

# One Way ANOVA

## Lecture 18

---

Alexandra Chronopoulou



### **COLLEGE OF LIBERAL ARTS & SCIENCES**

Department of Statistics  
101 Illini Hall, MC-374  
725 S. Wright St.  
Champaign, IL 61820-5710

© Alexandra Chronopoulou. Do not distribute without permission of the author.

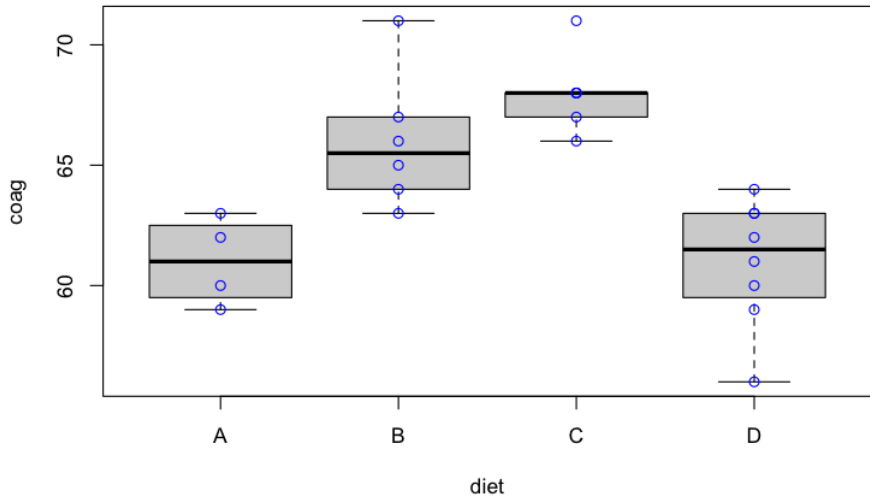
- A **comparative experiment** is intended to answer research questions regarding the differences between the effects of imposing two or more different conditions.
- The imposed conditions are the **treatments**, and they are imposed on the **experimental units**. The effects are measured using the **responses** (usually values of a single response variable).
- The way treatments are assigned to experimental units is called the **design** of the experiment. Some form of **randomization** is usually used. In that case, it is a **randomized experiment** (or sometimes **randomized study**).

# Blood Coagulation Example

- 24 animals were randomly assigned to 4 different diets with goal to study blood coagulation times.
- The samples were taken in a random order.
- This data set can be found in the *faraway* library.

##	coag	diet
## 1	62	A
## 2	60	A
## 3	63	A
## 4	59	A
## 5	63	B
## 6	67	B
## 7	71	B
## 8	64	B
## 9	65	B
## 10	66	B
## 11	68	C
## 12	66	C
## 13	71	C
## 14	67	C
## 15	68	C
## 16	68	C
## 17	56	D
## 18	62	D
## 19	60	D
## 20	61	D
## 21	63	D
## 22	64	D
## 23	63	D
## 24	59	D

# Blood Coagulation Example



- **Factor**: an Independent variable. They can be experimental or observational. In our example: *Diet*
- **Level**: A particular form of the factor. In our example: *Levels of the Diet: A, B, C, D*
- **Treatments**: Factor levels or factor level combinations (if the study contains more than one factors). They provide insights into mechanisms causing the variation being studied.  
Control treatments?
- **Complete Randomized Design**: Experimental units are randomly split into  $r$  groups, and  $r$  treatments are assigned, one per group.

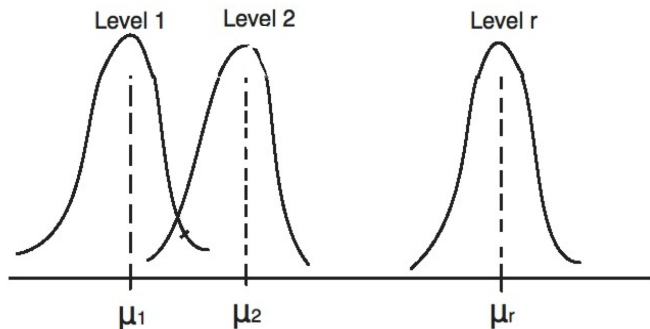
- Data:

group 1	$y_{11},$	$y_{12}$	$\dots$	$y_{1n_1}$
group 2	$y_{21},$	$y_{22}$	$\dots$	$y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
group $r$	$y_{r1},$	$y_{r2}$	$\dots$	$y_{rn_r}$

- $r$  is the number of groups
- $n_i$  denotes the number of obs in the  $i$ th group
- $n = \sum_{i=1}^r n_i$  is the total sample size
- $y_{ij}$  = observation  $j$  for the  $i$ th factor.

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, n_i$$

- $y_{ij}$ : the value of the response in the  $j$ th trial for the  $i$ th factor.
- $\mu_i$ : the population mean for the  $i$ th factor level (treatment).
- $\varepsilon_{ij} \sim^{iid} \mathcal{N}(0, \sigma^2)$





Define the effect of factor level  $i$  on the response, i.e. the **treatment effect** as

$$\alpha_i = \mu_i - \mu$$

where  $\mu$  is the overall mean.

## Factor Effects Model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, r; \quad j = 1, \dots, n_i$$

$$\varepsilon_{ij} \sim^{iid} \mathcal{N}(0, \sigma^2)$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The factor effects model has  $r + 1$  model parameters, i.e.

$$(\mu, \alpha_1, \dots, \alpha_r)$$

- In order for the  $\alpha$ 's to be (uniquely) estimated, we need to impose restrictions.
- The restrictions on the  $\alpha$ 's depend on how  $\mu$  is defined.

---

Model	$\mu$ Definition	$\alpha$ 's Restriction
Reference Cell	$\mu = \mu_1$	$\alpha_1 = 0$
Sum-to-Zero	$\mu = \frac{1}{r} \sum_i \mu_i$	$\sum_i \alpha_i = 0$
Weighted Sum-to-Zero	$\mu = \frac{1}{n} \sum_i n_i \mu_i$	$\sum_i n_i \alpha_i = 0$

---

- The default in R is the Reference Cell model.

## Coagulation Example: Reference Cell (default)

```
contrasts(diet)=contr.treatment(4)
g=lm(coag~diet)
summary(g)
```

```
##
## Call:
## lm(formula = coag ~ diet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.100e+01  1.183e+00  51.554 < 2e-16 ***
## diet2        5.000e+00  1.528e+00   3.273 0.003803 **
## diet3        7.000e+00  1.528e+00   4.583 0.000181 ***
## diet4        2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF, p-value: 4.658e-05
```

```
model.matrix(g)
```

```
##      (Intercept) dietB dietC dietD
## 1              1     0     0     0
## 2              1     0     0     0
## 3              1     0     0     0
## 4              1     0     0     0
## 5              1     1     0     0
## 6              1     1     0     0
## 7              1     1     0     0
## 8              1     1     0     0
## 9              1     1     0     0
## 10             1     1     0     0
## 11             1     0     1     0
## 12             1     0     1     0
## 13             1     0     1     0
## 14             1     0     1     0
## 15             1     0     1     0
## 16             1     0     1     0
## 17             1     0     0     1
## 18             1     0     0     1
## 19             1     0     0     1
## 20             1     0     0     1
## 21             1     0     0     1
## 22             1     0     0     1
## 23             1     0     0     1
## 24             1     0     0     1
```

## Coagulation Example: A coding that fits the Mean Model

```
g1=lm(coag~diet-1)
summary(g1)
```

```
##
## Call:
## lm(formula = coag ~ diet - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## dietA  61.0000     1.1832   51.55 <2e-16 ***
## dietB  66.0000     0.9661   68.32 <2e-16 ***
## dietC  68.0000     0.9661   70.39 <2e-16 ***
## dietD  61.0000     0.8367   72.91 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic: 4399 on 4 and 20 DF, p-value: < 2.2e-16
```

```
model.matrix(gl)
```

```
##      dietA dietB dietC dietD
## 1      1      0      0      0
## 2      1      0      0      0
## 3      1      0      0      0
## 4      1      0      0      0
## 5      0      1      0      0
## 6      0      1      0      0
## 7      0      1      0      0
## 8      0      1      0      0
## 9      0      1      0      0
## 10     0      1      0      0
## 11     0      0      1      0
## 12     0      0      1      0
## 13     0      0      1      0
## 14     0      0      1      0
## 15     0      0      1      0
## 16     0      0      1      0
## 17     0      0      0      1
## 18     0      0      0      1
## 19     0      0      0      1
## 20     0      0      0      1
## 21     0      0      0      1
## 22     0      0      0      1
## 23     0      0      0      1
## 24     0      0      0      1
```

## Coagulation Example: $\sum_i \alpha_i = 0$

```
contrasts(diet) = contr.sum(4)
g2 = lm(coag~diet)
summary(g2)
```

```
##
## Call:
## lm(formula = coag ~ diet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000    0.4979  128.537 < 2e-16 ***
## diet1       -3.0000    0.9736  -3.081 0.005889 **
## diet2        2.0000    0.8453   2.366 0.028195 *
## diet3        4.0000    0.8453   4.732 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```



```
model.matrix(g2)
```

```
##      (Intercept) diet1 diet2 diet3
## 1              1      1      0      0
## 2              1      1      0      0
## 3              1      1      0      0
## 4              1      1      0      0
## 5              1      0      1      0
## 6              1      0      1      0
## 7              1      0      1      0
## 8              1      0      1      0
## 9              1      0      1      0
## 10             1      0      1      0
## 11             1      0      0      1
## 12             1      0      0      1
## 13             1      0      0      1
## 14             1      0      0      1
## 15             1      0      0      1
## 16             1      0      0      1
## 17             1     -1     -1     -1
## 18             1     -1     -1     -1
## 19             1     -1     -1     -1
## 20             1     -1     -1     -1
## 21             1     -1     -1     -1
## 22             1     -1     -1     -1
## 23             1     -1     -1     -1
## 24             1     -1     -1     -1
```

- $E(y_{ij}) = \mu_i$
- $Var(y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$   
Thus, all observations have the same variance, regardless of factor level.
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$  and independent.
- $y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$  and independent.

We can re-state the model as

$y_{ij}$  are independent  $\mathcal{N}(\mu_i, \sigma^2)$

*Minimize* the sum of squared deviations of the observations around their expected values with respect to the parameters:

$$Q = \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \mathbb{E}(y_{ij}))^2$$

If we re-write  $Q$  we have

$$Q = \sum_j (y_{1j} - \mu_1)^2 + \sum_j (y_{2j} - \mu_2)^2 + \dots + \sum_j (y_{rj} - \mu_r)^2$$

So the **least squares estimator** of  $\mu_i$ , denoted by  $\hat{\mu}_i$  is

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Using the appropriate constraints, we can easily extract the estimators for  $\mu$  and  $\alpha_i$ .

Using the model 'g2' with constraint  $\sum_{i=1}^4 \alpha_i = 0$  we have:

```
g2$coef
```

## (Intercept)	diet1	diet2	diet3
## 64	-3	2	4

This implies that  $\hat{\mu} = 64$  and

$$\hat{\alpha}_1 = -3 \quad \hat{\mu}_1 = 64 - 3 = 61$$

$$\hat{\alpha}_2 = 2 \quad \hat{\mu}_2 = 64 + 2 = 66$$

$$\hat{\alpha}_3 = 4 \quad \hat{\mu}_3 = 64 + 4 = 68$$

The estimators for  $\alpha_4$  and the corresponding mean  $\mu_4$ , are obtained them using the constraints:

$$\hat{\alpha}_4 = -\hat{\alpha}_1 - \hat{\alpha}_2 - \hat{\alpha}_3 = 3 - 2 - 4 = -3 \text{ and } \hat{\mu}_4 = 64 - 3 = 61$$

- The LS fit for  $y_{ij}$  is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_i.$$

- Residuals

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i.$$

- RSS

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

i.e. the within-group variation.

Source of Variation	SS	df	MS
Between Groups	$FSS = \sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$r - 1$	$\frac{FSS}{r-1}$
Error (within groups)	$RSS = \sum \sum (y_{ij} - \bar{y}_{i.})^2$	$n - r$	$\frac{RSS}{n-r}$
Total	$TSS = \sum \sum (y_{ij} - \bar{y}_{..})^2$	$n - 1$	

- We want to test whether the means of the groups are really different. We can express this as

$$\left\{ \begin{array}{l} H_0 : \mu_1 = \mu_2 = \dots = \mu_r \\ H_\alpha : \text{not all } \mu_i, i = 1, \dots, r \text{ are equal} \end{array} \right.$$

- or in terms of models

$$\left\{ \begin{array}{l} H_0 : y_{ij} = \mu + \varepsilon_{ij} \\ H_\alpha : y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \end{array} \right.$$

- They are two nested models, so we can use the  $F$ -test

$$\frac{(RSS_0 - RSS_\alpha)/(r - 1)}{RSS_\alpha/(n - r)} \sim F_{r-1, n-r},$$

under  $H_0$ .

- The test statistic can also be written as

$$\frac{FSS/(r - 1)}{RSS/(n - r)} = \frac{\text{Between-group Variation}/(r - 1)}{\text{Within-group Variation}/(n - r)},$$

where  $FSS$ ,  $RSS$  are defined in the ANOVA table.



```
null = lm(coag ~ 1)
anova(null, g2)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet
##   Res.Df RSS Df Sum of Sq      F      Pr(>F)
## 1      23 340
## 2      20 112   3      228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- It does not matter which coding is used for the mean/effects. The results would be the same.

Equivalently, we can get the ANOVA table that contains the same  $F$  test and  $p$ -value:

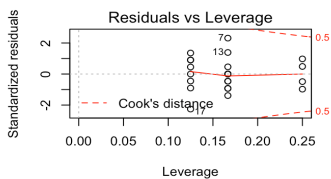
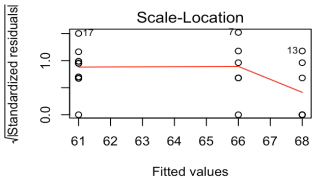
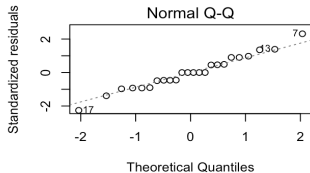
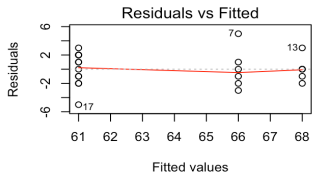
```
anova(g2)
```

```
## Analysis of Variance Table
##
## Response: coag
##      Df Sum Sq Mean Sq F value    Pr(>F)
## diet     3    228    76.0   13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The  $p$ -value is much less than  $\alpha = 5\%$ , so we reject the null and conclude that there are differences among the different types of diet.

- Check for outliers/ unusual observations.
- Check the residuals vs. fitted values plot for departures from the constant variance assumption.
- Check the Q-Q plot for departures from the normality assumption.

```
par(mfrow=c(2,2))
plot(g2)
```



### Levene's Test for Equality of Variances:

- Run Regression `abs(residuals) ~ X`, i.e. use `abs(residuals)` as the response in a new one-way ANOVA.
- If the  $p$ -value for the  $F$ -test is **greater** than 1% level, then we conclude that there is no evidence of a non-constant variance.

$H_0$  : All group variances are equal.

```
g2=lm(coag-diet)
summary(lm(abs(g2$res) ~diet))

##
## Call:
## lm(formula = abs(g2$res) ~ diet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.000  -1.000   0.000   0.625   3.000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.6250     0.3013   5.394 2.8e-05 ***
## diet1         -0.1250     0.5891  -0.212   0.834
## diet2          0.3750     0.5115   0.733   0.472
## diet3         -0.6250     0.5115  -1.222   0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.432 on 20 degrees of freedom
## Multiple R-squared:  0.09559,    Adjusted R-squared:  -0.04007
## F-statistic: 0.7046 on 3 and 20 DF,  p-value: 0.5604
```

- Since the  $p$ -value is greater than 0.01, there is no evidence of unequal variances.