ANOVA - Analysis of Variance

So far, we compared two sets of samples, or two groups

Let us develop an intuitive way of comparing across multiple groups

Imagine we have data of heights and weights of three different groups

Our goal is to say whether these three groups have statistically the same height/weight

ANOVA - Analysis of Variance

Setup 1

American Basketball players — Very low variance within this group

Indonesian college students — Very low variance within this group

Indian cricket team — Maybe not too low

Setup 2

Suppose we take all these three groups and sort their names alphabetically

Names from A to G

Names from H to N

Names from O to Z

Which setup will have higher F-ratio?

Setup 1 will have higher F-ratio

If there is a difference, then F-ratio will be high.

If there is no difference, then F-ratio will be small.

$$\text{F-ratio} = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

H_0 : all groups have same mean

Under H_0, F-ratio will be very low

If F-ratio is high, we reject H_0

iPhone sales in 3 stores

	A	B	C	
	25	30	18	
	25	30	30	
	27	25	29	
	30	24	29	
	23	26	24	
	20	28	26	
	25	26.5	26	25.83
	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}$

$$F = \frac{3.49}{14.9} = 0.23$$

$$F = \frac{MSB}{MSW}$$

H_0: All means are equal H_a: Means are different

Step 1 Compute individual group means $\bar{Y}_1 = 25$ $\bar{Y}_2 = 26.5$ $\bar{Y}_3 = 26.5$

Step 2 Compute mean of these 3 values $\bar{Y} = \frac{25 + 26.5 + 26}{3} = 25.83$

Step 3 Between groups

$$SSB = 6(25 - 25.83)^2 + 6(26.5 - 25.83)^2 + 6(26 - 25.83)^2 = 6.9$$

$$DF = 3 - 1 = 2$$

$$MSB = \frac{SSB}{DF} = \frac{6.9}{2} = 3.49$$

Step 4 Within groups

$$SSW = (25 - 25)^2 + (25 - 25)^2 + (27 - 25)^2 + \cdots + (20 - 25)^2$$
$$+$$
$$(30 - 26.5)^2 + (30 - 26.5)^2 + (25 - 26.5)^2 + \cdots + (28 - 26.5)^2$$
$$+$$
$$(18 - 26)^2 + (30 - 26)^2 + (29 - 26)^2 + \cdots + (26 - 26)^2$$

$$= 223$$

$$DF = 18 - 3 = 15$$

$$MSW = \frac{SSW}{DF} = \frac{223}{15} = 14.9$$

iPhone sales in 3 stores

	A	B	C	
	25	30	18	
	25	30	30	
	27	25	29	
	30	24	29	
	23	26	24	
	20	28	26	
	25	26.5	26	25.83
	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}$

$$F = \frac{3.49}{14.9} = 0.23$$

$$F = \frac{MSB}{MSW}$$

H_0: All means are equal H_a: Means are different

Critical region for 95% confidence

```
from scipy.stats import f
cr = f.ppf(0.95, dfn=2, dfd=15)
cr = 3.68
```

Fail to reject H_0 since observed F statistic 0.23 is less than 3.68

$\alpha = 0.05$

```
from scipy.stats import f_oneway
a = [25, 25, 27, 30, 23, 20]
b = [30, 30, 21, 24, 26, 28]
c = [18, 30, 29, 29, 24, 26]
f_stat, p_value = f_oneway(a,b,c)
f_stat = 0.234
p_value = 0.793
```

p_value > 0.1

Assumptions of ANOVA

Normality, independent, equal variances

Normality – that each sample is taken from a normally distributed population (Gaussian)

Independence - each sample is drawn independently of the other samples

Equal variance of data in different groups

When assumptions of ANOVA don't hold, we use the Kruskal Wallis test

```
from scipy.stats import f_oneway
a = [25, 25, 27, 30, 23, 20]
b = [30, 30, 21, 24, 26, 28]
c = [18, 30, 29, 29, 24, 26]
f_stat, p_value = f_oneway(a,b,c)
f_stat = 0.234
p_value = 0.793
```

```
from scipy.stats import kruskal
  a = [25, 25, 27, 30, 23, 20]
  b = [30, 30, 21, 24, 26, 28]
  c = [18, 30, 29, 29, 24, 26]
  kruskal_stat, p_value = kruskal(a, b, c)
  kruskal_stat = 0.679
  p_value = 0.711
```

Online Vs Offline shopping Does gender effect this?

Observed

	Male	Female		
Offline	527	72	599	66%
Online	206	102	308	34%
	733	174	907	

Expected

	Male	Female	
Offline	484	115	599
Online	249	59	308
	733	174	907

All these are observed values

To compute χ^2 test statistic, what do we need? The expected values

What percent people prefer offline? 66%

Among 733 males, how many are expected to prefer offline? 733 * 0.66 = 484

Among 174 females, how many are expected to prefer offline? 174 * 0.66 = 115

What percent people prefer online? 34%

Among 733 males, how many are expected to prefer online? 733 * 0.34 = 249

Among 174 females, how many are expected to prefer online? 174 * 0.34 = 59

Assumptions of Chi2 test

Variables are categorical

Observations are independent

Each cell is mutually exclusive

Expected value in each cell is greater than 5 (at least in 80% of cells)