INTRODUCTION TO DATA ANALYTICS

Class # 28

ANOVA

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology
IIIT Sri City

How ANOVA?

SOME TERMINOLOGIES

- Factor
  - A characteristic under consideration, thought to influence the measured observations.
- Level (also called treatment)
  - A value of the factor

Typical data for a **Single-Factor** Experiment

Level	Observations (RBC)				Total	Mean
1_Anemic	y_{11}	y_{12}	…	y_{1n1}		
2_Normal	y_{21}	y_{22}	…	y_{2n2}		
3-Erythrocytosis	…	…	…	…		
…	…	…	…	…		
…	…	…	…	…		
k	y_{k1}	y_{k2}	…	y_{knk}		

EXAMPLE 3: SINGLE-FACTOR ANOVA

- Draw a straight line of between 20cm and 25 cm on a sheet of plain white card (only you know its exact length).
  - Collect 6 to 10 volunteers from each of Class VII, Class X and Class XII. Ask each volunteer to estimate independently the length of the line.
- Do differences in class means appear to outweigh differences within class?

What is/ are the Factor(s) and Levels here?

Example 4 : Two-Factor ANOVA

- Make a list of 10 food/household items purchased regularly by your family.
- Obtain the current prices of the items in three different shops; preferably a small 'corner' shop, a small supermarket and a large supermarket or hyper market.

What is/ are the Factor(s) and Levels here?

- Compare total shop prices.

VARIANTS OF ANOVA

Based on the number of Independent Variables and Dependent Variables considered for the study, there are different variants of ANOVA

1. **One-way ANOVA:** Only one independent variable (factor) with greater than 2 levels.
2. **Two-way ANOVA:** Two independent variables (i.e., factors).
3. **Three-way ANOVA:** Three independent variables (i.e., factors).

4. **Multivariate ANOVA:** It is used to test the significance of the effect of more independent variables.

One-way ANOVA

One-way ANOVA

- The purpose of the procedure is to compare sample means of k populations.
- In general, One-way ANOVA technique can be used to study the effect of k ($>$ 2) levels of a single factor.

- To determine if different levels of the factor affect measured observations differently, the following hypotheses are tested.

$$H_0: \mu_i = \mu \quad \text{all} \quad i = 1,2,\ldots,k$$
$$H_1: \mu_i \neq \mu \quad \text{some} \quad i = 1,2,\ldots,k$$

That is, at least one equality is not satisfied

where μ_i is the population mean for a level i.

Assumptions

- When applying one-way analysis of variance, there are three key assumptions that should be satisfied as follows.
    1. The observations are obtained independently and randomly from the populations defined by the factor levels.
    2. The population at each factor level is (approximately) normally distributed.
    3. These normal populations have a common variance, σ^2.
- Thus, for factor level i, the population is assumed to have a distribution which is $N(\mu_i, \sigma^2)$.

ONE-WAY ANOVA

Level (Group)	Observations (One Factor)				Total	Average
1			………			
2 . . .			……… ……… ……… ………			
k						

An entry in the table (e.g., y_{ij}) represents the j^{th} observation taken under the factor at level i.

- There will be, in general, n observations under the i^{th} level.
- $y_{i.}$ represents the total of the observations under the i^{th} level.
- $\overline{y_{i.}}$ represent the average of the observation under the i^{th} level.
- y_g represent the grand total of all the observation under the $factor$.
- $\overline{y_g}$ represent the average grand total of all the observation under the factor.

ONE-WAY ANOVA

Expressed symbolically,

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad\qquad i = 1,2, \dots \dots, k$$

$$\bar{y}_{i..} = \frac{y_{i.}}{n_i}$$

$$y_g = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \qquad\qquad \overline{y_g} = {}^{y_g}\!/_{N}$$

Here, N is the total observations, that is, $N = n_i + n_2 + \dots + n_k$

OVERALL VARIABILITY IN DATA

The correlated sum of squares for each factor level

$$SS_i = \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 \quad \text{for } i = 1, 2, \ldots, k$$

OVERALL VARIABILITY IN DATA

The corrected sum of squares for each factor level

$$SS_i = \sum_{i=1}^{n_i} (y_{ij} - \overline{y_g})^2$$

Alternatively, it can be represented as,

$$SS_i = \sum_{j=1}^{n_i} y_{ij}^2 - \frac{(y_{i.})^2}{n_i}$$

OVERALL VARIABILITY IN DATA

We then calculate a pooled sum of squares

$$SS_p = \sum_{i=1}^{k} SS_i$$

Finally, the pooled sample of variance is

$$s_p = \frac{SS_p}{pooled\ degree\ of\ fredom} = \frac{SS_p}{(\sum n_i) - k}$$

Note that if the individual variances are available, the same can be computed as

$$s_p = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{(\sum n_i) - k}$$

where s_i^2 are the variances for each sample. This is also called **variance within samples** and also popularly be denoted as $\hat{\sigma}_W^2$

Example 5: Variance within Samples

- The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours, of samples of 60W electric light bulbs of three different brands.

Lifetime of bulb		
Blub 1	Blub 2	Blub 3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

Solution : Variance within Samples

- Here, there is one factor (lifetime of the blub) at three levels (blub 1, bulb 2 and bulb3). Also the sample sizes are all equal (to 5).

- The sample mean and variance (divisor $(n - 1)$) for each level are as follows.

	Lifetime of bulb		
	Blub 1	Blub 2	Blub 3
Sample Size	5	5	5
Sum	80	100	135
Sum of squares	1316	2040	3689
Mean	16	20	27
Variance	9	10	11

Solution : Variance within Samples

- A pooled estimate of variance then can be calculated as follows.

$$\hat{\sigma}_W^2 = \frac{(5-1)\times 9 + (5-1)\times 10 + (5-1)\times 11}{5+5+5-3} = 10$$

- This quantity is called the **variance within samples**.
- It is an estimate of σ^2 based on $v = 5+5+5-3 = 12$ degrees of freedom.

Heuristic Justification of ANOVA

- From the sampling distribution of the mean, we know that a sample mean computed from a random sample of size n from a population with mean μ and variance σ^2 is a random variable with mean μ and variance σ^2/n [Central Limit Theorem].

- Let us see, what we can conclude in case of k (where $k > 1$) populations, which may have different μ_i but have the same variance σ^2.

Heuristic Justification of ANOVA

- If the null hypothesis is true, that is, each of the μ_i has the same value, say, μ, then the distribution of each of the k sample means, $\overline{y_{i.}}$ will have mean μ and variance σ^2/n .

- It then follows that, if we calculate a variance using the sample means as observations,

$$\hat{\sigma}_B^2 = \Sigma(\bar{y}_{i.} - \bar{y}_g)^2/(k-1)$$

- Then the quantity is an estimate of σ^2/n .

- Hence, $n\hat{\sigma}_B^2$ is an estimate of σ^2.

Heuristic Justification of ANOVA

- Out of several sampling distributions, the F-distribution describes the ratio of two independent estimates of a common variance.

- The parameters of the distribution are the degrees of freedom of the numerator and denominator variances, respectively.

- If the null hypothesis of equal mean is true, then we can compute the two estimates of σ^2 namely

  $\hat{\sigma}_B^2 = \sum(\bar{y}_{i.} - \bar{y}_g)^2/(k-1)$ and s_p^2, the pooled variance

- Therefore, the ratio $\frac{\widehat{n\,\sigma_B^2}}{s_p^2}$ has the F-distribution with degrees of freedom (k-1) and $n-k$

Heuristic Justification of ANOVA

- Thus, the procedure for testing the hypothesis.

$$H_0: \mu_i = \mu \quad \text{all} \quad i = 1, 2, \ldots, k$$
$$H_1: \text{at least one equality is not satisfied}$$

- We are to reject H_0, if the calculated value of F $= \dfrac{\widehat{n\,\sigma}_B^2}{s_p^2}$ exceeds α (confidence level) of the F-distributions with (k-1) and $n - k$ degrees of freedom.

Example 6: F-Test

Set 1			Set 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$	$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$

- For both sets, the value of $n\hat{\sigma}_B^2$ is 101.67. However, for Set 1, $s_p^2 = 0.250$ while for Set 2, $s_p^2 = 10.67$. Thus, for Set 1, F = 406.67 and for Set 2, F = 9.53.

Example 7: Variance between Samples

- The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours, of samples of $60W$ electric light bulbs of three different brands.

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

- Assuming all lifetimes to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the three brands with respect to mean lifetime.

Solution : Variance between Samples

- The variability between samples may be estimated from the three sample means as follows.

	Brand		
	1	2	3
Sample Mean	16	20	27
Sum		63	
Sum of squares		1385	
Mean		21	
Variance		31	

- This variance (divisor $(n-1)$), denoted by $\hat{\sigma}_{\bar{B}}^2$ is called the **variance between sample means**. Since it calculated using sample means, it is an estimate of

$$\frac{\sigma^2}{5} \text{ (that is } \frac{\sigma^2}{n} \text{ in general)}$$

based upon $(3 - 1) = 2$ degrees of freedom, but only if the null hypothesis is true. If H_0 is false, then the subsequent 'large' differences between the sample means will result in $5\hat{\sigma}_{\bar{B}}^2$ being an inflated estimate of σ^2.

Solution : F-Test

- The two estimates of σ^2, $\widehat{n\sigma}_B^2$ and $\hat{\sigma}_W^2$, may be tested for equality using the F-test with

$$F = \frac{5\hat{\sigma}_B^2}{\hat{\sigma}_W^2}$$

  as lifetimes may be assumed to be normally distributed.

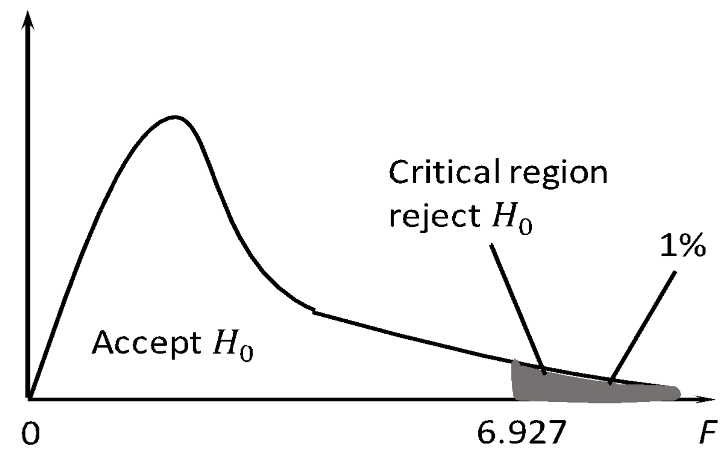
- Recall that the F-test requires the two variances to be independently distributed (from independent samples). Although this is by no means obvious here (both were calculated from the same data), $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$ are in fact independently distributed.

- The test is always one-sided, upper-tail, since if H_0 is false, $\hat{\sigma}_W^2$ is inflated whereas $5\hat{\sigma}_B^2$ is unaffected.

- **Thus in analysis of variance, the convention of placing the larger sample variance in the numerator of the F-statistic is NOT applied.**

Solution

- The solution is thus summarized and completed as follows.
  - H_0: $\mu_i = \mu$ all $i = 1, 2, 3$
  - H_1: $\mu_i \neq \mu$ some $i = 1, 2, 3$
  - Significance level, $\alpha = 0.01$
  - Degrees of freedom, $v_1 = 2$, $v_2 = 12$
  - Critical region is $F > 6.927$



  - Test statistic is $F = \frac{5\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = \frac{155}{10} = 15.5$

- This value does lie in the critical region. There is evidence, at the 1% significance level, that the true mean lifetimes of the three brands of bulb do differ.

Notation and computational formulae

- In essence, given a population a single factor of k levels, we have to calculate two estimations for σ^2.

- Sampling variance between groups with (k-1) degree of freedom

$$n\hat{\sigma}_B^2 = n\sum(\bar{y}_{i.} - \bar{y}_g)^2/(k-1).$$

- Sampling variance within groups with (n-k) degree of freedom

$$\hat{\sigma}_W^2 = \frac{\sum_{i=1}^{k} SS_i}{\sum n_i - k}$$

Notation and computational formulae

- The calculations undertaken in the previous example are somewhat cumbersome, and are prone to inaccuracy with non-integer sample means. They also require considerable changes when the sample sizes are unequal. Equivalent computational formulae are available which cater for both equal and unequal sample sizes.

- First, some notation.

Number of samples (or levels)

Total number of observations

Notation and computational formulae

- The computational formulae now follow.

Total sum of squares,
Between samples sum of squares,
Within samples sum of squares,

- A mean square (or unbiased variance estimate) is given by

(sum of squares) ÷ (degrees of freedom)

e.g. $$\hat{\sigma}^2 = \frac{(x-\bar{x})^2}{n-1}$$

Hence

Total mean square,
Between samples mean square,
Within samples mean square,

- **Note that for the degrees of freedom: $(k-1) + (n-k) = (n-1)$**

Example 8: F-Test using Formula

- For the previous example on 60W electric light bulbs, use these computational formulae to show the following.

(a) $SS_T = 430$

(b) $SS_B = 310$

(c) $MS_B = 155\ (5\hat{\sigma}_B^2)$

(d) $MS_W = 10\ (\hat{\sigma}_W^2)$

Note that $F = \dfrac{MS_B}{MS_W} = \dfrac{155}{10} = 15.5$ as previously.

One-way ANOVA Table

- It is convenient to summarize the results of an analysis of variance in a table. For a one factor analysis this takes the following form.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between samples				
Within samples				
Total				

Example 9: F-Test for unbalanced

- In a comparison of the cleaning action of four detergents, 20 pieces of white cloth were first soiled with India ink. The cloths were then washed under controlled conditions with 5 pieces washed by each of the detergents. Unfortunately three pieces of cloth were 'lost' in the course of the experiment. Whiteness readings, made on the 17 remaining pieces of cloth, are shown below.

Detergent			
A	B	C	D
77	74	73	76
81	66	78	85
61	58	57	77
76		69	64
69		63	

- Assuming all whiteness readings to be normally distributed with common variance, test the hypothesis of no difference between the four brands as regards mean whiteness readings after washing.

Solution

-
  - H_0: $\mu_i = \mu$ all $i = 1, 2, 3$
  - H_1: $\mu_i \neq \mu$ some $i = 1, 2, 3$
  - Significance level, $\alpha = 0.05$ (say)
  - Degrees of freedom, $v_1 = k - 1 = 3$,

    and $v_2 = n - k = 17 - 4 = 13$
  - Critical region is $F > 3.411$

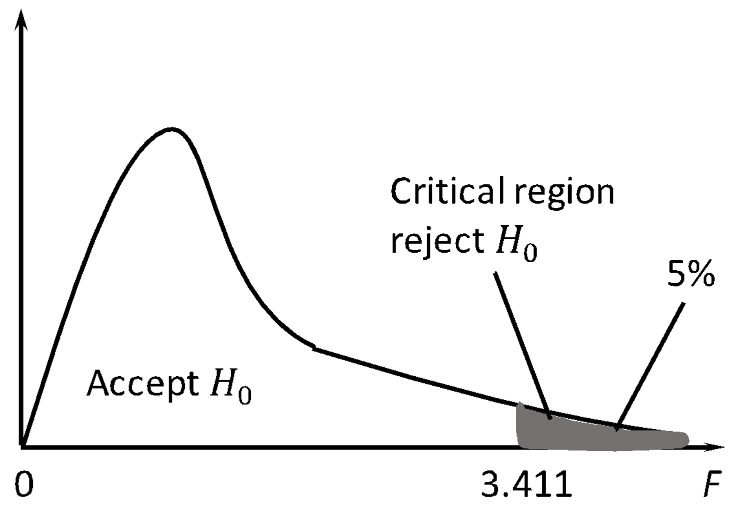


Table A.9 Critical Values for *F* Distributions *(cont.)*

		ν_1 = numerator df								
ν_2 = denominator df	α	1	2	3	4	5	6	7	8	9
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39

Solution

	A	B	C	D	Total
	5	3	5	4	
	364	198	340	302	

$$\sum_i \sum_j {y_{ij}}^2 = 86362$$

$$SS_T = 86362 - \frac{1204^2}{17} = 1090.47$$

$$SS_B = \left(\frac{364^2}{5} + \frac{198^2}{3} + \frac{340^2}{5} + \frac{302^2}{4}\right) - \frac{1204^2}{17} = 216.67$$

$$SS_W = 1090.47 - 216.67 = 873.80$$

Solution

- The ANOVA table is now as follows.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between detergents				
Within detergents				
Total				

- The F ratio of 1.07 does not lie in the critical region.

- Thus there is no evidence, at the 5% significance level, to suggest a difference between the four brands as regards mean whiteness after washing.

Two-way ANOVA

Two-way (factor) anova

- This is an extension of the one factor situation to take account of a second factor.

- The levels of this second factor are often determined by groupings of subjects or units used in the investigation. As such it is often called a blocking factor because it places subjects or units into homogeneous groups called blocks. The design itself is then called a randomised block design.

Example 10: Two-factor Analysis

- A computer manufacturer wishes to compare the speed of four of the firm's compilers. The manufacturer can use one of two experimental designs.

  a) Use 20 similar programs, randomly allocating 5 programs to each compiler.

  b) Use 4 copies of any 5 programs, allocating 1 copy of each program to each compiler.

- Which of (a) and (b) would you recommend, and why?

Solution

- In (a), although the 20 programs are similar, any differences between them may affect the compilation times and hence perhaps any conclusions. Thus in the 'worst scenario', the 5 programs allocated to what is really the fastest compiler could be the 5 requiring the longest compilation times, resulting in the compiler appearing to be the slowest! If used, the results would require a one factor analysis of variance; the factor being compiler at 4 levels.

- In (b), since all 5 programs are run on each compiler, differences between programs should not affect the results. Indeed it may be advantageous to use 5 programs that differ markedly so that comparisons of compilation times are more general. For this design, there are two factors; compiler (4 levels) and program (5 levels). The factor of principal interest is compiler whereas the other factor, program, may be considered as a blocking factor as it creates 5 blocks each containing 4 copies of the same program.

- **Thus (b) is the better designed investigation.**

SOLUTION

- The actual compilation times, in milliseconds, for this two factor (randomised block) design are shown in the following table.

	Compiler			
	1	2	3	4
Program A	29.21	28.25	28.20	28.62
Program B	26.18	26.02	26.22	25.56
Program C	30.91	30.18	30.52	30.09
Program D	25.14	25.26	25.20	25.02
Program E	26.16	25.16	25.26	25.46

Assumptions and Interaction

- The three assumptions for a two factor analysis of variance when there is only one observed measurement at each combination of levels of the two factors are as follows.
  1. The population at each factor level combination is (approximately) normally distributed.
  2. These normal populations have a common variance, σ^2 .
  3. The effect of one factor is the same at all levels of the other factor.

  Hence from assumptions 1 and 2, when one factor is at level i and the other at level j, the population has a distribution which is

  $$N(\mu_{ij}, \sigma^2)$$

- Assumption 3 is equivalent to stating that there is no interaction between the two factors.

ASSUMPTIONS AND INTERACTION

- Now interaction exists when the effect of one factor depends upon the level of the other factor. For example consider the effects of the two factors:
  sugar (levels none and 2 teaspoons), and stirring (levels none and 1 minute), on the sweetness of a cup of tea.

- Stirring has no effect on sweetness if sugar is not added but certainly does have an effect if sugar is added. Similarly, adding sugar has little effect on sweetness unless the tea is stirred.

- Hence factors sugar and stirring are said to interact.

- Interaction can only be assessed if more than one measurement is taken at each combination of the factor levels. Since such situations are beyond the scope of this text, it will always be assumed that interaction between the two factors does not exist.

ASSUMPTIONS AND INTERACTION

- Thus, for example, since it would be most unusual to find one compiler particularly suited to one program, the assumption of no interaction between compilers and programs appears reasonable.

Notation and Computational Formulae

- As illustrated earlier, the data for a two-way ANOVA can be displayed in a two-way table. It is thus convenient, in general, to label the factors as

  a **row factor** and a **column factor**.

- Notation, similar to that for the one factor case, is then as follows.

Number of levels of row factor	$= r$
Number of levels of column factor	$= c$
Total number of observations	$= rc$
Observation in (i j-th cell of table	$= x_{ij}$
(ith level of row factor and jth level of column factor)	i=1,2,…,r j=1,2,…,c

Notation and computational formulae

Sum of c observations in i-th row

Sum of r observations in j-th column

Sum of all rc observations

- These lead to the following computational formulae which again are similar to those for one-way ANOVA except that there is an additional sum of squares, etc. for the second factor.

Notation and computational formulae

Total sum of squares,

Between rows sum of squares,

Between columns sum of squares,

Error (residual) sum of squares,

What are the degrees of freedom for SST , SSR and SSC when there are 20 observations in a table of 5 rows and 4 columns?
What is the degrees of freedom of SSE ?

ANOVA Table and Hypothesis Test

For a two factor analysis of variance this takes the following form.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between rows		r- 1		
Between columns		c - 1		
Error (residual)		(r-1) (c-1)		
Total		rc- 1		

- **Notes :**
  1. The three sums of squares, SS_R , SS_C and SS_E are independently distributed.
  2. For the degrees of freedom:
     $(r\text{-}1)+ (c\text{-}1) +(r\text{-}1)+ (c\text{-}1) = rc\text{-} 1$

ANOVA Table and Hypothesis Test

- Using the F ratios, tests for significant row effects and for significant column effects can be undertaken.

H0: no effect due to row factor	**H0: no effect due to column factor**
H1: an effect due to row factor	H1: an effect due to column factor

Example 11: Two-way ANOVA

- Returning to the compilation times, in milliseconds, for each of five programs, run on four compilers.
- Test, at the 1% significance level, the hypothesis that there is no difference between the performance of the four compilers.
- Has the use of programs as a blocking factor proved worthwhile? Explain.
- The data, given earlier, are reproduced below.

	Compiler			
	1	2	3	4
Program A	29.21	28.25	28.20	28.62
Program B	26.18	26.02	26.22	25.56
Program C	30.91	30.18	30.52	30.09
Program D	25.14	25.26	25.20	25.02
Program E	26.16	25.16	25.26	25.46

Solution : Dataset

- To ease computations, these data have been transformed (coded) by

$$x = 100 \times (\text{time} - 25)$$

to give the following table of values and totals.

	Compiler				
	1	2	3	4	
Program A	421	325	320	362	**1428**
Program B	118	102	122	56	**398**
Program C	591	518	552	509	**2170**
Program D	14	26	20	2	**62**
Program E	116	14	26	46	**202**
	1260	**985**	**1040**	**975**	**4260 = T**

Solution : Parameters

- The sums of squares are now calculated as follows.
  (Rows = Programs, Columns = Compilers)

- $SS_T = 1757768 = \frac{4260^2}{20} = 850388$
- $SS_R = \frac{1}{4}(1428^2 + 398^2 + \ 2170^2 + 62^2 + 202^2)\text{-}\ \frac{4260^2}{20} = 830404$
- $SS_C = \frac{1}{5}(1260^2 + 985^2 + 1040^2 + 975^2)\ \text{-}\ \frac{4260^2}{20} = 10630$

- SS_E =850388 – 830404 – 10630 = 9354

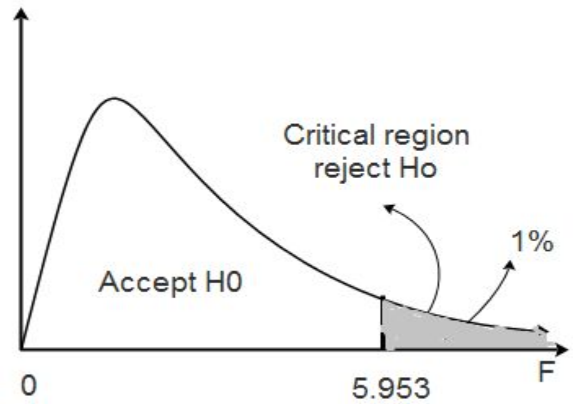
Solution: ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between programs	830404	4	207601.0	266.33
Between compilers	10630	3	3543.3	4.55
Error (residual)	9354	12	779.5	
Total	850388	19		

Solution : Hypothesis Test

- H0: no effect on compilation times due to compilers
- H1: an effect on compilation times due to compilers
- Significance level, $\alpha = 0.001$
- Degrees of freedom, $v1 = c - 1 = 3$

  and $v2 = (r - 1)(c - 1) = 4 \times 3 = 12$
- Critical region is $F > 5.953$
- Test statistic $FC = 4.55$



- This value does not lie in the critical region. Thus there is no evidence, at the 1% significance level, to suggest a difference in compilation times between the four compilers.

REFERENCE

- The detail material related to this lecture can be found in

Design and Analysis of Experiments (8th Edition), Douglas C. Montgomery, John Wiley & Sons, 2013.

Any question?