Estimation of Regression Line (Best-fit line)

$$y_e = b_0 + b_1 \cdot X$$

b_1 → slope of the line

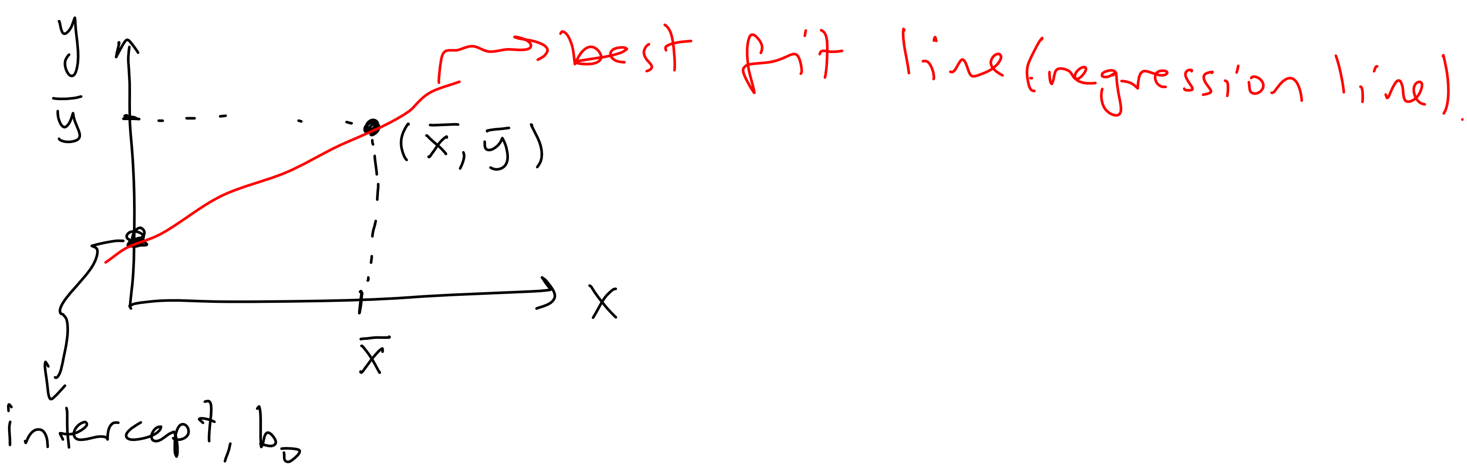
b_0 → intercept when $x = 0$

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \quad \text{or} \quad b_1 = \frac{n \cdot (\sum x \cdot y) - (\sum x)(\sum y)}{n \cdot (\sum x^2) - (\sum x)^2}$$

$\bar{x}, \bar{y}$ are the means of X and y variables.

$$b_0 = \bar{y} - b_1 \cdot \bar{X} \quad \text{or} \quad b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum x \cdot y)}{n \cdot (\sum x^2) - (\sum x)^2}$$

The regression line (best fit line) has to pass through the points ($\bar{x}, \bar{y}$) and intercept of y-axis when x = 0.



<u>**Example**</u>**: The following data were obtained from an experimental study on total colour change (TCC) of peach puree during drying at 110°C for 80 min.**

Time (min), t	0	20	40	60	80
TCC	0	2	4	6	9

a) Determine the regression equation [TCC = f(t)] of best fit line that represents the experimental data.

b) Draw the best-fit line represented by the regression equation on a hypothetical graph.

c) Find correlation coefficient.

d) Predict the TCC at 70th minutes of drying process.

e) What is the error of estimation when t = 60 min ?

<u>**Solution**</u>**:** $t \to x$, $TCC \to y$, $y_e = b_0 + b_1 \cdot x$

a) $n = 5$, $\sum t = 200$, $\sum TCC = 21$, $\bar{t} = 40$, $\overline{TCC} = 4.2$

$$\sum t^2 = 0^2 + 20^2 + 40^2 + 60^2 + 80^2 = 12000$$

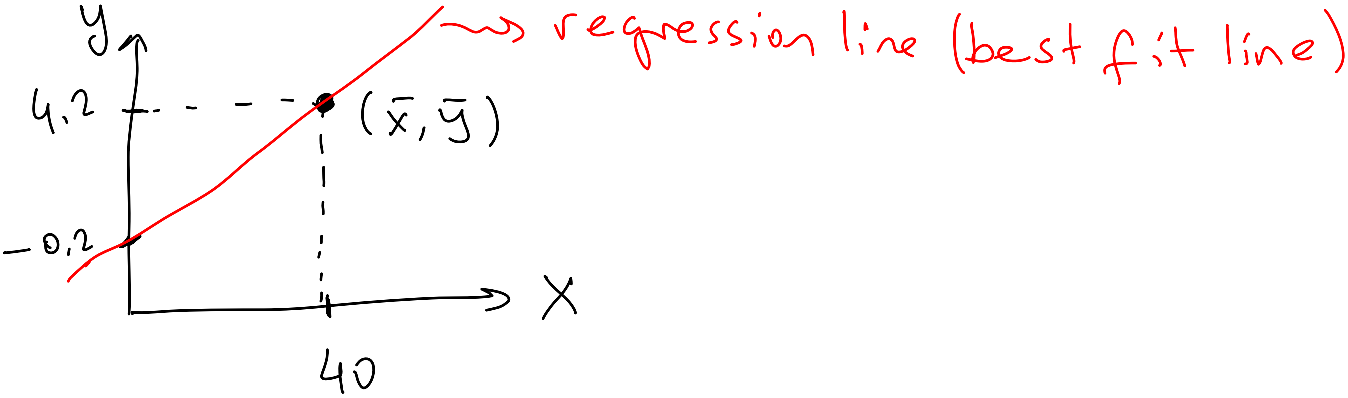
$$\sum x \cdot y = (0 \times 0) + (20 \times 2) + (40 \times 4) + (60 \times 6) + (80 \times 9) = 1280$$

$$b_1 = \frac{[\ldots]}{[\ldots]} = 0{,}11, \quad b_0 = \frac{[\ldots]}{[\ldots]} = -0{,}2$$

The regression equation =)

$$TCC = 0{,}11 \times t - 0{,}2 \quad \text{or} \quad TCC = \underbrace{-0{,}2}_{b_0} + \underbrace{0{,}11}_{b_1} \times t$$

b)



c) $r = \dfrac{(\ldots)}{(\ldots)} = +0.9959$

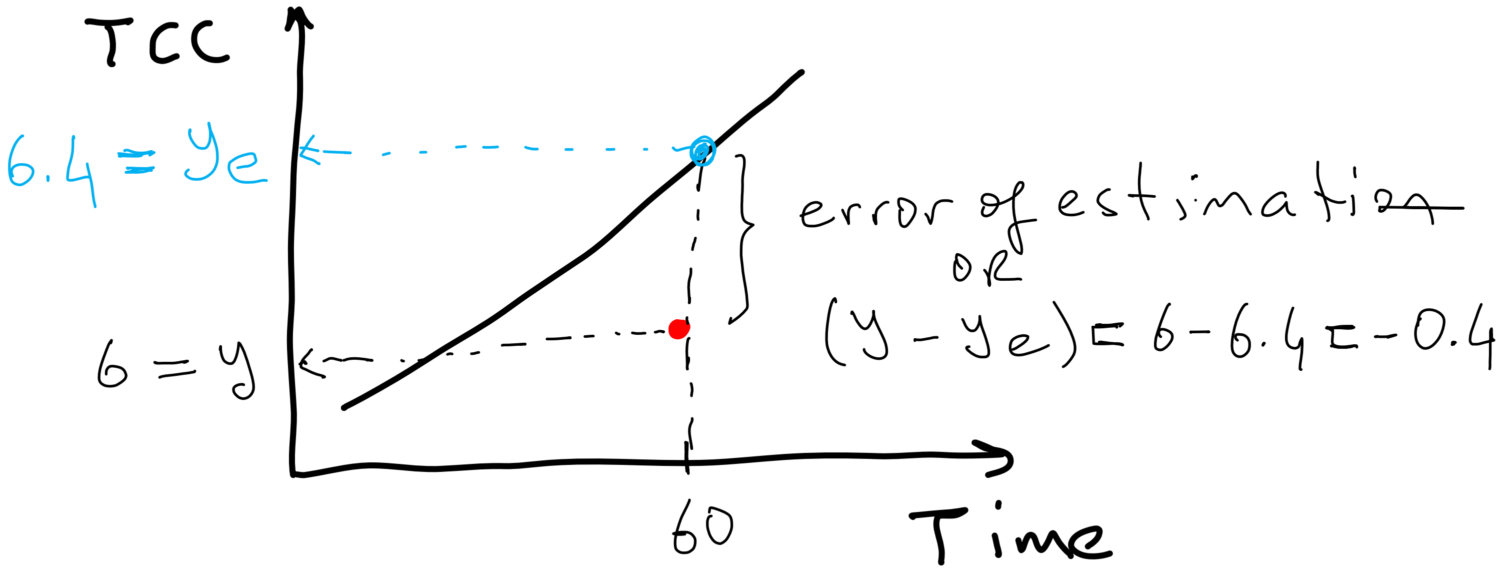
d) $TCC = ?$ when $t = 70$ min.

$TCC = 0.11 \times (70) - 0.2 = 7.5$

e) error $= (y - y_e) = ?$ when $t = 60$ min.

when $t = 60 \Rightarrow TCC = 6$, $y_e = ?$ $y_e = 0.11 \times 60 - 0.2 = 6.4$

error $= (6 - 6.4) = -0.4$.



ANALYSIS OF VARIANCE (ANOVA)

ANOVA technique is concerned with testing a hypothesis about several means (at least more than two means).

⊛ Response (R): It is the measured property.
e.g., vitamin C loss during storage of a food.

⊙ Factor (F): It is the property that is set.
e.g., temperature of the storage.

1) One-Way ANOVA

* One-way ANOVA ⇒ 1 R and 1 F.

$1R,\ 1F \rightarrow T \rightarrow T_1,\ T_2,\ T_3,\ \vdots$

e.g., color change of apple during drying at 60°C ⇒

Response: color change.

Factor: temp.

2) Two-Way ANOVA

$1R,\ 2\ \text{Factors} \rightarrow T \rightarrow T_1,\ T_2,\ T_3 \ldots$

$\rightarrow \text{Humidity} \rightarrow H_1,\ H_2,\ H_3$

3) Three-Way ANOVA

1 R, 3 Factors

- $T \rightarrow T_1, T_2, \ldots$
- $H \rightarrow H_1, H_2, \ldots$
- Pressure $\rightarrow P_1, P_2, \ldots$

Hypothesis:

$H_o: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n$ (all means are equal)

$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \neq \mu_n$ (all means are not equal).

Test statistic value: F_{calc}

Reject H_o if $F_{calc} > \underbrace{F_\alpha}_{F_{table}}$

<u>ANOVA With Equal Replications</u>

Number of observations must be the same at all factor levels.

<u>Example</u>: Color change of strawberry was measured during one week of storage at different temperatures. Does the storage temperature have significant effect on the color change of strawberry stored at various temperatures for one week ?

Test for significance at 0.05 significance level.

Factor (i): Temp. level	Replicate (j)				Row Total (T_i)	Mean ($\bar{x}$)
	1	2	3	4		
68°F	10	12	10	9	41	10.25
72°F	7	6	7	8	28	7
76°F	3	3	5	4	15	3.75

$\bar{X}_i$: the observed color change mean at level i

i : 1, 2, 3 for 68, 72, 76°F, respectively.

Solution:

✗ $H_0: \mu_{68} = \mu_{72} = \mu_{76}$ (No T effect)

✓ $H_A: \mu_{68} \neq \mu_{72} \neq \mu_{76}$ (T affects the color)

ANOVA Table

Source of variation	Sum of squares (SS)	d.f.	Mean Squares (MS)	F_{calc}
Among groups (A): (Factor)	SS_A	$d.f._A$	$MS_A = \dfrac{SS_A}{d.f._A}$	$\dfrac{MS_A}{MS_W}$
Within groups (W): (Error)	SS_W	$d.f._W$	$MS_W = \dfrac{SS_W}{d.f._W}$	
Total	SS_T	$d.f._T$		

Compare F_{calc} with F_{table}.

$$F_{table} = F(\underbrace{d.f._{MS_A}}_{\nu_1}, \underbrace{d.f._{MS_W}}_{\nu_2})$$

Sum of squares $= SS = \sum (x - \bar{x})^2$

$$SS_{Total} = \sum (x^2) - \frac{(\sum x)^2}{n}$$

$$\sum (x^2) = 10^2 + 12^2 + 10^2 + \dots + 5^2 + 4^2 = 682$$

$$\sum x = 10 + 12 + 10 + \dots + 5 + 4 = 84$$

$$SS_{Total} = 682 - \frac{(84)^2}{12} = 94$$

$$SS_{Total} = \underbrace{SS_{factor}}_{\text{Temp}} + \underbrace{SS_{error}}_{\text{within groups}}$$

$$SS_{Total} = 94 = SS_{Temp} + SS_{error}$$

$$SS_{factor} = \frac{\sum (T_i)^2}{c} - \frac{(\sum x)^2}{n}$$ measures the variation between the rows.

T_i: row totals

c: # of replicates for each levels $= 4$

n: # of data for total sample $= 3 \times 4 = 12$

$$\sum (T_i^2) = 41^2 + 28^2 + 15^2 = 2690$$

$$SS_{factor} = \frac{2690}{4} - \frac{(84)^2}{12} = 84.5$$

$$SS_{error} = \sum(x^2) - \frac{\sum(T_i^2)}{c}$$ ⇝ measures variation within the rows.

$$SS_{error} = 682 - \frac{2690}{4} = 9.50$$

$$d.f._{factor} = \underbrace{r}_{\text{\# of factor level}} - 1 = 3 - 1 = 2$$

$$d.f._{error} = r(c-1) = 3(4-1) = 9$$

$$d.f._{total} = n - 1 = 12 - 1 = 11$$

check SS and d.f. values ⇒

$$SS_{Total} = 94 \stackrel{?}{=} 84.5 + 9.5 = 94 \Rightarrow ok$$

$$d.f._{total} = 11 \stackrel{?}{=} df_{factor} + d.f._{error}$$

$$11 \stackrel{?}{=} 2 + 9 = 11 \Rightarrow ok.$$

$$MS_{factor} = \frac{SS_{factor}}{d.f._{factor}} = \frac{84.5}{2} = 42.25$$

$$MS_{error} = \frac{SS_{error}}{d.f._{error}} = \frac{9.5}{9} = 1.056$$

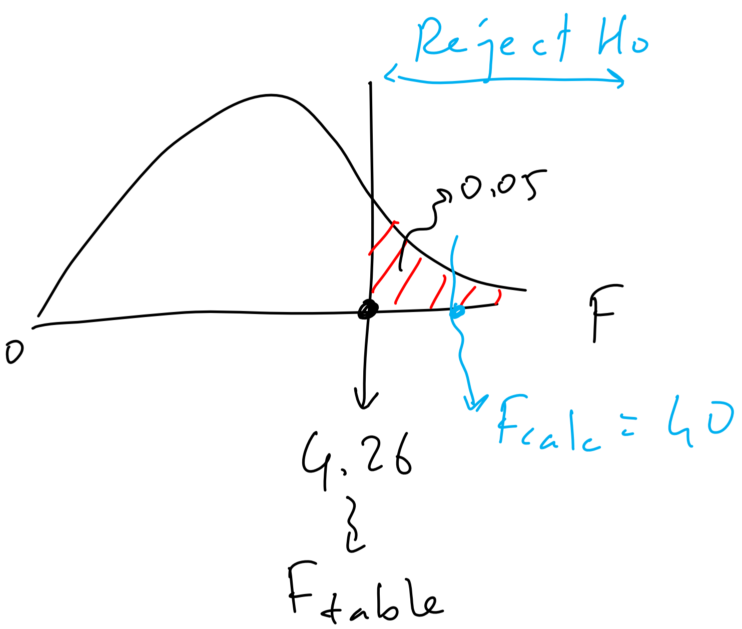
Test statistic value $= F_{calc} = \dfrac{MS_{factor}}{MS_{error}}$

$$F_{calc} = \frac{42.25}{1.056} = 40$$

ANOVA Table

Source of variation	Sum of squares (SS)	d.f.	Mean Squares (MS)	F_{calc}
Factor (Temp.)	84.5	2	42.25	40
Error	9.5	9	1.056	
Total	94	11		

$F_{table} = ?$ $\quad F_{0.05}(2,9) = 4.26$



Decision: Reject Ho

Conclusion: There is significant effect of temp. on the color change of strawberry during storage.

Example: The scores of shooting at a target by different sighting methods (right eye open, left eye open, both eyes are open) are shown in the table below. Test if there is no advantage in using one sighting method over the others. Use 0.05 significance level.

Sighting method	Replicate					Total
	1	2	3	4	5	
Right eye open	2	0	8	2	4	16
Left eye open	0	7	6	3	10	26
Both eyes open	6	4	6	1	11	28
Total	8	11	20	6	25	70

Solution:

✓ $H_0: \mu_R = \mu_L = \mu_B$ (no difference in methods)

X $H_A: \mu_R \neq \mu_L \neq \mu_B$ (there is difference).

$$d.f._{factor} = r-1 = 3-1 = 2$$

$$d.f._{error} = r(c-1) = 3(5-1) = 12$$

$$d.f._{total} = n-1 = 3\times 5 - 1 = 14$$

$$SS_{factor} = \frac{16^2+26^2+28^2}{5} - \frac{(70)^2}{15} = 16.53$$

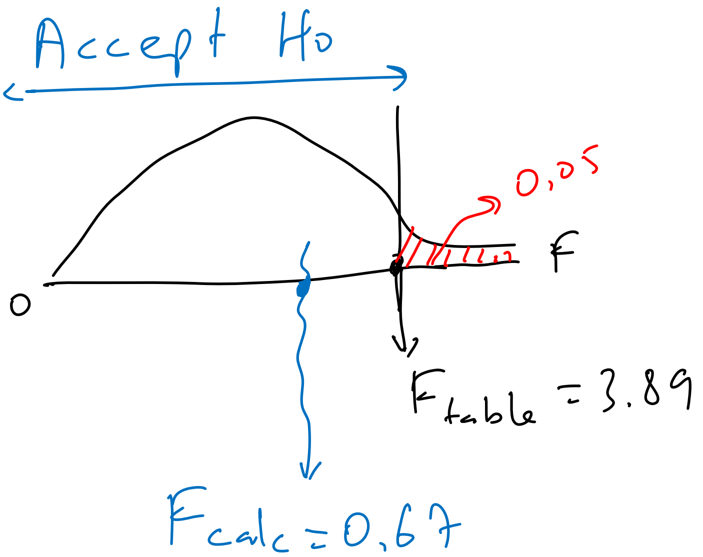
$$SS_{error} = (2^2+0^2+8^2+\dots+1^2+11^2) - \frac{16^2+26^2+28^2}{5} = 148.8$$

$$SS_{Total} = (2^2+0^2+8^2+\dots+1^2+11^2) - \frac{(70)^2}{15} = 165.33$$

ANOVA Table

Source of variation	Sum of squares (SS)	d.f.	Mean Squares (MS)	F_{calc}
Factor (Method)	16.53	2	$\frac{16.53}{2} = 8.265$	$\frac{8.265}{12.4} = 0.67$
Error	148.8	12	$\frac{148.8}{12} = 12.4$	
Total	165.33	14		

$F_{table} = F_{0,05}(2,12) = 3.89$



Decision: Do not reject Ho

Conclusion: There is no advantage of using any one of the sighting methods over the others.

ANOVA With Unequal Replications

In experimental work one often loses some of the desired observations. For example, an experiment might be conducted to determine if college students obtain different grades on the average for classes meeting at different semesters. It is entirely possible to conclude the experiment with unequal numbers of students in the different semesters.

A slight modification of sum of squares formulas is needed.

Group (Factor)	Number of replicates						Total
	1	2	3	.	.	n	T_i
1	X_{11}	X_{12}	X_{13}	.	.	X_{1n}	T_1
2	X_{21}	X_{22}	X_{23}	.	.	X_{2n}	T_2
3	X_{31}	X_{32}	X_{33}	.	.	X_{3n}	T_3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
r	X_{r1}	X_{r2}	X_{r3}	.	.	X_{rn}	T_r

$$\sum_{i=1}^{r} T_i$$

$$\sum_{i=1}^{r} T_i = T_1 + T_2 + T_3 + \cdots + T_r$$

N = total # of observations

$N = r \times n$ for equal replicates (sample size).

$$\text{But, } \underbrace{N = \underbrace{n_{group\,1}}_{\text{level }1} + \underbrace{n_{group\,2}}_{\text{level }2} + \cdots + \underbrace{n_{group\,r}}_{\text{level }r}}_{\text{for unequal replicates or sample size.}}$$

$$SS_{Total} = \sum_{i=1}^{N} (x_i)^2 - \frac{\left[\sum(x_i)\right]^2}{N}$$

$$SS_{factor} = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{(\sum x_i)^2}{N}$$

$$SS_{error} = \sum_{i=1}^{N} (x_i)^2 - \sum_{i=1}^{r} \frac{T_i^2}{n_i}$$

OR

$$SS_{error} = SS_{Total} - SS_{factor}$$

$$d.f._{Total} = N - 1$$

$$d.f._{factor} = \underbrace{r}_{\text{\# of factor levels}} - 1$$

$$d.f._{error} = (N-1) - (r-1) = N - r$$

<u>Example</u>: It is suspected that higher-priced automobiles are assembled with greater care than lower-prised automobiles. To investigate whether there is any basis for this feeling, a large luxury model A, a medium size sedan B and a subcompact hatchback C were compared for defects when they arrived at the dealer's showroom. All cars were manufactured by the same company. The numbers of defects for several of the three models are recorded in the following table:

Car Model	Number of defects (n_i)						Total T_i
	1	2	3	4	5	6	
A	4	7	6	6			23
B	5	1	3	5	3	4	21
C	8	6	8	9	5		36

$$\sum_{i=1}^{r} T_i = 23 + 21 + 36 = 80$$

Test the hypothesis at the 0.05 level of significance that the average number of defects is the same for the three models.

<u>Solution</u>: $n_A = 4,\ n_B = 6,\ n_C = 5 \Rightarrow N = 4+6+5 = 15$

$r = 3\ ,\ \alpha = 0{,}05$

x $H_0 : \mu_A = \mu_B = \mu_C$

✓ $H_A : \mu_A \neq \mu_B \neq \mu_C$

$$SS_{Total} = \sum_{i=1}^{N} (x_i)^2 - \frac{(\sum x_i)^2}{N}$$

$$\sum (x_i)^2 = 4^2 + 7^2 + 6^2 + \cdots + 8^2 + 9^2 + 5^2 = 492$$

$$\sum (x_i) = 4 + 7 + 6 + \cdots + 8 + 9 + 5 = 80 = \sum T_i$$

$$SS_{Total} = 492 - \frac{(80)^2}{15} = 65.333$$

$$SS_{factor} = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{(\sum x_i)^2}{N}$$

$$SS_{factor} = \frac{(23)^2}{4} + \frac{(21)^2}{6} + \frac{(36)^2}{5} - \frac{(80)^2}{15} = 38.283$$

$$SS_{error} = \sum_{i=1}^{N} (x_i)^2 - \sum_{i=1}^{r} \frac{T_i^2}{n_i}$$

$$SS_{error} = 492 - \left(\frac{23^2}{4} + \frac{21^2}{6} + \frac{36^2}{5} \right) = 27.05$$

OR

$$SS_{error} = SS_{Total} - SS_{factor}$$

$$= 65.333 - 38.283 = 27.05$$

$$d.f._{Total} = N - 1 = 15 - 1 = 14$$

$$d.f._{factor} = r - 1 = 3 - 1 = 2$$

$$d.f._{error} = N - r = 15 - 3 = 12$$

$$MS_{factor} = \frac{SS_{factor}}{d.f._{factor}} = \frac{38.283}{2} = 19.142$$

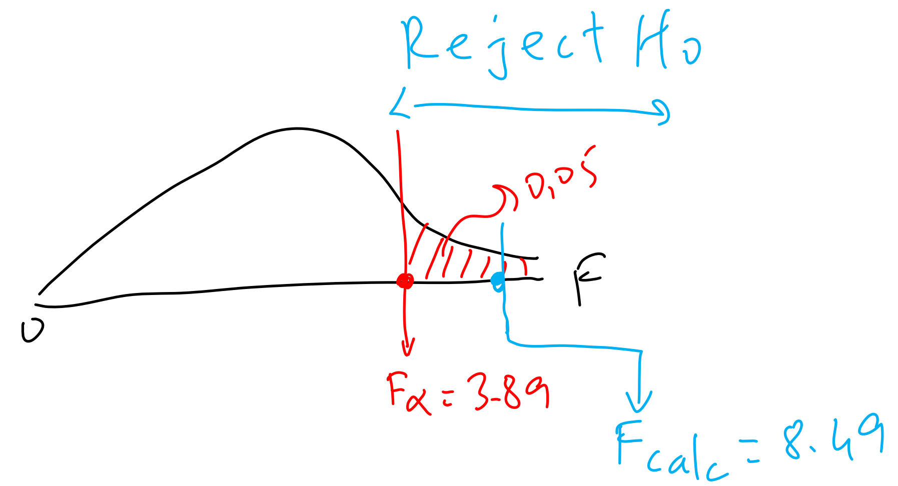
$$MS_{error} = \frac{SS_{error}}{d.f._{error}} = \frac{27.05}{12} = 2.254$$

$$F_{calc} = \frac{MS_{factor}}{MS_{error}} = \frac{19.142}{2.254} = 8.49$$

ANOVA Table

Source of variation	Sum of squares (SS)	d.f.	Mean Squares (MS)	F_{calc}
Factor (A): Car Model	38.283	2	19.142	8.49
Error (W)	27.05	12	2.254	
Total	65.333	14		

$$F_{table} = F_{0.05}(2, 12) = 3.89$$



Decision: Reject Ho

Conclusion: The average # of defects for the three models is not the same.