Measures of Spread

Range, Variance and Stnadard Deviation

- **Example Datasets:**

$$\boldsymbol{X} = (10, 0, 10, 20, 30) \qquad \boldsymbol{and} \qquad \boldsymbol{Y} = (8, 9, 10, 11, 12)$$

- **Population Mean:**

$$\boldsymbol{X} = \boldsymbol{\mu} = \frac{(10 + 0 + 10 + 20 + 30)}{5} = \boxed{\mathbf{10}}$$

$$\boldsymbol{Y} = \boldsymbol{\mu} = \frac{(8 + 9 + 10 + 11 + 12)}{5} = \boxed{\mathbf{10}}$$

- **Population Range:** The difference between the high & low datapoints in dataset.

$$\boldsymbol{X} = (\ 30 - (-10)\ ) = \boxed{\mathbf{40}} \qquad \boldsymbol{and} \qquad \boldsymbol{Y} = (\ 12 - 8\ ) = \boxed{\mathbf{4}}$$

- **Variance:** Sum of squared differences of datapoints divided by total.

$$\boldsymbol{X} = \boldsymbol{\sigma^2} = \frac{(-10 - 10)^2 + (0 - 10)^2 + (10 - 10)^2 + (20 - 10)^2 + (30 - 10)^2}{5}$$

$$= \frac{1000}{5} = \boxed{\mathbf{200^{u^2}}}$$

$$\boldsymbol{Y} = \boldsymbol{\sigma^2} = \frac{(8 - 10)^2 + (9 - 10)^2 + (10 - 10)^2 + (11 - 10)^2 + (12 - 10)^2}{5}$$

$$= \frac{10}{5} = \boxed{\mathbf{2^{u^2}}}$$

Summary: We can clearly see which dataset has the greater variance or spread. But the result σ^2 is a unit squared, meters, miles, etc. which brings us to finding the standard deviation σ.

- **Standard Deviation:**

$$\boldsymbol{SD} = \sqrt{variance} = \sqrt{\sigma^2} = \sigma \quad \boldsymbol{then,} \quad \sigma = \sqrt{200} = 10\sqrt{2} \quad \boldsymbol{and} \quad \sigma = \sqrt{2}$$

Sample Variance

Sample Variance and Stnadard Deviation

- **Sample Datasets:**

$$\boldsymbol{X} = (10, 20, 30) \qquad \boldsymbol{and} \qquad \boldsymbol{Y} = (10, 11, 12)$$

- **Sample Formula:**

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x}^2)}{n - 1}$$

Explanation Sum of all datapoints from $\boldsymbol{i = 1}$, denoted with lowercase $\boldsymbol{n}$, which is convention for the sample (uppercase $\boldsymbol{N}$ denoting the population), then taking each datapoint $\boldsymbol{x_i}$, and from it subtracting the sample mean $\bar{\boldsymbol{x}}$, then dividing it by one less than the total datapoints $\boldsymbol{n-1}$.

- **Sample Mean:**

$$\boldsymbol{X} = \bar{\boldsymbol{x}} = \frac{(10+20+30)}{2} = \boxed{\mathbf{30}}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} = \frac{(10+11+12)}{2} = \boxed{\mathbf{16.5}}$$

- **Sample Range:** The difference between the high & low datapoints in dataset.

$$\boldsymbol{X} = (\ 30 - 10\ ) = \boxed{\mathbf{20}} \qquad \boldsymbol{and} \qquad \boldsymbol{Y} = (\ 12 - 10\ ) = \boxed{\mathbf{2}}$$

- **Unbiased Sample Variance:** Sum of squared differences of datapoints divided by total minus one.

$$\boxed{\boldsymbol{S^2_{n-1}} = \frac{(10-30)^2 + (20-30)^2 + (30-30)^2}{2}}$$

$$= \quad \frac{500}{2} \quad = \quad \mathbf{250}^{\boldsymbol{u^2}}$$

$$\boxed{\boldsymbol{S^2_{n-1}} = \frac{(10-16.5)^2 + (11-16.5)^2 + (12-16.5)^2}{2}}$$

$$\approx \quad \frac{92.75}{2} \quad \approx \quad \mathbf{46.38}^{\boldsymbol{u^2}}$$

Summary: We can clearly see which dataset has the greater variance or spread. But the result σ^2 is a unit squared, meters, miles, etc. which brings us to finding the standard deviation σ.

- **Sample Standard Deviation:**

$$\boldsymbol{SSD} = \sqrt{variance} = \sqrt{S^2_{n-1}} = \boldsymbol{S}$$

$$\boldsymbol{then,}$$

$$\boldsymbol{S} = \sqrt{250} = 5\sqrt{10} \approx 15.81 \quad \boldsymbol{and} \quad \boldsymbol{S} \approx \sqrt{46.38} \approx 6.81$$

Note Because the square root function is non-linear, this actually is not a true unbiased estimate of the population standard deviation. We use this sample variance to calculate the sample standard deviation but it is NOT unbiased.