BITS Pilani
Pilani Campus

ANOVA

Akanksha Bharadwaj
Asst. Professor, CS/IS Department

Need for ANOVA

- In the machine operator example, **is it possible to analyze the four samples by using a *t* test** for the difference in two sample means?
- These four samples would require ${}^{4}C_{2} = 6$ individual *t* tests to accomplish the analysis of two groups at a time.
- Recall that if α = .05 for a particular test, there is a 5% chance of rejecting a null hypothesis that is true (i.e., committing a Type I error).
- If enough tests are done, eventually one or more null hypotheses will be falsely rejected by chance.
- Hence, α = .05 is valid only for one *t* test. In this problem, with six *t* tests, the error rate compounds, so when the analyst is finished with the problem there is a much greater than .05 chance of committing a Type I error.

Analysis of Variance

- When there are more than two groups to be compared, it is not correct to compare the groups in pairs, as this type of comparison will not take the within variability into consideration

- The Analysis procedure used in such comparisons is known as ANALYSIS OF VARIANCE

Example

- As an example of a completely randomized design, suppose a researcher decides to analyze the effects of the machine operator on the valve opening measurements of valves produced in a manufacturing plant, like those shown in Table below.

6.26	6.19	6.33	6.26	6.50
6.19	6.44	6.22	6.54	6.23
6.29	6.40	6.23	6.29	6.58
6.27	6.38	6.58	6.31	6.34
6.21	6.19	6.36	6.56	

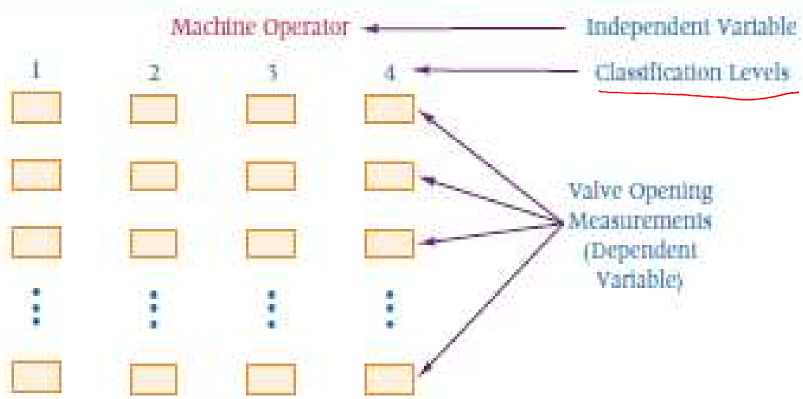
$\bar{x} = 6.34$ Total Sum of Squares Deviation $=$ SST $= \sum(x_i - \bar{x})^2 = .3915$

- The independent variable in this design is machine operator.

Example continued

independent variable

- Suppose further that four different operators operate the machines. These four machine operators are the levels of treatment, or classification, of the independent variable.
- The dependent variable is the opening measurement of the valve.
- Figure below shows the structure of this completely randomized design.
- Table below contains the valve opening measurements for valves produced under each operator.



Valve Openings by Operator

Machine operators

1	2	3	4
6.33	6.26	6.44	6.29
6.26	6.36	6.38	6.23
6.31	6.23	6.58	6.19
6.29	6.27	6.54	6.21
6.40	6.19	6.56	
	6.50	6.34	
	6.19	6.58	
	6.22		

One Way ANOVA

ANOVA → Testing equality of k group means against not equal

ANOVA

- Samples are drawn from normal population
- The population variances should be equal
- The sample size should be less than 30 (i.e., n < 30)
- Groups should be independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e., n > 30) ANOVA should be continue to apply, because of central limit theorem it approaches normal.

Hypothesis in ANOVA

- In general, if *k* samples are being analyzed, the following hypotheses are being tested in a one-way ANOVA.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

H_a: At least one of the means is different from the others.

- The null hypothesis states that the population means for all treatment levels are equal.

- Because of the way the alternative hypothesis is stated, if even one of the population means is different from the others, the null hypothesis is rejected.

Testing hypotheses

Testing these hypotheses by using one-way ANOVA is accomplished by partitioning the total variance of the data into the following two variances.

MO1 MO2 MO3 MO4

1. The variance resulting from the treatment (columns)

2. The error variance, or that portion of the total variance unexplained by the treatment

Total Sum of Squares of Variation

The error variation can be viewed at this point as variation due to individual differences within treatment groups.

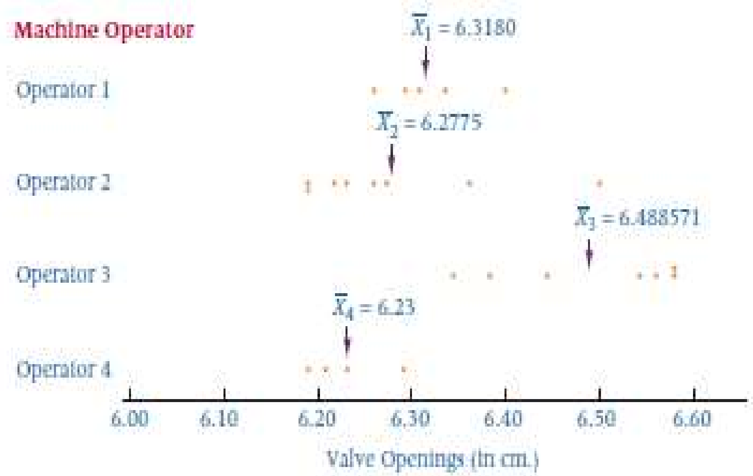
$$\underset{\sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij}-\bar{x})^2}{\text{SST}} = \underset{\sum_{j=1}^{C} n_j(\bar{x}_j-\bar{x})^2}{\text{SSC}} + \underset{\sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)^2}{\text{SSE}}$$

where

SST = total sum of squares
SSC = sum of squares column (treatment)
SSE = sum of squares error
i = particular member of a treatment level
j = a treatment level
C = number of treatment levels
n_j = number of observations in a given treatment level
$\bar{x}$ = grand mean
$\bar{x}_j$ = mean of a treatment group or level
x_{ij} = individual value

Example

- Figure below displays the data from the machine operator example in terms of treatment level.
- Note the variation of values (x) *within* each treatment level. Now examine the variation between levels 1 through 4 (the difference in the machine operators).



Assumptions

- Analysis of variance is used to determine statistically whether the variance between the treatment level means is greater than the variances within levels (error variance).
- Several important assumptions underlie analysis of variance:

1. Observations are drawn from normally distributed populations.
2. Observations represent random samples from the populations.
3. Variances of the populations are equal.

These assumptions are similar to those for using the *t* test for independent samples

Formula

FORMULAS FOR COMPUTING A ONE-WAY ANOVA

$$SSC = \sum_{j=1}^{C} n_j(\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)^2$$

$$SST = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x})^2$$

$$df_C = C - 1$$

$$df_E = N - C$$

$$df_T = N - 1$$

$$MSC = \frac{SSC}{df_C}$$

$$MSE = \frac{SSE}{df_E}$$

$$F = \frac{MSC}{MSE}$$

where

i = a particular member of a treatment level
j = a treatment level
C = number of treatment levels
n_j = number of observations in a given treatment level
$\bar{x}$ = grand mean
$\bar{x}_j$ = column mean
x_{ij} = individual value

- SST is the total sum of squares and is a measure of all variation in the dependent variable.
- As shown previously, SST contains both SSC and SSE and can be partitioned into SSC and SSE.
- MSC, MSE, and MST are the mean squares of column, error, and total respectively.
- Mean square is an average and is computed by dividing the sum of squares by the degrees of freedom.
- Finally, the *F* value is determined by dividing the treatment variance (MSC) by the error variance (MSE).
- As discussed earlier, the *F* is a ratio of two variances.
- In the ANOVA situation, the ***F* value** is *a ratio of the treatment variance to the error variance*.

Machine operator Example

innovate achieve lead

Machine Operator			
1	2	3	4
6.33	6.26	6.44	6.29
6.26	6.36	6.38	6.23
6.31	6.23	6.58	6.19
6.29	6.27	6.54	6.21
6.40	6.19	6.56	
	6.50	6.34	
	6.19	6.58	
	6.22		

Treatment levels

$$SSC = \sum_{j=1}^{n} n_j (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{c} (x_{ij} - \bar{x}_j)^2$$

$$SST = \sum_{i=1}^{n} \sum_{j=1}^{c} (x_{ij} - \bar{x})^2$$

T_j:	$T_1 = 31.59$	$T_2 = 50.22$	$T_3 = 45.42$	$T_4 = 24.92$	$T = 152.15$
n_j:	$n_1 = 5$	$n_2 = 8$	$n_3 = 7$	$n_4 = 4$	$N = 24$
$\bar{x}_j$:	$\bar{x}_1 = 6.318$	$\bar{x}_2 = 6.2775$	$\bar{x}_3 = 6.488571$	$\bar{x}_4 = 6.230$	$\bar{x} = 6.339583$

$$SSC = \sum_{j=1}^{C} n_j(\bar{x}_j - \bar{x})^2 = [5(6.318 - 6.339583)^2 + 8(6.2775 - 6.339583)^2 + 7(6.488571 - 6.339583)^2 + 4(6.230 - 6.339583)^2]$$

$$= 0.00233 + 0.03083 + 0.15538 + 0.04803$$

$$= \underline{0.23658}$$

$$SSE = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)^2 = [(6.33 - 6.318)^2 + (6.26 - 6.318)^2 + (6.31 - 6.318)^2 + (6.29 - 6.318)^2 + (6.40 - 6.318)^2 + (6.26 - 6.2775)^2 + (6.36 - 6.2775)^2 + \ldots + (6.19 - 6.230)^2 + (6.21 - 6.230)^2$$

$$= \underline{0.15492}$$

$df_C = C - 1 = 4 - 1 = 3$

$df_E = N - C = 24 - 4 = 20$

$df_T = N - 1 = 24 - 1 = 23$

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij} - \bar{x})^2 = [(6.33 - 6.339583)^2 + (6.26 - 6.339583)^2 + (6.31 - 6.339583)^2 + \ldots + (6.19 - 6.339583)^2 + (6.21 - 6.339583)^2 = 0.39150$$

$$df_C = C - 1 = 4 - 1 = 3$$

$$df_E = N - C = 24 - 4 = 20$$

$$df_T = N - 1 = 24 - 1 = 23$$

$$MSC = \frac{SSC}{df_C} = \frac{.23658}{3} = .078860$$

$$MSE = \frac{SSE}{df_E} = \frac{.15492}{20} = .007746$$

$$\boxed{F} = \frac{.078860}{.007746} = 10.18$$

From these computations, an analysis of variance chart can be constructed

Source of Variance	df	SS	MS	F
Between	3	0.23658	0.078860	10.18
Error	20	0.15492	0.007746	
Total	23	0.39150		

$\alpha = 0.05$

- Associated with every *F* value in the table are two unique df values: degrees of freedom in the numerator (df_C) and degrees of freedom in the denominator (df_E).
- For the machine operator example, df_C = 3 and df_E = 20, $F_{.05,3,20}$ is **3.10**. This value is the **critical** value of the *F* test.
- Analysis of variance tests are always one-tailed tests with the rejection region in the upper tail.
- The decision rule is to **reject** the null hypothesis if the observed *F* value is greater than the critical *F* value

Comparison of F and t Values

- Analysis of variance can be used to test hypotheses about the difference in two means.
- Analysis of data from two samples by both a *t* test and an ANOVA shows that the observed
- *F* value equals the observed *t* value squared.

  $F = t^2$ for $df_C = 1$

- The ***t* test of independent samples actually is a special case of one-way ANOVA** when there are only two treatment levels ($df_C = 1$).
- The *t* test is computationally simpler than ANOVA for two groups.
- However, some statistical computer software packages do not contain a *t* test.
- In these cases, the researcher can perform a one-way ANOVA and then either take the square root of the *F* value to obtain the value of *t* or use the generated probability with the *p*-value method to reach conclusions.

Exercise (HW)

A company has three manufacturing plants, and company officials want to determine whether there is a difference in the average age of workers at the three locations. The following data are the ages of five randomly selected workers at each plant. Perform a one-way ANOVA to determine whether there is a significant difference in the mean ages of the workers at the three plants. Use $\alpha = .01$ and note that the sample sizes are equal.

Plant (Employee Ages)		
1	2	3
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26

SSC $\quad df_C = 3-1 = 2$

SSE $\quad df_E = 15-3 = 12$

SST $\quad df_T = 15-1 = 14$

F

critical value of $F_{0.01,2,12} = 6.93$

Solution

HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: At least one of the means is different from the others.

TEST:

STEP 2. The appropriate test statistic is the F test calculated from ANOVA.

STEP 3. The value of α is .01.

STEP 4. The degrees of freedom for this problem are $3 - 1 = 2$ for the numerator and $15 - 3 = 12$ for the denominator. The critical F value is $F_{.01,2,12} = 6.93$.

Because ANOVAs are always one tailed with the rejection region in the upper tail, the decision rule is to reject the null hypothesis if the observed value of F is greater than 6.93.

$$T_j: \quad T_1 = 141 \quad T_2 = 160 \quad T_3 = 124 \quad T = 425$$
$$n_j: \quad n_1 = 5 \quad n_2 = 5 \quad n_3 = 5 \quad N = 15$$
$$\bar{x}_j: \quad \bar{x}_1 = 28.2 \quad \bar{x}_2 = 32.0 \quad \bar{x}_3 = 24.8 \quad \bar{x} = 28.33$$

$$\text{SSC} = 5(28.2 - 28.33)^2 + 5(32.0 - 28.33)^2 + 5(24.8 - 28.33)^2 = 129.73$$
$$\text{SSE} = (29 - 28.2)^2 + (27 - 28.2)^2 + \ldots + (25 - 24.8)^2 + (26 - 24.8)^2 = 19.60$$
$$\text{SST} = (29 - 28.33)^2 + (27 - 28.33)^2 + \ldots + (25 - 28.33)^2 + (26 - 28.33)^2 = 149.33$$
$$\text{df}_C = 3 - 1 = 2$$
$$\text{df}_E = 15 - 3 = 12$$
$$\text{df}_T = 15 - 1 = 14$$

Source of Variance	SS	df	MS	F
Between	129.73	2	64.87	39.80
Error	19.60	12	1.63	
Total	149.33	14		

ACTION:

STEP 7. The decision is to reject the null hypothesis because the observed F value of 39.80 is greater than the critical table F value of 6.93.

References

- Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics by Ken Black