October-28:

Measures of Dispersion:

1. Variance
2. Standard deviation
3. Range.

1. Variance:

Always variance should be less if we are getting higher variance it will reduce/affect our prediction accuracy.

population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

Sample variance

$$\delta^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

Degree of freedom.

Ex: $x=\{1,2,2,3,4,5\}$

$$\mu = \frac{17}{6} = 2.8$$

$$\sigma^2 = \frac{(1-2.8)^2+(2-2.8)^2+(2-2.8)^2+(3-2.8)^2+(4-2.8)^2+(5-2.8)^2}{6}$$

$$= \frac{3.24+0.64+0.64+0.04+1.44+4.84}{6}$$

$$= \frac{10.84}{6}$$

$$\boxed{\sigma^2 = 1.8}$$

$$\delta^2 = \frac{10.84}{5}$$

$$= 2.16.$$

* If variance is more than data spread is also more.

2. **Standard Deviation:**

population S.D(σ)

$SD = \sqrt{\text{Variance}}$

$\sigma = \sqrt{\sigma^2}$

Sample SD(s)

$SD = \sqrt{\text{Variance}}$

$s = \sqrt{s^2}$

Normalisation of the values because $\sigma^2 = 3.8$

$\sqrt{\sigma^2} = \sqrt{3.8}$

$\sigma = 1.94$

3. **Range:** The difference between the minimum and maximum.

Range = max - min.

Ex: {1,2,2,3,4,5}

range = 5-1
= 4.

* A small range values are close to each other (less than variation).

* A large range values are widely spread out (more variable)

* Outlier can make the range misleading
(outlier = out of the boundaries)

Percentiles and Quartiles:

Percentile is a value below with a certain percentage of observation lie/came.

Ex:
{2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

Total = 20 = 100%

40% of ~~to~~ records value is <7.

formula:

$$\text{percentile Rank of } x = \frac{\#\text{ of values below } x}{n} \times 100.$$

→ In next page

Removing outliers from the data:

[1, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29]

What is the % ranking of 10?

$= \frac{16}{\cancel{20}} \times \overset{5}{\cancel{100}}$

$= 80\%$ of data value is below 10.

$= \frac{8}{20} \times 100 = 8 \times 5 = 40\%$ of records value is <7

What value exists at percentile ranking of 90%

$$\text{Value} = \frac{\text{percentile}}{100} \times (n+1)$$

$$= \frac{90}{100} \times 21$$

$$= 18.9$$

$= 19$ → Index

~~value = 19~~ $\boxed{\text{value} = 19}$

Five number Summary:

1. Minimum
2. First Quartile (Q1) - 25%
3. Second | Median Quartile (Q2) - 50%
4. Third Quartile (75%) - Q3.
5. Maximum

<u>Note</u>: Choose these 5 numbers after removing outlier from the data.

Boundaries = [lower fence - Upper fence]

lower fence = Q1 - 1.5 (IQR)

Upper fence = Q3 + 1.5

IQR = Q3 - Q1

$Q_1 = \frac{25}{100} \times 20$

$Q_1 = 5$ 5 → Index

$Q_1 = 3$

$Q3 = \frac{\overset{13}{\cancel{75}}}{\underset{9}{\cancel{100}}} \times \cancel{20}$

$Q_3 = \frac{15}{100} \times 20 = 15$

Q3 = 7.

IQR = 7 - 3 = 4

lower fence = Q1 - ~~Q~~ 1.56 ξ IQR

= 3 - 1.5(4)

= -3

Upper fence = Q3 + 1.56 ξ IQR

= 7 + 1.5(4)

= 7 + 6

= 13

Anything ~~6~~ -3 & 13 is considered as outliers.

[1, 2, 2, 2, 3, 3, 4; 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29]

1. min → 1

2. Q1 → 3

3. Q2 → 5

4. Q3 → 7

5. ~~Q4~~ → 9.
Max.