

## STA305/1004 - Class 12

Test coverage:  
includes all material  
up to & including this  
class.

4 Questions:  
4-5 parts each question

Mid-term Test on March 2  
11:10-12:45

February 24, 2016

## Today's class


- ▶ ANOVA table
- ▶ ANOVA identity
- ▶ Degrees of freedom and ANOVA table
- ▶ Geometry of ANOVA
- ▶ Two estimates of the population variance
- ▶ Mean squares
- ▶ F statistic
- ▶ Assumptions

## Comparing more than two treatments

If interest is in designing an experiment to compare more than two treatments then the previous designs will need to be modified.

- ▶ A clinical trial comparing three drugs A, B, C to reduce duration of intubation for patients on mechanical ventilation.
- ▶ Coagulation time of blood samples for animals receiving four different diets A, B, C, D.

What are the null and alternative hypotheses in these two scenarios?


$$\begin{aligned}H_0: \mu_A &= \mu_B = \mu_C = \mu_D, \\ H_1: \mu_i &\neq \mu_j\end{aligned}$$

## Blood Coagulation Study

Coagulation times for blood samples drawn from 24 animals receiving four different diets A, B, C, and D.

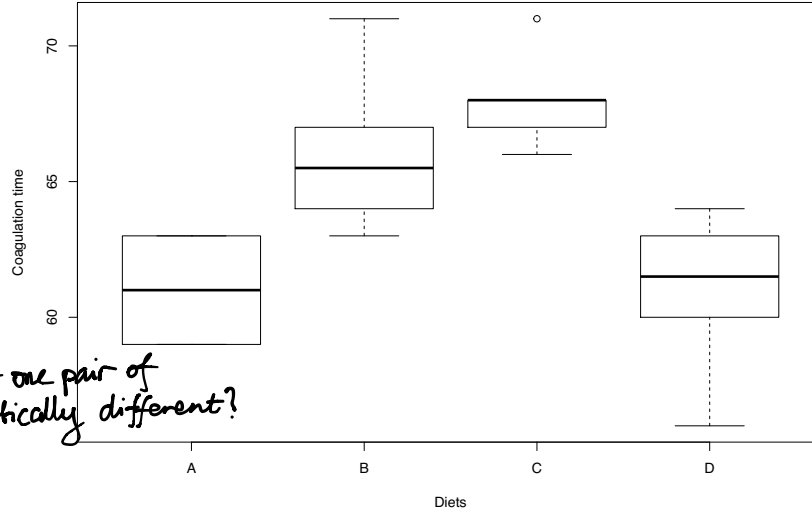
	4 diets			
	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

6 obs

average of 24.

# Blood Coagulation Study

Coagulation time from 24 animals randomly allocated to four diets



*H<sub>1</sub>: at least one pair of diets statistically different?*

Do the boxplots show evidence of a difference between diets?

## Analysis of Variance (ANOVA)

- ▶ An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.
- ▶ Fisher introduced the method in his 1925 book “Statistical Methods for Research Workers”.
- ▶ The statistical procedure enables experimenters to answer several questions at once.
- ▶ The prevailing method at the time was to test one factor at a time in an experiment.

## Analysis of Variance (ANOVA) table

- ▶ The between treatments variation and within treatment variation are two components of the total variation in the response.
- ▶ In the coagulation study data we can break up each observation's deviation from the grand mean into two components: treatment deviations; and residuals within treatment deviations.
- ▶ Let  $y_{ij}$  be the  $j$ th ( $j = 1, \dots, 6$ ) observation taken under treatment  $i = 1, 2, 3, 4$ .

total deviation from grand mean  $\nearrow$

$$y_{ij} - \bar{y}_{..} = \underbrace{(y_{i.} - \bar{y}_{..})}_{\text{treatment deviation}} + \underbrace{(y_{ij} - \bar{y}_{i.})}_{\text{residual deviation}}$$
$$y_{i.} = \sum_{j=1}^n y_{ij}, \quad \bar{y}_{i.} = y_{i.}/n,$$

$$y_{..} = \sum_{i=1}^a \sum_{j=1}^n y_{ij}, \quad \bar{y}_{..} = y_{..}/N,$$

$y_{ij}$  =  $j$ th obs in  
the  $i$ -th treat group  
 $j=1, \dots, 6$   
 $i=1, \dots, 4$

## Analysis of Variance (ANOVA) model

- ▶ Let  $y_{ij}$  be the  $j$ th observation taken under treatment  $i = 1, \dots, a$ .

$$E(y_{ij}) = \mu_i = \mu + \tau_i,$$

and  $\text{Var}(y_{ij}) = \sigma^2$  and the observations are mutually independent.

- ▶ The parameter  $\tau_i$  is the  $i$ th treatment effect.
- ▶ The parameter  $\mu$  is the overall mean.

$$\tau_i = \mu_i - \mu$$
$$\mu = \sum_{i=1}^4 \frac{\mu_i}{4}$$



## Analysis of Variance (ANOVA) model

e.g.  $a=4$

We are interested in testing if the  $a$  treatment means are equal.

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j, i \neq j.$$

There will be  $n$  observations under the  $i$ th treatment.

$$y_{i\cdot} = \sum_{j=1}^n y_{ij}, \quad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^a \sum_{j=1}^n y_{ij}, \quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

where  $N = an$  is the total number of observations. The “dot” subscript notation means sum over the subscript that it replaces.

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \quad \text{The ANOVA identity}$$

$$= \sum_i \sum_j \underbrace{(y_{ij} - \bar{y}_{i.})}_a + \underbrace{(\bar{y}_{i.} - \bar{y}_{..})}_b$$

The total sum of squares  $SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$  can be written as

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + 2(\sum_i \sum_j (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})) + \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$(a+b)^2 = a^2 + 2ab + b^2$

$$\sum_{i=1}^a \sum_{j=1}^n [(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2$$

by adding and subtracting  $\bar{y}_{i.}$  to  $SS_T$ .

works  $\because$  vectors  $b, c$  are orthogonal

It can be shown that

$$\text{Show } \sum_i \sum_j (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) = 0$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2$$

$\nwarrow$  deviations within treatments  
 $\nearrow$  treat mean  
 $\nwarrow$  deviations between treatment means

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = \underbrace{n \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2}_{\text{Sum of Squares Due to Treatment}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2}_{\text{Sum of Squares Due to Error}}$$

$$= SS_{Treat} + SS_E.$$

## The ANOVA identity

This is sometimes called the analysis of variance identity. It shows how the total sum of squares can be split into two sum of squares: one part that is due to differences between treatments; and one part due to differences within treatments.

## The ANOVA identity

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

- The decomposition of the first observation  $y_{11} = 60$  in diet A is

$$y_{11} - \bar{y}_{..} = (y_{1.} - \bar{y}_{..}) + (y_{11} - \bar{y}_{1.})$$

$$\underline{60 - 64} = (61 - 64) + (60 - 61)$$

$$\underline{-4} = \underline{-3} + \underline{-1} \quad \text{called residuals}$$

- If each observation is decomposed in this manner then there will be three tables of residuals: total residuals; between treatment residuals; and within treatment residuals.

## Example - Blood coagulation study ( $SS_T$ )

The deviations from the grand average ( $y_{ij} - \bar{y}_{..}$ ) are in the table below:

A	B	C	D
-4	1	7	-2
-1	2	2	-4
-5	3	4	-3
-1	-1	4	0
-2	0	3	-1
-5	7	4	-8

The total sum of squares is obtained by squaring all the entries in this table and summing:  $SS_T = (-4)^2 + (-1)^2 + \cdots + (-8)^2 = 340$ .

## Example - Blood coagulation study ( $SS_{Treat}$ )

The between treatment deviations ( $y_{i.} - \bar{y}_{..}$ ) are in the table below:

A	B	C	D
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3

The sum of squares due to treatment is obtained by squaring all the entries in this table and summing:  $SS_{Treat} = (-3)^2 + (2)^2 + \cdots + (-3)^2 = 228$ .

## Example - Blood coagulation study ( $SS_E$ )

The within treatment deviations ( $y_{ij} - \bar{y}_{i\cdot}$ ) are in the table below:

A	B	C	D
-1	-1	3	1
2	0	-2	-1
-2	1	0	0
2	-3	0	3
1	-2	-1	2
-2	5	0	-5

The sum of squares due to error ( $y_{ij} - \bar{y}_{i\cdot}$ ) is obtained by squaring the entries in this table and summing:  $SS_E = (-1)^2 + (2)^2 + \cdots + (-5)^2 = 112$ .

$$\underbrace{340}_{SS_T} = \underbrace{228}_{SS_{Treat}} + \underbrace{112}_{SS_E}.$$

Which illustrates the ANOVA identity for the blood coagulation study.

## ANOVA - degrees of freedom

### The deviations

- ▶  $SS_{Treat}$  is called the sum of squares due to treatments (i.e., between treatments), and  $SS_E$  is called the sum of squares due to error (i.e., within treatments).
- ▶ There are  $an = N$  total observations. So  $SS_T$  has  $N - 1$  degrees of freedom.
- ▶ There are  $a$  treatment levels so  $SS_{Treat}$  has  $a - 1$  degrees of freedom.
- ▶ Within each treatment there are  $n$  replicates with  $n - 1$  degrees of freedom. There are  $a$  treatments. So, there are  $a(n - 1) = an - a = N - a$  degrees of freedom for error.



## Geometry and the ANOVA Table

A	B	C	D
-4	1	7	-2
-1	2	2	-4
-5	3	4	-3
-1	-1	4	0
-2	0	3	-1
-5	7	4	-8

Table of Total residual

A	B	C	D
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3

Between treatment residuals

A	B	C	D
-1	-1	3	1
2	0	-2	-1
-2	1	0	0
2	-3	0	3
1	-2	-1	2
-2	5	0	-5

within treatment residuals

## Geometry and the ANOVA Table

- ▶ Let  $a$  be the vector of deviations from the grand mean,
- ▶ Let  $b$  be the vector of deviations of treatment deviations
- ▶ Let  $c$  be the vector of within-treatment deviations.

$$a = (-4, -1, -5, -1, -2, -5, 1, 2, 3, -1, 0, 7, 7, 2, 4, 4, 3, 4, -2, -4, -3, 0, -1, -8),$$

$$b = (-3, -3, -3, -3, -3, -3, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, -3, -3, -3, -3, -3, -3),$$

$$c = (-1, 2, -2, 2, 1, -2, -1, 0, 1, -3, -2, 5, 3, -2, 0, 0, -1, 0, 1, -1, 0, 3, 2, -5).$$

## Geometry and the ANOVA Table

- ▶ The dot product of  $b$  and  $c$ ,  $b \cdot c$ , is

$b * c$  ← residual deviations

treatment  
deviations

$$b = (b_1, b_2, \dots, b_6)$$

$$c = (c_1, c_2, \dots, c_6)$$

$$b \cdot c = b_1 c_1 + b_2 c_2 + \dots + b_6 c_6$$

||

A	B	C	D
3	-2	12	-3
-6	0	-8	3
6	2	0	0
-6	-6	0	-9
-3	-4	-4	-6
6	10	0	15

`sum(b*c)` =  $b \cdot c$

= dot product of  $b$  &  $c$

[1] 0

- ▶ Therefore, the vectors  $b$  and  $c$  are orthogonal.
- ▶ Thus, the vector  $a$  is the hypotenuse of a right triangle with sides  $b$  and  $c$ .

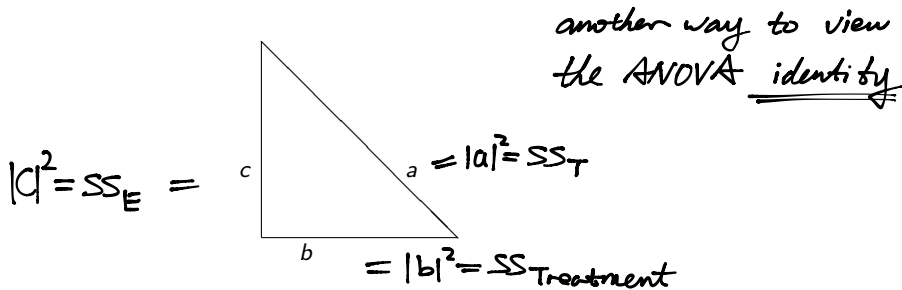
## Geometry and the ANOVA Table

Pythagoras' theorem in  $n$  dimensions is  $|a|^2 = |b|^2 + |c|^2$ , where  $|a| = \sqrt{a_1^2 + \dots + a_n^2}$ .

The ANOVA identity can be seen using Pythagoras' theorem since

$$|a|^2 = SS_T, |b|^2 = SS_{Treat}, |c|^2 = SS_E.$$

If there were only three observations then the vectors would be as shown below.



The degrees of freedom are the dimensions in which the vectors are free to move given the constraints.

## ANOVAs Two Estimates of the Population Variance ( $\sigma^2$ )

$$SS_E = \sum_{i=1}^a \left[ \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2 \right]$$

If the term inside the brackets is divided by  $n - 1$  then it is the sample variance for the  $i$ th treatment

$$S_i^2 = \frac{\sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2}{n - 1}, \quad i = 1, \dots, a.$$

Combining these  $a$  variances to give a single estimate of the common population variance

$$\frac{(n - 1)S_1^2 + \dots + (n - 1)S_a^2}{(n - 1) + \dots + (n - 1)} = \frac{SS_E}{N - a}.$$

Thus,  $SS_E$  is a pooled estimate of the common variance  $\sigma^2$  within each of the  $a$  treatments.

## ANOVAs Two Estimates of the Population Variance ( $\sigma^2$ )

If there were no differences between the  $a$  treatment means  $\bar{y}_{i\cdot}$ , we could use the variation of the treatment averages from the grand average to estimate  $\sigma^2$ .

$$\frac{n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{a - 1} = \frac{SS_{Treat}}{a - 1}$$

is an estimate of  $\sigma^2$  when the treatment means are all equal.

*two estimates of  $\sigma^2$*

## ANOVAs Two Estimates of the Population Variance ( $\sigma^2$ )

- ▶ The analysis of variance identity gives two estimates of  $\sigma^2$ .
- ▶ One is based on the variability within treatments and one based on the variability between treatments.
- ▶ If there are no differences in the treatment means then these two estimates should be similar.
- ▶ If these estimates are different then this could be evidence that the difference is due to differences in the treatment means.

## ANOVA - Mean square error

The mean square for treatment is defined as

$$MS_{Treat} = \frac{SS_{Treat}}{a - 1}$$

and the mean square for error is defined as

$$MS_E = \frac{SS_E}{N - a}.$$

$$\frac{\text{sum of squares}}{df}$$



## ANOVA - F statistic

$$\sum_{i=1}^V z_i^2 = Q_1 + Q_2 + \dots + Q_s$$

$SS_T$   $S \leq V \rightarrow SS_{Treat} + SS_E$

- ▶  $SS_{Treat}$  and  $SS_E$  are independent.
- ▶ It can be shown that  $SS_{Treat}/\sigma^2 \sim \chi_{a-1}^2$  and  $SS_E/\sigma^2 \sim \chi_{N-a}^2$ .
- ▶ Thus, if  $H_0 : \mu_1 = \dots = \mu_a$  is true then the ratio

$$F = \frac{MS_{Treat}}{MS_E} \sim F_{a-1, N-a}.$$

*assume  
normality*

*To prove this Cochran's  
Theorem (But not responsible  
for the proof).*

## ANOVA - F statistic

- ▶ In Fisher's 1925 book that introduced ANOVA he included one F table for various numerator and denominator degrees of freedom.
- ▶ The table gave the critical values for only the 5% points.
- ▶ As use of the method spread so did the use of the 5% level. (Stigler, 2008)

		1	2	3	numerator df ... 20
denom df	1				
	2				
	3				
	⋮				
	10				

## ANOVA Table - Blood coagulation study

The ANOVA table for the coagulation data can be calculated in R.

```
aov.diets <- aov(y~diets,data=tab0401)
summary(aov.diets)
```

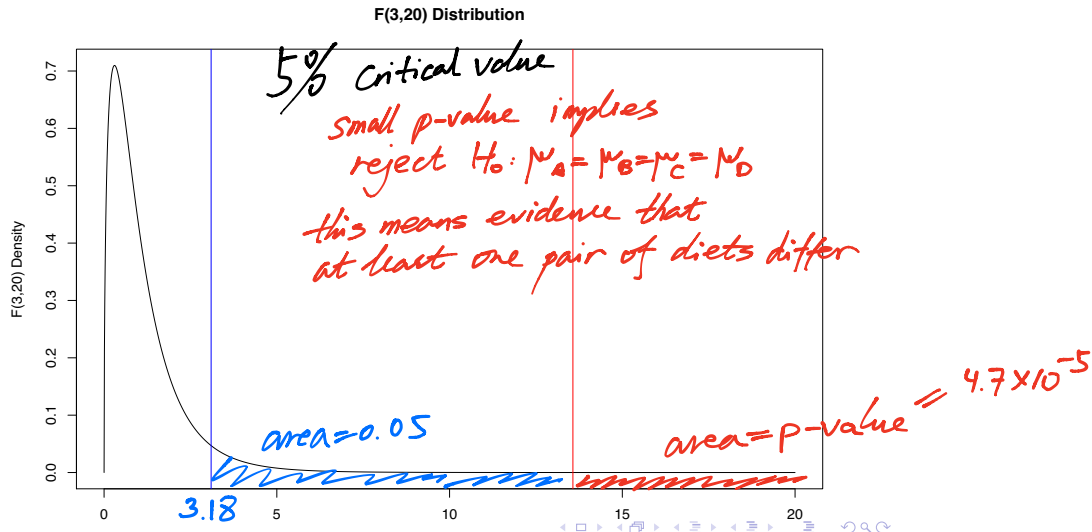
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diets	3	228	76.0	13.57	4.66e-05 ***
Residuals	20	112	5.6		
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

In this example  $a - 1 = 3$ ,  $N - a = 20$ ,  $SS_{Treat} = 228$ ,  $SS_E = 112$ ,  $MS_{Treat} = 228/3 = 76.0$ ,  $MS_E = 112/20 = 5.6$ ,  $F = 76/5.6 = 13.57$ .

$a=4$ ,  $N=24$ . if we compare 13.57 to  $F_{3,20}$

## ANOVA Table - Blood coagulation study

The observed  $F$  value of 13.57 is shown on the  $F_{3,20}$  distribution. The p-value of the test is the area under the density to the right of 13.57 (red line). The 95% critical value of the  $F_{3,20}$  is 3.10 (blue line). In other words,  $P(F_{3,20} > 3.10) = 0.05$ .



## ANOVA Table - Blood coagulation study

The p-value could also be calculated directly using the cdf of the  $F_{3,20}$  distribution.

```
1-pf(q = 13.57, df1 = 3, df2 = 20)
```

$$F(x) = \int_0^x f(x) dx$$

$\hookrightarrow$  density of  $f$

[1] 4.66169e-05

- ▶ The small p-value indicates that the difference between at least one pair of the treatment means is significantly different from 0.
- ▶ The p-value does not indicate which pairs are significantly different.

$F_{3,20}$  dist'n.

$$= P(F \leq x)$$

$$P(F > x) = 1 - P(F \leq x)$$

## General ANOVA

The general form of the ANOVA table is

Source of variation	df	Sum of squares	Mean square	F
Between treatments	$a - 1$	$SS_{Treat}$	$MS_{Treat}$	$F = \frac{MS_{Treat}}{MS_E}$
Within treatments	$N - a$	$SS_E$	$MS_E$	

$$MS_{Treat} = SS_{Treat} / (a - 1)$$

$$MS_E = SS_E / (N - a)$$

## ANOVA Assumptions

The calculations that make up an ANOVA table require no assumptions. You could write 24 numbers in the ANOVA table and complete the table using the ANOVA identity and definitions of mean square and F statistic. However, using these numbers to make inferences about differences in treatment means will require certain assumptions.

*Geometry or Algebra*

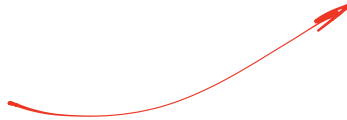
## ANOVA Assumptions - Additive Model

### 1. Additive model.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

The parameters  $\tau_i$  are interpreted as the treatment effect of the  $i^{th}$  mean. That is, if  $\mu_i$  is the mean of  $i^{th}$  group and  $\mu$  is the overall mean then  $\tau_i = \mu_i - \mu$ .

$\tau_i =$





## ANOVA Assumptions - iid with common variance

2. Under the assumption that the errors  $\epsilon_{ij}$  are independent and identically distributed (iid) with common variance  $\text{Var}(\epsilon_{ij}) = \sigma^2$ , for all  $i, j$  then

$$E(MS_{Treat}) = \sum_{i=1}^a \tau_i^2 + \sigma^2, \quad E(MS_E) = \sigma^2.$$

If there are no differences between the treatment means then  $\tau_1 = \dots = \tau_4 = 0$  and  $\sum_{i=1}^a \tau_i^2 = 0$  then both  $MS_{treat}$  and  $MS_E$  would be estimates  $\sigma^2$ .

## ANOVA Assumptions - errors are normally distributed

3. If  $\epsilon_{ij} \sim N(0, \sigma^2)$  then  $MS_{Treat}$  and  $MS_E$  are independent. Under the null hypothesis that  $\sum_{i=1}^a \tau_i^2 = 0$  the ratio

$$F = \frac{MS_{Treat}}{MS_E}$$

is the ratio of two independent estimates of  $\sigma^2$ . Therefore,

$$\frac{MS_{Treat}}{MS_E} \sim F_{a-1, N-a}.$$

## Example - checking the assumptions in the blood coagulation study - additivity

The additive model assumption seems plausible since the observations from each diet can be viewed as the sum of a common mean plus a random error term.

*Can not be directly checked.*

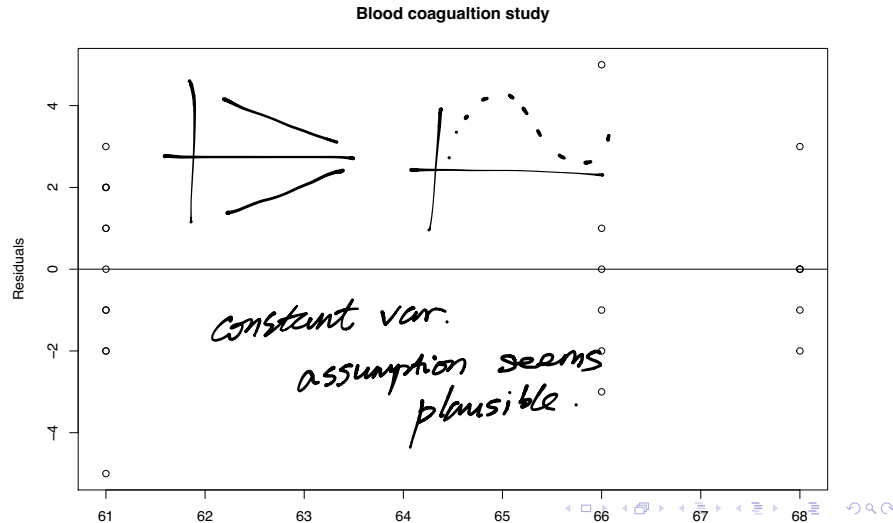
*But can make an argument based on context of expect.*

## Example - checking the assumptions in the blood coagulation study - constant variance

- ▶ The common variance assumption can be investigated by plotting the residuals versus the fitted values of the ANOVA model.
- ▶ A plot of the residuals versus fitted values can be used to investigate the assumption that the residuals are randomly distributed and have constant variance.
- ▶ If the points fall randomly on both sides of 0, with no recognizable patterns in the points then this is an indication that the assumption is satisfied.

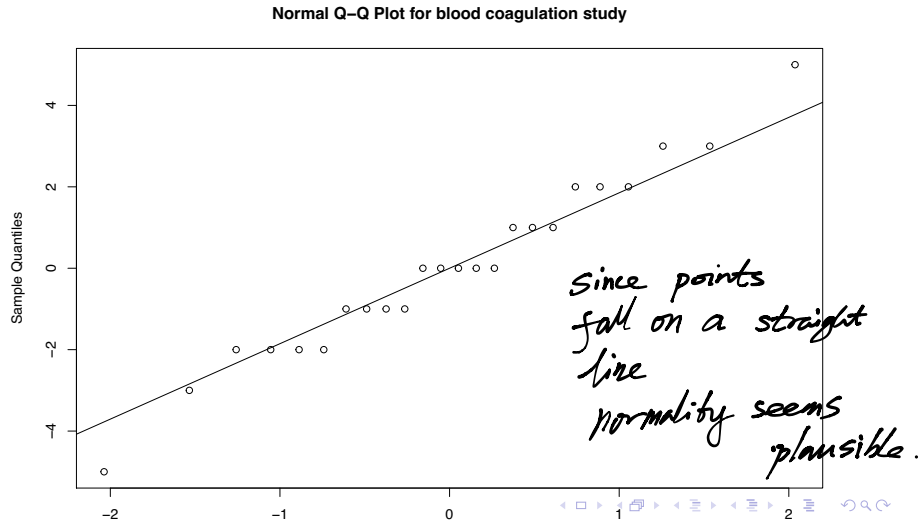
## Example - checking the assumptions in the blood coagulation study - constant variance

```
plot(aov.diets$fitted.values,aov.diets$residuals,ylab="Residuals",  
     xlab="Fitted",main="Blood coagulation study")  
abline(h=0)
```



## Example - checking the assumptions in the blood coagulation study - normality

```
qqnorm(aov.diets$residuals,  
       main="Normal Q-Q Plot for blood coagulation study")  
qqline(aov.diets$residuals)
```



END