VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CSE3506 Essentials of Data Analytics (2 0 2 4 4)

B.Tech. Computer Science and Engineering

Winter 22-23

[SCOPE]

Regression Modelling

Model Estimation and Evaluation

- We assume that the true relationship between X and Y takes the form **Y = f(X) + ε** for some unknown function f, where ε is a mean-zero random error term
- If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- ✓ β_0 is the intercept, that is the expected value of Y when X = 0
- ✓ β_1 is the slope—the average increase in Y associated with a one-unit increase in X
- ✓ ε the error term is independent of X

Model Estimation and Evaluation

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- This model defines the **population regression line** which is the best linear approximation to the true relationship between X and Y
- **Population mean = μ** which is unknown

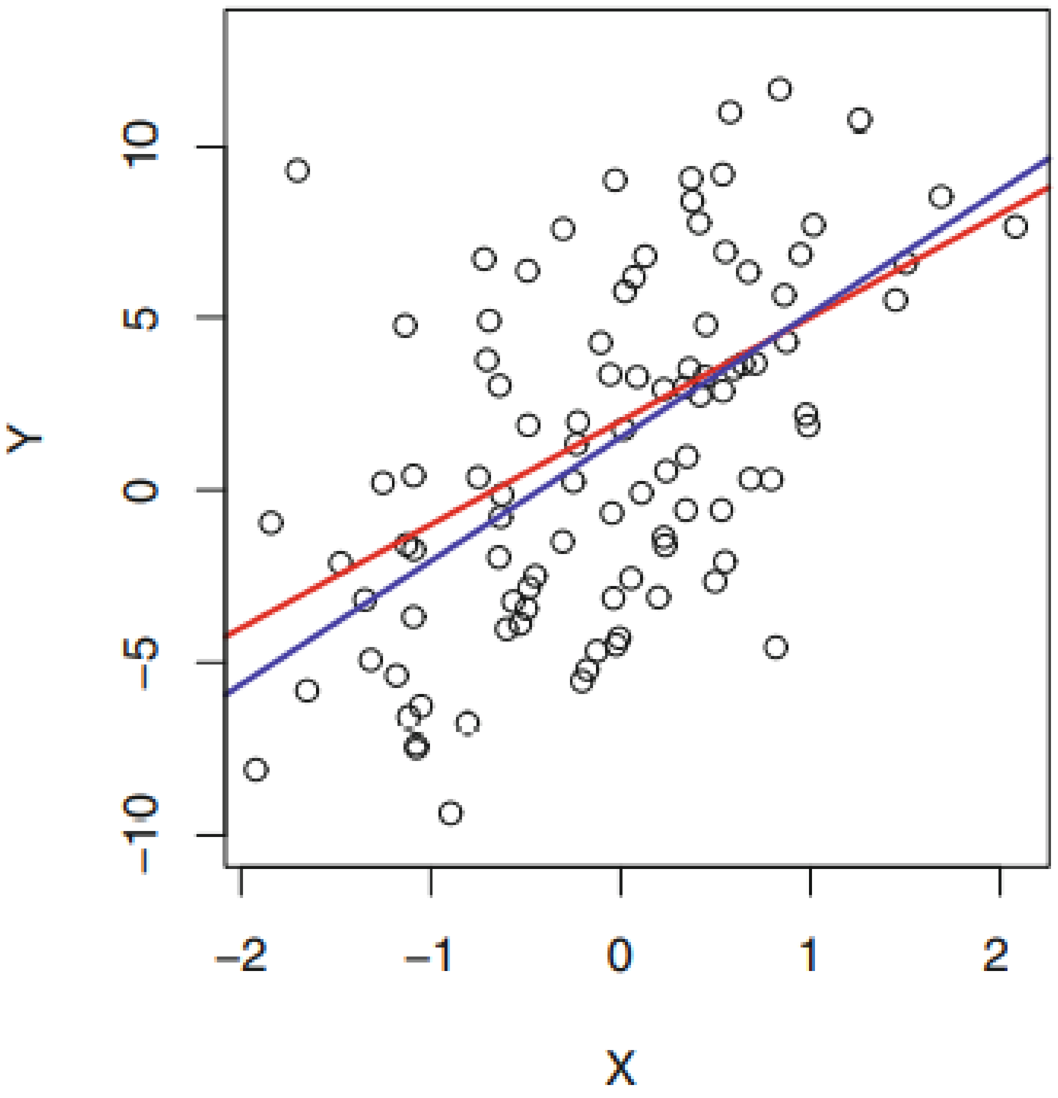
Model Estimation and Evaluation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$\hat{\mu} = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean

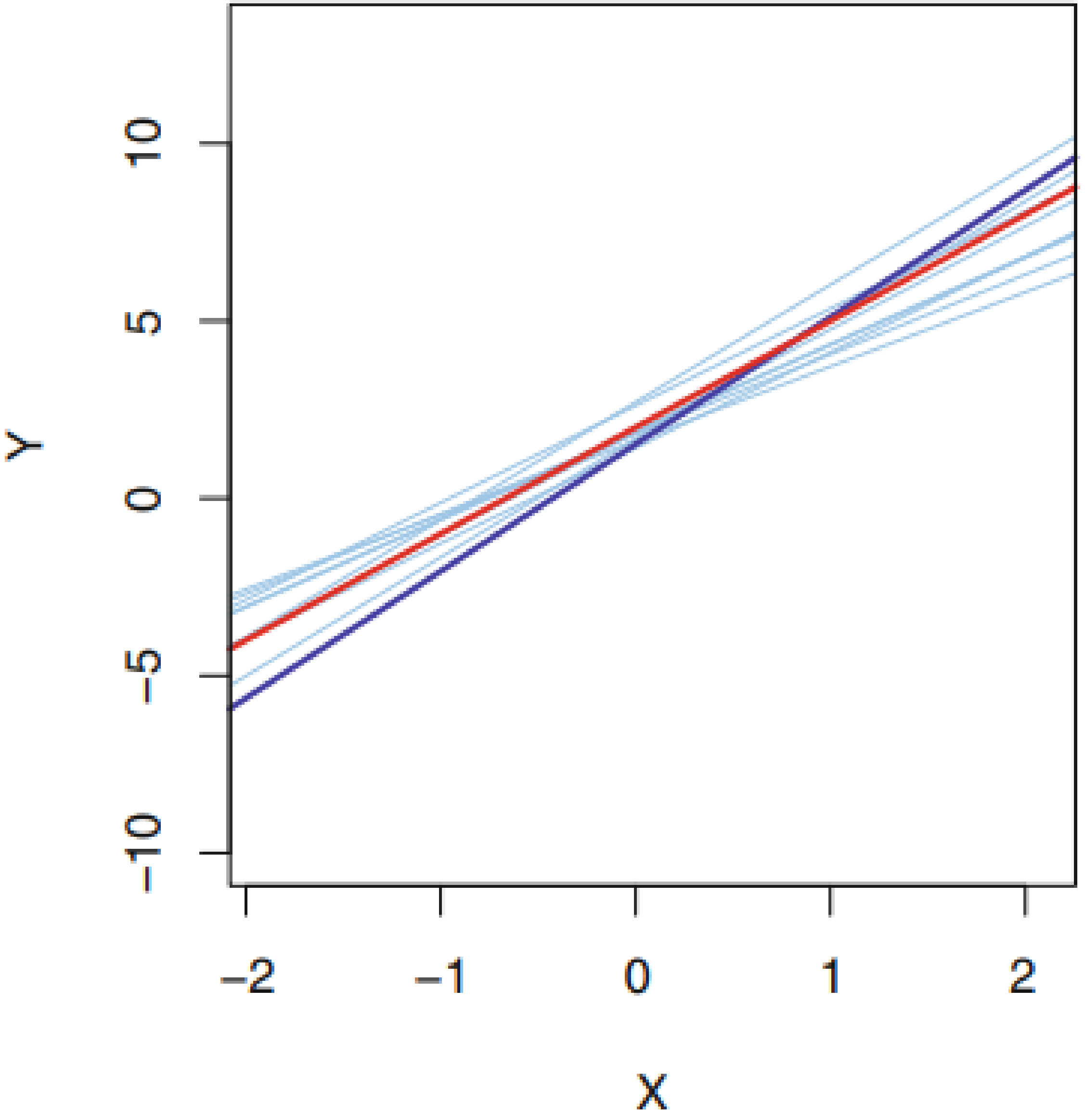
- This model defines the **least squares line** estimated from least square coefficients
- In real applications, set of observations is used to compute the least squares line
- **Sample mean =** $\widehat{\mu}$
- The sample mean and the population mean are different, but in general the sample mean will provide a good estimate of the population mean

Model Estimation and Evaluation



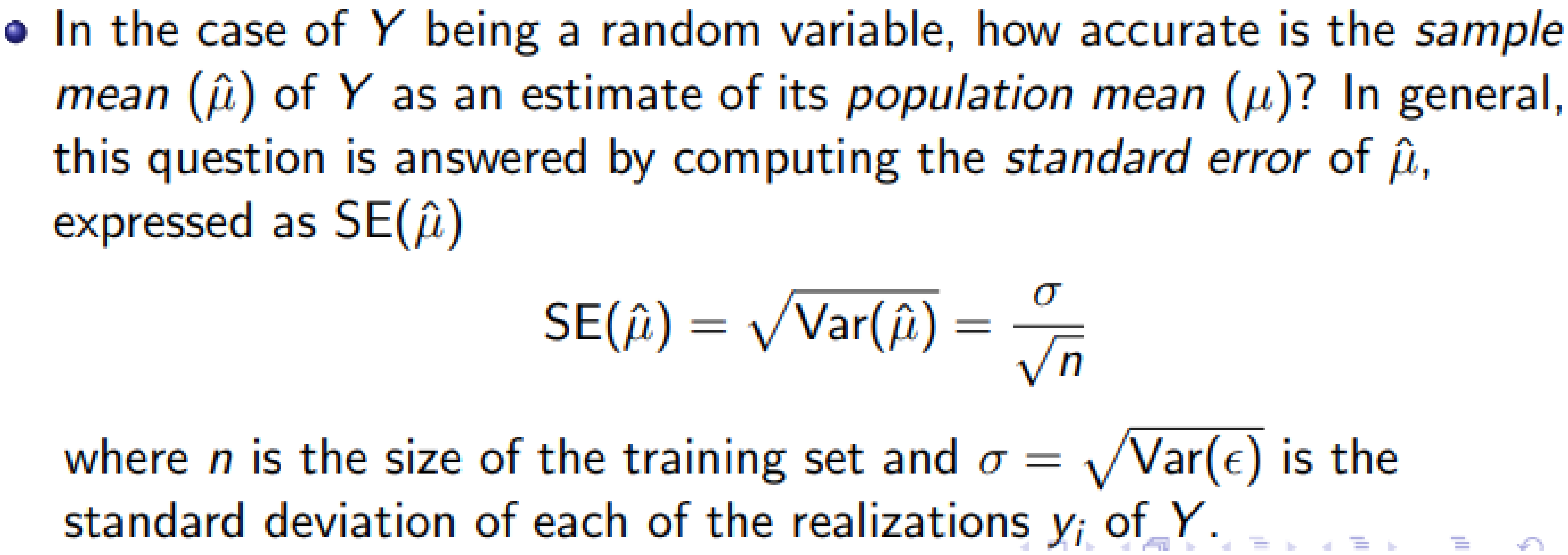
- Figure shows a simulated data set
- The **red line** represents the true relationship, **f(X)=2+3X,** which is known as the **population regression line**
- The **blue line** is the **least squares line**; it is the least squares estimate for f(X) based on the observed data, shown in black

Model Estimation and Evaluation



- Figure shows a simulated data set
- The population regression line is shown in red, and the least squares line in dark blue
- In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations from f(X)=2+3X + ε
- Each least squares line is different, but on average, the least squares lines are quite close to the population regression line

Model Estimation and Evaluation



- In the case of Y being a random variable, how accurate is the *sample mean* ($\hat{\mu}$) of Y as an estimate of its *population mean* (μ)? In general, this question is answered by computing the *standard error* of $\hat{\mu}$, expressed as $\mathrm{SE}(\hat{\mu})$

$$\mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$$

where n is the size of the training set and $\sigma = \sqrt{\mathrm{Var}(\epsilon)}$ is the standard deviation of each of the realizations y_i of Y.

- The standard error tells us the average amount that this estimate $\widehat{\mu}$ differs from the actual value of μ. The standard error equation tells us how this deviation shrinks with n – the more observations we have, the smaller the standard error of $\widehat{\mu}$

Model Estimation and Evaluation

- Assuming the errors ϵ_i for each observation are uncorrelated with common variance σ^2, the *standard errors* associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as

$$\mathrm{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$\mathrm{SE}(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- In general, $\sigma = \sqrt{\mathrm{Var}(\epsilon)}$ is not known, but can be estimated from the data. This estimate is known as the *residual standard error* (RSE), and is expressed as

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}}.$$

Model Estimation and Evaluation

- Standard errors can be used to compute **confidence intervals**
- A 95% confidence interval is defined as a range of values such that with 95% interval probability, the range will contain the true unknown value of the parameter
- The range is defined in terms of lower and upper limits computed from the sample of data

Model Estimation and Evaluation

- For linear regression, the 95% confidence interval for β_0 approximately takes the form

$$\hat{\beta}_0 \pm 2\,\mathrm{SE}(\hat{\beta}_0).$$

- That is, there is approximately a 95 % chance that the interval

$$[\hat{\beta}_0 - 2\,\mathrm{SE}(\hat{\beta}_0)\ ,\ \hat{\beta}_0 + 2\,\mathrm{SE}(\hat{\beta}_0)]$$

will contain the true value of β_0

Model Estimation and Evaluation

- Similarly, a confidence interval for β_1 approximately takes the form

$$[\hat{\beta}_1 - 2\,\mathrm{SE}(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + 2\,\mathrm{SE}(\hat{\beta}_1)]$$

will contain the true value of β_1

- The word 'approximately' is included mainly because
  - ✓ The errors are assumed to be Gaussian and
  - ✓ The factor '2' in front of $\mathrm{SE}(\hat{\beta}_1)$ term will vary slightly depending on the number of observations 'n' in the linear regression

Model Estimation and Evaluation

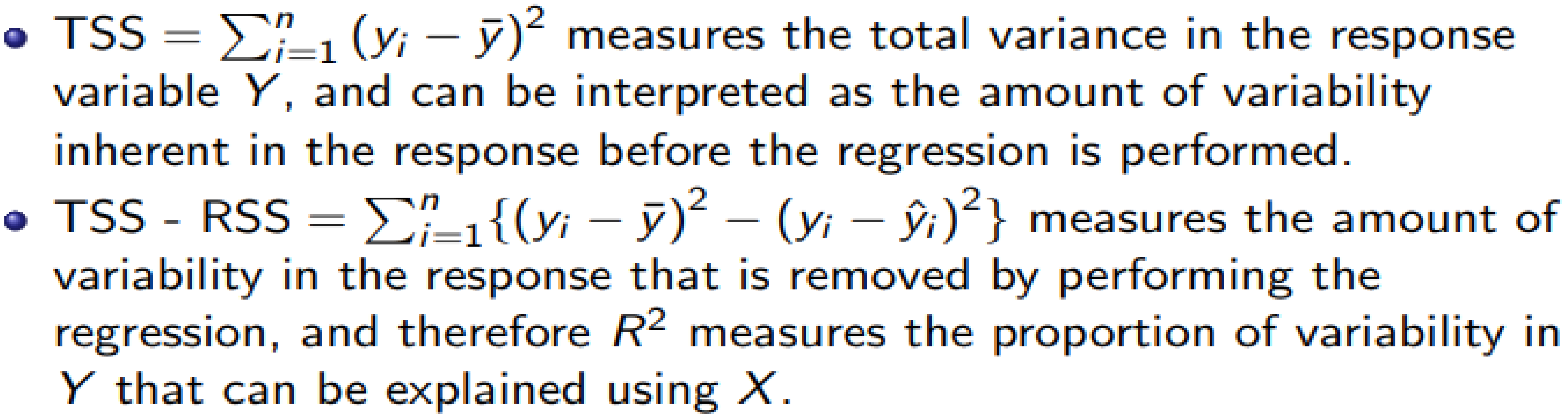
- The RSE provides an absolute measure of lack of fit of the model to the data. A small RSE indicates that the model fits the data well whereas a large RSE indicates that the model doesn’t fit the data well. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE
- The **R^2 statistic** provides an alternative measure of fit. It takes the form of a proportion of variance, expressed as

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares*.

- Note that R^2 statistic is independent of the scale of Y, and it always **takes a value between 0 and 1**

Model Estimation and Evaluation



- TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ measures the total variance in the response variable Y, and can be interpreted as the amount of variability inherent in the response before the regression is performed.
- TSS - RSS $= \sum_{i=1}^{n}\{(y_i - \bar{y})^2 - (y_i - \hat{y}_i)^2\}$ measures the amount of variability in the response that is removed by performing the regression, and therefore R^2 measures the proportion of variability in Y that can be explained using X.

- An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been taken care by the regression.
- The **R^2 statistic** is also a measure of the linear relationship between X and Y and it is closely related to **correlation between X and Y**

Model Estimation and Evaluation

Question-3:

Consider the following five training examples

$$X = [2\ 3\ 4\ 5\ 6]$$

$$Y = [12.8978\ 17.7586\ 23.3192\ 28.3129\ 32.1351]$$

We want to learn a function f(x) of the form f(x) = ax + b which is parameterized by (a, b).

(a) Find the best linear fit

(b) Evaluate the standard errors associated with $\hat{a}$ and $\hat{b}$.

(c) Determine the 95% confidence interval for a and b

(d) Compute R^2 statistic

Simple Linear Regression

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$
	2	12.8978	-2	-9.9869	19.9738	4
	3	17.7586	-1	-5.1261	5.1261	1
	4	23.3192	0	0.4345	0.0000	0
	5	28.3129	1	5.4282	5.4282	1
	6	32.1351	2	9.2504	18.5008	4
Sum	20	114.4236	0	0.0000	49.0289	10
Mean	4	22.88472				

The best linear fit is

Y = 3.2732 + 4.9029X

Substituting in the formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \mathbf{4.9029}$$

$$\hat{\beta}_0 = \mathbf{3.2732}$$

Simple Linear Regression

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
	2	12.8978	-2	-9.9869	19.9738	4	13.0789	0.0328
	3	17.7586	-1	-5.1261	5.1261	1	17.9818	0.0498
	4	23.3192	0	0.4345	0.0000	0	22.8847	0.1888
	5	28.3129	1	5.4282	5.4282	1	27.7876	0.2759
	6	32.1351	2	9.2504	18.5008	4	32.6905	0.3085
Sum	20	114.4236	0	0.0000	49.0289	10	**RSS**	**0.8558**
Mean	4	22.88472						

Y predicted is calculated using the best linear fit

Y = 4.9029 + 3.2732 X

RSS_{min} = 0.8558

Model Estimation and Evaluation

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}.$$

Substituting **RSS = 0.8558** and n = 5, then **RSE = 0.5341**.

Standard error for a is

σ = RSE

$$\text{SE(a)} = \sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \quad = 0.1689$$

Standard error for b is

$$\text{SE(b)} = \sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad = 0.7186$$

Model Estimation and Evaluation

95% confidence interval for standard error for a is

$$[a - 2\ SE(a)\ ,\ a + 2\ SE(a)] = [4.5651, 5.2407]$$

95% confidence interval for standard error for b is

$$[b - 2\ SE(b)\ ,\ b + 2\ SE(b)] = [1.8400, 4.7063]$$

Model Estimation and Evaluation

	X	Y	$(Y-Y_{mean})^2$
	2	12.8978	99.73857
	3	17.7586	26.27711
	4	23.3192	0.188773
	5	28.3129	29.46514
	6	32.1351	85.56953
Sum	20	114.4236	241.2391
Mean	4	22.88472	

To find R^2 value, first find TSS

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad = 241.2391$$

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \quad = 0.9965$$

Correlation

Correlation

- A correlation is a relationship between two variables.
- Is there a relationship between the number of employee training hours and the number of jobs produced?
- Is there a relationship between the number of hours a student spends studying for a Mathematics test and the student's score on that test?
- Let x to be the independent variable and y to be the dependent variable. Data is represented by a collection of ordered pairs (x, y)
- Mathematically, the strength and direction of a linear relationship between two variables is represented by the correlation coefficient.

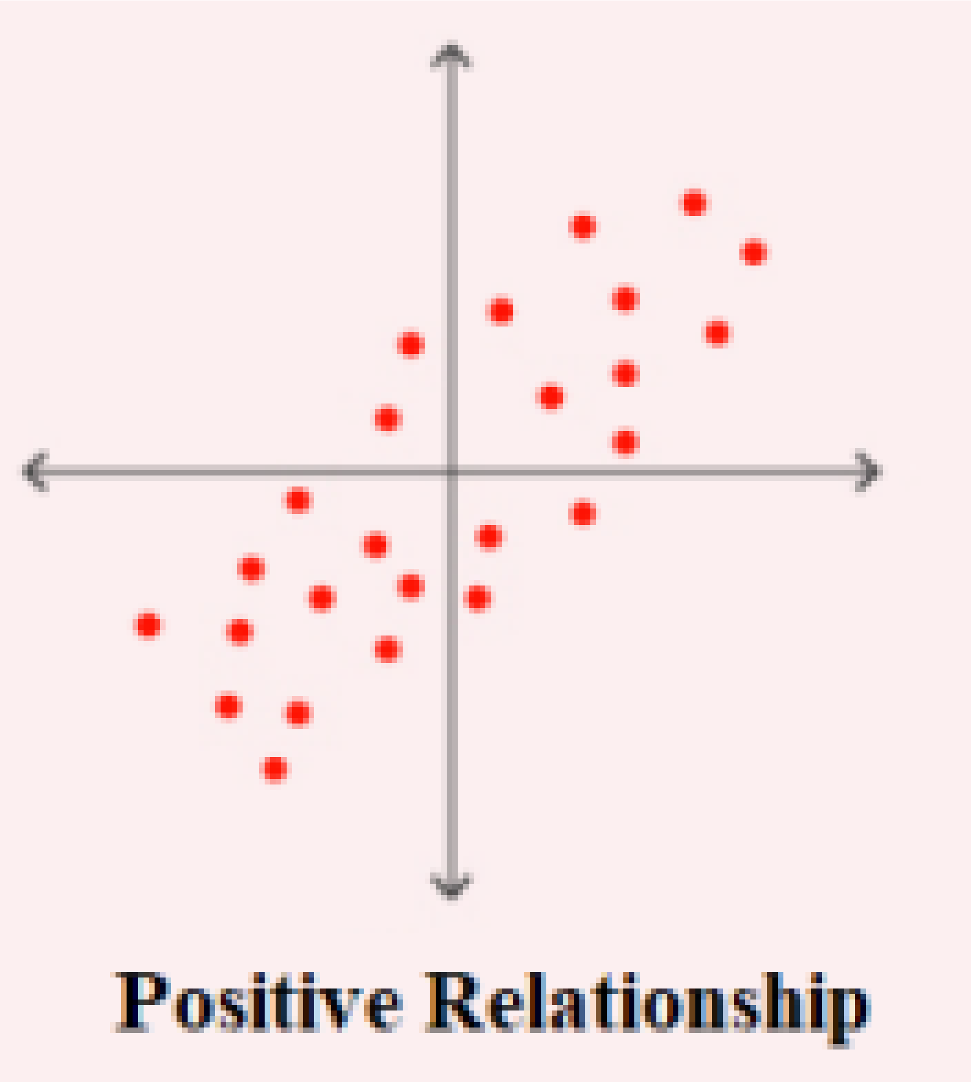
Correlation

- The correlation coefficient r is given by

$$r = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

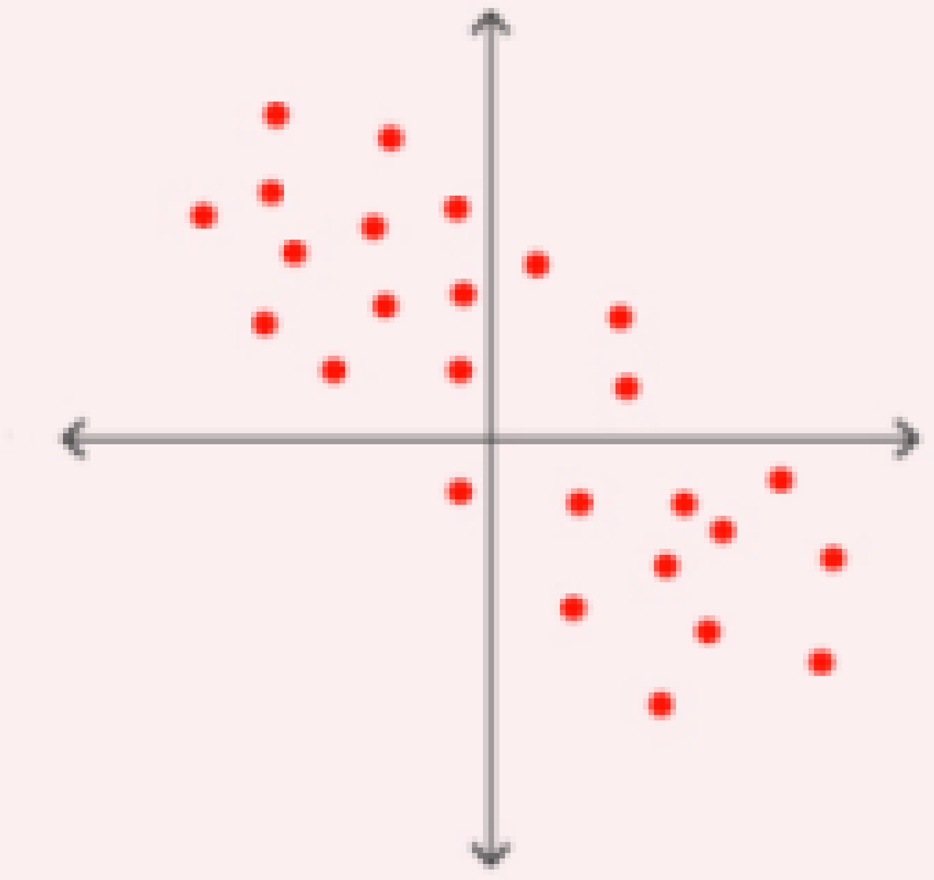
- **This will always be a number between -1 and 1 (inclusive).**

Correlation



- If r is close to 1, the variables are positively correlated ➔ there is likely a strong linear relationship between the two variables, with a positive slope.

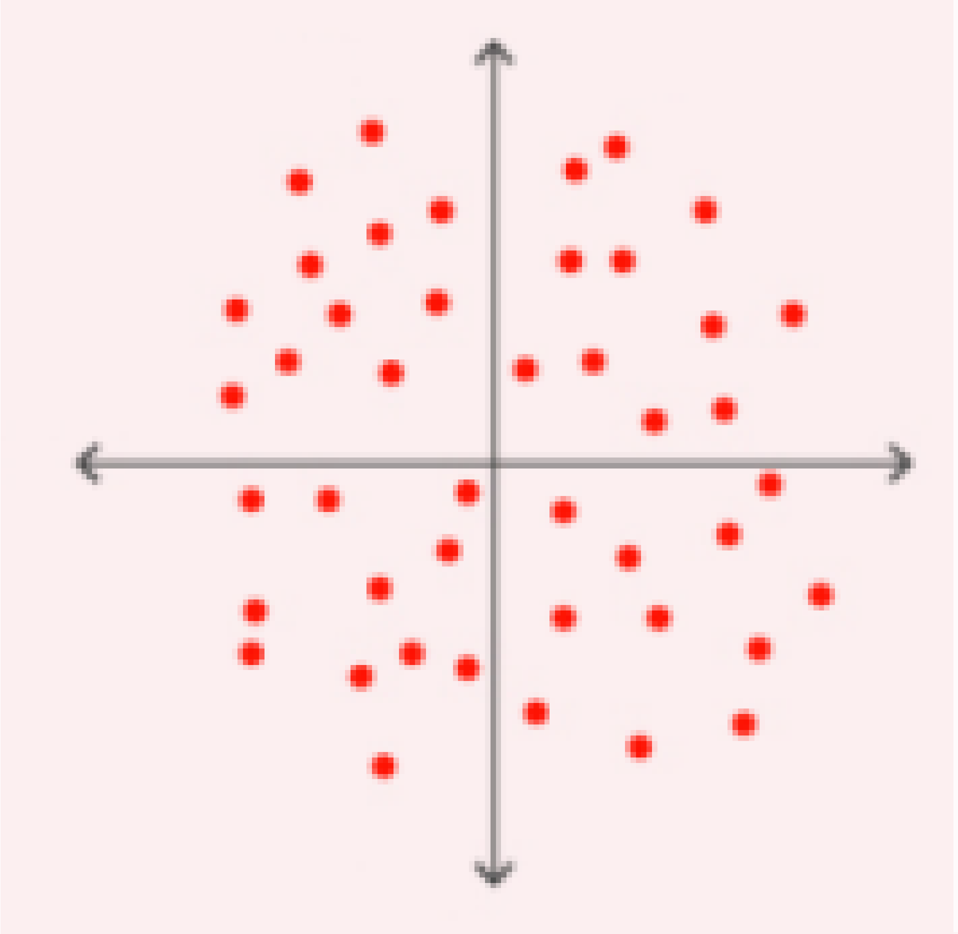
Correlation



Negative Relationship

- If r is close to -1, the variables are negatively correlated ➔ there is likely a strong linear relationship between the two variables, with a negative slope.

Correlation



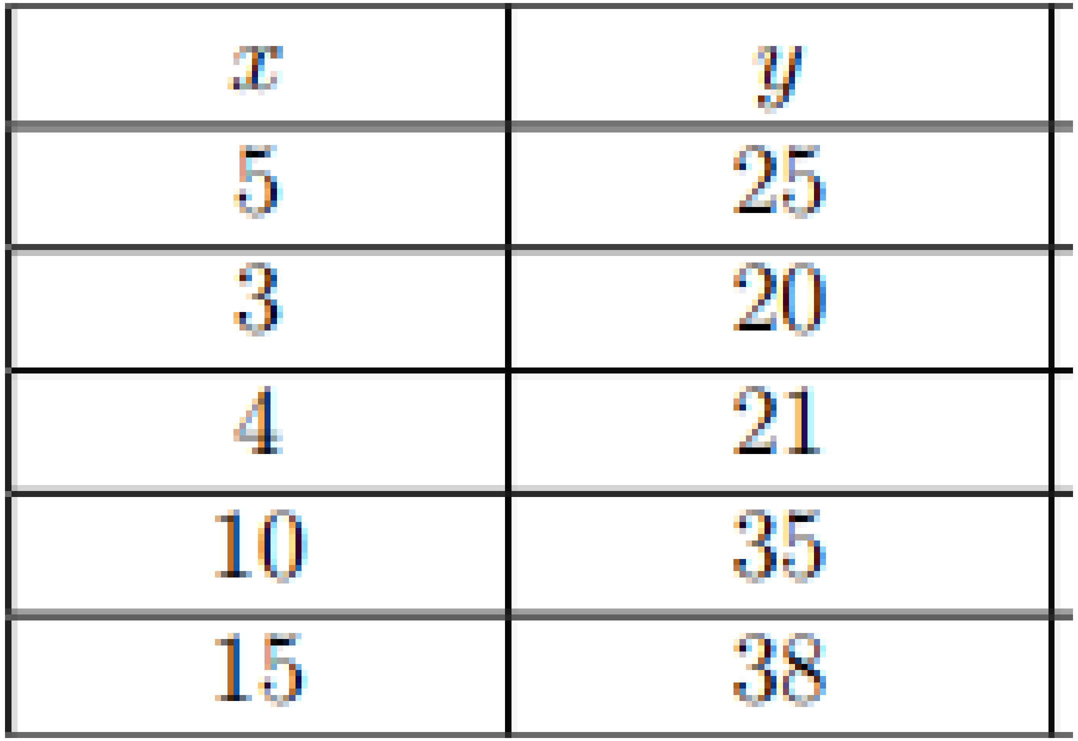
No Relationship

- If r is close to 0, the variables are not correlated ➔ that there is likely no linear relationship between the two variables, however, the variables may still be related in some other way.

Correlation

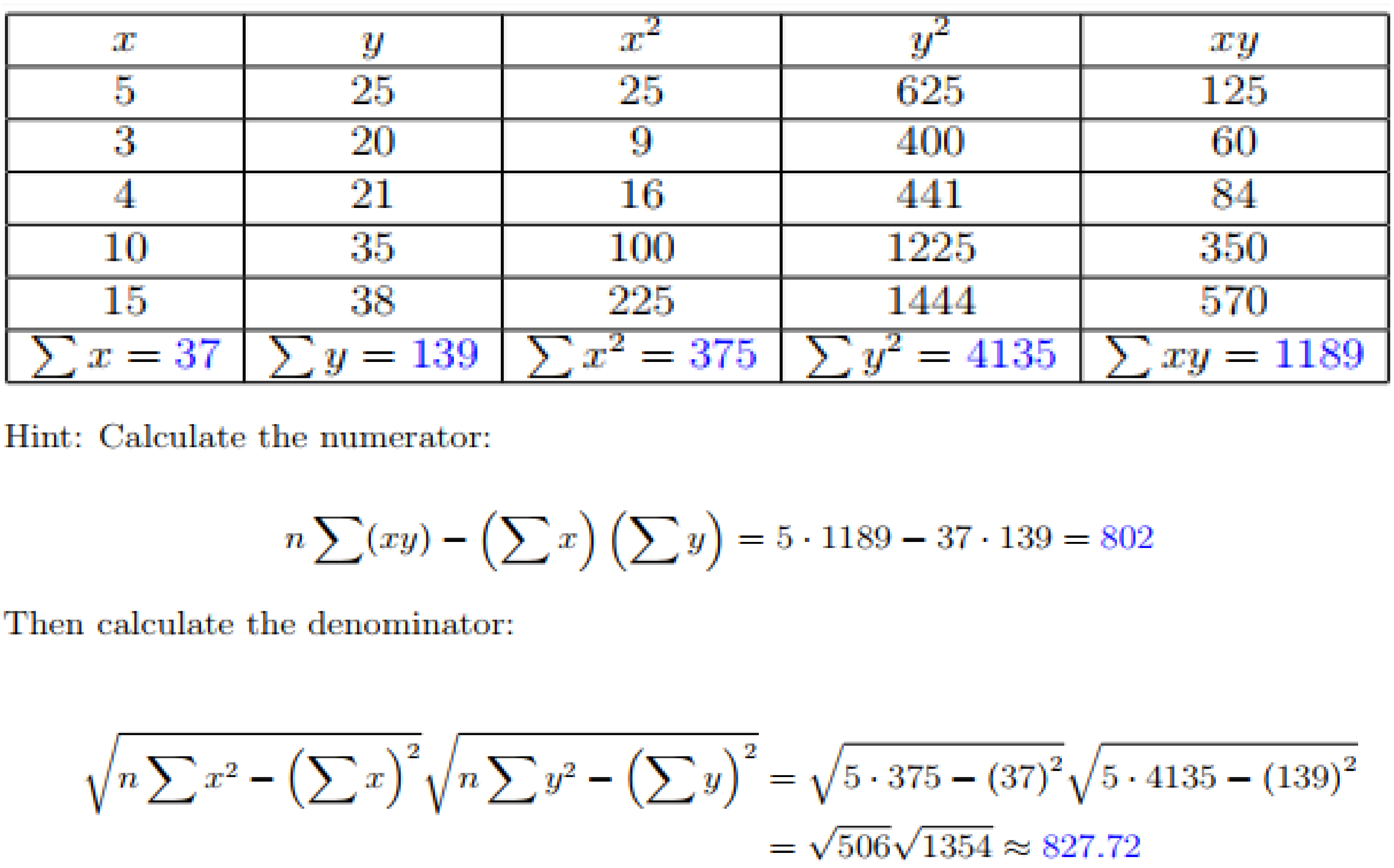
Question:

- The time x in years that an employee spent at a company and the employee’s hourly pay, y, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient r



x	y
5	25
3	20
4	21
10	35
15	38

Correlation



x	y	x^2	y^2	xy
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570
$\sum x = 37$	$\sum y = 139$	$\sum x^2 = 375$	$\sum y^2 = 4135$	$\sum xy = 1189$

Hint: Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 5 \cdot 1189 - 37 \cdot 139 = 802$$

Then calculate the denominator:

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{5 \cdot 375 - (37)^2}\sqrt{5 \cdot 4135 - (139)^2}$$
$$= \sqrt{506}\sqrt{1354} \approx 827.72$$

Correlation

Now, divide to get $r \approx \dfrac{802}{827.72} \approx 0.97$.

- **Interpret this result:** There is a strong positive correlation between the number of years and employee has worked and the employee’s salary, since r is very close to 1

ANOVA

Population

Sampling

- IoE inspection to get feedback from students/faculty/parent/Alumni/Industry
- Quality control (Statistical Quality Control)
    - 100% inspection
    - Sample inspection
- Conducting Experiments

Note:

There should not be significant variation between the sample mean and the population mean.

This is to be proved statistically.

Why ANOVA?

Helps us to understand how different sample groups respond.

ANOVA

- **ANOVA – ANalysis of Variance**
- **Variance:**
- The variance measures the average degree to which each data point is different from the mean.
- The variance is greater when there is a wider range of numbers in the group.
- The calculation of variance uses squares because it weighs outliers more heavily than data point closer to the mean.
- This prevents differences above the mean from canceling out those below, which would result in a variance of zero.
- Thus variance is the average of the squared differences from the mean.
- **ANOVA** is a **hypothesis testing procedure** that is used to evaluate differences between 2 or more samples

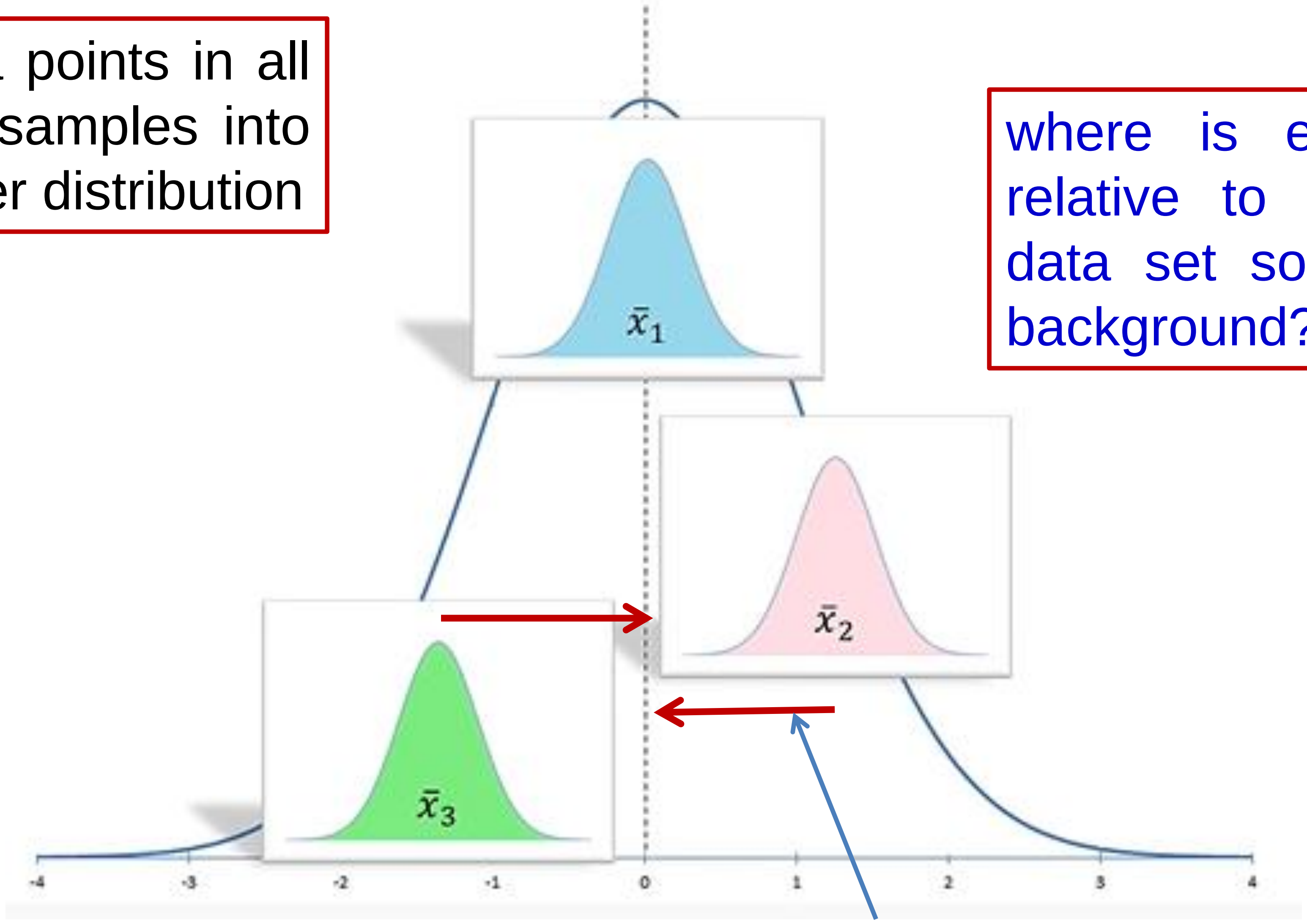
Standard Deviation:

- Standard Deviation tells how far the data points are from the mean.
- It is the square root of variance
- These two statistical concepts are closely related
- For Data analysts, these two mathematical concepts are of paramount importance as they are used to measure volatility of data distribution.
- In stock trading, if the standard deviation is less, it indicates the investment is less risky.

ANOVA

Put all the data points in all of the THREE samples into a common larger distribution

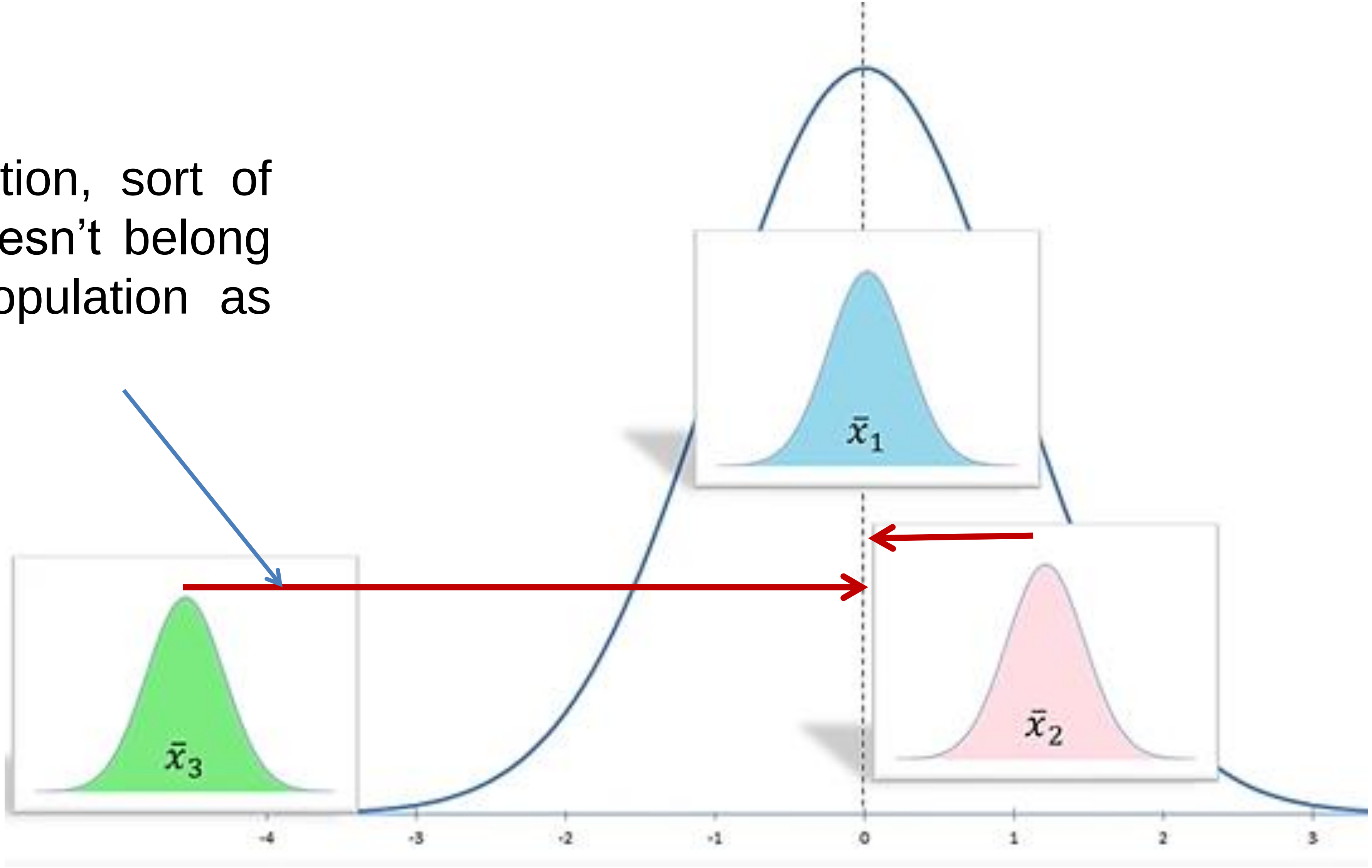
where is each mean relative to the overall data set sorted in the background?



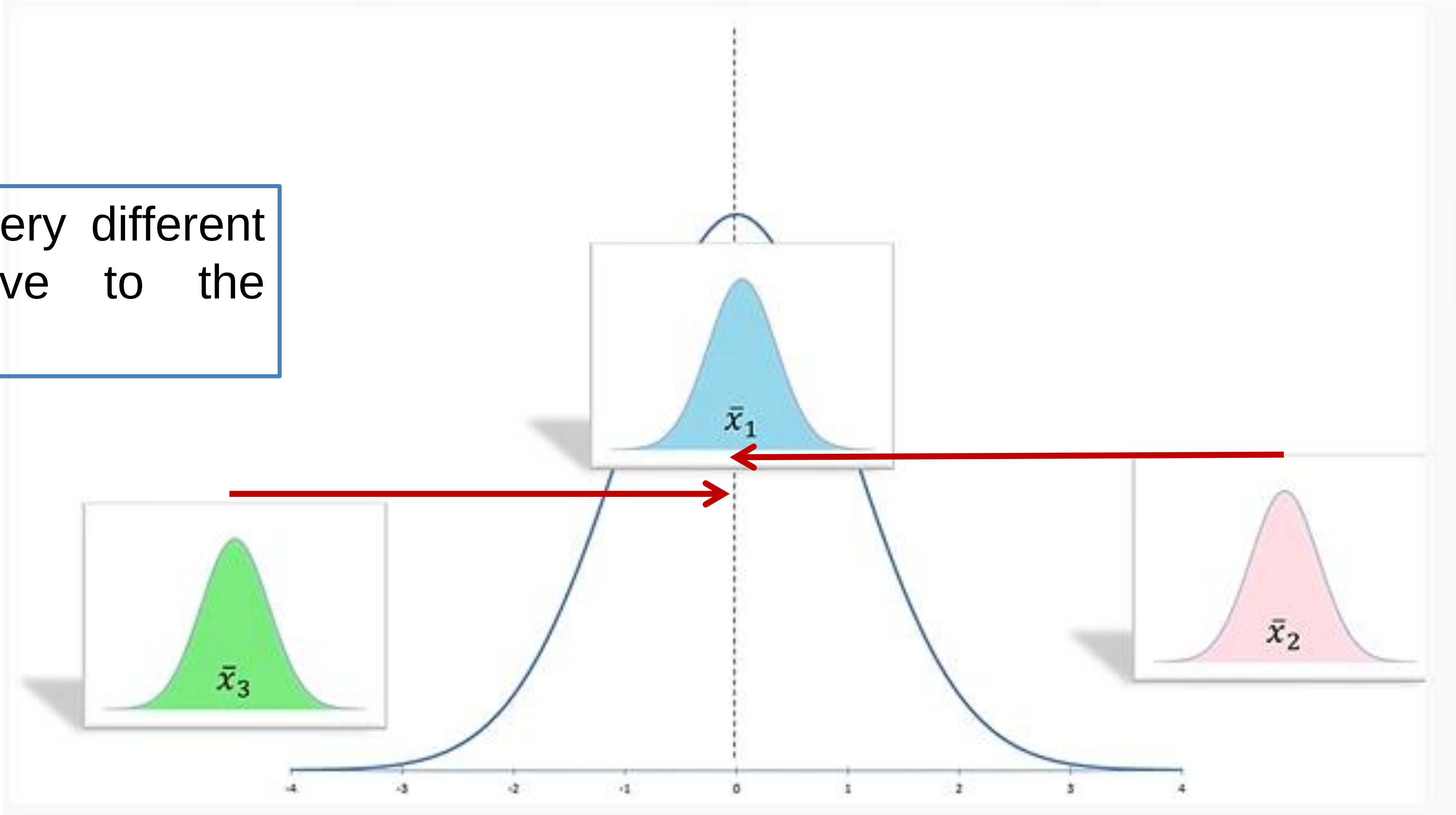
Shows how far the mean it is away from the mean of the larger sort of combined population

ANOVA

Oddball distribution, sort of the one that doesn't belong in the same population as the other two



Means are in very different locations relative to the overall mean



Step1:

Setting the hypothesis (Null hypothesis or alternate hypothesis)

- Null Hypothesis (H0: $\mu1=\mu2=\mu3$)
- Alternate Hypothesis (Ha: Alteast one difference among the means)

And

- Fixing the confidence interval (90%, 95%)

α=0.1 or 0.05

Step2: Find the df

- df between the groups/columns
- df within the groups/columns
- df_total

Step3:Calculating the Means

- Means for each group and
- Grand mean

Step4: All variability across the columns/groups

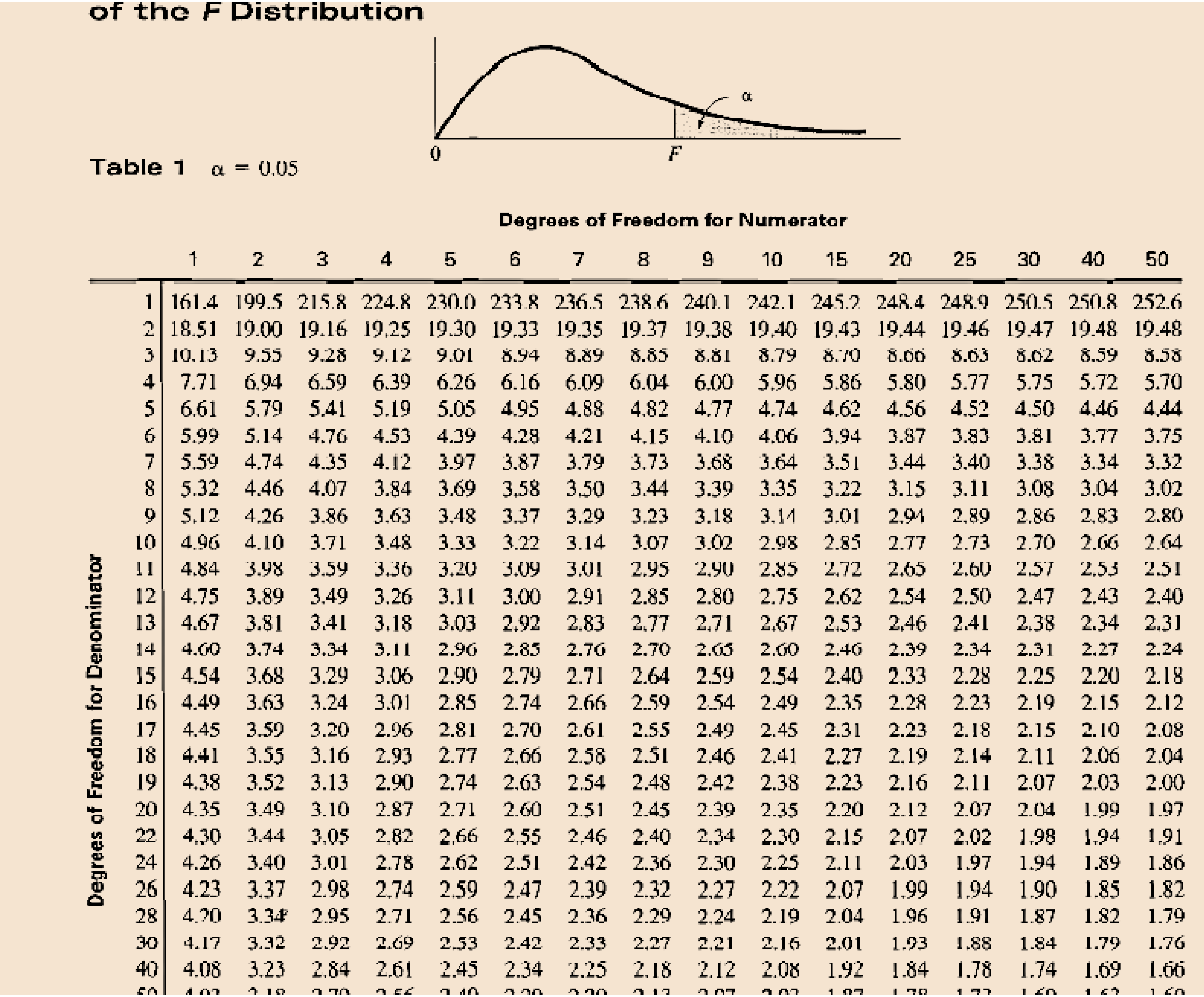
- SST
- SSC (Sum of Squares between/Columns)
- SSE(Sum of Squares within/Errors)

Step5: To calculate the variance between and within

- Mean Squares_between = $\frac{SS\_between}{df\_between}$
- Mean Squares_within = $\frac{SS\_within}{df\_within}$

Step 6: To perform F test (To calculate F_ratio)

- F_statistic = Mean Square_between / Mean Square_within
- F_critical from F distribution table (Corr to df_numerator and df_denominator)



of the *F* Distribution

Table 1 α = 0.05

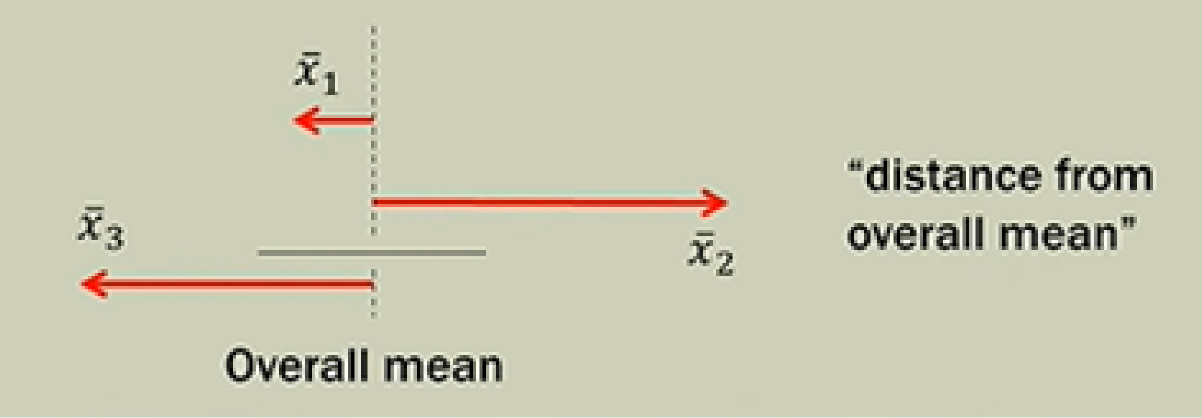
Degrees of Freedom for Numerator

Degrees of Freedom for Denominator	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.73	1.69	1.63	1.60

F_statistic < F_critical

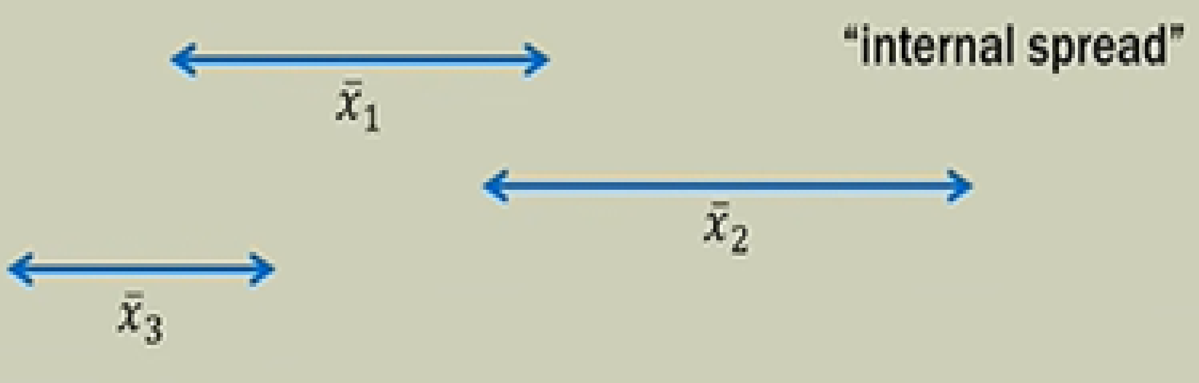
ANOVA: Analysis of Variance is a *variability ratio*

Variability
AMONG / BETWEEN
the means.



"distance from overall mean"

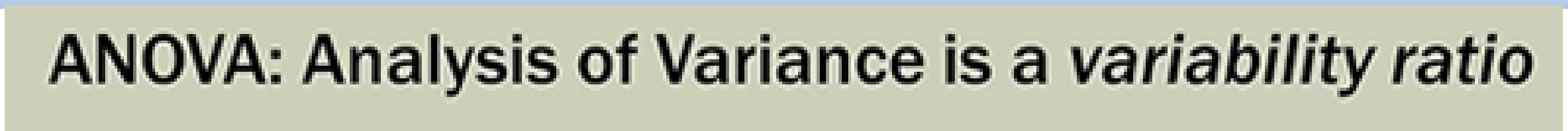
Variability
AROUND / WITHIN the
distributions.



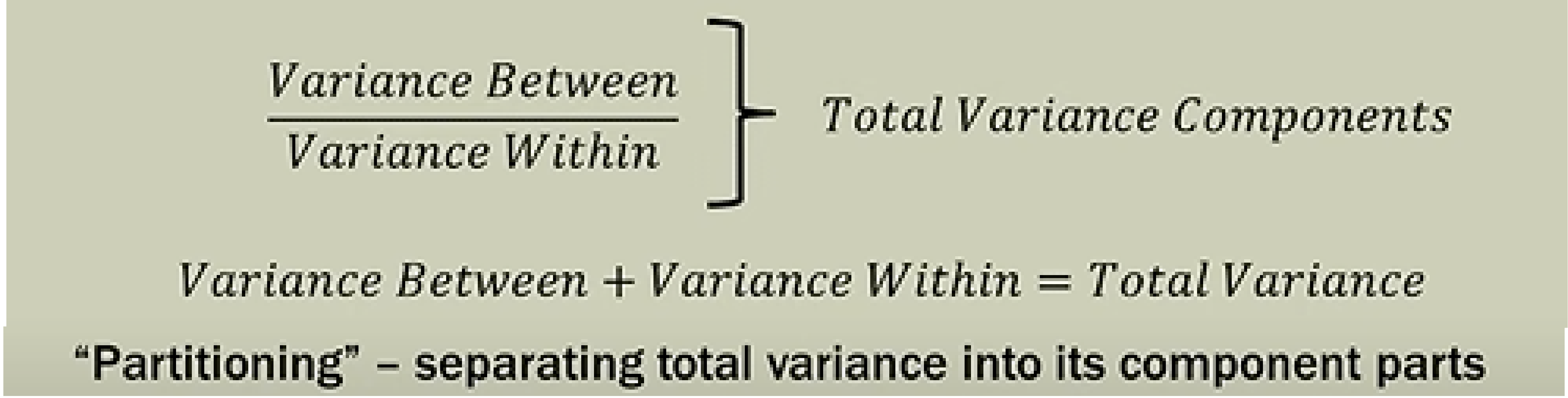
"internal spread"

$$= \frac{\text{Variance Between}}{\text{Variance Within}}$$

ANOVA



ANOVA: Analysis of Variance is a *variability ratio*

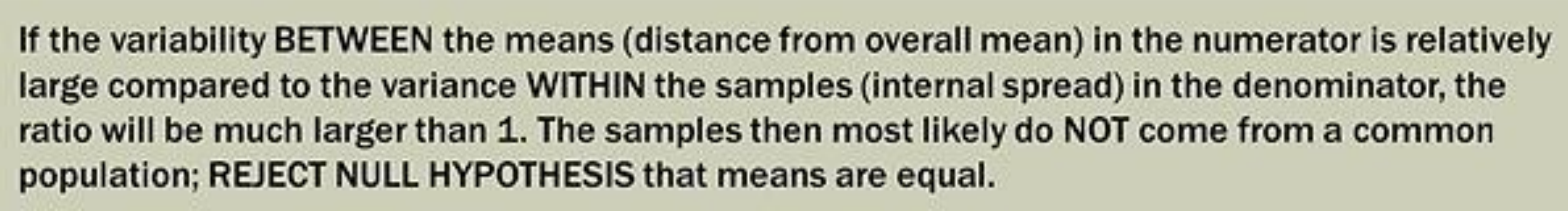


$$\left.\frac{\textit{Variance Between}}{\textit{Variance Within}}\right\} \quad \textit{Total Variance Components}$$

$$\textit{Variance Between} + \textit{Variance Within} = \textit{Total Variance}$$

“Partitioning” – separating total variance into its component parts

This is One way ANOVA/ Single Factor ANOVA



If the variability BETWEEN the means (distance from overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ratio will be much larger than 1. The samples then most likely do NOT come from a common population; REJECT NULL HYPOTHESIS that means are equal.

ANOVA

ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{Variance\ Between}{Variance\ Within}$$

$$\frac{LARGE}{small} = Reject\ H_0$$

At least one mean is an outlier and each distribution is narrow; distinct from each other

$$\frac{similar}{similar} = Fail\ to\ Reject\ H_0$$

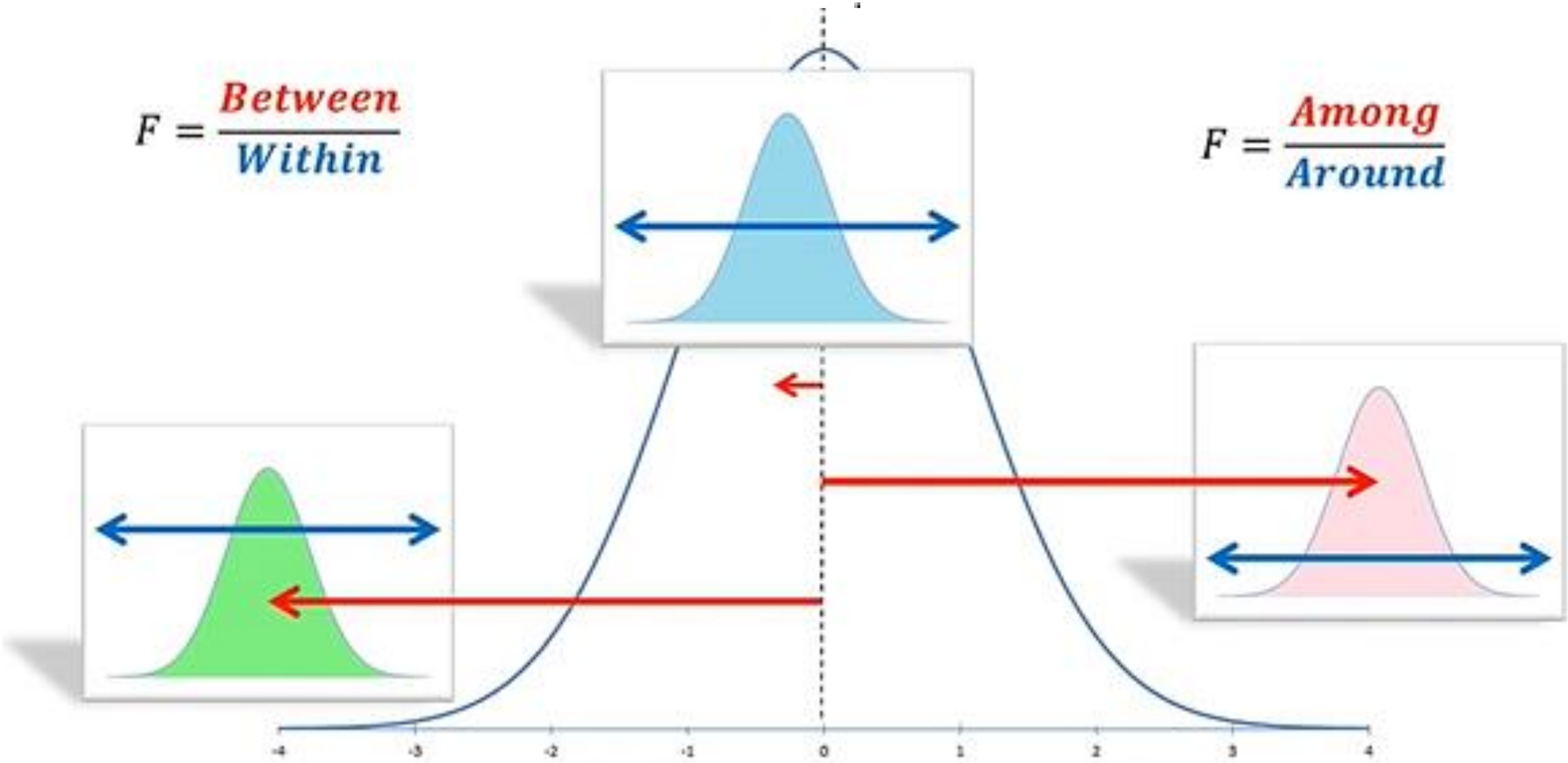
Means are fairly close to overall mean and/ or distributions overlap a bit; hard to distinguish

$$\frac{small}{LARGE} = Fail\ to\ Reject\ H_0$$

The means are very close to overall mean and/ or distribution “melt” together

ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$



Question-4:

18 students (six each from first year to third year) were selected for an informal study about their understanding skill level. The evaluation was done for a score of 100. Using One-way ANOVA technique, find out whether or not a difference exists somewhere between the three different year levels

Scores		
First Year	**Second Year**	**Third Year**
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

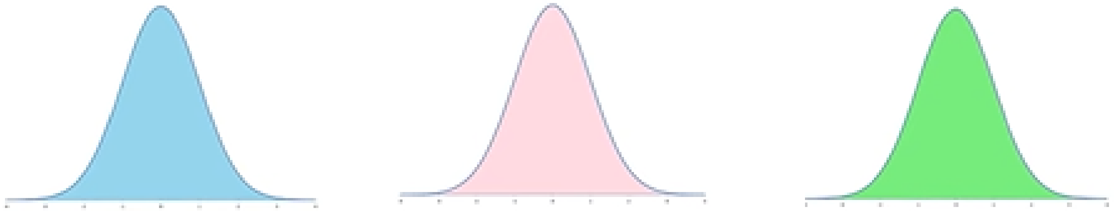
ANOVA

Groups/ Columns

Random Sample within each group

Scores		
First Year	**Second Year**	**Third Year**
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

Calculate the mean of each column



	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Calculate Grand Mean/ Overall Mean $\bar{\bar{x}}$

The mean of all 18 scores is

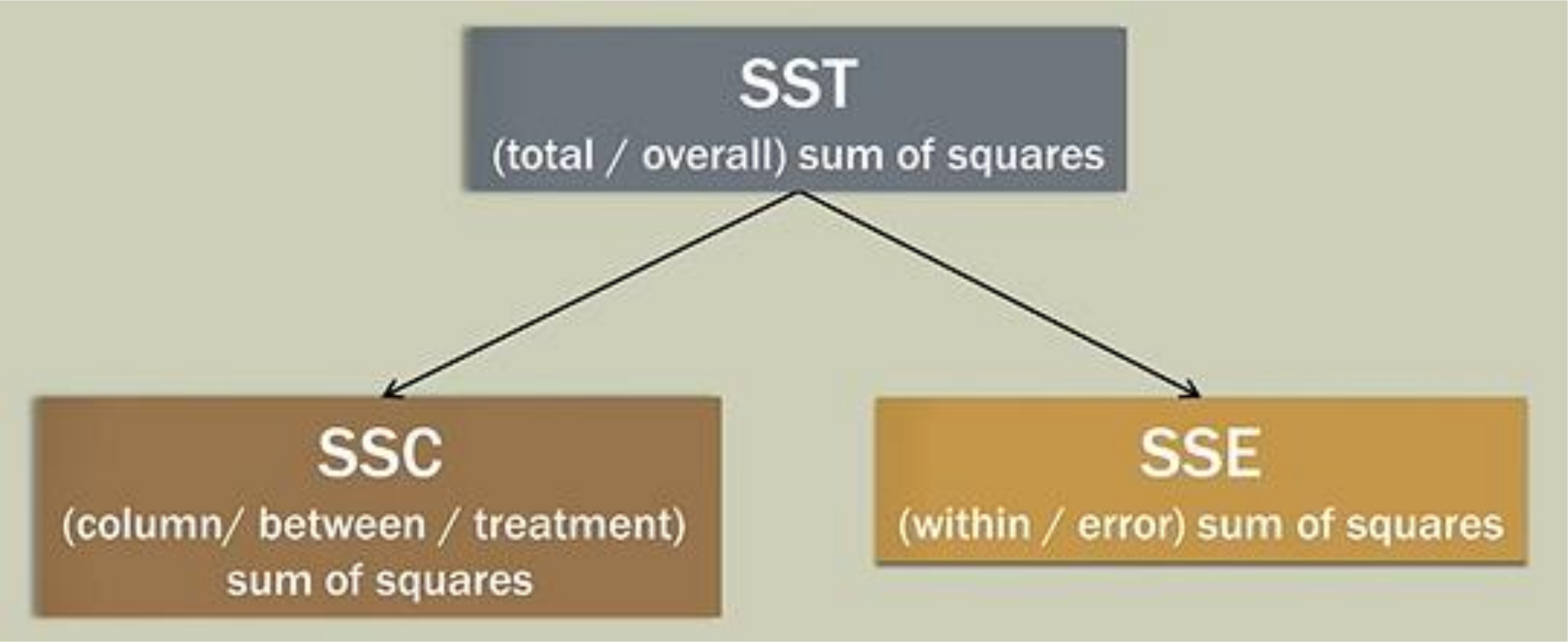
$$\bar{\bar{x}} = 74.28$$

Sum of Squares (SS)

Sum of squares of the difference of the dependent variable and its mean

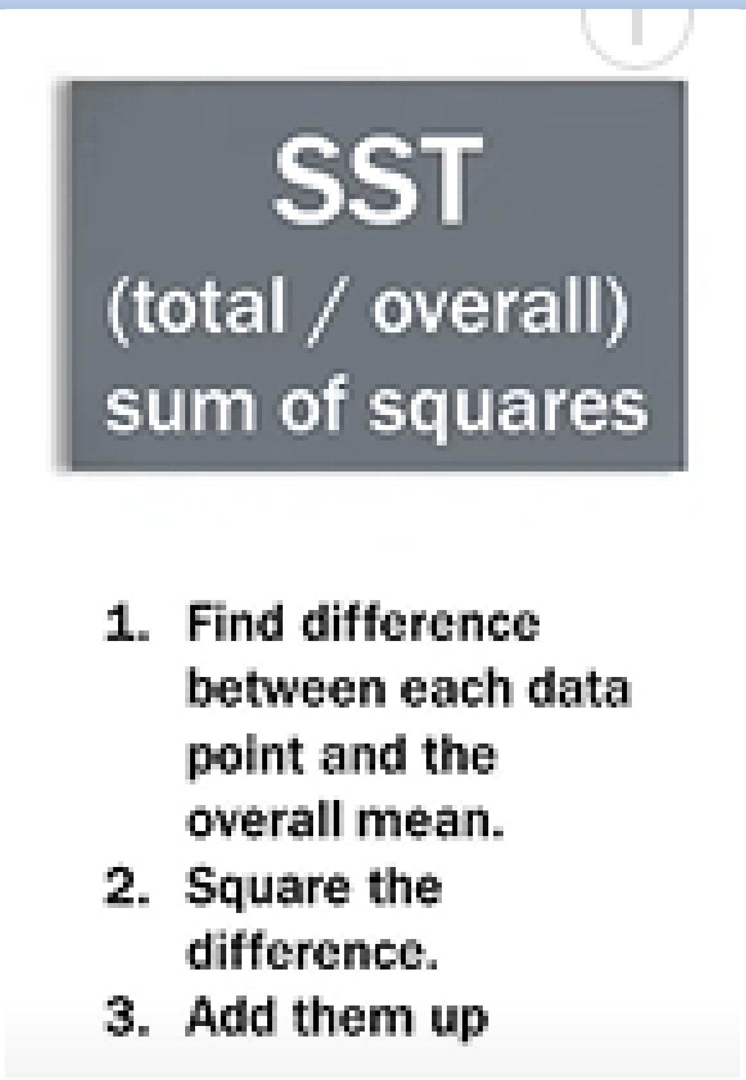
$$SS = \sum(x - \bar{x})^2$$

Partitioning Sum of Squares


SST
(total / overall) sum of squares
SSC
(column/ between / treatment) sum of squares
SSE
(within / error) sum of squares

ANOVA

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**



SST

(total / overall) sum of squares

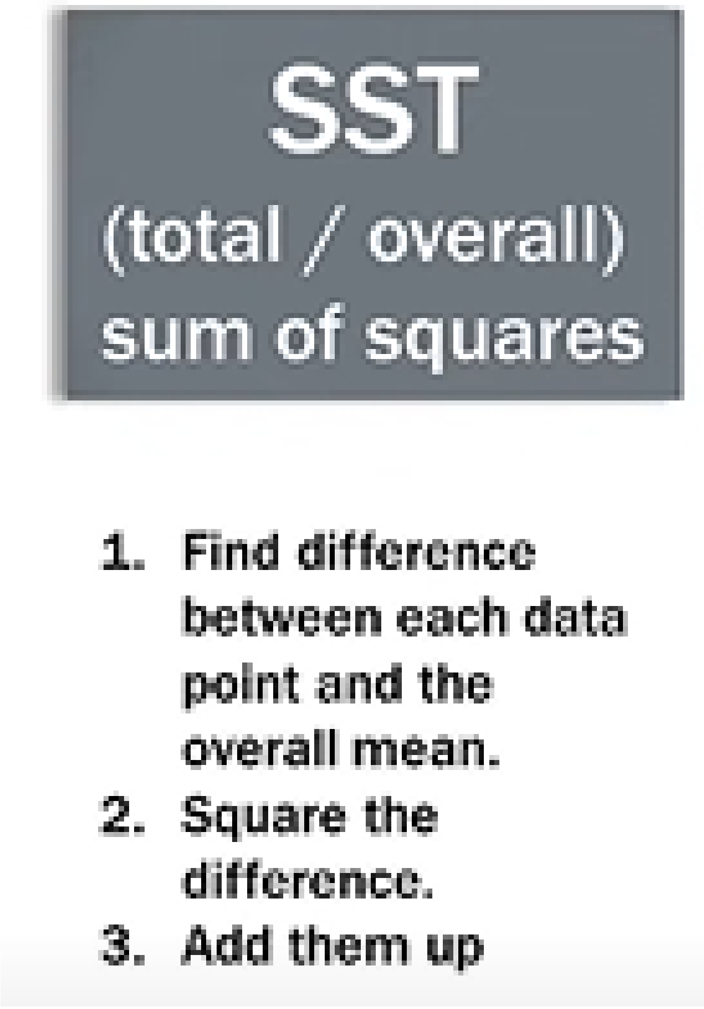
1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}}$ = 74.28

ANOVA

	Scores					
	First Year	Second Year	Third Year	(X_A-X_{mean})^2	(X_B-X_{mean})^2	(X_c-X_{mean})^2
	82	62	64	59.633	150.744	105.633
	93	85	73	350.522	114.966	1.633
	61	94	87	176.299	388.966	161.855
	74	78	91	0.077	13.855	279.633
	69	71	56	27.855	10.744	334.077
	53	66	78	452.744	68.522	13.855
Sum	432	456	449	**1067.130**	**747.796**	**896.685**
Mean	72	76	74.83			

SST = 1067.130 + 747.796 + 896.685 = 2711.611



SST

(total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}}$ **= 74.28**

Sum of Squares_between

	Scores		
	First Year	**Second Year**	**Third Year**
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

1. **Find difference between each group mean and the overall mean**
2. **Square the deviations**
3. **Multiply with no. of values of each column**
4. **Add them up**

$\bar{\bar{x}}$ **= 74.28**

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

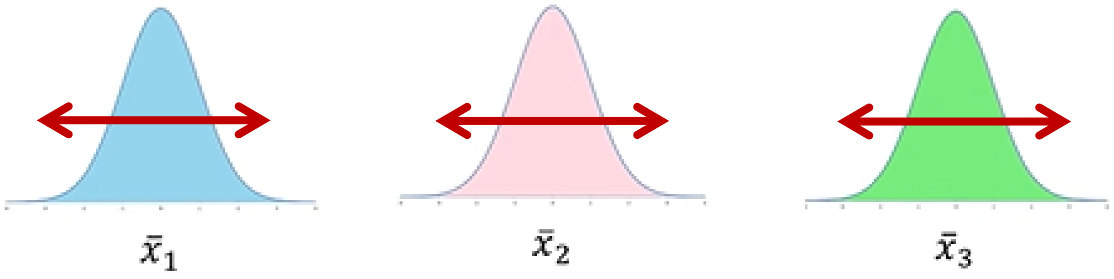
Sum of Squares_between

1. **Find difference between each group mean and the overall mean**
2. **Square the deviations**
3. **Multiply with no. of values of each column**
4. **Add them up**

$\bar{\bar{x}} = 74.28$

SSC = 6(72 – 74.28)^2 + 6(76 – 74.28)^2 +6 (74.83 – 74.28)^2 = 50.778

ANOVA



	Scores		
	First Year	**Second Year**	**Third Year**
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_between

1. Find difference between each data point and its column mean.
2. Square each deviation.
3. Add them up the squared deviations.

ANOVA

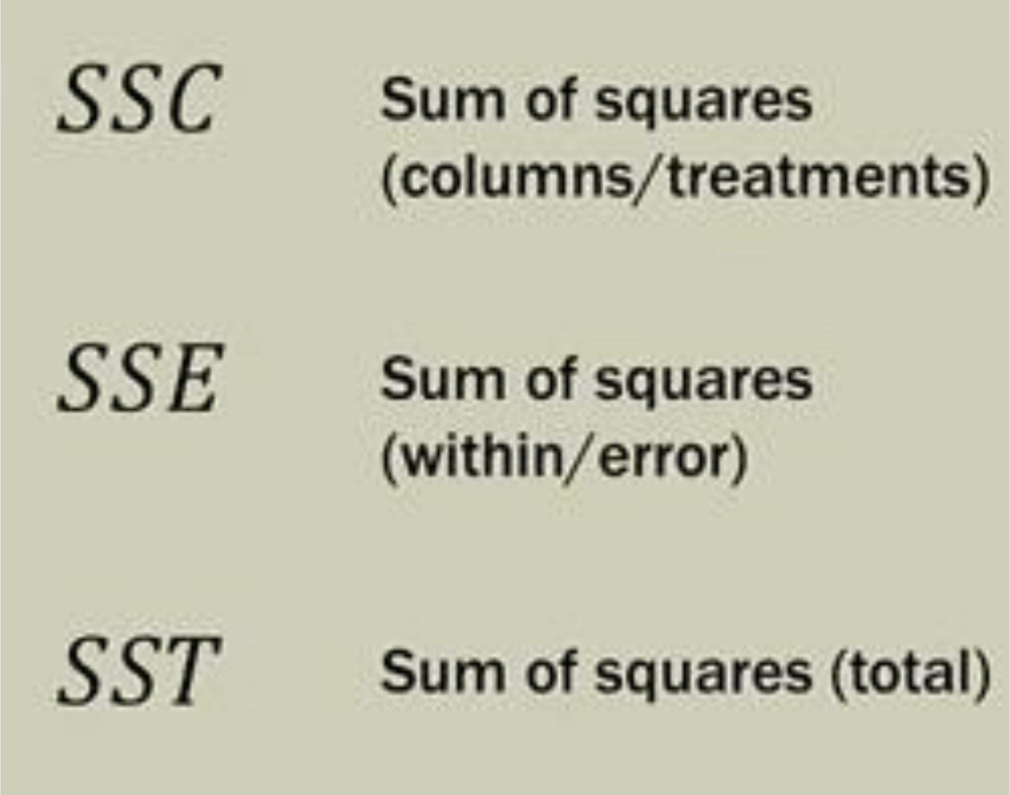
	Scores					
	First Year	**Second Year**	**Third Year**	**$(X_A-x_{a\_mean})$^2**	**$(X_B-x_{b\_mean})$^2**	**$(X_C-x_{c\_mean})$^2**
	82	62	64	100	196	117.361
	93	85	73	441	81	3.361
	61	94	87	121	324	148.028
	74	78	91	4	4	261.361
	69	71	56	9	25	354.694
	53	66	78	361	100	10.028
Sum	432	456	449	**1036**	**730**	**894.833**
Mean	72	76	74.83			

SSE = 1036 + 730 + 894.833 = 2660.833

Sum of Squares_within

1. Find difference between each data point and its column mean.
2. Square each deviation.
3. Add them up the squared deviations.

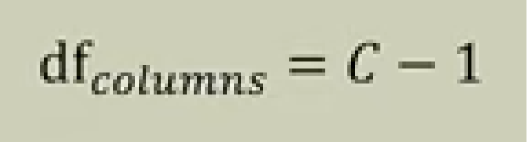
Formulas for One-Way ANOVA



SSC	Sum of squares (columns/treatments)
SSE	Sum of squares (within/error)
SST	Sum of squares (total)

df = Degrees of Freedom

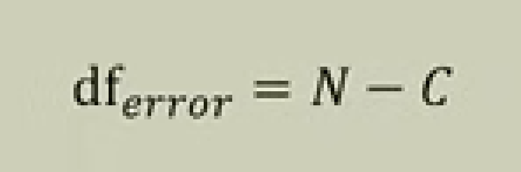
1. DoF b/w the columns



$$\mathrm{df}_{columns} = C - 1$$

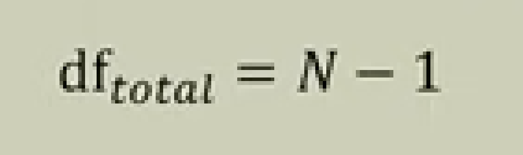
$$\text{Mean Squares\_between} = \frac{\text{SS\_between}}{df\_between}$$

2. DoF within the columns



$$\mathrm{df}_{error} = N - C$$

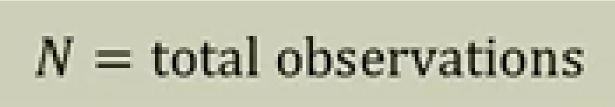
$$\text{Mean Squares\_within} = \frac{\text{SS\_}within}{df\_within}$$



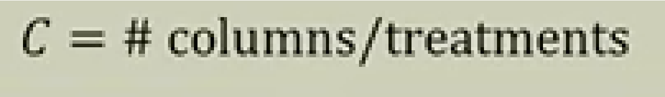
$$\mathrm{df}_{total} = N - 1$$

ANOVA F - statistic

$$F = \frac{\text{Mean Squares\_between}}{\text{Mean Squares\_within}}$$



N = total observations



C = # columns/treatments

MSC = Mean Square Columns/ Treatments

MSE = Mean Square Error/ Within

Substituting the values

Formula to calculate Critical Value in Excel:

F.INV.RT(ALPHA,NUMERATOR DOF, DENOMINATOR DOF)

$$\text{Mean Squares\_between} = \frac{50.778}{3-1} = \mathbf{25.389}$$

$$\text{Mean Squares\_within} = \frac{2660.833}{18-3} = \mathbf{177.389}$$

$$F = \frac{\text{MSC}}{\text{MSE}} = \frac{25.389}{177.389} = \mathbf{0.1431}$$

- F-statistic value is less than F_{critical}
- Null hypothesis is accepted.
- **It means there is no significant difference in mean values**

Critical value of F: $F_{\alpha,\ dfc,\ dfe} = F_{0.05,\ 2,\ 15} = \mathbf{3.68}$

Thank You!