

Descriptive Statistics

~15K students appear in the entrance exam for PUCIT each year.

- └ F.Sc per subject
- └ Matriculation per subject
- └ Entrance per area

Lots of data!

Humans are not good with lots of numbers.

They might be good with a few numbers (~10) but not many.

Summarize the data

Humans are very good at this

Visual

Bar Chart
Line graph
Pictogram
Pie Chart
Histograms

Numerical

Location

Mean
Median
Quartiles
Percentiles
Box-Whisker Plot

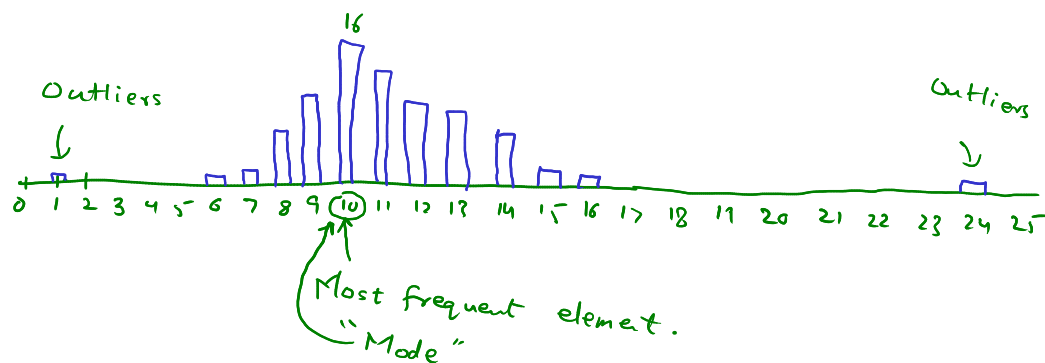
Spread

Range
Inter-quartile range
Std. Deviation

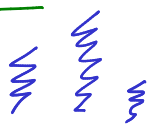
Frequency Table

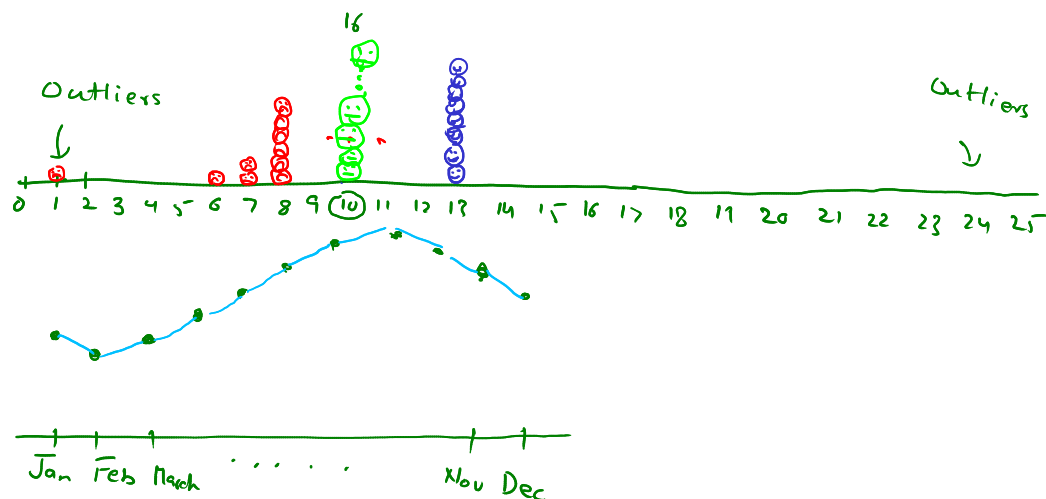
Score	1	6	7	8	9	10	11	12	13	14	15	16	24
Frequency	1	1	2	6	10	16	13	9	8	5	2	1	1

Bar Chart



Pictogram

Electricity Consumption 
Money 



Line Graph

- Time-series data
- Y-axis values are too large.



Pie Chart

Total

$$75 - 360$$

$$\frac{1}{75} \times 360$$

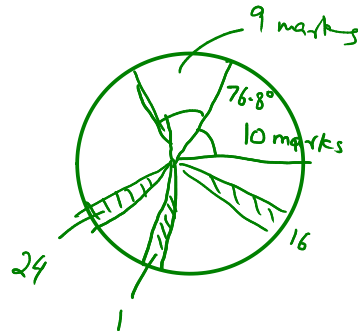
Score	1	6	7	8	9	10	11	12	13	14	15	16	24
Frequency	1	1	2	6	10	16	13	9	8	5	2	1	1

75

Angle

$$\frac{1}{75} \times 360 \quad \frac{1}{75} \times 360 \quad \frac{2}{75} \times 360 \quad \frac{6}{75} \times 360 \quad \dots \dots \dots$$

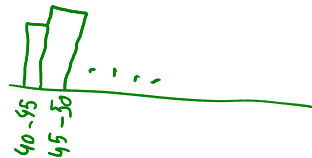
$$\frac{10}{75} \times 360$$



Histogram

For continuous variables.

40K ⁺ - 45K	35
45K ⁺ - 50K	40
⋮	⋮



Numerical Summary

Location

① Mean $x_1 \quad x_2 \quad \dots \quad x_n$

$$\frac{\sum_{i=1}^n x_i}{n} \quad \leftarrow \text{average}$$

② Median: That number from which half the numbers are larger and half are smaller.

Mean
= 7.6

→ 7 1 9 16 5

① Sort in ascending order

1 5 7 9 16

② Median location

When n is odd

$$\begin{aligned} \text{Median location} &= \frac{(n+1)}{2} \\ &= \frac{(5+1)}{2} \\ &= 3 \end{aligned}$$

When n is even

$$\begin{aligned} \text{Median location} &= \text{avg. of } V_{\frac{n}{2}} \text{ and } V_{\frac{n}{2}+1} \\ &= \text{avg. of } V_{\frac{5}{2}} \text{ and } V_{\frac{5}{2}+1} \\ &= \text{avg. of } V_3 \text{ and } V_4 \\ &= \frac{5+7}{2} = 6 \end{aligned}$$

7	1	9	16	5	5
1	5	5	7	9	16
V_1	V_2	V_3	V_4	V_5	V_6

③ Median = number at median location in sorted list
= 7

③ Percentiles

The p -th percentile of a data set is a number X_p such that $p\%$ of the data is smaller than or equal to X_p .

1 5 7 9 16

 ↑

 Median

 } 50th percentile

Step 1 Sort the data in ascending order.
Let $d_1, d_2, \dots, d_N$ represent the sorted data.

Step 2 Percentile location

$$L = \frac{P}{100} * N$$

Step 3 Compute %tile

L is whole number
↓
avg. of d_L and d_{L+1}

L is not a whole number

$$\begin{array}{l} d_{L^+} \leftarrow L^+ = \text{round-up}(L) \\ \downarrow \\ d_{\lfloor L \rfloor} \end{array}$$

80th %ile of Morning section $\Rightarrow$ score for which only 20% ^{Morning} students got higher marks

_____ Afternoon section $\Rightarrow$ _____ Afternoon _____

④ Quartile

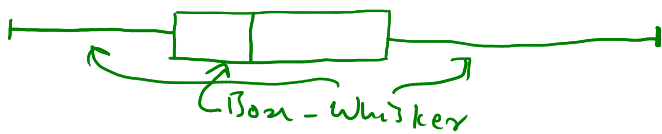
$Q_1 = 1^{st} \text{ Quartile} = 25^{th} \% \text{ tile}$

$Q_2 = \text{2nd Quartile} = 50\% \text{ tile} = \text{Median}$

$Q_3 = 3rd \text{ —————} = 75^{th} \% \text{ tile}$

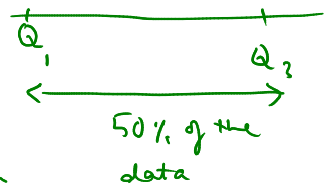
⑤ Box-Whisker Plot

min	Q_1	Median Q_2	Q_3	max
-----	-------	-----------------	-------	-----



Spread

① Inter-quartile range
 $= Q_3 - Q_1$



② Range = max - min

③ Std. Deviation

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$
 always ≥ 0 b.c. of squaring.