Measures of Location: Second Moment

[John S Butler (TU Dublin)](https://johnsbutler.netlify.com)

# Introduction

A different aspects of a distribution of data can be summarised by the measures of location:

1. The First Moment: Middle.
2. **The Second Moment: Spread.**
3. The Third Moment: Symmetry.

All that being said, I would always recommend plotting the data first before anything else.

**A picture (histogram) is worth a thousand words.**

# Second Moment: Spread

## Variance

### Definition:

The variance is the spread of the data around the mean . The formula is given by

where is each element, is the average of the elements and is the number of elements.

### Example

Given the list of 7 ages at a concert {19,18,20,18,18,18,20}, the mean is :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 20 | 18 | 19 | 18 | 18 | 18 | 20 |
|  | 20-18.71429 | 18-18.71429 | 19-18.71429 | 18-18.71429 | 18-18.71429 | 18-18.71429 | 20-18.71429 |
|  | 1.65306122 | 0.51020408 | 0.08163265 | 0.51020408 | 1.65306122 | 0.51020408 | 1.65306122 |

Or you could just use the command Var().

In R code:

Age=c(20, 18,19, 18,18,18,20) # List of 7 numbers  
## Cacluate the variance in the long form  
Age-mean(Age) # Each element minus the mean

## [1] 1.2857143 -0.7142857 0.2857143 -0.7142857 -0.7142857 -0.7142857  
## [7] 1.2857143

(Age-mean(Age))^2# Each element minus the mean

## [1] 1.65306122 0.51020408 0.08163265 0.51020408 0.51020408 0.51020408  
## [7] 1.65306122

sum((Age-mean(Age))^2)/6 #Sum the elements and divide by (7-1)=6

## [1] 0.9047619

# Or just use the function  
var(Age)

## [1] 0.9047619

## Standard Deviation

### Definition:

The standard deviation , is the square root of the variance.

where is each element, is the average of the elements and is the number of elements. In R-code

Age=c(20, 18,19, 18,18,18,20) # List of 7 numbers  
sd(Age) # Each element minus the mean

## [1] 0.9511897

### Standard Deviation and Variance Pros and Cons

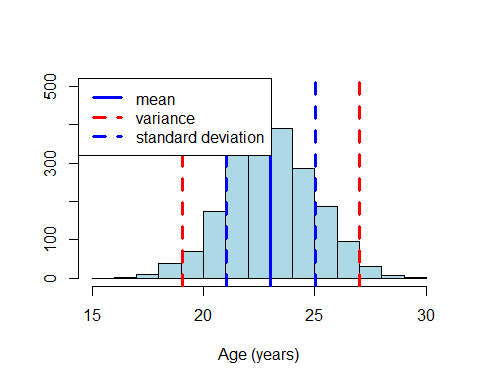
Pros of the standard deviation and variance;

* Takes all data into account.
* Lends itself to computation of other stable measures (and is a prerequisite for many of them).

Cons of the mean;

* Hard to interpret.
* Can be influenced by extreme scores.

### Graphical representation of the variance/standard deviation:

The figure below shows the histogram from 2000 Age observations at a concert: 

## Range

### Definition:

The range is the difference between the smallest and largest observations.

### Example

Given the list of 7 ages at a concert {19,18,20,18,18,18,20}.

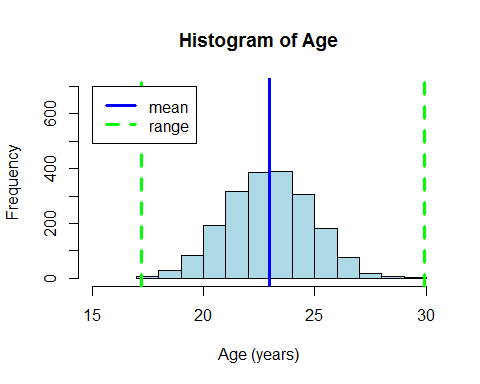
1. First the list has to be ordered {18,18,18,18,19,20,20}.
2. Then the range is 20-18=2  
   The range is 2.

In R code:

Age<-rnorm(2000,23,2)  
range(Age)

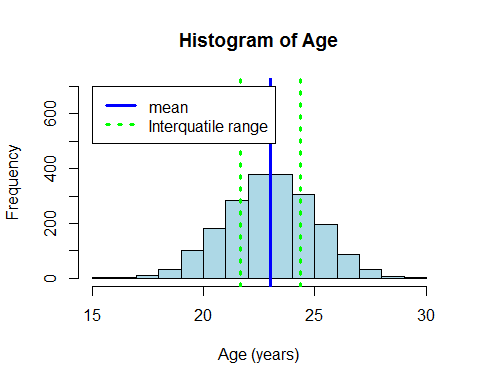
## [1] 17.21867 29.94106

### Graphical representation of the range:



## Interquartile Range

The interquartile range brackets 50% of the observations.



### Range Pros and Cons

Pros of the range and interquartile range:

Pros;

* Fairly easy to compute.
* Scores exist in the data set.
* Eliminates influence of extreme scores.

Cons;

* Discards much of the data.

## Mean and Standard Deviation together

As both the first and second moments give different information it makes sense to use them together, the most commmon combination is the mean and standard deviation. The mean and standard deviation are an efficient way to describe a distribution with just two numbers. It also allows a direct comparison between distributions that are on different scales.

### Coefficient of Variation

The Coefficient of Variation uses both the mean and standard to describe the distribution

Pros;

* The Coefficient of Variation is unitless and therefore can be used to compare across different variables.

Cons;

* The Coefficient of Variation does not have a direct meaning to the original data.

## All in one plot

The figure below shows a histogram of 3000 attendees age observations at a Wiggles concert: 