Lecture 15
The Analysis of Variance

Chao Song

College of Ecology

Lanzhou University

December 8, 2025

The analysis of variance

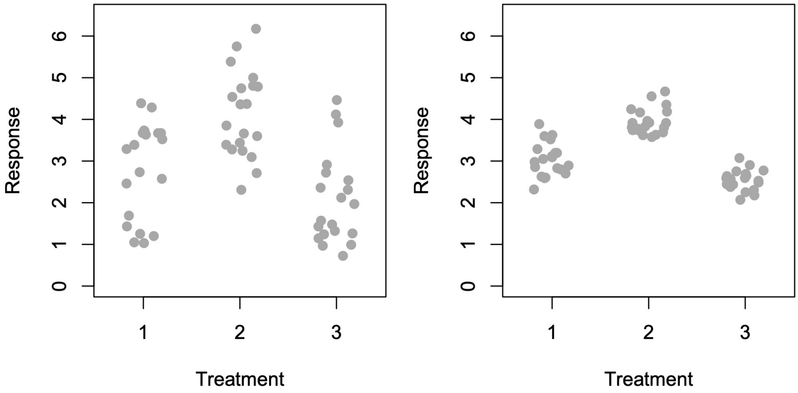
A common scenario we encounter in data analysis is comparing a response variable under several treatments

- Comparing plant growth rate under three levels of fertilizations;
- Comparing soil respiration under ambient temperature and warming;
- Evaluating student performance under different teaching methods.

In these examples, we refer to the independent variable that defines the groups as **factors**, and different values of the factor is called its **levels**. To analyze the difference among group means, we typically use a method called **analysis of variance**, or simply **ANOVA**.

The analysis of variance

To examine whether the means in each treatment are all the same or not, an intuitive approach is to compare the variation among treatment means and the variation within a treatment group.



One-way ANOVA

Suppose we have k groups of observations, each sampled from a normal population with means $\mu_1, \mu_2, \ldots, \mu_k$ and a common variance σ^2. Each group has n_k observations. To test H$_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$, we compare variation among group means and variation within a group.

Variation among group means, often referred to as SST or sum square of treatments, is calculated as

$$SST = \sum_{i=1}^{k} n_i(\overline{Y_{i.}} - \overline{Y})^2$$

Variation within a group, which is referred to as sum square of errors or SSE, is calculated as

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i.}})^2$$

One-way ANOVA

The hypothesis H$_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$ is tested by a F-statistic

$$F = \frac{SST/(k-1)}{SSE/(n-k)} \sim F_{k-1,n-k}$$

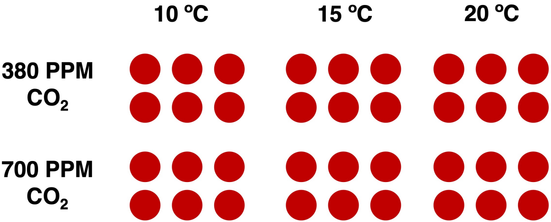
Conventionally, the results of ANOVA are presented in an ANOVA table

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	TSS		

Two-way ANOVA

We can consider the effects of more than one factor. Suppose we have two factors *A* and *B*. They have *a* and *b* levels respectively. We assume that each observation is $N(\mu_{ij}, \sigma^2)$. This type of data typically allows us to examine

- The main effects of *A* and *B*;
- The interaction between *A* and *B*.



10 ºC

15 ºC

20 ºC

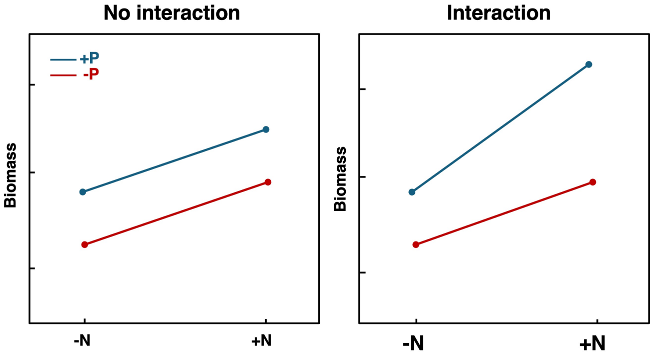
380 PPM CO_2

700 PPM CO_2

(A typical experiment set up for two-way ANOVA)

Two-way ANOVA

Two-way ANOVA allows us to identify interactions, which means that the effect of one factor depends on the level of another factor.



(An illustration of interaction in two-way ANOVA.)

Two-way ANOVA

The main effect and interaction can be tested by comparing sum squares of each factor and their interaction with the sum squares of error.

Let Y_{ijk} be the kth replicates at level i of factor A and level j of factor j. Suppose there are a levels of factor A, b levels of factor B, and n_{ij} replicates within each A and B treatment combination.

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{i..}} - \overline{Y})^2$$

$$SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{.j.}} - \overline{Y})^2$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\overline{Y_{ij.}} - \overline{Y_{i..}} - \overline{Y_{.j.}} + \overline{Y})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y_{ij.}})^2$$

Two-way ANOVA

The results of two-way ANOVA can be presented in an ANOVA table.

Source	df	SS	MS	F
A	$a - 1$	SSA	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
B	$b - 1$	SSB	$MSB = \frac{SSB}{b-1}$	$F = \frac{MSB}{MSE}$
AB	$(a - 1)(b - 1)$	SSAB	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F = \frac{MSAB}{MSE}$
Error	$n - ab$	SSE	$MSE = \frac{SSE}{n-ab}$	
Total	$n - 1$	TSS		

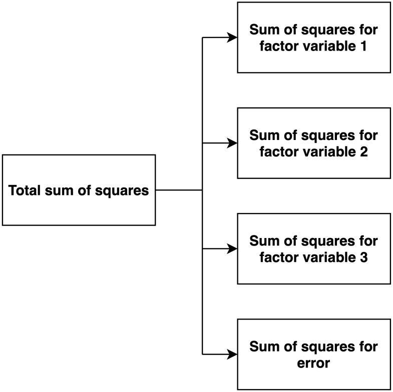
Two-way ANOVA

Example: An ecologists grew three varieties of oats under four nitrogen fertilization levels. She measured the crop yield and used a two-way ANOVA to analyze the effects of oats variety and nitrogen fertilization level. What conclusions would you draw based on the results below?

Source	Df	SS	MS	F	P-value
Variety	2	1786.4	893.2	1.7949	0.175
Nitrogen	3	20020.5	6673.5	13.41	8.367×10^{-7}
Variety:Nitrogen	6	321.7	53.6	0.1078	0.9952
Error	60	29857.3	497.6		

The analysis of variance

The intuitive approach we have introduced so far can be extended to cases with more than two factors. In essence, ANOVA decomposes the total sum of squares into sum of squares for each factor.


Total sum of squares
Sum of squares for factor variable 1
Sum of squares for factor variable 2
Sum of squares for factor variable 3
Sum of squares for error

ANOVA as a linear model

ANOVA uses decomposition of sum of squares to compare means in groups defined by factors. Similarly, we have seen that we use the F-statistics constructed from sum of squares for hypothesis testing in linear models. This similarity is not coincidence. In fact, ANOVA is a type of linear model.

Recall a linear model is defined as a model where the response variable is a linear function of parameters, i.e.,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

ANOVA can be written as a linear model by defining x as a dummy or indicator variable.

ANOVA as a linear model

A one-way ANOVA with k groups can be written as

$$y_{ij} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Here, x_i is an indicator variable where $x_i = 1$ if the observation is in group i and $x_i = 0$ if the observation is not in group i.

A few comments on the model parameters:

- The model can be written more concisely as $y_{ij} = \beta_i, \ i = 1, 2, \ldots, k$;
- There are $k + 1$ parameters in the model;
- The model can also be written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

ANOVA as a linear model

Example: We compare the biomass of plants grown under two nitrogen levels and observed the following data:

N level	Biomass
Low	12, 11, 10
High	20, 22, 21

Here, we can write the model as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{y} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 20 \\ 22 \\ 21 \end{bmatrix}; \ \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}.$$

ANOVA as a linear model

There are multiple ways to write the same linear model. The model in the above example can also be written as as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\boldsymbol{y} = \begin{bmatrix} 12 \\ 11 \\ 10 \\ 20 \\ 22 \\ 21 \end{bmatrix} ; \ \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} ; \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} ; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix} .$$

Here, β_1 is the mean of group 1 and β_2 is the difference in group mean between group 2 and group 1. This type of coding is referred to as the reference level coding and is used in most statistical software.

ANOVA as a linear model

A two-way ANOVA model with interaction can be written as

$$y_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

where α_i is the effect of level i of factor A, β_j is the effect of level j of factor B, and γ_{ij} is the interaction effect.

Comments on the two-way ANOVA model

- A two-way ANOVA do not necessarily has all terms as in the model above. For example, if there are reasons to suggest that there are no interactions, you do not need the interaction term.
- Hierarchical principle: if you include a interaction term, you should usually include the main effects involved in that interaction.

ANOVA as a linear model

Since ANOVA is a linear model, the procedures for parameter estimation and hypotheses testing we derived for linear models can be readily applied here.

- Parameters can be estimated as $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{Y}$;
- Hypothesis can be tested by F-tests. The results are the same as the ANOVA table we presented in previous slides.

$$SSH = (\Lambda\hat{\beta} - \boldsymbol{d})^T(\Lambda(\boldsymbol{X}^T\boldsymbol{X})^{-}\Lambda^T)^{-}(\Lambda\hat{\beta} - \boldsymbol{d})$$

$$SSE = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$$

$$\frac{\frac{1}{\sigma^2}SSH/m}{\frac{1}{\sigma^2}SSE/(n-k-1)} \sim F_{m,n-k-1}$$

Hypothesis testing in ANOVA

In a one-way ANOVA, the factor has 3 levels and 3 replicates in each group. Using reference level coding, how do you write the model in matrix form?

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix} ; \ \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} ; \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} .$$

Hypothesis testing in ANOVA

How do you test the hypothesis that the factor has no effects?

If the factor has no effects, the group means are equal. Therefore,

$$H_0 : \beta_2 = 0 \textit{ and } \beta_3 = 0$$

$$H_A : \beta_2 \neq 0 \textit{ or } \beta_3 \neq 0.$$

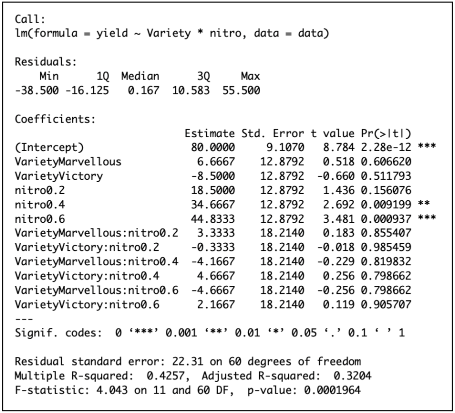
In matrix form, this hypothesis is written as $\boldsymbol{\Lambda\beta} = \mathbf{0}$, where

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Understanding the output of ANOVA

ANOVA is essentially a linear model with categorical independent variables. Thus, you can interpret the ANOVA output just like linear regression.

Example: Parameter estimates from an ANOVA model comparing yield among 3 oat varieties under 3 nitrogen levels.

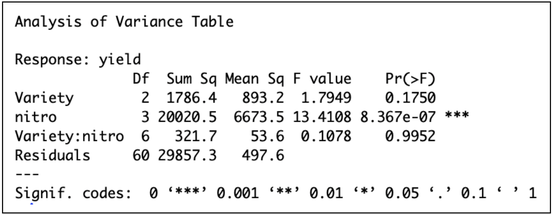


```
Call:
lm(formula = yield ~ Variety * nitro, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-38.500 -16.125   0.167  10.583  55.500

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 80.0000     9.1070   8.784 2.28e-12 ***
VarietyMarvellous            6.6667    12.8792   0.518 0.606620
VarietyVictory              -8.5000    12.8792  -0.660 0.511793
nitro0.2                    18.5000    12.8792   1.436 0.156076
nitro0.4                    34.6667    12.8792   2.692 0.009199 **
nitro0.6                    44.8333    12.8792   3.481 0.000937 ***
VarietyMarvellous:nitro0.2   3.3333    18.2140   0.183 0.855407
VarietyVictory:nitro0.2     -0.3333    18.2140  -0.018 0.985459
VarietyMarvellous:nitro0.4  -4.1667    18.2140  -0.229 0.819832
VarietyVictory:nitro0.4      4.6667    18.2140   0.256 0.798662
VarietyMarvellous:nitro0.6  -4.6667    18.2140  -0.256 0.798662
VarietyVictory:nitro0.6      2.1667    18.2140   0.119 0.905707
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.31 on 60 degrees of freedom
Multiple R-squared:  0.4257,	Adjusted R-squared:  0.3204
F-statistic: 4.043 on 11 and 60 DF,  p-value: 0.0001964
```

Understanding the output of ANOVA

In addition to parameter estimates, a common goal of ANOVA is to test whether levels of categorical predictors are the same. This is often presented in an ANOVA table.

Example: ANOVA table for the model comparing oat yield among 3 varieties under 3 nitrogen levels.



```
Analysis of Variance Table

Response: yield
              Df  Sum Sq Mean Sq F value    Pr(>F)    
Variety        2  1786.4   893.2  1.7949    0.1750    
nitro          3 20020.5  6673.5 13.4108 8.367e-07 ***
Variety:nitro  6   321.7    53.6  0.1078    0.9952    
Residuals     60 29857.3   497.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```