STAT 420 Fall 2015

Examples for 11/10/15

ANOVA (Analysis of Variance)

Population 1	Population 2		Population J
mean μ_1	mean μ_2		mean μ_J
std. dev. σ	std. dev. σ	• • •	std. dev. σ
$\Downarrow$	$\Downarrow$		$\Downarrow$
$y_{11}, y_{21}, \cdots, y_{n_1 1}$	$y_{12}, y_{22}, \cdots, y_{n_2 2}$		$y_{1J}, y_{2J}, \cdots, y_{n_J J}$
$\bar{y}_1,\ s_1^2$	$\bar{y}_2,\ s_2^2$		$\bar{y}_J,\ s_J^2$

H_0: $\mu_1 = \mu_2 = \ldots = \mu_J$ H_1: not all of the μ_j are equal.

(We want to test *simultaneously* for differences among the means of all J populations.)

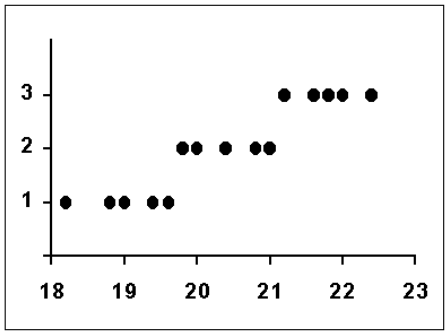
Example: There is a great deal of interest in comparing the mileage rating of different makes of automobiles. Other things being equal, customers buy car that gets the best mileage. Suppose we wish to compare the mean fuel consumption for $J = 3$ different makes of automobile, Car 1, Car 2, and Car 3. Suppose 5 cars of each make are selected randomly and the gasoline mileage (in miles per gallon) is recorded for each car. Suppose that

$\bar{y}_1 = 19.0,$ $\bar{y}_2 = 20.4,$ $\bar{y}_3 = 21.8.$

Is there enough evidence that the mean fuel consumption is not the same for Car 1, Car 2 and Car 3?

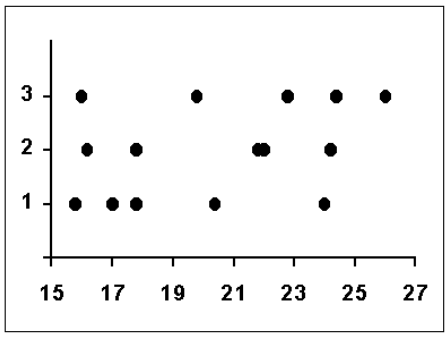
Data Set A:

	Car 1	Car 2	Car 3
	18.2	19.8	21.2
	19.4	21.0	21.8
	19.6	20.0	22.4
	19.0	20.8	22.0
	18.8	20.4	21.6
Sample mean	19.0	20.4	21.8
Sample std. dev.	0.5477	0.5099	0.4472
Sample variance	0.300	0.260	0.200



Data Set B:

	Car 1	Car 2	Car 3
	17.0	24.2	26.0
	20.4	22.0	19.8
	24.0	17.8	24.4
	15.8	16.2	16.0
	17.8	21.8	22.8
Sample mean	19.0	20.4	21.8
Sample std. dev.	3.26	3.29	3.97
Sample variance	10.66	10.84	15.76



The Analysis of Variance Idea:

Analysis of Variance (ANOVA) compares the variation due to specific sources (between groups) with the variation among individuals who should be similar (within groups). In particular, ANOVA tests whether several populations have the same mean by comparing how far apart the sample means are with how much variation there is within the samples.

ANOVA Assumptions:

- We have ***J* independent simple random samples**, one from each of *J* populations.

- The jth population has a **normal distribution** with unknown mean μ_j. The means may be different in the different populations. The ANOVA F test statistic tests the null hypothesis that all of the populations have the same mean:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_J$$

$$H_1: \text{ not all of the } \mu_j \text{ are equal.}$$

- All of the populations have the **same standard deviation σ**, whose value is unknown.

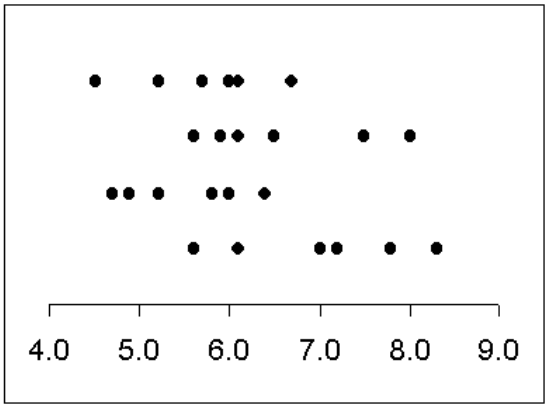
  The results of ANOVA F test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

ANOVA table:

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Test Statistic*
Source	**SS**	**DF**	**MS**	**F**
Between	SSB	$J-1$	$\frac{\text{SSB}}{J-1}$	$\frac{\text{MSB}}{\text{MSW}}$
Within	SSW	$N-J$	$\frac{\text{SSW}}{N-J}$	
Total	SSTot	$N-1$		

1. Six samples of each of four types of cereal grain grown in a certain region were analyzed to determine thiamin content, resulting in the following data (μg/g):

Wheat	5.2	4.5	6.0	6.1	6.7	5.7
Barley	6.5	8.0	6.1	7.5	5.9	5.6
Maize	5.8	4.7	6.4	4.9	6.0	5.2
Oats	8.3	6.1	7.8	7.0	5.6	7.2



Does this data suggest that at least two of the grains differ with respect to true average thiamin content? Use a level $\alpha = 0.05$ test.

	n_j	$\bar{y}_j$	s_j	s_j^2
Wheat	6	5.7	0.7668	0.588
Barley	6	6.6	0.9508	0.904
Maize	6	5.5	0.6693	0.448
Oats	6	7.0	1.0139	1.028

$$y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$$

$$\begin{bmatrix} 5.2 & 4.5 & 6.0 & 6.1 & 6.7 & 5.7 \\ 6.5 & 8.0 & 6.1 & 7.5 & 5.9 & 5.6 \\ 5.8 & 4.7 & 6.4 & 4.9 & 6.0 & 5.2 \\ 8.3 & 6.1 & 7.8 & 7.0 & 5.6 & 7.2 \end{bmatrix}$$

$$= \begin{bmatrix} 6.2 & 6.2 & 6.2 & 6.2 & 6.2 & 6.2 \\ 6.2 & 6.2 & 6.2 & 6.2 & 6.2 & 6.2 \\ 6.2 & 6.2 & 6.2 & 6.2 & 6.2 & 6.2 \\ 6.2 & 6.2 & 6.2 & 6.2 & 6.2 & 6.2 \end{bmatrix}$$

$$+ \begin{bmatrix} -0.5 & -0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\ -0.7 & -0.7 & -0.7 & -0.7 & -0.7 & -0.7 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \end{bmatrix}$$

$$+ \begin{bmatrix} -0.5 & -1.2 & 0.3 & 0.4 & 1.0 & 0.0 \\ -0.1 & 1.4 & -0.5 & 0.9 & -0.7 & -1.0 \\ 0.3 & -0.8 & 0.9 & -0.6 & 0.5 & -0.3 \\ 1.3 & -0.9 & 0.8 & 0.0 & -1.4 & 0.2 \end{bmatrix}$$

$$(y_{ij} - \bar{y}) \quad = \quad (\bar{y}_j - \bar{y}) \quad + \quad (y_{ij} - \bar{y}_j)$$

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}\right)^2 \quad = \quad \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{y}_j - \bar{y}\right)^2 \quad + \quad \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2$$

SS Total	SS Between	SS Within
	$\sum_{j=1}^{J} n_j\left(\bar{y}_j - \bar{y}\right)^2$	$\sum_{j=1}^{J}\left(n_j - 1\right)s_j^2$
$N - 1$ degrees of freedom	$J - 1$ degrees of freedom	$N - J$ degrees of freedom

a) How many different groups (treatments) are there?

b) What was the total sample size, N?

c) Compute the overall average $\bar{y}$.

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + ... + n_J \cdot \bar{y}_J}{N}$$

d) Compute SSB.

$$\text{SSB} = n_1 \cdot \left(\bar{y}_1 - \bar{y}\right)^2 + n_2 \cdot \left(\bar{y}_2 - \bar{y}\right)^2 + ... + n_J \cdot \left(\bar{y}_J - \bar{y}\right)^2$$

e) Find the number of degrees of freedom that is associated with SSB.

f) Compute MSB.

$$\text{MSB} = \frac{\text{SSB}}{J-1}$$

g) Compute SSW.

$$\text{SSW} = \left(n_1 - 1\right) \cdot s_1^2 + \left(n_2 - 1\right) \cdot s_2^2 + \ldots + \left(n_J - 1\right) \cdot s_J^2$$

h) Find the number of degrees of freedom that is associated with SSW.

i) Compute MSW.

$$\text{MSW} = \frac{\text{SSW}}{N - J}$$

j) Compute SSTot.

$$\text{SSTot} = \text{SSB} + \text{SSW}$$

k) Find the number of degrees of freedom that is associated with SSTot.

l) Compute the value of the test statistic F.

$$\text{F} = \frac{\text{MSB}}{\text{MSW}}$$

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. H_1: not all of the μ_j are equal.

Reject H_0 if $F > F_\alpha(J-1, N-J)$, where

$F_\alpha(J-1, N-J)$ is the critical value of F distribution

with probability α lying to its right,

degrees of freedom in the numerator = $J-1$.

degrees of freedom in the denominator = $N-J$.

m) What is the critical value, $F_\alpha(J-1, N-J)$, at $\alpha = 0.05$?

$\alpha = 0.05$, $J-1 = 3$, $N-J = 20$.

$F_\alpha(J-1, N-J) = F_{0.05}(3, 20) =$ _____.

n) What is your conclusion regarding the null hypothesis at $\alpha = 0.05$?

o) What is the critical value, $F_\alpha(J-1, N-J)$, at $\alpha = 0.01$?

$\alpha = 0.01$, $J-1 = 3$, $N-J = 20$.

$F_\alpha(J-1, N-J) = F_{0.01}(3, 20) =$ _____.

p) What is your conclusion regarding the null hypothesis at $\alpha = 0.01$?

$J = \mathbf{4}$. $\qquad N = n_1 + n_2 + \ldots + n_J = 6 + 6 + 6 + 6 = \mathbf{24}$.

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + \ldots + n_J \cdot \bar{y}_J}{N} = \frac{6 \cdot 5.7 + 6 \cdot 6.6 + 6 \cdot 5.5 + 6 \cdot 7.0}{24} = \mathbf{6.2}.$$

$$\text{SSB} = n_1 \cdot \left(\bar{y}_1 - \bar{y}\right)^2 + n_2 \cdot \left(\bar{y}_2 - \bar{y}\right)^2 + \ldots + n_J \cdot \left(\bar{y}_J - \bar{y}\right)^2$$

$$= 6 \cdot (5.7 - 6.2)^2 + 6 \cdot (6.6 - 6.2)^2 + 6 \cdot (5.5 - 6.2)^2 + 6 \cdot (7.0 - 6.2)^2 = \mathbf{9.24}.$$

$$\text{MSB} = \frac{\text{SSB}}{J-1} = \frac{9.24}{3} = \mathbf{3.08}.$$

$$\text{SSW} = \left(n_1 - 1\right) \cdot s_1^2 + \left(n_2 - 1\right) \cdot s_2^2 + \ldots + \left(n_J - 1\right) \cdot s_J^2$$

$$= 5 \cdot 0.588 + 5 \cdot 0.904 + 5 \cdot 0.448 + 5 \cdot 1.028 = \mathbf{14.84}.$$

$$\text{MSW} = \frac{\text{SSW}}{N-J} = \frac{14.84}{20} = \mathbf{0.742}.$$

SSTot = SSB + SSW = **24.08**.

$$\text{F} = \frac{\text{MSB}}{\text{MSW}} = \frac{3.08}{0.742} = \mathbf{4.151}.$$

ANOVA table:

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Test Statistic*
Source	**SS**	**DF**	**MS**	**F**
Between	9.24	3	3.08	4.151
Within	14.84	20	0.742	
Total	24.08	23		

$F_{0.05}(3, 20) = 3.10$. Reject H_0 at $\alpha = 0.05$.

$F_{0.01}(3, 20) = 4.94$. Do NOT Reject H_0 at $\alpha = 0.01$.

Practice Problem

Construct the ANOVA tables and perform the ANOVA F test at $\alpha = 0.05$ for Data Set A and Data Set B.

Answers:

Data Set A:

Source	SS	DF	MS	F
Between	19.60	2	9.8	38.68421
Within	3.04	12	0.253333	
Total	22.64	14		

$F_{0.05}(2, 12) = 3.88$. **Reject H_0.**

Data Set B:

Source	SS	DF	MS	F
Between	19.60	2	9.8	0.78905
Within	149.04	12	12.42	
Total	168.64	14		

$F_{0.05}(2, 12) = 3.88$. **Do NOT Reject H_0.**