



Data Science



By:
Dr. Shikha Deep



Descriptive Statistics

Part A:

Descriptive Statistics & Data Preparation

- Descriptive Statistics
- Measures of Central Tendency: Mean, Median, Mode
- Measures of Spread: Range, Variance, Standard Deviation, IQR
- Skewness & Kurtosis (Measuring Asymmetry)
- Python with `pandas.describe()`, `scipy.stats`, `numpy`



1. Measures of Central Tendency

- These describe the "center" of the data.

(A) Mean (Average):

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

(B) Median: Middle value when data is sorted. If even number of observations, median = average of two middle values.

(C) Mode: Most frequent value(s).



2. Measures of Dispersion (Spread)

These describe how spread out the data is.

(A) Range: Max – Min.

(B) Variance (σ^2): Average squared deviation from mean.

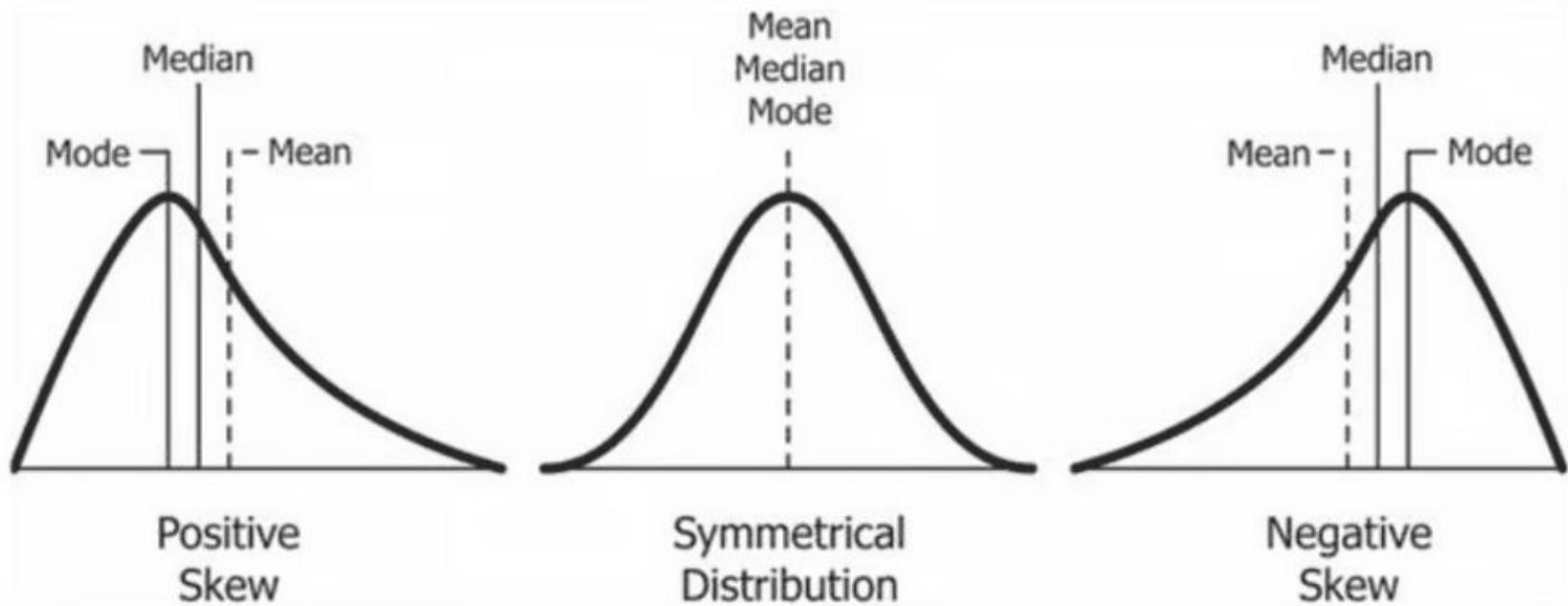
$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{\sigma^2}$$

(C) Standard Deviation (σ): Square root of variance.

(D) IQR (Interquartile Range): Q3 – Q1 (middle 50% of data).

3. Measuring Asymmetry

- **Skewness:** Tells if data is symmetric or skewed.
 - Positive skew → long right tail
 - Negative skew → long left tail
- **Kurtosis:** Tells if data is flat or peaked compared to normal distribution.
 - High kurtosis → heavy tails
 - Low kurtosis → light tails

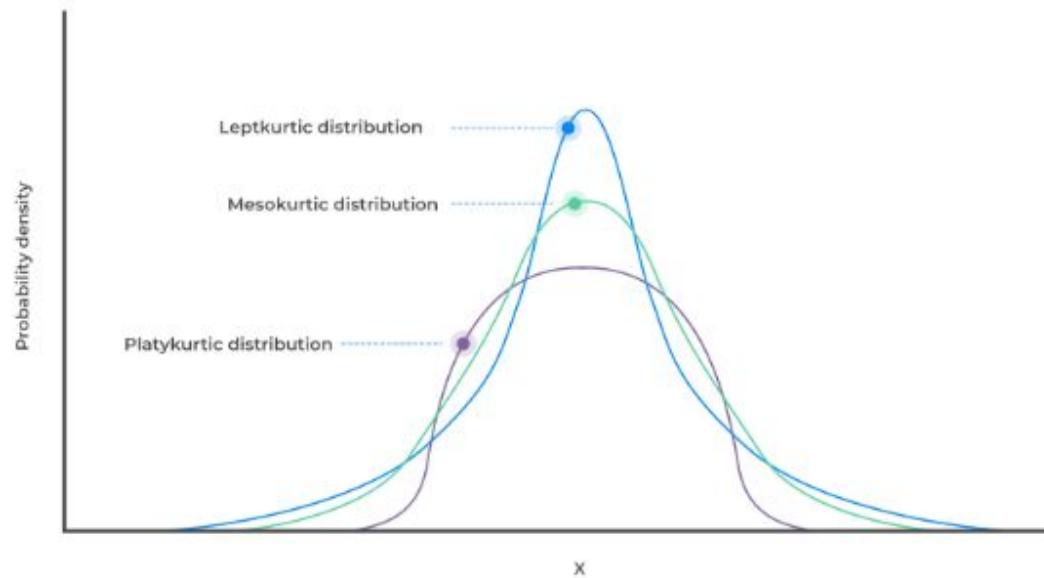


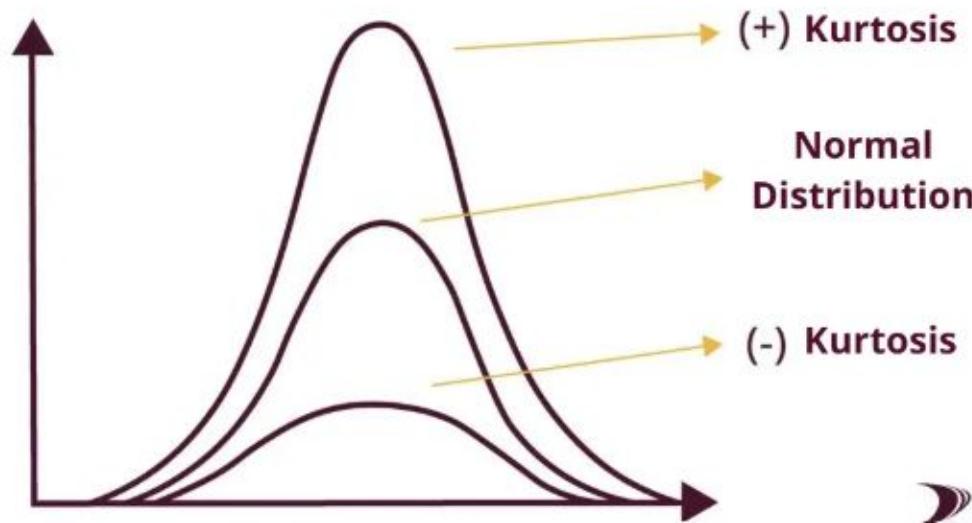


Kurtosis



Kurtosis





Code 1: Mean, Median, Mode, Skweness, Kurtosis

```
import numpy as np  
import pandas as pd  
from scipy import stats
```

```
# Sample dataset  
data = [45, 50, 55, 60, 65, 70, 70, 80, 85, 90]
```

```
# Convert to pandas Series  
s = pd.Series(data)
```

```
# Measures of Central Tendency  
mean = s.mean()  
median = s.median()  
mode = s.mode().tolist()
```



Measures of Spread

```
variance = s.var()
```

```
std_dev = s.std()
```

```
iqr = s.quantile(0.75) - s.quantile(0.25)
```

Skewness & Kurtosis

```
skewness = s.skew()
```

```
kurtosis = s.kurt()
```



```
print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
print("Variance:", variance)
print("Standard Deviation:", std_dev)
print("Interquartile Range (IQR):", iqr)
print("Skewness:", skewness)
print("Kurtosis:", kurtosis)
```



OUTPUT:

Mean: 67.0

Median: 67.5

Mode: [70]

Variance: 206.67

Standard Deviation: 14.38

Interquartile Range (IQR): 25.0

Skewness: -0.07 # almost symmetric

Kurtosis: -1.34 # flatter than normal

Skewness and Kurtosis Visualization

CODE:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import skew, kurtosis
```

```
# Sample dataset
data = [45, 50, 55, 60, 65, 70, 70, 80, 85, 90]
s = pd.Series(data)
```

```
# Calculate skewness and kurtosis
```

```
skewness = skew(s)
```

```
kurt = kurtosis(s)
```

```
print("Skewness:", skewness)
```

```
print("Kurtosis:", kurt)
```

```
# Plot histogram with KDE
```

```
plt.figure(figsize=(8,5))
```

```
sns.histplot(s, bins=8, kde=True, color="skyblue",  
edgecolor="black")
```

```
plt.title(f"Distribution of Data\nSkewness = {skewness:.2f},  
Kurtosis = {kurt:.2f}", fontsize = 14)
```



```
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.show()
```

NOTE:

- Both list and Series can be plotted.
- But **Pandas Series is more powerful** because it has built-in **statistics functions**.



Thanks!