

# MATH 4720 / MSCS 5720

---

**Instructor: Mehdi Maadooliat**

## **Chapter 8 (Part B)**



**Department of Mathematics, Statistics and Computer Science**



# BACK TO ANOVA: WHAT IF EQUALITY OF THE VARIANCES **FAIL**?

- The assumption that the sample are generated from normal distribution is not very important as long as the total sample size is **large**.
- Note that conceptually the test statistic  $F = \frac{SS_B/df_B}{SS_E/df_E}$  still makes sense.
- The major problem is with the assumption  $\sigma_1 = \sigma_2 = \dots = \sigma_t$ . If this cannot be assumed, F- test **must not be used**.
- If  $H_0: \sigma_1 = \sigma_2 = \dots = \sigma_t$  is **rejected**, then one approach is to **transform** the data if the variances  $\sigma^2$  is a function of the mean  $\mu$ .

# EQUALITY OF VARIANCES **FAIL** CONT'D

- Transforming the data:**

Treatment Levels				
<b>1</b>	<b>2</b>	<b>3</b>	<b>.</b>	<b>t</b>
$y_{11}$	$y_{21}$	$y_{31}$	$\cdot$	$y_{t1}$
$y_{12}$	$y_{22}$	$y_{32}$	$\cdot$	$y_{t2}$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$y_{1n_1}$	$y_{2n_2}$	$y_{3n_3}$	$\cdot$	$y_{tn_t}$
=====				
$N(\mu_1, \sigma_1^2)$	$N(\mu_2, \sigma_2^2)$	$N(\mu_3, \sigma_3^2)$	$\cdot$	$N(\mu_t, \sigma_t^2)$

- If  $\sigma^2 \propto \mu$ , then use  $Y_T = \sqrt{Y}$  or  $\sqrt{Y + 0.375}$**
- If  $\sigma^2 \propto \mu^2$ , then use  $Y_T = \ln(Y)$  or  $\ln(Y + 1)$**
- If  $\sigma^2 \propto \mu(1 - \mu)$ , then use  $Y_T = \sin^{-1} \sqrt{Y}$**



## BOOK EXAMPLE 8.4

- **Biologists believe that Mississippi river causes the oxygen level to be depleted near the Gulf of Mexico. To test this hypothesis water samples are taken at different distances from the mouth of Mississippi river, and the amounts of dissolve oxygen (in ppm) are recorded**

Oxygen Content	Distance			
	1 KM	5 KM	10 KM	20 KM
	1	4	20	37
	5	8	26	30
	.	.	.	.
	.	.	.	.
	2	3	24	33
	$\bar{y}_{1.} = 2.2$	$\bar{y}_{2.} = 4.6$	$\bar{y}_{3.} = 21.2$	$\bar{y}_{4.} = 31.4$
	$s_1 = 1.476$	$s_2 = 2.119$	$s_3 = 4.7333$	$s_4 = 5.522$

- **Minitab**

## EXAMPLE 8.4 CONT'D

- $H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_t$

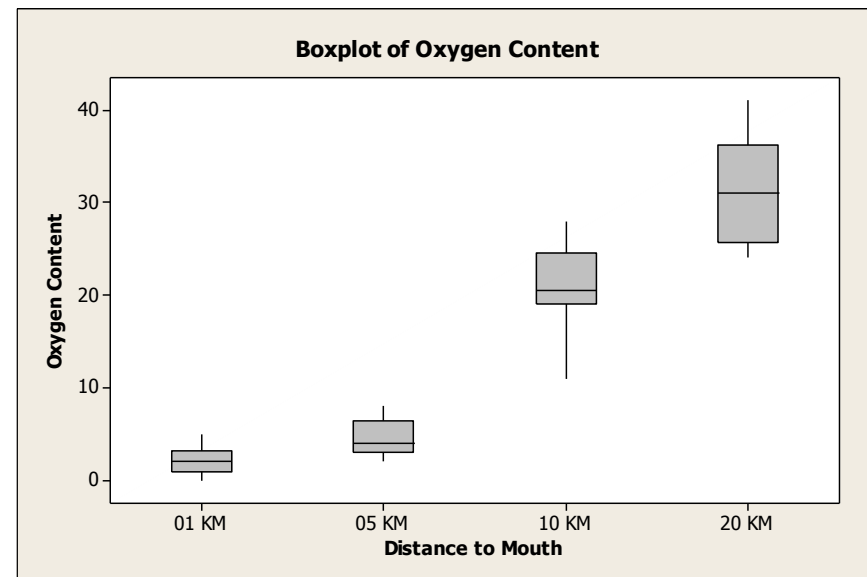
### Test for Equal Variances: Oxygen Content versus Distance to Mouth

95% Bonferroni confidence intervals for standard deviations

Distance to Mouth	N	Lower	StDev	Upper
01 KM	10	0.92361	1.47573	3.2636
05 KM	10	1.32602	2.11870	4.6855
10 KM	10	2.96214	4.73286	10.4668
20 KM	10	3.45583	5.52167	12.2113

Bartlett's Test (Normal Distribution)  
Test statistic = 17.17, p-value = 0.001

Levene's Test (Any Continuous Distribution)  
Test statistic = 3.70, p-value = 0.020



- Levene's test  $p$  - value is 0.02.
- **Reject** the  $H_0$  : Equality of variances
- Let's calculate  $\frac{s_i^2}{\bar{y}_i}$  for  $i = 1, 2, 3, 4$

## EXAMPLE 8.4 CONT'D

- |                      |                      |                       |                       |
|----------------------|----------------------|-----------------------|-----------------------|
| $\bar{y}_{1.} = 2.2$ | $\bar{y}_{2.} = 4.6$ | $\bar{y}_{3.} = 21.2$ | $\bar{y}_{4.} = 31.4$ |
| $s_1 = 1.476$        | $s_2 = 2.119$        | $s_3 = 4.7333$        | $s_4 = 5.522$         |
- $\frac{s_1^2}{\bar{y}_{1.}} = 0.99$      $\frac{s_2^2}{\bar{y}_{2.}} = 0.97$      $\frac{s_3^2}{\bar{y}_{3.}} = 1.06$      $\frac{s_4^2}{\bar{y}_{4.}} = 0.97$
- So, nearly, Variance  $\propto$  Mean.**
- We use the transformation  $Y_T = \sqrt{Y + 0.375}$**

- Now, the ANOVA on this transformed data can be performed.**

Distance				
Transformed Data	1 KM	5 KM	10 KM	20 KM
	1.173	2.092	4.514	6.114
	2.318	2.894	5.136	5.511
	.	.	.	.
	.	.	.	.
	1.541	1.837	4.937	5.777
	$\bar{y}_{1.} = 1.54$	$\bar{y}_{2.} = 2.19$	$\bar{y}_{3.} = 4.62$	$\bar{y}_{4.} = 5.62$
	$s_1 = 0.24$	$s_2 = 0.22$	$s_3 = 0.29$	$s_4 = 0.24$

# A COMPREHENSIVE MODELING APPROACH (USEFUL FOR MODEL ASSESSMENT AND EXTENSIONS)



- To generalize the ANOVA, it is easier to think of one-factor ANOVA in the following way:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, t$$

- Here,
  - $\tau_i$  is the effect due to  $i^{th}$  treatment,
  - $\mu$  is the overall effect irrespective of the treatment,
  - $\epsilon_{ij}$ s are the random errors
- Assumption:
  - The random errors  $\epsilon_{ij}$ s are independent and normally distributed
  - $Var(\epsilon_{ij}) = \sigma^2$  (**a constant value**)

# HOW ABOUT THE HYPOTHESIS TEST?

- To test the effect of the treatment, we can test
  - $H_0: \tau_i = 0$ , for all  $i = 1, 2, \dots, t$
  - $H_a: \tau_i \neq 0$ , for some  $i$
- Test Statistics and decision rule are same as before
- **TS:**  $F = \frac{SS_B/df_B}{SS_E/df_E}$
- **Decision Rule:** Reject  $H_0$  in favor of  $H_a$  if
  - $F > F_\alpha(df_B, df_E)$



# CHECKING THE ASSUMPTIONS

- To check the assumption, we first estimate the errors  $\epsilon_{ij}$  by  $r_{ij}$  (called residuals)

$$r_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i$$

- To test, **normal distribution of errors**  $\epsilon_{ij}$ , we look at the normal probability plot of  $r_{ij}$ .
- To test that the **Var( $\epsilon_{ij}$ ) = constant**, we look at the scatter plot of the residuals  $r_{ij}$  and the predicted values  $\hat{y}_{ij}$ , where

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i$$

## GOING BACK TO EXAMPLE 8.4

- **Biologists believe that Mississippi river causes the oxygen level to be depleted near the Gulf of Mexico. To test this hypothesis water samples are taken at different distances from the mouth of Mississippi river, and the amounts of dissolve oxygen (in ppm) are recorded**

Oxygen Content	Distance			
	1 KM	5 KM	10 KM	20 KM
	1	4	20	37
	5	8	26	30
	.	.	.	.
	.	.	.	.
	2	3	24	33
	$\bar{y}_{1.} = 2.2$	$\bar{y}_{2.} = 4.6$	$\bar{y}_{3.} = 21.2$	$\bar{y}_{4.} = 31.4$
	$s_1 = 1.476$	$s_2 = 2.119$	$s_3 = 4.7333$	$s_4 = 5.522$

- **Minitab**

# WRONG ANOVA

- Minitab

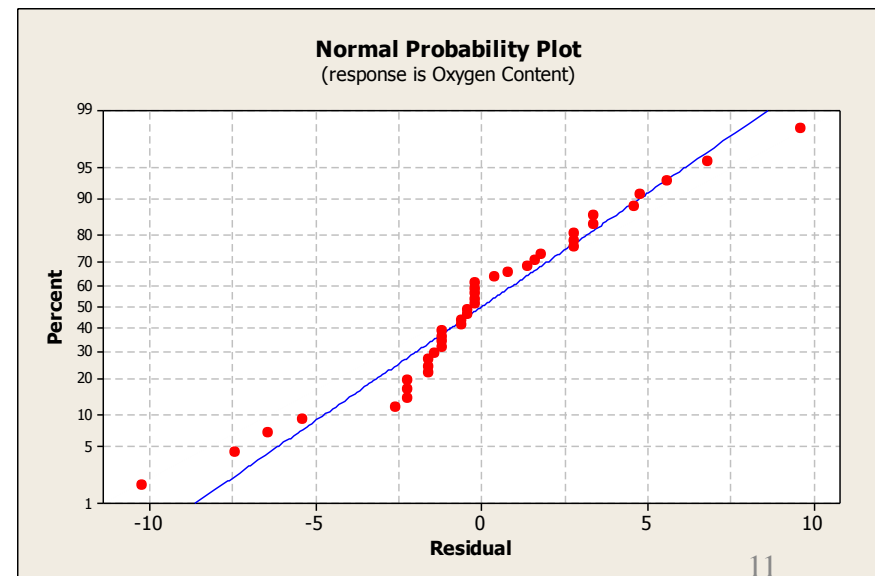
Stat → ANOVA → One-Way

One-way ANOVA: Oxygen Content versus Distance to Mouth

Source	DF	SS	MS	F	P
Distance to Mouth	3	5793.1	1931.0	129.70	0.000
Error	36	536.0	14.9		
Total	39	6329.1			

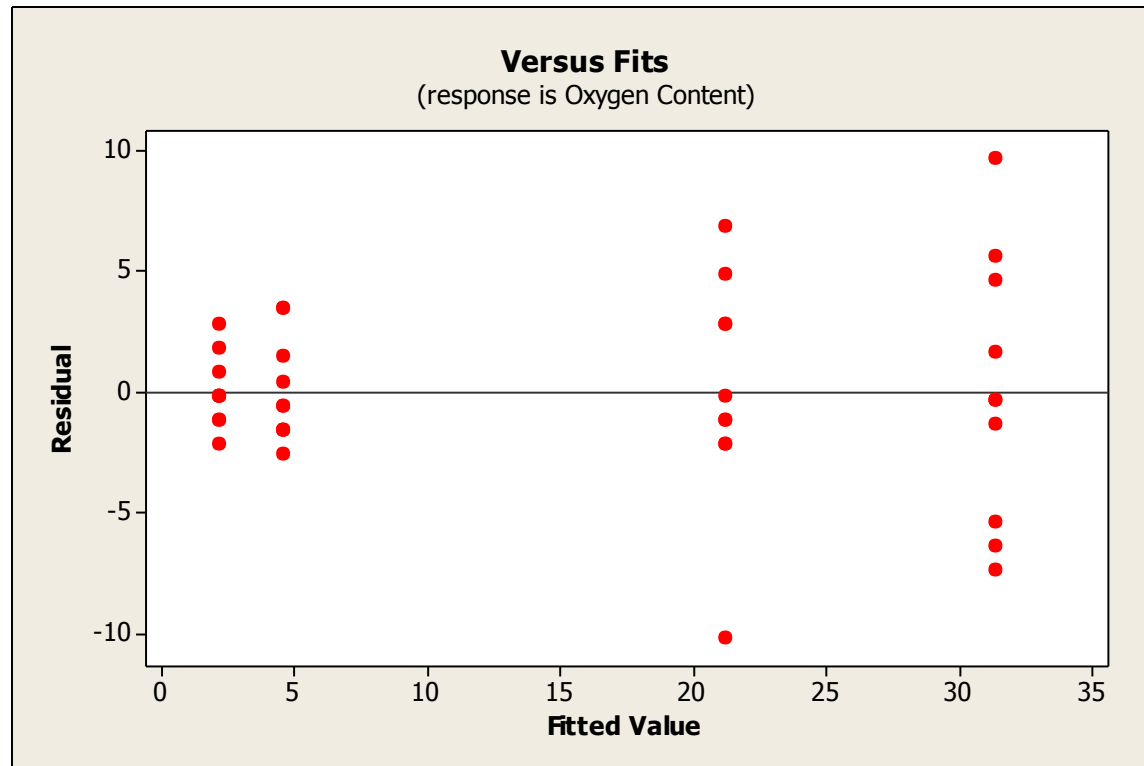
S = 3.859    R-Sq = 91.53%    R-Sq(adj) = 90.83%

- Although the normal probability plot is not very closed to straight line
- But we have relatively large total sample size  $n = 40$



# WRONG ANALYSIS CONT'D

- **Fitted values versus Residuals:**
  - Scatterplot of  $\hat{y}_{ij}$  versus  $r_{ij}$



- Due to cone shape, we can conclude that  $Var(\epsilon_i)$  is not constant.
- We can further say that this variance is a function of the mean of  $y_i$ .

# CORRECT ANALYSIS BASED ON TRANSFORMED DATA

- $Var(y_i) = Var(\epsilon_i) \propto E(y_i)$  **the mean of  $y_i$ .**
- $Y_T = \sqrt{Y + 0.375}$

## One-way ANOVA: Y\_t versus Distance to Mouth

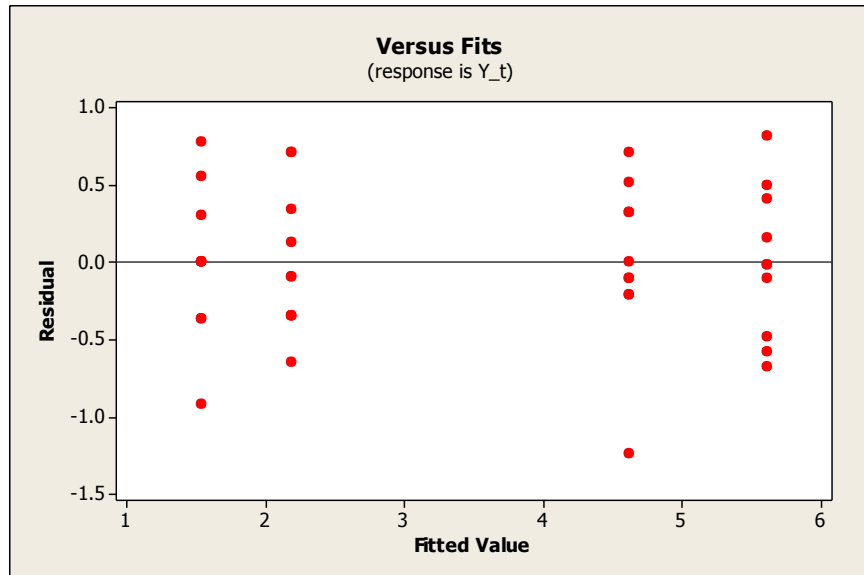
Source	DF	SS	MS	F	P
Distance to Mouth	3	113.095	37.698	153.30	0.000
Error	36	8.853	0.246		
Total	39	121.948			

S = 0.4959    R-Sq = 92.74%    R-Sq(adj) = 92.14%

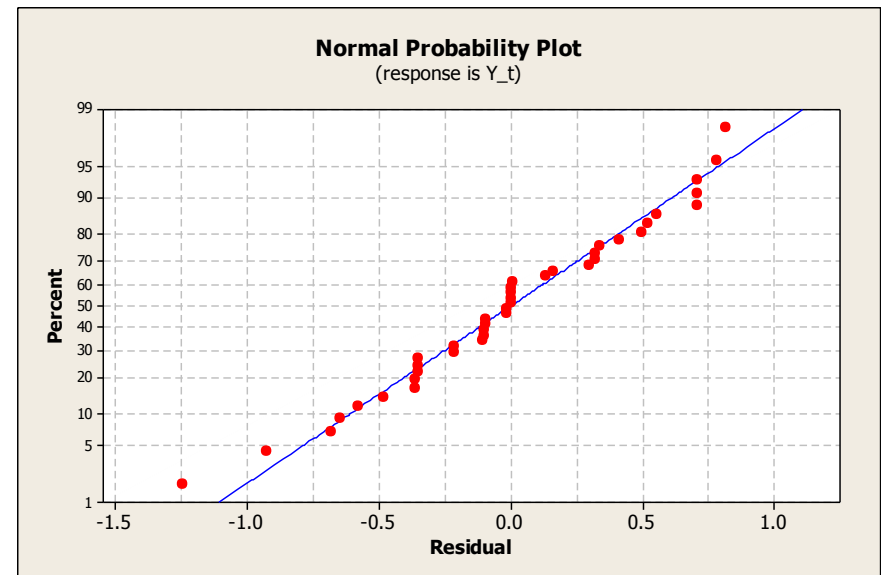
- **For the transformed variable ANOVA we need to check the assumptions**
  - **Normal distributions of the errors**
  - **$Var(\epsilon_i)$  = constant are satisfied.**

# CHECKING THE ASSUMPTIONS BASED ON TRANSFORMED DATA

- $Var(\epsilon_i) = \text{constant}$



- Normal distributions of the errors

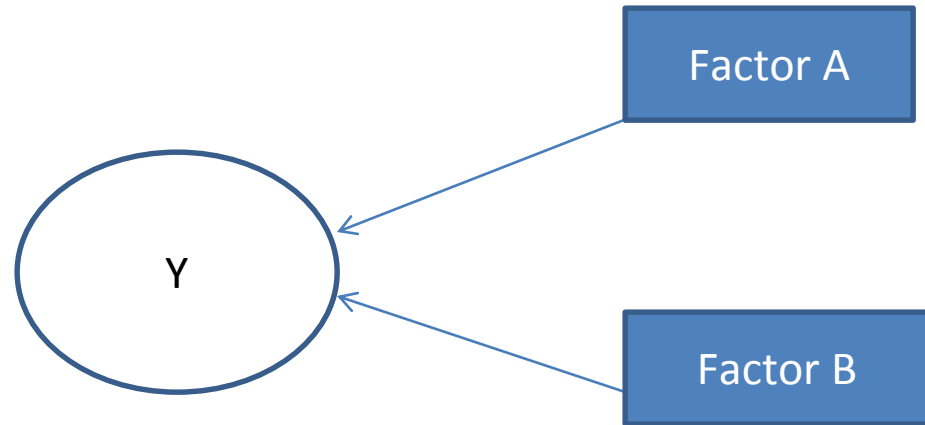


# TWO-FACTOR ANALYSIS OF VARIANCE

- **Two Way ANOVA**

- **Minitab**

**Stat → ANOVA → Two-Way**



**Levels**

- **Factor A:    Low                      High**
- **Factor B    Low      Medium      High**

# A COMPREHENSIVE MODELING APPROACH

- The observation Y is affected by two factors A and B

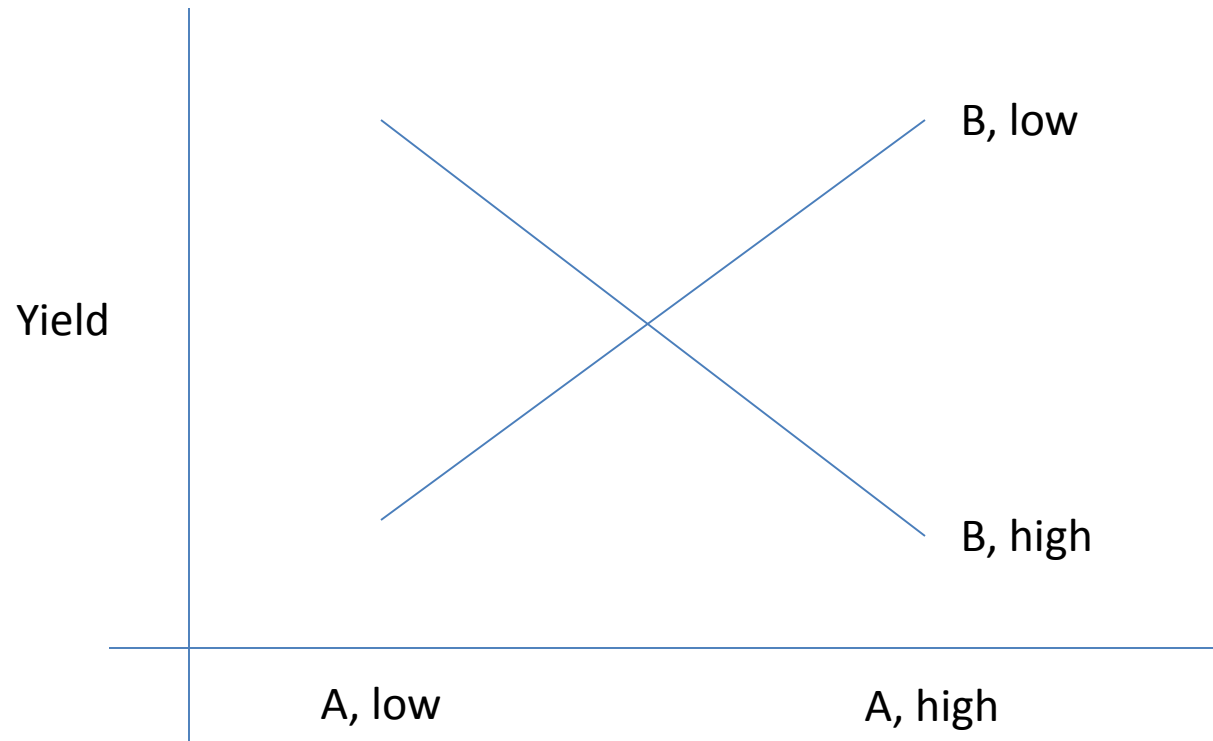
$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

- Here
  - $\alpha_i$  - effect of  $i^{th}$  level of factor A
  - $\beta_j$  - effect of  $j^{th}$  level of factor B
  - $\gamma_{ij}$  - called the **interaction** effect of A and B
  - $\epsilon_{ijk}$  - random errors
- Assumptions:
  - (1) Errors are normally distributed
  - (2)  $Var(\epsilon_{ijk}) = \text{Constant}$ .



# WHAT IS INTERACTION EFFECT?

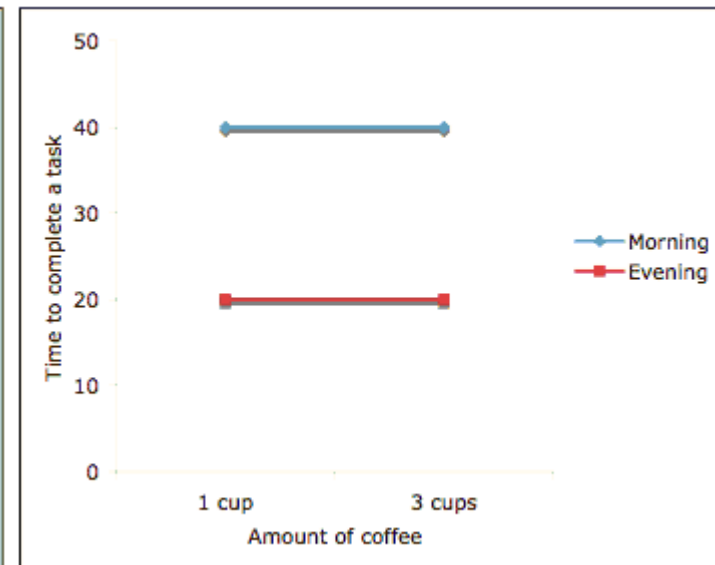
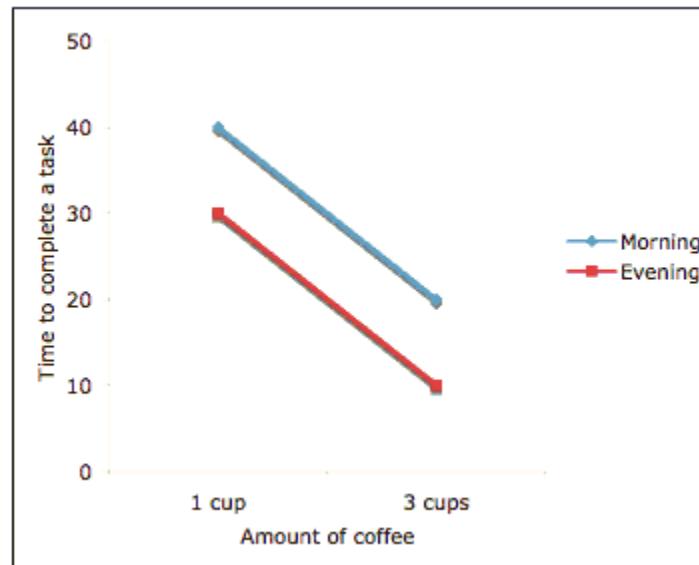
- Meaning of **Interaction** Effect



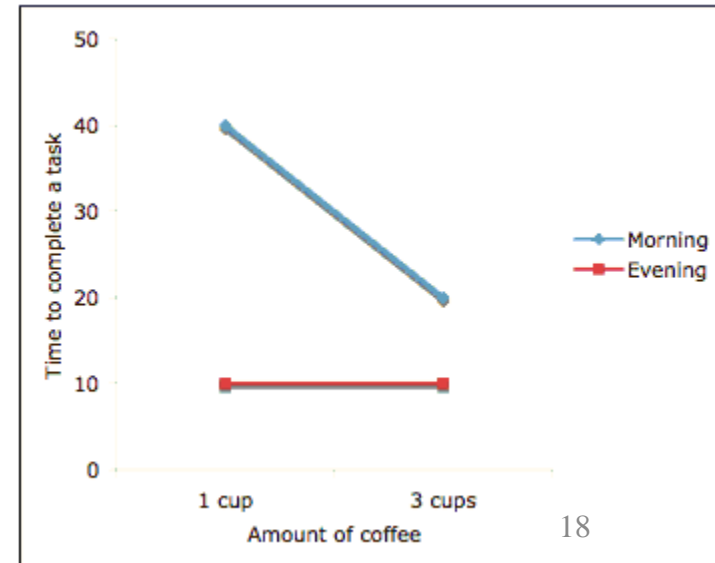
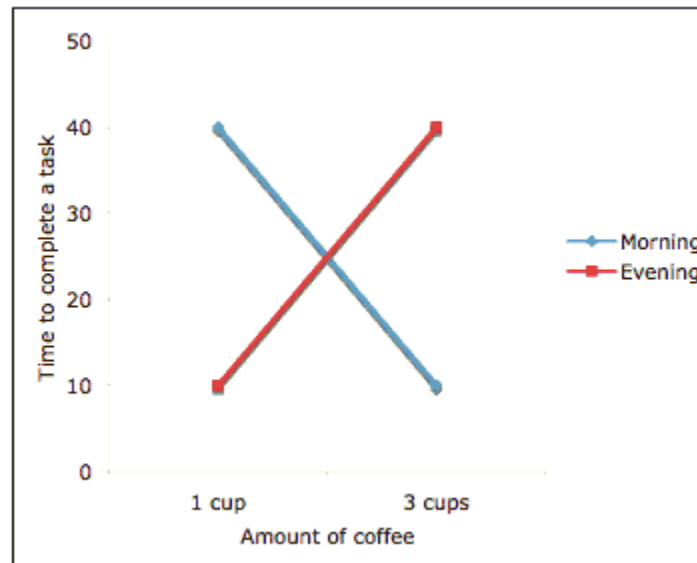


# WHAT IS INTERACTION EFFECT? CONT'D

- **No Interaction**



- **Interaction**



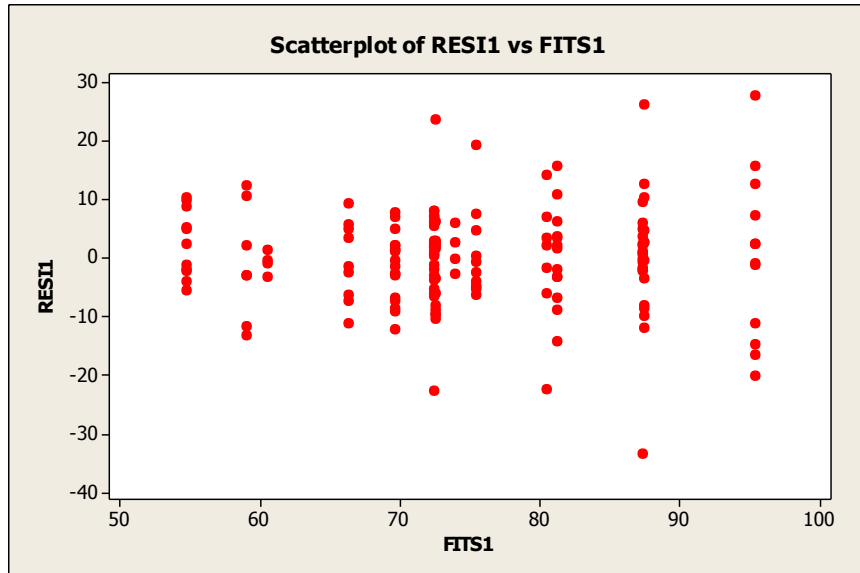
# AUSTRALIAN INSTITUTE OF SPORT EXAMPLE

- **Response Variable: Weight**
- **Factor A: Gender**
- **Factor B: Sport**
- **Two-way ANOVA: yield versus Factor A, Factor B**

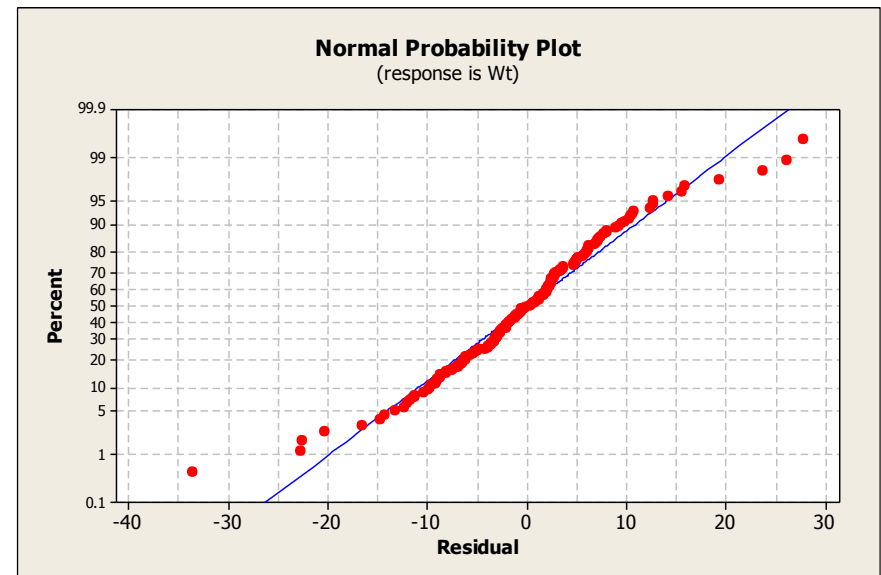
	<b>DF</b>	<b>SS</b>	<b>MS</b>	<b>F value</b>	<b>Pr (&gt;F)</b>
<b>Gender</b>	1	7424	7424	95.845	<2e-16 ***
<b>Sport</b>	6	10975	1829	23.614	<2e-16 ***
<b>Interaction</b>	6	185	31	0.398	0.879
<b>Error</b>	144	11155	77		

# CHECKING THE ASSUMPTIONS BASED ON TRANSFORMED DATA

- $Var(\epsilon_i) = \text{constant}$



- Normal distributions of the errors



# ANOVA RESULTS BASED ON THE TRANSFORMED DATA

- **Assumptions of errors seems to be satisfied**
- $H_0: \gamma_{ij} = 0$  vs.  $H_a: \gamma_{ij} \neq 0$ 
  - **TS.**  $F = 0.398$ , p-value = 0.897
  - **There is no significant interaction**
- $H_0: \alpha_i = 0$  vs.  $H_a: \alpha_i \neq 0$ 
  - **TS.**  $F = 95.845$ , p-value  $< 2 * 10^{-16}$
  - **Significant effect of Gender**
- $H_0: \beta_j = 0$  vs.  $H_a: \beta_j \neq 0$ 
  - **TS.**  $F = 23.614$ , p-value  $< 2 * 10^{-16}$
  - **Significant effect of Sport**

# INTERACTION



**Interaction Plot for Wt**  
Data Means

