Numerical Summary of Level and Spread

Numerical summarization of **level** and **spread** is fundamental in data analysis to describe the data's **central tendency** (level) and **variability** (spread). These metrics help understand the overall structure of the dataset, including its typical values and the range of variability.

Key Metrics

1. Level (Central Tendency)

- **Mean**: The average value of the data.
- **Median**: The middle value when the data is sorted.
- **Mode**: The most frequently occurring value.

2. Spread (Variability)

- **Range**: The difference between the maximum and minimum values.
- **Variance**: The average of the squared differences from the mean.
- **Standard Deviation**: The square root of the variance, showing the spread in the same units as the data.
- **Interquartile Range (IQR)**: The range of the middle 50% of the data, calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

Syntax for Key Metrics

Central Tendency

```
data.mean()       # Mean
data.median()     # Median
data.mode()       # Mode (returns a series)
```

Spread

```
data.max() - data.min()   # Range
data.var()                # Variance
data.std()                # Standard deviation
data.quantile(0.25)       # 25th percentile (Q1)
data.quantile(0.75)       # 75th percentile (Q3)
Q3 - Q1                   # Interquartile Range (IQR)
```

Example

```
import pandas as pd
# Sample Dataset
data = {
```

```
    "Scores": [45, 50, 67, 68, 75, 80, 85, 90, 92, 100]
}

# Create DataFrame
df = pd.DataFrame(data)
# Numerical Summaries
mean = df["Scores"].mean()
median = df["Scores"].median()
mode = df["Scores"].mode()[0]  # Accessing the first mode if multiple exist
data_range = df["Scores"].max() - df["Scores"].min()
variance = df["Scores"].var()
std_dev = df["Scores"].std()
Q1 = df["Scores"].quantile(0.25)
Q3 = df["Scores"].quantile(0.75)
iqr = Q3 - Q1
# Display Results
print("Numerical Summary of Level and Spread:")
print(f"Mean: {mean}")
print(f"Median: {median}")
print(f"Mode: {mode}")
print(f"Range: {data_range}")
print(f"Variance: {variance}")
print(f"Standard Deviation: {std_dev}")
print(f"Interquartile Range (IQR): {iqr}")
```

Output

```
Numerical Summary of Level and Spread:
Mean: 75.2
Median: 75.0
Mode: 45
Range: 55
Variance: 303.51111111111106
Standard Deviation: 17.429855544366473
Interquartile Range (IQR): 25.0
```

1. **Dataset**:
   - The Scores column represents a numerical dataset of student scores.
2. **Central Tendency**:
   - **Mean**: The average score is calculated using .mean().
   - **Median**: The middle value (75) is obtained using .median().
   - **Mode**: The most frequent score is retrieved using .mode().
3. **Spread**:
   - **Range**: The difference between the highest (100) and lowest (45) values.

- **Variance**: Indicates how much the data points differ from the mean on average (measured in squared units).
- **Standard Deviation**: Shows the spread in the same units as the data, making it easier to interpret.
- **IQR**: Highlights the range of the middle 50% of the scores, calculated as Q3 - Q1.

Key Insights

1. **Level**:
   - **Mean** (75.2) and **Median** (75.0) are close, suggesting the data is symmetrically distributed.
   - The **Mode** (45) occurs less frequently, as it's the only repeating value in the dataset.
2. **Spread**:
   - **Range** (55) shows the data spans widely.
   - **Variance** and **Standard Deviation** reflect moderate variability in the scores.
   - **IQR** (25.0) indicates that the middle 50% of the data is tightly clustered between the 25th and 75th percentiles.

Use Cases

- **Mean** and **Median** help identify the central tendency for reporting a typical value.
- **Variance** and **Standard Deviation** are essential for understanding data dispersion, especially in statistical modeling.
- **IQR** is particularly useful for detecting outliers.