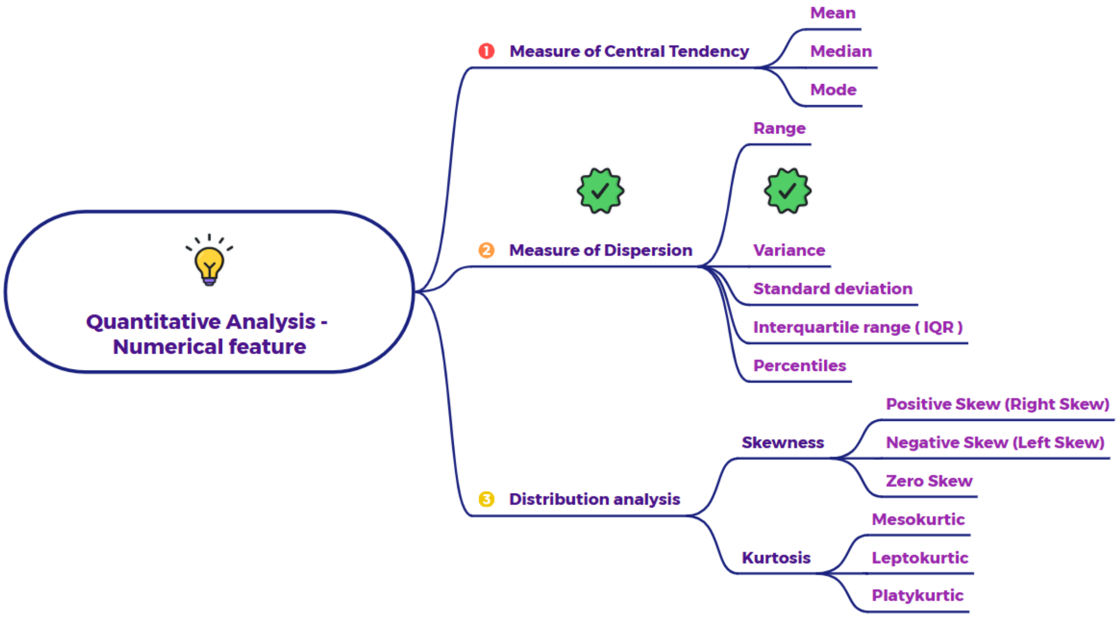
Explain measure of dispersion – Variance



Concept:

Variance essentially measures the average of the squared differences between each data point and the mean of the dataset. Squaring the differences is crucial because it handles negative differences (values below the mean) by making them positive, and it also gives more weight to values that are farther from the mean.

Formulas:

There are slightly different formulas for population variance (σ2) and sample variance (s2):

- **Population Variance (σ2)**: Used when you have data for the entire population.

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

Where:

  - xi represents each individual data point in the population.
  - μ (mu) is the population mean.
  - N is the total number of data points in the population.

- ∑ (sigma) denotes the sum.

- **Sample Variance (s2)**: Used when you have data from a sample of the population.

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

Where:

- xi represents each individual data point in the sample.
- x¯ (x-bar) is the sample mean.
- n is the total number of data points in the sample.
- The denominator is n−1 (Bessel's correction), which provides a less biased estimate of the population variance when using a sample.

Detailed Examples:

Let's use our math quiz scores again to illustrate variance. Assume these 10 scores represent a sample of students.

Scores: 12, 15, 18, 16, 15, 10, 15, 19, 14, 15

Step 1: Calculate the Sample Mean (x¯)

As we calculated before, the sample mean (x¯) is 14.9.

Step 2: Calculate the Squared Differences from the Mean

For each score, we subtract the mean and then square the result:

- (12−14.9)^2= (−2.9)^2=8.41
- (15−14.9)^2 = (0.1)^2=0.01
- (18−14.9)^2= (3.1)^2=9.61
- (16−14.9)^2=(1.1)^2=1.21
- (15−14.9)^2= (0.1)^2=0.01
- (10−14.9)^2=(−4.9)^2=24.01
- (15−14.9)^2= (0.1)^2=0.01

- $(19-14.9)^2=(4.1)^2=16.81$
- $(14-14.9)^2=(-0.9)^2=0.81$
- $(15-14.9)^2=(0.1)^2=0.01$

Step 3: Sum the Squared Differences

Sum of squared differences =
$8.41+0.01+9.61+1.21+0.01+24.01+0.01+16.81+0.81+0.01=60.9$

Step 4: Divide by (n-1) for Sample Variance

Since we are treating this as a sample ($n=10$), we divide by $n-1=10-1=9$:

Sample Variance ($s2$) = $960.9 \approx 6.77$

Interpretation:

The sample variance of the math quiz scores is approximately 6.77. A higher variance indicates a greater spread of the data points around the mean. In this case, the variance gives us a numerical measure of how much the individual scores deviate from the average score of 14.9.

Example 2: Comparing Two Datasets with the Same Mean

Consider two sets of test scores, both with a mean of 70:

- **Set A:** 65, 75, 70, 68, 72
- **Set B:** 40, 100, 70, 55, 85

Let's calculate the sample variance for each:

Set A:

- Mean ($x^{-}A$) = 70
- Squared differences: $(65-70)^2=25,(75-70)^2=25,(70-70)^2=0,(68-70)^2=4,(72-70)^2=4$
- Sum of squared differences = $25+25+0+4+4=58$
- Sample variance ($sA2$) = $5-158=458=14.5$

Set B:

- Mean ($\bar{x}_B$) = 70
- Squared differences: $(40-70)^2=900, (100-70)^2=900, (70-70)^2=0, (55-70)^2=225, (85-70)^2=225$
- Sum of squared differences = $900+900+0+225+225=2250$
- Sample variance (s_B^2) = 5–12250=42250=562.5

Interpretation:

Although both datasets have the same mean (70), Set B has a much higher variance (562.5) compared to Set A (14.5). This indicates that the scores in Set B are much more spread out from the average than the scores in Set A, which are clustered more closely around the mean.

Key Points about Variance:

- **Considers all data points:** Unlike the range, variance uses every value in the dataset, providing a more comprehensive measure of spread.
- **Squared units:** The variance is in squared units of the original data (e.g., if the data is in scores, the variance is in squared scores). This can make it less intuitive to interpret directly.
- **Sensitivity to outliers:** Similar to the mean, variance is also sensitive to outliers because the squared differences from the mean become very large for extreme values, heavily influencing the overall variance.

In summary, variance is a valuable measure of dispersion that quantifies the average squared deviation of data points from the mean. A higher variance signifies greater variability in the data. While it provides a more detailed understanding of spread than the range, its squared units can sometimes make direct interpretation challenging, which leads us to the standard deviation (the square root of the variance) as a more commonly used and interpretable measure.