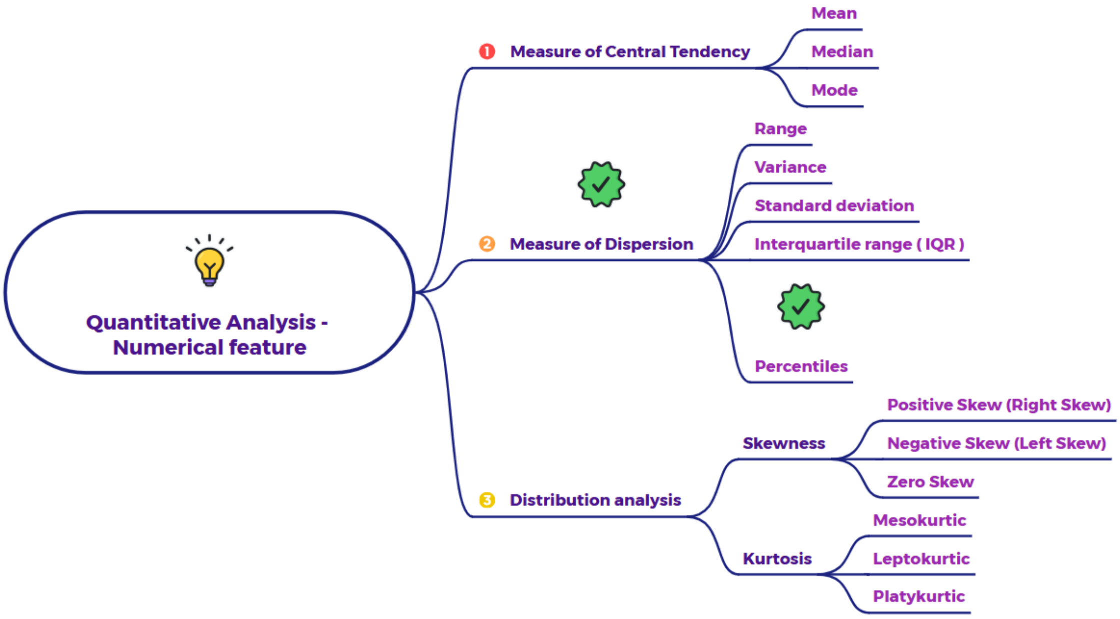
Explain measure of dispersion – Percentile



Concept:

The Pth percentile is the value below which P% of the data falls. For example, the 25th percentile is the value below which 25% of the data lies, the 50th percentile is the median (50% of the data is below it), and the 75th percentile is the value below which 75% of the data lies.

Percentiles help us understand the relative standing of a particular data point within the entire distribution. They are particularly useful for comparing values across different datasets or for understanding the distribution's shape.

Finding Percentiles:

The method for calculating percentiles can vary slightly depending on the convention used. A common approach involves the following steps for a dataset of n ordered values:

- **Order the data:** Arrange the data from smallest to largest.
- **Calculate the rank:** For the Pth percentile, the rank R is calculated as: $R=P/100\times(n+1)$
- **Determine the percentile value:**
  - If R is an integer, the Pth percentile is the value at the Rth position in the ordered data.

- If R is not an integer, the Pth percentile is found by interpolating between the values at the integer parts of R (lower integer L and upper integer $U=L+1$). The value is:
  Percentile=(Value at position L)+(R−L)×(Value at position U−Value at position L)

Detailed Examples:

Let's use a new set of exam scores for 12 students to illustrate percentiles:

Scores: 60, 65, 70, 72, 75, 78, 80, 82, 85, 88, 90, 95

Example 1: Finding the 25th Percentile (Q1)

- $n=12$
- $P=25$
- $R=25/100\times(12+1)=0.25\times13=3.25$

Since R is not an integer, we interpolate between the 3rd and 4th values in the ordered list.

- Value at position 3 = 70
- Value at position 4 = 72
- $L=3$, $U=4$, $R-L=3.25-3=0.25$

25th Percentile = $70+(0.25)\times(72-70)=70+0.25\times2=70+0.5=70.5$

Interpretation: 25% of the students scored below 70.5. This is also the first quartile (Q1).

Example 2: Finding the 50th Percentile (Median, Q2)

- $n=12$
- $P=50$
- $R=50/100\times(12+1)=0.5\times13=6.5$

Since R is not an integer, we interpolate between the 6th and 7th values.

- Value at position 6 = 78
- Value at position 7 = 80
- $L=6$, $U=7$, $R-L=6.5-6=0.5$

50th Percentile = $78+(0.5)\times(80-78)=78+0.5\times2=78+1=79$

Interpretation: 50% of the students scored below 79, and 50% scored above. This is the median of the dataset.

Example 3: Finding the 75th Percentile (Q3)

- $n=12$
- $P=75$
- $R=75/100\times(12+1)=0.75\times13=9.75$

Since R is not an integer, we interpolate between the 9th and 10th values.

- Value at position 9 = 85
- Value at position 10 = 88
- $L=9$, $U=10$, $R-L=9.75-9=0.75$

75th Percentile = $85+(0.75)\times(88-85)=85+0.75\times3=85+2.25=87.25$

Interpretation: 75% of the students scored below 87.25. This is also the third quartile (Q3).

How Percentiles Relate to Dispersion:

While individual percentiles tell us about the position of a value, looking at the difference between certain percentiles gives us insights into the spread of the data:

- **Interpercentile Range:** The difference between two percentiles. The IQR, as we discussed, is a specific interpercentile range (Q3 - Q1, or the 75th - 25th percentile).
- **Range (as the difference between the 100th and 0th percentile):** Although sometimes the minimum and maximum are considered the 0th and 100th percentiles respectively, the calculation can be less straightforward with some definitions. However, the idea of the total spread is related to the extremes of the percentile distribution.
- **Understanding Skewness:** The relative positions of the mean, median (50th percentile), and other percentiles can indicate the skewness of the distribution. For example, if the median is much lower than the mean, it suggests a positive skew.

- **Identifying Outliers:** Extreme low percentiles (e.g., below the 5th) or high percentiles (e.g., above the 95th) can help identify potential outliers in the data.

Example of Using Percentiles for Comparison:

Suppose we have the test scores of two different classes on the same exam:

- **Class A (n=15):** 65, 70, 72, 75, 78, 80, 82, 85, 88, 90, 92, 94, 96, 98, 100
- **Class B (n=15):** 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 100, 100, 100, 100

Let's look at the 25th, 50th, and 75th percentiles:

- **Class A:**
  - $R25=0.25\times16=4\Longrightarrow Q1=75$
  - $R50=0.50\times16=8\Longrightarrow Median=85$
  - $R75=0.75\times16=12\Longrightarrow Q3=94$
  - $IQRA=94-75=19$
- **Class B:**
  - $R25=0.25\times16=4\Longrightarrow Q1=65$
  - $R50=0.50\times16=8\Longrightarrow Median=85$
  - $R75=0.75\times16=12\Longrightarrow Q3=100$
  - $IQRB=100-65=35$

Interpretation:

Both classes have the same median score (85). However, the IQR for Class B (35) is larger than that of Class A (19), indicating that the middle 50% of scores in Class B are more spread out than in Class A. Also, the 75th percentile is much higher in Class B, suggesting a cluster of high-achieving students, while the lower percentiles show a wider range of performance.

In summary, percentiles provide a detailed way to understand the distribution of numerical data by indicating the value below which a certain percentage of the data falls. While individual percentiles describe the position of values, the differences between them (like the IQR) give us valuable insights into the spread and shape of the distribution, and can help in identifying skewness and potential outliers.