

Data and Inference

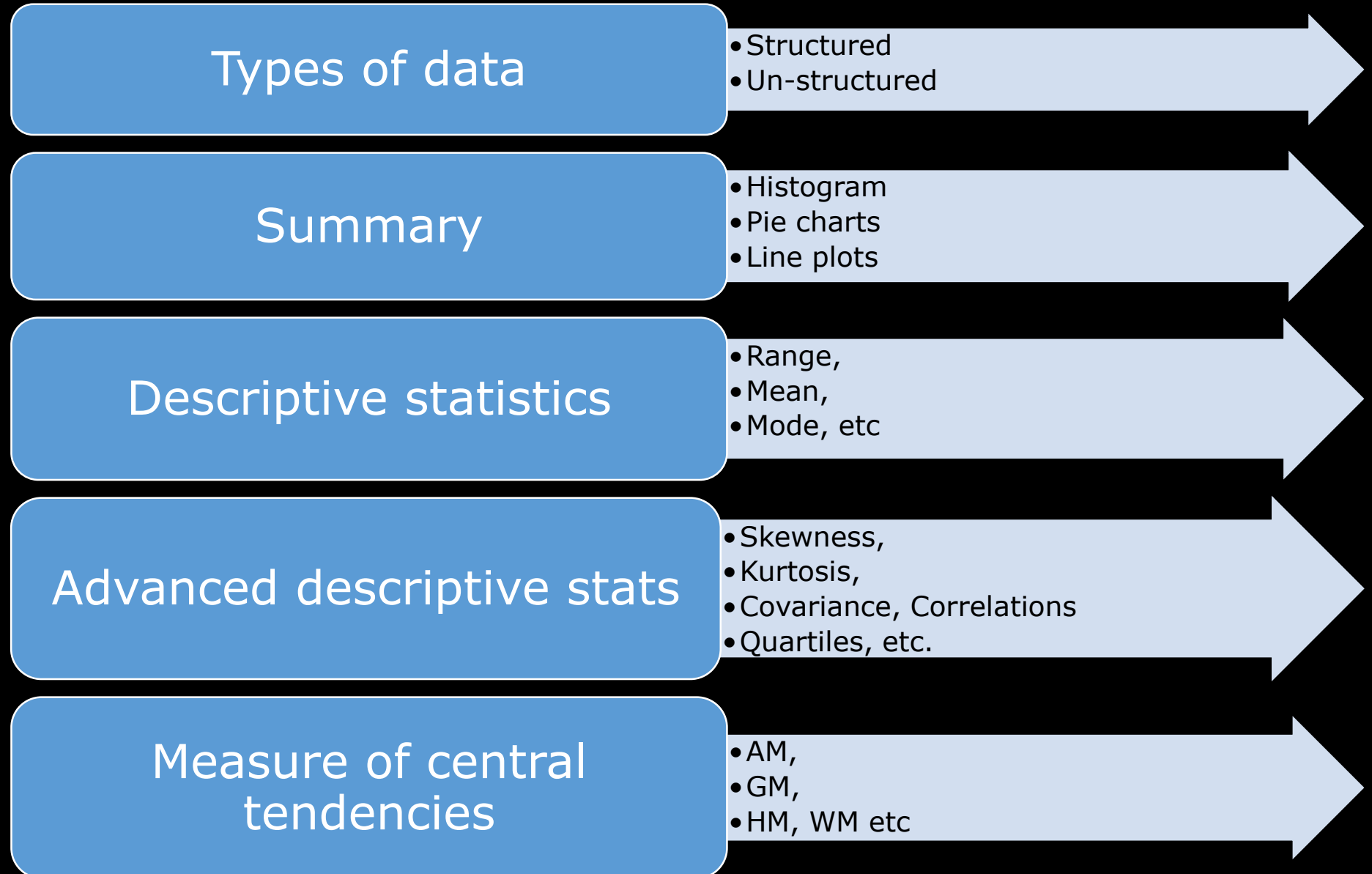
By

Kalpesh R. Patil

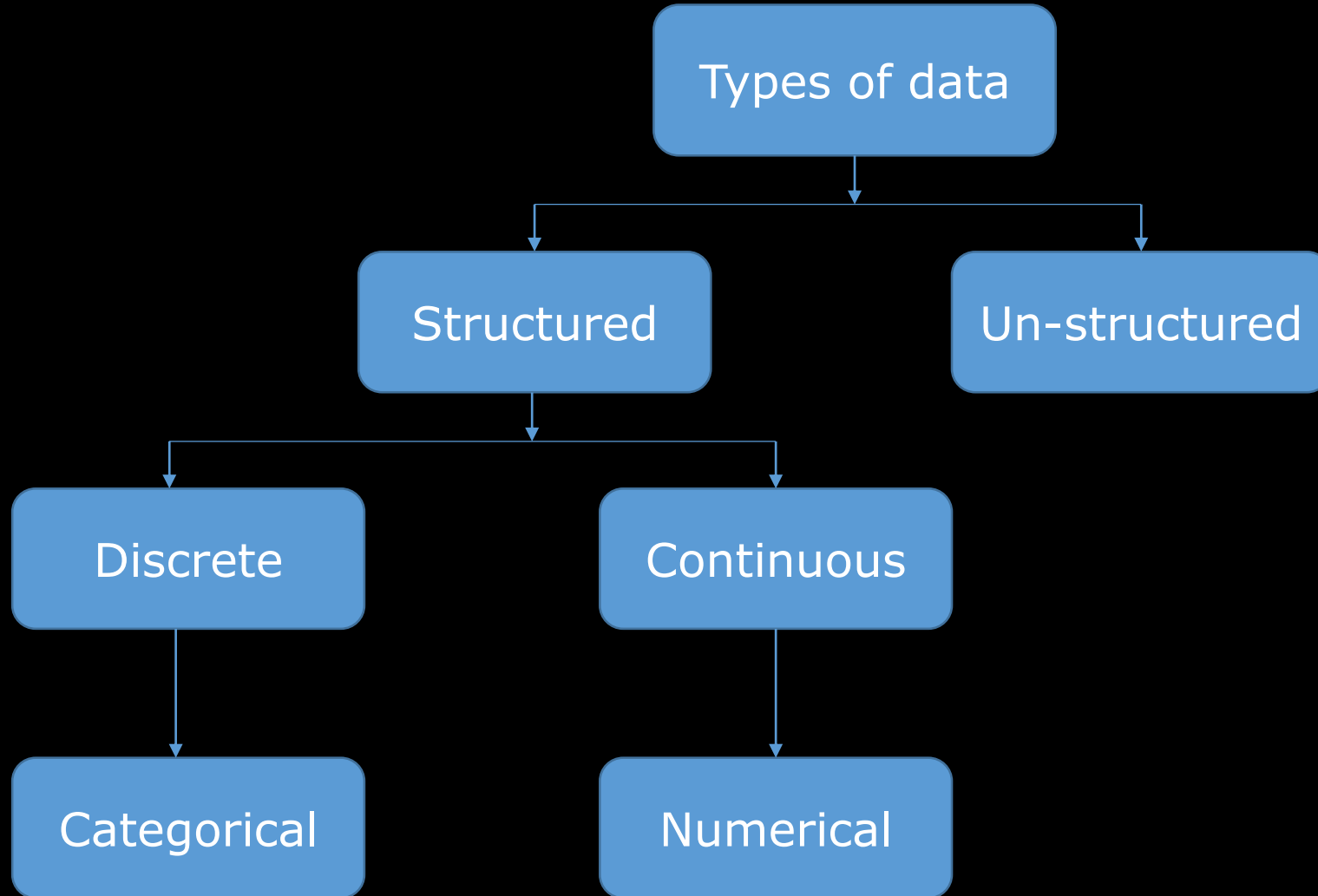
Research Scholar,

IIT Bombay

Contents overview



Data



Structured data

Data which have definite fields

Number of instances/entries of fixed Same fields

Information about data is readily available for processing

Classroom

Name	Age	Percentage	Special Interests
ABC	18	75%	Football
XYZ	21	85%	Script writing

Bank

Name	Account type	Loans availed	Max Trans Limit	Customer type
ABC	Savings	1	2 lacs	Gold
XYZ	Current	3	25 lacs	Premium

Telecom

Name	Age	Avg Bill Per month	Data limit	Type	Do Not disturb
ABC	19	149/-	3 GB/ Day	Regular	Disabled
XYZ	29	339/-	1 GB/ Day	Corporate	Enabled

Give me
Some
Examples ?

Structured data

Categorical

Division of students based on percentage

- Distinction, First, Second, Third, Fail

Movie/product/usage review

- 5, 4, 3, 2, 1 star

Gender

- Male, Female, Trans

Loans slabs in Banks

- 0-5 lacs, 6-15 lacs, 26 to 50 lacs etc.

Numerical

- Daily average temperatures
 - Monthly savings
 - Daily expenses
 - Daily rainfall
- Number of vehicle crossing a particular intersection
- Number of visitors in shopping mall/movie theatre
 - Stock indices

Un-structured data

Data which does not have definite fields

Number of instances/entries are not fixed

Information has to be processed

**Reviews
Online / Offline**

**Customer Care
Recordings**

Emails



Phone is simply superb in all aspects...low light performance of the camera is outstanding...you simply cannot go wrong with this phone



Horrible experience. Dirty bus inside and pathetic halt stop.



It's a disaster .I hope amir khan don't make this type of movie in future. This movie is torture for movie fans . It's a copy of pirates of carebian.

Social Media posts

- Photos
- Videos
- Comments

Give me
Some
Examples ?

What type of data is this?

Traffic survey

Name	Age	Gender	Health	Transportation mode	Frequency	Purpose of travel
ABC	20	M	PWD	Private/public	Daily	Coaching class, college, gymkhana
XYZ	15	F	Good	Public	Daily	School, Class,
PQR	55	M	Paralytic	Private	Weekly/Monthly	Hospital, Private meetings, social gatherings

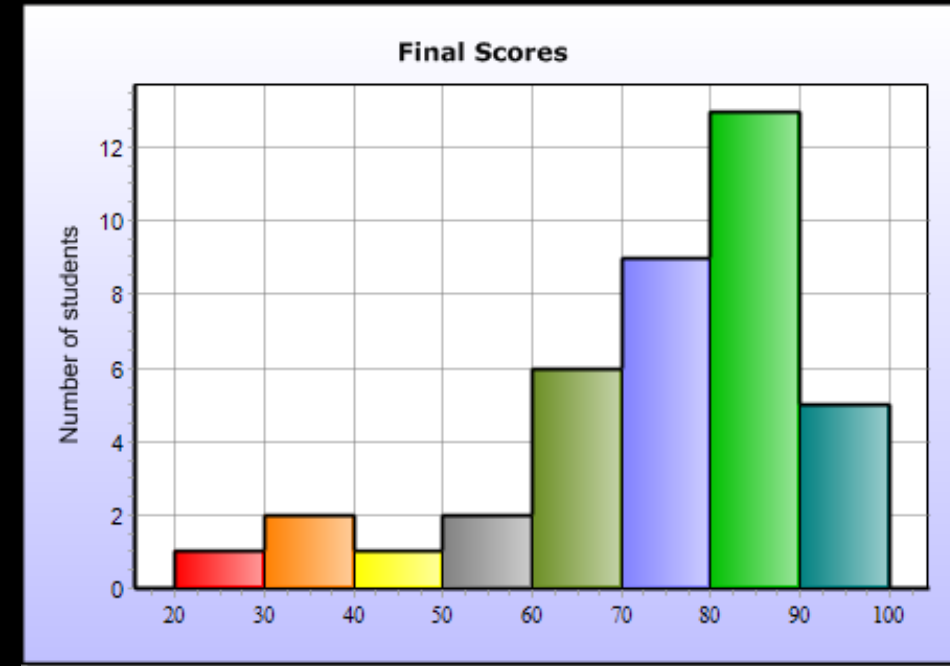
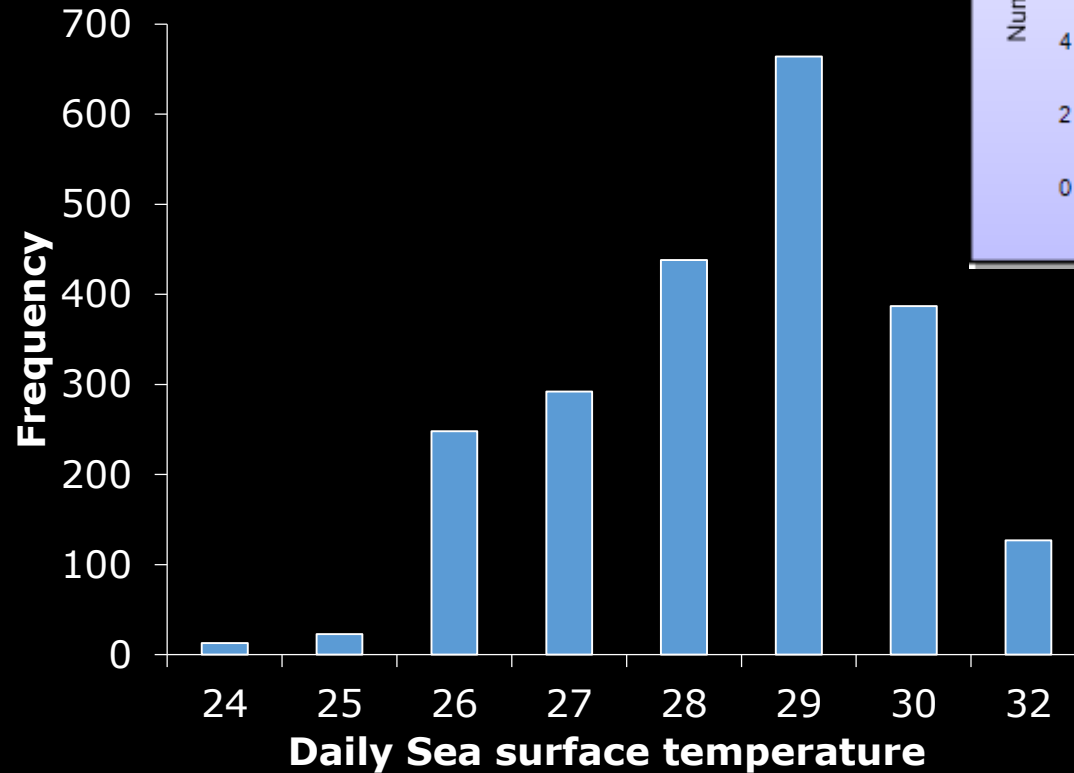
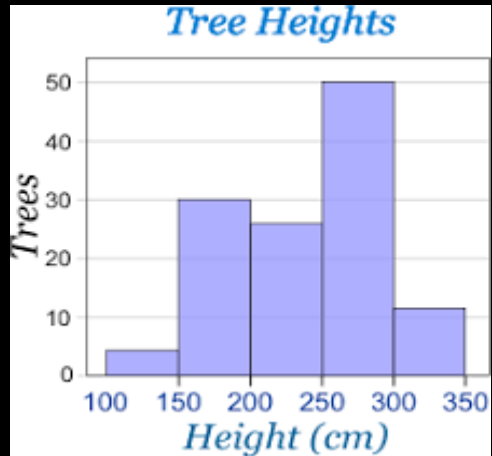
Examples
?

Summary of data

Histogram

Number of occurrence of
values from a group

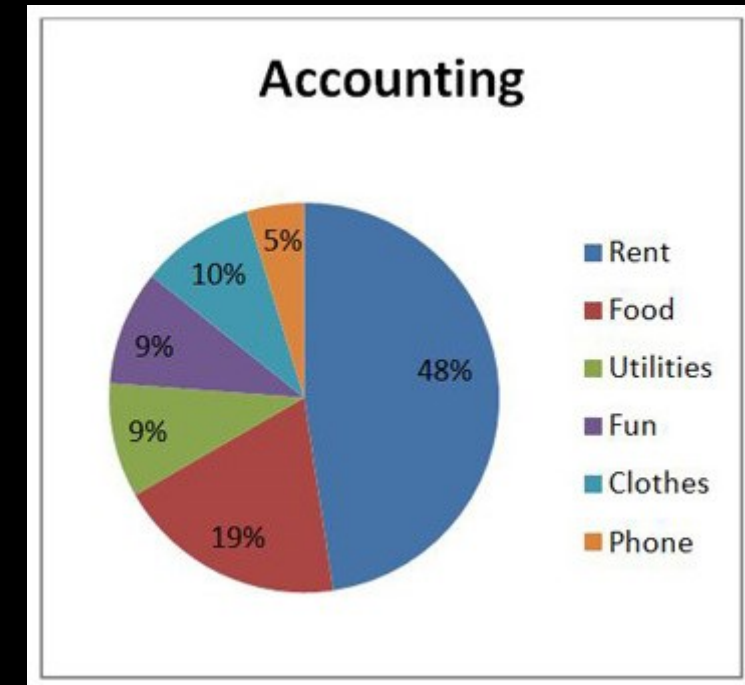
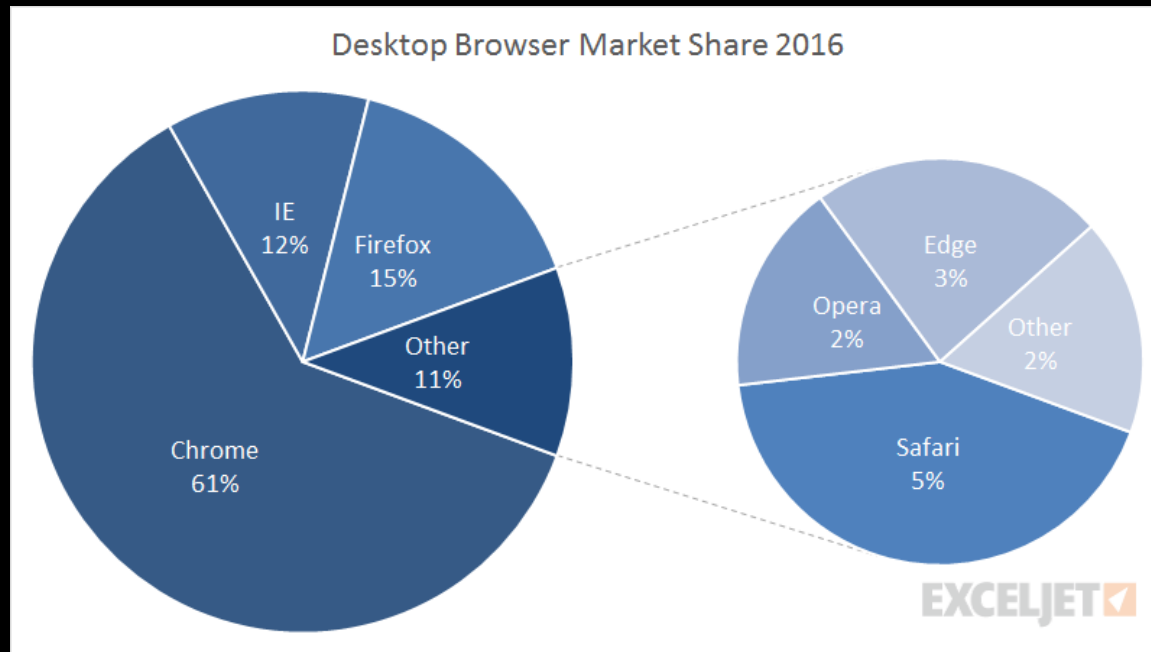
Group name / batch



Summary of data

Pie chart

- Breakdown of total dataset in some smaller parts
- Converting numerical data to categorical data

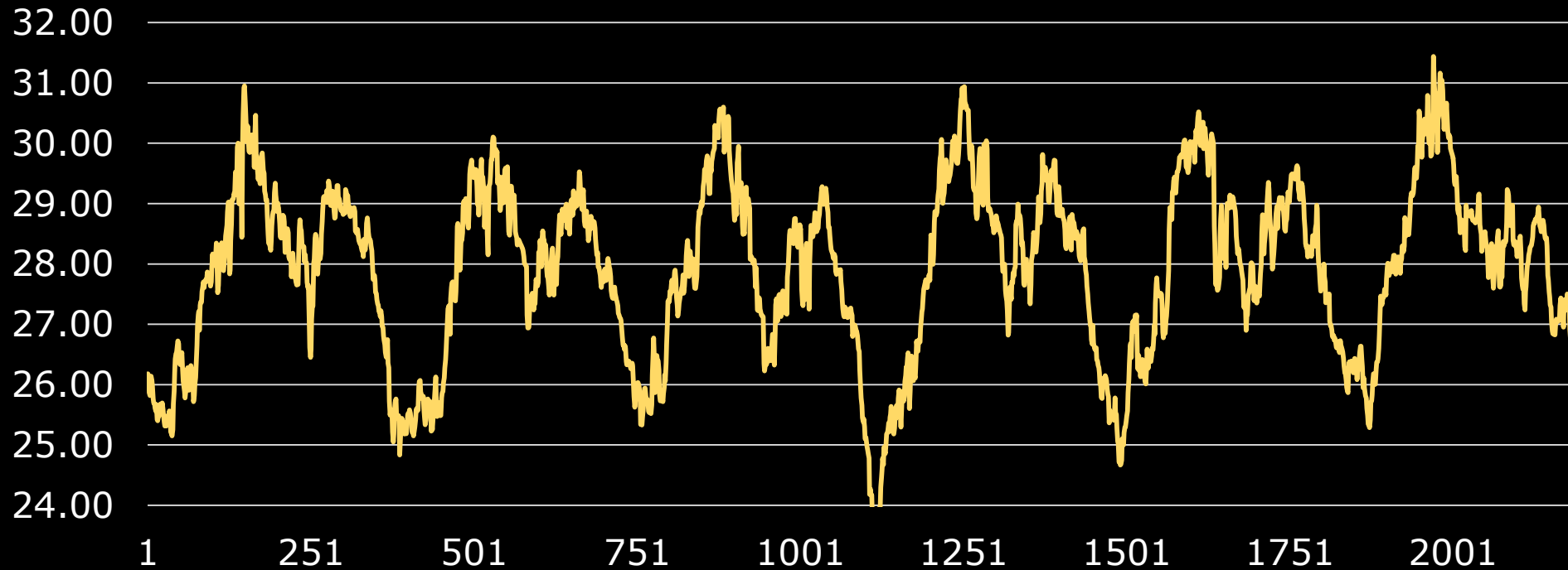


Summary of data

Line plots

- Variation of each individual values in particular range

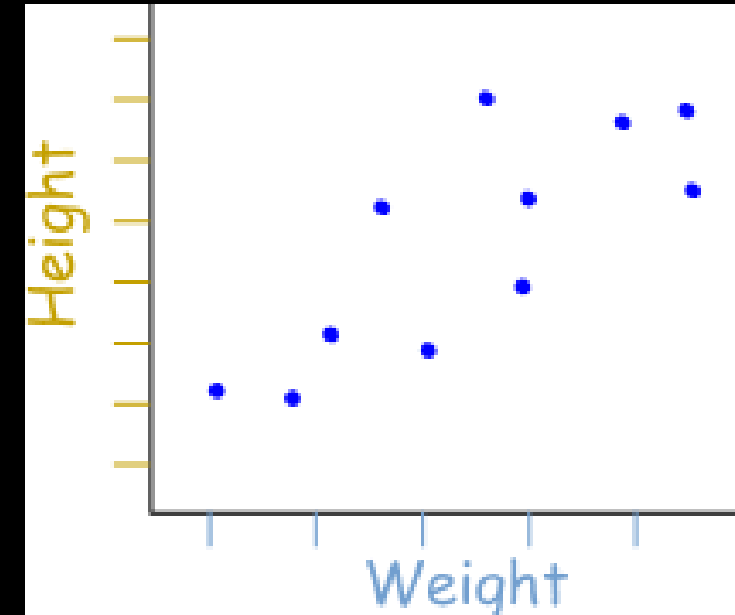
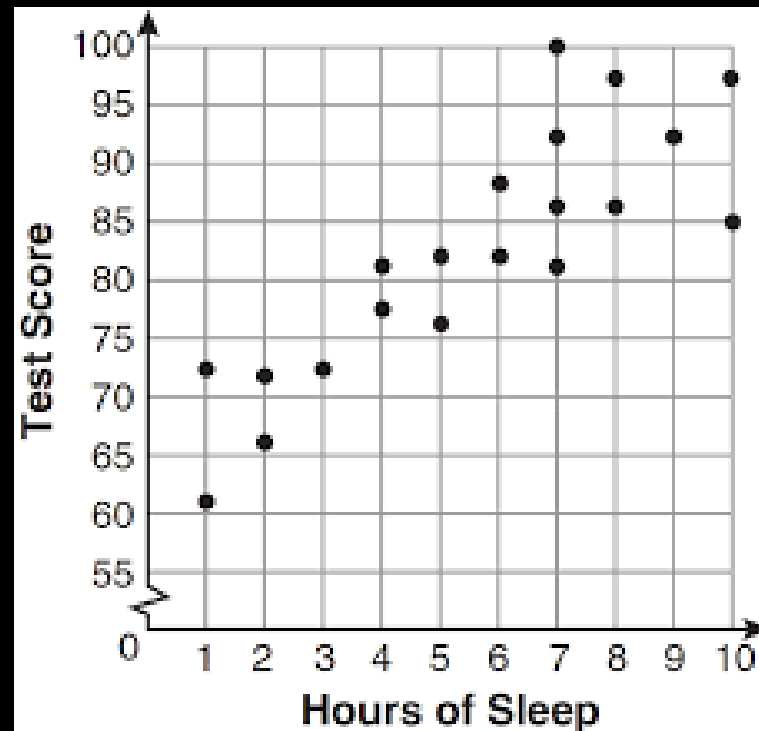
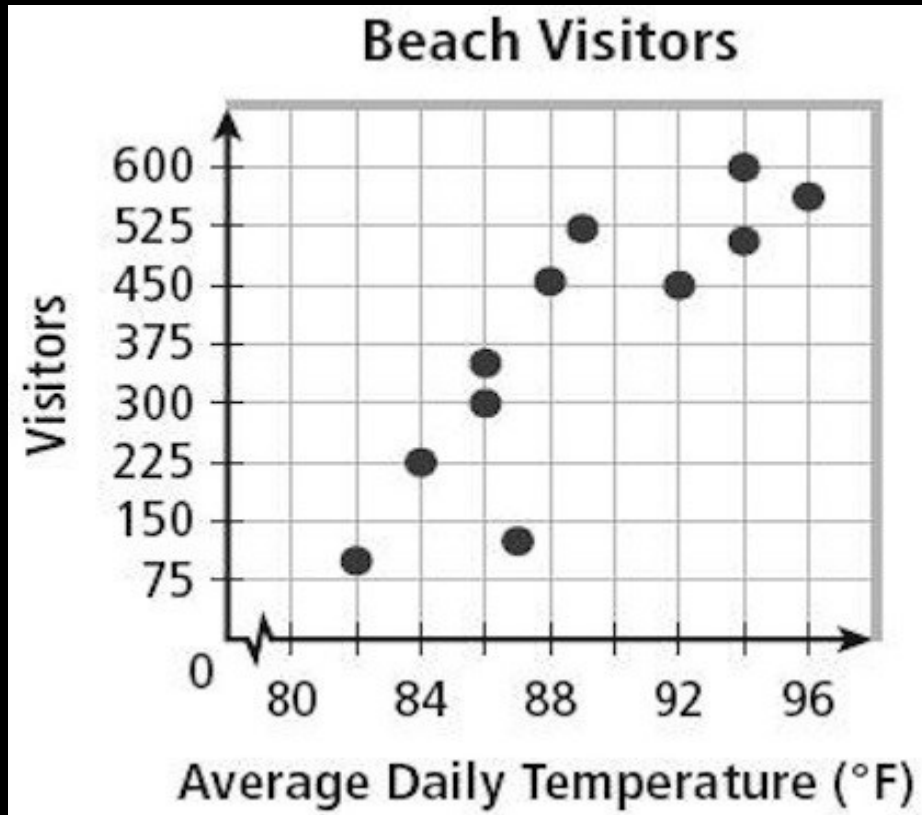
Daily Sea surface temp



Summary of data

Scatter plots


- Dependency of one data on another related data



Descriptive statistics

Range

- Expected extend of a variable
- Range = Max – min

 Range,
more the variability in data

Mean / average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mean = $\frac{\text{sum of all the scores}}{\text{number of scores}}$

For the scores 1, 2, 3, 4, 5, 5, 6, 9, 10

$$\begin{aligned}\text{Mean} &= \frac{1+2+3+4+5+5+6+9+10}{9} \\ &= 5\end{aligned}$$

Descriptive statistics

Median

- Center of the dataset
- Independent of values in dataset
- Dependent only on number of values
- This value tells half of observations are below and above this value.

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$
= **4.5**

Mode

- the most occurred value in dataset

4, 8, 1, 3, 4, 3, 3, 2, 4, 4

ORDER

1, 2, 3, 3, 3, 4, 4, 4, 4, 8

Mode = 4

Descriptive statistics

Mode

- the most occurred value in dataset

Class Interval	Exclusive Interval	Frequency (f)
10-19	9.5-19.5	10
20-29	19.5-29.5	12
30-39	29.5-39.5	18
40-49	39.5-49.5	30
50-59	49.5-59.5	16
60-69	59.5-69.5	6
70-79	69.5-79.5	8

By inspection, we observe that the modal class is 39.5-49.5 as it has the highest frequency. Then, $l_1 = 39.5$, $f_0 = 18$, $f_1 = 30$, $f_2 = 16$ and $c = 10$

\therefore

$$\text{Mode, } M_o = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

$$= 39.5 + \frac{30 - 18}{2 \times 30 - 18 - 16} \times 10 = 39.5 + \frac{12}{26} \times 10$$

\Rightarrow

$$\underline{M_o = 44.11}$$

Descriptive statistics

Variance

- 'Average' of 'departure of individual values' from 'mean' in the dataset

$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

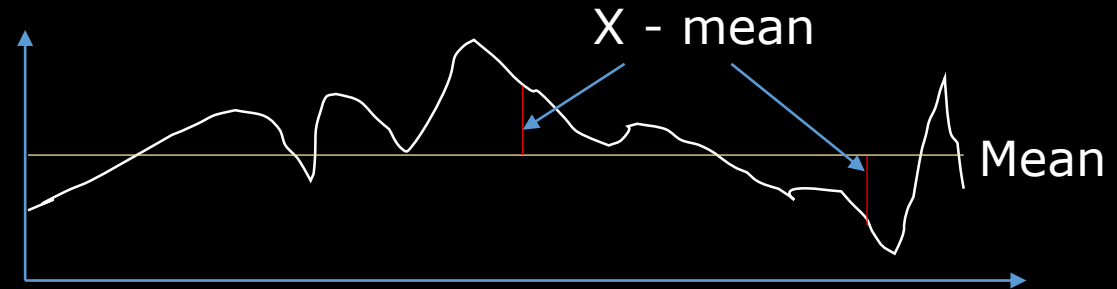
$\sigma^2 = \text{variance}$

$X_i = \text{the value of the } i\text{th element}$

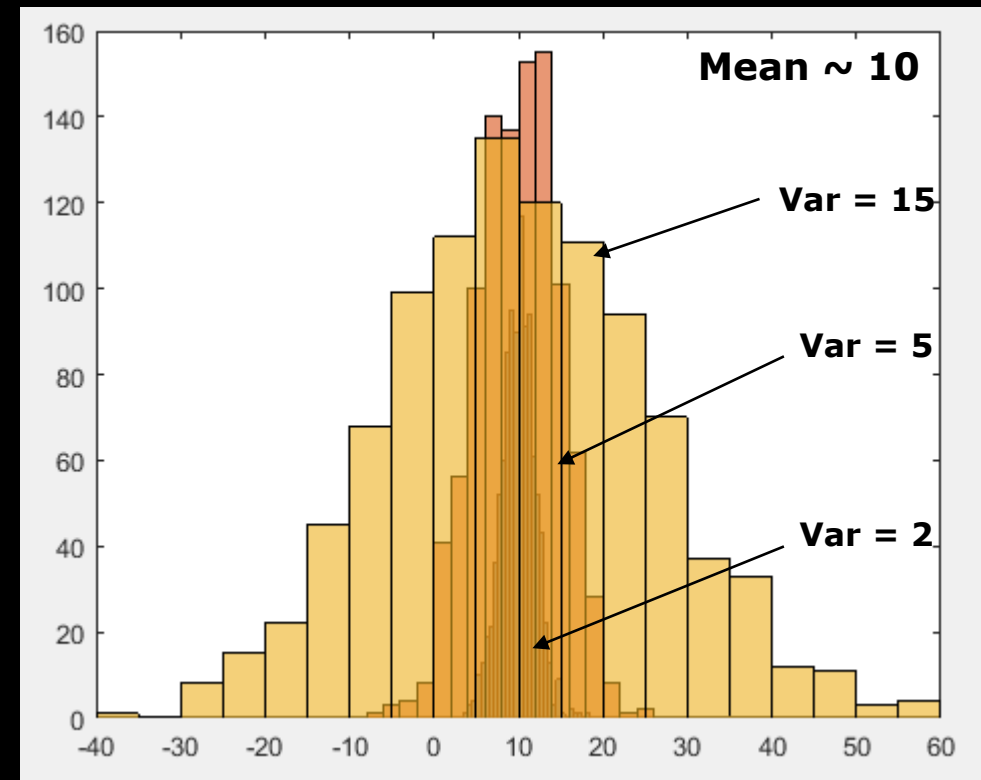
$\bar{X} = \text{the mean of } X$

$N = \text{the number of elements}$

- Explains the distributions of values in fixed range
- Higher variance means values in dataset are more equally distributed in each group



Please note a extend of a base with respect to CV



Descriptive statistics

Variance - example

Contribution for a get-together

Case - A Case - B

1	150	850
2	250	750
3	350	820
4	1500	950
5	2500	850
6	3500	1150
7	50	1000
8	0	550
9	60	550
10	6	896
Mean	836.6	836.6
Total	8366	8366
Var	1396464.04	31372
std dev	1181.7208	177.1
CV	1.41252785	0.212



Variance or std. deviation → Data distribution is 'unequal',
higher uncertainty in process



Variance or std. deviation → Data distribution is 'equal',
Lesser uncertainty in process

**In this example mean was same, if not use
'CV'**

But when to un-related datasets are to be compared
We can use coefficient of variation (CV)

- $CV = \text{std. Dev} / \text{Mean}$ (Dimensionless quantity)

Which condition
is better ?
Any example ?

Descriptive statistics

Variance - example

Consider two customers (A and B)
Process – whether to call or not for providing any offer (loan/mobile plan) information

How will you decide whom to call in this two?

Provided: You have two records.

- Information about picking unknown numbers.
- Response towards call for promotion offers.

Which condition is better ?
Any example ?

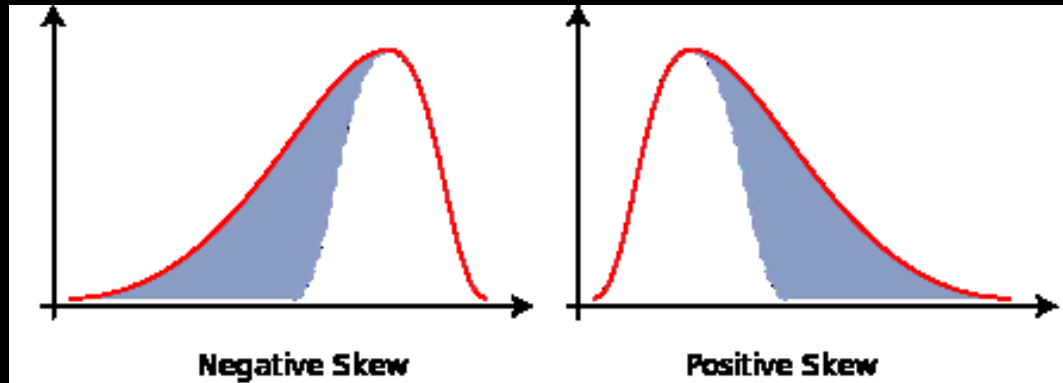
Map the call picking to
0 – Not answered, 1 – answered.

If CV for 'A' is higher he/she is uncertain to respond to calls hence call 'B'.

Advanced Descriptive statistics

Skewness

Strong Indication about higher value occurrence

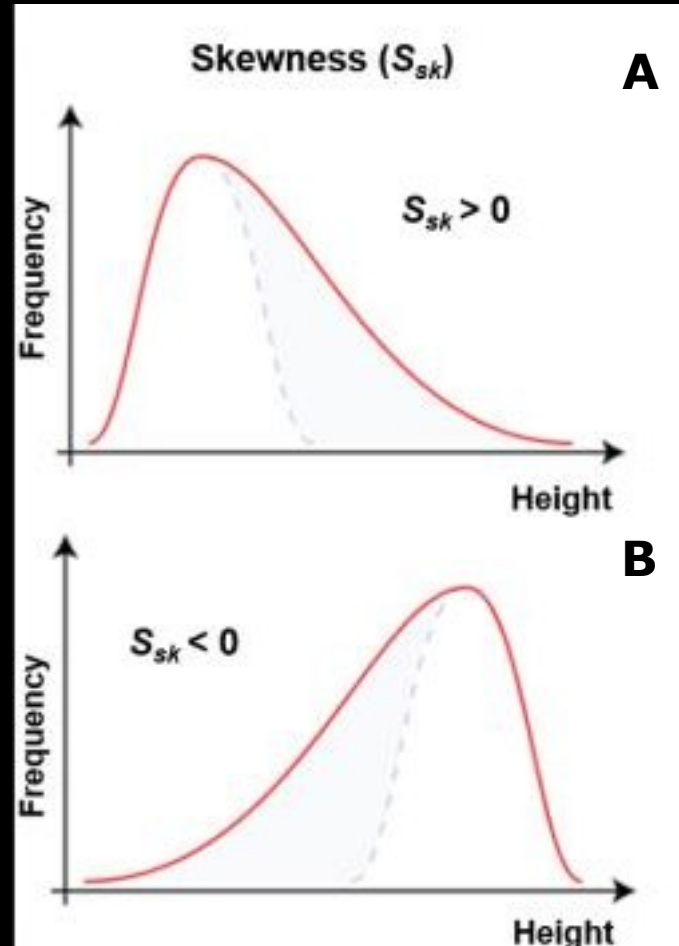


Higher values are occurring more

Higher values are occurring less

$$= \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3 / N}{s^3}$$

Example



Number of participants grouped in two to be sent for Olympics for high jump.

Which group is better?
Why?

Replace height by
rainfall at tourist place?

Does this change the
scenario?

Advanced Descriptive statistics

Kurtosis

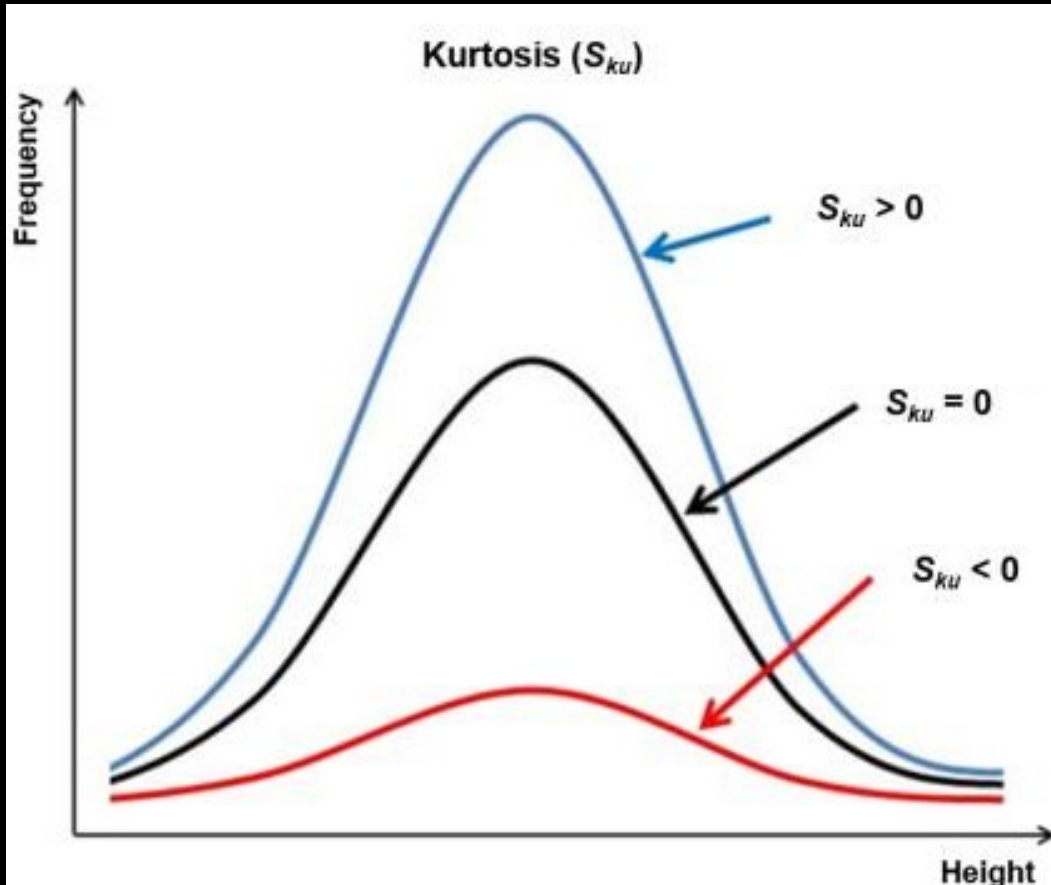
- +ve – Leptokurtic
- 0 – Mesokurtic
- ve – platokurtic

General measure for data behavior compared to normal distribution

Value close to '0' is preferable.

Kurtosis is difficult to relate in practical terms.

It has to be looked along with other statistics.



Population Kurtosis Formula

$$K = n \frac{\sum_{i=1}^n (X_i - X_{avg})^4}{(\sum_{i=1}^n (X_i - X_{avg})^2)^2}$$

Sample Kurtosis Formula

$$K = \frac{n(n+1)(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^n (X_i - X_{avg})^4}{(\sum_{i=1}^n (X_i - X_{avg})^2)^2}$$

Advanced Descriptive statistics

A B C

0	0	1
0	1	1
1	1	1
0	1	1
1	0	1
0	0	0
0	0	1
1	1	1
0	1	1
0	1	0
0	0	1
0	1	1
1	0	0
0	1	1
1	0	1
0	1	0
0	0	1
1	1	1
0	0	1
0	1	1

You are calling a friend for a movie

Process:

Uncertain – Not coming for movie (0)

Certain – Coming for movie (1)

Now try to note all statistics, with their practical significance.



CV



certainty

-ve skewness

Kurtosis near zero

B



CV



certainty

+ve skewness

Kurtosis away from zero

A

After looking all stats which friend you will call for movie?

NB: This inference will vary for each dataset

Mean	0.3	0.55	0.8
Total	6	11	16
Var	0.21	0.248	0.16
std dev	0.458	0.497	0.4
CV	1.528	0.905	0.5
skewness	0.873	-0.201	-1.5
kurtosis	-1.242	-2.183	0.699

Advanced Descriptive statistics

Daily rainfall (in mm of water) statistics for different tourist places in monsoon. Which one is preferred?

Or involves lower risk?

#observation were equal for every dataset.

Justify the answer.

Location	A	B	C	D	E
std dev	13.14	21.31	19.43	21.69	13.90
Mean	5.16	9.50	8.92	9.31	5.10
CV	2.54	2.24	2.18	2.33	2.73
skewness	5.29	4.38	3.92	4.77	7.36
kurtosis	45.79	29.37	22.73	37.17	125.74

CV – suggests 'B' & 'C'

Skewness – suggests 'A' & 'E'

Kurtosis – suggests 'B' & 'C'

Mean – suggests 'A' & 'E'

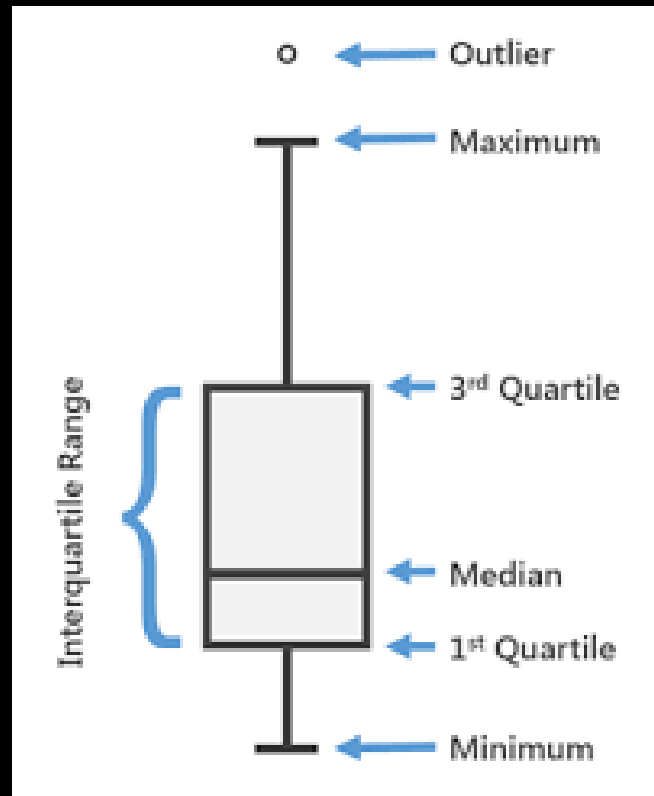
Advanced Descriptive statistics

Quartiles

Quartiles give information about the range, variance, min, max and important population stats in a single plot

Also helps in finding the outliers.

Outlier are extremes in data.



1st quartile – 25% observations are lesser than this value.

2nd quartile is the values which divides the data in two halves.

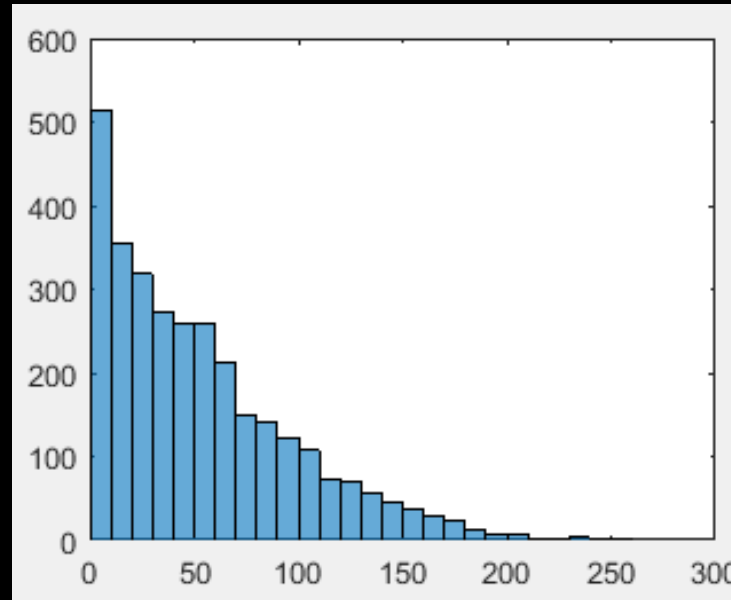
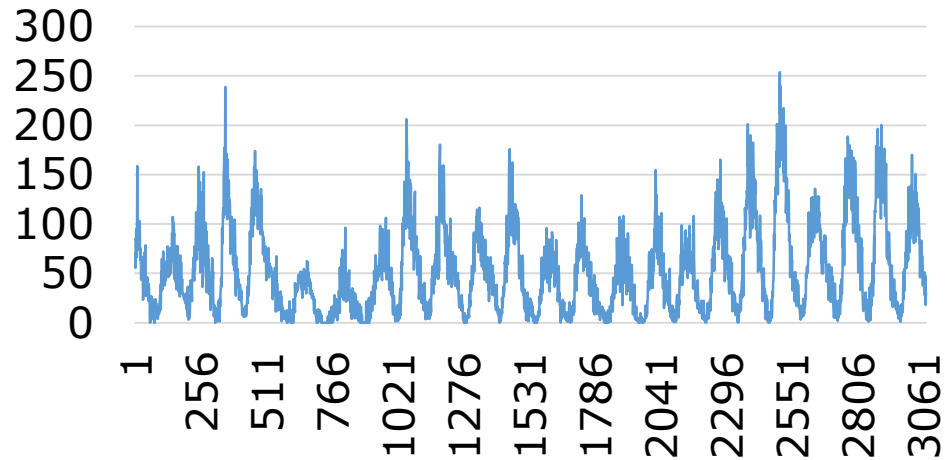
Sound similar

Its Median – 2nd quartile.

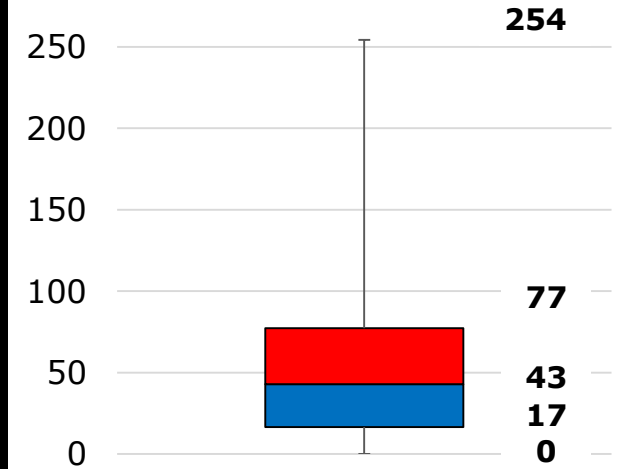
3rd quartile – 75% observations are lesser than this value.
25% observations are greater than this value.

Advanced Descriptive statistics

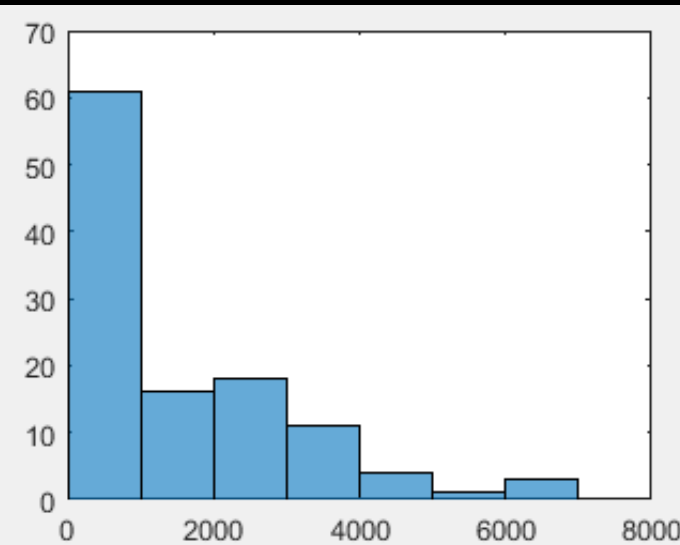
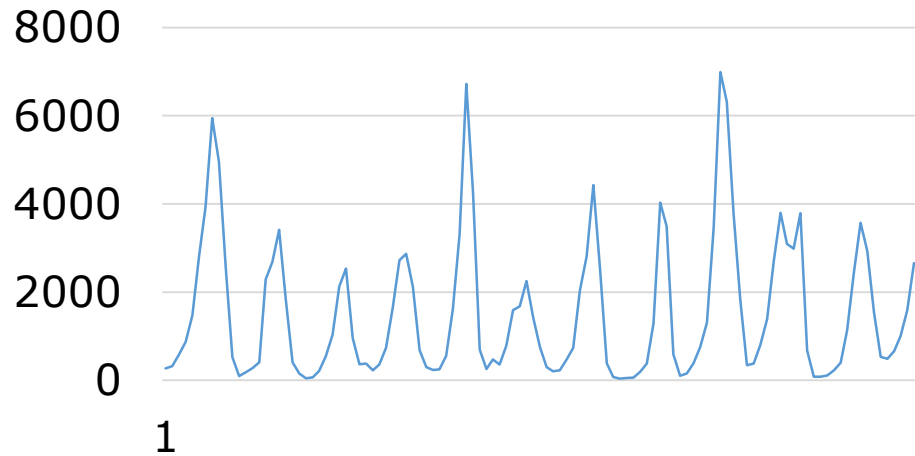
Daily Sunspot series



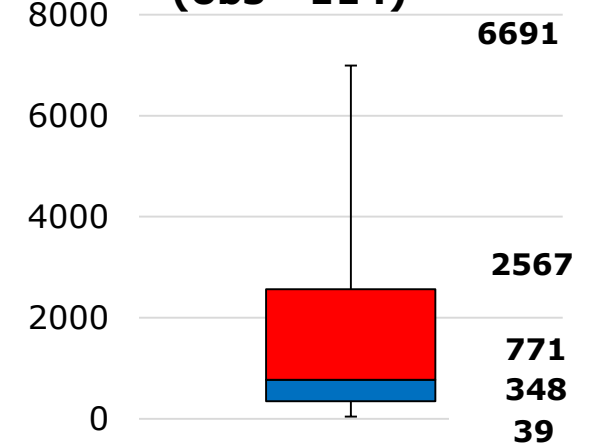
Daily Sunspot
(obs - 3074)



Canadian Lynx

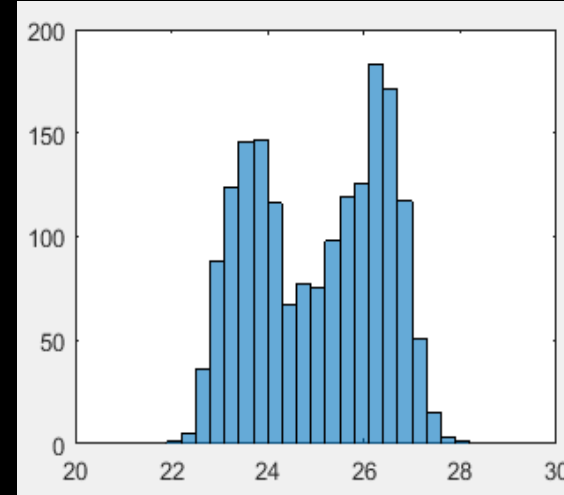
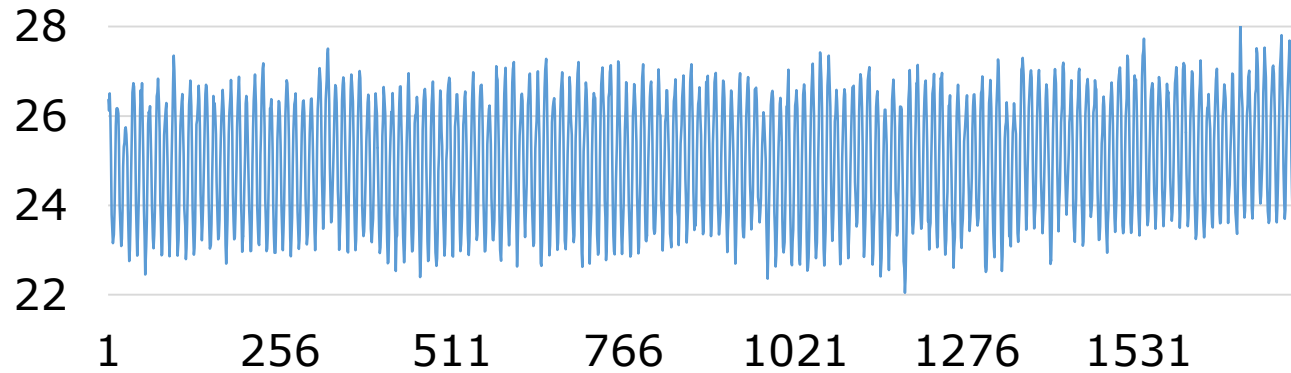


Canadian Lynx
(obs - 114)

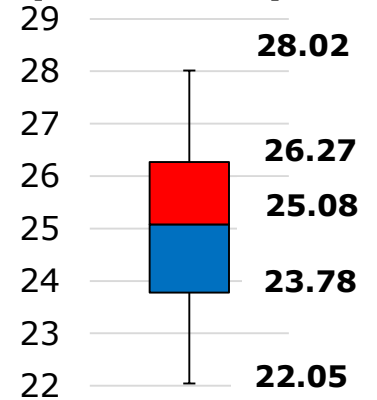


Advanced Descriptive statistics

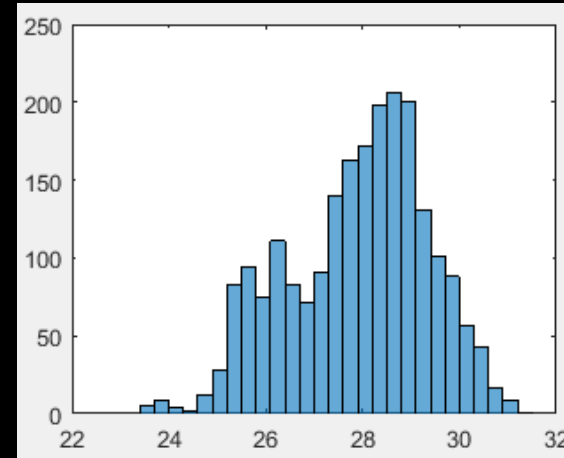
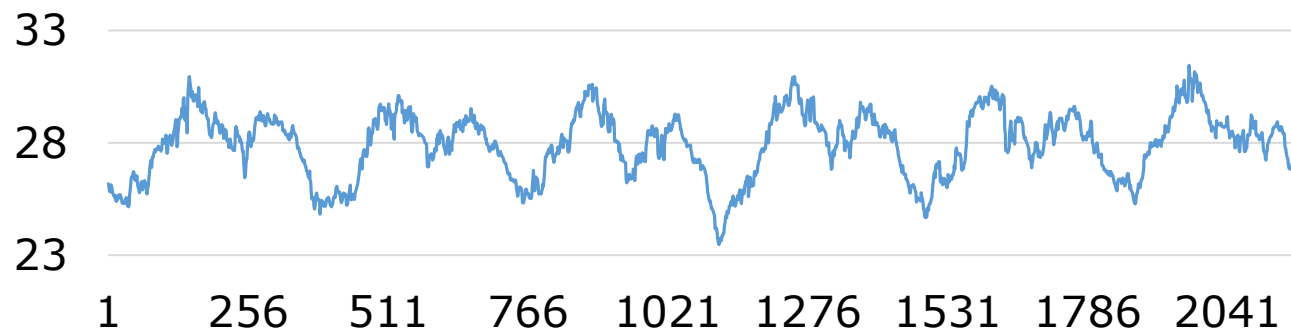
Monthly Sea surface temp



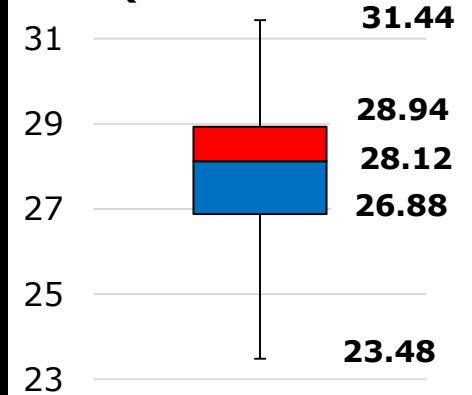
Monthly SST
(obs - 1766)



Daily Sea surface temp

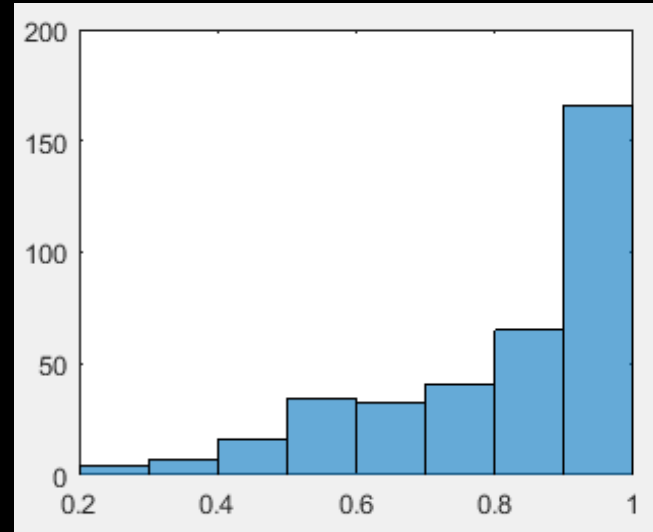
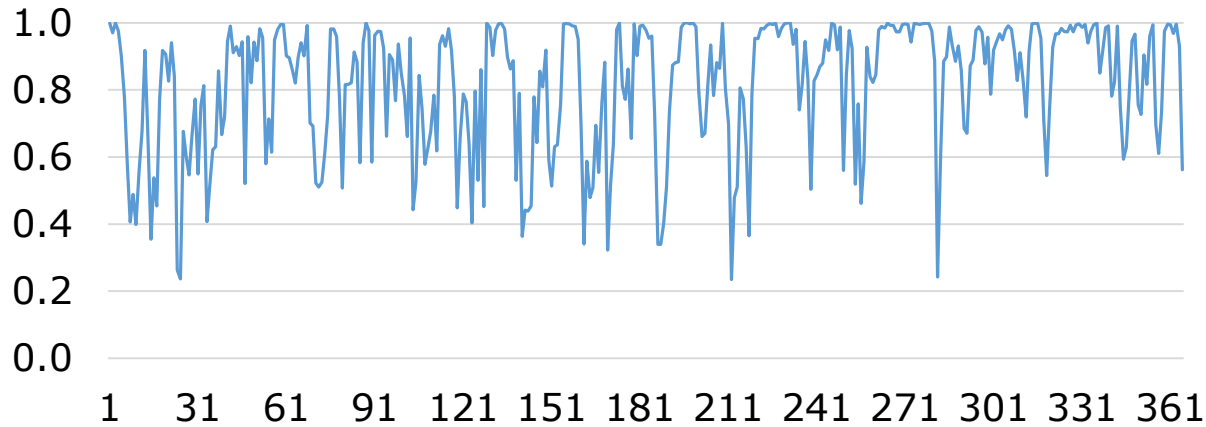


Daily SST
(obs - 2192)

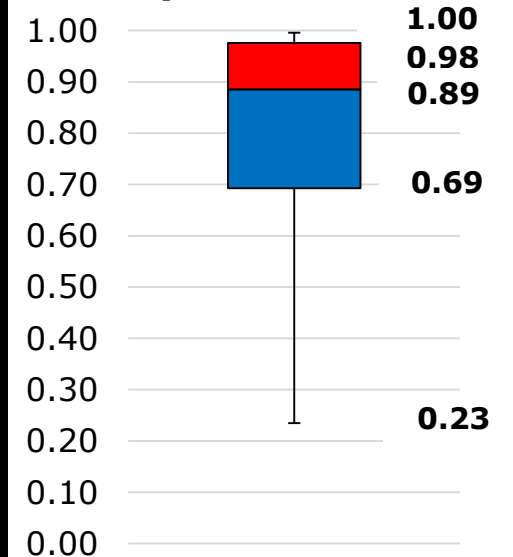


Advanced Descriptive statistics

Total Cloud Cover



Total cloud cover
(obs - 365)



	Daily Sunspot series	Canadian Lynx	Monthly Sea surface temp	Daily Sea surface temp	Total Cloud Cover
Max	253.80	6991.00	28.02	31.44	1.00
Min	0.00	39.00	22.05	23.48	0.23
Range	253.80	6952.00	5.97	7.95	0.77
average	52.82	1538.02	25.04	27.91	0.82
variance	1968.58	2492840.39	1.83	2.08	0.04
Median	43.00	771.00	25.22	28.12	0.89
mode	0.00	409.00	25.06	29.14	1.00
skewness	1.05	1.37	-0.11	-0.36	-1.02
kurtosis	0.73	1.58	-1.32	-0.45	0.11
CV	0.84	1.03	0.05	0.05	0.23

Co-variance

Co-variance

$$\text{cov}(X, Y) = \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

- Average of
- Products of
- Departures of two variables with respect to their mean
 - Range: - inf to + inf
- It explains the dependency of variance of two variables on each other

Condition

Both variables should have same number of observations

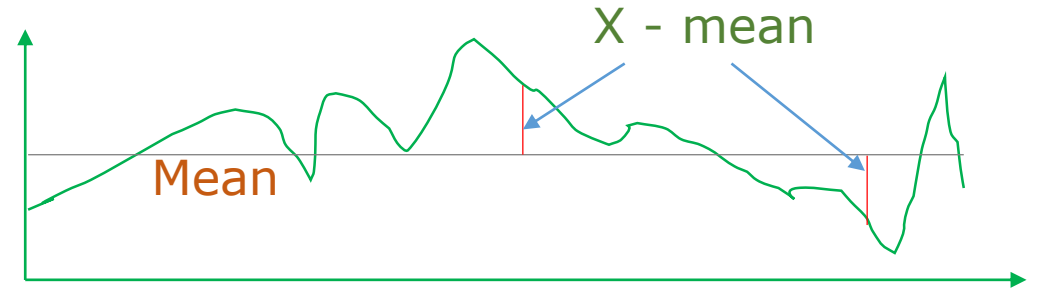
Caution

A mathematical property hence will always yield some results even for distantly related datasets. Therefore use for related datasets.

E.g. Cov(Height of student, marks of student) – meaning less
Cov(attendance of student, marks obtained) – meaning full

Co-variance example

- Q1: x = 2, y = 10
- Q2: x = 3, y = 14
- Q3: x = 2.7, y = 12
- Q4: x = 3.2, y = 15
- Q5: x = 4.1, y = 20



The average x value equals 3, and the average y value equals 14.2. To calculate the covariance, the sum of the products of the x_i values minus the average x value, multiplied by the y_i values minus the average y values would be divided by (n-1), as follows:

$$\text{Cov}(x, y) = ((2 - 3) \times (10 - 14.2) + (3 - 3) \times (14 - 14.2) + \dots + (4.1 - 3) \times (20 - 14.2)) / 4 = (4.2 + 0 + 0.66 + 0.16 + 6.38) / 4 = 2.85$$

Inference

- Higher positive value confirms the better agreement between two variables, i.e. One increase other also increases (rate of increase may be different)
- Higher negative values confirms the agreement in reverse way.
 - Near zero values indicates very poor agreement.
 - $\text{Cov}(X, X) = \text{variance of 'X'}$
 - $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

Co-variance

example

	Daily Sunspot series	Canadian Lynx	Monthly Sea surface temp	Daily Sea surface temp	Total Cloud Cover
Daily Sunspot series	871.96	-1968.55	-1.23	-12.15	-1.62
Canadian Lynx	-1968.55	2492840.39	-241.01	271.93	-21.12
Monthly Sea surface temp	-1.23	-241.01	1.66	0.00	0.01
Daily Sea surface temp	-12.15	271.93	0.00	0.93	0.05
Total Cloud Cover	-1.62	-21.12	0.01	0.05	0.04

- Give me some inferences
- Give examples of meaningful and meaningless covariance

Correlation

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Why to divide by Std.dev ?

- Numerator in 'Cov' is variance term – square term
- To neutralize that we need to divide by another square term.
- Hence division by std dev of product of two variables – its also a square term
- This will cancel the units and gives a unit less ratio which will have fixed range unlike 'Cov'
 - Range: -1 to +1

Covariance vs correlation

'Cov' had unbounded range, hence it was difficult to judge dependency between two variables of different units.

'Corr' has fixed range for all real variables.

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

	Daily Sunspot series	Canadian Lynx	Monthly Sea surface temp	Daily Sea surface temp	Total Cloud Cover
Daily Sunspot series	1.00	-0.04	-0.03	-0.43	-0.29
Canadian Lynx	-0.04	1.00	-0.12	0.18	-0.07
Monthly Sea surface temp	-0.03	-0.12	1.00	0.00	0.04
Daily Sea surface temp	-0.43	0.18	0.00	1.00	0.25
Total Cloud Cover	-0.29	-0.07	0.04	0.25	1.00

Correlation

Pearson correlation

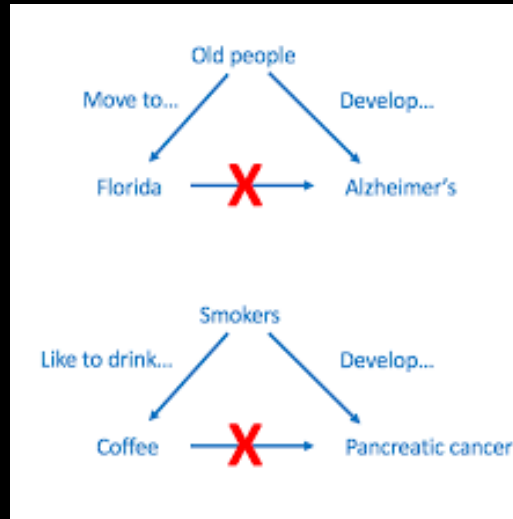
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

Based on actual values in dataset

Causation

- One event leads to other
 - E.g. Population -> Unemployment
- In correlation two event may or may depend on each other



Spearman correlation (rank correlation)

- Similar formula but calculated based on the rank of individual values in data
- Generally used for categorical datasets

English	Math	Rank (E)	Rank(M)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

Different types of means

Arithmetic mean

$$\bar{x}_{\text{arithm}} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- When data is closely behaving as normally distributed data,
- Sum is more important
- Academic sectors - Marks

Geometric mean

$$\bar{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Simple difference, instead of addition we need take a product of all values in dataset

- Product is more important
- Finance sectors – Average outcome of recurring investment

Harmonic mean

$$\text{Harmonic Mean} = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right)}$$

Inverse of averages of reciprocals

- When extreme values are more frequent
- -vely skewed data
- Rarely used, because cant be used when data a single 0 value in it

Weighted mean

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i},$$

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}.$$

- Higher values are having more weights and vice-a-versa, seems logical.
- Arithmetic mean - is a special case of the weighted mean where all data have equal weights
- Weights can assigned with various methods.
- Value/sum of all values – simplest method

- When extreme values are more + sample size is limited.
- To know who is pulling the mean

Different types of means

Various datasets

	Daily Sunspot series	Canadian Lynx	Monthly Sea surface temp	Daily Sea surface temp	Total Cloud Cover
AM	52.82	1538.02	25.04	27.91	0.82
GM	-	801.06	24.96	27.60	0.79
HM	-	343.88	24.97	27.84	0.75

Get together contribution example

	Case A	Case B
AM	836.6	836.6
GM	-	816.4
HM	-	794.9
WM	2505.81353	874.1

← Clearly indicates presence of higher values.

Questions?

Doubt clearing session

Thank you.