 0.061138
```

In p-value=0.4823, the conclusion is to accept the null hypothesis. That is, there is no evidence that the gene expresses differently for patients in the three different disease stages.

Graphical Display of The Two Examples.

The following plot shows the gene expression values in the two examples above, by disease stage, and plots the three group means (for the three stages) with red colored “-” symbols.



The visual display helps us understand the information from the data. We can see that between groups, variation appears much bigger in the 1866_g_at gene than the 1242_at gene. This agrees with our conclusions in the examples: the 1866_g_at gene express differently while the 1242_at gene express similarly in all groups.

Seeing can also be deceiving. We need to understand the graph correctly. In the graph on the left, the red arrow (between group variation) is much shorter than the black arrow (within group variation). The arrows should not be directly compared against each other since they have different degrees of freedom. The variance of $Y_{i,j}$ is σ^2 while the variance of $\bar{Y}_j$ is only σ^2 / n_j. So usually, the red arrow will be much shorter than the black arrow. The comparison should be based on MSB and MSW that adjusts the degrees of freedom.

Generally, the graphs show the pattern (the picture on the left is likely to have more differences in group means than the picture on the right). But for quantitative conclusions, we need to apply statistical tests (F-test here) based on the probability theory.

Relationship to the Two Sample T-test.

The ANOVA tests equal means in g groups. When g=2, this reduces to the two-sample comparison problem in previous modules.

For ANOVA, we made the assumption that the variances are the same in all groups. When g=2, the ANOVA F-test is equivalent to the two-sample t-test with equal variances. To see this, notice that the square of a t-distributed random variable follows an F-distribution with the numerator degree of freedom=1.

$(t_{df})^2 \sim F_{1,df}$

You can check this equivalence by comparing the p-values and test statistics from t.test(y~ group, var.equal=T) and anova(lm(y~group)) on any data set with only two groups.

Summary

In this lesson, we considered testing the equal means for more than two groups. That is, to test
$H_0: \mu_1 = ... = \mu_g$ versus $H_A:$ at least two of the group means are different.

We separated the variation in data into two components: between group variation and within group variation. After adjusting for their degrees of freedom, we arrive at the ANOVA F-test for $H_0: \mu_1 = ... = \mu_g$.

We showed how to use R to conduct the ANOVA, and how to read and interpret the ANOVA table.

The group identifiers are coded in one (factor) variable. Hence the analysis is called one-way ANOVA. We will consider two factor variables, thus two-way ANOVA in a later lesson of this module.

In the next lesson, we go over linear model representations of the ANOVA model, and consider the determination of the group mean difference.

Lesson 2: Linear Model and Testing For Pairwise Difference

Objectives

By the end of this lesson you will have had the opportunity to:

- Distinguish the parameters for linear models with/without intercept term
- Do pairwise group means comparison to locate the difference

Description

We will now discuss more details about the linear model representation of the ANOVA model. The linear model is a basic representation used in many statistical procedures. We use the one-way ANOVA to illustrate its notations. It will be used in the later module on regression.

We present two commonly used representations of ANOVA in linear model, one with intercept term and one without intercept term. And compare their R implementation and outputs.

Finally, we will discuss the pairwise comparison of group means, and multiple testing adjustment to those comparisons.

Linear Model

To get the ANOVA table in R, we used the command anova(lm()). The inside function lm() stands for linear model. A basic form of the linear model is

$Y_i = x_i \beta + \varepsilon_i$, for $i = 1, ..., n,$

where Y_i is an observable variable, x_i is a fixed variable with known values, β is an unknown weight (parameter), and ε_i is an unobservable error variable. The $\varepsilon_1, ..., \varepsilon_n$ are generally assumed to be random noise $\sim N(mean = 0, sd = \sigma)$.

This is called a **linear model** because the equation is linear in the unknown parameter β which is the focus of the statistical inference. In general application, we often need x_i to be a vector $(x_{i,1}, ..., x_{i,k})$ and k unknown parameters $\beta_1, ..., \beta_k$. So the general form of the linear model is

$Y_i = x_{i,1}\beta_1 + ... + x_{i,k}\beta_k + \varepsilon_i$, for $i = 1, ..., n.$

The linear regression is the most often used linear model. We will cover linear regression in the next module. Now, we look at how the ANOVA model can also be written as a linear model.

ANOVA as A Linear Model

The ANOVA model is
$Y_{i,j} = \mu_j + \varepsilon_{i,j},$ for $j = 1,...,g,$ and $i = 1,...,n_j$ with $\varepsilon_{i,j} \sim N(mean = 0, sd = \sigma)$.

There are total of $N = n_1 + ... + n_g$ observations. We can reorganize the data to save Ys in a vector $(Y_1,...,Y_N)$, and a corresponding group indicators vector $(group_1,..., group_N)$.

To set this up in the linear model form, we turn the group indicator "group" into g variables $x_1,...,x_g$ each as an indicator for one group. That is, $x_{i,j} = I\{group_i = j\},$ for $i = 1,...,N$ and $j = 1,...,g$. Then the ANOVA model becomes a linear model
(A) $Y_i = x_{i,1}\beta_1 + ... + x_{i,g}\beta_g + \varepsilon_i,$ with $\beta_j = \mu_j$ for $j = 1,...,g$.
The null hypothesis $H_0 : \mu_1 = ... = \mu_g$ is equivalent to $H_0 : \beta_1 = ... = \beta_g$ under this linear model representation.

A more common form of linear model has the first variable $x_{i,1} \equiv 1$ always, so that
$Y_i = \beta_1 + x_{i,2}\beta_2 + ... + x_{i,g}\beta_g + \varepsilon_i.$
Here β_1 is called the "intercept" term. The default option in R (as well as all other statistical packages) for linear model assumes an intercept term. To express the ANOVA model in this standard form, we can let $\beta_1 = \mu_1$ and $\beta_j = \mu_j - \mu_1$ for $j = 2,...,g$. The with $x_{i,j} = I\{group_i = j\},$ for $i = 1,...,N$ and $j = 2,...,g$, we have
(B) $Y_i = \beta_1 + x_{i,2}\beta_2 + ... + x_{i,g}\beta_g + \varepsilon_i.$
The null hypothesis $H_0 : \mu_1 = ... = \mu_g$ is equivalent to $H_0 : \beta_2 = ... = \beta_g = 0$ under this linear model representation.

Next we use the small artificial gene expression data set earlier to illustrate these two forms of linear model representation of ANOVA model.

ANOVA as a Linear Model Without Intercept Term

Suppose we have the following artificial gene expressing values 2,3,1 of Group 1, and 3,2,4 of Group 2, and 2,0,1 of Group 3. We store the data and the group indicator as

```
y<-c(2,3,1,3,2,4,2,0,1)
group<c(1,1,1,2,2,2,3,3,3)
```

The linear model representation of ANOVA model without intercept term is
(A) $Y_i = x_{i,1}\beta_1 + ... + x_{i,g}\beta_g + \varepsilon_i$,
with $\beta_j = \mu_j$ and $x_{i,j} = I\{group_i = j\}$. To see three new group indicator variables, we show them together with original data in R as

```
> x1<-(group==1) #new dummy variable =1 for group 1, =0 for other groups.
> x2<-(group==2) #new dummy variable =1 for group 2, =0 for other groups.
> x3<-(group==3) #new dummy variable =1 for group 3, =0 for other groups.
> cbind(y,group,x1,x2,x3) #Combine the variables as columns and print out.
     y group x1 x2 x3
 [1,] 2    1  1  0  0
 [2,] 3    1  1  0  0
 [3,] 1    1  1  0  0
 [4,] 3    2  0  1  0
 [5,] 2    2  0  1  0
 [6,] 4    2  0  1  0
 [7,] 2    3  0  0  1
 [8,] 0    3  0  0  1
 [9,] 1    3  0  0  1
```

These new group indicator variables $x_{i,j}$ needed for the linear model can be automatically created in R with a model statement `y~ group -1` where '-1' indicates no intercept term. We demonstrate this in the next item.

ANOVA as A Linear Model Without Intercept Term in R

The model statement generates the $x_{i,j}$ values and put them in a matrix, the so called “model matrix”.

```
> group<-as.factor(group) #The groups should be a factor, no intrinsic numerical
values
> model.matrix(y~ group -1) #Let R build the linear model automatically, “-1”
means no intercept term
  group1 group2 group3
1     1      0      0
2     1      0      0
3     1      0      0
4     0      1      0
5     0      1      0
6     0      1      0
7     0      0      1
8     0      0      1
9     0      0      1
```

The R generated variables ‘group1’, ‘group2’ and ‘group3’ are exactly the same as ‘x1’, ‘x2’ and ‘x3’ in the last page.

Fitting this linear model in R will also give point estimations for β_j s which, in this model, are the group means μ_j s. These estimates are shown in the summary of the lm() fit

```
> summary(lm(y~ group -1)) #summary of the anova results
```

… (some omitted outputs)

```
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
group1   2.0000     0.5774   3.464  0.01340 *
group2   3.0000     0.5774   5.196  0.00202 **
group3   1.0000     0.5774   1.732  0.13397
```

… (more outputs omitted)

ANOVA as A Linear Model With Intercept Term in R

The linear model representation of ANOVA model with intercept term is

(B) $Y_i = \beta_1 + x_{i,2}\beta_2 + ... + x_{i,g}\beta_g + \varepsilon_i$,

where $\beta_1 = \mu_1$ and $\beta_j = \mu_j - \mu_1$ for $j = 2,...,g$. This is the default option for lm() in R. We can see the model matrix and the point estimates as following.

```
> model.matrix(y~ group) #Let R build the linear model automatically
  (Intercept) group2 group3
1           1      0      0
2           1      0      0
3           1      0      0
4           1      1      0
5           1      1      0
6           1      1      0
7           1      0      1
8           1      0      1
9           1      0      1
> summary(lm(y~ group)) #summary of the anova results
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0000     0.5774   3.464   0.0134 *
group2        1.0000     0.8165   1.225   0.2666
group3       -1.0000     0.8165  -1.225   0.2666
```

Unlike the linear without intercept term, the $\hat{\beta}_j$s here are not all group means. For example, to get the sample mean in the second group, we calculate by $\hat{\mu}_2 = \hat{\mu}_1 + (\mu_2 - \mu_1) = \hat{\beta}_1 + \hat{\beta}_2 = 2 + 1 = 3$.

The Two Linear Models For ANOVA

The two linear model representations of ANOVA are equivalent to each other. However, the R command produces different ANOVA tables. The anova(lm(y~ group -1) produces the ANOVA table for testing if all group means equal zero. The anova(lm(y~ group) produces the ANOVA table for testing if all group means equal (may be a non-zero value).

You should learn the linear model representations here to understand the model specification in lm(). We will use lm() again in the next module on regression. Now, you should be able to interpret the outputs of the lm() fits correctly for these two equivalent specifications.

Example R Outputs for the Two Linear Models

We look at the lm() fits for the SKI-like oncogene “1866_g_at” in the “ALL” data in the last lesson. We use the linear models with and without intercept term for ANOVA on the gene expression values on three disease stages: B1, B2, and B3.

```
> library(ALL);data(ALL)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in three stages
> y <- exprs(ALLB123)["1866_g_at",] #gene 1866_g_at expression values
> summary(lm(y ~ ALLB123$BT)) #summary of the anova results with intercept
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.58222    0.08506  53.873  < 2e-16 ***
ALLB123$BTB2 -0.43689    0.10513  -4.156 8.52e-05 ***
ALLB123$BTB3 -0.72193    0.11494  -6.281 2.00e-08 ***
```

```
> summary(lm(y ~ ALLB123$BT -1)) #summary of the anova results, no intercept
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
ALLB123$BTB1  4.58222    0.08506   53.87   <2e-16 ***
ALLB123$BTB2  4.14533    0.06179   67.09   <2e-16 ***
ALLB123$BTB3  3.86029    0.07731   49.94   <2e-16 ***
```

The second model outputs provide the group means estimates directly, while the first model outputs provides estimates for differences in group means. For example, the B3 group mean is 3.86 from the second output, and is calculated as 4.58-0.72=3.86 from the first output. For the difference between the B3 group and B1 group means, the first output provide -0.72 directly, and need to be calculated in the second output as 3.86-4.48.

Determining Which Group Differ
When the ANOVA analysis says there are differences among the group means, then we usually wants to ask next which groups are different. The between group variation component sums over differences from all groups, and does not distinguish which group really contributes. We need to look at the point estimators from each group for this information.

The ANOVA linear model without interaction term shows the differences of each group mean from the first group mean. For example, we take another look at the lm() outputs before for the 1866_g_at gene on the three disease stages:

```
> summary(lm(y ~ ALLB123$BT)) #summary of the anova results
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            4.58222    0.08506  53.873  < 2e-16 ***
ALLB123$BTB2 -0.43689    0.10513  -4.156 8.52e-05 ***
ALLB123$BTB3 -0.72193    0.11494  -6.281 2.00e-08 ***
```

Besides the point estimator, the outputs also contain the corresponding t-statistics and the p-value. For example, for difference between B2 and B1 group means, the point estimator is $(\mu_2 - \mu_1) = (-0.43689)$, its t-statistics is $t_{obs} = (-0.43689)/0.10513 = -4.156$ with p-value $P(t \ge |t_{obs}|) = 8.52 \times 10^{-5}$.

However, this table does not provide all information on the group mean differences we want. The first two is the B1 group mean, and the t-test for if the B1 group mean is zero. We generally are not interested in that t-test as it is not related to the group means difference. Also, the differences between group means other than B1 are not displayed directly. For example, the t-test for difference between B2 and B3 group means need further calculation. There is another R function pairwise.t.test() conducts test for comparison between all pairs of groups.

Pairwise Group Means Comparison in R

We use pairwise.t.test() on the earlier example of one-way ANOVA on the 1866_g_at gene on the three disease stages.

```
> pairwise.t.test(y, ALLB123$BT) #Do all pairwise comparison of group means
        Pairwise comparisons using t tests with pooled SD
data:  y and ALLB123$BT
   B1      B2
B2 0.00017 -
B3 6e-08   0.00518
P value adjustment method: holm
```

The p-values for the pairwise comparison t-tests are listed in a table. We can see that the p-values are 0.00017, 6×10^{-8}, 0.00518 for testing $H_0:\mu_{B1}=\mu_{B2}$, $H_0:\mu_{B1}=\mu_{B3}$ and $H_0:\mu_{B2}=\mu_{B3}$ respectively.

Notice that the p-values here are adjusted for multiple testing here. Hence, for example, the p-value for testing $H_0:\mu_{B1}=\mu_{B2}$ (0.00017) is different from the value (0.0000852) displayed in the last page. The default Holm adjustment method controls the family wise error rate (FWER). It is an improvement over the Bonferroni adjustment for the ANOVA model. We can also do the false discovery rate (FDR) control by specifying the method in pairwise.t.test().

```
> pairwise.t.test(y,ALLB123$BT,p.adjust.method='fdr') #FDR-adjusted pairwise tests
        Pairwise comparisons using t tests with pooled SD
data:  y and ALLB123$BT

   B1      B2
B2 0.00013 -
B3 6e-08   0.00518

P value adjustment method: fdr
```

In this case the p-values from FDR adjustment is almost identical to the p-values from the Holm adjustment, because there are only three pairwise hypotheses. As we discussed in earlier module, the FDR adjustment will be preferred when the number of hypotheses are big.

Lesson Summary

We introduced the linear model notations on the one-way ANOVA model. You should know the meaning of the parameters in the two representations (with and without the intercept term). You should be able to fit the linear model in R. You need to be able to find the point estimates, from the R outputs, of group means and their differences.

We taught how to conduct pairwise group means comparison in R. Using these pairwise comparisons, we can locate the different group means.

In the next lesson, we introduce two-way ANOVA.

Lesson 3: Two-way ANOVA

Objectives

By the end of this lesson you will have had the opportunity to:

- Perform two-way ANOVA in R and interpret the results

Description

In this lesson, we will teach two-way ANOVA. The one-way ANOVA data are separated into groups according to one factor variable. When there are two factors, the group membership is decided by both factors. We want to study effects on the group means by each of the two factors.

We describe the linear model representation of the two-way ANOVA. We teach how to conduct in R the two-way ANOVA with and without interaction.

Two-way ANOVA

The one-way ANOVA model can be written as
$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$,
where μ is the overall mean, α_j is the effect of the j-th level of the factor on the group mean, and $\sum_{j=1}^{g} \alpha_j = 0$.

When there are two factors, we can extend the above model as
$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$,
where $\sum \alpha_i = \sum \beta_j = \sum (\alpha\beta)_{ij} = 0$. Here α_i and β_j are the effects by the first factor and the second factor on group means, and are called the *main effects*.
$\varepsilon_{i,j} \sim N(mean = 0, sd = \sigma)$ is random noise. $(\alpha\beta)_{ij}$ is called the *interaction*, reflecting the effect by the combination of these two factors not explained by the sum of these factor effects. You may think about two factors as taking two drugs. Certain combination of the doses of the two drugs may have a much stronger effects than taking these two drugs separately, and may kill you. That is why the pharmacists always need to know what other drugs you are taking, to avoid taking combination of drugs with interactions.

Often the researcher may just want to know the effects of the factors separately, or there is reason to think the interaction is small. In such cases, we can fit the linear model without interaction term
$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$.

In R, the full two-way ANOVA model with interaction is specified by y~factor1*factor2. The two-way ANOVA model without interaction is specified by y~factor1+factor2. We illustrate fitting two-way ANOVA in R with an example next.

Example of Fitting Two-way ANOVA

We analyze the expression values of NEDD4 binding protein 1 with probe id 32069_at in the ALL data from Chiaretty et al. (2004). We study effects from two factors. The first factor is the disease stage, we consider B cell patients in four groups: B1, B2, B3 and B4. For the second factor, we select from the molecular biology of the patients assigned to BCR/ABL and NEG. The two-way ANOVA (with interaction) is conducted as

```
> library("ALL"); data(ALL)
> ALLBm <- ALL[,which(ALL$BT %in% c("B1","B2","B3","B4") &
ALL$mol.biol %in% c("BCR/ABL","NEG"))] #select patients
> y<-exprs(ALLBm)["32069_at",] #gene 32069_at expression values
> Bcell<-ALLBm$BT          # B-cell stages
> molb<-ALLBm$mol.biol   #molecular biology types
> anova(lm(y~ Bcell*molb)) #full two-way ANOVA
Analysis of Variance Table
Response: y
                    Df  Sum Sq Mean Sq F value   Pr(>F)
Bcell                3  3.4983 1.16610  4.5051 0.006127 **
molb                 1  2.0182 2.01819  7.7971 0.006814 **
Bcell:molb           3  0.8561 0.28538  1.1025 0.354324
Residuals           67 17.3421 0.25884
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that both factor affects the gene expression, and there is no statistical significant interaction. We may also fit the two-way ANOVA without interaction as

```
> anova(lm(y~ Bcell+molb)) #Additive (no interaction) two-way ANOVA
Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
Bcell      3  3.4983 1.16610  4.4854 0.006148 **
molb       1  2.0182 2.01819  7.7630 0.006857 **
Residuals 70 18.1982 0.25997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Two-way ANOVA Example (continued)

To check the factor effects explicitly, we can look at the R summary of the linear model fit.

```
> summary(lm(y~ Bcell+molb)) #summary of anova fit
Call:
lm(formula = y ~ Bcell + molb)
Residuals:
    Min       1Q   Median       3Q      Max
-0.99530 -0.30786  0.01379  0.25288  1.32842

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.27630    0.20384  30.791  < 2e-16 ***
BcellB2      0.43105    0.19825   2.174  0.03306 *
BcellB3      0.11444    0.20427   0.560  0.57711
BcellB4      0.05017    0.25475   0.197  0.84444
molbNEG     -0.35273    0.12660  -2.786  0.00686 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5099 on 70 degrees of freedom
Multiple R-squared:  0.2326,    Adjusted R-squared:  0.1888
F-statistic: 5.305 on 4 and 70 DF,  p-value: 0.0008638
```

Notice here the baseline group is of those patients with “B1” stage and molecular biology of BCR/ABL. For those patients the mean gene expression value is 6.27630. For patients of “B1” stage and NEG molecule, the mean gene expression value is 6.27630-0.35273=5.92357; For patients of “B2” stage and BCR/ABL molecule, the mean is 6.27630+0.43105=6.70735; For patients of “B2” stage and NEG molecule, the mean is 6.27630+0.43105-0.35273=6.35462.

While the disease stage has a statistical significant effect on the 32069_at gene expression value, the detailed analysis shows that only “B2” stage group is different from the other stages group.

Lesson Summary

In this lesson, we introduced the two-way ANOVA. You learned to conduct the fit using anova() and lm() in R. Also, you should learned to interpret the R outputs.

In the next lesson, we consider some diagnostic tests to check assumptions in the ANOVA model. Also, we introduce some robust ANOVA tests when the assumptions are violated.

Lesson 4: Diagnostic Tests, And Robust Tests.

By the end of this lesson you will have had opportunity to:

- Check the ANOVA assumptions by Shapiro-Wilk test and Breusch-Pagan test
- Conduct ANOVA tests without normality or homoscedasticity assumption

Description

The data analyst should always check the model assumptions to make sure that the statistical analysis is appropriate. For ANOVA, the two main assumptions are the normality or homoscedasticity assumptions. We introduce Shapiro-Wilk test and Breusch-Pagan test for these two assumptions.

We also introduce robust ANOVA tests when these two assumptions are violated: the Welch test and the Kruskal-Wallis test. Finally, we teach how to do the nonparametric permutation tests in ANOVA.

Tests for ANOVA Model Assumptions

Recall the one-way ANOVA model:
$Y_{i,j} = \mu_j + \varepsilon_{i,j}$, for $j = 1,...,g$, and $i = 1,...,n_j$ with $\varepsilon_{i,j} \sim N(mean = 0, sd = \sigma)$.

There are two important assumptions made. First, the errors are assumed to be independent and normally distributed, and, second, the error variances are assumed to be equal for each group. The equal variance assumption is often referred as the homoscedasticity assumption in literature.

In previous modules, we used the Shapiro-Wilk test to check normality of data. Hence we can check the first assumption here by applying the Shapiro-Wilk test on the residuals. The homoscedasticity assumption can be tested by the Breusch and Pagan (1979) test on the residuals. This BP test is a generalization of the *F*-test for equal variances we used before in two samples comparison.

Next, we will illustrate these tests on the previous example of analyzing the SKI-like oncogene “1866_g_at” in the “ALL” data.

Testing normality of the residuals. We test the normality of the residuals from the estimated linear model on the B-cell ALL data from 1866_g_at as follows.

```
> data(ALL,package="ALL");library(ALL)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in three stages
> y <- exprs(ALLB123)["1866_g_at",] #gene 1866_g_at expression values
> shapiro.test(residuals(lm(y ~ ALLB123$BT))) #normality test on residuals
        Shapiro-Wilk normality test
data:  residuals(lm(y ~ ALLB123$BT))
W = 0.9346, p-value = 0.0005989
```

Since the *p*-value 0.0005989 is very small, we reject the null-hypothesis of normally distributed residuals. Therefore, the normality assumption does not hold. And we need other more robust tests which we will mention later.

Testing homoscedasticity of the residuals. We use the function bptest() from the ‘lmtest’ package to test the homoscedasticity assumption.

```
> library(ALL); data(ALL); library(lmtest)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in three stages
> y <- exprs(ALLB123)["1866_g_at",] #gene 1866_g_at expression values
> bptest(lm(y ~ ALLB123$BT), studentize = FALSE) #test equal variances
        Breusch-Pagan test
data:  lm(y ~ ALLB123$BT)
BP = 8.7311, df = 2, p-value = 0.01271
```

From the *p*-value 0.01271, the conclusion follows to reject the null hypothesis of equal variances (homoscedasticity). Therefore, we need a more advanced test not assuming equal variance.

ANOVA Test with Unequal Variances

When the ANOVA model assumptions are violated, an alternative testing procedure is called for. In case only homoscedasticity is violated, we are in a situation quite similar to that of *t*-testing with unequal variances. The null hypothesis $H_0 : \mu_1 = ... = \mu_g$ of equal means can be tested without assuming equal variances by a test proposed by Welch (1951). This is implemented in R by oneway.test().

```
> data(ALL,package="ALL");library(ALL)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")]
> y <- exprs(ALLB123)["1866_g_at",]
> oneway.test(y ~ ALLB123$BT)
        One-way analysis of means (not assuming equal variances)
data:  y and ALLB123$BT
F = 14.1573, num df = 2.000, denom df = 36.998, p-value = 2.717e-05
```

In this case, the Welch test also has a small **p-value = 2.717e-05**, the conclusion follows to reject the hypothesis of equal means.

ANOVA Test without Normality Assumption

In case normality is violated a rank type of test is more appropriate. In particular, to test the null-hypothesis of equal distributions for all groups, the Kruskal-Wallis rank sum test is recommended. This test is a generalization of the two-sample Wilcoxon test. Because it is based on ranking the data, it is highly robust against non-normality, it, however, does not estimate the size of experimental effects. It is implemented as kruskal.test() in R.

```
> data(ALL,package="ALL");library(ALL)
> ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")]
> y <- exprs(ALLB123)["1866_g_at",]
> kruskal.test(y ~ ALLB123$BT)
        Kruskal-Wallis rank sum test
data:  y by ALLB123$BT
Kruskal-Wallis chi-squared = 30.6666, df = 2, p-value = 2.192e-07
```

In this case, the Kruskal-Wallis test has a small **p-value = 2.192e-07**, we reject the null-hypothesis of equal distributions of expression values among patient groups.

Permutation Tests for ANOVA

For the nonparametric Kruskal-Wallis test, we are testing the null hypothesis that the data in all groups come from a common distribution. Therefore, permuting the data do not change the distribution under null hypothesis. We can therefore use permutation tests for the nonparametric testing (without normality assumption).

The following program implements the permutation test using the ANOVA F-statistic.

```
data(ALL,package="ALL");library(ALL)
ALLB123 <- ALL[,ALL$BT %in% c("B1","B2","B3")] #patients in 3
stages
data<- exprs(ALLB123)["1866_g_at",] #gene 1866_g_at expression values
group<-ALLB123$BT[,drop=T] #drop unused levels, keep only B1,B2,B3
n<-length(data)  #sample size n
T.obs<- anova(lm(data~group))$F[1]  #Observed statistic = F-statistic
n.perm=2000   # we will do 2000 permutations
T.perm = rep(NA, n.perm) #A vector to save permutated statistic
for(i in 1:n.perm) {
    data.perm = sample(data, n, replace=F)   #permute data
    T.perm[i] = anova(lm(data.perm~group))$F[1]    #Permuted statistic
}
mean(T.perm>=T.obs)  #p-value
```

Here “extreme” means big F-statistic values (all equal group means lead to smaller F values). So p-value= $P(F \ge F_{obs})$. Run this R script, and we get p-value of zero.

```
[1] 0
```

As we mentioned earlier, the good thing about permutation test is that it is very easy to program for any test statistic. For example, if we wish to instead use the maximum difference among group means, we only need to change two steps in the above program (on calculating the observed statistic and permutated statistic):

```
T.obs<- max(by(data,group,mean))-min(by(data,group,mean))#Observed statistic
T.perm[i] = max(by(data.perm,group,mean))-min(by(data.perm,group,mean))
#Permuted statistic
```

Lesson Summary

This lesson taught how to check ANOVA assumptions. We taught the Shapiro-Wilk test for normality assumption and the Breusch-Pagan test for homoscedasticity assumption. We also taught robust ANOVA tests: the Welch test and the Kruskal-Wallis test. You should know when to use these tests and how to do these tests in R. You should also know how to program the nonparametric permutation tests for ANOVA.

Module Summary

This module covered ANOVA. You should know the setup of one-way and two-way ANOVA. You should know the linear model representation of the ANOVA model. You should know how to get the estimates from the linear model fits. Also, you should know how to locate the group means difference by pairwise comparison with multiple testing adjustments.

We also taught how to check ANOVA model assumptions. You should know how to do the model checking and how to apply appropriate robust tests when the assumptions are violated. Finally, you should be able to do the permutation tests in R for ANOVA model.

In the next module, we will discuss linear regression, which is also a linear model as the ANOVA.