STATISTICS FOR DATA SCIENCE

Statistics Types and Summary

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Descriptive Statistics

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Measures of Dispersion: Quartiles

As the name '**quartile**' suggests, we want to divide the data into **four equal parts**

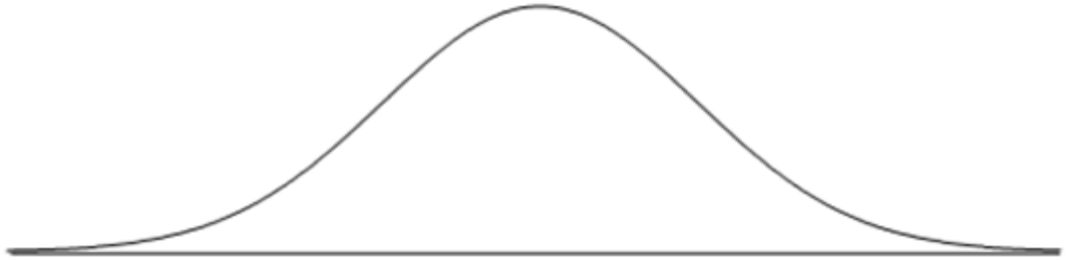


Figure : Graph representing heights of adult males.

The first quartile is the 25th percentile

The median is the 50th percentile

The third quartile is the 75th percenile

In the above graph, we want to divide the area under our curve into four equal areas.

First put the list of numbers in order

Then cut the list into four equal parts

The **Quartiles** are at the "**cuts**"

Measures of Dispersion: Quartiles

The first quartile is the 25^{th} percentile

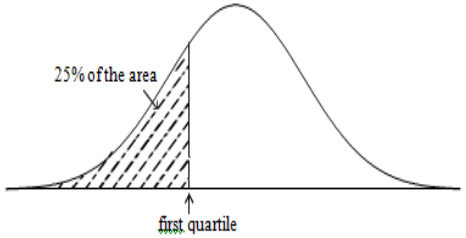
The median is the 50^{th} percentile

The third quartile is the 75^{th} percenile

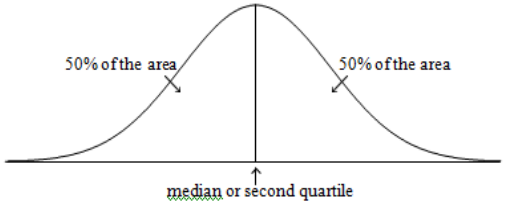
STATISTICS FOR DATA SCIENCE

Measures of Dispersion: Quartiles

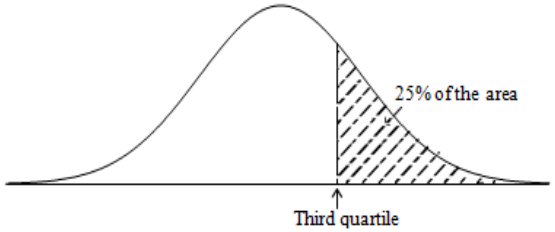
The first quartile is the 25th percentile



The median is the 50th percentile



The third quartile is the 75th percentile

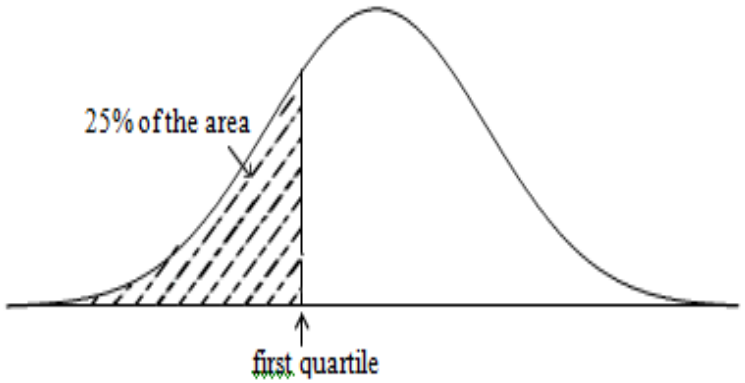


The first quartile

The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it.

This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile.

The first quartile is also called the 25th percentile.



Measures of Dispersion: Quartiles

To find the first quartile, compute the value 0.25(n +1).

If this is an integer, then the sample value in that position is the first quartile.

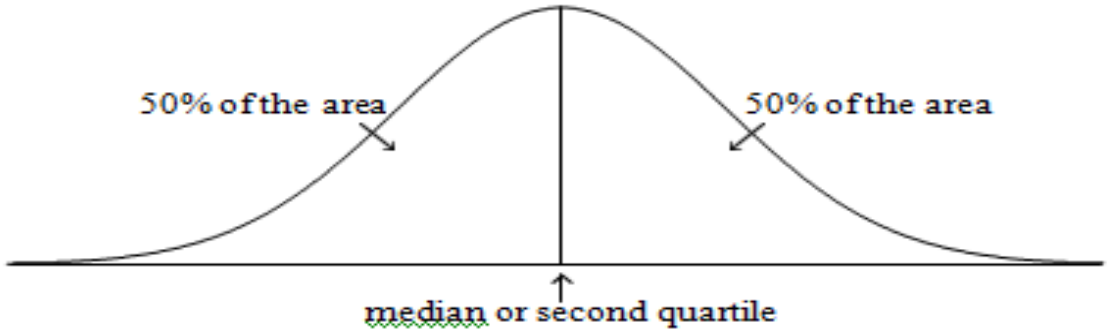
If not, then take the average of the sample values on either side of this value.

The second quartile or median

It is easy to see how to divide the area in Figure 9 into two equal parts, since the graph is symmetric.

The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median

Second quartile is calculated using the value 0.5(n+1)



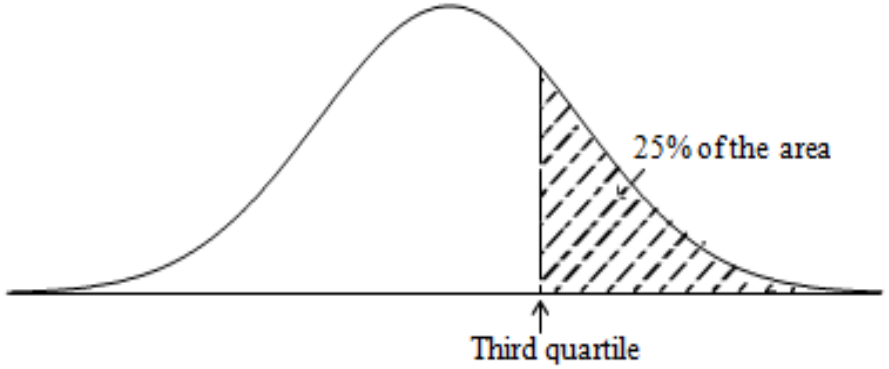
The third quartile

The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it.

This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile.

The third quartile is also called the 75th percen

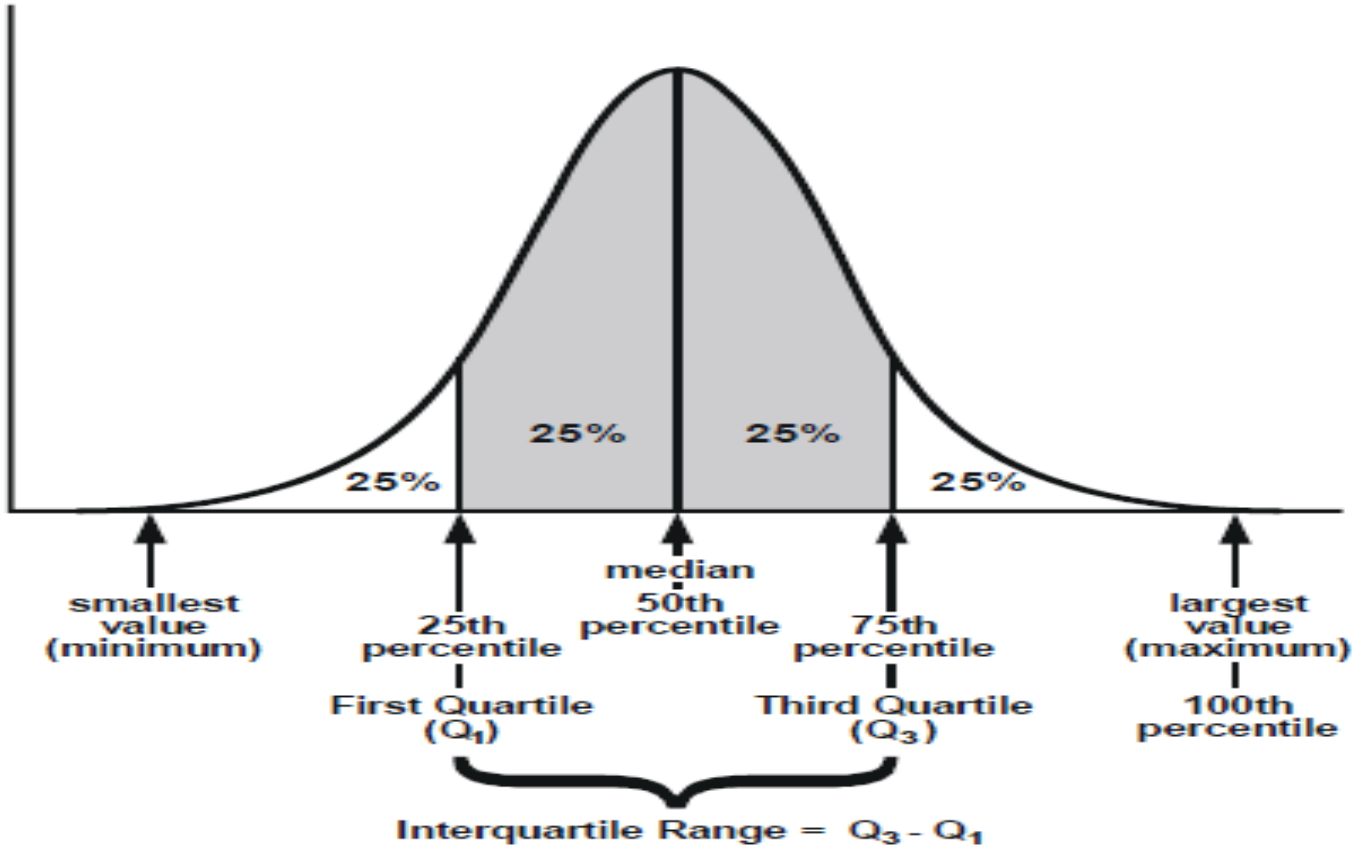
The third quartile is computed in the same way the value 0.75(n+1) is used.



STATISTICS FOR DATA SCIENCE

Measures of Dispersion: Quartiles Summary

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts.



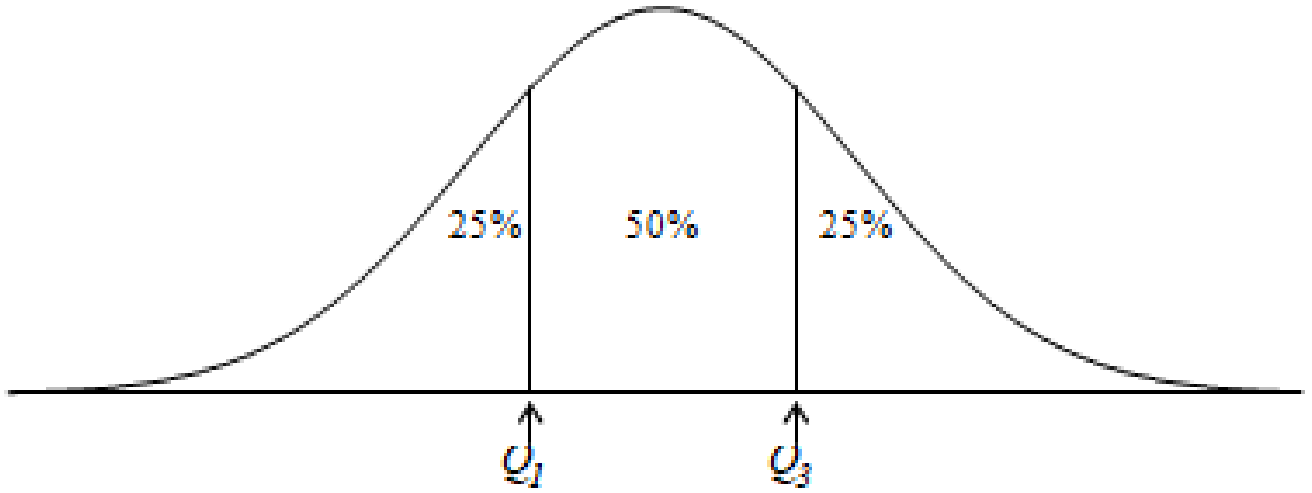
Source: https://medium.com/analytics-vidhya/descriptive-statistics-in-data-science-with-illustrations-in-python-efd5ccc152f1

Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1

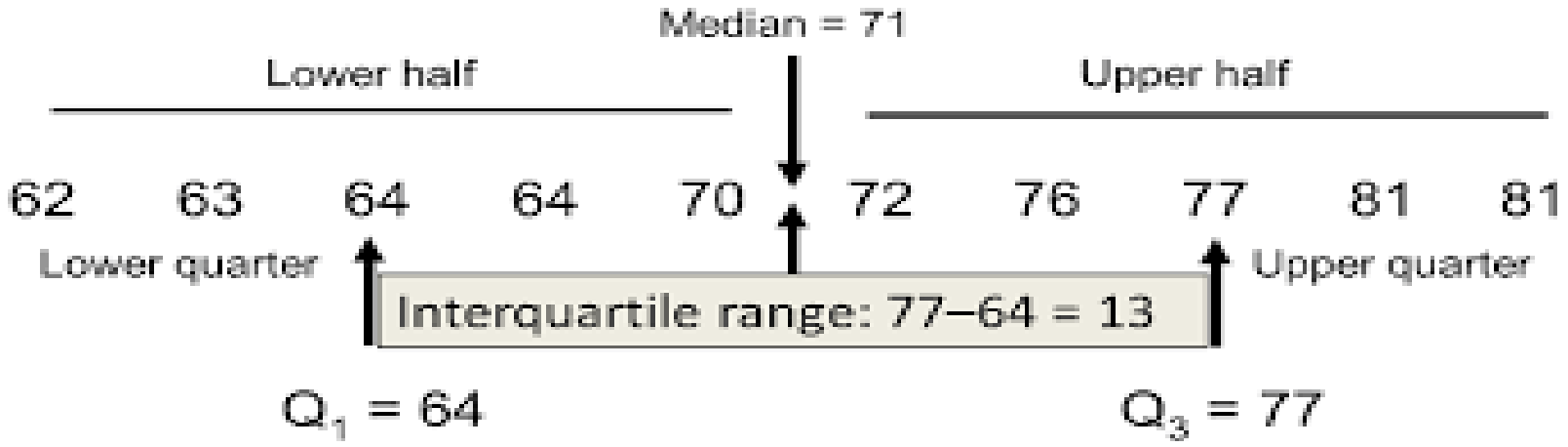
The interquartile range quantifies the difference between the third and first quartiles.



Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1



Source: sphweb.bumc.bu.edu/

Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1

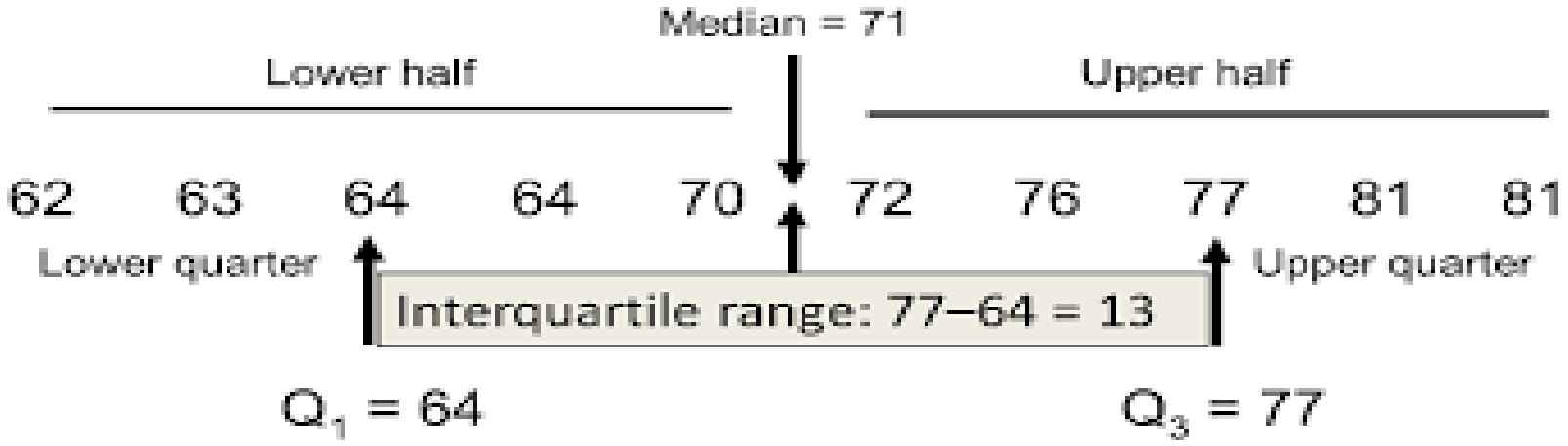
Steps to find IQR :

1. Arrange the data scores in ascending order.
2. Find the median of the data set(the number in the middle).
3. Find the median of the lower half of the scores (Q1).
4. Find the median of the upper half of the scores (Q3).

Note: If the number of scores is even, the median is the average of the two middle scores.

STATISTICS FOR DATA SCIENCE

Measures of Spread: IQR-Example



Source: sphweb.bumc.bu.edu/

Measures of Spread: IQR-Example

For the following data sets, calculate the quartiles and find the interquartile range.

The following numbers represent the time in minutes that twelve employees took to get to work on a particular day.

18 34 68 22 10 92 46 52 38 29 45 37

STATISTICS FOR DATA SCIENCE

Measures of spread: Variance

Average of the distance that each score is from the mean (Squared deviation from the mean).

1. Find the mean value of the given data values.
2. Subtract mean from each data value.
3. Square each value that is obtained from step2.
4. Find the sum of all values that is obtained from step 3.
5. Divide the result that is obtained from step4 by N(for population) and n-1(for sample).

Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

54

Source: standard-deviation-calculator.com

Measures of spread: Variance And Standard Deviation

Consider these two list of numbers:-
28,29,30,31,32 and
10,20,30,40,50. Find their means.

What did u found ??

Both the list have same mean i.e. 30
But,
Clearly list differs which is not captured by mean

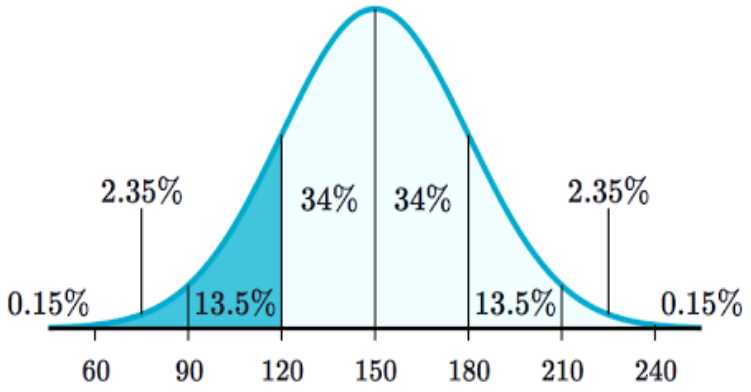
Source: standard-deviation-calculator.com

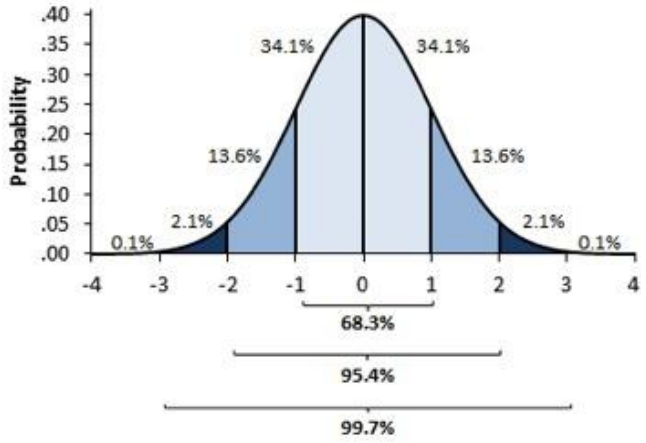
STATISTICS FOR DATA SCIENCE

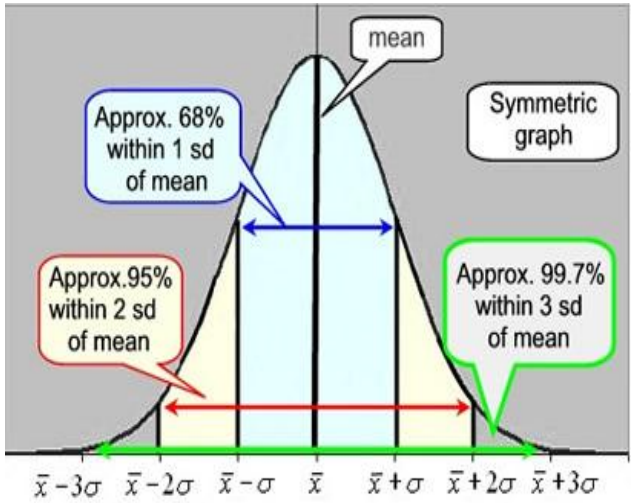
Measures of spread: Standard Deviation

Standard deviation signifies the deviation of the terms from the mean value of the distribution.

It quantifies the amount of variation of a set of data values.







Source: exceluser.com, MathBitsNotebook

STATISTICS FOR DATA SCIENCE

Measures of spread: Standard Deviation

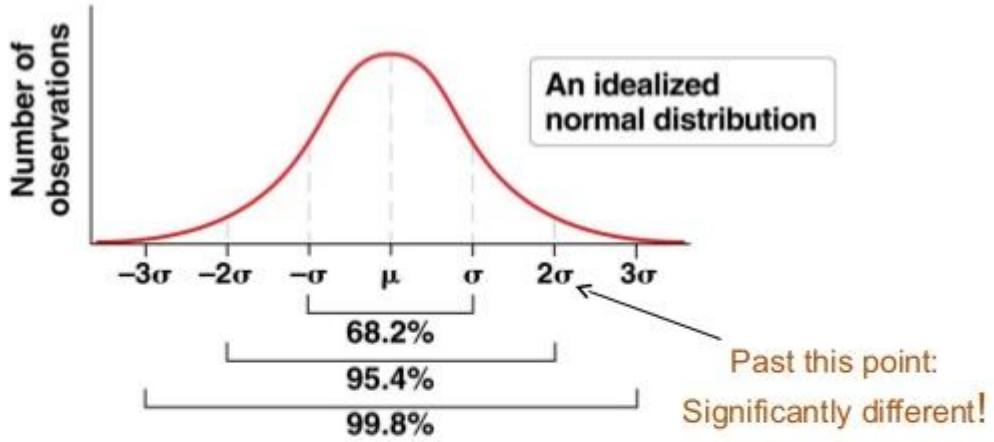
Standard deviation signifies the deviation of the terms from the mean value of the distribution.

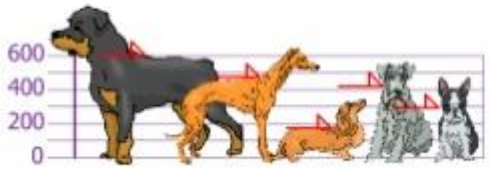
It quantifies the amount of variation of a set of data values.

Source: exceluser.com, MathBitsNotebook

STATISTICS FOR DATA SCIENCE

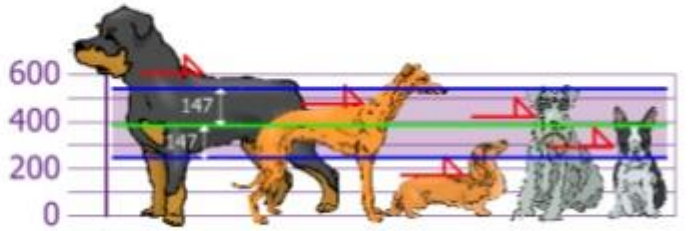
Measures of spread: Standard Deviation





Measure the height of dogs

www.mathisfun.com



Standard Deviation (purple):

Outside of that area you know you have an extra tall/small dog

Source: standard-deviation-calculator.com

Measures of spread: Variance and Standard Deviation

Standard Deviation = Square root of Variance

Example

Find the standard deviation and variance

x	$x - \bar{x}$	$(x - \bar{x})^2$
30	4	16
26	0	0
22	-4	16
78	Mean = 26	32

Sum = 0

The variance

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = 32 / 2 = 16$$

The standard deviation

$$S = \sqrt{16} = 4$$

Source: standard-deviation-calculator.com

Measures of spread: Standard Deviation

Problem:

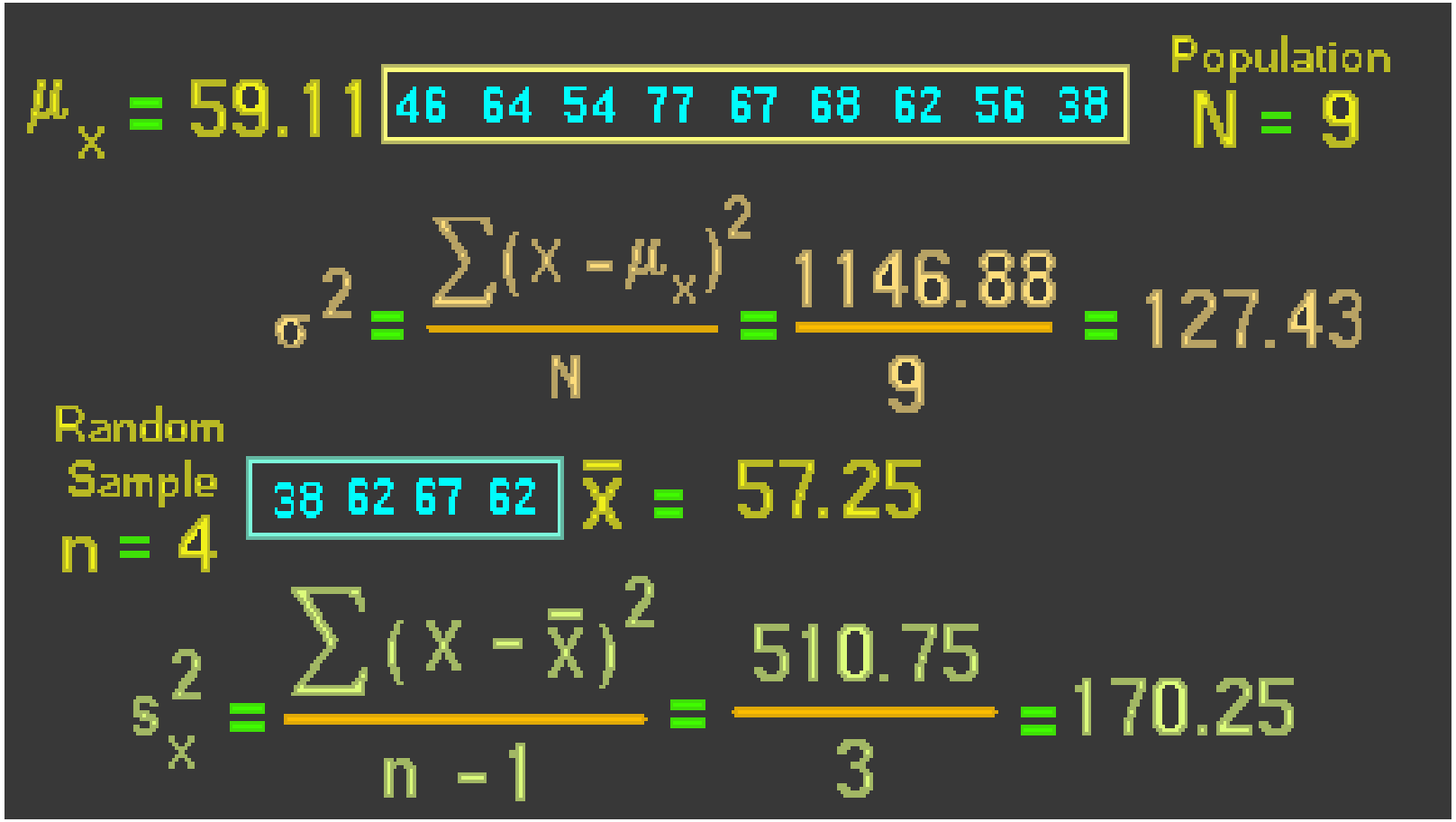
The heights of the players (in centimeters) from a basketball team are represented by the table:

Height	[170, 175)	[175, 180)	[180, 185)	[185, 190)	[190, 195)	[195, 2.00)
No. of players	1	3	4	8	5	2

Calculate standard deviation.

STATISTICS FOR DATA SCIENCE

Measures of spread : Variance - Example



Source: Chilimath

STATISTICS FOR DATA SCIENCE

Trimmed Mean

The trimmed mean is computed by arranging the sample values in order, “trimming” an equal number of them from each end, and computing the mean of those remaining.

If p% of the data are trimmed from each end, the resulting trimmed mean is called the “p% trimmed mean.”

There are no hard-and-fast rules on how many values to trim.

The most commonly used trimmed means are the 5%, 10%, and 20% trimmed means.

Source: exceluser.com, MathBitsNotebook

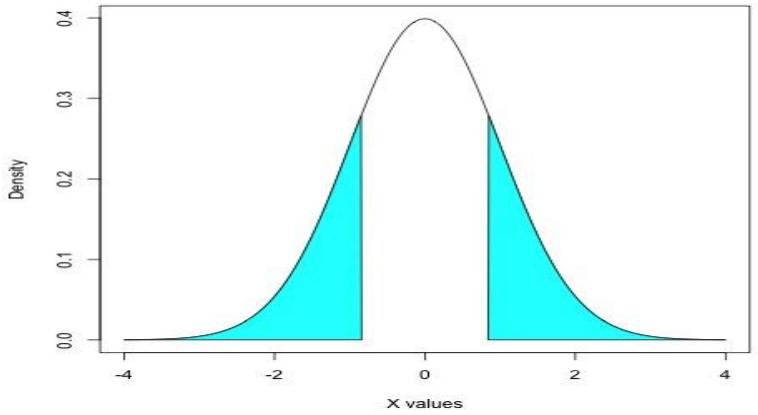
STATISTICS FOR DATA SCIENCE

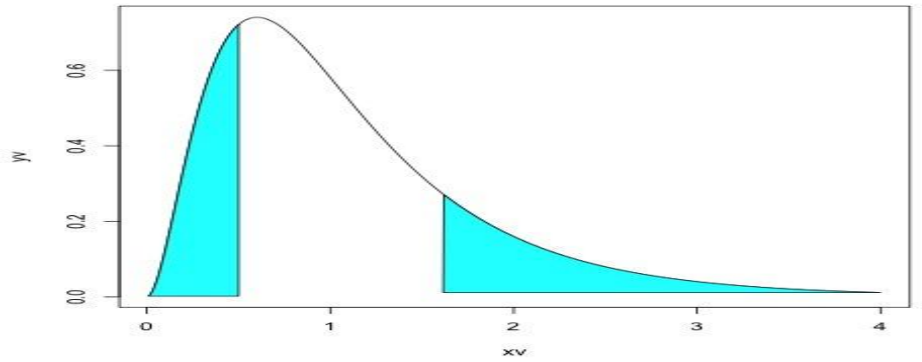
Trimmed Mean

If the sample size is denoted by n, and a p% trimmed mean is desired, the number of data points to be trimmed is np/100

It is used to reduce the effects of outliers on the calculated average.

This method is best suited for data with large, erratic deviations or extremely skewed distributions





Source: exceluser.com, MathBitsNotebook

STATISTICS FOR DATA SCIENCE

Trimmed Mean

For the following data

30 75 79 80 80 105 126 138 149 179 179 191

223 232 232 236 240 242 245 247 254 274 384 470

Compute the mean, median, and the 5%, 10%, and 20% trimmed means.

Solution:-

……

Source:

STATISTICS FOR DATA SCIENCE

Trimmed Mean

The mean is found by averaging together all 24 numbers, which produces a value of 195.42.

The median is the average of the 12th and 13th numbers, which is (191 + 223)/2 = 207.00.

To compute the 5% trimmed mean, we must drop 5% of the data from each end. This comes to (0.05)(24) = 1.2 observations.

We round 1.2 to 1, and trim one observation off each end.

Source:

STATISTICS FOR DATA SCIENCE

Trimmed Mean

The 5% trimmed mean is the average of the remaining 22 numbers:
75 + 79 +⋯+ 274 + 384/22= 190.45

To compute the 10% trimmed mean, round off (0.1)(24) = 2.4 to 2.

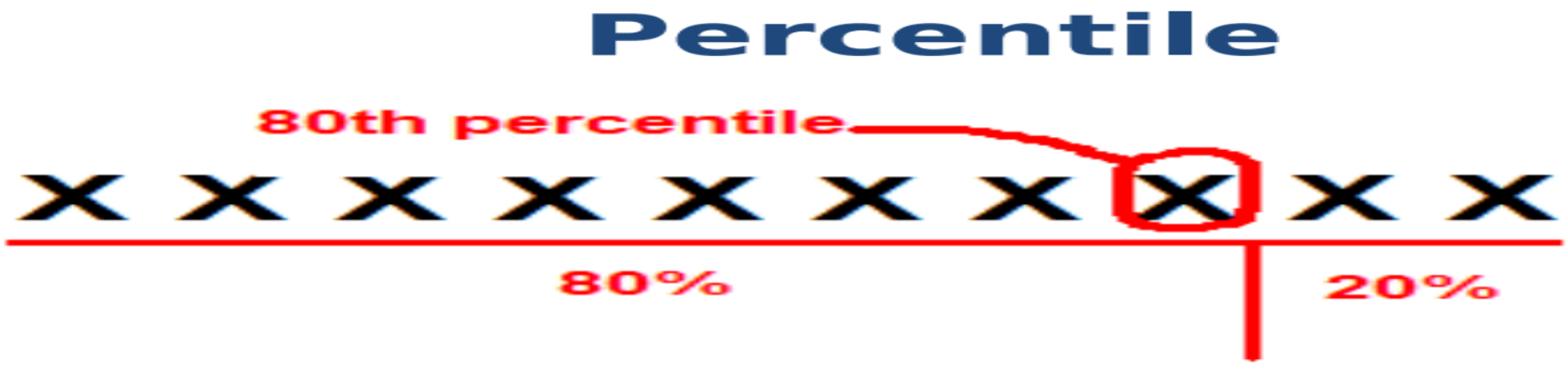
Drop 2 observations from each end, and then average the remaining 20:
79 + 80 +⋯+ 254 + 274/20= 186.55

To compute the 20% trimmed mean, round off (0.2)(24) = 4.8 to 5.
Drop 5 observations from each end, and then average the remaining 14:

105 + 126 +⋯+ 242 + 245/14= 194.07

Source:

Percentile



The pth percentile of a sample, for a number p between 0 and 100, divides the sample such that

1. p% of the sample values are less than the pth percentile

2. And (100-p%) are greater.

Source:

Percentile

Steps:-

Order the n samples values from smallest to largest

Compute the quantity (p/100)(n+1), where n is the sample size.

If this quantity is an integer, the sample value in this position is the percentile.

Otherwise, average the two sample values pn either side.

STATISTICS FOR DATA SCIENCE

Tertile and Decile

Steps:-

Order the n samples values from smallest to largest

Compute the quantity (p/100)(n+1), where n is the sample size.

If this quantity is an integer, the sample value in this position is the percentile.

Otherwise, average the two sample values pn either side.

Source:

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering