Data collection, Analysis and Inference

Subject Code: CPE-RPE,

May 2021,
SRM Univeristy-AP, Andhrapradesh

Lecture- 8: Large Sample Tests and Small Sample Tests: F-test

Aim: Interpret the F probability distribution as the number of groups and the sample size change.

- Conduct and interpret one-way ANOVA.

Why F-test ?

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups.

- For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water.
- A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing.
- A consumer looking for a new car might compare the average gas mileage of several models.

ANOVA

For hypothesis tests comparing averages between more than two groups, statisticians have developed a method called "Analysis of Variance" (ANOVA).

- We will study the simplest form of ANOVA called single factor or one-way ANOVA and the F distribution, used for one-way ANOVA,
- and also, the test of two variances.

One-Way ANOVA

This test is to determine *the existence of a statistically significant difference among* **several group means** by making use of variances.

There are five basic assumptions to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.

2. All samples are randomly selected and independent.

3. The populations are assumed to have equal standard deviations (or variances).

4. The factor is a categorical variable.

5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same.

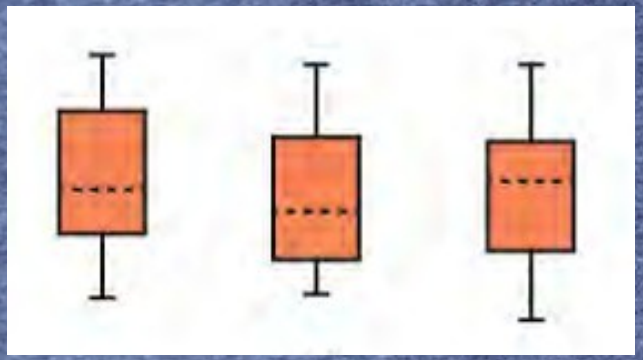
The alternative hypothesis is that at least one pair of means is different.

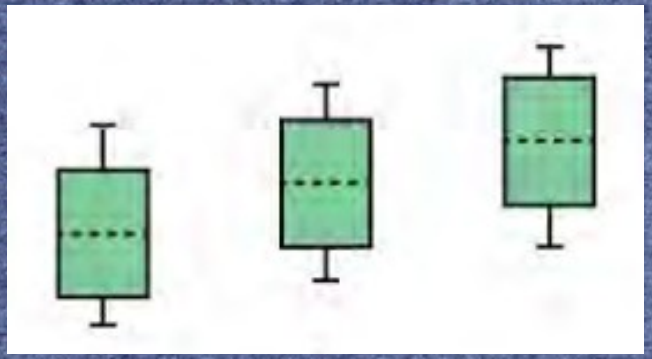
- For example, if there are k groups:

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

H_a : At least two of the group means μ_1 , μ_2 ,μ_3 , ..., μ_k are not equal.

- Box plots help in the understanding of the hypothesis test.The group means indicated by a horizontal line.
- In the first graph, the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.





If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means.

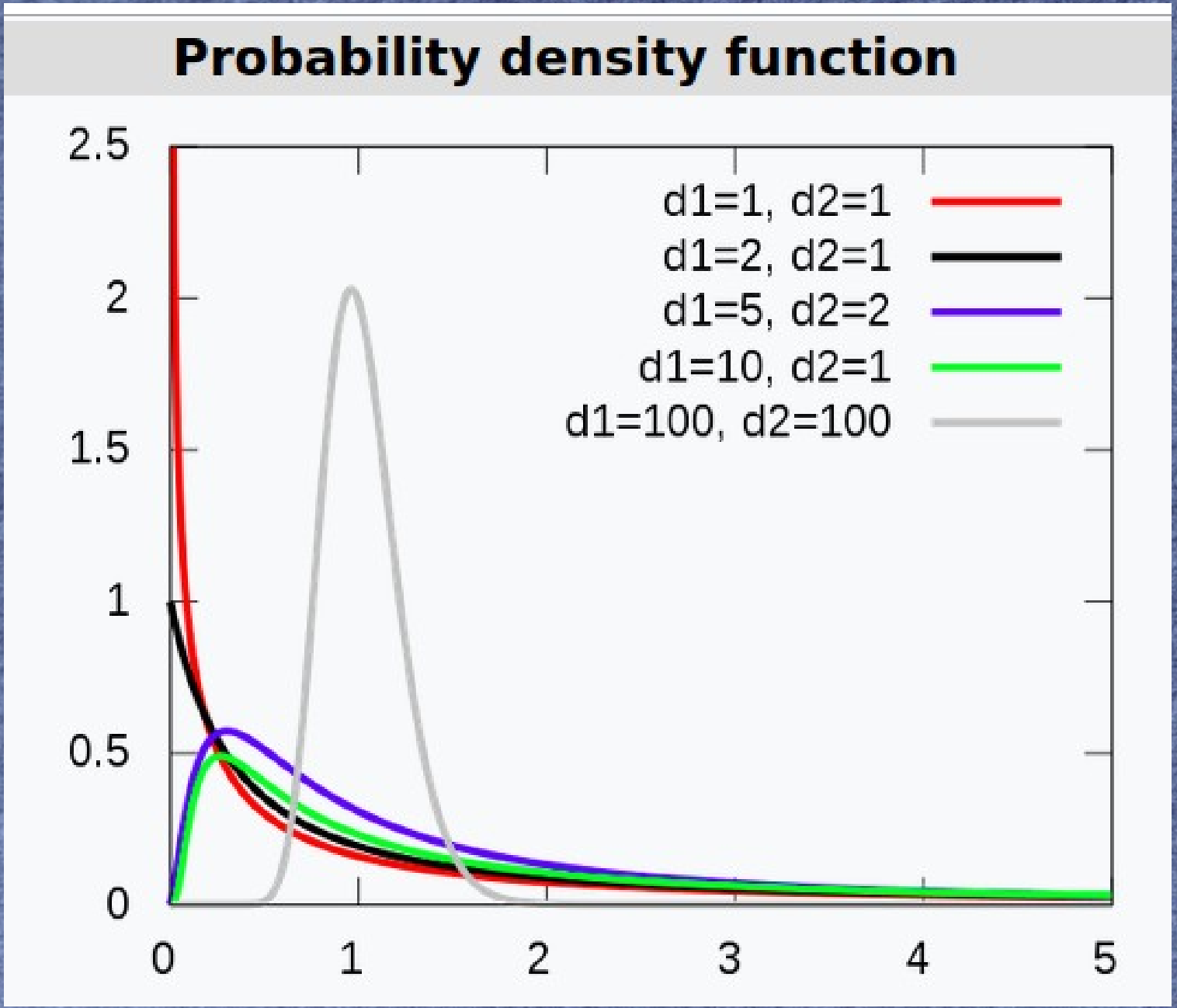
Fisher-Snedecor's F-Distribution

- If a random variable X has an F-distribution with parameters d_1 and d_2, we write $X \sim F(d_1, d_2)$ then the probability density function (pdf) for X is given by

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\ d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

for real $x > 0$. Here B is the beta function.

Did you observe the Normal ?



F-ratio, F-score, F-statistic

- The F statistic is a ratio. There are two sets of degrees of freedom; one for the numerator and one for the denominator.
- One-Way ANOVA expands the t-test for comparing more than two groups.

Note: The F distribution is derived from the Student's t-distribution. The values of the F distribution are squares of the corresponding values of the t-distribution.

Various Variations

To calculate the F-ratio, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.).

- If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes.
- The variance is also called *variation due to treatment or explained variation*.

2. Variance within samples: An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance).

- When the sample sizes are different, the variance within samples is weighted. The variance is also called *the variation due to error or unexplained variation.*

• $SS_{between}$ = the sum of squares that represents the variation among the different samples.

• SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total sample size: $\sum n_j$

Sum of squares of all values from every group combined: $\sum x^2$

Hence,

Between group variability is=SS_{total}= $\sum x^2 - \frac{(\sum x)^2}{n}$

Explained variation: sum of squares representing variation among the different samples:

$$SS_{between} = \sum \left[\frac{(sj)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n}$$

- Unexplained variation: sum of squares representing variation within samples due to chance:

$$SS_{within} = SS_{total} - SS_{between}$$

- df for different groups (for numerator): $df_{between} = k-1$

df for errors within samples (for denominator): $df_{within} = n-k$

Mean square (variance estimate) explained by the different groups:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$$

- Mean square (variance estimate) that is due to chance (unexplained):

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{n-k}$$

F-Ratio or F Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

Data and One-Way ANOVA results are typically put into a table for easy viewing.

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor)	$k - 1$	*MS*(Factor) = *SS*(Factor)/($k - 1$)	*F* = *MS*(Factor)/*MS*(Error)
Error (Within)	*SS*(Error)	$n - k$	*MS*(Error) = *SS*(Error)/($n - k$)	
Total	*SS*(Total)	$n - 1$		

Example: Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in Table.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

Sum of the values in each group :

$s_1 = 16.5, s_2 = 15, s_3 = 15.7$

Following are the calculations needed to fill in the one-way ANOVA table.

$$SS(between) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1 + s_2 + s_3)^2}{10}$$

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} - \frac{(16.5 + 15 + 15.5)^2}{10}$$

$$SS(between) = 2.2458$$

$$S(total) = \sum x^2 - \frac{(\sum x)^2}{n} = \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right)$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9$$

$$SS(total) = 23.1$$

SS(within) = SS(total) − SS(between) = 23.1 − 2.2458

= 20.8542

where $n_1 = 4$, $n_2 = 3$, $n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$

One-way ANOVA table

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	SS(Factor) = SS(Between) = 2.2458	$k - 1$ = 3 groups – 1 = 2	MS(Factor) = SS(Factor)/(k – 1) = 2.2458/2 = 1.1229	F = MS(Factor)/MS(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	SS(Error) = SS(Within) = 20.8542	$n - k$ = 10 total data – 3 groups = 7	MS(Error) = SS(Error)/(n – k) = 20.8542/7 = 2.9792	
Total	SS(Total) = 2.2458 + 20.8542 = 23.1	$n - 1$ = 10 total data – 1 = 9		

F-table

The F - Distribution with α = 0.05								
$\nu_2 \setminus \nu_1$	2	3	4	5	6	7	8	9
2	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18

With significance level β=0.05
The P-value=P($F_{2,7}$>0.3769) is Greater than 0.05

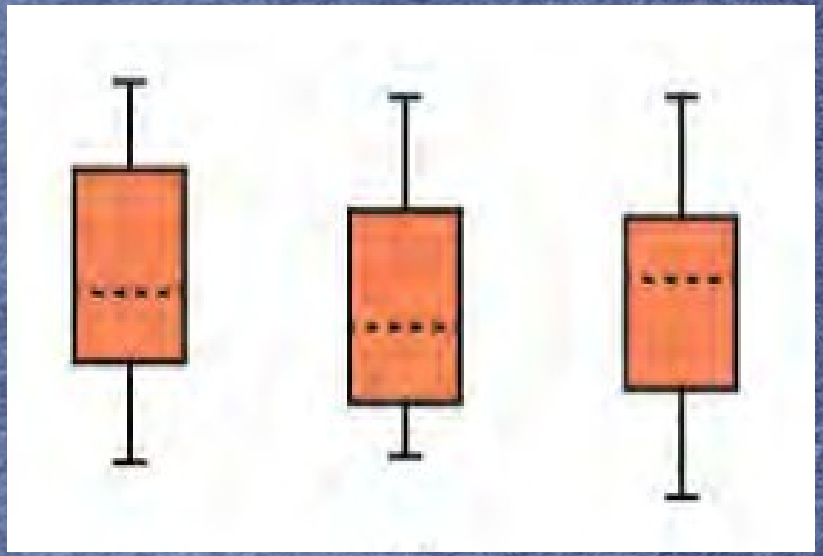
- Decision: Do not reject the Null hypothesis.

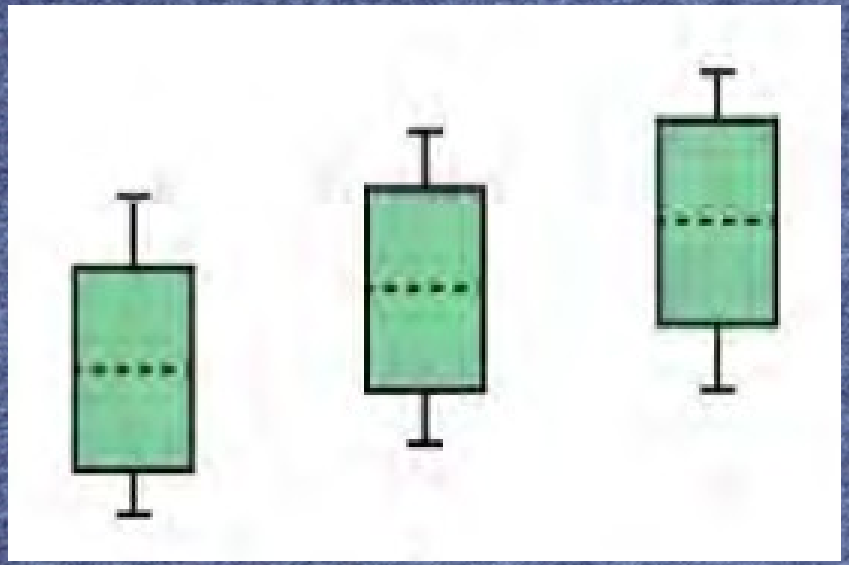
Conclusion: All group means are equal for β=0.05

- The one-way ANOVA test depends on the fact that $MS_{between}$ can be influenced by *population differences among means* of the several groups.
- Since MS_{within} compares values of each group to its own group mean, the fact that *group means might be different* does not affect MS_{within}.
- The null hypothesis says that all groups are samples from populations having the same normal dist.
- The alternate hypothesis says that at least two of the sample groups come from populations with different normal dist's.

An important remark

If the null hypothesis is true, $MS_{between}$ and MS_{within} should both estimate the same value.





- END