

# Descriptive Statistics

Dr Amin Karami

[a.karami@uel.ac.uk](mailto:a.karami@uel.ac.uk)

[www.aminkarami.com](http://www.aminkarami.com)

CN5209 – Week 3  
14 October 2019

# Outline

- Quantitative and Qualitative Analysis
- Descriptive Statistics
  - ✓ Numerical representations (today)
  - ✓ Graphical representations (next session)
- Several MATLAB examples

# Learning Outcomes

- Understand what descriptive statistics are
- Understand different types of analysis over raw data
- Be able to describe and differ descriptive statistical methods
- A hands-on experience on using descriptive statistics on data

# Quality vs Quantity

- **Qualitative Observations:**

- Use our senses to observe the results (sight, smell, touch, taste and hear) based on characteristics and **quality**.
- Such as colour, size, texture.
- Example: The school is hot, the shirt is blue.

- **Quantitative Observations:**

- observations are made with instruments such as rulers, balances, beakers, and thermometers. These results are measurable and include **numbers (quantity)**.
- Such as weight, height, length.
- Example: There is one shirt. It is 91 degrees.



# Discrete & Continuous Data

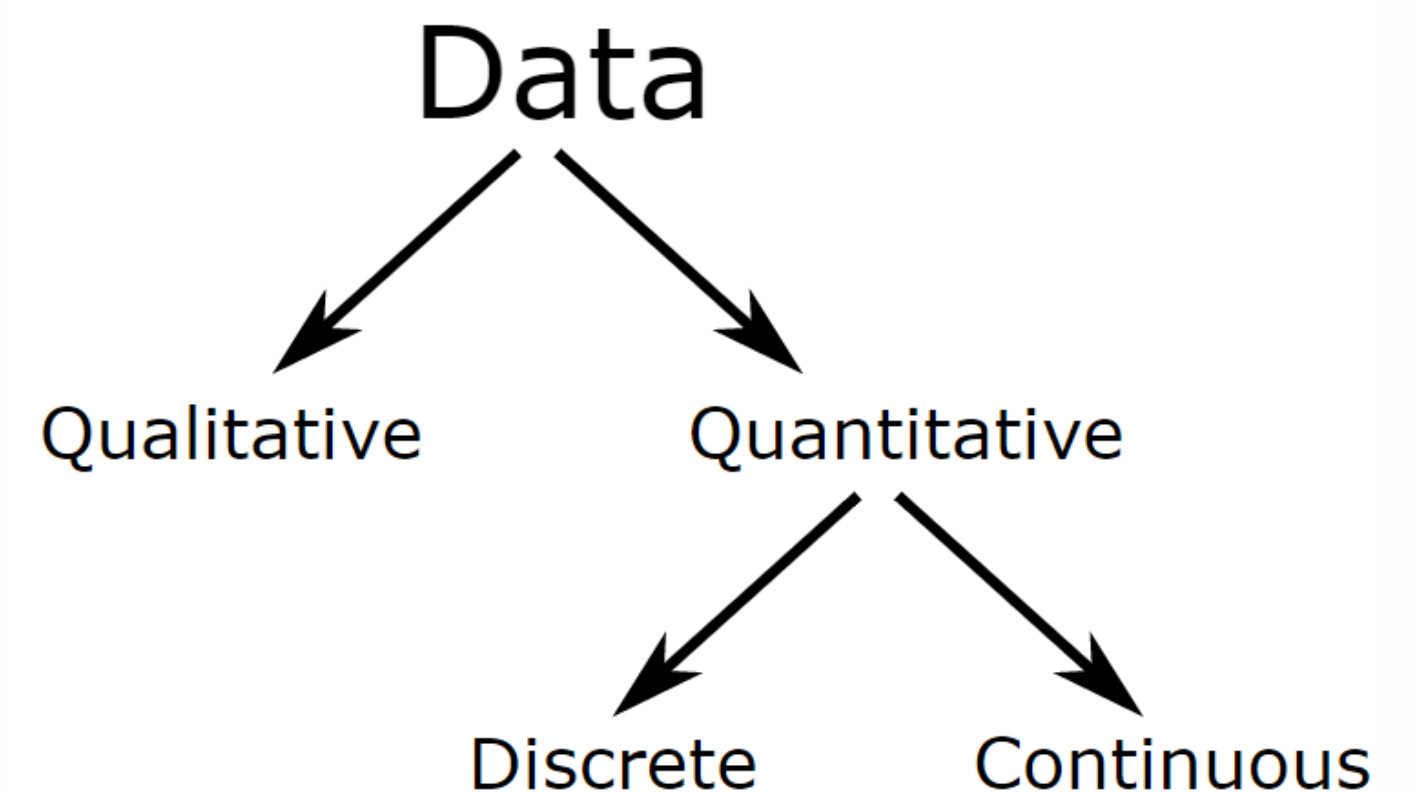
**Discrete data** have finite values with certain values. Can count them.

**Sample:**

*Number of children in a household*

*Number of languages a person speaks*

*Number of people sleeping in stats class*



**Continuous data** have an infinite number of steps, which form a continuum. They come from measurements and can take any value within a given range.

**Sample:**

*Height of children*

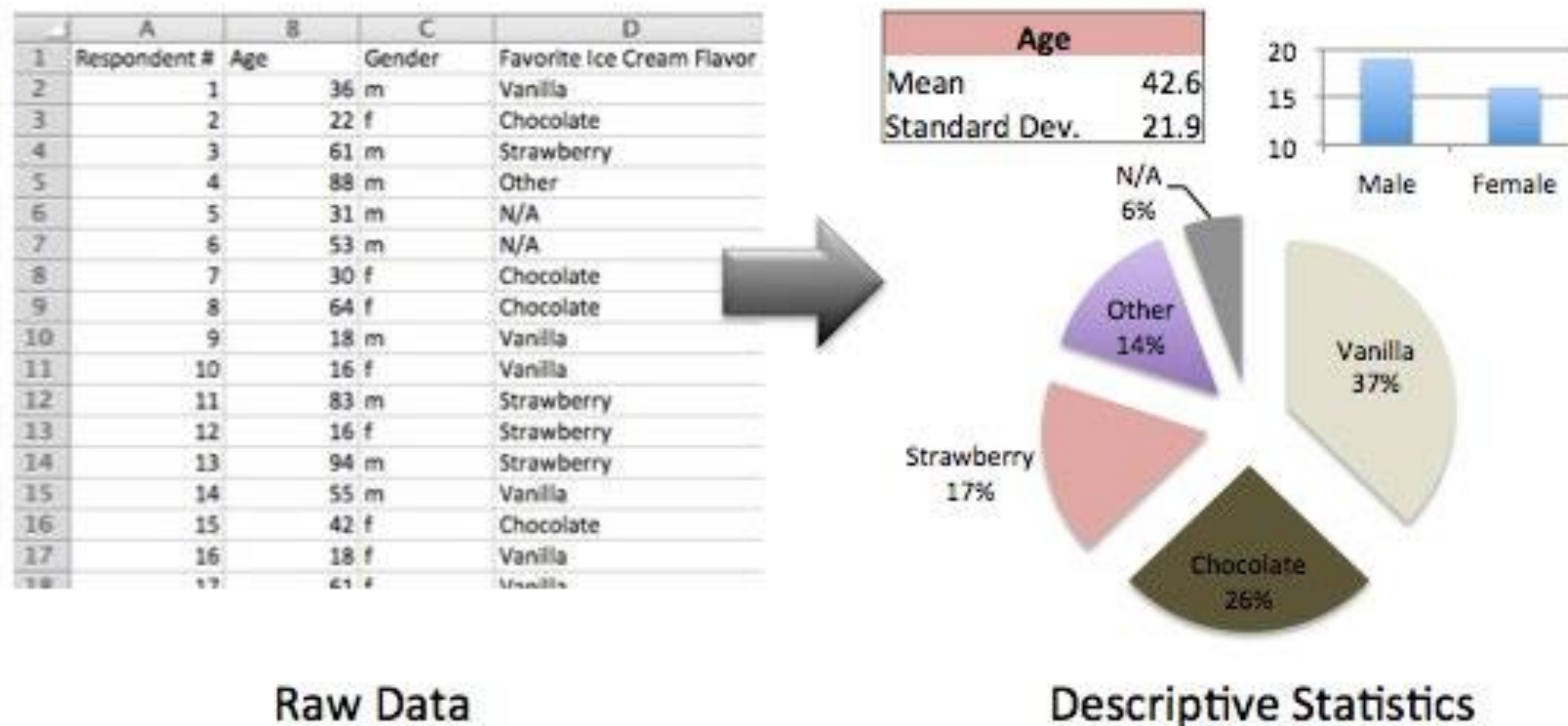
*Time to wake up in the morning*

*Speed of the train*





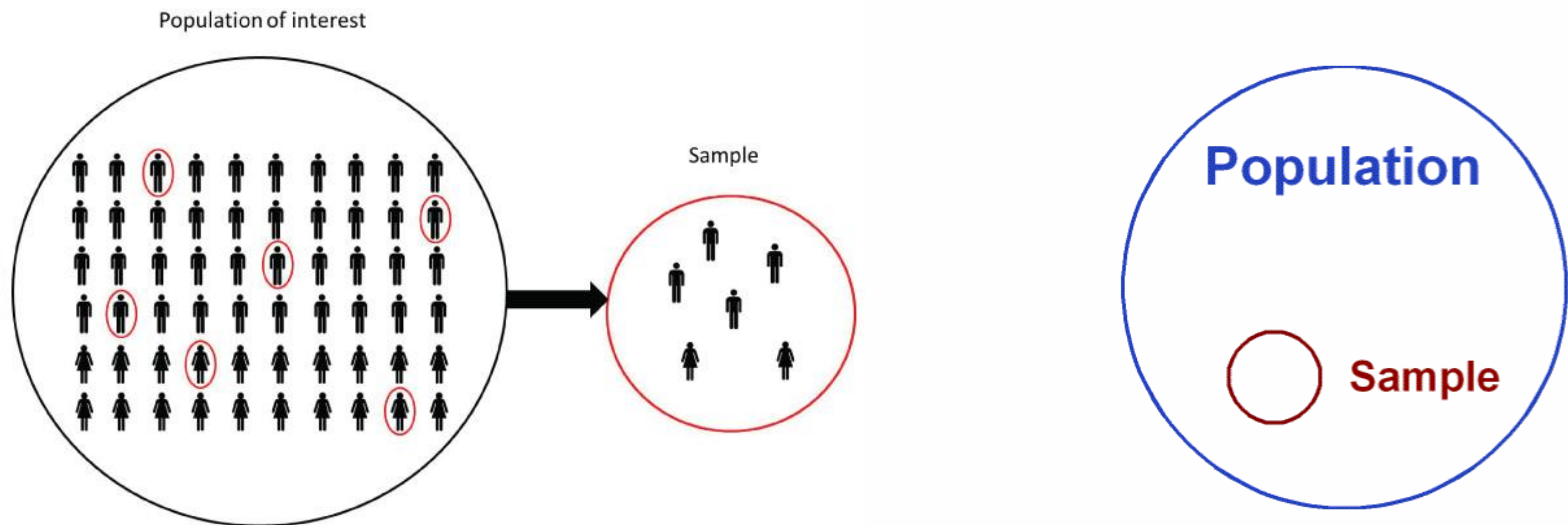
# What are Descriptive Statistics?



- Descriptive Statistics are used to present quantitative descriptions in a manageable form.
- They help us to simplify large amounts of data in a sensible way.
- Each descriptive statistic reduces lots of data into a simpler summary.

# Definitions

- The measures are computed for data from a sample, they are called **sample statistics**.
- If the measures are computed for data from a population, they are called **population parameters**.



# Descriptive Statistics

## **1. Numerical Measures for data representation (Measures of Location)**

- Mean
- Median
- Mode
- Percentiles
- Quartiles







# A Central Value: Mean

- The Mean is a measure of *central value* (average)
  - Sum of a set of numbers divided by the number of numbers in the set.

$$\frac{1+2+5+4+11+4+7+8+12+10}{10} = \frac{64}{10} = 6.4 \qquad \text{mean} = \frac{\sum x}{n}$$

```
>> A = [1 2 5 4 11 4 7 8 12 10];  
>> mean(A)
```

```
ans =  
  
6.4000
```

Name ^	Value
 A	[1,2,5,4,11,4,7,8,12,10]
 ans	6.4000



# A Central Value: Median

- Middlemost or the most central item in the set of ordered numbers;  
**it separates the distribution into two equal halves**
- If *odd n*, middle value of sequence
  - if  $X = [1, 2, 4, 6, 9, 10, 12, 14, 17]$
  - then **9** is the median
- If *even n*, average of 2 middle values
  - if  $X = [1, 2, 4, 6, 9, 10, 11, 12, 14, 17]$
  - then **9.5** is the median; i.e.,  $(9+10)/2$

```
>> median(A)
```

```
ans =
```

```
6
```



# A Central Value: Mode

- It is the most frequently occurring number in a distribution
  - if  $X = [1, 2, 4, 7, 7, 7, 8, 10, 12, 14, 17]$
  - then 7 is the mode
- Easy to see in a simple frequency distribution

```
>> mode(A)
```

```
ans =
```

```
4
```

Command Window		Workspace	
>> B = [3 3 1 4; 0 0 1 1; 0 1 2 4]		Name^	Value
B =		B	3x4 double
		M	[3;0;0]
3 3 1 4			
0 0 1 1			
0 1 2 4			
>> M = mode(B, 2)			
M =			
3			
0			
0			



# Percentiles

- A percentile provides information about **how the data are spread** over the interval from the smallest value to the largest value.
- The  $p$ th percentile of a data set is a value such that at least  $p$  percent of the items take on this value or less.
- Example: Admission test scores for colleges and universities are frequently reported in terms of percentiles.

# Percentiles

▷ Arrange the data in ascending order.

▷ Compute index  $i$ , the position of the  $p$ th percentile.

$$i = (p/100)n$$

▷ If  $i$  is not an integer, round up. The  $p$ th percentile is the value in the  $i$ th position.

▷ If  $i$  is an integer, the  $p$ th percentile is the average of the values in positions  $i$  and  $i+1$ .





# Percentiles: example

■ Find 25<sup>th</sup> and 80<sup>th</sup> percentiles from the sample below:

$P = [1, 2, 5, 4, 11, 4, 7, 8, 12, 10]$

(1) Sort:  $P = [1 \quad 2 \quad 4 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 11 \quad 12]$

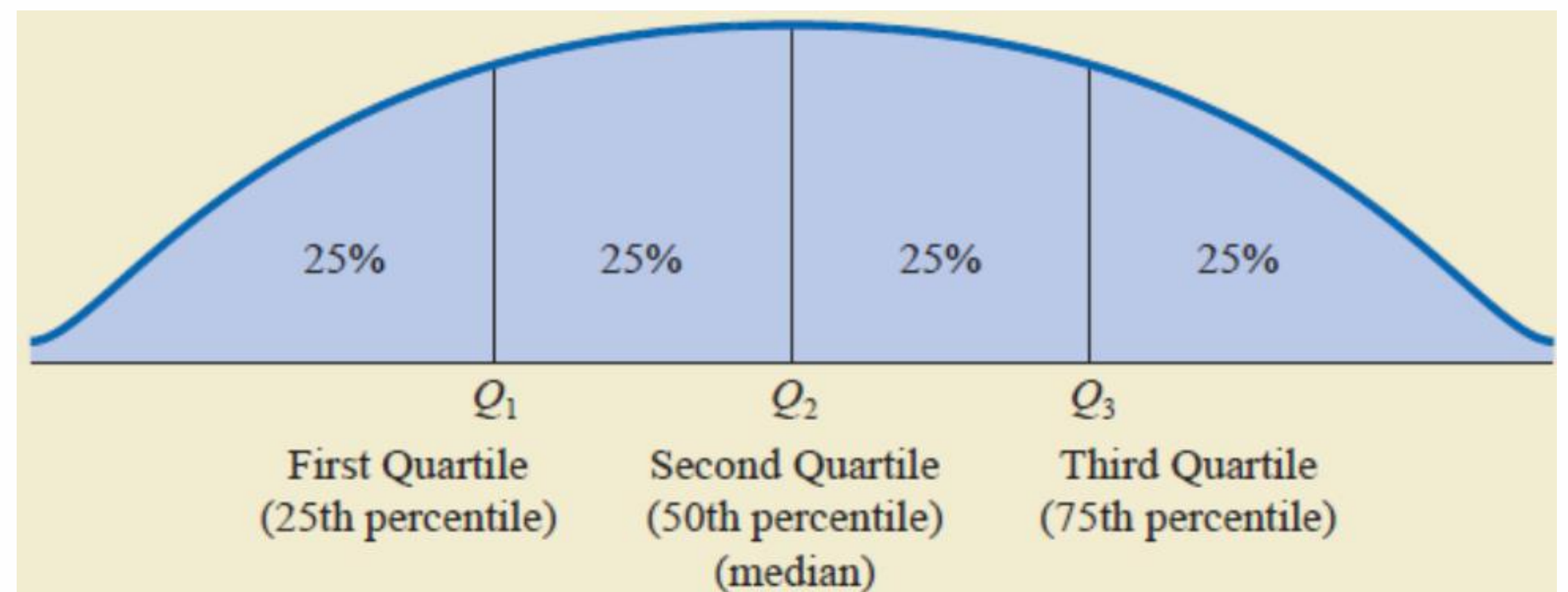
(2) 25<sup>th</sup> percentiles:  $(25/100) * 10 = 2.5 \approx 3$ ;  $P(3) = 4$ ; at least 25% of items take on a value of 4 or less.

(3) 80<sup>th</sup> percentiles:  $(80/100) * 10 = 8$ ;  $P(8) = 10.5$ ; at least 80% of items take on a value of 10.5 or less.

Command Window		Workspace	
		Name^	Value
<pre>&gt;&gt; P = [1, 2, 5, 4, 11, 4, 7, 8, 12, 10]; &gt;&gt; prctile(P,25)</pre>		ans	10.5000
<pre>ans =  4</pre>		P	[1,2,5,4,11,4,7,8,12,10]
<pre>&gt;&gt; prctile(P,80)</pre>			
<pre>ans =  10.5000</pre>			

# Quartiles

- Quartiles are just specific percentiles; thus, the steps for computing percentiles can be applied directly in the computation of quartiles.
- Quartiles are specific percentiles.
  - *First Quartile = 25th Percentile*
  - *Second Quartile = 50th Percentile = Median*
  - *Third Quartile = 75th Percentile*



# Descriptive Statistics

## 2. Measures of Variability

- Variation (or Summary of Differences Within Groups)
  - Range
  - Min/Max
  - Variance
  - Standard Deviation



# Range

- The simplest measure of variability is the **range**.

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

- Example: The largest starting salary is \$3925 and the smallest is \$3310. The range is  $3925 - 3310 = 615$

Graduate	Monthly Starting Salary (\$)	Graduate	Monthly Starting Salary (\$)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480



# Range Calculation in MATLAB

```
Command Window
>> Salary = [3450, 3550, 3650, 3480, 3355, 3310, 3490, 3730, 3540, 3925, 3520, 3480];
>> [max(Salary) min(Salary)]

ans =

    3925    3310

>> Range = max(Salary) - min (Salary)

Range =

    615
```

Name^	Value
ans	[3925,3310]
Range	615
Salary	1x12 double

- `max (A)`: Largest elements in array
- `min (A)`: Smallest elements in array
- **`max (A, [], dim)`**: the largest elements along dimension *dim*
- **`min (A, [], dim)`**: the smallest elements along dimension *dim*



# Min/Max Examples

Command Window

```
>> A = [2 8 4; 7 3 9; 4 5 1]
```

A =

2	8	4
7	3	9
4	5	1

```
>> Max = max(A)
```

Max =

7	8	9
---	---	---

```
>> Min = min(A)
```

Min =

2	3	1
---	---	---

```
>> max(max(A))
```

ans =

9

```
>> min(min(A))
```

ans =

1

Command Window

```
>> A
```

A =

2	8	4
7	3	9
4	5	1

```
>> Max = max(A, [], 2)
```

Max =

8
9
5

```
>> Min = min(A, [], 2)
```

Min =

2
3
1

# Min/Max Examples

## Max & Min Element Indices

Command Window		Workspace	
>> [Max MaxInd] = max(A)		Name ^	Value
Max =		A	[2,8,4;7,3,9;4,5,1]
7      8      9		Max	[7,8,9]
		MaxInd	[2,1,2]
		Min	[2,3,1]
		MinInd	[1,2,3]
MaxInd =			
2      1      2			
>> [Min MinInd] = min(A)			
Min =			
2      3      1			
MinInd =			
1      2      3			

## Largest and Smallest Element Comparison

>> A		Name ^	Value
A =		A	[2,8,4;7,3,9;4,5,1]
2      8      4		B	5
7      3      9		C_MAX	[5,8,5;7,5,9;5,5,5]
4      5      1		C_Min	[2,5,4;5,3,5;4,5,1]
>> B = 5;			
>> C_MAX = max(A,B)			
C_MAX =			
5      8      5			
7      5      9			
5      5      5			
>> C_Min = min(A,B)			
C_Min =			
2      5      4			
5      3      5			
4      5      1			



# Variance

- Variance is a measure of how spread out a data set is. In other words, they are measures of variability.
- The variance is computed as the average squared deviation of each number from its mean.

- A sample of a population:

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

- If the data are for a population, the average of the squared deviations is called the population variance.

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{n}$$



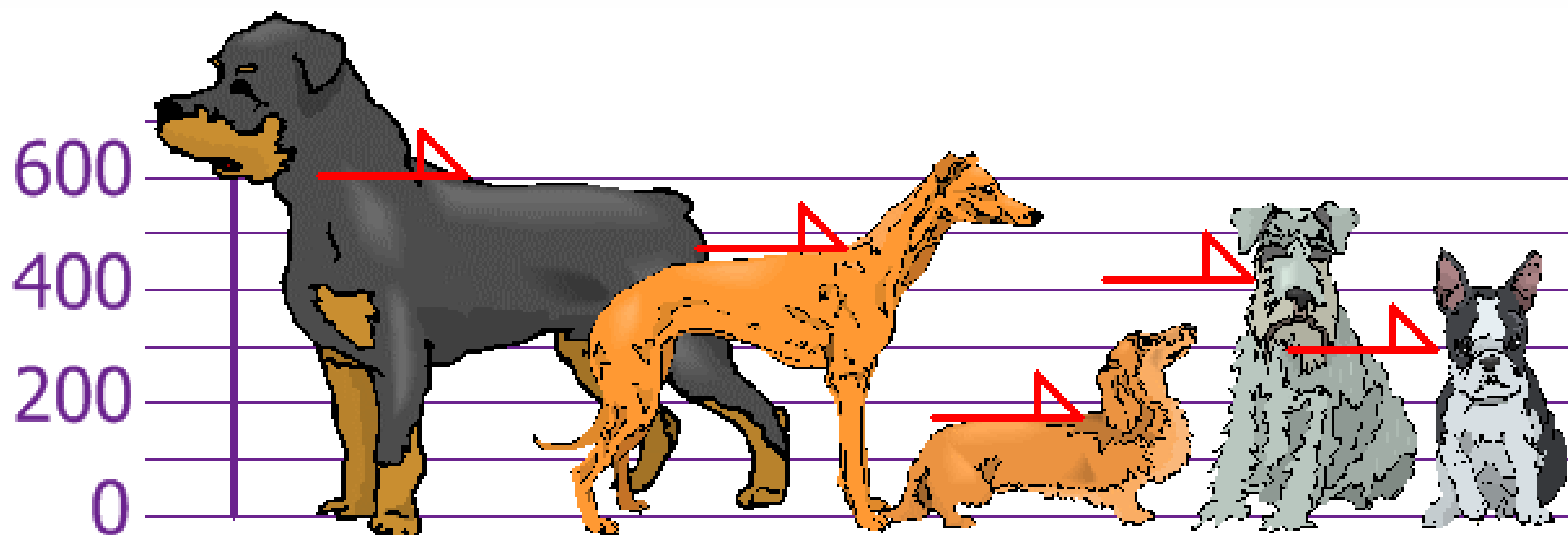
# Standard Deviation (Std.)

- Deviation just means how far from the normal.
- Standard Deviation (SD) is a measure of how spread out numbers are.
- The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread.

$$\sigma = \sqrt{\textit{Variance}}$$

# Example

- You and your friends have just measured the heights of your dogs (in millimetres).
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.



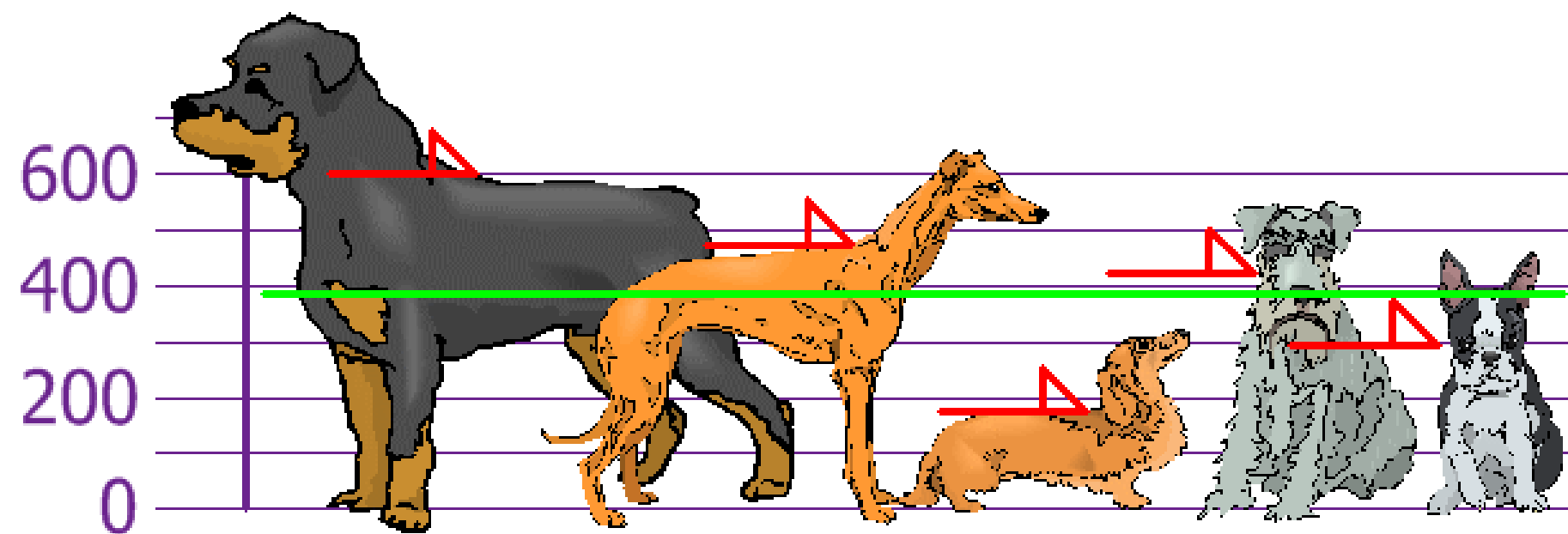
source: <https://www.mathsisfun.com/data/standard-deviation.html>



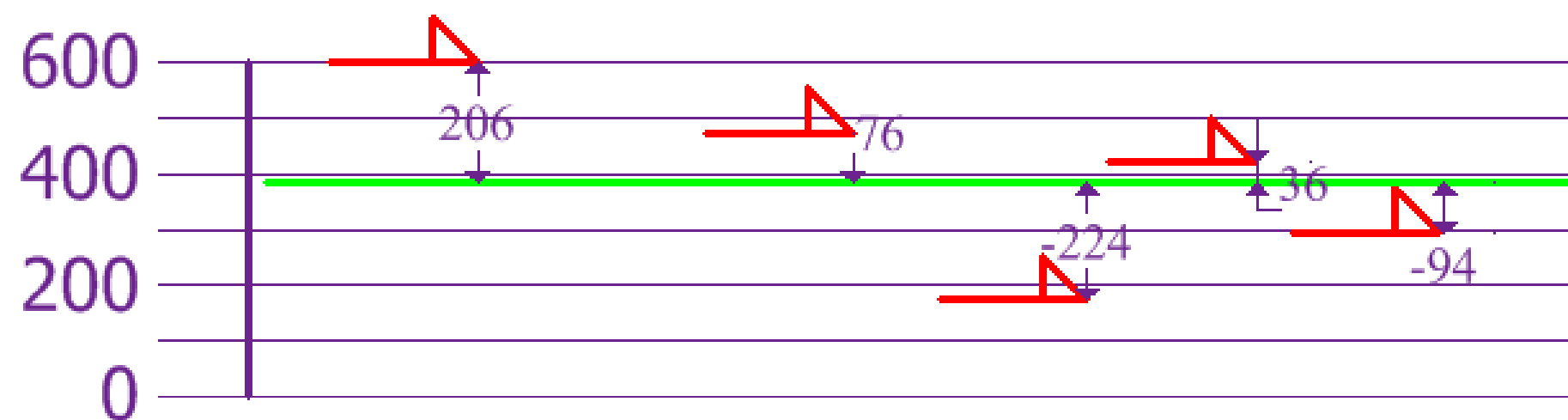
# Example

■ Your first step is to find the Mean:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = 394$$



Now we calculate each dog's difference from the Mean:

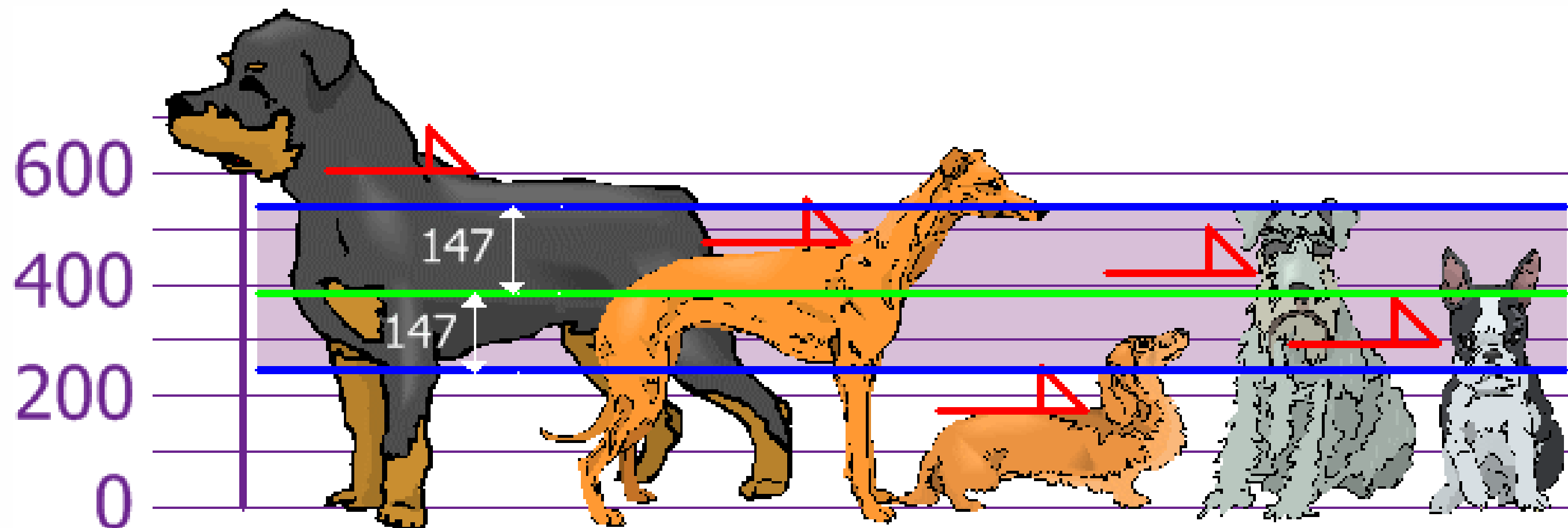


# Example

- Calculate Variance, which is 21,704 mm

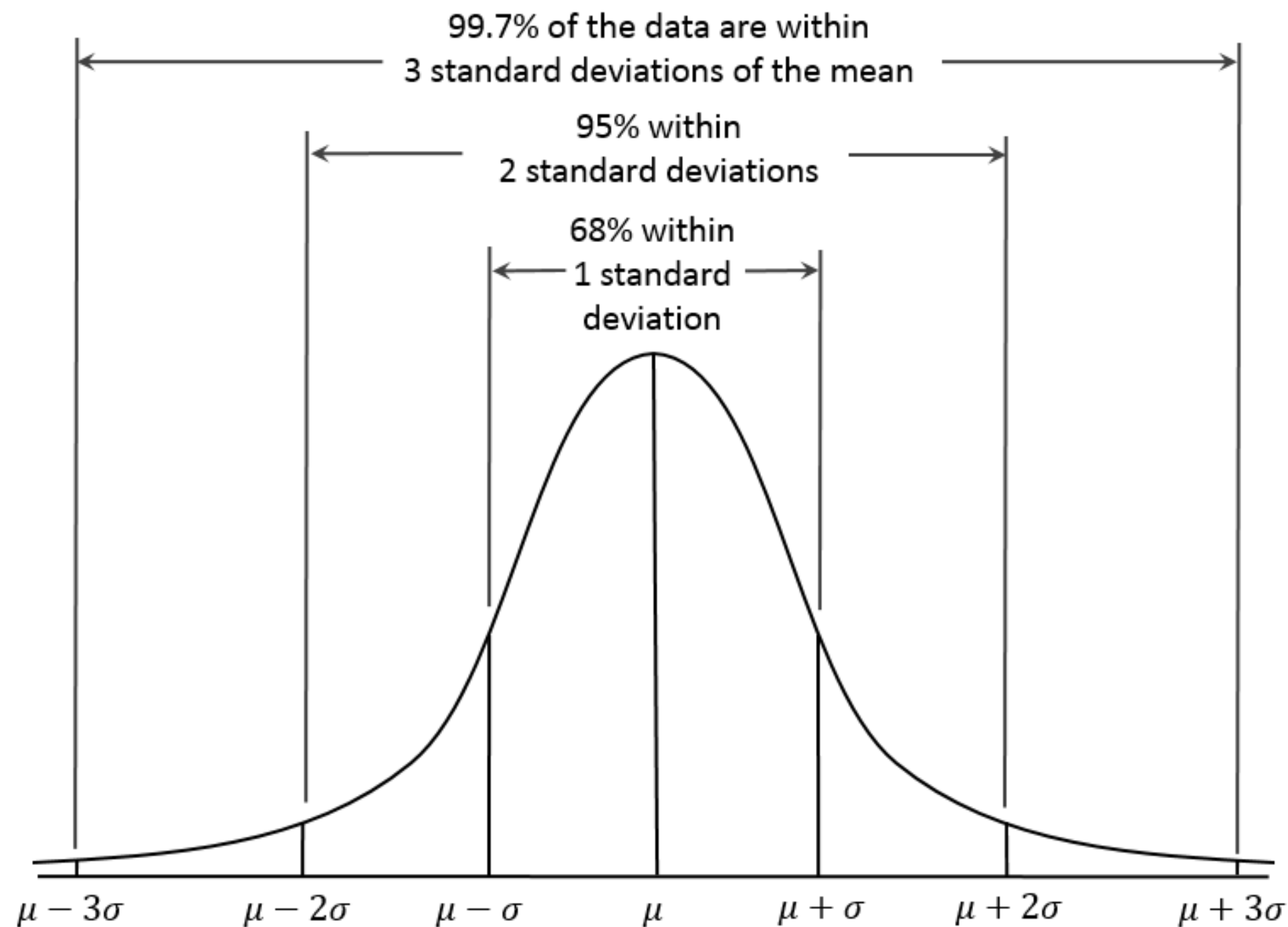
$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = 21704$$

- Calculate Standard Deviation:  $\sigma = \sqrt{21704} = 147.32$
- Now we can show which heights are within Standard Deviation (147mm) of the Mean:



# Normal Distribution

- For the normal distribution, the values within one standard deviation of the mean account for 68.27% of the set; while within two standard deviations account for 95.45%; and within three standard deviations account for 99.73%.



# Var. and Std. in MATLAB

## Command Window

```
>> Dog = [600, 470, 170, 430, 300];  
>> var(Dog) % A sample of population
```

```
ans =
```

```
27130
```

```
>> std(Dog) % A sample of population
```

```
ans =
```

```
1.647118696390761e+02
```

```
>> var(Dog,1) % Population
```

```
ans =
```

```
21704
```

```
>> std(Dog,1) % Population
```

```
ans =
```

```
1.473227748856232e+02
```

- You have measured the heights (at the shoulders) of your dogs in millimetres: 600, 470, 170, 430, 300.
- MATLAB default is **the sample of population** for Variance calculation.



University of  
East London

# Example

Command Window		Workspace	
>> X = rand(10,1)*10		Name^	Value
X =		ans	[11.9614,3.4585,0.9754,9.6489]
8.1472		X	[8.1472;9.0579;1.2699;9.1338;6.3236;0.9754;2.7850;5.4688;9.5751;9.6489]
9.0579			
1.2699			
9.1338			
6.3236			
0.9754			
2.7850			
5.4688			
9.5751			
9.6489			
>> [var(X) std(X) min(X) max(X)]			
ans =			
11.9614	3.4585	0.9754	9.6489



# Example

Command Window				Workspace	
>> Y = rand(7,4)*10				Name ^	Value
Y =				ans	5x4 double
				Y	7x4 double
0.1190      6.0198      2.2898      4.4268					
3.3712      2.6297      9.1334      1.0665					
1.6218      6.5408      1.5238      9.6190					
7.9428      6.8921      8.2582      0.0463					
3.1122      7.4815      5.3834      7.7491					
5.2853      4.5054      9.9613      8.1730					
1.6565      0.8382      0.7818      8.6869					
>> [var(Y); std(Y); min(Y); max(Y); mean(Y)]					
ans =					
6.8623      6.0557      14.8181      14.9411					
2.6196      2.4608      3.8494      3.8654					
0.1190      0.8382      0.7818      0.0463					
7.9428      7.4815      9.9613      9.6190					
3.3013      4.9868      5.3331      5.6811					



# Example

app1.m

```
1- clc;
2- clear;
3- D = rand(7,4) * 10; % Generate a random matrix
4- Var = var(D); STD = std(D); MIN = min(D); MAX = max(D);
5-
6- D
7-
8- disp(['The Var of D is --> ' num2str(Var)]);
9- disp(['The STD of D is --> ' num2str(STD)]);
10- disp(['The Min value of D is --> ' num2str(MIN)]);
11- disp(['The Max value of D is --> ' num2str(MAX)]);
12-
13- [Var; STD; MIN; MAX]
14-
```

Name	Value
ans	4x4 double
D	7x4 double
MAX	[8.7393,9.4793,9.3120,9.8440]
MIN	[1.2393,0.8207,0.5208,0.6340]
STD	[2.7742,3.0297,3.3758,3.1200]
Var	[7.6963,9.1789,11.3957,9.7345]

Command Window

D =  
  
5.3580    2.0846    1.0571    7.2866  
4.4518    5.6498    1.4204    7.3784  
1.2393    6.4031    1.6646    0.6340  
4.9036    4.1703    6.2096    8.6044  
8.5300    2.0598    5.7371    9.3441  
8.7393    9.4793    0.5208    9.8440  
2.7029    0.8207    9.3120    8.5894  
  
The Var of D is --> 7.69631        9.1789        11.3957        9.73451  
The STD of D is --> 2.7742        3.0297        3.3758        3.12  
The Min value of D is --> 1.2393        0.82071        0.52078        0.63405  
The Max value of D is --> 8.7393        9.4793        9.312        9.844  
  
ans =  
  
7.6963    9.1789    11.3957    9.7345  
2.7742    3.0297    3.3758    3.1200  
1.2393    0.8207    0.5208    0.6340  
8.7393    9.4793    9.3120    9.8440

# Numerical Measures

## 3. Exploratory Data Analysis

- **Five-number Summary:** It is a set of descriptive statistics that provide information about a dataset. It consists of the five most important sample percentiles.

# Five-number Summary

■ In a five-number summary, the following five numbers are used to summarize the data:

1. **the sample minimum (smallest observation)**
2. **the lower quartile or first quartile (Q1)**
3. **the median (the middle value) (Q2)**
4. **the upper quartile or third quartile (Q3)**
5. **the sample maximum (largest observation)**

■ The easiest way to develop a five-number summary is to first place the data in ascending order. Then it is easy to identify the smallest value, the three quartiles, and the largest value.



# Five-number Summary example

■ We have final marks for a class as follows:

[48 57 58 65 68 69 71 73 73 74 75 77 78 78 78 79 80 85 87 88 89 89 89 95 96 97 99]

■ Illustrate the descriptive analysis over the data through Five-number summary method.

**Step 0:** Put your numbers in ascending order (from smallest to largest)

**Step 1: Smallest value:** Find the minimum for your data. **Answer = 48**

**Step 2: First Quartile (Q1)**

**Step 3: Median (Q2):** The median is the middle number. **Answer = 78**

**Step 4: Third Quartile (Q3)**

**Step 5: Largest value:** Find the maximum for your data. **Answer = 99**

# Five-number Summary example

Place parentheses around the numbers above and below the median

(48 57 58 65 68 69 71 73 73 74 75 77 78) **78**

(78 79 80 85 87 88 89 89 89 95 96 97 99)

**Step 2: First Quartile (Q1):** the median in the lower half of the data.

(48 57 58 65 68 69 **71** 73 73 74 75 77 78)

**Step 4: Third Quartile (Q3):** the median in the upper half of the data.

(78 79 80 85 87 88 **89** 89 89 95 96 97 99)

Summary:

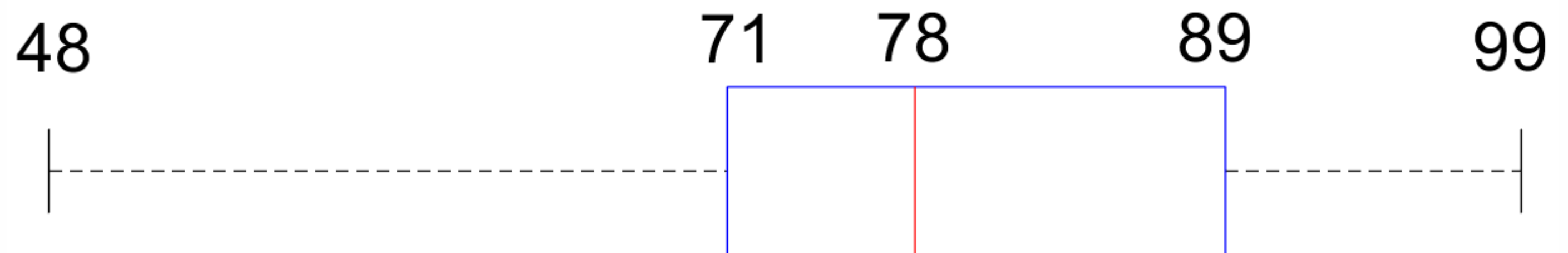
minimum=**48**;

Q1=**71**;

median=**78**,

Q3=**89**;

maximum=**99**.



University of  
East London



# Box Plot

- A box plot is a graphical summary of data that is based on a five-number summary.

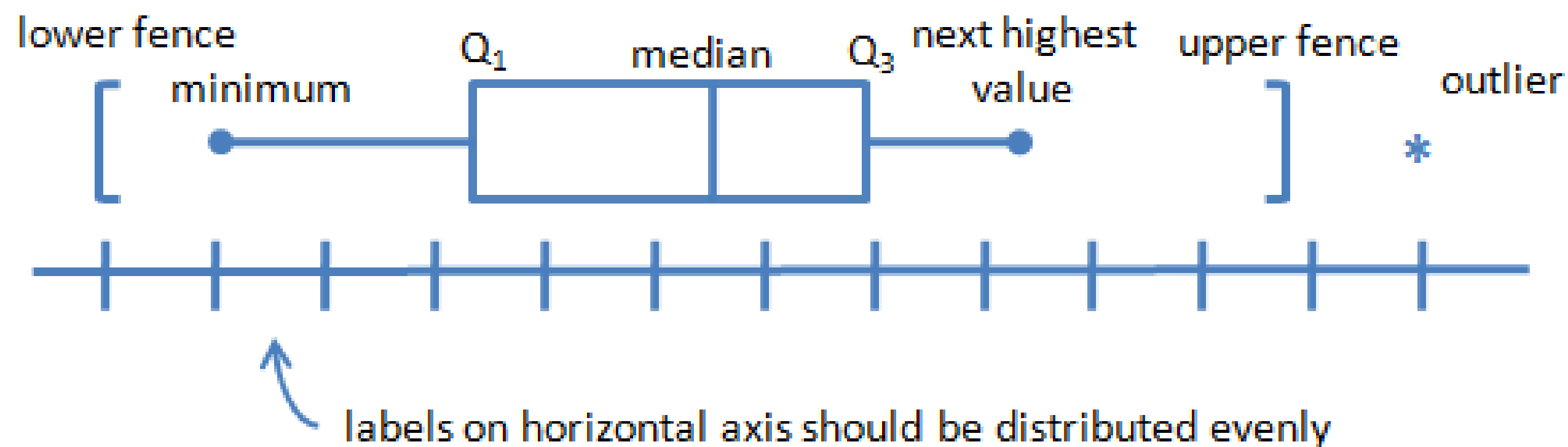
**Step 1:** Determine the five-number summary and the lower and upper fences.

**Step 2:** Draw a horizontal line and label it with an appropriate scale.

**Step 3:** Draw vertical lines at  $Q_1$ ,  $M$ , and  $Q_3$ . Enclose these vertical lines in a box.

**Step 4:** Draw a line from  $Q_1$  to the smallest data value that is within the lower fence. Similarly, draw a line from  $Q_3$  to the largest value that is within the upper fence.

**Step 5:** Any values outside the fences are outliers and are marked with an asterisk (\*).



# Summary

- Explained Descriptive Statistics
- Discussed the differences between Qualitative and Quantitative Data
- Explained Discrete and Continuous Data types
- Understood and practiced many statistical methods in MATLAB