Lecture 6

Describing numerical data

Histogram

Numerical data is represented in the form of a histogram

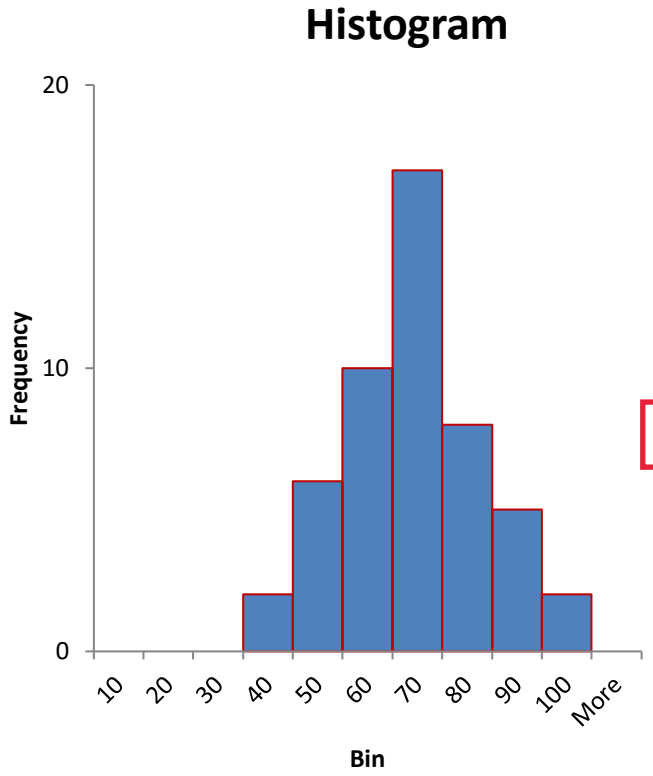
Most common plot of the distribution of a numerical variable

Marks obtained by studnts in an exam

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Data converted into a range with frequency

Interval	*Frequency*
0-10	0
10-20	0
20-30	0
30-40	2
40-50	6
50-60	10
60-70	17
70-80	8
80-90	5
90-100	2



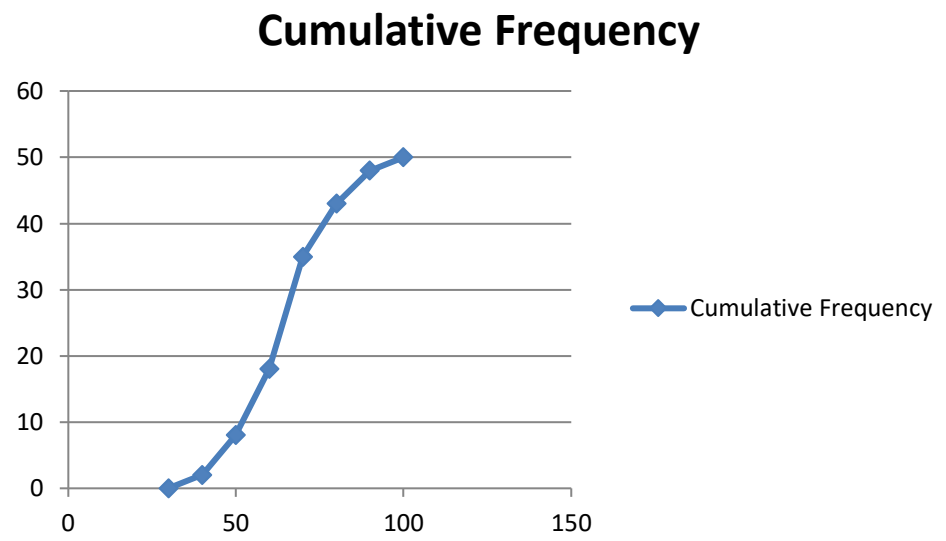
Same colour bec just a variable is represented

Difference b/w histogram and bar chart

Histogram positions bars with no gaps

Tells that there is some data between 70 and 80

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32



Ogive or Cumulative line graph

Cumulative frequency distribution

Range	*Frequency*	Cumulative
0 – 30	0	0
30 - 40	2	2
40 - 50	6	8
50 - 60	10	18
60 - 70	17	35
70 - 80	8	43
80 - 90	5	48
90 - 100	2	50

Note the difference from the previous graph is that the we had 0-10,10-20,20-30 as Zero so here we have clubbed all the three together and shown it as cumulative frequency

Also note that with cumulative frequency we can also evaluate the data less than 40,50,60... and the last value of cumulative frequncy will equal the total of Frequency distribution table

Stem and Leaf

Only change from frequency distribution table is that here we also show the data split into stem and a leaf. Leaf being attached to a stem like '8' is attached to '3' in '38'

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Cumulative Frequency	Stem	Leaf
2	3	2 8
8	4	4 6 7 7 7 8
17	5	2 3 3 4 6 7 8 9 9
35	6	0 2 2 2 2 3 3 4 4 6 6 6 6 7 7 7 8 8
44	7	1 1 2 3 3 3 7 7 8
48	8	3 8 8 9
50	9	0 2 4

This shows the Cumulative frequency as any other table.

Stem is the 10ns Digit ('3' in case of 32 and 38; '4' in case of 46,44,47,48)

Scatter Plots

Scatter Plots locates a point for each pair of two variables that represent an observation

Salary vs CGPA for 10 MBA students

CGPA	Salary
6.5	7
7	7.6
7.2	8.1
7.5	8.8
8	10
8.2	11
8.4	12.5
8.6	14
9	15
9.2	17

This point shows the that at the CGPA of 6.5 the salary that an employee gets is 7 alkhs.
-- Thus in Scatter plot you locate a point which represents a pair of data (here, CGPA and Salary)

Measures of central tendency

Mean, Median, Mode

$$Sample\ mean\ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mean = 64.5

Median = 64

Mode = 66

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Sorted data

-- Sigma = Summation = which has i and n to it. I = 1, which means that the sum will start from the 1st item in the data and will go onto 'N' item.

-- So here i = 94 and n is 50th item i.e, 32.
All it says is start from 'i' and continue till 'n' and sum all these together

Exercise

Pay package in lakhs for 50 students of DoMS is given below:

18	11	10.2	8.5	7.7
17.4	11	10.2	8.5	7.7
16.5	10.6	9.9	8.4	7.7
15.9	10.6	9.9	8.4	7.7
11.2	10.6	9.9	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.5	9.6	8.3	6.9
11.2	10.5	9.3	8.2	6.9
11	10.5	9.3	7.7	6

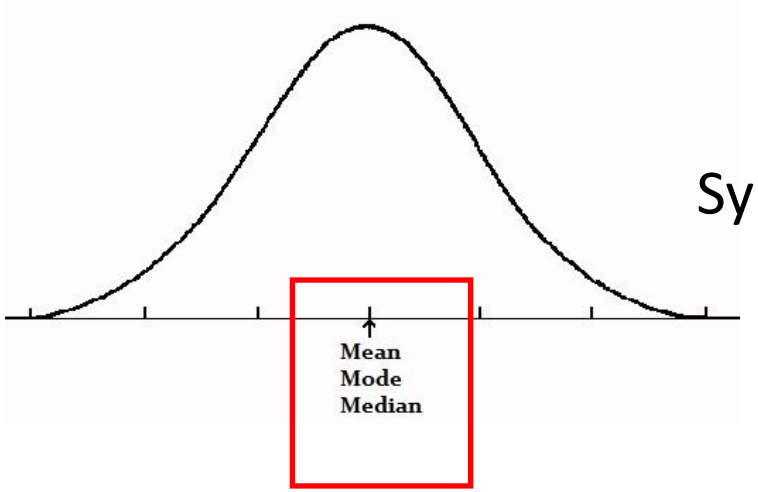
Compute the mean, median and mode

Median = 9.9+9.6 = 9.75

Mode - 7.7

Shape of the distribution

We can represent the data also in the form of curve and can draw some conclusions from the shape of the distribution

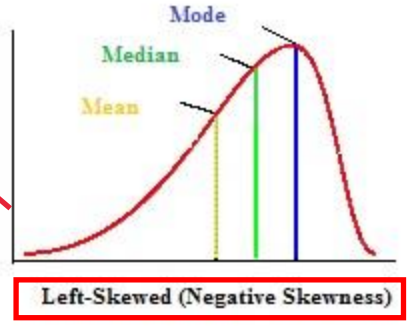


Symmetric

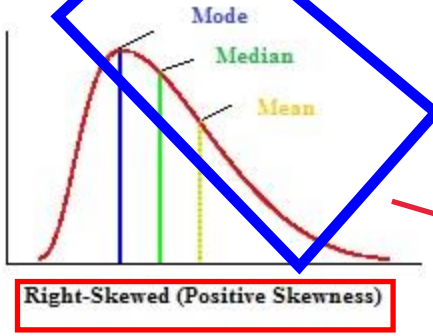
Here, Mean, Median and mode are all in the middle

skewed

Left Skewed distribution. Longer part comes to the left



marks



salary

Longer part of the Curve

Right Skewed distribution.
As the longer part comes to the right

Measures of variation (dispersion)

Range = L-S =
94-32 = range

Range, Inter quartile range

Variance and Standard deviation

Coefficient of Variation

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Sorted data

Five number summary of data

Median and Inter quartile range

53	52	63	59	62
48	47	66	54	67
62	72	46	53	68
58	77	38	66	83
66	60	78	90	88
73	88	62	32	73
89	94	68	47	62
92	73	67	64	59
66	71	67	56	44
57	64	71	63	47

Total marks of 50 students in a course

Sorted in ascending order

32	53	62	67	73
38	54	62	67	77
44	56	63	67	78
46	57	63	68	83
47	58	64	68	88
47	59	64	71	88
47	59	66	71	89
48	60	66	72	90
52	62	66	73	92
53	62	66	73	94

Mode = 62, 66 ← Bimodal mode

Median is the middle value
Median is the 50th percentile of the data
It is the second quartile of the data

Percentile and Quartile

Percentile is the value below which a given percentage of observations in a group of observations fall.

P^{th} percentile of the data is the smallest value in the list (in ascending order) such that no more than P% of the data points is strictly less than the value and at least P% is less than or equal to that value.

Calculate percentiles .

Find $\frac{P \times N}{100}$. If it is a fraction rank = upper integer value (n). If it is an integer,

rank = average of n and n+1 values

There are 4 quartiles. The first quartile is the 25% percentile, the second is the 50^{th} percentile (median), the third is 75^{th} percentile and the fourth is the last point which is 100^{th} percentile

If it is an Integer then rank will be the average of the n+(n+1) term of the data.

Say P = 25, N = 50 then P*N/100 = 12.5, thus the rank is a fraction so we will take the upper integer value = 13. So the 25%ile of the N is the 13th term (Sorted in ascending or descending order)

32	53	62	67	73
38	54	62	67	77
44	56	63	67	78
46	57	63	68	83
47	58	64	68	88
47	59	64	71	88
47	59	66	71	89
48	60	66	72	90
52	62	66	73	92
53	62	66	73	94

Minimum = 32

Ordinal rank of 25% percentile

Lower quartile =12.5; n = 13

25% percentile value = 56

Median = 50th Percentile = rank is the avg of 25th and 26th item. Thus the median is 64

Ordinal rank of 50% percentile = 25

Median = (64 + 64)/2 = 64

Ordinal rank of 75% percentile = 37.5

n = 38; 75th percentile = 72

75*50/100 =37.5 term or 38th term. =72

Minimum, 1st Quart, Median, 3rd Quart, Max.

Five number summary of data

32, 56, 64, 72, 94

IQR = 16

Inter Quartile range = 16

Maximum = 94

The difference b/w the 1st and the 4th Quartile. (72-56=16)

Exercise

Pay package in lakhs for 50 students is given below:

18	11	10.2	8.5	7.7
17.4	11	10.2	8.5	7.7
16.5	10.6	9.9	8.4	7.7
15.9	10.6	9.9	8.4	7.7
11.2	10.6	9.9	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.5	9.6	8.3	6.9
11.2	10.5	9.3	8.2	6.9
11	10.5	9.3	7.7	6

Compute the Five number summary of data, IQR and range?

25*50/100 = 12.5 = n = 13 = 10.6

Minimum = 6

Lower quartile = 8.3

50*50/100 = 25. n=25+26th term = 9.9+9.6/2 = 9.75

Median = 9.75

75^{th} percentile = 10.6

75*50/100 = 37.5 = n=38 =8.3

Maximum = 18

IQR = 2.3

Range = 12

Measures of variation (dispersion)

Inter quartile range

Variance and Standard deviation

Coefficient of Variation

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Sorted data

Median = 64
Mode = 66
Lower quartile = 72.5
Upper quartile = 56.5
IQR = 16

Minimum = 32
Lower quartile = 56.5
Median = 64
75^{th} percentile = 72.5
Maximum = 94

Five number summary of data

Mean of data from last page table

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \cdots + y_n}{n}$$

$\bar{y} = 64.5$ marks

Variance

--Each data is subtracted out of mean to find how they deviate from the mean.
--The difference is then squared to create a positive value.
The sum of all the Squared no. is the variance.
--Unit of variance is say Marks squared (For the data of marks scored), rupees squared(For the data about money)

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}$$

Square root of variance is SD. Unit is marks, rupees, etc.

Variance = 195.85

Standard deviation = 13.99

Coefficient of variation = 0.217

--The co-efficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.
-- The co-efficient of variation (CV) indicates the size of a standard deviation in relation to its mean. The higher the co-efficient of variation, the greater the dispersion level around the mean.

SD / Mean = COV

Coefficient literally means the factor that measures a particular property.

Six months earnings of a businessman is given: 5.4, 7.3, 10.9, 3.2, 4.7, 11.4. Find the mean and variance?

Month	Earnings	Deviation	Squared
1	5.4	5.4 – 7.15 = -1.75	3.0625
2	7.3	7.3 – 7.15 = 0.15	0.0225
3	10.9	3.75	14.0625
4	3.2	-3.95	15.6025
5	4.7	-2.45	6.0025
6	11.4	4.25	18.0625
Sum	42.9	0	56.815

$$Mean = \frac{42.9}{6} = \mathbf{7.15} \qquad variance = \frac{56.815}{5} = \mathbf{11.363}$$

Variance

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}$$

$$s^2 = \frac{9596.5}{49} = 195.85\ marks\ squared$$

$$s^2 = \frac{310.7402}{49} = 6.342\ lakhs\ squared$$

How do I understand lakhs squared?

Take the square root so that the unit of measurement is the same

$$s = \sqrt{s^2} = \sqrt{195.85} = 13.99\ marks$$

$$s = \sqrt{6.342} = 2.518\ lakhs$$

Role of standard deviation

100 gems chocolate balls were weighed and the mean weight was found to be 2.54 grams. The standard deviation = 0.022
How many pieces are in a 50 gram packet?

Mean = 2.502, number = 50/2.54 = 19.98 = 20.
Due to standard deviation some packets may either way less than 50 g if you put 20 chocolates or we have to pack more than 20 to take care of variation

Reduce process variation. Introduces to the concept called 6 sigma

Lecture 7

Describing numerical data

Role of standard deviation – Calculating Risk

Year	Stock A	Stock B
1	10.8	9
2	12	14.2
3	13	16
4	12	8.3
5	12.2	12.5
Average	12	12
Std Dev	0.787	3.308

Mean = 12 for both the shares. Share B has a higher standard deviation than share A. It has higher risk.

Variance (or standard deviation) is a measure of risk

What happens when the averages are different?

Scores of a cricketer in the last 10 innings;

62, 0, 81, 10, 147, 48, 13, 38, 98, 0

Find the mean and standard deviation? How is the dispersion comparable to the average?

Total = 497; n = 10; average = 497/10 = **49.7** s = **45.7538**

In calculating s we divided by 10

$$Coefficient\ of\ variation\ C_v = \frac{\sigma}{\bar{X}} \times 100 = \frac{45.7538}{49.7} \times 100 = 0.92$$

C_v has no units.
It is appropriate when mean is not close to zero.
$C_v > 1$ means there is considerable variation

Scores of second cricketer in the last 10 test innings;

35, 141, 19, 1, 69, 54, 147, 46, 14, 103

Find the mean and standard deviation? How is the dispersion comparable to the average?

Total = 629; n = 10; average = **62.9** s = **49.1**

$$Coefficient\ of\ variation\ C_v = \frac{\sigma}{\bar{X}} \times 100 = \frac{49.1}{62.9} \times 100 = 0.78$$

Can you say who is better?

Data – Marks of 50 students

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Describing the data

Summary Statistics	
Mean	64.5
Standard Error	1.979126
Median	64
Mode	66
Standard Deviation	13.99453
Sample Variance	195.8469
Kurtosis	-0.02813
Skewness	0.16293
Range	62
Minimum	32
Maximum	94
Sum	3225
Count	50

$$Standard\ error\ of\ mean = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

$$Skewness\ \gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$ Measure of asymmetry of data

$$Kurtosis\ Kurt[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$ Measure of tailedness of data

Measures of relationship between variables

Covariance
Correlation coefficient

Year	Stock A	Stock B
1	10.8	9
2	12	14.2
3	13	16
4	12	8.3
5	12.2	12.5
Average	12	12
Std Dev	0.787	3.308

$$Covariance\ (X,Y)\sigma_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$Correlation\ coefficient\ r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r = \frac{1.54}{0.704 \times 2.959} = 0.739$$

X	Y	(X-12)	(Y-12)	Product
10.8	9	-1.2	-3	3.6
12	14.2	0	2.2	0
13	16	1	4	4
12	8.3	0	-3.7	0
12.2	12.5	0.2	0.5	0.1
Sum				7.7
Covariance				1.54

r lies between +1 and -1. When covariance is negative, correlation coefficient becomes negative.

Example – Scores of 2 players

	Player 1	Player 2	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	62	35	12.3	-27.9	151.29	778.41	-343.17
2	0	141	-49.7	78.1	2470.09	6099.61	-3881.57
3	81	19	31.3	-43.9	979.69	1927.21	-1374.07
4	10	1	-39.7	-61.9	1576.09	3831.61	2457.43
5	147	69	97.3	6.1	9467.29	37.21	593.53
6	48	54	-1.7	-8.9	2.89	79.21	15.13
7	13	147	-36.7	84.1	1346.89	7072.81	-3086.47
8	38	46	-11.7	-16.9	136.89	285.61	197.73
9	98	14	48.3	-48.9	2332.89	2391.21	-2361.87
10	0	103	-49.7	40.1	2470.09	1608.01	-1992.97
Average	49.7	62.9			20934.1	24110.9	0
SD	45.7538	49.1			45.7538	49.10285	-9776.3
							Covariance **= -977.63**

$$r = \frac{-977.63}{45.75 \times 49.1} = -0.435$$

Negative covariance reduces risk and results in negative correlation

Fitting a linear relationship

Practice Tests taken (X)	Marks obtained in final (Y)
4	62
5	74
6	80
7	86
8	95

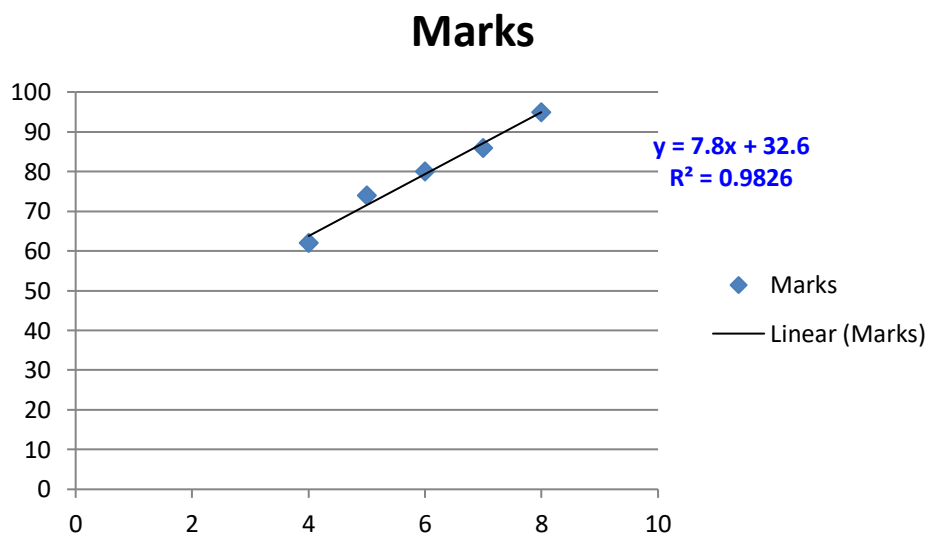
$\bar{X} = 6$ $\bar{Y} = 79.4$

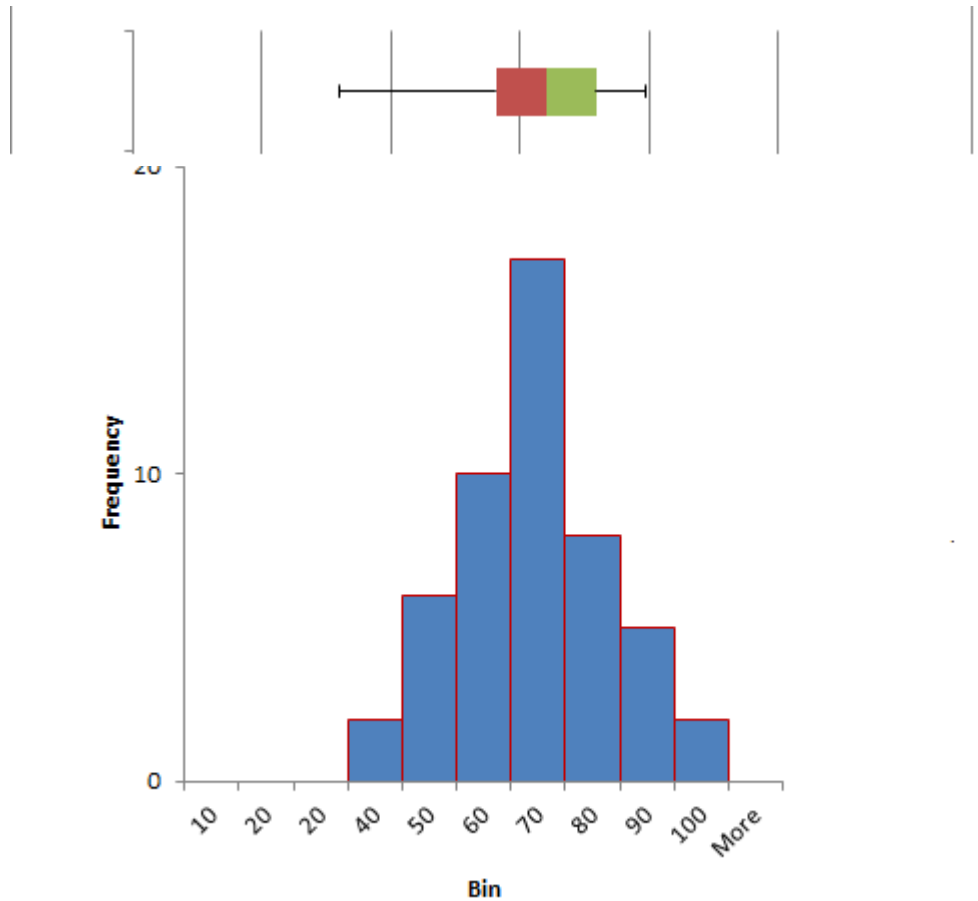
$\sigma_x = 1.414$ $\sigma_y = 11.128$

In calculating s we divided by 5

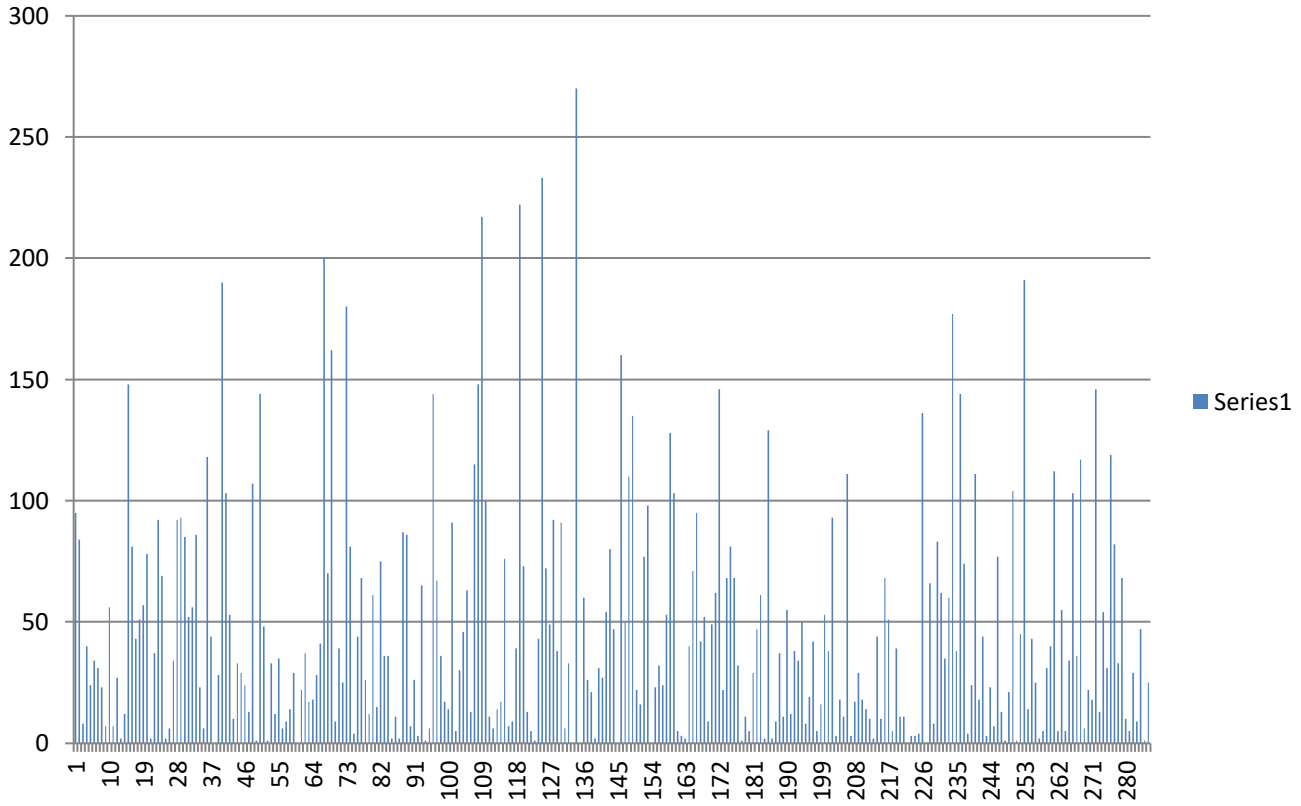
$\sigma_{xy} = 15.6$

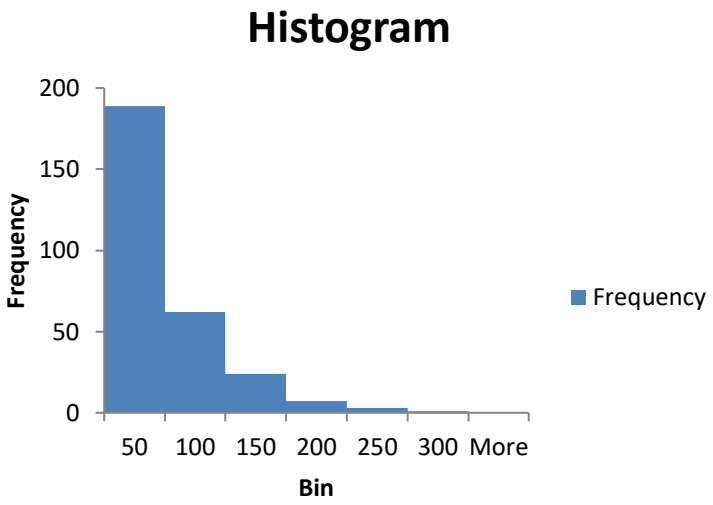
$r = 0.991$

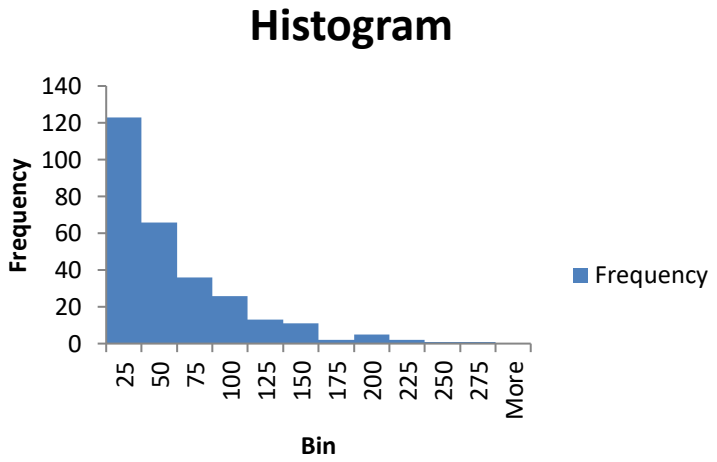


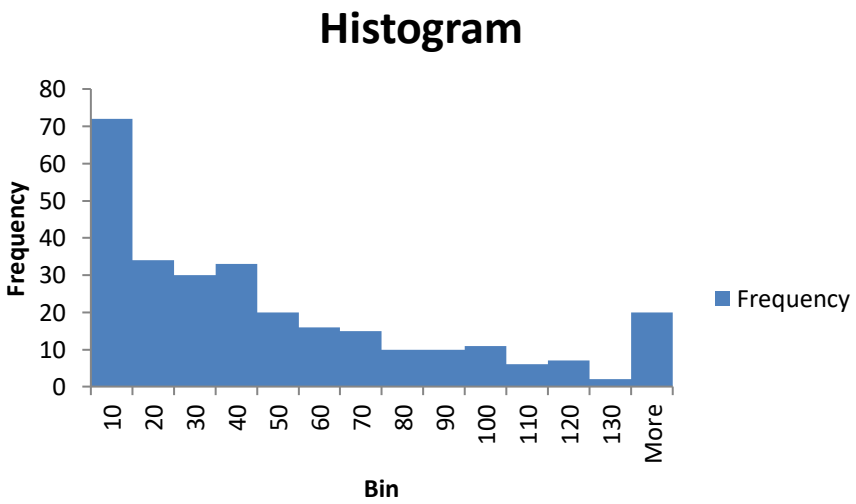

Frequency
10
0
10
20
20
40
50
60
70
80
90
100
More
Bin

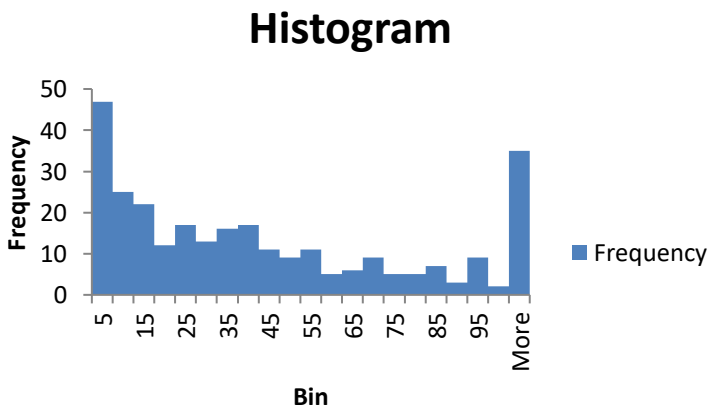
286 innings – Test Scores of a cricketer

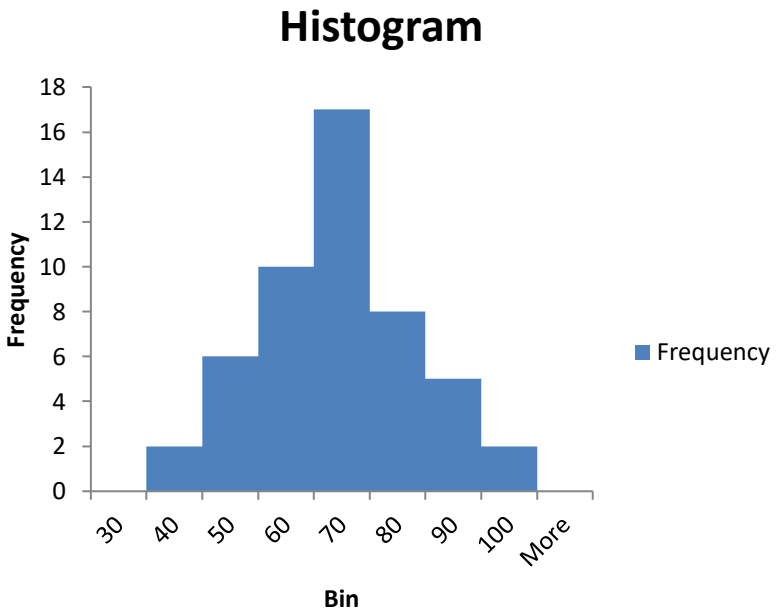


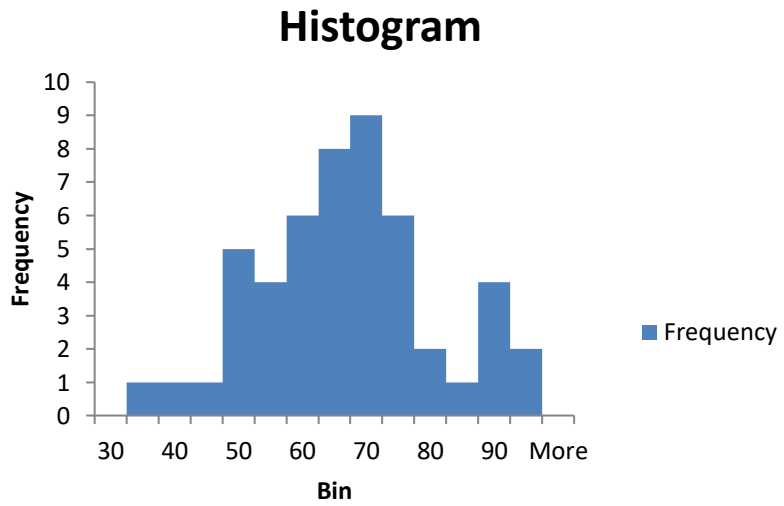

Histogram
200
150
100
50
0
Frequency
50 100 150 200 250 300 More
Bin
Frequency


Histogram
140
120
100
80
60
40
20
0
Frequency
25 50 75 100 125 150 175 200 225 250 275 More
Bin
Frequency

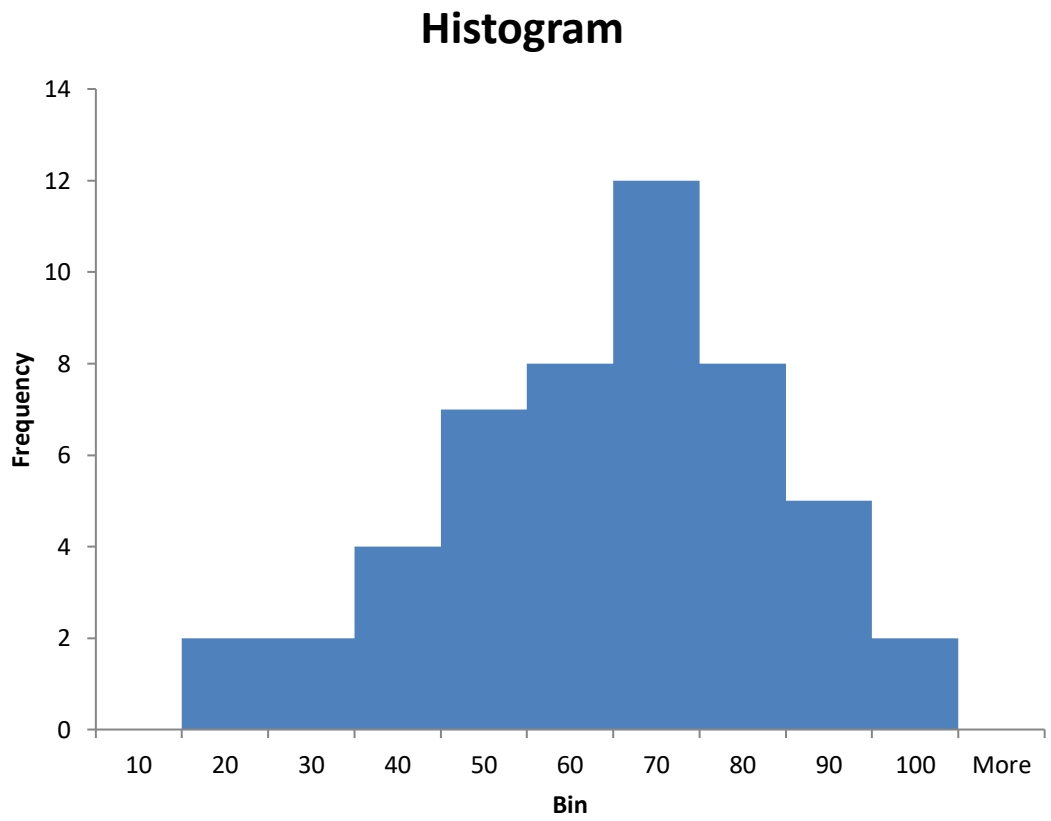

Histogram
80
70
60
50
40
30
20
10
0
Frequency
10 20 30 40 50 60 70 80 90 100 110 120 130 More
Bin
Frequency


Histogram
50
40
30
20
10
0
Frequency
5 15 25 35 45 55 65 75 85 95 More
Bin
Frequency

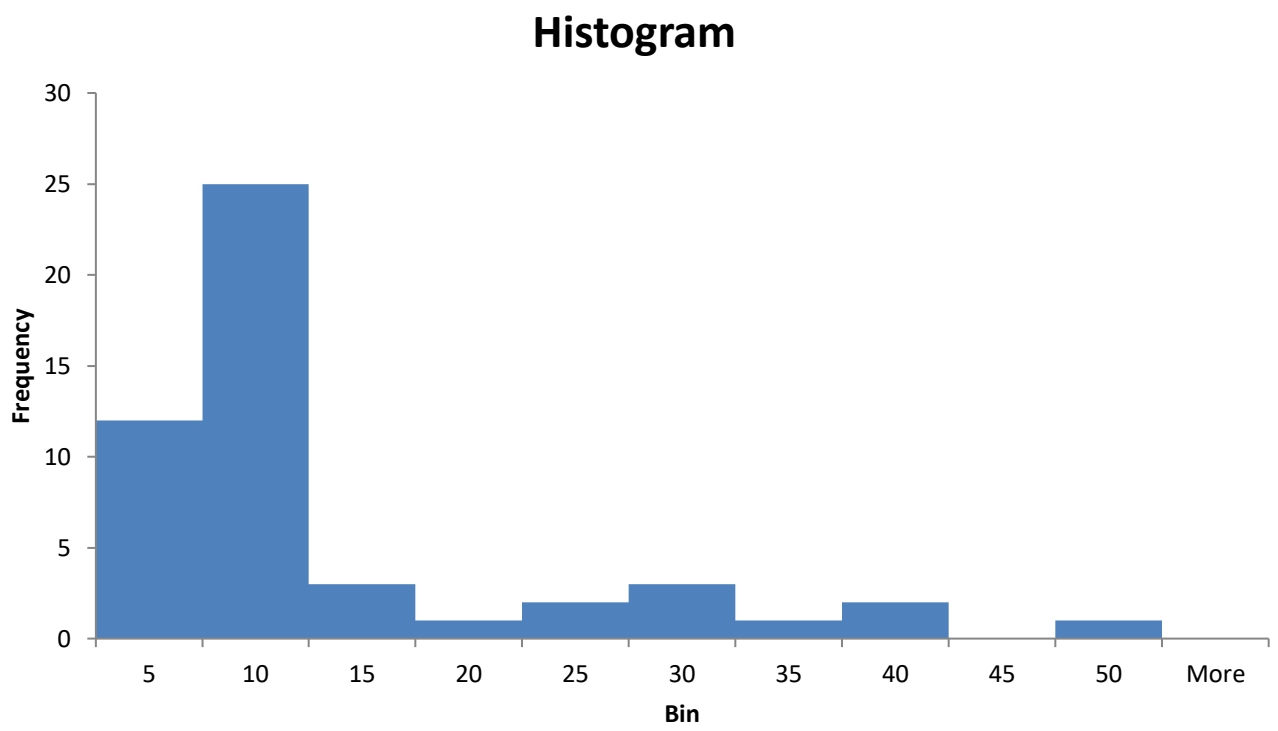




For the same data, narrow intervals produce multi modes

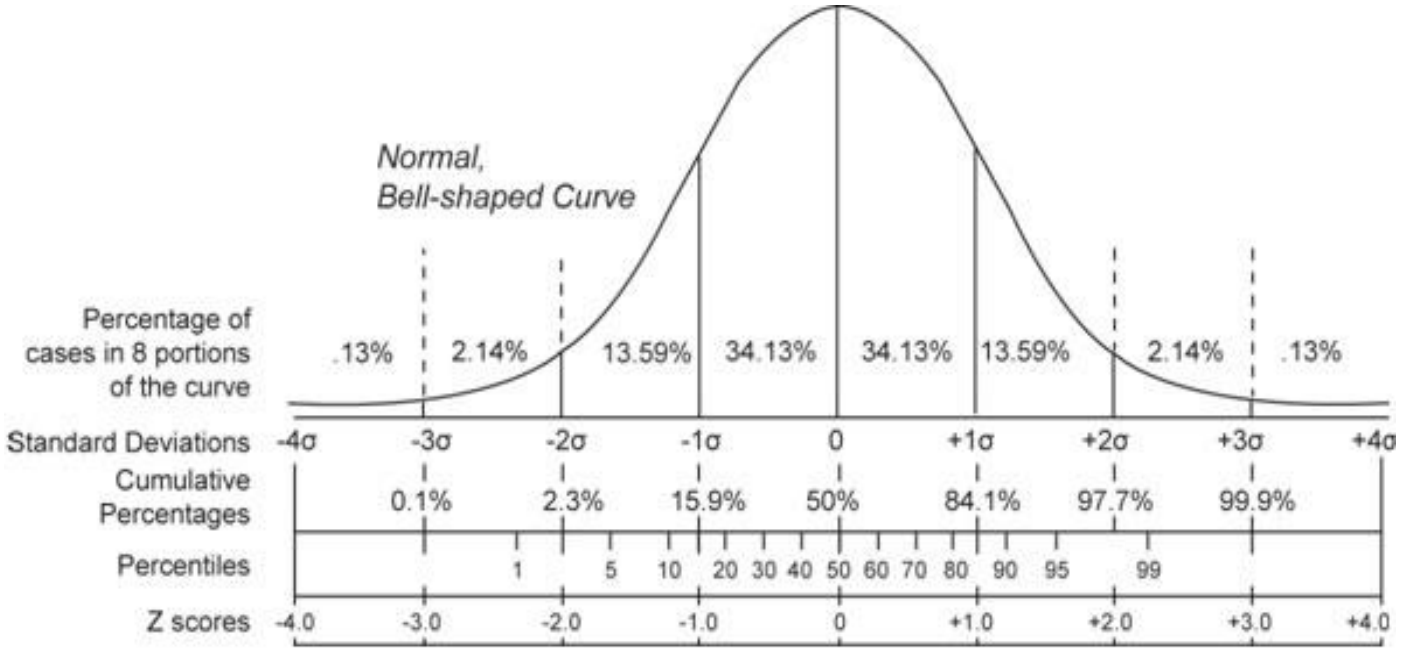


Marks distribution is Skewed to the left



Salary distribution is Skewed to the right

Bell shaped distribution and empirical rule



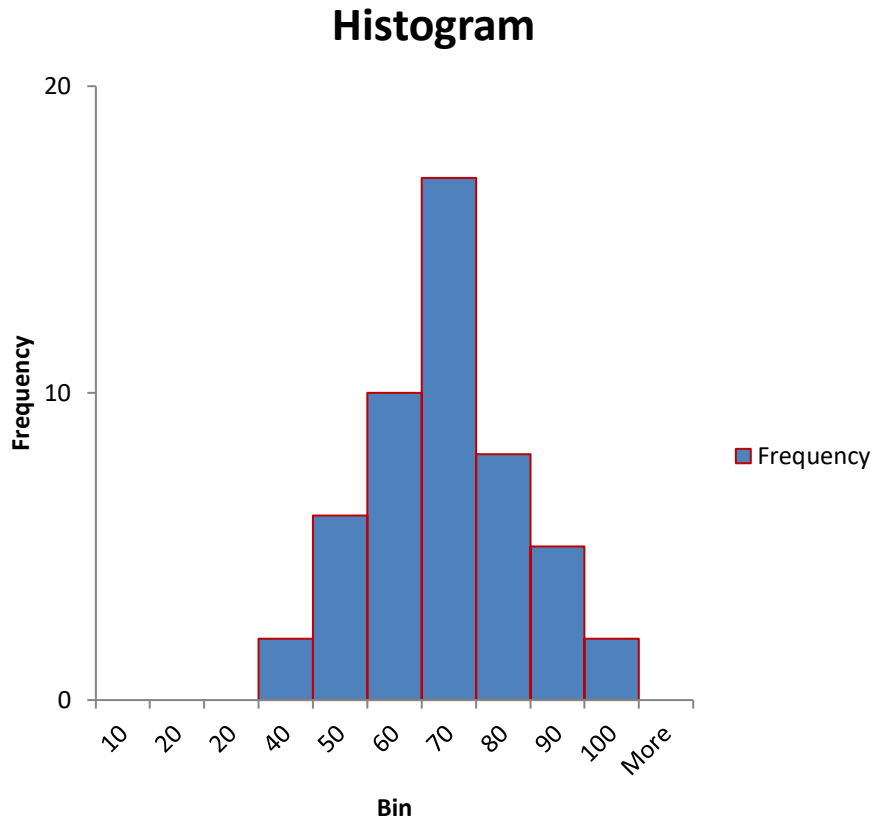
68% lies within y - s to y + s

95% lies within y - 2s to y + 2s

99.7% lies between y – 3s to y + 3s

Histogram

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32



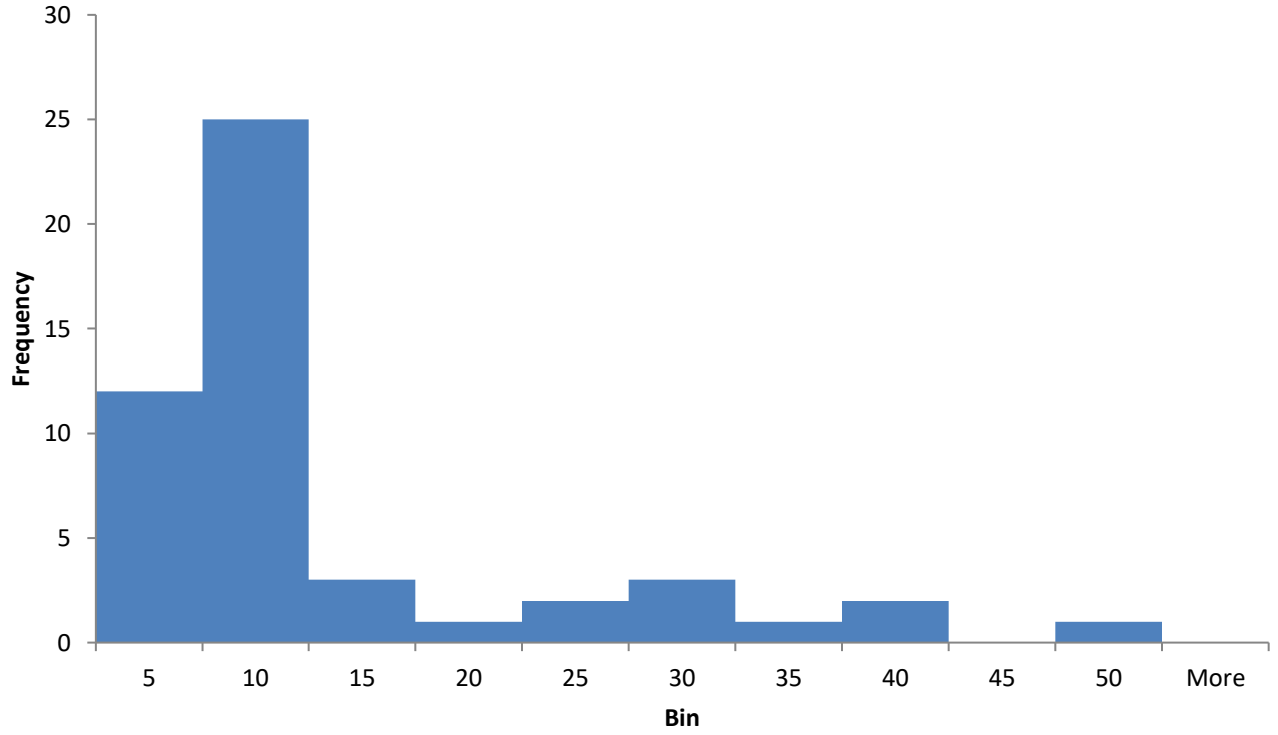
y = 64.5; s = 14

y – s to y + s = 51 to 79; 35 points vs 68% of 50 is 34

y – 2s to y + 2s = 37 to 93; 48 points vs 95% of 50 is 48

y – 3s to y + 3s = 23 to 100; 50 points vs 99.7% of 50 is 50

Salary distribution is Skewed to the right



y = 12; s = 10

y – s to y + s = 2 to 22; 42 points vs 68% of 50 is 34

y – 2s to y + 2s = 0 to 32; 47 points vs 95% of 50 is 48

y – 3s to y + 3s = 0 to 100; 50 points vs 99.7% of 50 is 50

Lecture 8

Discussion

Describing numerical data

Match the following

No.	Column A	Column B
1	Position of the peak	Median
2	Half the values are smaller	Standard deviation
3	Length of box in a boxplot	Interquartile range
4	Histogram with a long right tail	z score
5	Average squared deviation from the average	2/3
6	Square root of variance	mode
7	Number of standard deviations from the mean	variance
8	Proportion of bell shaped curve within one s d from mean	skewed

Mode

Median

Inter quartile range

Skewed

variance

Standard deviation

z score

2/3

True or false

1. Box plot shows mean plus one standard deviation of data
2. If data is right skewed, mean is larger than median
3. Removal of an outlier with z = 4 decreases the mean
4. Variance increases as the number of observations increases
5. If standard deviation is zero. Mean = median

1. False. Shows lower quartile, median, upper quartile and whiskers. Whiskers are roughly 1.5 times IQR. Small number of data are outside the whiskers
2. True
3. True
4. False
5. True

Question 1

- The median size of hundred files is 2 MB. Will they fit into a 2GB pen drive? Does SD play a role here?

- Cant say. Plays a role

Question 2

The mean time taken by students to prepare for the exams is 20 hours with standard deviation of 5 hours. You spoke to one of your friends and he said that he spent 26 hours preparing for the exam. Would it be a surprise to you?

Z = 1.2 Not surprising

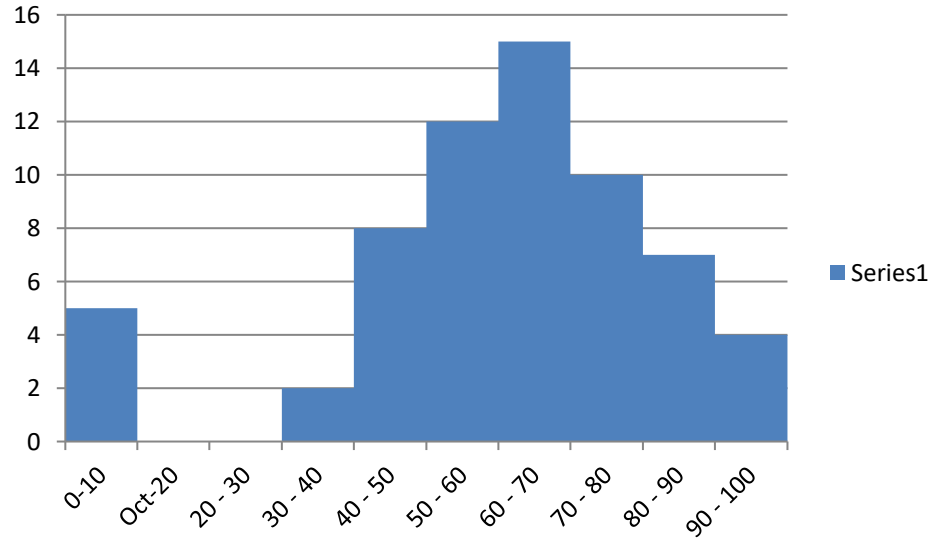
Question 3

Would you expect the distribution of the following to be uniform, unimodal, bimodal, symmetric or skewed?

1. Number of songs in the computer of 100 students
2. Heights of students in a class of 50 students
3. Exact weight of 500 gram biscuit packets in a factory
4. Bill value in a supermarket

1. Number of songs in the computer of 100 students – Right skewed with a single peak at zero
2. Heights of students in a class of 50 students – bimodal with men/women
3. Exact weight of 500 gram biscuit packets in a factory - normal
4. Bill value in a supermarket – Right skewed with one mode

Question 4



1. Which is larger – mean or median?
2. How many students have got marks between 30 to 50?
3. Find the mean?
4. Is the standard deviation close to 20 or 50? Why?

1. median
2. How many students have got marks between 30 to 50?
3. Find the mean?
4. Is the standard deviation close to 20 or 50? Why?

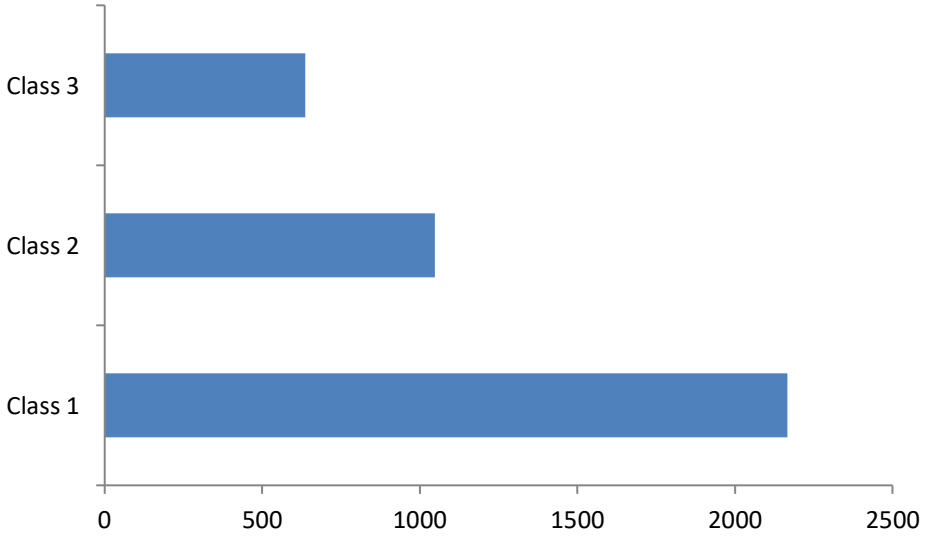
Association between categorical variables

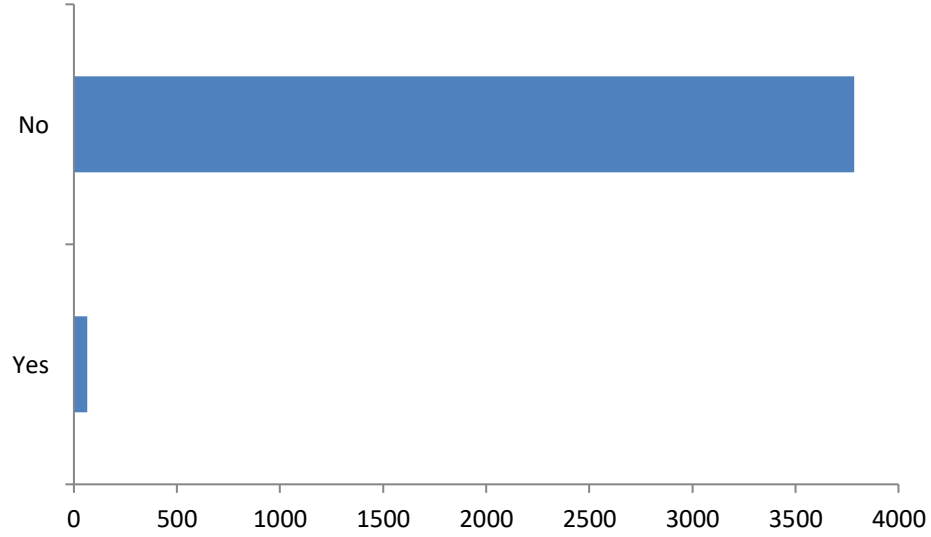
Chapter 5

Getting into a management school (data imaginary)

Preparation class	Number
Class 1	2166
Class 2	1047
Class 3	636
Total	3849

Out of the 3849 applicants, 65 joined





Contingency table shows counts of cases of one categorical variable contingent on the value of another

Joined		Preparation class			
		Class 1	Class 2	Class 3	Total
	Yes	37	18	10	65
	No	2129	1029	626	3784
	Total	2166	1047	636	3849

The cells of the Contingency table are mutually exclusive. Each case appears exactly in one cell.

The right margin shows the frequency distribution of the selected people. It is also called marginal distribution

Percentages

10 students from Class 3 joined the program
This is 0.26% of all the students who applied
This is 1.57% of the students who went to Class 3
This is 15.38% of the students who joined the program

Joined		Preparation class			
		Class 1	Class 2	Class 3	Total
	Yes	37	18	10	65
		0.96%	0.47%	0.26%	1.69%
		1.71%	1.72%	1.57%	
		56.92%	27.69%	15.38%	
	No	2129	1029	626	3784
		55.31%	26.73%	16.26%	98.31%
		98.29%	98.28%	98.43%	
		56.26%	27.19%	16.54%	
	Total	2166	1047	636	3849
		56.27%	27.2%	16.52%	

We are interested in knowing which preparation class produces the highest proportion of students joining

Joined		Preparation class			
		Class 1	Class 2	Class 3	Total
	Yes	37 1.71%	18 1.72%	10 1.57%	65 1.69%
	No	2129 98.29%	1029 98.28%	626 98.43%	3784 98.31%
	Total	2166	1047	636	3849

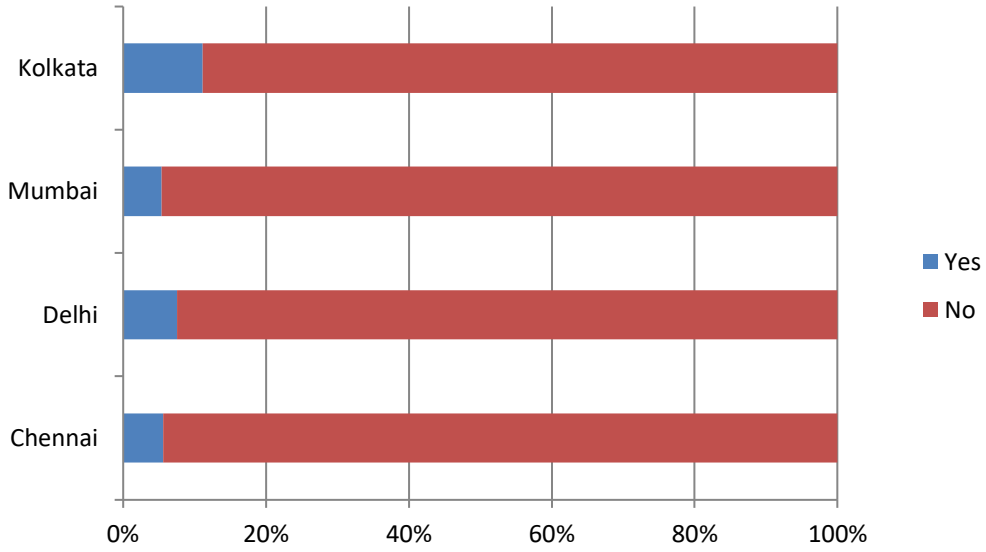
The distribution of a variable that is restricted to cases satisfying a condition is called conditional distribution.

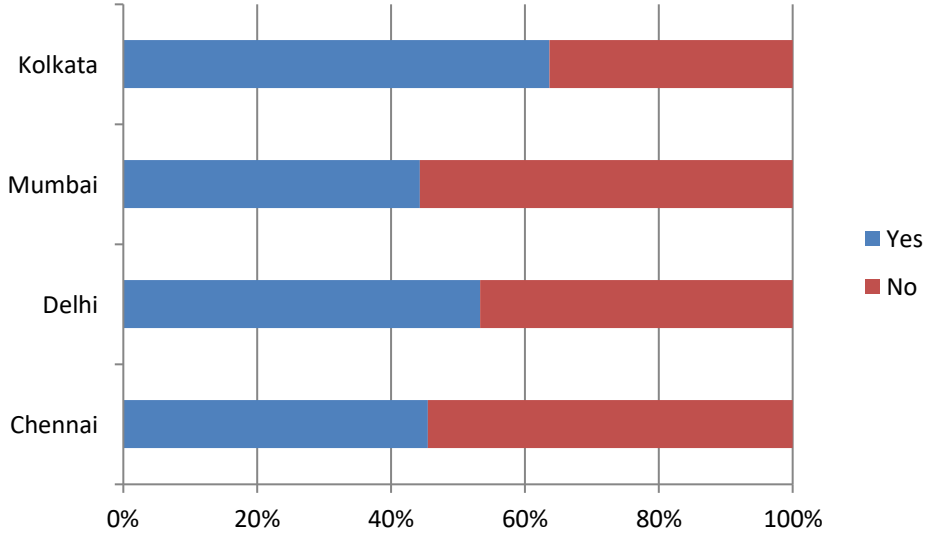
Conditional distribution restricts itself to a row or column

We are interested in knowing which preparation class produces the highest proportion of students joining

	Interview Zone				
	Chennai	Delhi	Mumbai	Kolkata	Total
Yes	18 5.63%	23 7.54%	14 5.39%	10 11.11%	65 6.66%
No	302 94.37%	282 92.46%	246 94.61%	80 88.89%	910 93.33%
Total	320	305	260	90	975

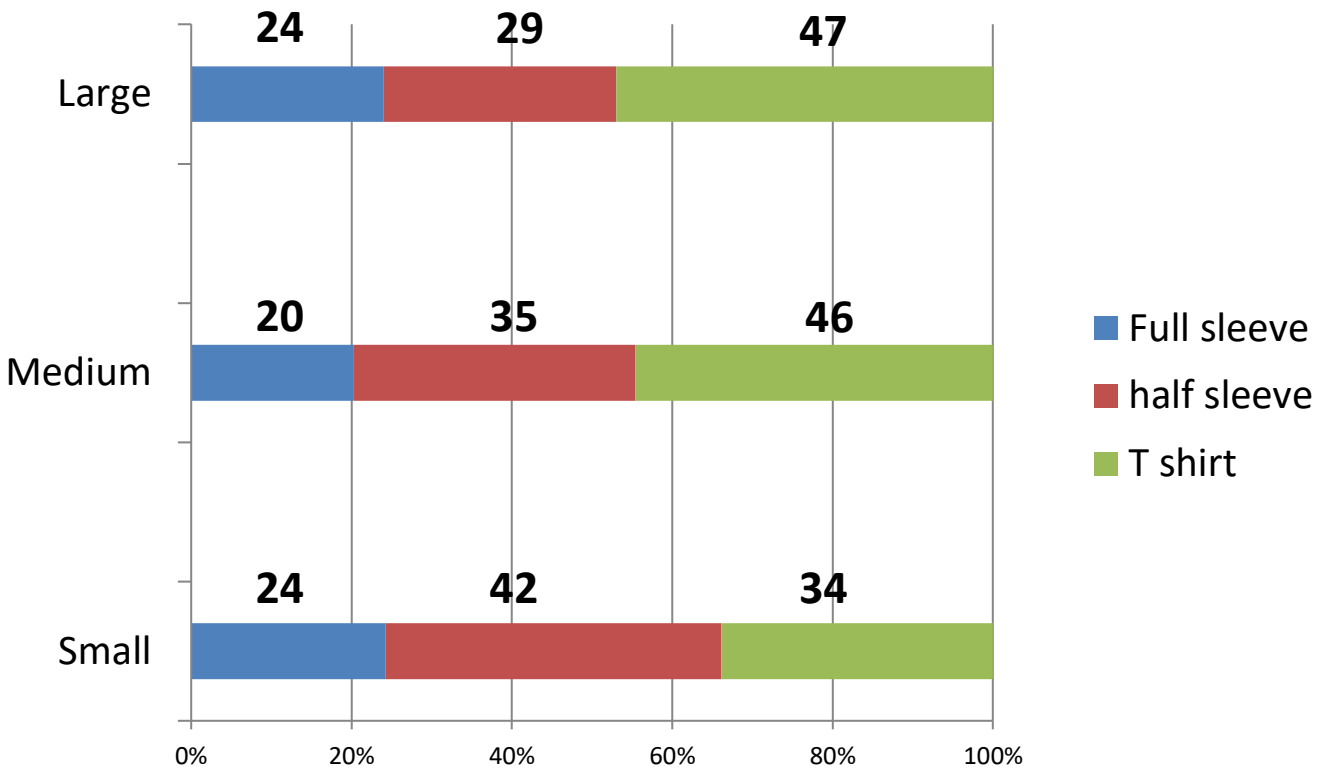
Conditional distribution restricts itself to a row or column



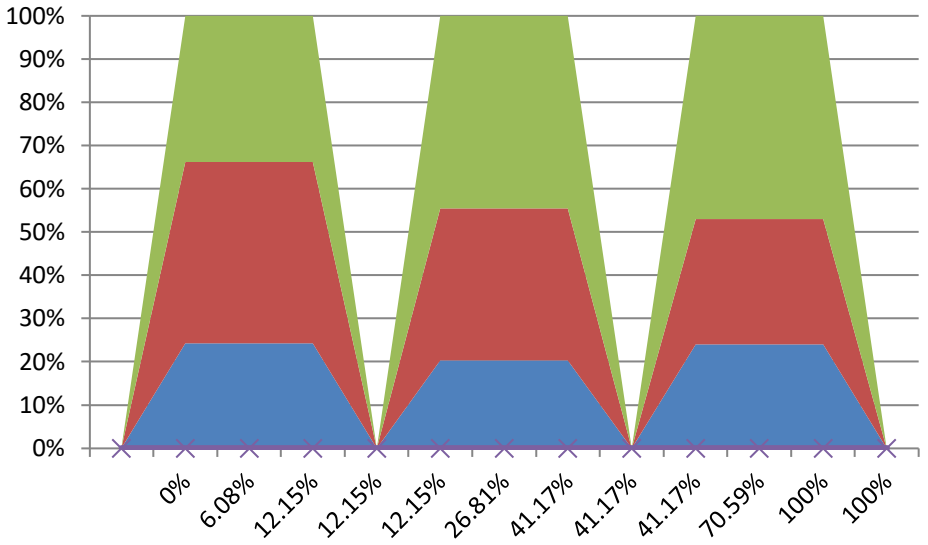


(based on % in yes/no)

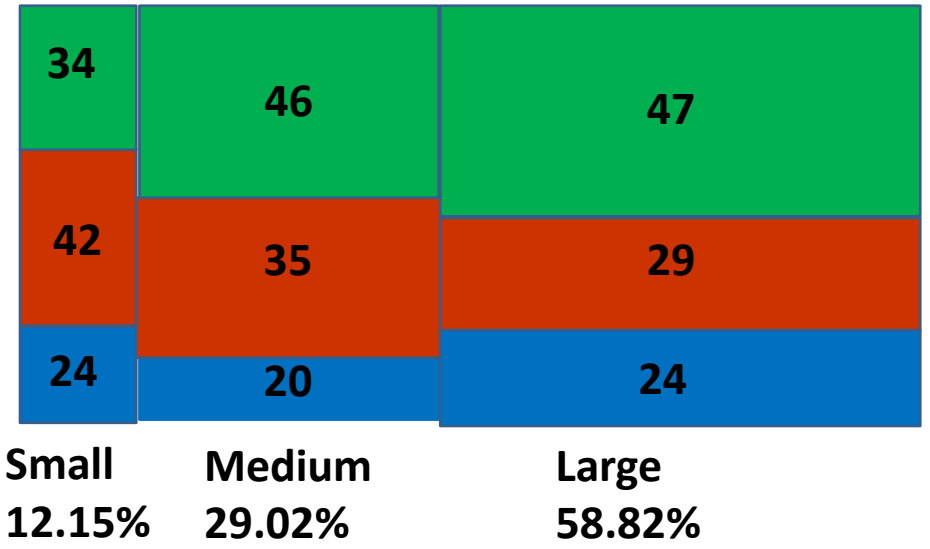
	Type of shirt			
	Full sleeve	Half sleeve	T shirt	Total
Small	15	26	21	62 12.15%
Medium	30	52	66	148 29.02%
Large	72	87	141	300 58.82%
Total	117	165	228	510



Segmented Bar Chart



From Excel



Mosaic Plot

Lecture 9

Association between categorical variables

	Airline XX	Airline YY	Total
On time	86 72%	81 81%	167 76%
Delay	34 28%	19 19%	53 24%
Total	120	100	220

YY has a better on time departure performance

Consider type of flight
Point to point hopping

	Airline XX		Airline YY		Total
	Hopping	Point to point	Hopping	Point to point	
On time	50	36	45	36	167
	81%	62%	78%	86%	76%
Delay	12	22	13	6	53
	19%	38%	22%	14%	24%
Total	62	58	58	42	220

If it is a hopping flight, XX has a better performance

The external variable that influences the performance is called a lurking variable and the change is called “Simpson’s paradox”

Exercise

Two cricketers A and B; Scores <50 and >50

Examples of lurking variables

Day match vs day-night
Team is batting first vs batting second
Batting position opening/middle order

	AA	BB	Total
≥ 30	23 46%	18 37%	41 41%
≤ 30	27 54%	31 63%	58 59%
Total	50	49	99

	AA		BB		Total
	First	Second	First	Second	
≥ 30	11	12	14	4	41
	48%	44%	52%	18%	41%
≤ 30	12	15	13	18	58
	52%	56%	48%	72%	59%
Total	23	27	27	22	99

Measuring association among categorical variables

Attitude towards attending classes when instructor does not take attendance

	Attend	Skip	
Fresh graduates	12	17	29
Work experience	28	15	43
Total	40	32	72

	Attend	Skip	
Fresh graduates	$= \frac{29 \times 40}{72}$	$= \frac{29 \times 32}{72}$	29
Work experience	$= \frac{40 \times 43}{72}$	$= \frac{32 \times 43}{72}$	43
Total	40	32	72

12	17
28	15

Data

16	13
24	19

Artificial (based on proportions)

-4	4
4	-4

Difference

$$\chi^2 = \frac{(12-16)^2}{16} + \frac{(17-13)^2}{13} + \frac{(28-24)^2}{24} + \frac{(15-19)^2}{19}$$

$$\chi^2 = 1 + 1.23 + 0.66 + 0.84 = 3.74$$

$$Cramer's\ V = \sqrt{\frac{\chi^2}{n \times min(r-1, c-1)}} \qquad Cramer's\ V = \sqrt{\frac{3.74}{72 \times 1}} = 0.23$$

2	27
38	5

Data

16	13
24	19

Artificial (based on proportions)

-14	14
14	-14

Difference

$$\chi^2 = 45.81$$

$$Cramer's\ V = \sqrt{\frac{\chi^2}{n \times min(r-1, c-1)}} \quad Cramer's\ V = \sqrt{\frac{45.81}{72 \times 1}} = 0.7976$$

V ≤ 0.25 shows weak association and ≥ 0.75 shows strong association

Association between scores

23	18
27	31

Data

21	20
29	29

Artificial (based on proportions)

2	-2
-2	2

Difference

$\chi^2 = 0.2758$

$$Cramer's\ V = \sqrt{\frac{\chi^2}{n \times min(r-1, c-1)}} \quad Cramer's\ V = \sqrt{\frac{0.2758}{99 \times 1}} = 0.053$$

V ≤ 0.25 shows weak association and ≥ 0.75 shows strong association

Discussion
on

Association between categorical variables

Match the following

No.	Column A	Column B
1	Table of cross classified counts	No association
2	Shown as bar chart	Cramers V
3	Measure of association between categorical variables	Contingency table
4	Measure of association between categorical variables (lies between 0 and 1)	Chi squared
5	Produced by a lurking variable	associated
6	Conditional distribution matches marginal distribution	cell
7	Percentage within row differs from marginal percentages	Marginal distribution
8	Cases that match two categorical variables	Simpson's paradox

Contingency table

Marginal distribution

Chi squared

Cramer's V

Simpson's paradox

No association

association

cell

True or false

1. We can fill cells of contingency table from marginal counts if the variables are not associated

2. The value of chi square depends on the number of observations in the contingency table

3. Cramer's V is zero when the variables are not associated

4. The value of chi squared depends on which two variables define the rows and which two define the columns

5. If male and female are values of a variable and if the percentage female is higher, there is association between variables

True, True, True, False, False

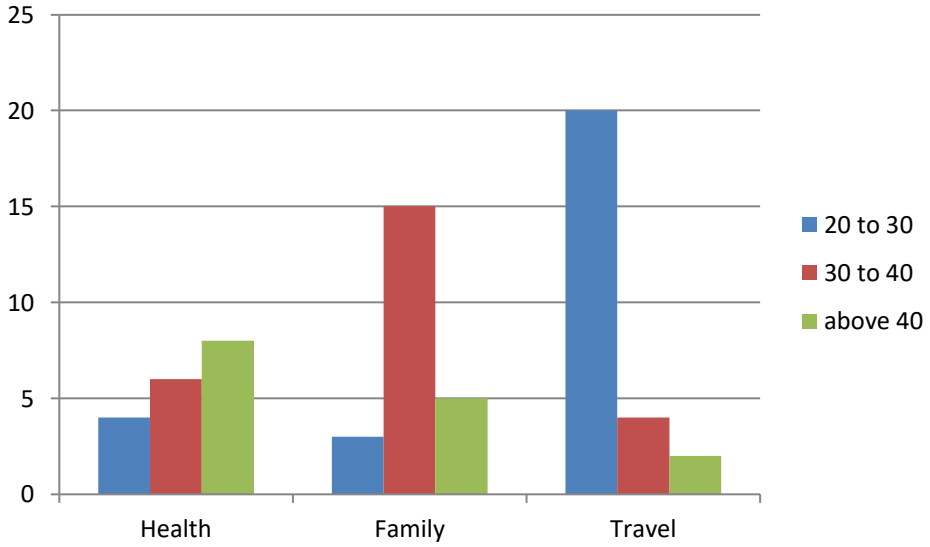
Question 1

- Customers were asked to give preferences for colour and shape of a product. Two teams were created by the company – each to determine the colour and shape of the product. Is it necessary to check if the two variables are associated?

Necessary to check. Good if there is no association. Otherwise the association has to be factored

Question 2

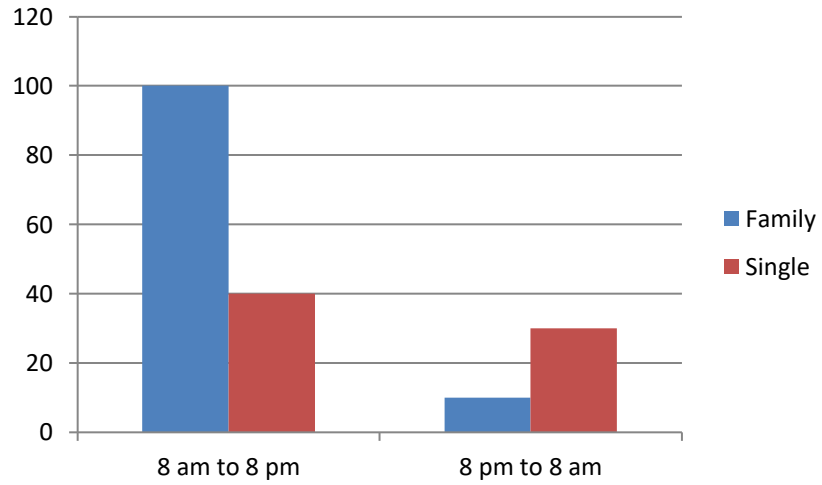
A survey was conducted to understand the reasons for absence of students in a research university. Three main reasons were identified and there were three broad categories of students. In which group, medical reasons dominated? Are the variables associated?



	Health	Family	Travel
20 to 30	4	3	20
30 to 40	6	15	4
above 40	8	5	2

Question 3

A survey was conducted in a 24x7 supermarket where two variables were considered – time of purchase (8 am to 8 pm and 8 pm to 8 am) and whether the buyers were single or family. The data is given below



Would you expect association in the data?

Yes. More families would come during day time.

Question 4

A survey indicated that the most popular colour for all cars is white. Should a dealer in cars stock all items in white?

Check association between types of buyers and colour.

Question 5

Find Cramers V for the following data?

	Red	Blue	White	
More than 30 lakh	20	30	40	90
Between 15 – 30 lakh	10	15	20	45
Less than 15 lakh	40	60	80	180
	70	105	140	315

	Red	Blue	White
More than 30 lakh	$= \frac{90 \times 70}{315} = 20$		
Between 15 – 30 lakh			
Less than 15 lakh			

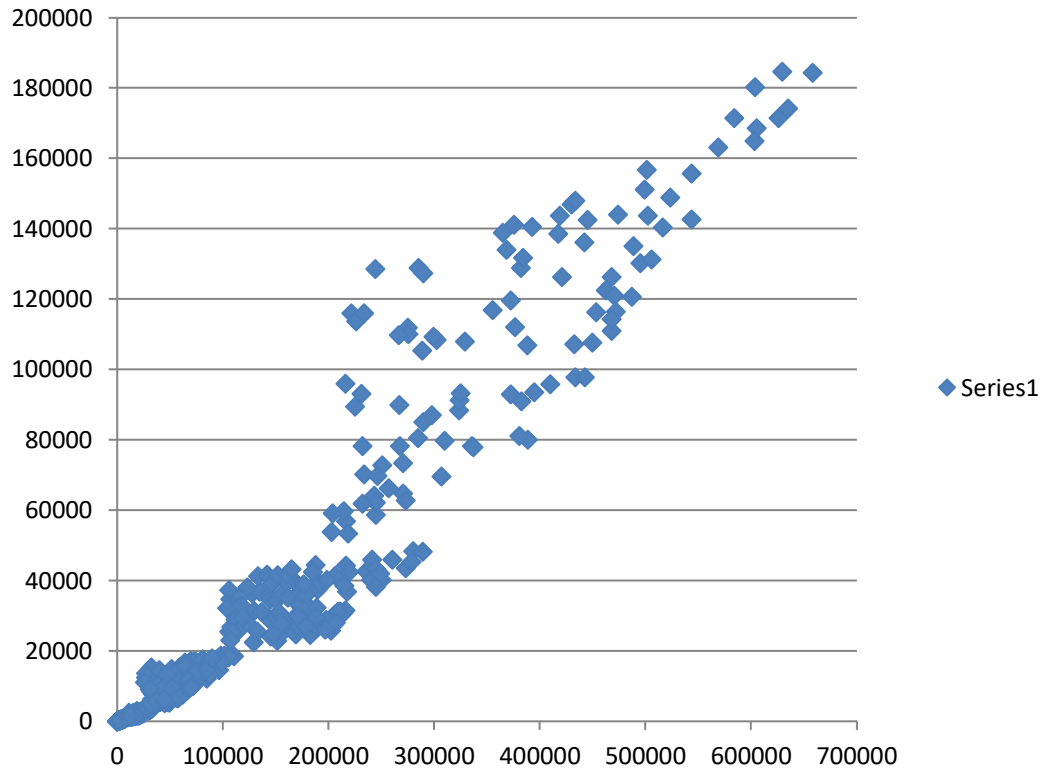
Cramers V = 0

Lecture 10

Association between quantitative variables

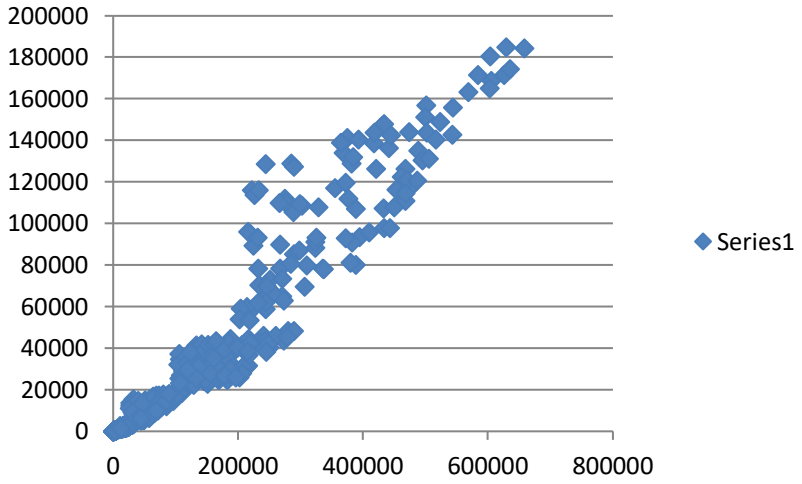
Scatter Plot

Salary vs saving

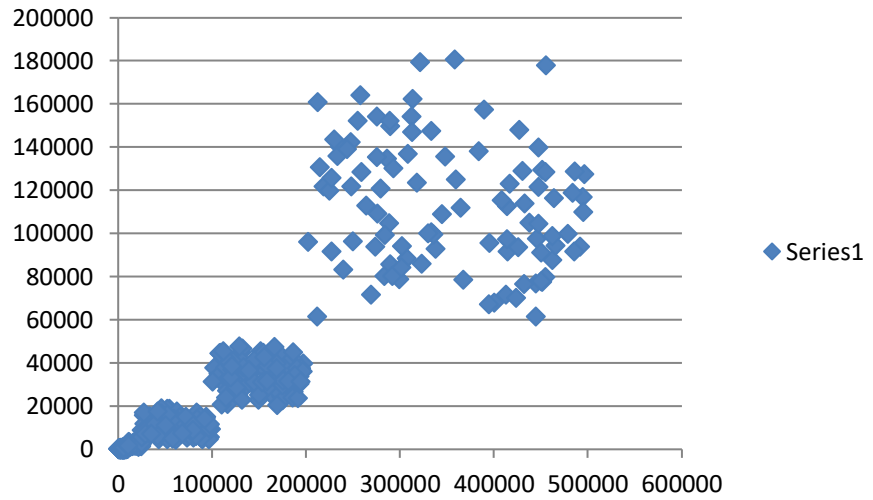


Which is x axis and which is y axis?

explanatory variable is x axis. Response is y axis



Data



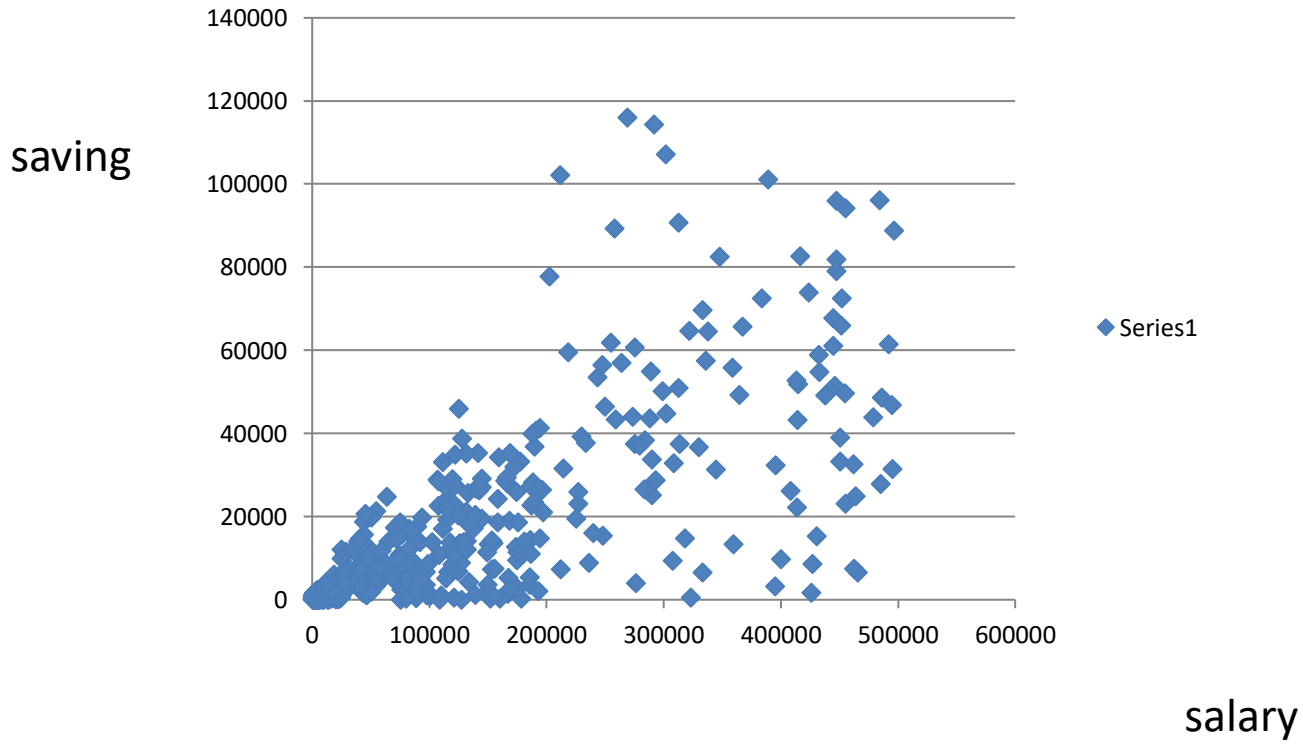
Random

Data and random look different. There is some association

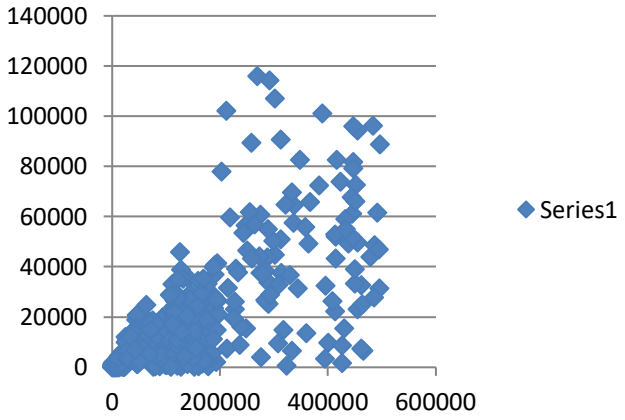
Describing association

1. *Trend* – upward or downward?
2. *Curvature* – Is it linear or does it show a curve?
3. *Variation* – Are points tightly clustered along the pattern?
4. *Outliers and surprises* – Are there outliers?

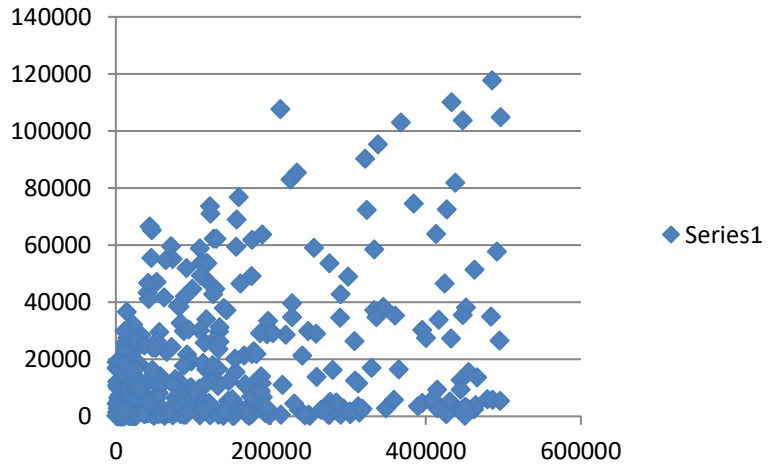
Salary and saving



Salary vs saving – another set of data



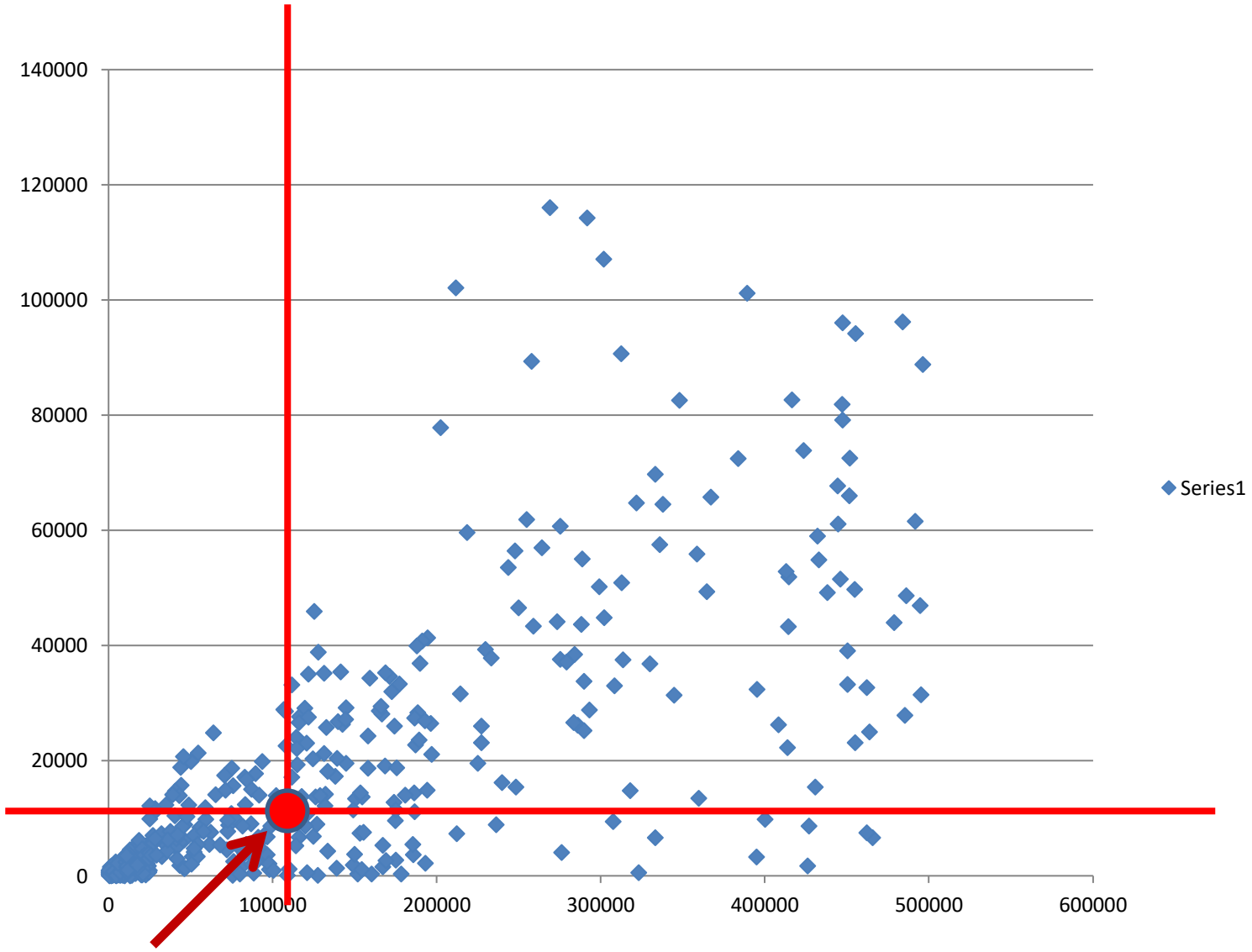
Data



Random

Do data and random look different?

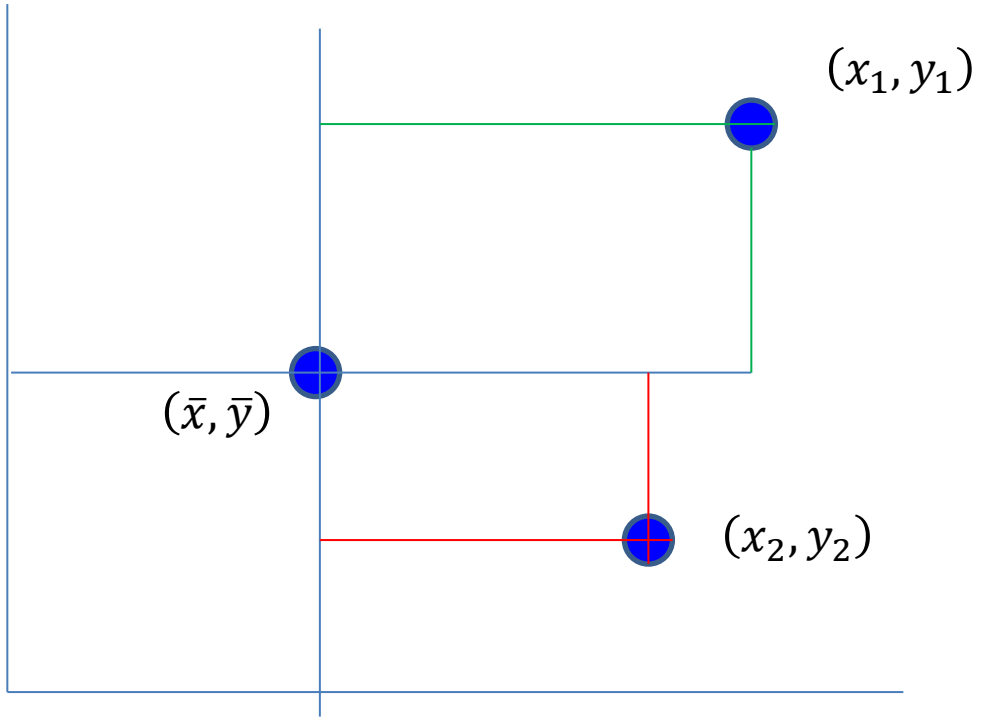
Is there an association? How do we compute the association?


140000
120000
100000
80000
60000
40000
20000
0
0
100000
200000
300000
400000
500000
600000
Series1

$(\bar{x}, \bar{y})$

Covariance

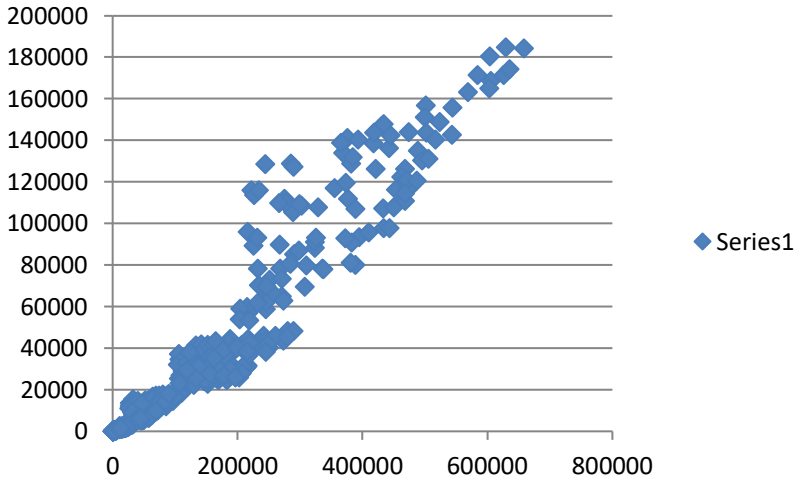
Covariance quantifies the strength of linear association between two numerical variables. Measures the degree to which data concentrates along the diagonal.

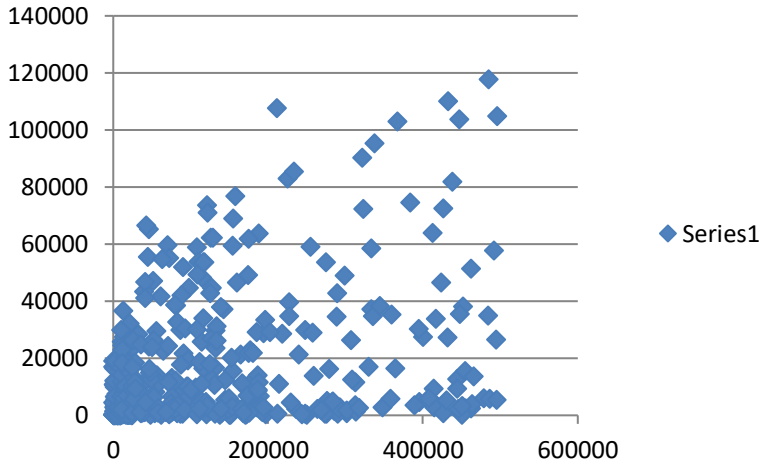


$$\boldsymbol{covariance} = \sum_{\boldsymbol{n}} \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}})\,(\boldsymbol{y} - \bar{\boldsymbol{y}})}{\boldsymbol{n}}$$

Is the denominator n or n-1?


200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
0
100000
200000
300000
400000
500000
600000
700000
Series1





$\bar{x} = 125523$

$\bar{y} = 31978$

$\bar{x} = 113403$

$\bar{y} = 14382$

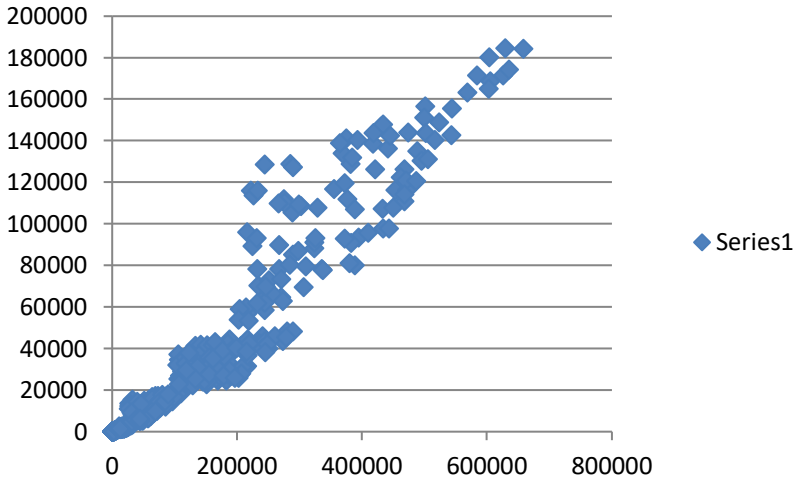
$$\boldsymbol{covariance} = \sum_{\boldsymbol{n}} \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}})\,(\boldsymbol{y} - \bar{\boldsymbol{y}})}{\boldsymbol{n}}$$

$covariance = 6282326586$

$covariance = 527175843$

What can we make out from these numbers?

Correlation



Correlation is a more easily interpreted measure of association from the covariance.

$$\bar{x} = 125523$$

$$\bar{y} = 31978$$

$$Correlation = \frac{covariance}{product\ of\ standard\ deviations}$$

$$\boldsymbol{covariance} = \sum_{\boldsymbol{n}} \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}})\ (\boldsymbol{y} - \bar{\boldsymbol{y}})}{\boldsymbol{n}}$$

$$covariance = 6282326586$$

$$Corr = \frac{cov}{s_x s_y}$$

$$s_x = 148035$$

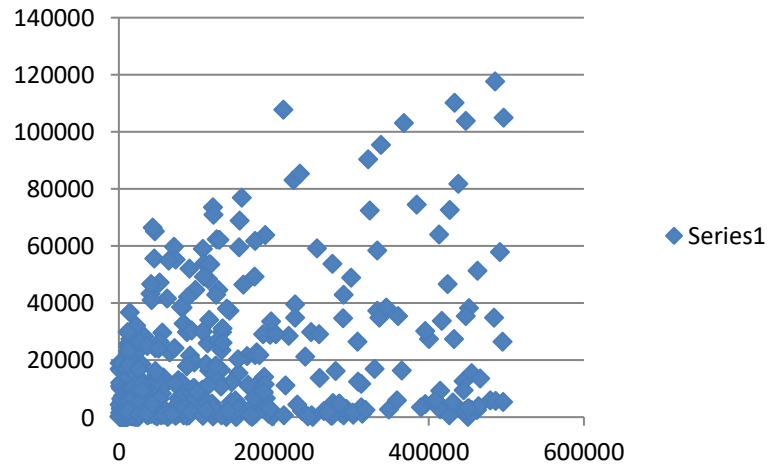
$$s_y = 44851.54$$

$$r = \frac{6282326586}{148035 \times 44851.54}$$

$$\boldsymbol{r = 0.9462}$$

Correlation

1. Correlation measures the strength of linear association
2. r is always between -1 and +1 ($-1 \leq r \leq 1$)
3. r does not have units.



$\bar{x} = 113403 \qquad \bar{y} = 14382$

$s_x = 134858 \qquad s_y = 44243.6$

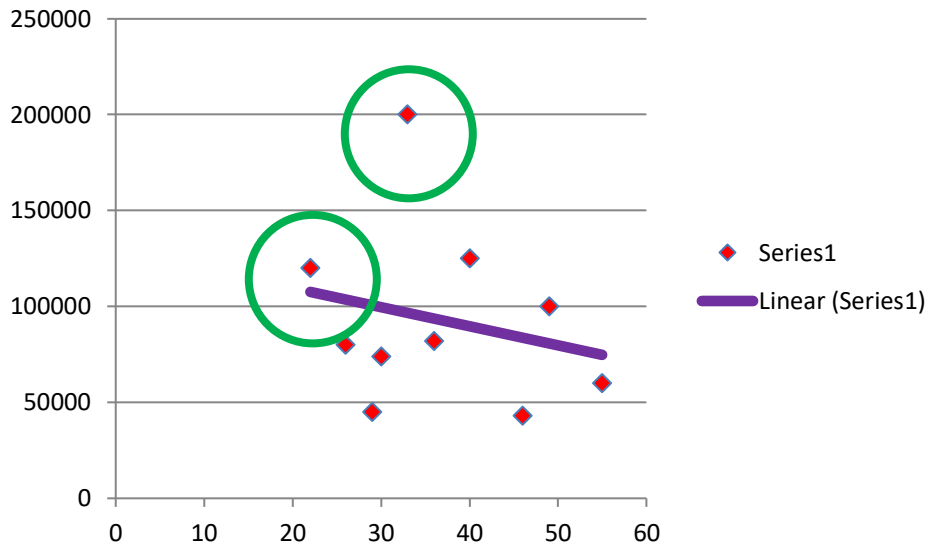
$$\boldsymbol{covariance} = \sum_{\boldsymbol{n}} \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}})\,(\boldsymbol{y} - \bar{\boldsymbol{y}})}{\boldsymbol{n}}$$

$$covariance = 527175843$$

$$Corr = \frac{cov}{s_x s_y}$$

$$\boldsymbol{r = 0.088}$$

Age	Salary
26	80000
22	120000
30	74000
36	82000
29	45000
55	60000
46	43000
49	100000
40	125000
33	200000



Is r positive or negative?

r = -0.22705

Age	Height	Weight
11	152	38
12	153	40
13	160	43
14	168	52
15	170	61
16	183	76
17	176	72
18	180	78
19	178	81
20	180	69

Correlation Matrix

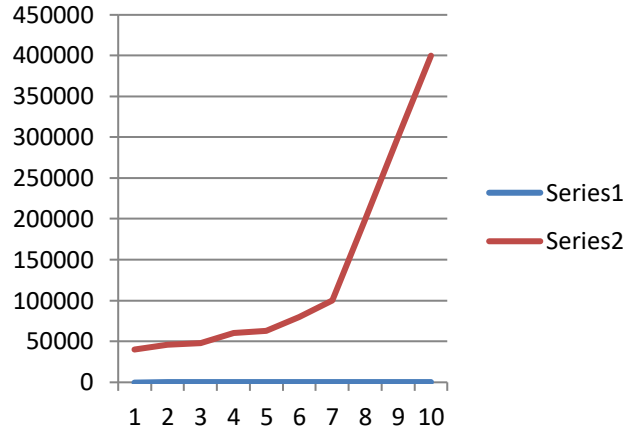
	Age	Height	Weight
Age	1.00	0.9015	
Height			
Weight			

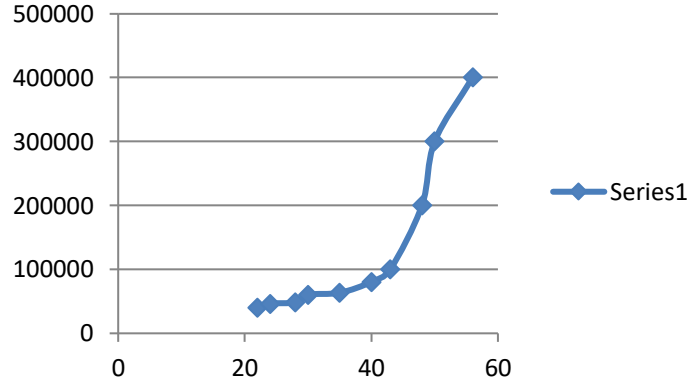
Maths	English	Science
77	75	69
82	80	80
46	66	43
62	56	52
59	64	61
100	77	76
92	80	72
87	85	78
56	51	61
64	42	69

	Maths	English	Science
Maths	1.00	0.9015	
English			
Science			

Age	Salary
22	40000
24	46000
28	48000
30	60000
35	63000
40	80000
43	100000
48	200000
50	300000
56	400000

Corr r = 0.8806

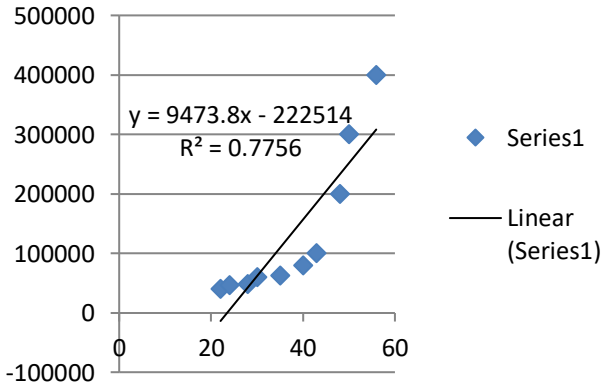




$$y = 9473.8x - 222514$$

$$R^2 = 0.7756$$

$$R = r = \sqrt{0.7756} = 0.8806$$



Correlation

Correlation measures the strength of linear association between variables
Larger $|r|$ becomes more closely the data cluster along a line
We can use r to find the equation of this line
We can predict y for a given x

Consider the z score of the two variables.
z score is the deviation from the mean divided by standard deviation.
Correlation converts z score of one variable into z score of another.

Correlation

$$z_x = \frac{(x - \bar{x})}{s_x} \qquad z_y = \frac{(y - \bar{y})}{s_y}$$

The equation of the line is $\hat{z}_y = r z_x$

$$\frac{(\hat{y} - \bar{y})}{s_y} = \frac{r(x - \bar{x})}{s_x}$$

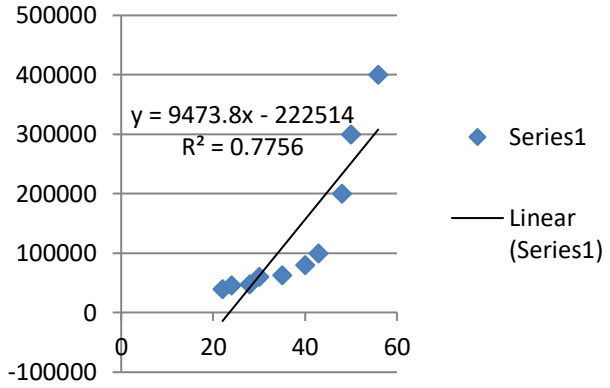
$$\hat{y} = \bar{y} + \frac{r s_y (x - \bar{x})}{s_x} = \left(\bar{y} - \frac{r s_y \bar{x}}{s_x}\right) + \frac{x r s_y}{s_x}$$

$$\hat{y} = a + bx$$

$$a = \bar{y} - b\bar{x} \qquad b = \frac{r s_y}{s_x}$$

Age	Salary
22	40000
24	46000
28	48000
30	60000
35	63000
40	80000
43	100000
48	200000
50	300000
56	400000

Corr r = 0.8806



$$\bar{x} = 37.6$$

$$\bar{y} = 133700$$

$$s_x = 11.047$$

$$s_y = 118841.1$$

$$r = 0.8806$$

$$y = 9473.8x - 222514$$

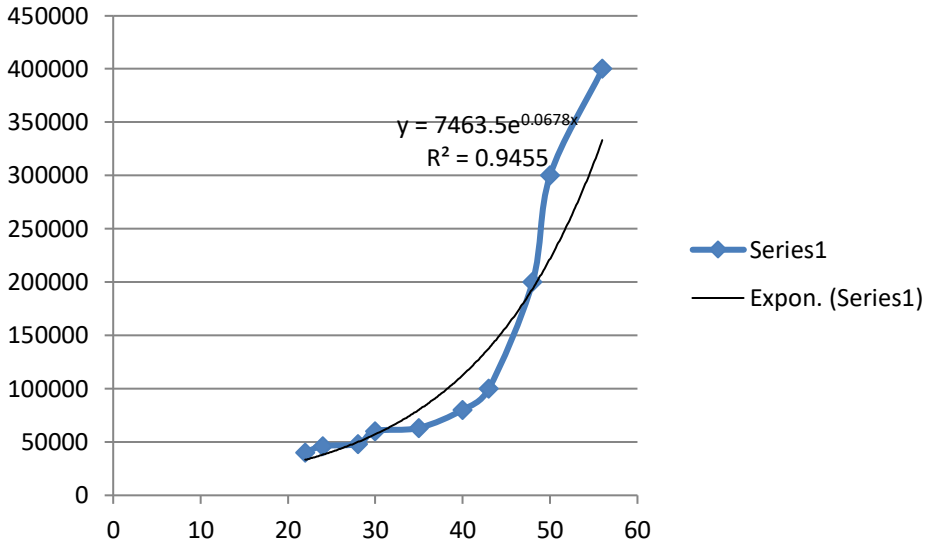
$$R^2 = 0.7756$$

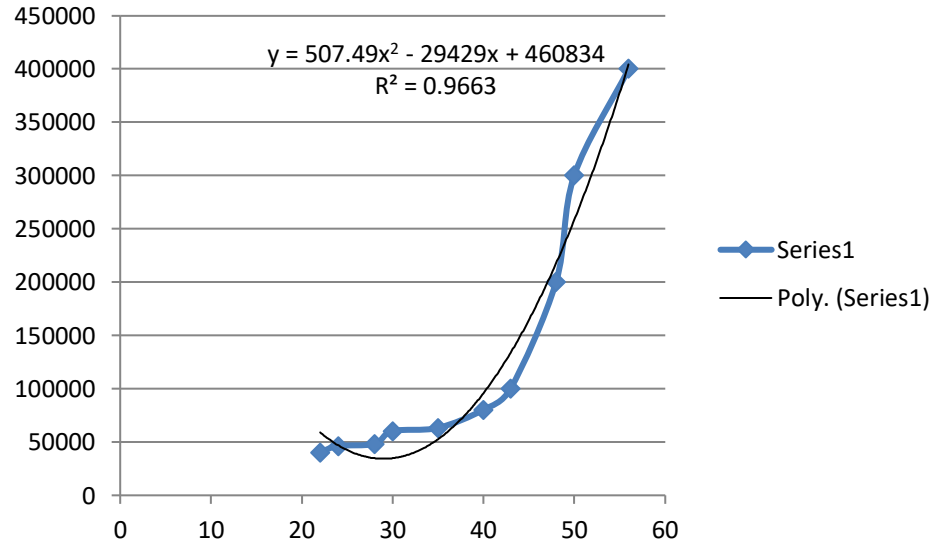
$$R = r = \sqrt{0.7756} = 0.8806$$

$$b = \frac{rs_y}{s_x} = \frac{0.8806 \times 118841}{11.047} = 9473.28$$

$a = \bar{y} - b\bar{x}$ = 133700 – 9473.28×37.6 = -222495

Age	Salary
22	40000
24	46000
28	48000
30	60000
35	63000
40	80000
43	100000
48	200000
50	300000
56	400000





Is it linear or non linear?

The analysis with r is only for linear

Correlation

Scatter plots and correlation reveal association and not causation

Better knowledge of variables can help us understand causation

46	95
27	93
24	90
30	96
42	98
50	98

Corr r = 0.7834

CAT Scores of 6 admitted students

Scores of an Indian cricketer

Lecture 11

Discussion on

Association between numerical variables

True or false

1. The x axis of the scatter plot has the explanatory variable.

2. The presence of a pattern indicates that the response variable as the explanatory variable increases

3. The net profit is about 10% of the sales. The scatter plot should be thought of as a line

4. If the correlation of a stock with the economy is 1, it is good to buy the stock when there is recession

5. The covariance between employees and production is computed with daily data. It is expected to increase if the data was aggregated to monthly

True, False, True, False, True

Question 1

Find the explanatory variable and the response variable

1. Marks and hours of study
2. Number of workers and units produced
3. Time to run and weight of the person
4. Total revenue and items sold
5. Exercise and body weight

Question 2

Correlation between number of customers and sales (in rupees) is 0.8. Does the correlation change if the sales is measured in thousands of rupees?

Question 3

Would correlation change if we add a constant to a variable? If we multiply by a constant?

Question 4

Cramer’s V measures association of categorical variables. Correlation does it for numerical variables. Can Cramers’s V be negative? Why or why not?

Question 5

Ten students took a test and after studying for a week took another test with the same portion. The marks are given below

60	66
45	50
72	78
77	77
56	60
64	70
66	70
58	62
42	47
50	55

1) Would you expect the scores to be associated?
2) What is the relationship between the marks?
3) The student with the highest score in the first has not got the highest in the second test. Is it an indication that he has not performed very well?