Distribution and Variable in Univariate Analysis

In **univariate analysis**, the focus is on analyzing and summarizing a single variable to understand its distribution, patterns, and central tendencies.

Key Concepts

1. Variable

- A variable is a characteristic or attribute that can take on different values.
- Variables can be:
    - **Numerical (Quantitative)**: Continuous or discrete values, e.g., age, height, scores.
    - **Categorical (Qualitative)**: Labels or categories, e.g., gender, region, colors.

2. Distribution

- The **distribution** of a variable describes how its values are spread or distributed across a range.
- Key aspects of a distribution include:
    - **Frequency**: How often each value appears.
    - **Shape**: Skewness, symmetry, kurtosis (peakedness).
    - **Outliers**: Extreme values that deviate from other observations.

Analyzing Distribution

- For **numerical variables**, use histograms, boxplots, and descriptive statistics.
- For **categorical variables**, use bar plots and frequency tables.

Syntax for Distribution Analysis

Numerical Variables

```
# Descriptive Statistics
data.describe()

# Histogram
data.hist(bins=10)

# Boxplot
data.boxplot()
```

Categorical Variables

```
# Frequency Table
data.value_counts()

# Bar Plot
```

```
data.value_counts().plot(kind='bar')
```

Example

```
import pandas as pd
import matplotlib.pyplot as plt

# Sample Dataset
data = {
    "Scores": [45, 50, 67, 68, 75, 80, 85, 90, 92, 100],
    "Grade": ["C", "C", "B", "B", "B", "A", "A", "A", "A", "A"]
}

# Create DataFrame
df = pd.DataFrame(data)

# Numerical Variable: Distribution Analysis
print("Numerical Summary of Scores:")
print(df["Scores"].describe())  # Summary statistics

plt.figure(figsize=(12, 5))

# Histogram
plt.subplot(1, 2, 1)
plt.hist(df["Scores"], bins=5, color='skyblue', edgecolor='black')
plt.title("Histogram of Scores")
plt.xlabel("Scores")
plt.ylabel("Frequency")

# Boxplot
plt.subplot(1, 2, 2)
plt.boxplot(df["Scores"], vert=False, patch_artist=True, boxprops=dict(facecolor='orange'))
plt.title("Boxplot of Scores")
plt.xlabel("Scores")

plt.tight_layout()
plt.show()

# Categorical Variable: Distribution Analysis
print("\nFrequency of Grades:")
print(df["Grade"].value_counts())

# Bar Plot for Grades
df["Grade"].value_counts().plot(kind="bar", color='purple', edgecolor='black')
plt.title("Bar Plot of Grades")
plt.xlabel("Grades")
plt.ylabel("Frequency")
plt.show()
```

Output

Numerical Summary (Scores):

```
Numerical Summary of Scores:
count     10.000000
mean      75.200000
std       17.429856
min       45.000000
25%       67.000000
50%       75.000000
75%       90.000000
max      100.000000
```

Frequency of Grades:

```
Frequency of Grades:
A    5
B    3
C    2
Name: Grade, dtype: int64
```

1. **Dataset**:
   - The dataset contains two variables:
     - **Scores**: A numerical variable (student scores).
     - **Grade**: A categorical variable (grades assigned to students).
2. **Numerical Variable (Scores)**:
   - **Descriptive Statistics**:
     - Using .describe(), we obtain metrics like mean, median, min, max, and quartiles.
   - **Histogram**:
     - Displays the frequency distribution of scores in bins.
     - Shows the spread and shape of the data.
   - **Boxplot**:
     - Visualizes the spread and highlights outliers, median, and quartiles.
3. **Categorical Variable (Grade)**:
   - **Frequency Table**:
     - Counts the occurrences of each grade.
   - **Bar Plot**:
     - Displays the distribution of grades visually.

Key Insights

1. **Numerical Variable (Scores)**:
   - The histogram shows the data distribution (e.g., symmetric, skewed).
   - The boxplot highlights the median (75) and spread, with no visible outliers.
2. **Categorical Variable (Grade)**:
   - The frequency table and bar plot indicate that most students scored "A" grades.

Use Cases

- **Numerical Distribution**: Helps detect outliers, skewness, or specific patterns in the data.
- **Categorical Distribution**: Useful for understanding class imbalance, e.g., in classification tasks.