F-Test: ANOVA

-Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Application of F test

- Test assumption of equal variances that was made in using the t-test
- Interest in actually comparing the variance of two populations

F test

- While dealing with t test we assumed that variance of two population are equal. This assumption can be verified by testing hypothesis $H_0 = {\sigma_1}^2 > {\sigma_2}^2$. This test is called variance ratio test.
- If n_1, n_2 be the sizes of the two samples and ${S_1}^2, {S_2}^2$ be the corresponding estimates of the population variance based on the two samples then F statistic is

$$F = \frac{{S_1}^2}{{S_2}^2}$$

- This is F distribution with $(n_1 - 1, n_2 - 1)$ DF, population is normally distributed.

- F-tests are named after its test statistic, F, which was named in honor of Sir Ronald Fisher. **The F-statistic is simply a ratio of two variance.**
- F-statistics are based on the ratio of mean squares. The term "mean squares" may sound confusing but it is simply an estimate of population variance that accounts for the degrees of freedom (DF) used to calculate that estimate.

Testing Procedure: To compare Two Variances

1. Formulate the null and alternate hypotheses.

   H_0: $\sigma_1^2 = \sigma_1^2$

   H_a: $\sigma_1^2 > \sigma_2^2$

   [Note that we might also use $\sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 =/ \sigma_2^2$]

2. Calculate the F ratio.

   $F = s_1^2/s_1^2$

   [where s_1 is the largest or the two variances]

3. Reject the null hypothesis of equal population variances if $F(v_1-1, v_2-1) > F\alpha$

   [or $F_{\alpha/2}$ in the case of a two tailed test]

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Example

The variability in the amount of impurities present in a batch of chemicals used for a particular process depends on the length of time that the process is in operation.

Suppose a sample of size 25 is drawn from the normal process which is to be compared to a sample of a new process that has been developed to reduce the variability of impurities.

	Sample 1	**Sample 2**
n	**25**	**25**
s^2	**1.04**	**0.51**

Continued…

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 > \sigma_2^2$

$F(24,24) = s_1^2/s_2^2 = 1.04/.51 = 2.04$

Assuming $\alpha = 0.05$

cv = 1.98 < 2.04

Thus, reject H_0 and conclude that the variability in the new process (Sample 2) is less than the variability in the original process.

ANOVA

(Analysis of Variance)

- Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken.
- Most of the time ANOVA is used to compare the equality of three or more means, however when the means from two samples are compared using ANOVA it is equivalent to using a t-test to compare the means of independent samples.

Assumptions of ANOVA:

(i) All populations involved follow a normal distribution.

(ii) All populations have the same variance (or standard deviation).

(iii) The samples are randomly selected and independent of one another.

- To use the F-test to determine whether group means are equal, it's just a matter of including the correct variances in the ratio. In one-way ANOVA, the F-statistic is this ratio:

$$F = \frac{variation\ between\ sample\ means}{variation\ within\ the\ samples}$$

One-Way ANOVA

One-Way ANOVA

- The one-way analysis of variance is used to test the claim that three or more population means are equal
- This is an extension of the two independent samples t-test

One-Way ANOVA

- The *response* variable is the variable you're comparing
- The *factor* variable is the categorical variable being used to define the groups
  - We will assume *k* samples (groups)
- The *one-way* is because each value is classified in exactly one way
  - Examples include comparisons by gender, race, political party, color, etc.

One-Way ANOVA

- The null hypothesis is that the means are all equal

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

- The alternative hypothesis is that at least one of the means is different

One-Way ANOVA

- The statistics classroom is divided into three rows: front, middle, and back
- The instructor noticed that the further the students were from him, the more likely they were to miss class or use an instant messenger during class
- He wanted to see if the students further away did worse on the exams

One-Way ANOVA

The ANOVA doesn't test that one mean is less than another, only whether they're all equal or at least one is different.

$$H_0 : \mu_F = \mu_M = \mu_B$$

One-Way ANOVA

- A random sample of the students in each row was taken
- The score for those students on the second exam was recorded
  - Front: 82, 83, 97, 93, 55, 67, 53
  - Middle: 83, 78, 68, 61, 77, 54, 69, 51, 63
  - Back: 38, 59, 55, 66, 45, 52, 52, 61

One-Way ANOVA

The summary statistics for the grades of each row are shown in the table below

Row	Front	Middle	Back
Sample size	7	9	8
Mean	75.71	67.11	53.50
St. Dev	17.63	10.95	8.96
Variance	310.90	119.86	80.29

One-Way ANOVA

- Variation
  - Variation is the sum of the squares of the deviations between a value and the mean of the value
  - Sum of Squares is abbreviated by SS and often followed by a variable in parentheses such as SS(B) or SS(W) so we know which sum of squares we're talking about

One-Way ANOVA

- Are all of the values identical?
  - No, so there is some variation in the data
  - This is called the total variation
  - Denoted SS(Total) for the total Sum of Squares (variation)
  - Sum of Squares is another name for variation

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

One-Way ANOVA

- Are all of the sample means identical?
  - No, so there is some variation between the groups
  - This is called the between group variation
  - Sometimes called the variation due to the factor
  - Denoted SS(B) for Sum of Squares (variation) between the groups

One-Way ANOVA

- Are each of the values within each group identical?
  - No, there is some variation within the groups
  - This is called the within group variation
  - Sometimes called the error variation
  - Denoted SS(W) for Sum of Squares (variation) within the groups

One-Way ANOVA

- There are two sources of variation
  - the variation between the groups, SS(B), or the variation due to the factor
  - the variation within the groups, SS(W), or the variation that can't be explained by the factor so it's called the error variation

One-Way ANOVA

- Here is the basic one-way ANOVA table

Source	SS	df	MS	F	p
Between					
Within					
Total					

One-Way ANOVA

- Grand Mean
  - The grand mean is the average of all the values when the factor is ignored
  - It is a weighted average of the individual sample means

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

$$\bar{\bar{x}} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \cdots + n_k \bar{x}_k}{n_1 + n_2 + \cdots + n_k}$$

One-Way ANOVA

- Grand Mean for our example is 65.08

$$\bar{\bar{x}} = \frac{7(75.71)+9(67.11)+8(53.50)}{7+9+8}$$

$$\bar{\bar{x}} = \frac{1562}{24}$$

$$\bar{\bar{x}} = 65.08$$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

One-Way ANOVA

- Between Group Variation, SS(B)
  - The between group variation is the variation between each sample mean and the grand mean
  - Each individual variation is weighted by the sample size

$$SS\left(B\right)=\sum_{i=1}^{k}n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2$$

$$SS\left(B\right)=n_1\left(\bar{x}_1-\bar{\bar{x}}\right)^2+n_2\left(\bar{x}_2-\bar{\bar{x}}\right)^2+\cdots+n_k\left(\bar{x}_k-\bar{\bar{x}}\right)^2$$

One-Way ANOVA

The Between Group Variation for our example is SS(B)=1902

$$SS\left(B\right)=7\left(75.71-65.08\right)^{2}+9\left(67.11-65.08\right)^{2}+8\left(53.50-65.08\right)^{2}$$

$$SS\left(B\right)=1900.8376\approx1902$$

One-Way ANOVA

- Within Group Variation, SS(W)
  - The Within Group Variation is the weighted total of the individual variations
  - The weighting is done with the degrees of freedom
  - The df for each sample is one less than the sample size for that sample.

One-Way ANOVA

Within Group Variation

$$SS(W) = \sum_{i=1}^{k} df_i s_i^2$$

$$SS(W) = df_1 s_1^2 + df_2 s_2^2 + \cdots + df_k s_k^2$$

One-Way ANOVA

- The within group variation for our example is 3386

$$SS(W) = 6(310.90) + 8(119.86) + 7(80.29)$$

$$SS(W) = 3386.31 \approx 3386$$

One-Way ANOVA

- After filling in the sum of squares, we have …

Source	SS	df	MS	F	p
Between	1902				
Within	3386				
Total	5288				

One-Way ANOVA

- Degrees of Freedom, df
  - A degree of freedom occurs for each value that can vary before the rest of the values are predetermined
  - For example, if you had six numbers that had an average of 40, you would know that the total had to be 240. Five of the six numbers could be anything, but once the first five are known, the last one is fixed so the sum is 240. The df would be 6-1=5
  - The df is often one less than the number of values

One-Way ANOVA

- The between group df is one less than the number of groups
  - We have three groups, so df(B) = 2
- The within group df is the sum of the individual df's of each group
  - The sample sizes are 7, 9, and 8
  - df(W) = 6 + 8 + 7 = 21
- The total df is one less than the sample size
  - df(Total) = 24 – 1 = 23

One-Way ANOVA

- Filling in the degrees of freedom gives this ...

Source	SS	df	MS	F	p
Between	1902	2			
Within	3386	21			
Total	5288	23			

One-Way ANOVA

- Variances
  - The variances are also called the Mean of the Squares and abbreviated by MS, often with an accompanying variable MS(B) or MS(W)
  - They are an average squared deviation from the mean and are found by dividing the variation by the degrees of freedom
  - MS = SS / df

$$Variance = \frac{Variation}{df}$$

One-Way ANOVA

- MS(B) = 1902 / 2 = 951.0
- MS(W) = 3386 / 21 = 161.2
- MS(T) = 5288 / 23 = 229.9
  - Notice that the MS(Total) is NOT the sum of MS(Between) and MS(Within).
  - This works for the sum of squares SS(Total), but not the mean square MS(Total)
  - The MS(Total) isn't usually shown

One-Way ANOVA

- Completing the MS gives …

Source	SS	df	MS	F	p
Between	1902	2	951.0		
Within	3386	21	161.2		
Total	5288	23	229.9		

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

One-Way ANOVA

- Special Variances
  - The MS(Within) is also known as the pooled estimate of the variance since it is a weighted average of the individual variances
    - Sometimes abbreviated s_p^2
  - The MS(Total) is the variance of the response variable.
    - Not technically part of ANOVA table, but useful none the less

One-Way ANOVA

- F test statistic
  - An F test statistic is the ratio of two sample variances
  - The MS(B) and MS(W) are two sample variances and that's what we divide to find F.
  - F = MS(B) / MS(W)
- For our data, F = 951.0 / 161.2 = 5.9

One-Way ANOVA

- Adding F to the table …

Source	SS	df	MS	F	p
Between	1902	2	951.0	5.9	
Within	3386	21	161.2		
Total	5288	23	229.9		

One-Way ANOVA

- The F test is a right tail test
- The F test statistic has an F distribution with df(B) numerator df and df(W) denominator df
- The p-value is the area to the right of the test statistic
- $P(F_{2,21} > 5.9) = 0.009$

One-Way ANOVA

- Completing the table with the p-value

Source	SS	df	MS	F	p
Between	1902	2	951.0	5.9	0.009
Within	3386	21	161.2		
Total	5288	23	229.9		

One-Way ANOVA

- The p-value is 0.009, which is less than the significance level of 0.05, so we reject the null hypothesis.
- The null hypothesis is that the means of the three rows in class were the same, but we reject that, so at least one row has a different mean.

One-Way ANOVA

- There is enough evidence to support the claim that there is a difference in the mean scores of the front, middle, and back rows in class.
- The ANOVA doesn’t tell which row is different, you would need to look at confidence intervals or run post hoc tests to determine that

Independent Random Samples from Two Populations of Serum Uric Acid values

	Sample 1	Sample 2
	1.2	1.7
	0.8	1.5
	1.1	2.0
	0.7	2.1
	0.9	1.1
	1.1	0.9
	1.5	2.2
	0.8	1.8
	1.6	1.3
	0.9	1.5
Sum	10.6	16.1
Mean	1.06	1.61

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Serum Acid SS (Total) Worksheet

Person	x	x^2
1	1.2	1.44
2	0.8	0.64
3	1.1	1.21
4	0.7	0.49
5	0.9	0.81
6	1.1	1.21
7	1.5	2.25
8	0.8	0.64
9	1.6	2.56
10	0.9	0.81
11	1.7	2.89
12	1.5	2.25
13	2.0	4.00
14	2.1	4.41
15	1.1	1.21
16	0.9	0.81
17	2.2	4.84
18	1.8	3.24
19	1.3	1.69
20	1.5	2.25
Sum	26.7	39.65
Mean	1.34	
Sum^2/n	35.64	
SS(Total)	4.01	
Variance	0.21	
SD	0.46	

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

SS (Within) and SS (Among) worksheet

	x	x^2	x	x^2
	1.2	1.44	1.7	2.89
	0.8	0.64	1.5	2.25
	1.1	1.21	2.0	4.00
	0.7	0.49	2.1	4.41
	0.9	0.81	1.1	1.21
	1.1	1.21	0.9	0.81
	1.5	2.25	2.2	4.84
	0.8	0.64	1.8	3.24
	1.6	2.56	1.3	1.69
	0.9	0.81	1.5	2.25
Sum	10.6	12.06	16.1	27.59
Mean	1.06		1.61	
Sum^2/n	11.236		25.921	
SS	0.824		1.669	
Variance	0.092		0.185	
SD	0.303		0.431	

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

SS (Within) = SS (sample 1) + SS (sample 2)

= 0.824 + 1.669

= 2.490

$\boxed{\text{SS (Within)} = 2.49}$

$$SS(\text{Among}) = \frac{\text{sum}_1^{\,2}}{n_1} + \frac{\text{sum}_2^{\,2}}{n_2} - \frac{\text{total}^2}{20}$$

$$= \frac{(10.6)^2}{10} + \frac{(16.1)^2}{10} - \frac{(26.7)^2}{20}$$

$$= 11.236 + 25.921 - 35.64$$

$$= 1.51$$

$$\boxed{SS(\text{Among}) = 1.51}$$

1. The hypothesis: H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$

2. The assumptions: Independent random samples, normal distributions, $\sigma_1^2 = \sigma_2^2$

3. The α-level : $\alpha = 0.05$

4. The test statistic: ANOVA

5. The rejection region: Reject H0: $\mu_1 = \mu_2$ if

$$F = \frac{MS(Among)}{MS(Within)} > F_{0.95(1,18)} = 4.41$$

Where MS(Among)=SS(Among)/DF(Among)

MS(Within)=SS(Within)/DF(Within)

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

6. The result:

ANOVA				
Source	df	SS	MS	F
Among	1	1.52	1.52	10.86
Within	18	2.49	0.14	
Total	19	4.01		

7. The conclusion: Reject H_0: $\mu_1 = \mu_2$

Since $F = 10.86 > F_{0.95}(1,18) = 4.41$

Testing the Hypothesis that the Two Serum Uric Acid Populations have the Same Mean

1. The hypothesis: $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

2. The α-level: $\alpha = 0.05$

3. The assumptions: Independent Random Samples Normal Distribution $\sigma_1^2 = \sigma_2^2$

4. The test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

5. The reject region: Reject H_0 if t is not between ± 2.1009

6. The result:

$$t = \frac{0.55}{0.37(0.45)} = 3.30$$

7. The conclusion: Reject $H_0 : \mu_1 = \mu_2$ since t is not between ± 2.1009

Example

Independent Random Samples from Three Populations of Serum Uric Acid Values

	Sample		
	1	2	3
	1.2	1.7	1.3
	0.8	1.5	1.5
	1.1	2.0	1.4
	0.7	2.1	1.0
	0.9	1.1	1.8
	1.1	0.9	1.4
	1.5	2.2	1.9
	0.8	1.8	0.9
	1.6	1.3	1.9
	0.9	1.5	1.8
Sum	10.6	16.1	14.9
Mean	1.06	1.61	1.49

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Independent Random Samples from Three Populations of Serum Uric Acid Values

ANOVA Worksheet

	1		2		3		Combined Total	
	x	x^2	x	x^2	x	x^2	x	x^2
	1.2	1.44	1.7	2.89	1.3	1.69		
	0.8	0.64	1.5	2.25	1.5	2.25		
	1.1	1.21	2.0	4.00	1.4	1.96		
	0.7	0.49	2.1	4.41	1.0	1.00		
	0.9	0.81	1.1	1.21	1.8	3.24		
	1.1	1.21	0.9	0.81	1.4	1.96		
	1.5	2.25	2.2	4.84	1.9	3.61		
	0.8	0.64	1.8	3.24	0.9	0.81		
	1.6	2.56	1.3	1.69	1.9	3.61		
	0.9	0.81	1.5	2.25	1.8	3.24		
Sum	10.6	12.06	16.1	27.59	14.9	23.37	41.6	63.020
n	10		10		10		30	
Mean	1.06		1.61		1.49		1.39	
Sum^2/n	11.236		25.921		22.201		57.685	
SS	0.824		1.669		1.169		5.335	
Variance	0.092		0.185		0.130		0.184	
SD	0.303		0.431		0.360		0.429	

SS(Among) = 11.236 + 25.921 + 22.201 - 57.685

= 1.673

SS(Within) = 0.824 + 1.669 + 1.169

= 3.662

SS(Total) = 1.673 + 3.662 = 5.335

Testing the Hypothesis that the Three populations have the same Average Serum Uric Acid Levels

1. **The hypothesis:** $H_0: \mu_1 = \mu_2 = \mu_3$, vs. $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

2. **The assumptions:** Independent random samples
   normal distributions $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

3. **The α-level :** $\alpha = 0.05$

4. **The test statistic:** ANOVA

5. The Rejection Region: Reject H_0: $\mu_1 = \mu_2 = \mu_3$ if

$$F = \frac{MS(Among)}{MS(Within)} > F_{0.95(2,27)} = 3.35$$

where

$$MS(Among) = \frac{SS(Among)}{df\,(Among)}, \quad MS(Within) = \frac{SS(Within)}{df\,(Within)}$$

6. The Result:

ANOVA

Source	df	SS	MS	F
Among	2	1.67	0.84	6.00
Within	27	3.66	0.14	
Total	29	5.33		

7. The Conclusion: Reject H_0: Since $F = 6.00 > F_{0.95}(2, 27) = 3.35$.

Example

A random sample of n = 10 was taken from each of three populations of young males. Systolic blood pressure measurements were taken on each child. The measurements are listed below.

	Group		
	1	2	3
	100	104	105
	102	88	112
	96	100	90
	106	98	104
	110	102	96
	110	92	110
	120	96	98
	112	100	86
	112	96	80
	90	96	84
Sum	1,058	972	965
Mean	105.8	97.2	96.5

Independent Random Samples from Three Populations of Blood Pressure Levels

ANOVA Worksheet

	1		2		3		Combined Total	
	x	x^2	x	x^2	x	x^2	x	x^2
	100	10,000	104	10,816	105	11,025		
	102	10,404	88	7,744	112	12,544		
	96	9,216	100	10,000	90	8,100		
	106	11,236	98	9,604	104	10,816		
	110	12,100	102	10,404	96	9,216		
	110	12,100	92	8,464	110	12,100		
	120	14,400	96	9,216	98	9,604		
	112	12,544	100	10,000	86	7,396		
	112	12,544	96	9,216	80	6,400		
	90	8,100	96	9,216	84	7,056		
Sum	1,058	112,644	972	94,680	965	94,257	2,995	301,581
n	10		10		10		30	
Mean	105.8		97.2		96.5		99.8	
Sum^2/n	111,936		94,478		93,123		299,001	
SS	708		202		1135		2580	
Variance	78.6		22.4		126.1		89.0	
SD	8.9		4.7		11.2		9.4	

SS(Among) = 111,936 + 94,478 + 93,123 - 299,001

= 536.47

SS(Within) = 708 + 202 + 1,134

= 2,043.70

SS(Total) = 536 + 2,043 = 2,580.17

Testing the Hypothesis That the Three Populations Have the Same Average Blood Pressure Levels

1. **The hypothesis:** $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

2. **The assumptions:** Independent random samples normal distributions $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

3. **The α-level :** $\alpha = 0.05$

4. **The test statistic:** ANOVA

5. The Rejection Region: Reject H_0: $\mu_1 = \mu_2 = \mu_3$ if

$$F = \frac{\text{MS(Among)}}{\text{MS(Within)}} > F_{0.95(2,27)} = 3.35$$

where

$$\text{MS(Among)} = \frac{\text{SS(Among)}}{df\text{ (Among)}}, \quad \text{MS(Within)} = \frac{\text{SS(Within)}}{df\text{ (Within)}}$$

6. The Result:

ANOVA

Source	DF	SS	MS	F
Among	2	536.47	268.23	3.54
Within	27	2043.70	75.69	
Total	29	2580.17		

7. The Conclusion: Reject H_0: $\mu_1 = \mu_2 = \mu_3$, since $F = 3.54 > F_{0.95}(2, 27) = 3.35$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION