IIT ROORKEE

FREE ONLINE EDUCATION swayam

NPTEL ONLINE CERTIFICATION COURSE

ANOVA

Dr. A. Ramesh
DEPARTMENT OF MANAGEMENT STUDIES

Effect of Teaching Methodology

Group 1 Black Board	Group 2 Case Presentation	Group 3 PPT
4	2	2
3	4	1
2	6	3

ANOVA with Python

```
In [15]: a=[4,3,2]

In [16]: b=[2,4,6]

In [17]: c=[2,1,3]

In [18]: stats.f_oneway(a,b,c)

Out[18]: F_onewayResult(statistic=1.5, pvalue=0.2962962962962962)
```

Pandas.melt command

- Pd.melt allows you to ‘unpivot’ data from a ‘wide format’ into a ‘long format’, data with each row representing a data point.

Jupyter code

In [22]:
```
import pandas as pd
import numpy as np
import math
from scipy import stats
import scipy
import statsmodels.api as sm
from statsmodels.formula.api import ols
from matplotlib import pyplot as plt
```

In [23]:
```
data=pd.read_excel('oneway.xlsx')
```

In [24]:
```
data
```

Out[24]:

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3

```
In [26]: data_new=pd.melt(data.reset_index(),id_vars=['index'], value_vars=['Teachin Method1','Teachin Method2','Teachin Method3'])
         data_new.columns=['index','Treatments','value']

In [27]: data_new
```

Transforming table

```
In [27]: data_new
```

Out[27]:

	index	Treatments	value
0	0	Teachin Method1	4
1	1	Teachin Method1	3
2	2	Teachin Method1	2
3	0	Teachin Method2	2
4	1	Teachin Method2	4
5	2	Teachin Method2	6
6	0	Teachin Method3	2
7	1	Teachin Method3	1
8	2	Teachin Method3	3

```
4]: data
```

4]:

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3

In [31]:
```python
model=ols('value ~ C(Treatments)',data=data_new).fit()
```

In [32]:
```python
anova_table=sm.stats.anova_lm(model, typ=1)
```

In [33]:
```python
anova_table
```

Out[33]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Treatments)	2.0	6.0	3.0	1.5	0.296296
Residual	6.0	12.0	2.0	NaN	NaN

Analysis of Variance: A Conceptual Overview

- Analysis of Variance (ANOVA) can be used to test for the equality of three or more population means

- Data obtained from observational or experimental studies can be used for the analysis

- We want to use the sample results to test the following hypotheses:

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

H_a: Not all population means are equal

Analysis of Variance: A Conceptual Overview

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

H_a: Not all population means are equal

- If H_0 is rejected, we cannot conclude that all population means are equal
- Rejecting H_0 means that at least two population means have different values

Analysis of Variance: A Conceptual Overview

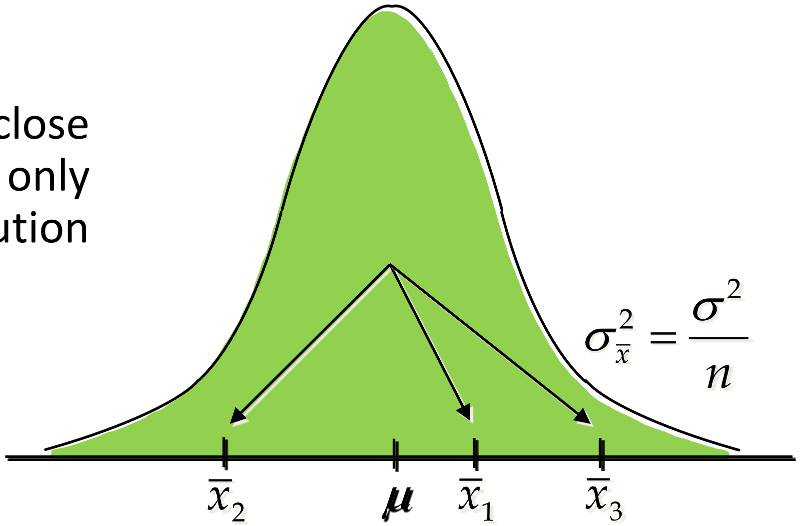
Assumptions for Analysis of Variance

- For each population, the response (dependent) variable is normally distributed
- The variance of the response variable, denoted σ^2, is the same for all of the populations
- The observations must be independent

Analysis of Variance: A Conceptual Overview

- Sampling Distribution of $\bar{x}$ Given H_0 is True

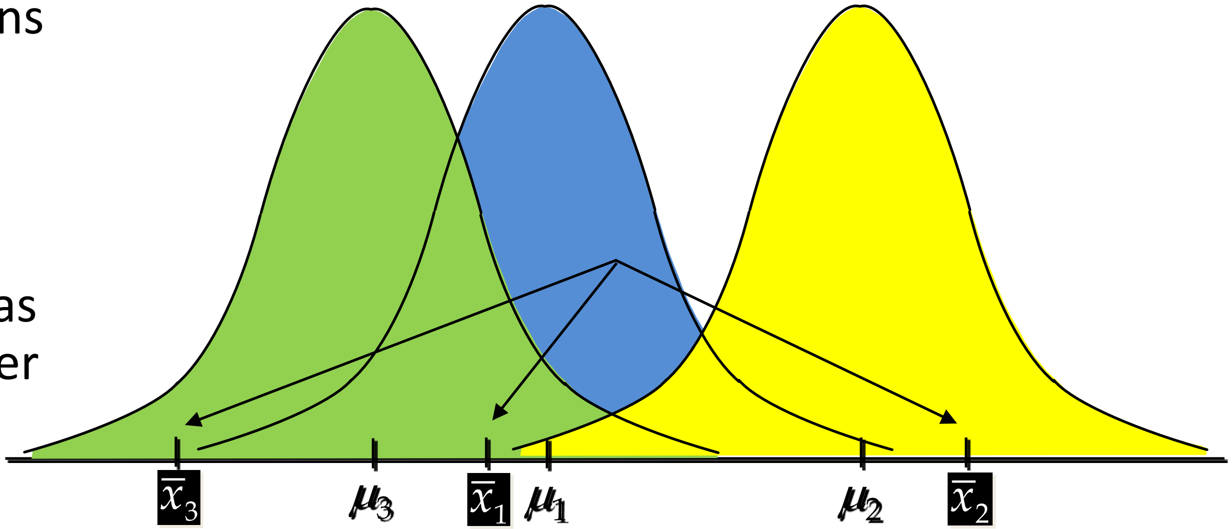
Sample means are close together because there is only one sampling distribution when H_0 is true.

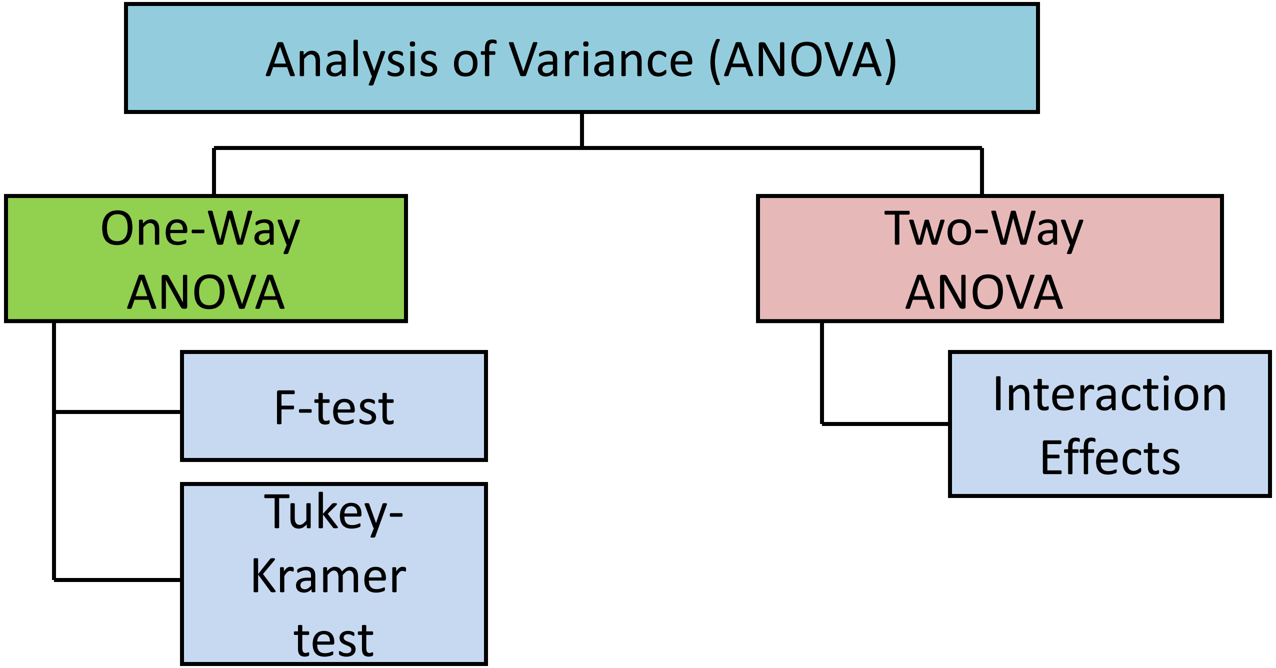


Analysis of Variance: A Conceptual Overview

- Sampling Distribution of $\bar{x}$ Given H_0 is False

Sample means come from different sampling distributions and are not as close together when H0 is false.




Analysis of Variance (ANOVA)
One-Way ANOVA
Two-Way ANOVA
F-test
Tukey-Kramer test
Interaction Effects

General ANOVA Setting

- Investigator controls one or more factors of interest
  - Each factor contains two or more levels
  - Levels can be numerical or categorical
  - Different levels produce different groups
  - Think of the groups as populations
- Observe effects on the dependent variable
  - Are the groups the same?
- Experimental design: the plan used to collect the data

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to the different levels (groups)
  - Subjects are assumed homogeneous
- Only one factor or independent variable
  - With two or more levels (groups)
- Analyzed by one-factor analysis of variance (one-way ANOVA)

Analysis of Variance and the Completely Randomized Design

- Between-Treatments Estimate of Population Variance
- Within-Treatments Estimate of Population Variance
- Comparing the Variance Estimates: The F Test
- ANOVA Table

Analysis of Variance and the Completely Randomized Design

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

H_a: Not all population means are equal

where

μ_j = mean of the j^{th} population

Analysis of Variance and the Completely Randomized Design

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

H_a: Not all population means are equal

- Assume that a simple random sample of size n_j has been selected from each of the *k* populations or treatments. For the resulting sample data, let

x_{ij} = value of observation *i* for treatment *j*

n_j = number of observations for treatment *j*

$\bar{x}_j$ = sample mean for treatment *j*

s_j^2 = sample variance for treatment *j*

s_j = sample standard deviation for treatment *j*

Between-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample means is called the mean square due to treatments and is denoted by MSTR

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

Denominator is the degrees of freedom associated with SSTR

Numerator is called the sum of squares due to treatments (SSTR)

Between-Treatments Estimate of Population Variance σ^2

- Mean Square due toTreatments (MSTR)

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

Where:

k = number of groups

n_j = sample size from group j

$\bar{x}_j$ = sample mean from group j

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Within-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample observations within each sample is called the <u>mean square error</u> and is denoted by <u>MSE</u>

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k}$$

Numerator is called the <u>sum of squares due to error</u> (SSE)

Denominator is the <u>degrees of freedom</u> associated with SSE

Within-Treatments Estimate of Population Variance σ^2

- Mean Square Error (MSE)

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k}$$

Where:

k = number of groups

n_j = number of observations for treatment *j*

s_j^2 = sample variance for treatment *j*

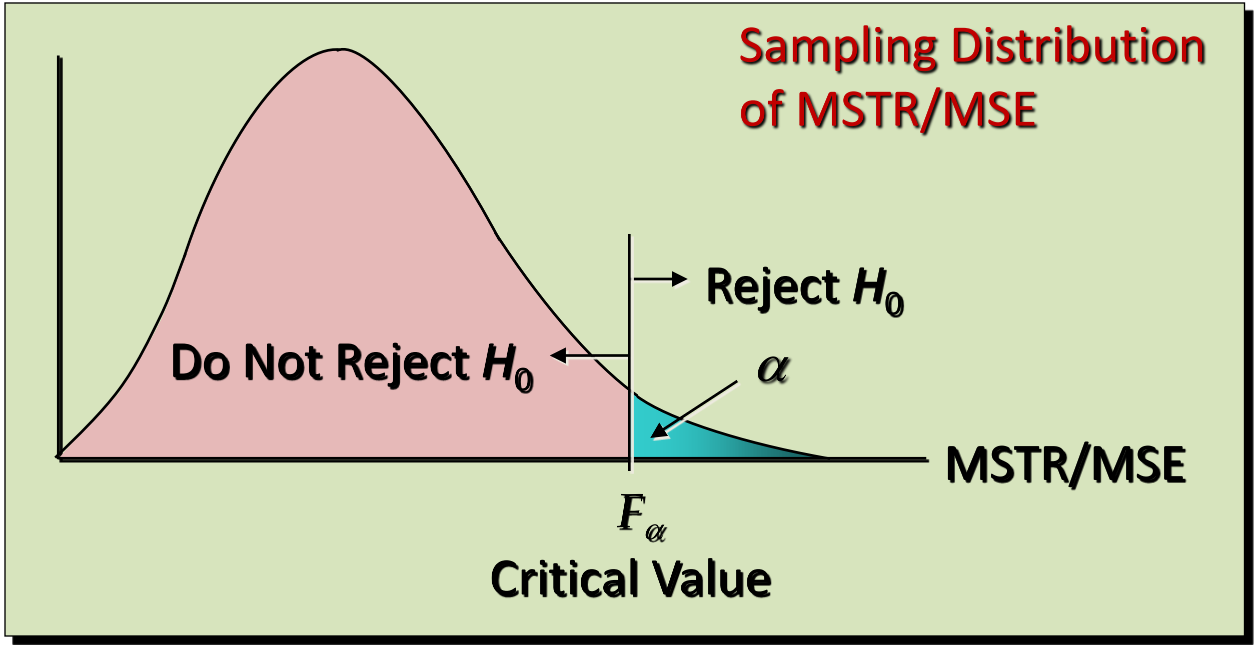
Comparing the Variance Estimates: The *F* Test

- If the null hypothesis is true and the ANOVA assumptions are valid, the sampling distribution of MSTR/MSE is an *F* distribution with MSTR d.f equal to k - 1 and MSE d.f. equal to $n_T - k$.

- If the means of the k populations are not equal, the value of MSTR/MSE will be inflated because MSTR overestimates σ^2

- Hence, we will reject H_0 if the resulting value of MSTR/MSE appears to be too large to have been selected at random from the appropriate *F* distribution

Comparing the Variance Estimates: The *F* Test



ANOVA Table for a Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-Value
Treatments	SSTR	k - 1	$\text{MSTR} = \dfrac{\text{SSTR}}{k\text{-}1}$	$\dfrac{\text{MSTR}}{\text{MSE}}$	
Error	SSE	n_T - k	$\text{MSE} = \dfrac{\text{SSE}}{n_T\text{-}k}$		
Total	SST	n_T - 1			

SST is partitioned into SSTR and SSE.

SST's degrees of freedom (d.f.) are partitioned into SSTR's d.f. and SSE's d.f.

ANOVA Table for a Completely Randomized Design

- SST divided by its degrees of freedom $n_T - 1$ is the overall sample variance that would be obtained if we treated the entire set of observations as one data set.
- With the entire data set as one sample, the formula for computing the total sum of squares, SST, is:

$$\text{SST} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 = \text{SSTR} + \text{SSE}$$

ANOVA Table for a Completely Randomized Design

- ANOVA can be viewed as the process of partitioning the total sum of squares and the degrees of freedom into their corresponding sources: treatments and error

- Dividing the sum of squares by the appropriate degrees of freedom provides the variance estimates and the *F* value used to test the hypothesis of equal population means.

Test for the Equality of k Population Means

- Hypotheses

$$H_0\!: \ \mu_1 = \mu_2 = \mu_3 = \cdot \ \cdot \ \cdot = \mu_\kappa$$

H_a: Not all population means are equal

- Test Statistic

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

Test for the Equality of k Population Means

p- Value Approach

Reject H_0 if p-value $\leq \alpha$

Critical Value Approach

Reject H_0 if $F \geq F_\alpha$

Where the value of F_α is based on an F distribution with k - 1 numerator d.f. and n_T - k denominator d.f.

Thank You