07|01|21 dec-05

→ Variance, std. devn; Z-score

→ Graphic displays

→ Similarity & dissimilarity measure

1. for nominal attributes
2. for binary attributes

Variance : $\boxed{\sigma^2 = \frac{\sum(x-\mu)^2}{n}}$

Standard deviation $\sigma = \sqrt{\text{Variance}}$

→ The smaller the σ value, the closer are the values to the mean.

Eg: find mean, variance and standard deviation for the following set of numbers.

Eg 1. $\{1, 2, 3, 4, 5, 6, 7\}$

$\mu = 04$

$\sigma^2 = 04$

$\sigma = 02$

2 $\{1, 2, 3, 4, 5, 6\}$

$\mu = 3.5$

$\sigma^2 = 2.92$

$\sigma = \sqrt{2.92}$

$= 1.71$

Significance of std. devn :

1. This is a way of measuring spread. It measures how far typical value are from mean. If σ is low, it means values tend to close to mean.
2. σ can be 0, if all the values are same.
3. σ unit will be same as the unit of the data.

Assignment :

Consider the performance of 3 players A, B and C. The mean for each of them is 10. Find out which player is more reliable to your team ?

player A

scores	7	9	10	11	13
f	1	2	4	2	1

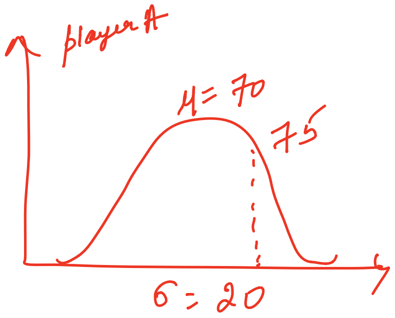
player B

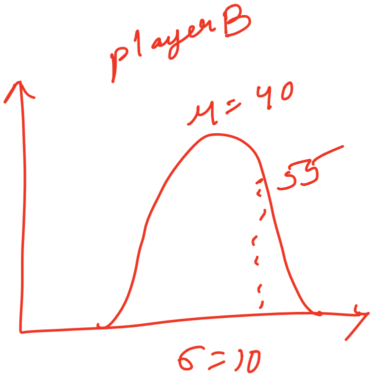
7	8	9	10	11	12	13
1	1	2	2	2	1	1

player C

3	6	7	10	11	13	30
2	1	2	3	1	1	1

Standard score / Z-score :





Standard score gives a measure for comparing values across different sets of data where the mean and std. devn. differ.

→ mean

$$\boxed{Z = \frac{x-\mu}{\sigma}}$$

→ std. devn.

$$Z_A : \frac{75-70}{20} = \frac{5}{20} = 0.25$$

$$Z_B = \frac{55-40}{10} = \frac{15}{10} = 1.5$$

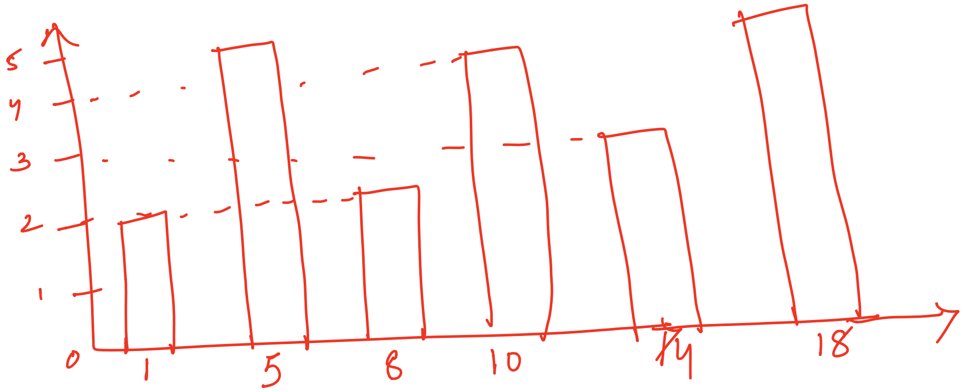
Graphic displays : → 1. Histogram (iii) q-q-plot

2. Quantile plot (iv) scatter plot

(v) scatter plot matrix

1) Histogram : (Height of bar indicate the frequency or count of X value}

{1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 14, 14, 14, 18, 18, 18, 18, 18}

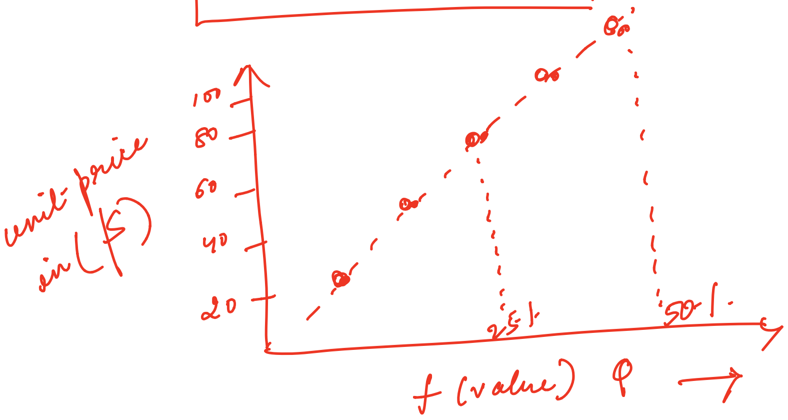


2) Quantile plot :

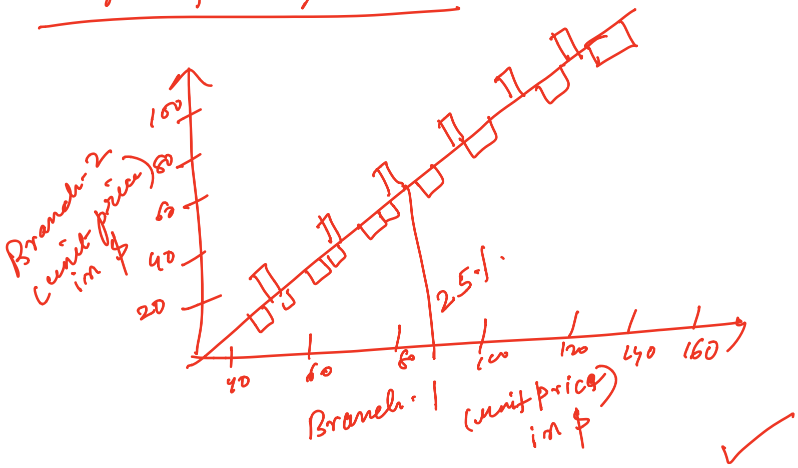
(for univariate data distribution)

→ Each observation x_i is paired with a percentage f_i which indicate that approximately $f_i \times 100\%$ of data are below x_i

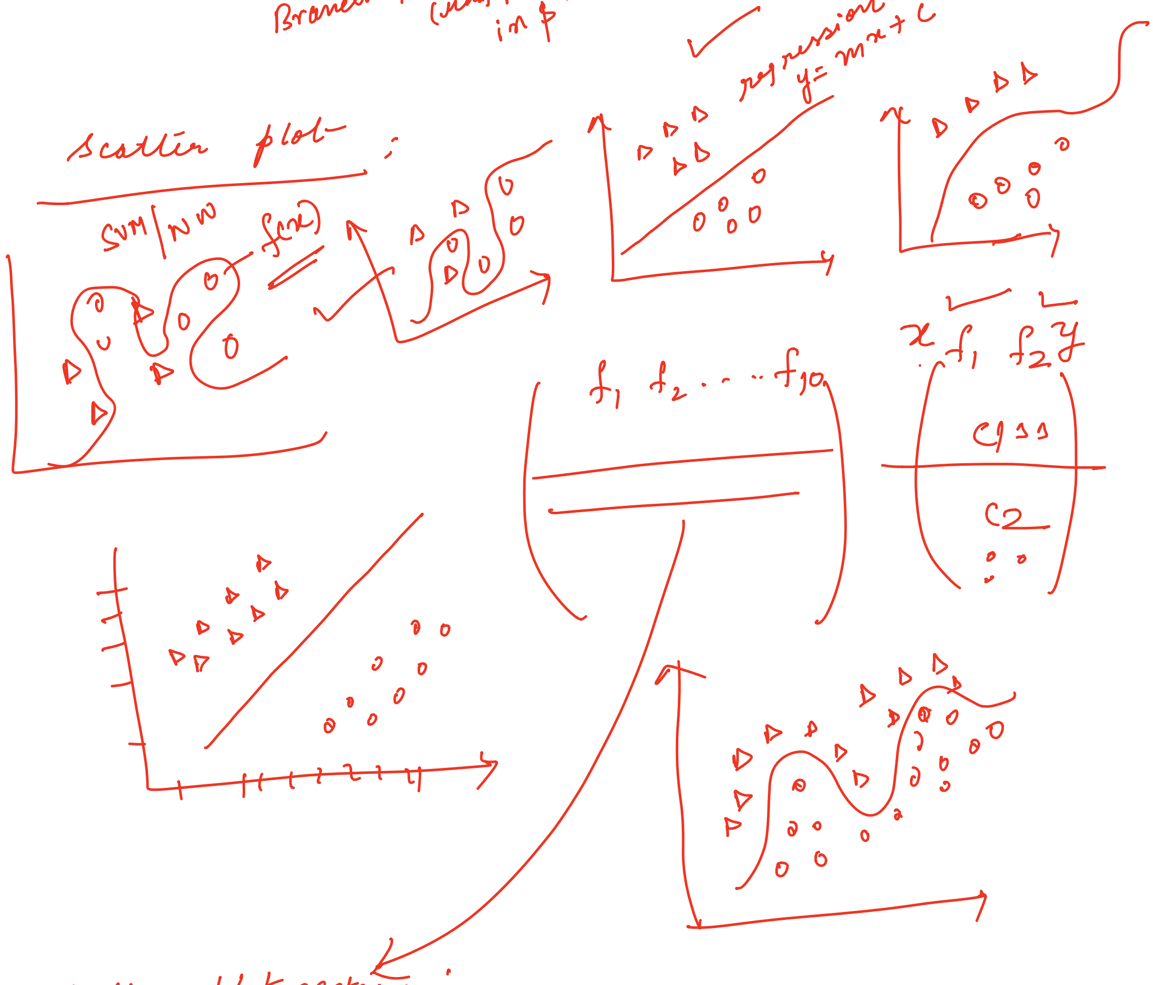
$$\boxed{f_i = \frac{i - 0.5}{N}}$$



3) Q- Q- plot :

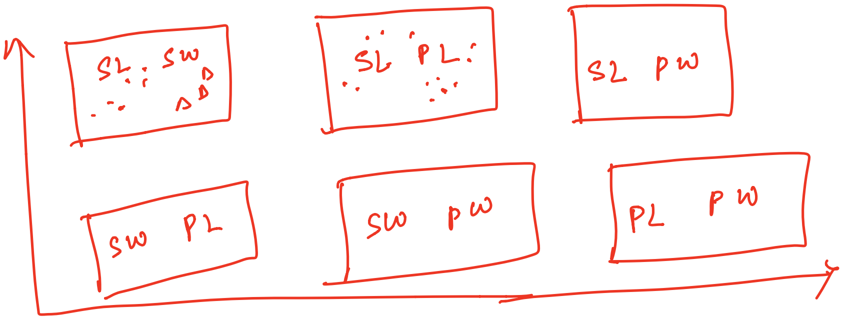


4) Scatter plot :



5) Scatter plot matrix :

$$\begin{pmatrix} SL \quad SW \quad PL \quad PW \\ C1 \\ \hline C2 \end{pmatrix}$$


SL SW
SL PL
SL PW
SW PL
SW PW
PL PW