1. What is Statistics
2. Data types in statistics

* Numerical
* Discrete
* Conti
* Categorical

1. Levels of data

* Nominal level
* Ordinal level
* Interval level : No zero scale
* Ratio level : Has Zero scale

1. Population
2. Sample
3. Population vs sample
4. Inferential statistics
5. Frequency table
6. Relative frequency table
7. Bar chart
8. Pie chart
9. Frequency distribution table
10. Histogram
11. Distribution plot
12. Central tendency

* Mean
* Median
* Mode

1. Mean vs Median
2. Outlier
3. Distribution Uni mode bi mode
4. Types of Distribution
5. Negative skewed : Left Skewed : Mean < Median<Mode
6. Positive skewed : Right Skewed: Mean > Median > Mode
7. No skew : Bell shaped : Normal distribution : Mean= Median =mode
8. Asymptotes : will never touch the real axis

In order to understand the distribution

Central tendency : Middle point analysis

Data flow or Data distribution

1. Range
2. Mean deviation
3. Absolute mean deviation
4. Variance
5. Standard deviation
6. Range

1 to 100

Min Max

Range= Max-Min= 100-1= 99

Data divide into 99 parts

Disadvantage :

This method concentrates only on end values rather than middle values

1. Mean deviation:

How much each observation or data point or sample or tuple is deviated from mean value

In a data we have rows : Observations, Data points , sample , tuple

Columns: Variables , features, Dimensions, Predictors

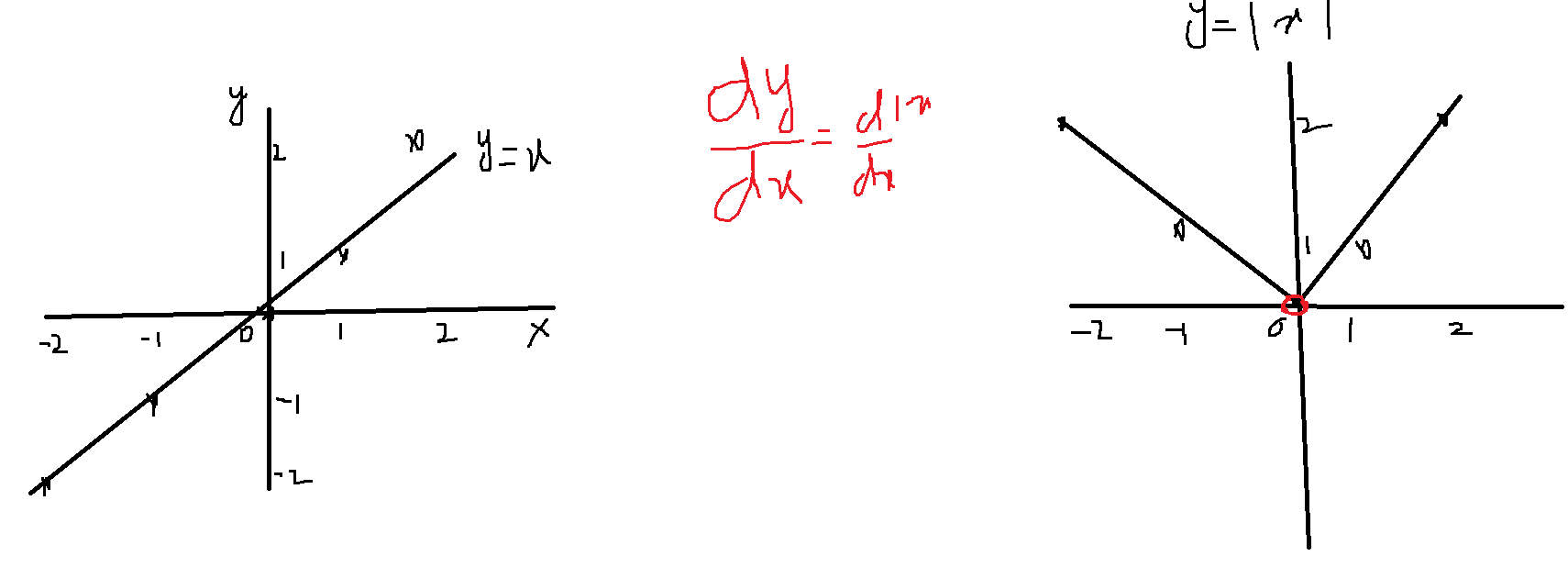
Assume that we have 5 observations are there

Anand kumar says : Mod

1. Absolute mean deviation:

Assume that we have 5 observations are there

|  |  |  |
| --- | --- | --- |
| X | Y=x | Y=|x| |
| -2 | -2 (-2,-2) | 2 (-2,2) |
| -1 | -1 (-1,-1) | 1 (-1,1) |
| 0 | 0 (0,0) | 0 (0,0) |
| 1 | 1 (1,1) | 1 (1,1) |
| 2 | 2 (2,2) | 2 (2,2) |



Draw back:

Absolute mean deviation related to absolute value

4. Variance:

Assume that we have 5 observations are there

1. Error is increasing
2. Units also raise

Km Fare

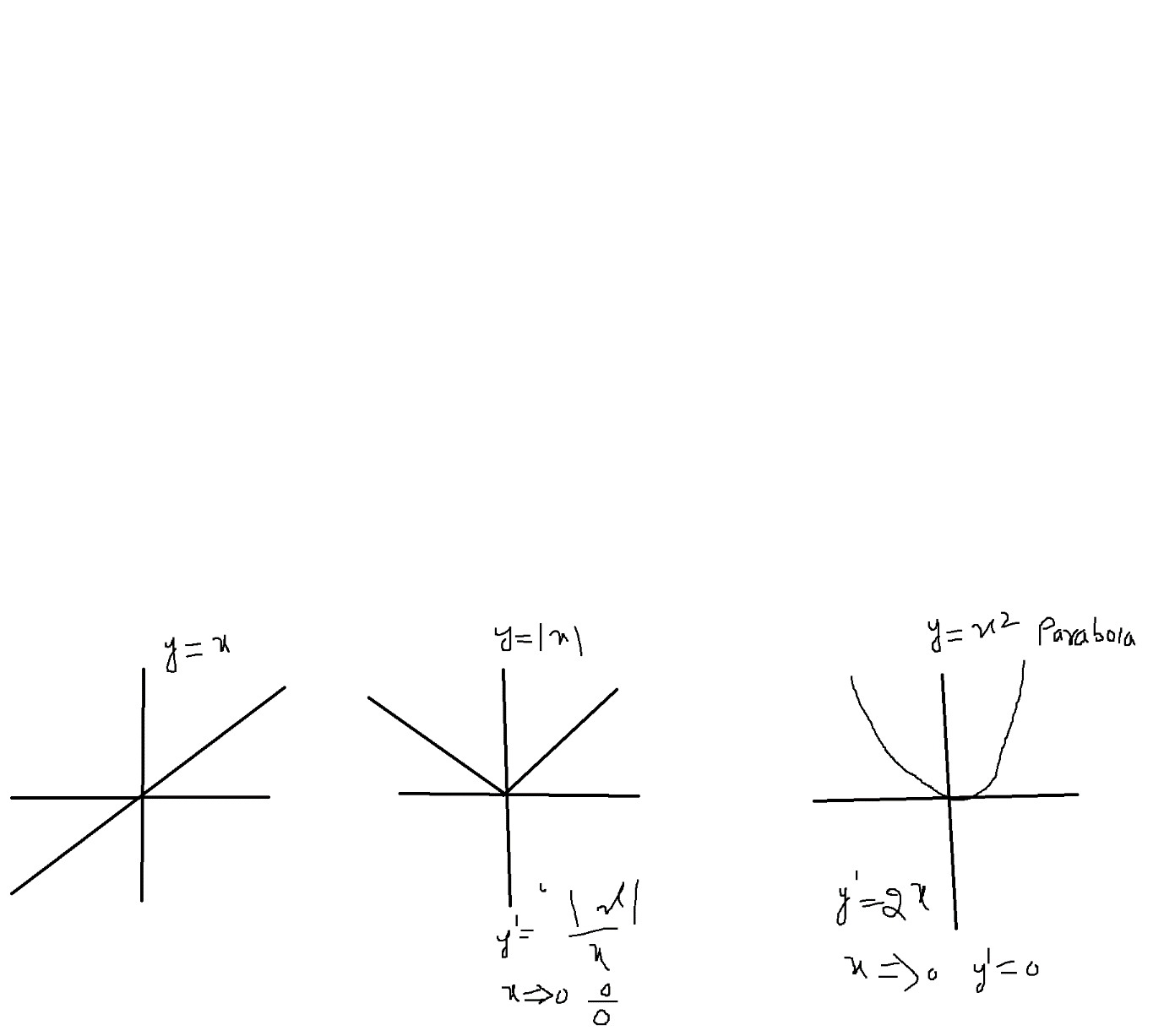
2 10

3 11

(2km-2.5km)^2 = 0.5km^2 km^2

5. Standard Deviation

|  |  |  |  |
| --- | --- | --- | --- |
| X | Y=x | Y=|x| | Y=x^2 |
| -2 | -2 (-2,-2) | 2 (-2,2) | 4 |
| -1 | -1 (-1,-1) | 1 (-1,1) | 1 |
| 0 | 0 (0,0) | 0 (0,0) | 0 |
| 1 | 1 (1,1) | 1 (1,1) | 1 |
| 2 | 2 (2,2) | 2 (2,2) | 4 |



Y=mx+c

Y= |x|

Y=x^2

Y=logx loge log10 log2

Normal scale vs Log scale

Y= sinx cosx tanx

Covariance:

Co : two

How a variable is varying w.r.t to another variable

Age Salary

30 50k

31 55k

32 70k

Var(Salary)

50 55 70 mean= (50+55+70)/3= 175/3= 58

|  |  |  |
| --- | --- | --- |
| Salary | (xi-xbar)^2 |  |
| 50 | (50-58)^2 | 64 |
| 55 | (55-58)^2 | 9 |
| 70 | (70-58)^2 | 144 |

Step-1: Calculate the mean : 58

Step-2: Xi-xbar : 50-58=-8

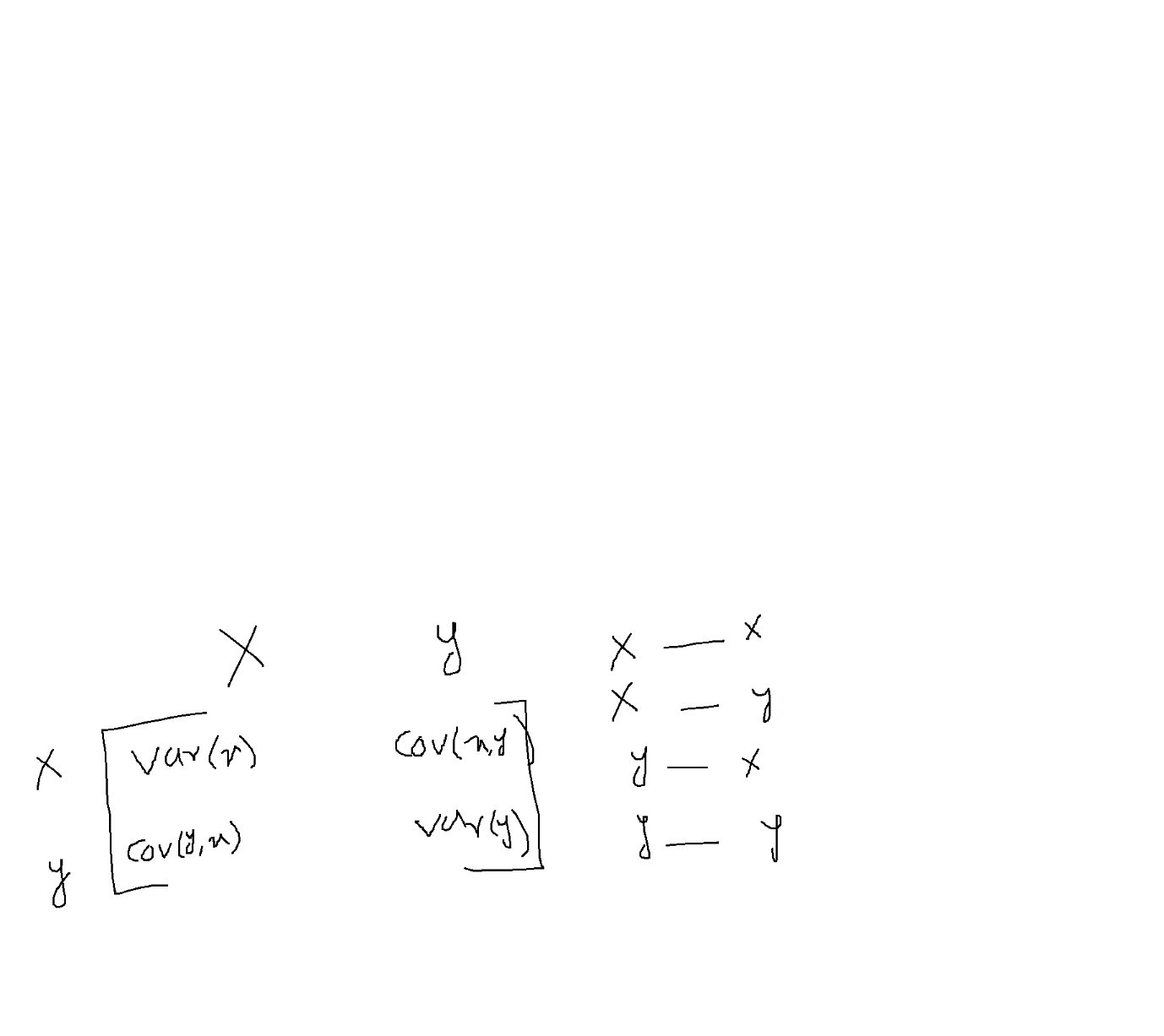
Step-3: square of that step-2 value : 64

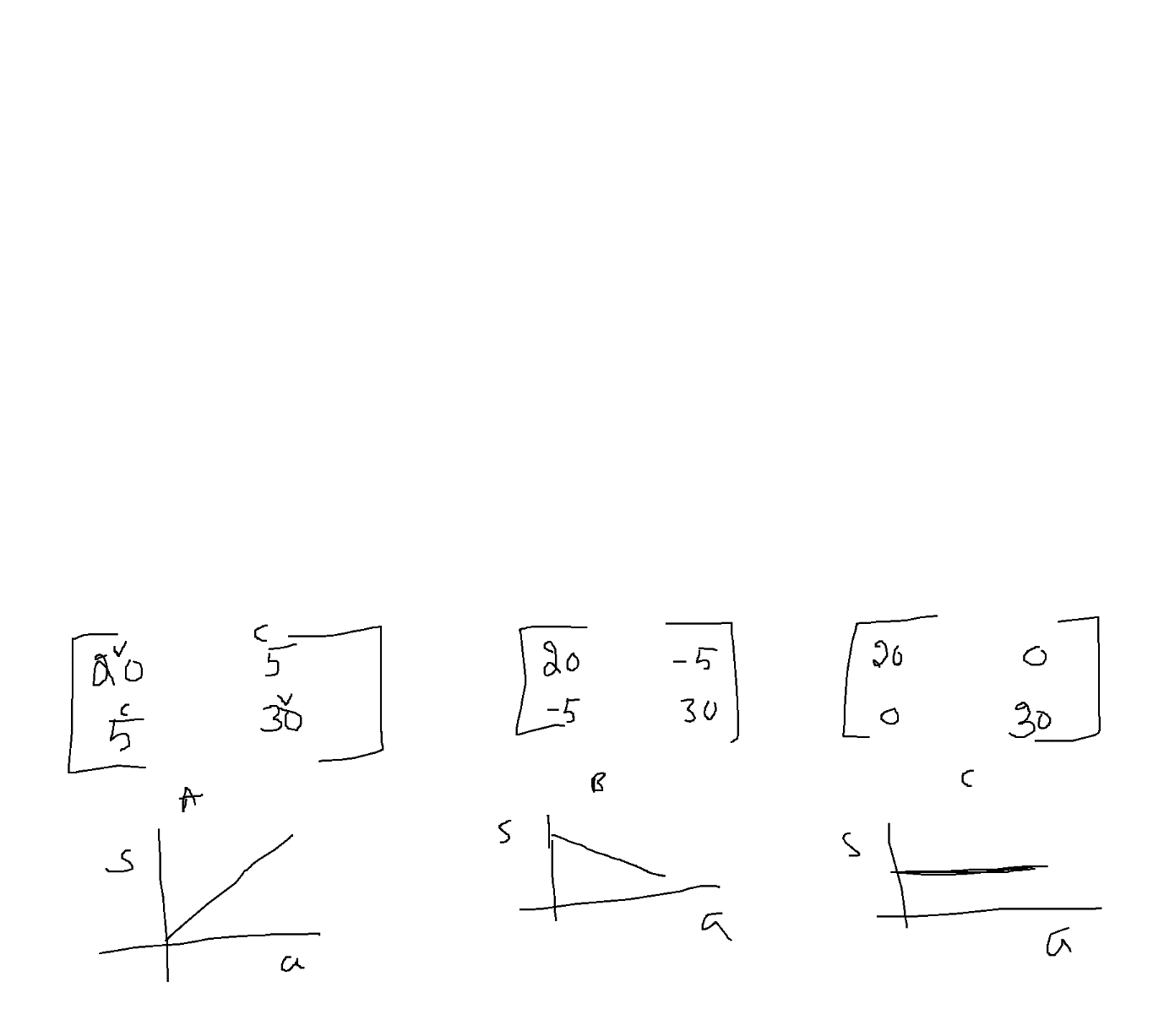
Step-4: summation: 64+9+144= 217

Step-5: Divide by num observations= 217/3=72

Covariance:

Covariance Matrix:





Covariance gives relation between two variables

Positively related :

Negatively related

No relation