Variation

By

Vaibhav Vasani

Variation

- Variation is a measure of how spread out the data is around the centre of the data.

- Measures of variation are statistics of how far away the values in the observations (data points) are from each other.

- There are different measures of variation. The most commonly used are:

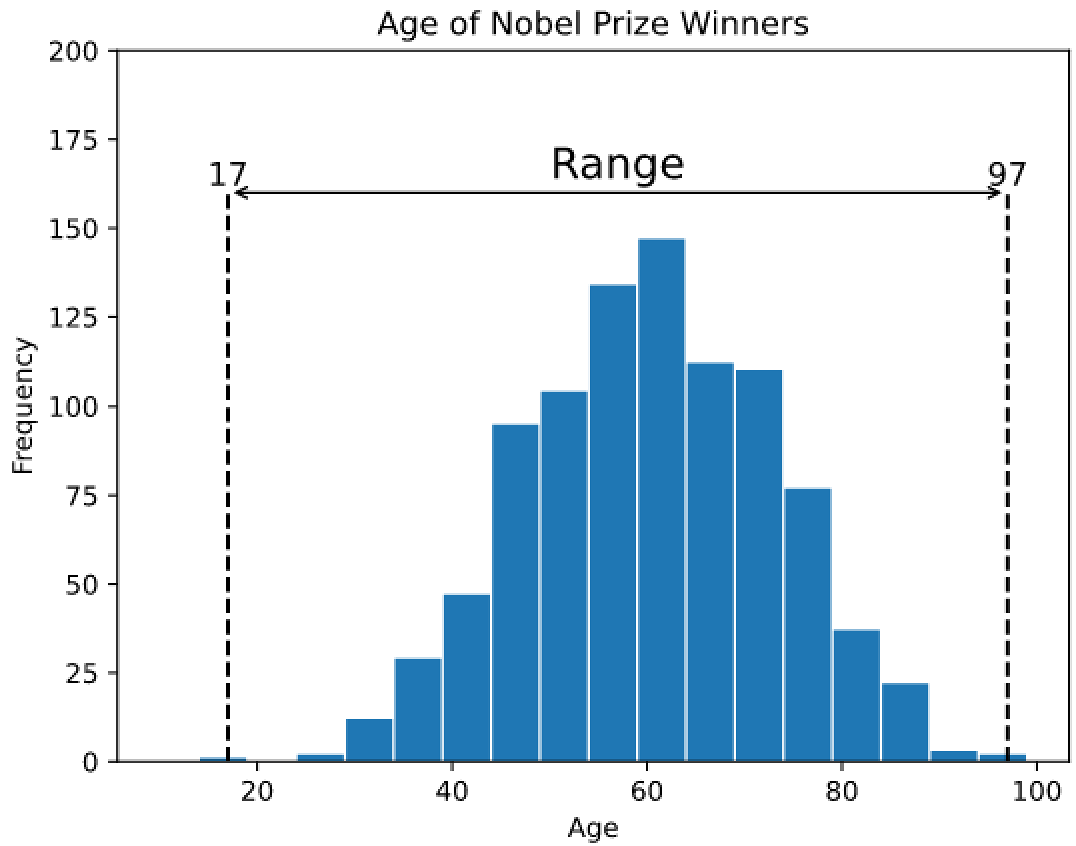
- Range
- Quartiles and Percentiles
- Interquartile Range
- Standard Deviation

- Measures of variation combined with an average (measure of centre) gives a good picture of the distribution of the data.
- **Note:** These measures of variation can only be calculated for numerical data.
-

Range

- The range is the difference between the smallest and the largest value of the data.
- Range is the simplest measure of variation.

Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing the range:

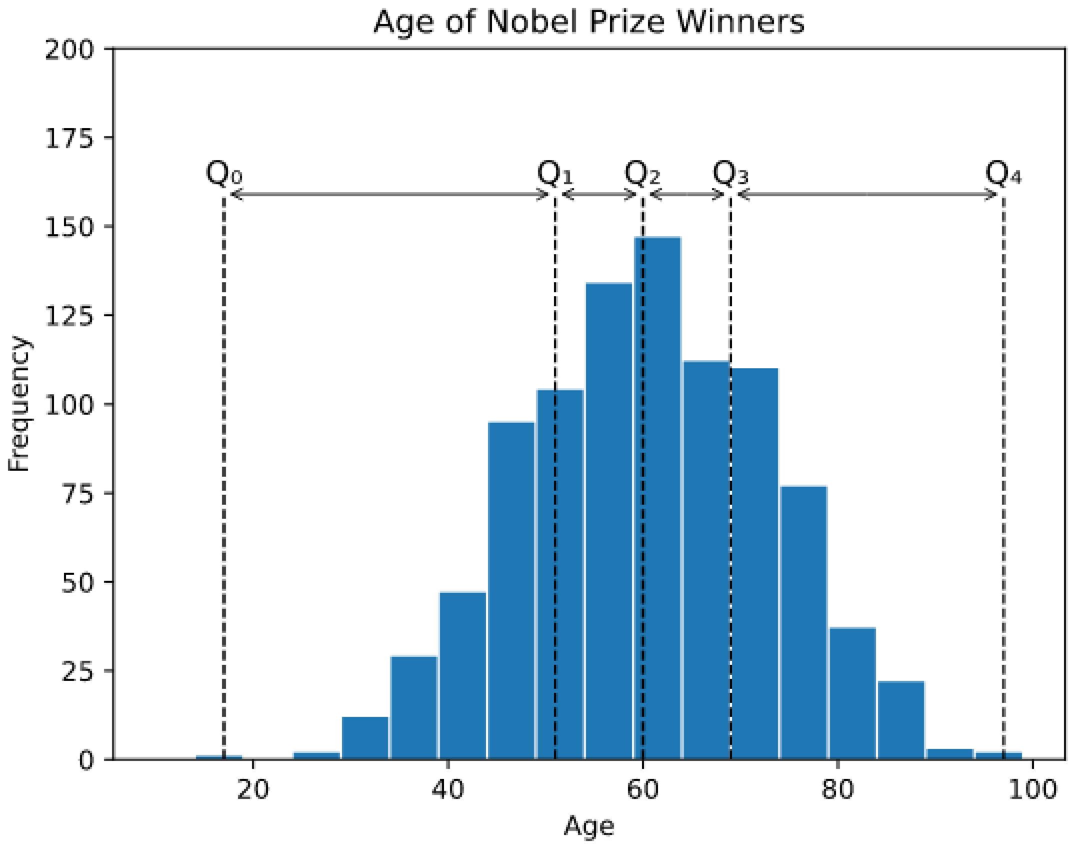


The youngest winner was 17 years and the oldest was 97 years. The range of ages for Nobel Prize winners is then 80 years.

Quartiles and Percentiles

- Quartiles and percentiles are ways of separating equal numbers of values in the data into parts.
- **Quartiles** are values that separate the data into four equal parts.
- **Percentiles** are values that separate the data into 100 equal parts.

Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing the **quartiles**:

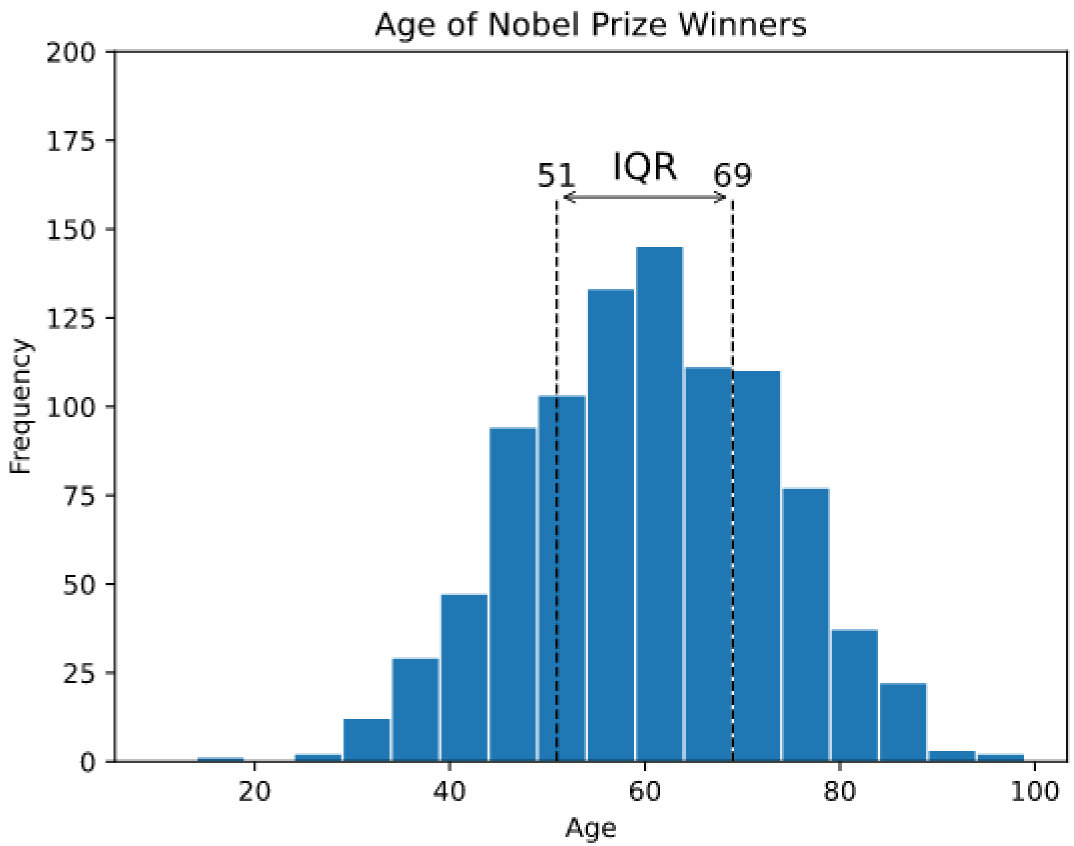


- The quartiles (Q_0,Q_1,Q_2,Q_3,Q_4) are the values that separate each quarter.
- Between Q_0 and Q_1 are the 25% lowest values in the data. Between Q_1 and Q_2 are the next 25%. And so on.
  - Q_0 is the smallest value in the data.
  - Q_2 is the middle value (median).
  - Q_4 is the largest value in the data.

Interquartile Range

- Interquartile range is the difference between the first and third quartiles (Q_1 and Q_3).
- The 'middle half' of the data is between the first and third quartile.

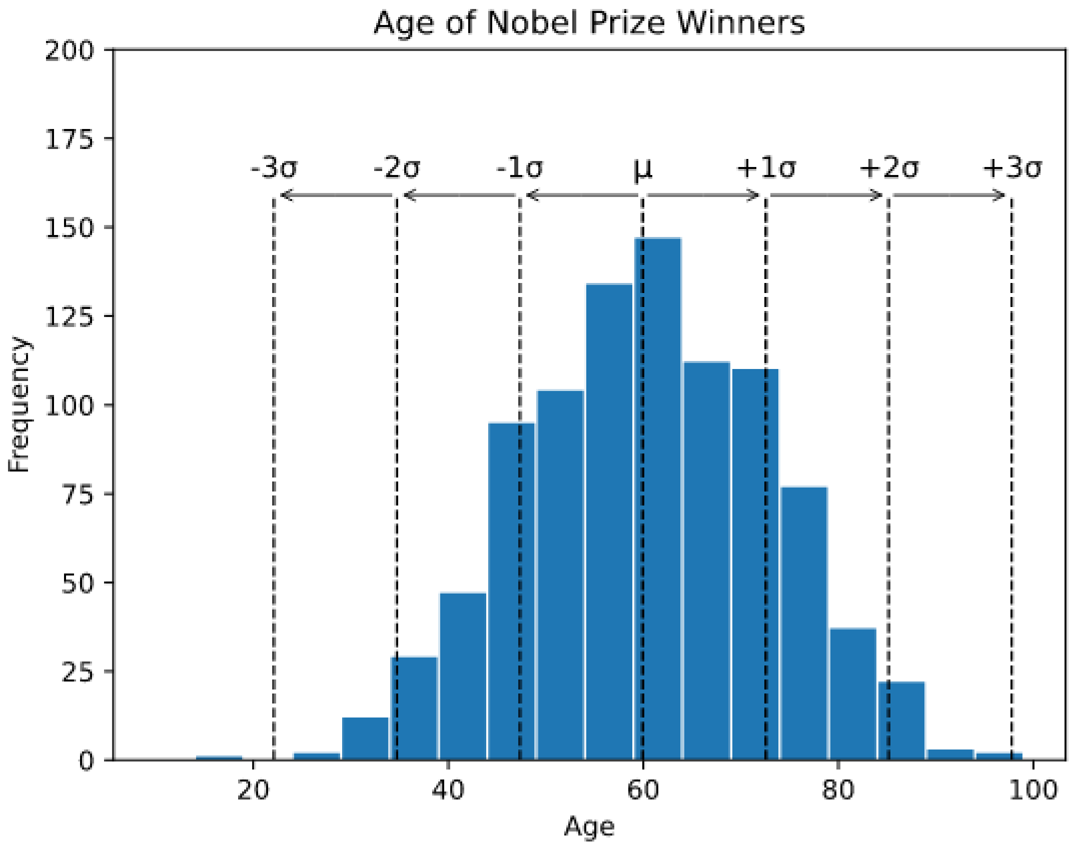
- Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing the **interquartile range (IQR)**:



Here, the middle half of is between 51 and 69 years. The interquartile range for Nobel Prize winners is then 18 years.

Standard Deviation

- Standard deviation is the most used measure of variation.
- Standard deviation (σ) measures how far a 'typical' observation is from the average of the data (μ).
- Standard deviation is important for many statistical methods.

- Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing **standard deviations**:
- 

Note: Values within one standard deviation (σ) are considered to be typical.
Values outside three standard deviations are considered to be **outliers**.

- Calculating the Range
- The range can only be calculated for numerical data.
- First, find the smallest and largest values of this example:
  - 13, 21, 21, 40, 48, 55, 72
- Calculate the difference by subtracting the smallest from the largest:
  - 72 - 13 = 59

Calculating the Range with Programming

- The range can easily be found with many programming languages.
- Using software and programming to calculate statistics is more common for bigger sets of data, as finding it manually becomes difficult.

- Example
- With Python use the NumPy library ptp() method to find the range of the values 13, 21, 21, 40, 48, 55, 72:
- import numpy

  values = [13,21,21,40,48,55,72]

  x = numpy.ptp(values)

  print(x)

- Example
- Use the R min() and max() functions to find the range of the values 13, 21, 21, 40, 48, 55, 72:
- values <- c(13,21,21,40,48,55,72)

  max(values) - min(values)

- **Note:** The range() function in R returns the smallest and largest values.

Calculating Quartiles with Programming

- Example
- With Python use the NumPy library quantile() method to find the quartiles of the values 13, 21, 21, 40, 42, 48, 55, 72:
- import numpy

```
values = [13,21,21,40,42,48,55,72]

x = numpy.quantile(values, [0,0.25,0.5,0.75,1])

print(x)
```

- Example
- Use the R quantile() function to find the quantiles of the values 13, 21, 21, 40, 42, 48, 55, 72:
- values <- c(13,21,21,40,42,48,55,72)

  quantile(values)

- Percentiles
- **Percentiles** are values that separate the data into 100 equal parts.
- For example, The 95th percentile separates the lowest 95% of the values from the top 5%
- The 25th percentile ($P_{25\%}$) is the same as the first quartile (Q_1).
- The 50th percentile ($P_{50\%}$) is the same as the second quartile (Q_2) and the median.
- The 75th percentile ($P_{75\%}$) is the same as the third quartile (Q_3)
-

Calculating Percentiles with Programming

- Percentiles can easily be found with many programming languages.

- Example
- With Python use the NumPy library percentile() method to find the 65th percentile of the values 13, 21, 21, 40, 42, 48, 55, 72:
- import numpy

```
values = [13,21,21,40,42,48,55,72]

x = numpy.percentile(values, 65)

print(x)
```

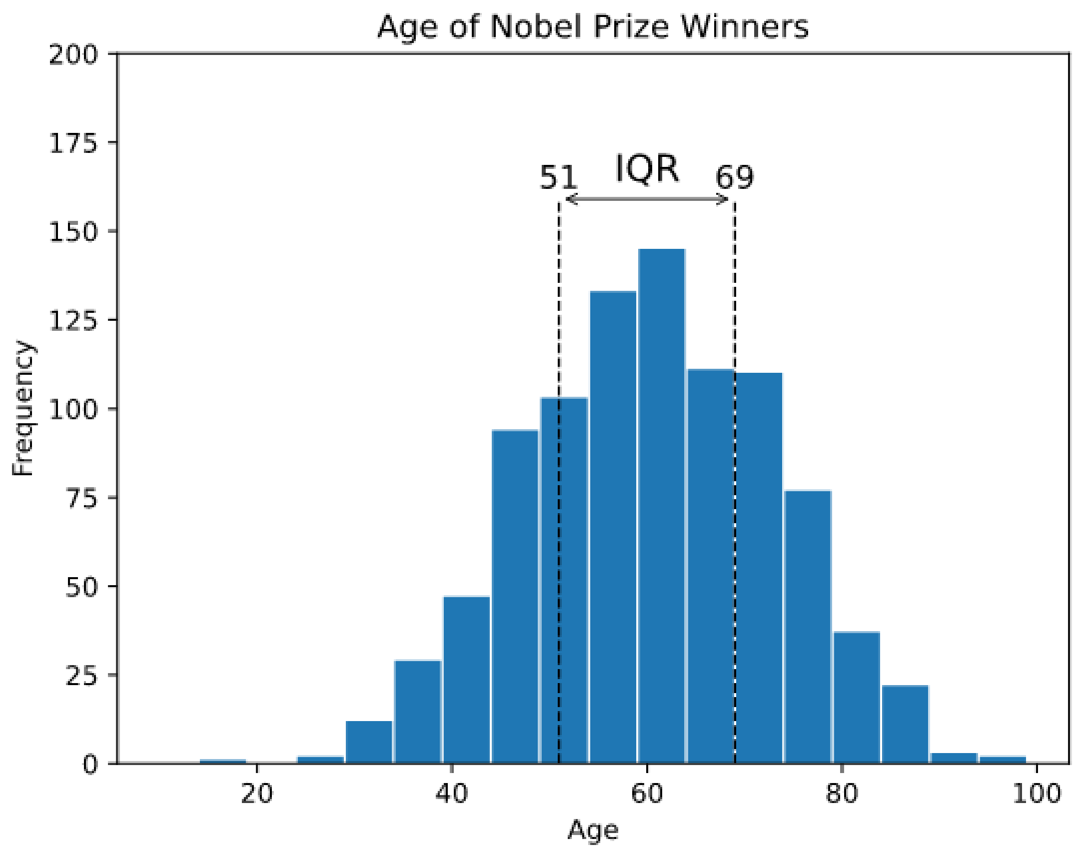
- Example
- Use the R quantile() function to find the 65th percentile (0.65) of the values 13, 21, 21, 40, 42, 48, 55, 72:
- values <- c(13,21,21,40,42,48,55,72)

  quantile(values, 0.65)

Interquartile range is a measure of variation, which describes how spread out the data is.

- Interquartile Range
- Interquartile range is the difference between the first and third quartiles (Q1 and Q3).
- The 'middle half' of the data is between the first and third quartile.
- The first quartile is the value in the data that separates the bottom 25% of values from the top 75%.
- The third quartile is the value in the data that separates the bottom 75% of the values from the top 25%

- Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing the **interquartile range (IQR)**:



Here, the middle half of is between 51 and 69 years. The interquartile range for Nobel Prize winners is then 18 years.

Calculating the Interquartile Range with Programming

- The interquartile range can easily be found with many programming languages.
-

- Example
- With Python use the SciPy library iqr() method to find the interquartile range of the values 13, 21, 21, 40, 42, 48, 55, 72:
- from scipy import stats

  values = [13,21,21,40,42,48,55,72]

  x = stats.iqr(values)

  print(x)

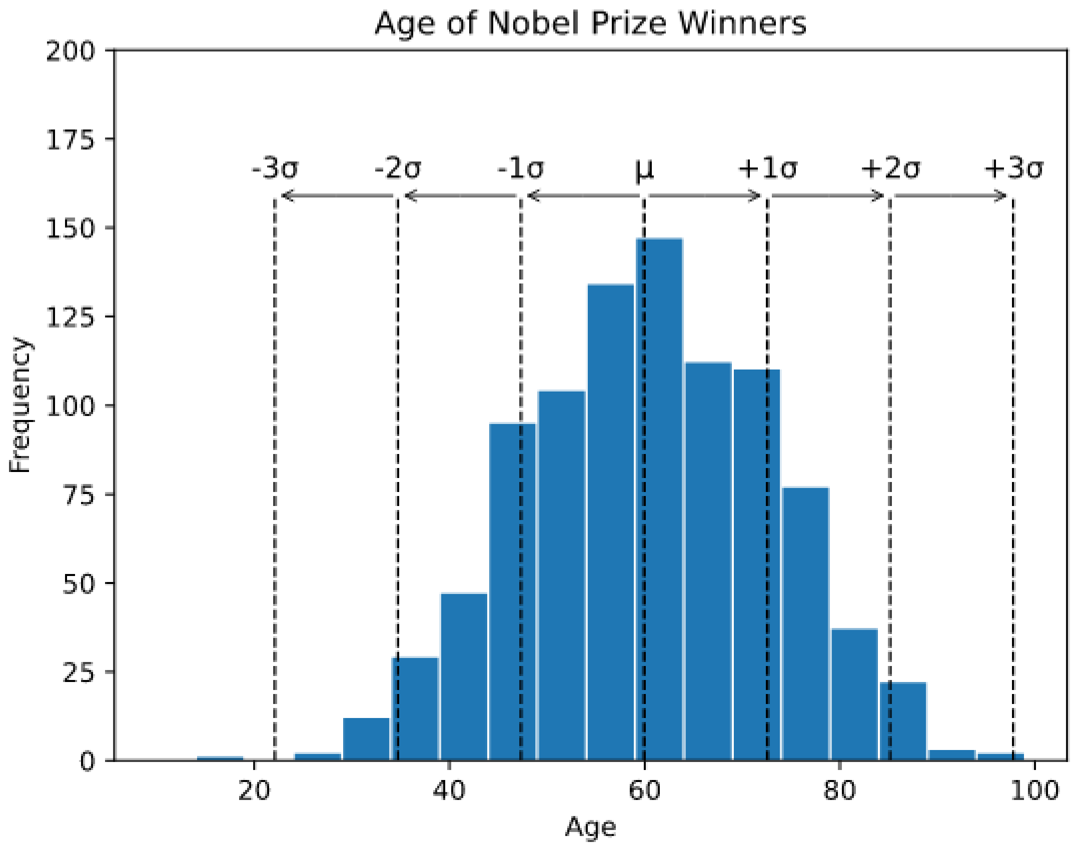
- Example
- Use the R IQR() function to find the interquartile range of the values 13, 21, 21, 40, 42, 48, 55, 72:
- values <- c(13,21,21,40,42,48,55,72)

  IQR(values)

Standard Deviation

- Standard deviation is the most commonly used measure of variation, which describes how spread out the data is.
- Standard Deviation
  - Standard deviation (σ) measures how far a 'typical' observation is from the average of the data (μ).
  - Standard deviation is important for many statistical methods.

- Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing **standard deviations**:



Each dotted line in the histogram shows a shift of one extra standard deviation.
If the data is **normally distributed:**
Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Note: A **normal** distribution has a "bell" shape and spreads out equally on both sides.

Calculating the Standard Deviation

- We can calculate the standard deviation for both the population and the sample.
- The formulas are **almost** the same and uses different symbols to refer to the standard deviation (σ) and **sample** standard deviation (s).

- Calculating the **standard deviation** (σ) is done with this formula:

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

- Calculating the **sample standard deviation** (s) is done with this formula:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

- n is the total number of observations.
- $\sum$ is the symbol for adding together a list of numbers.
- xi is the list of values in the data: $x1, x2, x3$,...
- μ is the population mean and x̄ is the sample mean (average value).
- $(xi - \mu)$ and $(xi - \bar{x})$ are the differences between the values of the observations (xi) and the mean.
- Each difference is squared and added together.
- Then the sum is divided by n or $(n - 1)$ and then we find the square root.

- Using these 4 example values for calculating the **population standard deviation**:
- 4, 11, 7, 14
- We must first find the mean:

$$\mu = \frac{\sum x_i}{n} = \frac{4 + 11 + 7 + 14}{4} = \frac{36}{4} = \underline{9}$$

Then we find the difference between each value and the mean $(x_i - \mu)$:

- $4 - 9 \quad = -5$
- $11 - 9 = 2$
- $7 - 9 \quad = -2$
- $14 - 9 = 5$

Each value is then squared, or multiplied with itself $(x_i - \mu)^2$:

- $(-5)^2 = (-5)(-5) = 25$
- $2^2 \qquad = 2 * 2 \qquad = 4$
- $(-2)^2 = (-2)(-2) = 4$
- $5^2 \qquad = 5 * 5 \qquad = 25$

All of the squared differences are then added together $\sum(x_i - \mu)^2$:

$$25 + 4 + 4 + 25 = 58$$

Then the sum is divided by the total number of observations, n:

$$\frac{58}{4} = 14.5$$

Finally, we take the square root of this number:

$$\sqrt{14.5} \approx \underline{3.81}$$

So, the standard deviation of the example values is roughly: 3.81

Calculating the Standard Deviation with Programming

- Population Standard Deviation
-
  Example
- With Python use the NumPy library std() method to find the standard deviation of the values 4,11,7,14:
- import numpy

  values = [4,11,7,14]

  x = numpy.std(values)

  print(x)

- Example
- Use an R formula to find the standard deviation of the values 4,11,7,14:
- values <- c(4,7,11,14)

  sqrt(mean((values-mean(values))^2))

- Sample Standard Deviation
- Example
- With Python use the NumPy library std() method to find the **sample** standard deviation of the values 4,11,7,14:
- import numpy

  values = [4,11,7,14]

  x = numpy.std(values, ddof=1)

  print(x)

- Example
- Use the R sd() function to find
  the **sample** standard deviation of the values
  4,11,7,14:
- values <- c(4,7,11,14)

  sd(values)