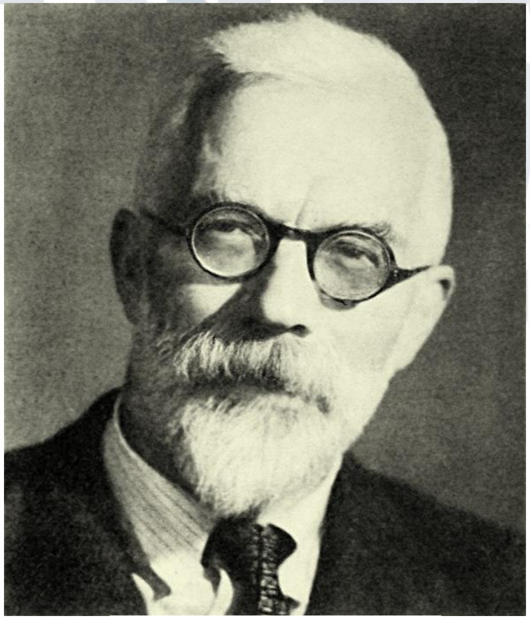
Analysis of Variance

What is ANOVA?

- **ANOVA** (ANalysis Of VAriance) is a statistical method for testing the equality of several population means.
  - ANOVA is designed to detect differences among means from populations subject to different groups often called as *treatments*
  - ANOVA tests for the equality of several population means by calculating and analyzing the two estimators of the population variance. Hence, the name *analysis of variance*.
- This technique was developed by Statistician Prof. Ronald Fisher



Ronald Fisher

One-Way

ANOVA

1-way ANOVA Model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Overall Effect

Treatment Effect

Error

- In 1-way, we think of any observation value(univariate) to be comprised of
  - An overall effect
  - Group or treatment effect
  - Error

Example

- Consider an agricultural experiment, in which we check the yield of a crop planted on a plot of land.
- Suppose that we divide the plot in 4 parts in the interest of applying 4 different treatments (fertilizers) to the parts.
- In the four parts, suppose that we are able to plant 6, 7, 5 and 6 plants respectively.

I	II	III	IV
y_{11}	y_{21}	y_{31}	y_{41}
y_{12}	y_{22}	y_{32}	y_{42}
y_{13}	y_{23}	y_{33}	y_{43}
y_{14}	y_{24}	y_{34}	y_{44}
y_{15}	y_{25}	y_{35}	y_{45}
y_{16}	y_{26}		y_{46}
	y_{27}		

$where\ ,\ y_{ij}$: yield (kg) of j^{th} plant from i^{th} part of the plot

Sane's STATS Academy of Statistics

Example

- After a certain period (an year), we note down the yields of all the plants as follows:

I	II	III	IV
23.4	34.2	23.8	36.7
24.1	45.2	24.5	39.5
19.6	24.9	29.3	43.2
23.9	40.3	18.3	50.2
29.4	39.4	19.4	47.2
21.9	35.3		34.1
	38.4		

Statements of Hypothesis

- The hypothesis test of analysis of variance:

  H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$

  H_1: Not all μ_i (i = 1, ..., r) are equal

- In our example,

  H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

  H_1: Not all μ_i (i = 1,2,3,4) are equal

Hypothesis Test of ANOVA

- In an analysis of variance:
  - We have r independent random samples, each one corresponding to a population subject to a different treatment.
  - We have:
    - $n = n_1 + n_2 + n_3 + \ldots + n_r$ total observations.
    - r sample means: $x_1, x_2, x_3, \ldots, x_r$
    - r sample variances: $s_1^2, s_2^2, s_3^2, \ldots, s_r^2$
      - These sample variances can be used to find a pooled estimator of the population variance.

Result of ANOVA

Sources of Variation	Sums of Squares	Degrees of freedom	Mean Square	F Ratio	P-Value
Treatment	SSTR	r – 1	MSTR=SSTR / (r – 1)	MSTR/MSE	
Error	SSE	n – r	MSE = SSE / (n – r)		
Total	SST	n – 1			

$$SSTR = \sum_i \frac{(\sum_j y_{ij})^2}{n_i} - \frac{(\sum_j \sum_i y_{ij})^2}{n}$$

$$SSE = \sum_j \sum_i y_{ij}^2 - \sum_i \frac{\left(\sum_j y_{ij}\right)^2}{n_i}$$

$$SST = \sum_j \sum_i y_{ij}^2 - \frac{(\sum_j \sum_i y_{ij})^2}{n}$$

Example

I	II	III	IV
23.4	34.2	23.8	36.7
24.1	45.2	24.5	39.5
19.6	24.9	29.3	43.2
23.9	40.3	18.3	50.2
29.4	39.4	19.4	47.2
21.9	35.3		34.1
	38.4		

- In our example, r = 4 , n = 6+7+5+6 = 24
- Our Python, function anova_lm() calculates not only the means and variances but also all the sums of squares

ANOVA in Python

Syntax :

anova_lm**(***args**,** ****kwargs**)**

Where

args : fitted linear model results instance

One or more fitted linear models

scale : float

Estimate of variance, If None, will be estimated from the largest model. Default is None.

test : str {"F", "Chisq", "Cp"} or None

Test statistics to provide. Default is "F".

typ : str or int {"I","II","III"} or {1,2,3}

The type of ANOVA test to perform.

R Program and Output

```
In [39]: import pandas as pd
    ...:
    ...: from statsmodels.stats.anova import anova_lm
    ...: from statsmodels.formula.api import ols
    ...: ######################Example 1#################################
    ...: agr = pd.read_csv("G:/Statistics (Python)/Datasets/Yield.csv")
    ...: agrYield = ols('Yield ~ Treatments', data=agr).fit()
    ...: table = anova_lm(agrYield, typ=2)
    ...: print(table)
                 sum_sq    df          F    PR(>F)
Treatments  1551.607762   3.0  18.293252  0.000006
Residual     565.457238  20.0        NaN       NaN
```

As p-value < 0.01, we can reject H_0 at 1% level of significance. Hence, we conclude that the yields are significantly different for all the 4 treatments.

Assumptions

- We assume *independent random sampling* from each of the *r* populations
- We assume that the *r* populations under study:
  - are *normally distributed*,
  - with means μ_i that may or may not be equal,
  - but with *equal variances*, σ_i^2.

Statements of Hypothesis

- The hypothesis test of analysis of variance:

  H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$

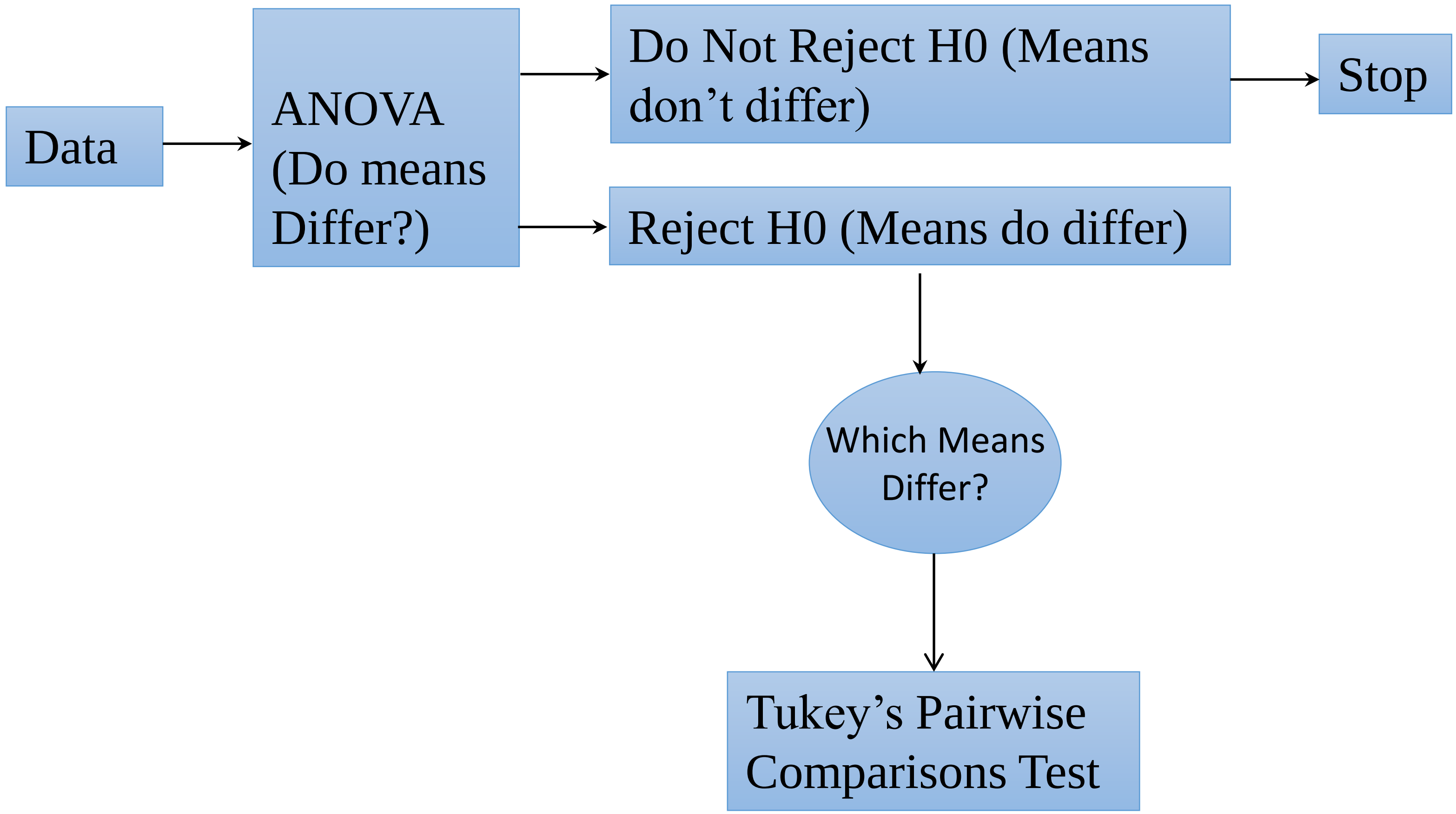
  H_1: Not all μ_i (i = 1, ..., r) are equal

- In our example,

  H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

  H_1: Not all μ_i (i = 1,2,3,4) are equal

Further Analysis



Tukey's Test in Python

In [9]:

```
######## Post Hoc Tukey HSD ################
from statsmodels.stats.multicomp import pairwise_tukeyhsd

compare = pairwise_tukeyhsd(agr.Yield, agr.Treatments, alpha=0.05)
pd.DataFrame(compare._results_table.data)
```

Out[9]:

	0	1	2	3	4	5
0	group1	group2	meandiff	lower	upper	reject
1	I	II	13.0976	4.8174	21.3779	True
2	I	III	-0.6567	-9.6689	8.3556	False
3	I	IV	18.1	9.5072	26.6928	True
4	II	III	-13.7543	-22.469	-5.0396	True
5	II	IV	5.0024	-3.2779	13.2826	False
6	III	IV	18.7567	9.7444	27.7689	True

The p-values for pair-wise comparisons namely II & I , IV & I , III & II , IV & III indicate that they have significant differences.

Further Studies

- Two way ANOVA : The way we analyzed the effect of one factor variable, we can also analyze the effects of two factor variables with interaction or without interactions.
- The design we saw is called Completely Randomized Design
- There are also following designs in this field of study of statistics:
    - Factorial Design
    - Lattice Design
    - Split Plot Design
    - Repeated Measures Design
    - Multivariate Analysis of Variance

Case: Funds

- *A magazine reports percentage returns and expense ratios for stock and bond funds.* The data FUNDS.csv are the expense ratios for 10 midcap stock funds, 10 small-cap stock funds, 10 hybrid stock funds, and 10 specialty stock funds.
- Test for any significant difference in the mean expense ratio among the four types of stock funds.

Sane's STATS Academy of Statistics

Thank You