# measures_of_variability

September 30, 2024

# 1 Measures of Variability: Understanding Spread and Dispersion

In this notebook, we will explore **measures of variability** in statistics. These measures help us understand how spread out or dispersed the data is. We will calculate and visualize the following key measures of dispersion: - **Range** - **Variance** - **Standard Deviation** - **Interquartile Range (IQR)**

We will use a random dataset and visualize the results to gain insights into the spread of the data.

### ***By Blake Zenuni***

- My github: https://github.com/BlakeBelisarius**
- Project in this repo: https://github.com/BlakeBelisarius/MacroMarketPulse**
- Yellowbrick from DistrictDataLabs Repo to produce visualizations for your machine learning workflow: https://github.com/BlakeBelisarius/yellowbrick**

---

[2]:
```python
# Importing necessary libraries

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

## 1.1 Generating Random Data

Let's generate a random dataset using a **normal distribution**. This dataset will be used to compute the measures of variability. As before: - `loc=50`: The mean of the distribution is 50. - `scale=10`: The standard deviation is 10. - `size=1000`: We will generate 1000 data points.

[3]:
```python
# Generating a random dataset
np.random.seed(0)
data = np.random.normal(loc=50, scale=10, size=1000)
```

## 1.2 1. Calculating the Range

The **range** is the simplest measure of variability and is calculated as the difference between the maximum and minimum values in the dataset.

[4]:

```
# Calculate the range
data_range = np.ptp(data)  # Peak-to-peak (max - min)

# Print the range for reference
print(f"Range: {data_range:.2f}")
```

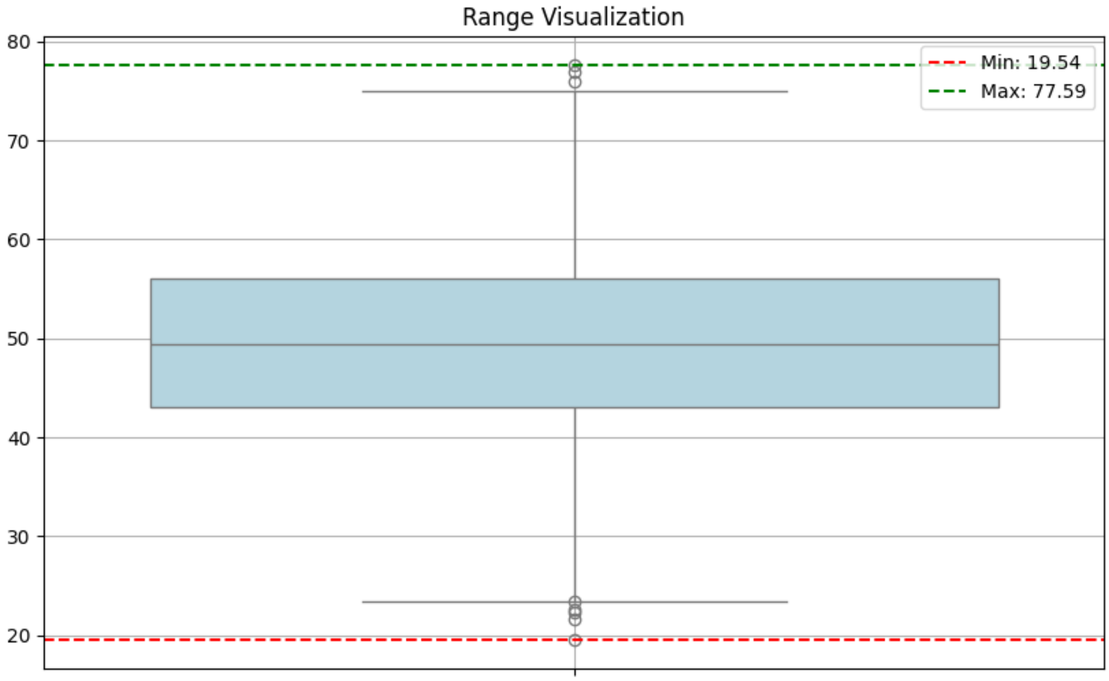
```
Range: 58.05
```

### 1.2.1 Visualization: Range

The **range** only takes into account the extremes of the data. We will visualize the entire dataset using a box plot and highlight the range using lines.

[5]:

```
# Create the plot for range
plt.figure(figsize=(10, 6))
sns.boxplot(data=data, color='lightblue')

# Highlight the range using max and min
plt.axhline(np.min(data), color='red', linestyle='dashed', label=f'Min: {np.
  ↪min(data):.2f}')
plt.axhline(np.max(data), color='green', linestyle='dashed', label=f'Max: {np.
  ↪max(data):.2f}')

# Add labels and title
plt.title('Range Visualization')
plt.legend()
plt.grid(True)
plt.show()
```

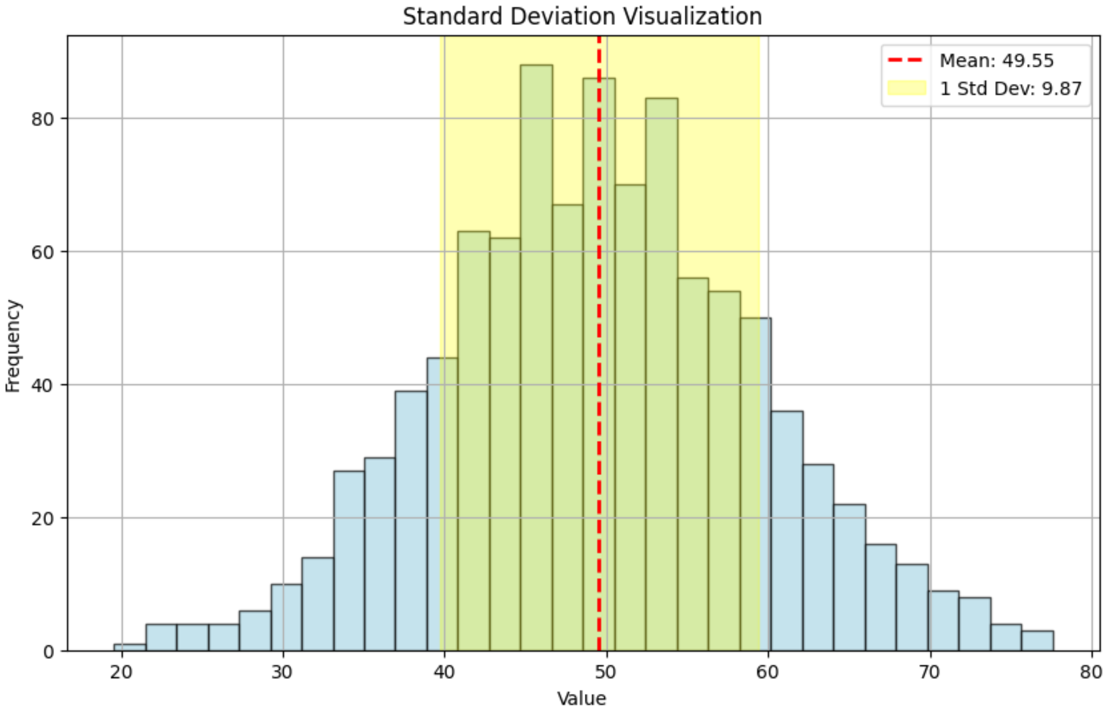


### 1.2.2 Visualization: Variance and Standard Deviation

We'll plot the dataset distribution and highlight the **mean** along with a shaded area representing one standard deviation above and below the mean. This shows how much the data deviates from the center.

[6]:

```
# Create the plot for standard deviation visualization
plt.figure(figsize=(10, 6))

# Plot the histogram of the data
plt.hist(data, bins=30, color='lightblue', edgecolor='black', alpha=0.7)



# Calculate the mean and standard deviation
mean = np.mean(data)
std_dev = np.std(data)  # Define the standard deviation

plt.axvline(mean, color='red', linestyle='dashed', linewidth=2, label=f'Mean:␣
  ↪{mean:.2f}')
plt.axvspan(mean - std_dev, mean + std_dev, color='yellow', alpha=0.3,␣
  ↪label=f'1 Std Dev: {std_dev:.2f}')

# Add labels and title
plt.title('Standard Deviation Visualization')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()

plt.grid(True)
plt.show()
```



## 1.3 3. Interquartile Range (IQR)

The **Interquartile Range (IQR)** is the range between the 25th percentile (Q1) and the 75th percentile (Q3). This measure ignores the extremes and focuses on the middle 50% of the data.

[7]:
```
# Calculate IQR
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
iqr = q3 - q1

# Print the IQR for reference
print(f"IQR: {iqr:.2f}")
```

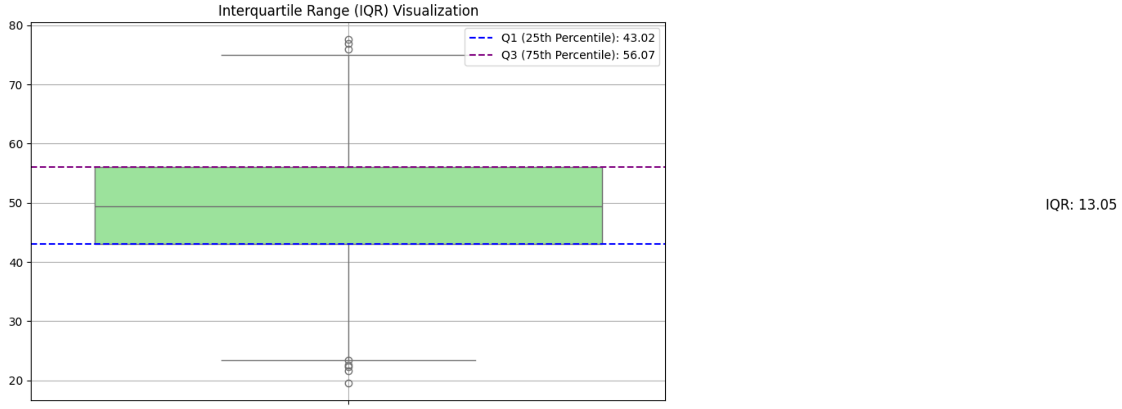
```
IQR: 13.05
```

### 1.3.1 Visualization: IQR

We will use a box plot to visualize the **Interquartile Range (IQR)**. The box plot highlights the quartiles, the median, and any potential outliers.

[8]:
```
# Create the box plot for IQR visualization
plt.figure(figsize=(10, 6))

# Plot the box plot
sns.boxplot(data=data, color='lightgreen')
```

```
# Highlight Q1, Q3, and IQR
plt.axhline(q1, color='blue', linestyle='dashed', label=f'Q1 (25th Percentile):␣
  ↪{q1:.2f}')
plt.axhline(q3, color='purple', linestyle='dashed', label=f'Q3 (75th␣
  ↪Percentile): {q3:.2f}')
plt.text(1.1, (q1+q3)/2, f'IQR: {iqr:.2f}', verticalalignment='center',␣
  ↪color='black', fontsize=12)

# Add labels and title
plt.title('Interquartile Range (IQR) Visualization')
plt.legend()
plt.grid(True)
plt.show()
```



## 1.4 Conclusion

In this notebook, we calculated and visualized the **measures of variability** for a random dataset: - **Range**: The difference between the maximum and minimum values. - **Variance**: The average squared deviation from the mean. - **Standard Deviation**: The square root of variance, which brings dispersion back to the original units. - **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, focusing on the middle 50% of the data.

These measures provide us with a deeper understanding of the spread and dispersion in our dataset.