Statistics Session-2

Course Info

Lesson Plan

STATISTICS BASICS

The goal of this course is to provide a comprehensive overview of the basics of statistics you will need to start your data science journey.

Custodian : Jason-Acad.Coord. (jason@clarusway.com)

In-class Sessions : 7 In-classes / 21 hours (*Part-1 → 3 In-classes | Part-2 → 4 In-classes*)

Lab Sessions : 3 Labs / 3 hours (*Part-1 → 1 Lab | Part-2 → 2 Labs*)

Certification Requirements:

1. Attend at least 70% of in-class sessions (at least 5 sessions of attendance)
2. Successfully complete and submit assignments (at least 2 assignments)

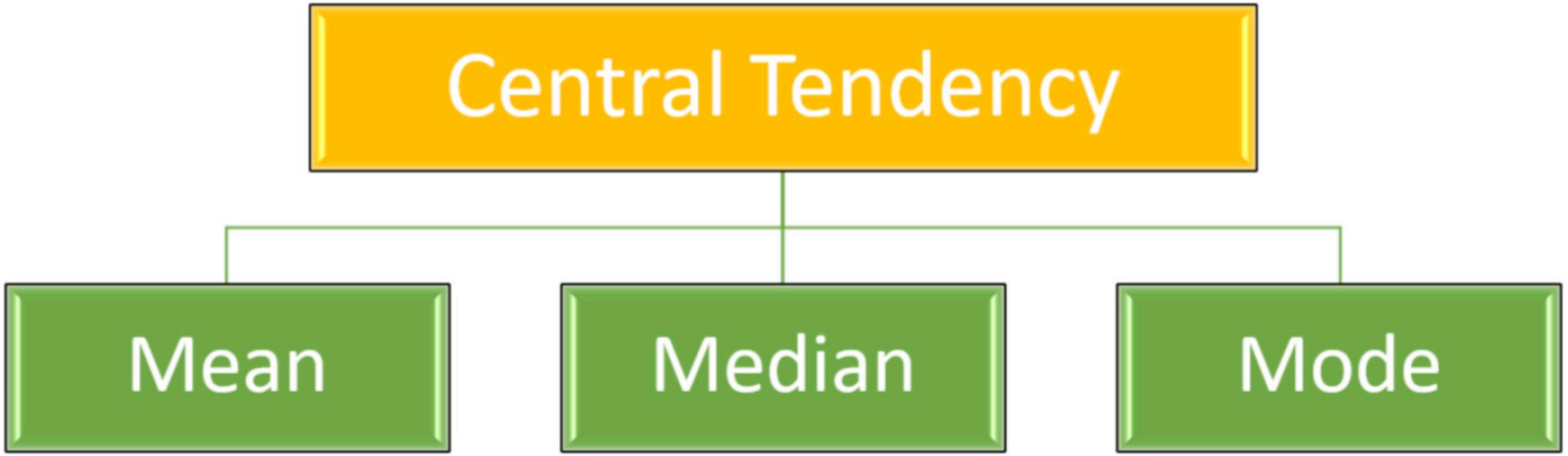
Table of Contents

1

Central Tendency (Measure of Centre)

Central Tendency (Measure of Centre)

The central tendency concept is that one single value can best describe the data.



Mean

The mean is equal to the sum of the values in the dataset divided by the number of values.

1. Find the sum of all values in a group of values
2. Divide the sum by the number of values in the group.

$$Mean = \frac{Sum\ of\ all\ values}{Number\ of\ values}$$

Mean *Notation*

Population Mean

$\sum X = Sum\ of\ all\ X\ values$

$N = Number\ of\ X\ values$

$\mu = Population\ mean$

$$\mu = \frac{\sum X}{N}$$

Mean *Notation*

Population Mean

$\sum X = Sum\ of\ all\ X\ values$

$N = Number\ of\ X\ values$

$\mu = Population\ mean$

$$\mu = \frac{\sum X}{N}$$

Sample Mean

$\sum x = Sum\ of\ all\ x\ values$

$n = Number\ of\ x\ values$

$\bar{x} = Sample\ mean$

$$\bar{x} = \frac{\sum x}{n}$$

Mean *Example*

Compute the mean age.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47

$$\mu = \frac{\sum X}{N}$$

Mean *Example*

Compute the mean age.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47

$$\mu = \frac{\sum X}{N}$$

$$\sum X = 49 + 64 + 36 + 64 + 47 = 260$$

$$N = 5$$

$$\mu = \frac{260}{5} = 52$$

Median

The median is the middle score for a dataset that has been sorted from small to large.

1. List scores from smallest to largest

2. With an odd number of scores, the median is the middle score.

3. With an even number of scores, the median is the sum of the middle two scores divided by 2.

Median *Example 1*

Find the median age, given an odd number of scores.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47

1. List scores in ascending order.

36 47 49 64 64

Median *Example 1*

Find the median age, given an odd number of scores.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47

1. List scores in ascending order.

   36 47 49 64 64

2. With an odd number of scores, the median is the middle score.

   36 47 49 64 64

Median age

Median *Example 2*

Find the median age, given an even number of scores.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
Martha L. Fox	47

1. List scores in ascending order.

   36 47 49 64

Median *Example 2*

Find the median age, given an even number of scores.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
Martha L. Fox	47

1. List scores in ascending order.

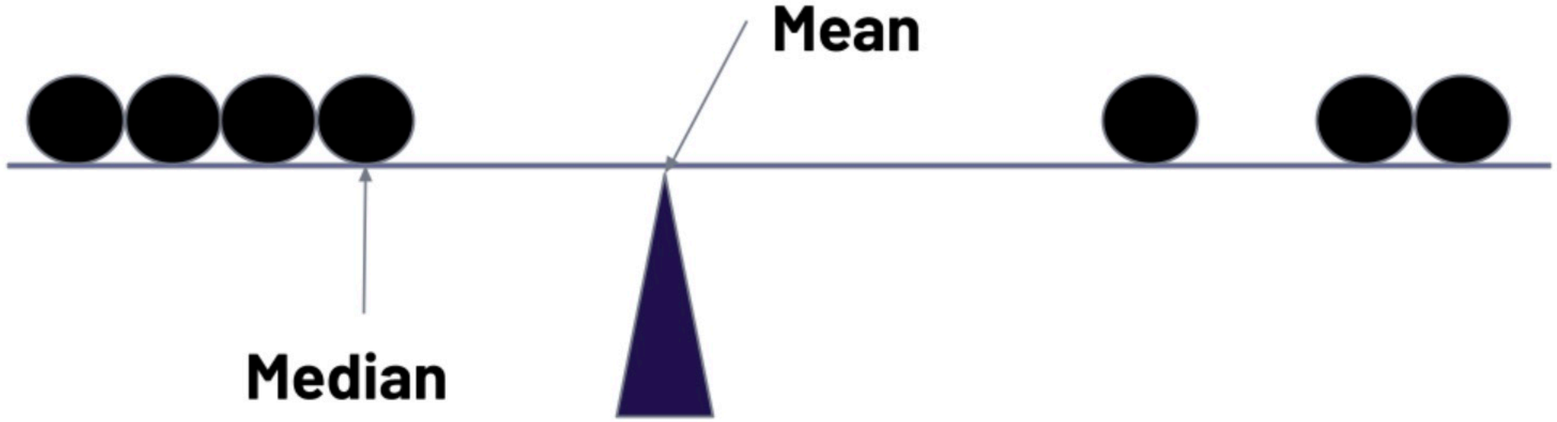
   36 47 49 64

2. With an even number of scores, the median is the sum of the middle two scores divided by 2.

$$Median = \frac{47 + 49}{2} = 48$$

Mean vs. Median

- The median is better if a small set of scores has an outlier.
- The mean is better if a large set of scores does not have an outlier.



Selection of Imputation Method for Missing Values

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Selection of Imputation Method for Missing Values

	First Name	Gender	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	97308.0	6.945	True	Marketing
1	Thomas	Male	61933.0	NaN	True	NaN
2	Jerry	Male	NaN	9.340	True	Finance
3	Dennis	n.a.	115163.0	10.125	False	Legal
4	NaN	Female	NaN	11.598	NaN	Finance
5	Angela	NaN	NaN	18.523	True	Engineering
6	Shawn	Male	111737.0	6.414	False	NaN
7	Rachel	Female	142032.0	12.599	False	Business Development
8	Linda	Female	57427.0	9.557	True	Client Services
9	Stephanie	Female	36844.0	5.574	True	Business Development
10	NaN	NaN	NaN	NaN	NaN	NaN

Mean vs. Median *Example*

$4000 **$15.000** **$20.000** **$33.000** **$1.800.000**

Find the mean and the median of car prices.

Mean vs. Median *Example*

$4000 $15.000 $20.000 $33.000 $1.800.000

Mean:

$$\mu = \frac{\sum X}{N}$$

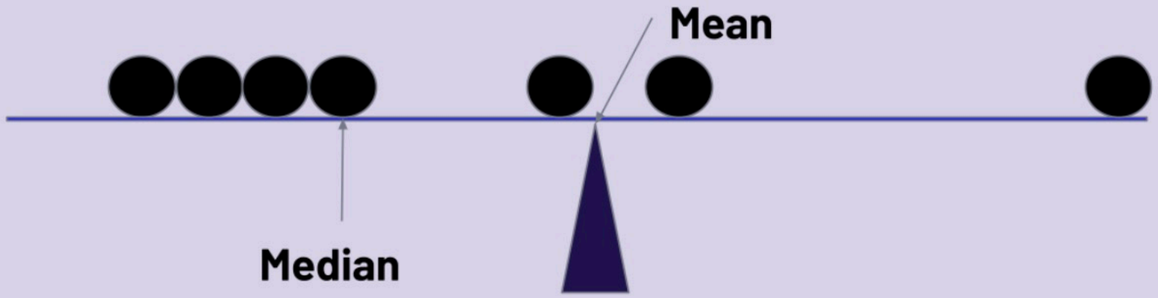
$$\mu = \frac{\$4000 + \$15000 + \$20000 + \$33000 + \$1800000}{5}$$

$$\mu = \frac{\$1872000}{5} = \$374400$$

Median: $20000

Let's Practice

_____ is resistant to outliers.

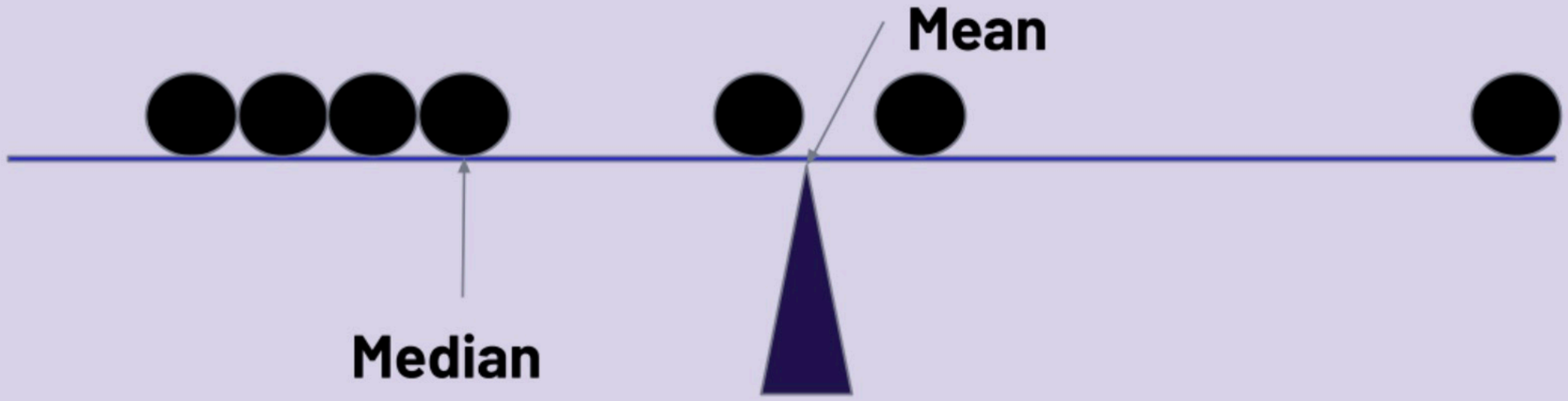


Students choose an option

No Multiple Choice Response
You didn't answer this question

Let's Practice Answer

Median is resistant to outliers.

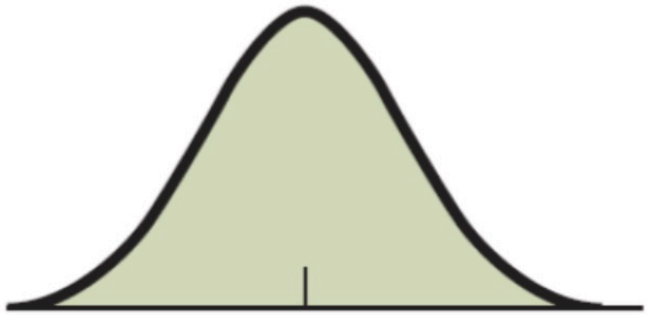


Mean vs. Median

Generally, if the shape is

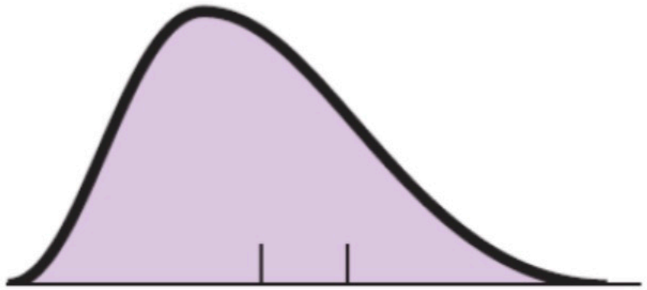
- Perfectly symmetric, the mean equals the median.
- Skewed to the right, the mean is larger than the median.
- Skewed to the left, the mean is smaller than the median.

Symmetric Distribution



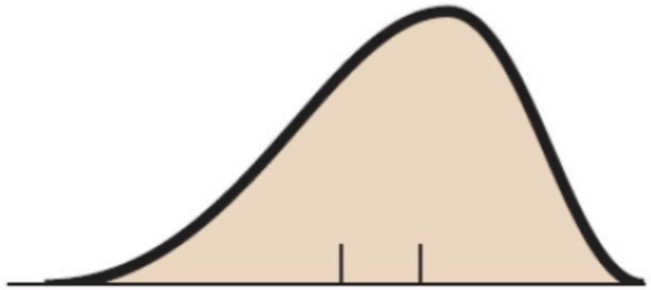
Mean = Median

Right-Skewed Distribution



Median Mean

Left-Skewed Distribution



Mean Median

According to Statistics Canada, in 2004 the median household income in Canada was $58,000 and the mean was $76,000. What would you predict about the shape of the distribution?

Students, write your response!

Pear Deck Interactive Slide
Do not remove this bar

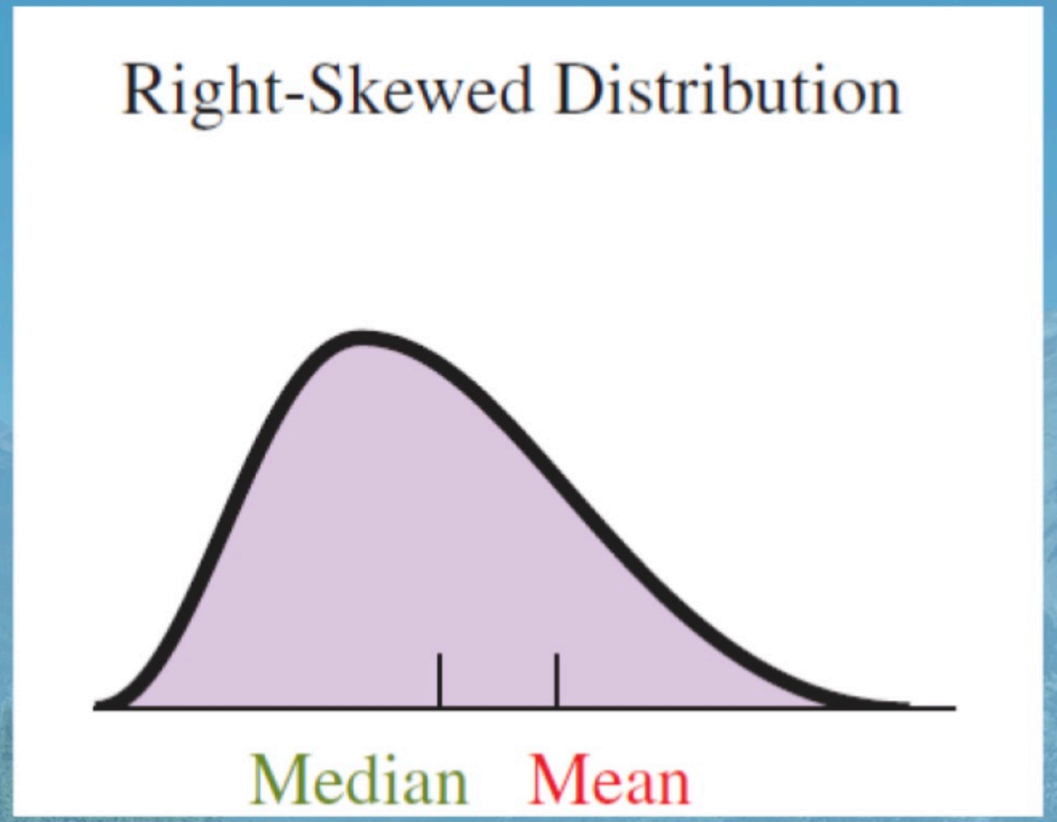
No Text Response
You didn't answer this question

median = $58,000

mean = $76,000

mean > median

Right-skewed

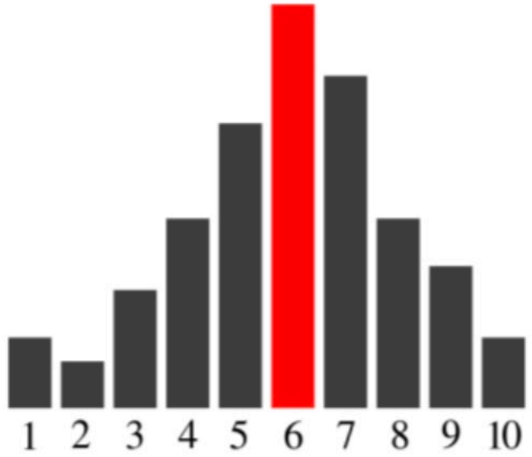


Mode

The mode is the most frequent score in a dataset.

1 List scores from smallest to largest.

2 Count how many of each number. A number that appears most often is the mode.



Mode *Example*

Find the mode.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47
Tim Berners-Lee	64

1. List scores from smallest to largest.

36 47 49 64 64 64

Mode *Example*

Find the mode.

Name	Age
Elon Musk	49
Bill Gates	64
Mark Zuckerberg	36
G. van Rossum	64
Martha L. Fox	47
Tim Berners-Lee	64

1. List scores from smallest to largest.

36 47 49 64 64 64

2. Count how many of each number.

1 1 1 3

36 47 49 64 64 64

Mode

Pro's and Con's of Mode

Advantages:

- The mode is easy to understand and calculate.
- The mode is not affected by extreme values.
- The mode is useful for categorical data.

Disadvantages:

- The mode is not defined when there are no repeats in a data set.
- The mode is not based on all values.
- The mode is unstable when the data consist of a small number of values.
- Sometimes data have one mode, more than one mode, or no mode at all.

Selection of Imputation Method for Missing Values

Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

Mode = Ford

Ford
Ford
Fiat
BMW
Ford
Kia
Ford
Fiat
Ford
Ford
Kia

Selection of Imputation Method for Missing Values

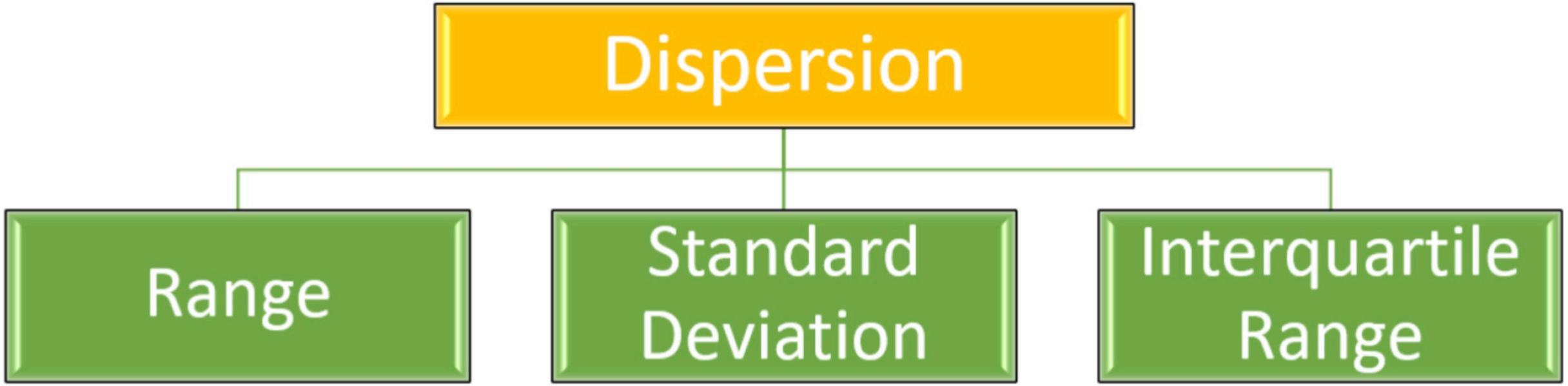
	id	A	B	C	D
0	0	10.0	A	NaN	NaN
1	1	9.0	B	BB	20.0
2	2	8.0	A	CC	18.0
3	3	7.0	A	BB	22.0
4	4	NaN	NaN	BB	18.0
5	5	NaN	B	CC	17.0
6	6	20.0	A	AA	19.0
7	7	15.0	B	BB	NaN
8	8	12.0	NaN	NaN	17.0
9	9	NaN	A	AA	23.0

2

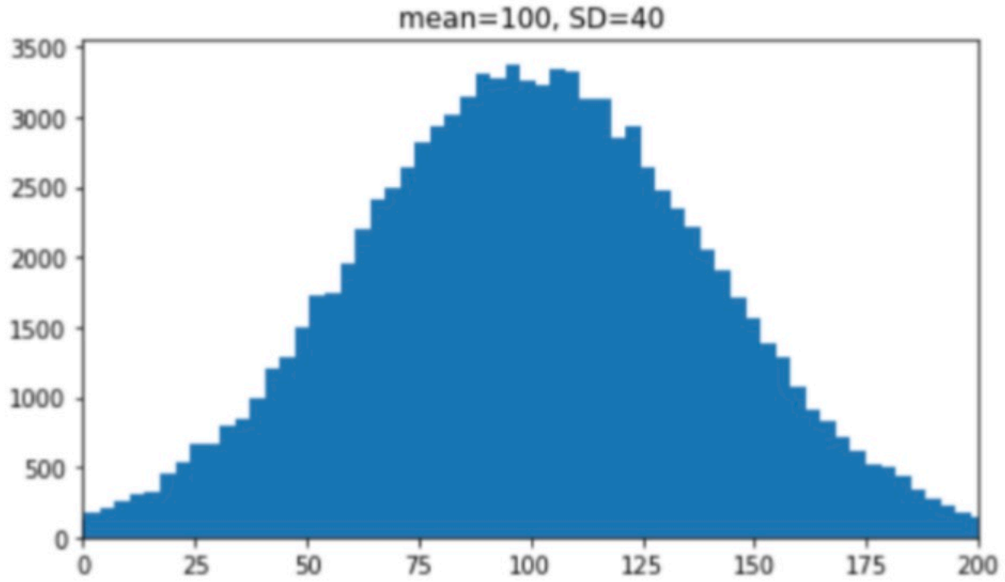
Dispersion (Measure of Spread)

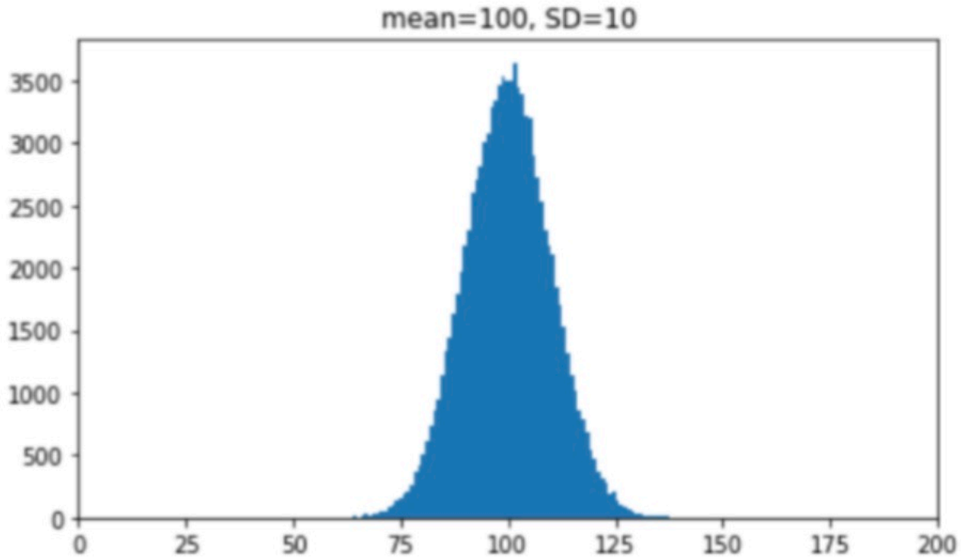
Dispersion (Measure of Spread)

The most common measures of variability are the range, the interquartile range (IQR), variance, and standard deviation.



Dispersion vs Central Tendency





Means are the same but standard deviations are quite different. Dispersion gives idea about distributions.

Range

The range is the difference between the largest and smallest values in a set of values.

Example:

2 4 9 5 7 3

Range = Largest - Smallest = 9 - 2 = 7

The Interquartile Range (IQR)

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles.

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called the first, second, and third quartiles.
- First, second, and third quartiles are denoted by Q1, Q2, and Q3, respectively.

IQR *Example*

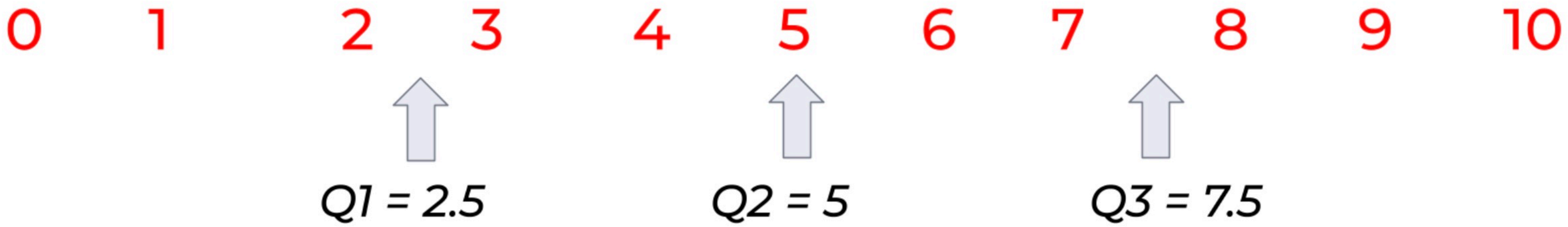
Ordered data set.

0 1 2 3 4 5 6 7 8 9 10

Q2 = 5

IQR *Example*

Ordered data set.

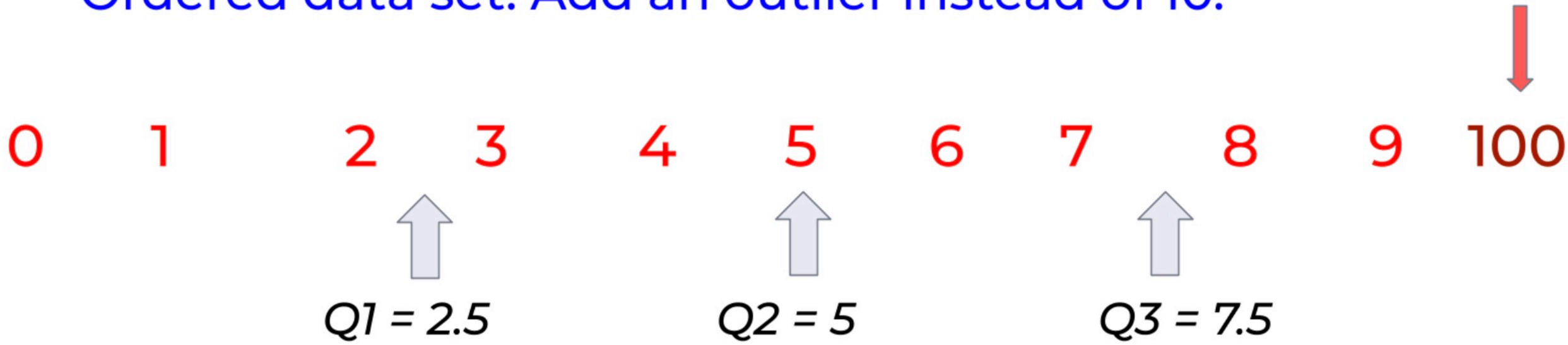


Interquartile Range = 7.5 - 2.5

IQR = 5

IQR *Example*

Ordered data set. Add an outlier instead of 10.



Interquartile Range = 7.5 - 2.5

IQR = 5

IQR *Practice*

Question:

What is the interquartile range of these numbers?

8 9 10 10 12 13 14 15 16

Students, enter a number!

No Number Response

You didn't answer this question

NEXT SLIDE

IQR Practice Answer

Question:

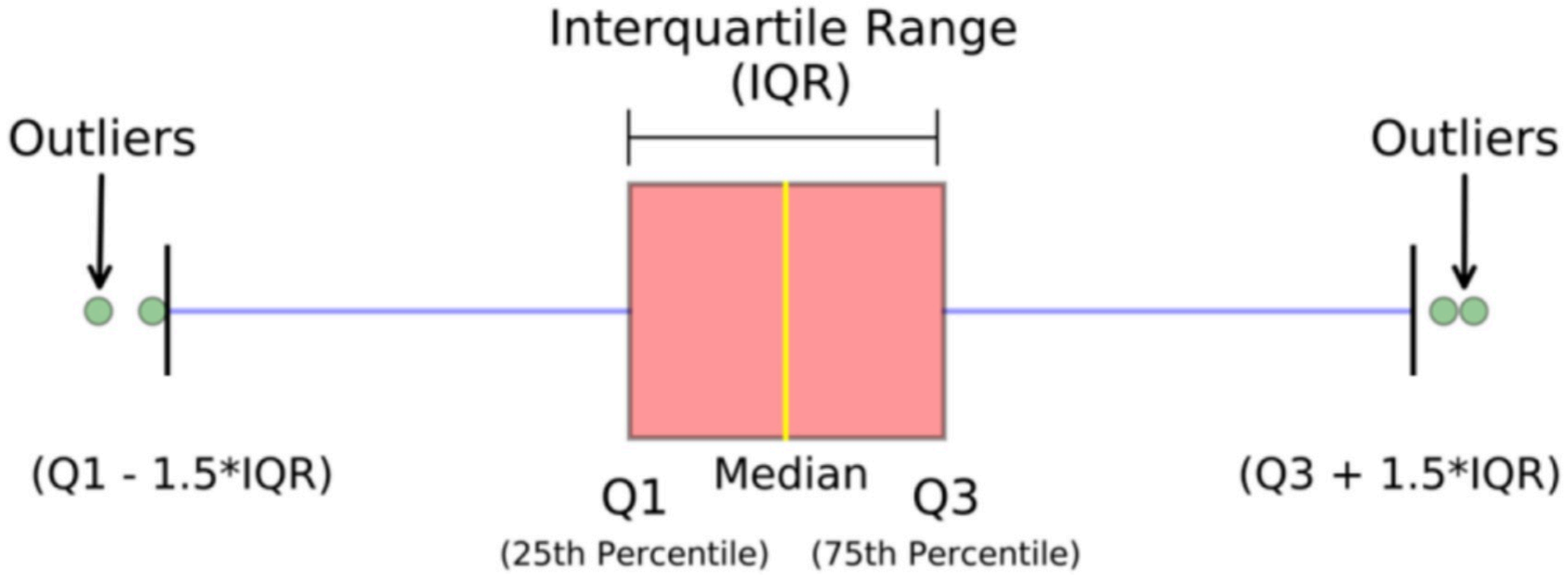
What is the interquartile range of these numbers?

8 9 (10) 10 12 13 (14) 15 16

IQR = Q3 - Q1
IQR = 14 - 10
IQR = 4

1.5xIQR Rule for Outliers

1.5 IQR Rule: If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an outlier.



Outlier Practice

Question:

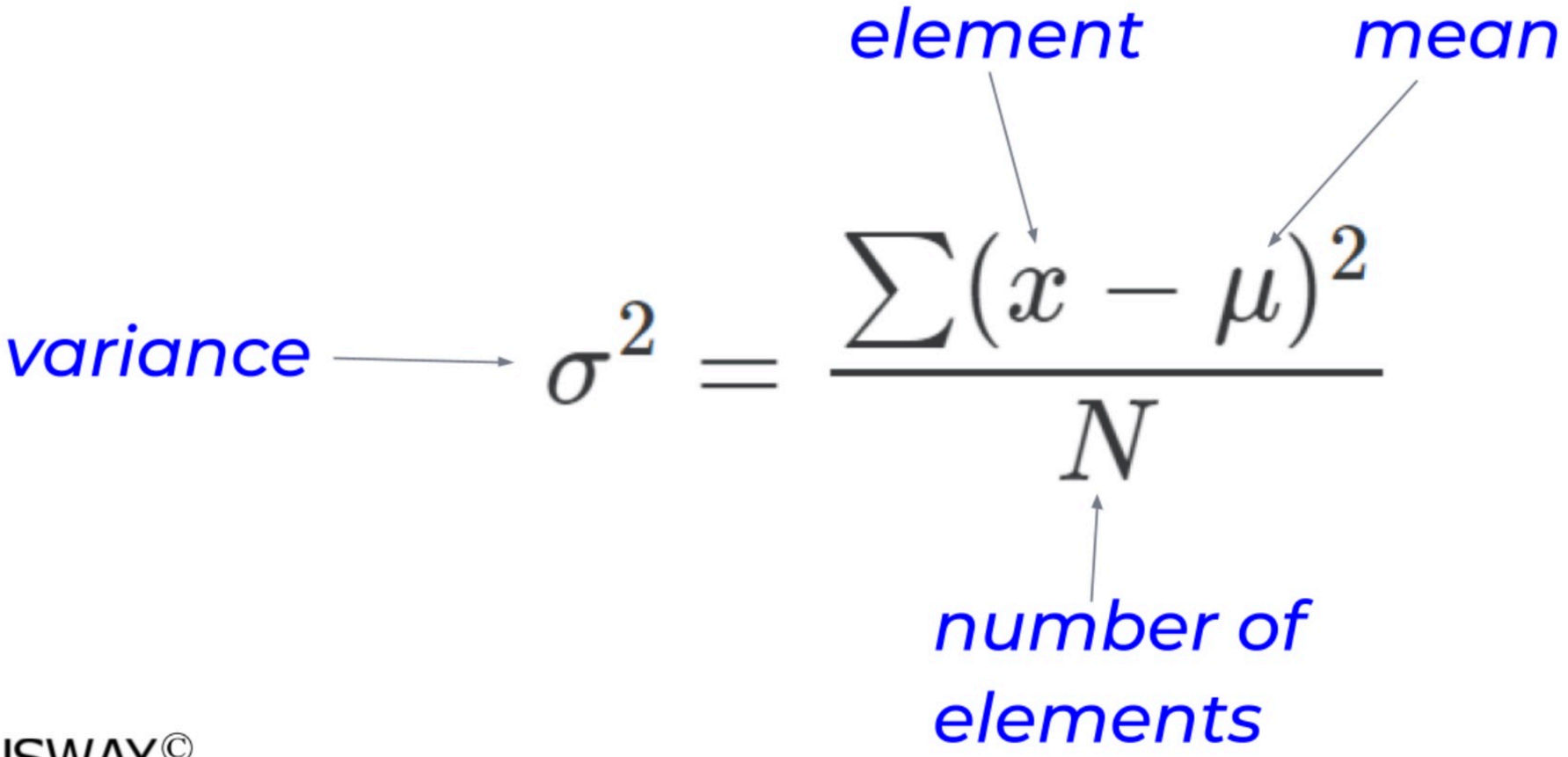
Are there any outliers?

8 9 (10) 10 12 13 (14) 15 22

IQR = Q3 - Q1
IQR = 14 - 10
IQR = 4

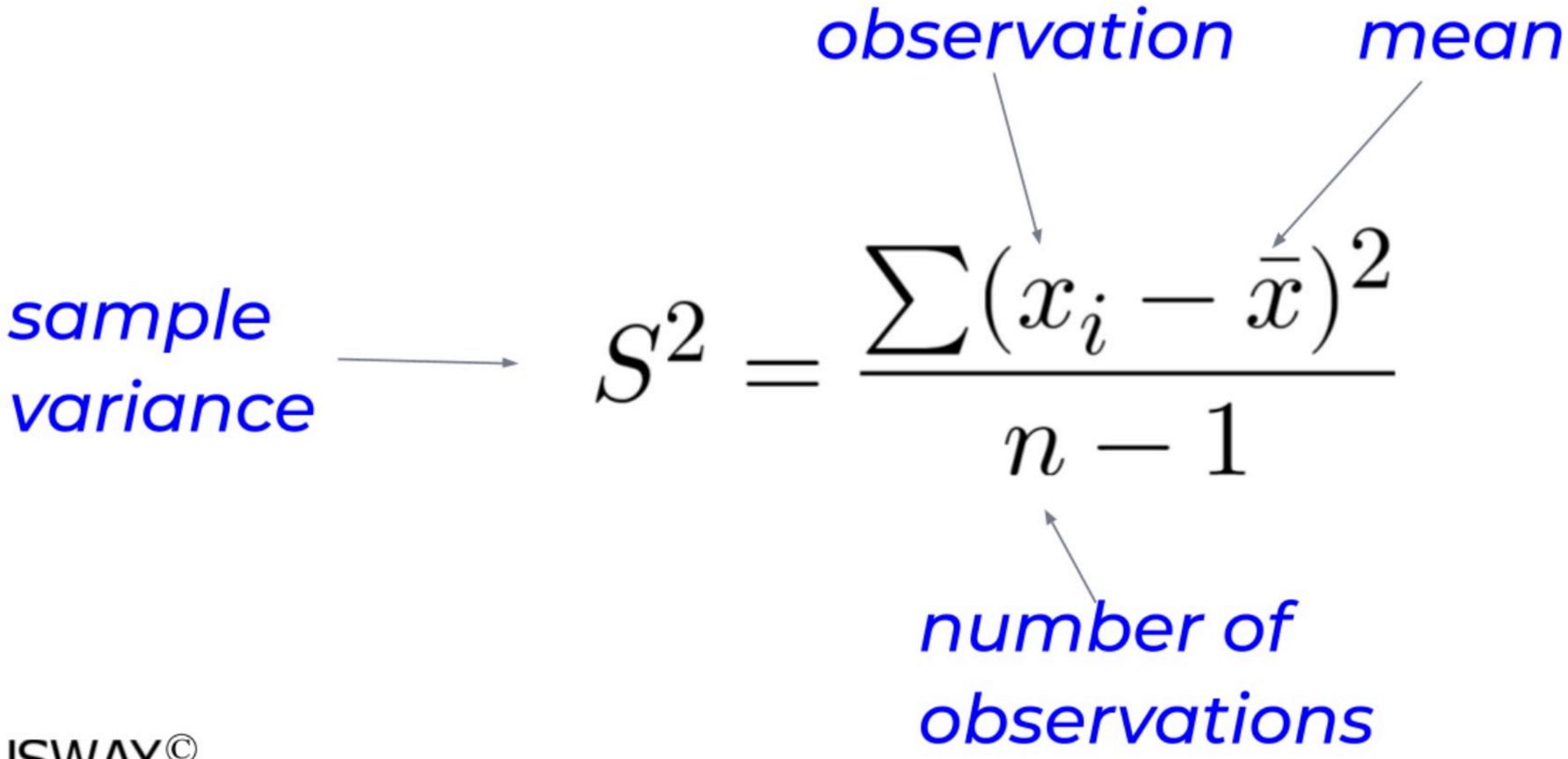
Variance (Population)

Variance is the average squared deviation from the mean.



Variance (Sample)

Variance is the average squared deviation from the mean.



Variance *Example*

Find the Variance.

0 1 5 6

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

Variance *Example*

Find the Variance.

0 1 5 6

Mean: $\mu = \frac{\sum X}{N} = \frac{0+1+5+6}{4} = \frac{12}{4} = 3$

Dev Sum of Squares: $SS = \sum(X - \mu)^2$

$SS = (0-3)^2 + (1-3)^2 + (5-3)^2 + (6-3)^2$

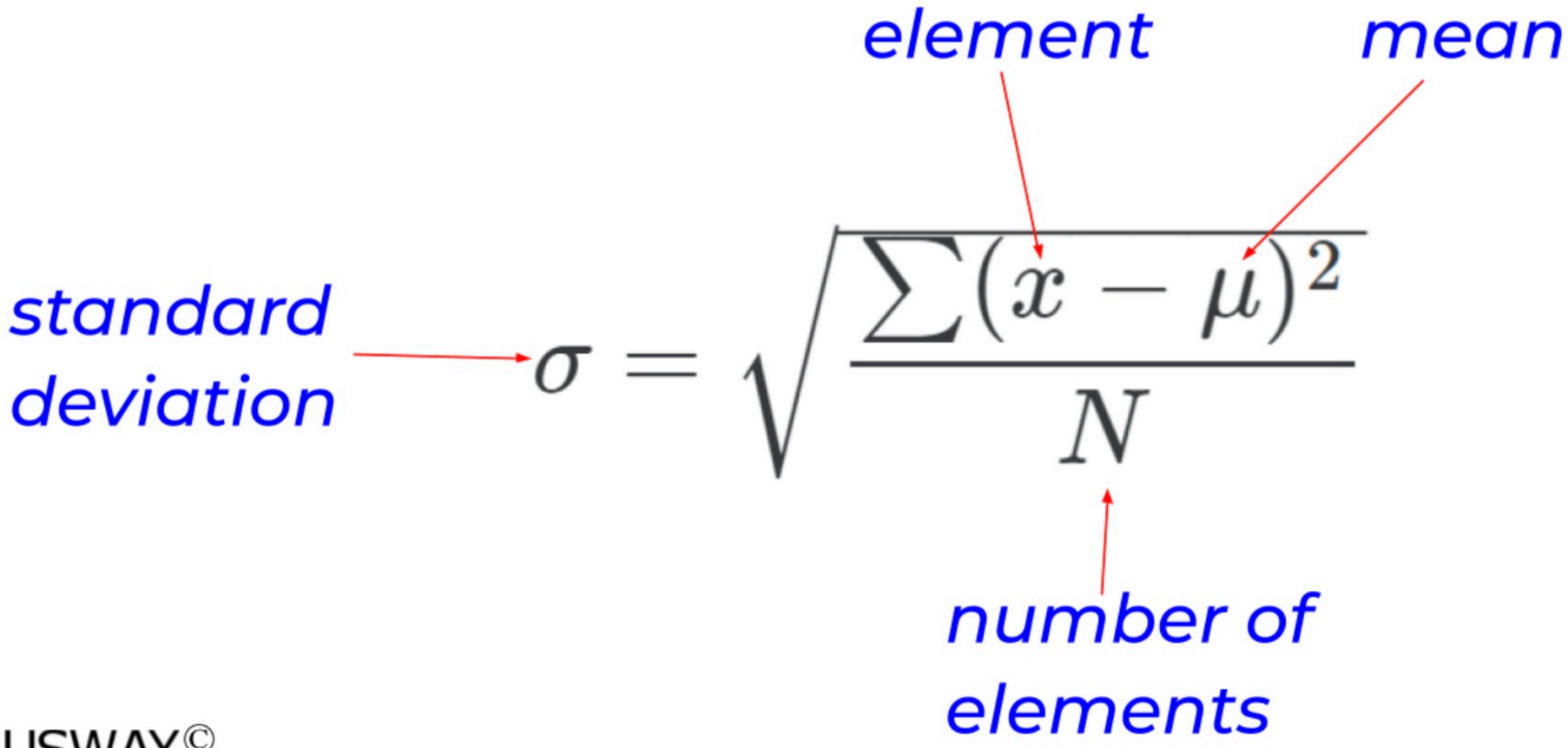
$SS = 9 + 4 + 4 + 9 = 26$

Variance: $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$

$\sigma^2 = \frac{26}{4} = 6.5$

Standard Deviation

Standard deviation is the square root of the variance.



Standard Deviation *Example*

Students in a class were asked on a questionnaire at the beginning of the course,
"How many children do you think is ideal for a family?"
The observations, classified by student's gender, were

Men	:	0	0	0	2	4	4	4
Women	:	0	2	2	2	2	2	4

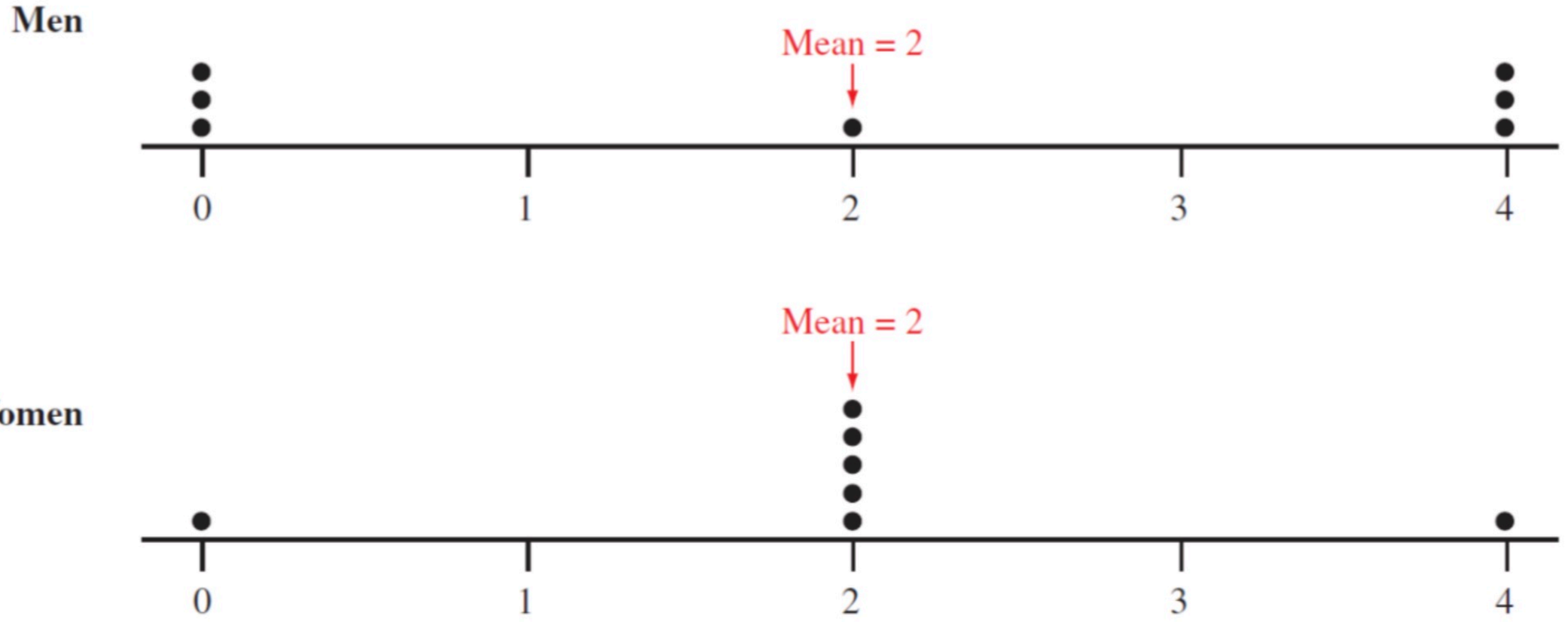
Both men and women have a mean of 2 and a range of 4.

Do the distributions of data have the same amount of variability around the mean?

Standard Deviation *Example*

Men : 0 0 0 2 4 4 4

Women : 0 2 2 2 2 2 4



Standard Deviation *Example*

Men : 0 0 0 2 4 4 4

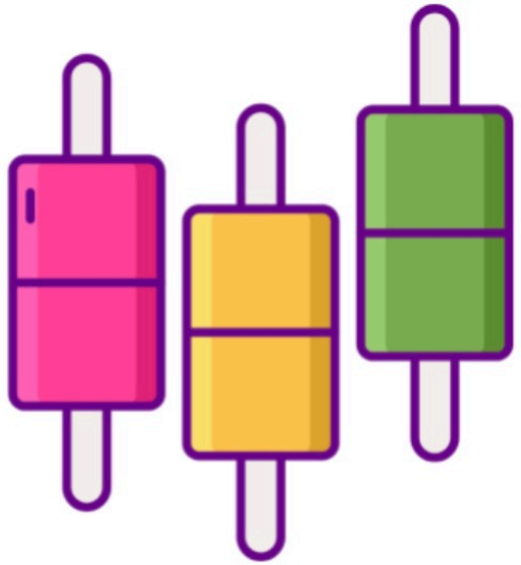
Women : 0 2 2 2 2 2 4

Men: $$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{24}{6}} = \sqrt{4} = 2.0$$

Women: $s = 1.2$

The observations for males tend to be farther from the mean than those for females, as indicated by $s = 2.0 > s = 1.2$. In summary, the men's observations vary more around the mean.

Box Plot

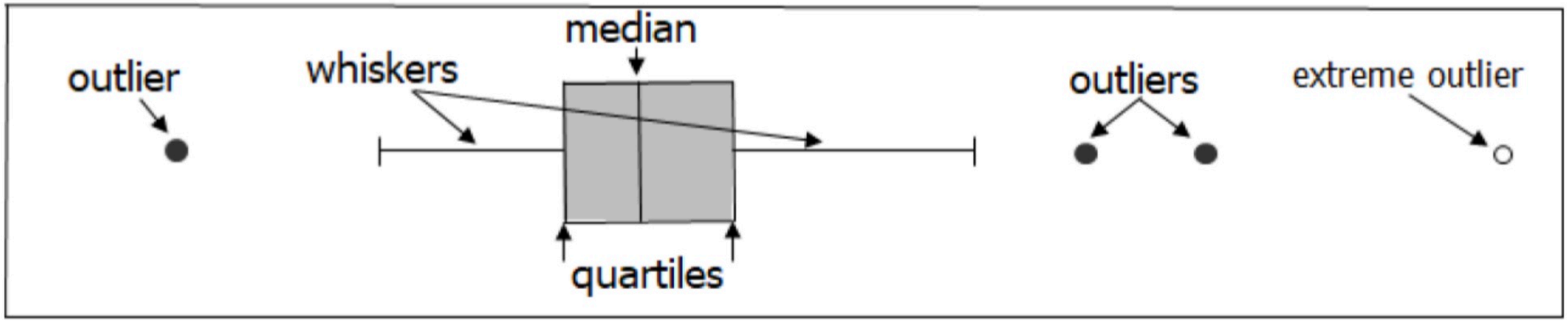


What is Box Plot

A **box plot** is a method for graphically depicting groups of numerical data through their *quartiles*.

A box plot generally shows

- ★ median
- ★ 1st quartile (Q1)
- ★ 3rd quartile (Q3)
- ★ outliers

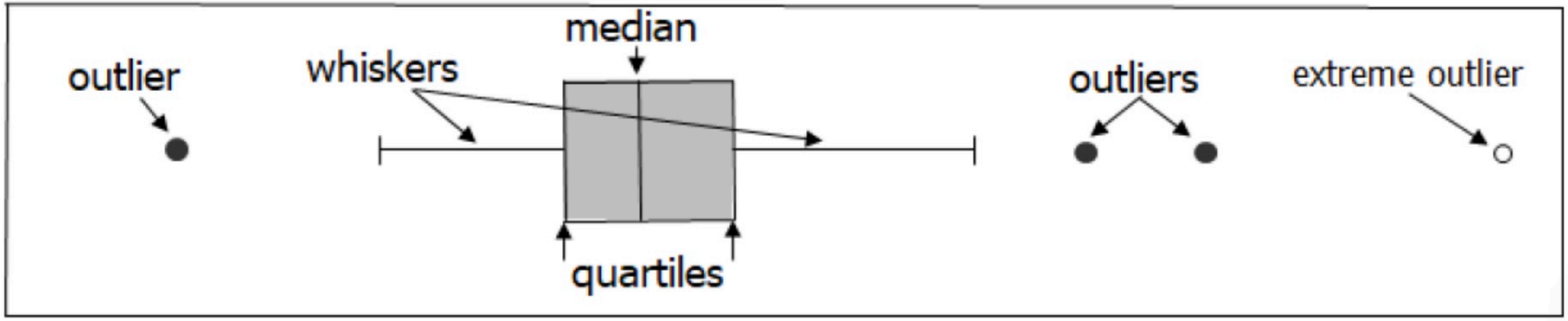


Box Plot

Boxplots show distribution in one dimension

- ★ Only useful for continuous variables
- ★ Good for comparing distributions of a continuous variable between categorical groups

Box plots are also known as **box and whisker plots**.



How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

Step 3: Find the quartiles.

25 28 29 29 30 34 35 35 37 38

Q1 = 29 *Q3 = 35*

How to make a Box Plot (Min & Max)

Weight, kg
38
25
37
28
35
29
35
29
34
30

Step 1: Order the data from smallest to largest.

25 28 29 29 30 34 35 35 37 38

Step 2: Find the median.

25 28 29 29 30 34 35 35 37 38

Median = 32

Step 3: Find the quartiles.

25 28 29 29 30 34 35 35 37 38

Q1 = 29 *Q2 = 35*

Step 4: Find the min and the max.

Min = 25 *Max = 38*

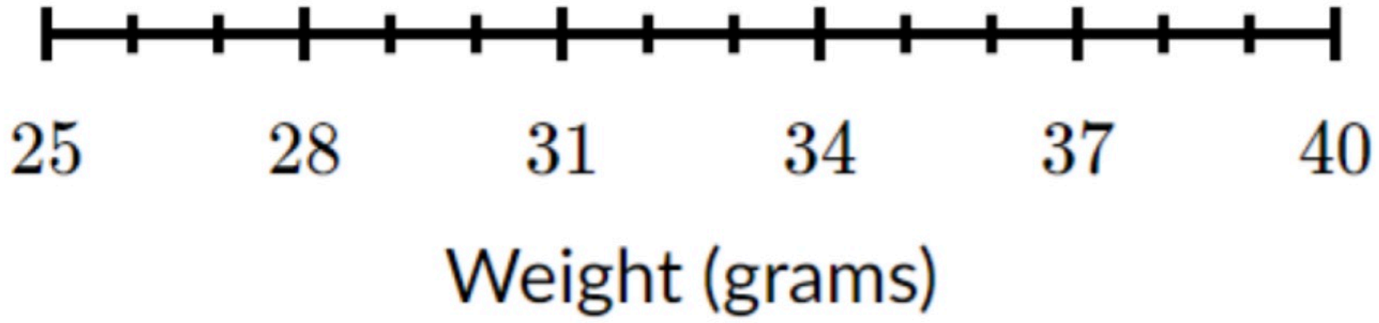
How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 1: Scale / label an axis that fits the five-number.

25 28 29 29 30 34 35 35 37 38

Min = 25 *Max = 38*

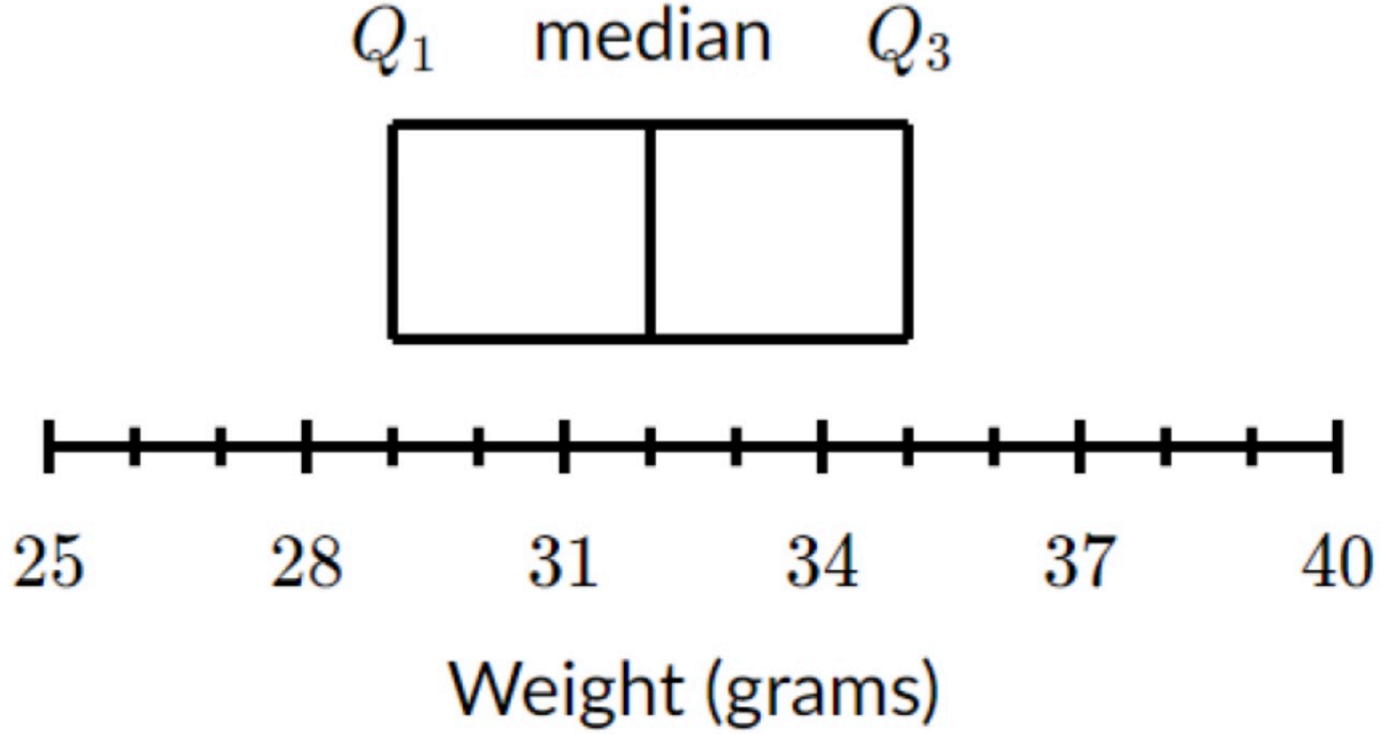


How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 2: Draw a box from *Q1* to *Q3* with a vertical line through the *median*.

25 28 29 29 30 34 35 35 37 38

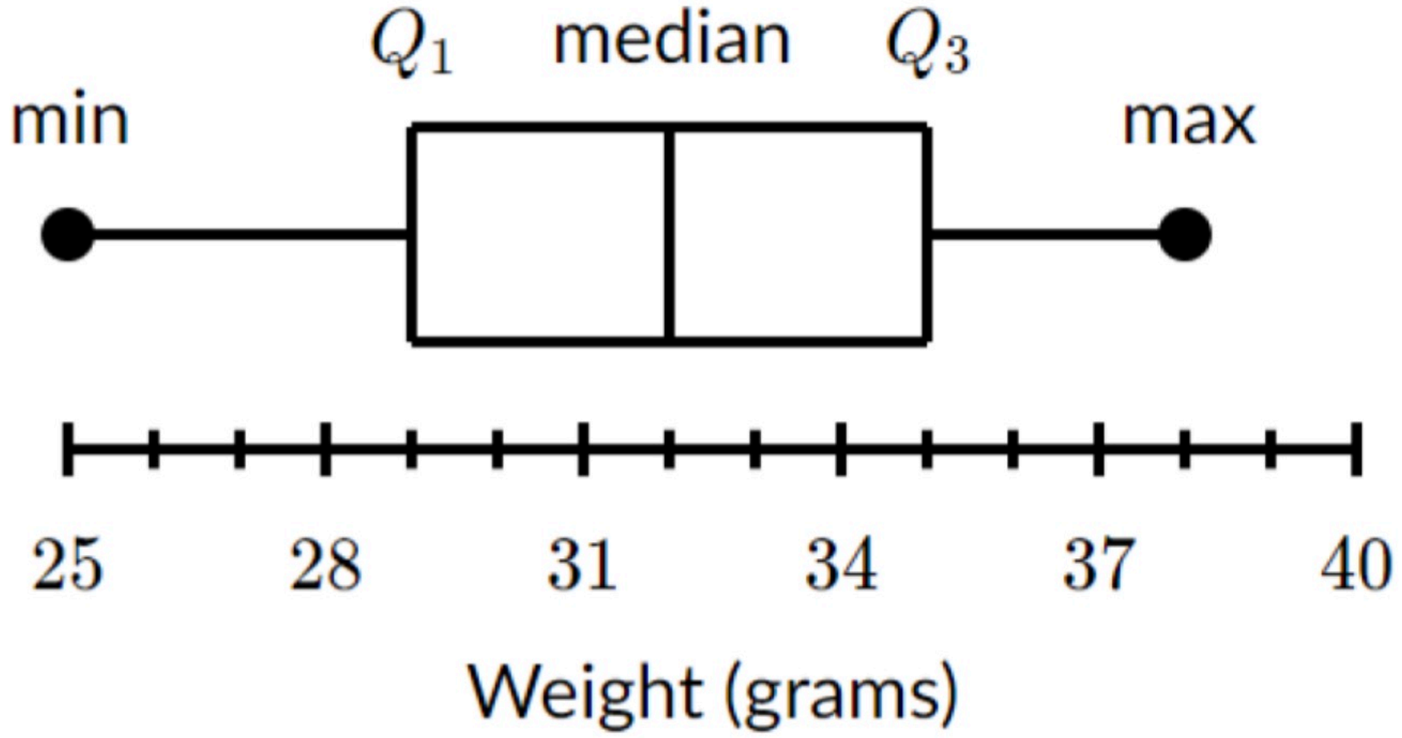


How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 3: Draw a whisker from *Q1* to the *min* and from *Q3* to the *max*.

25 28 29 29 30 34 35 35 37 38

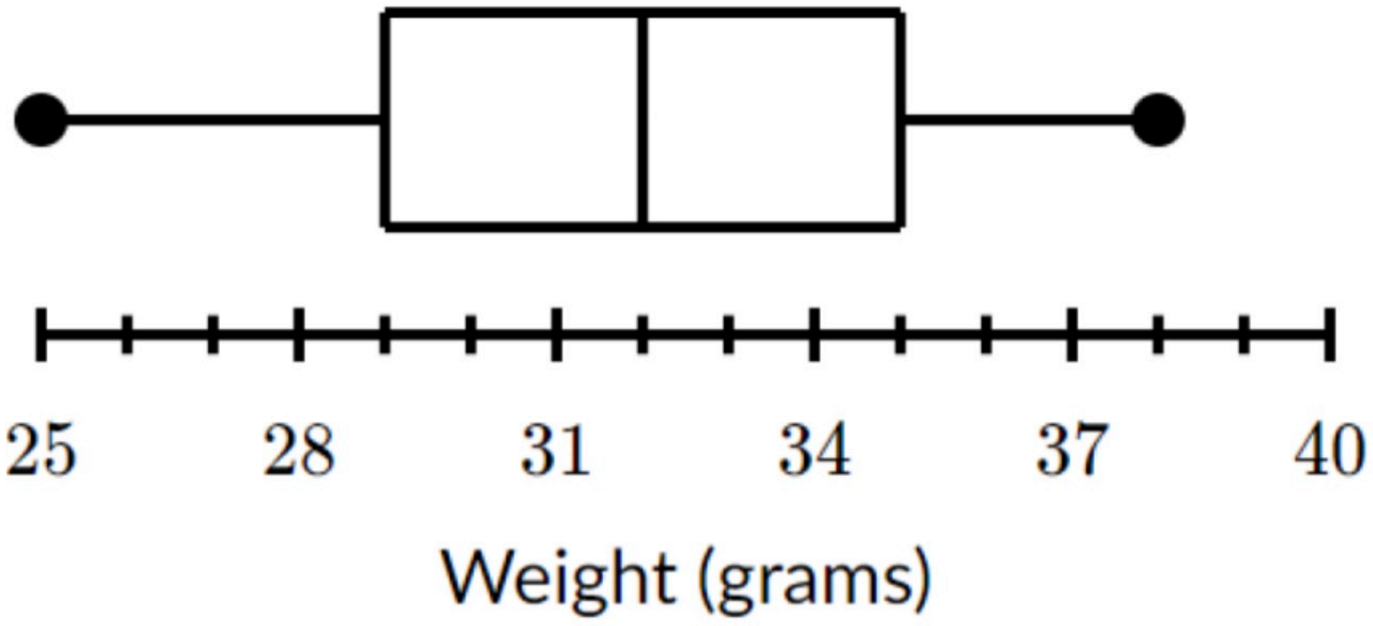


How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

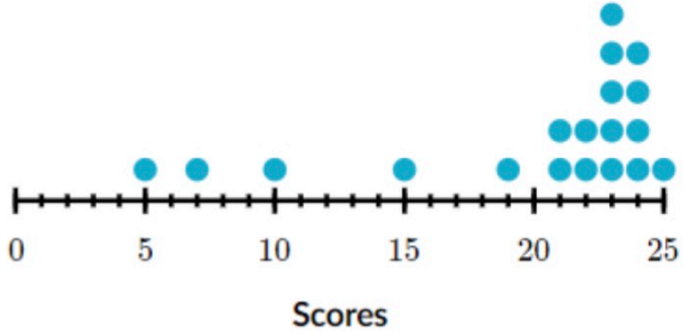
We don't need the labels on the final product:

25 28 29 29 30 34 35 35 37 38



Identifying Outliers

An **outlier** is a data point that lies outside the overall pattern in a distribution.



Question:

How many outliers do you see?

Students, enter a number!

No Number Response

You didn't answer this question

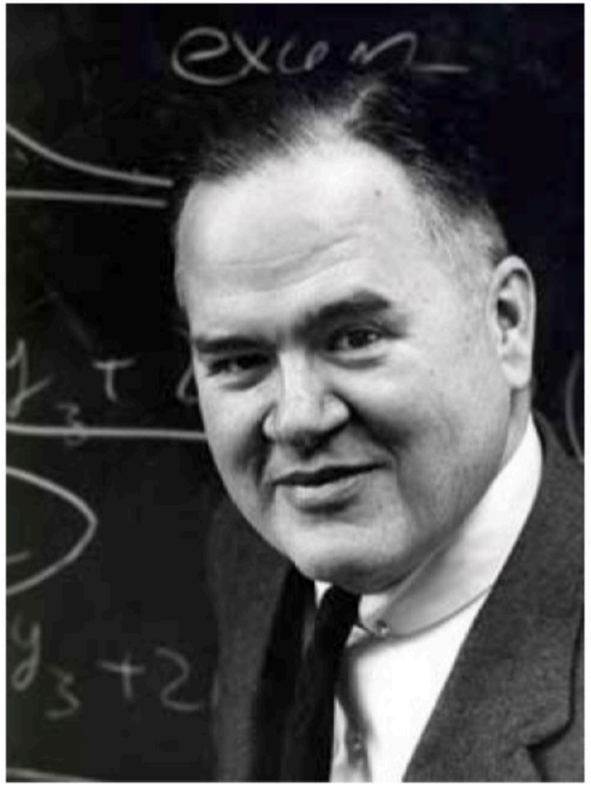
Identifying Outliers (1.5xIQR Rule)

1.5 IQR Rule: If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an outlier.

According to John Tukey,

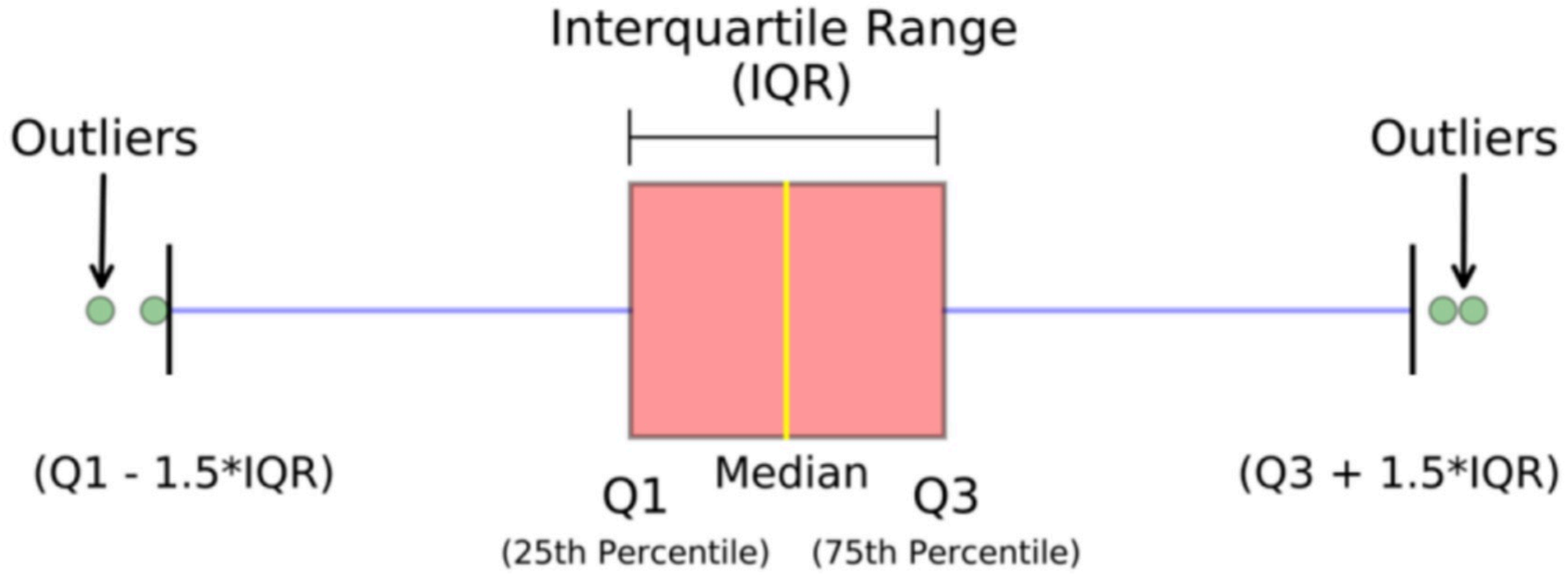
1 IQR seemed like too little and

2 IQRs seemed like too much.



Boxplot with 1.5xIQR Rule

1.5 IQR Rule: If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an outlier.



How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 1: Find the median, quartiles, and interquartile range.

Median = 23

Q1 = 20

Q3 = 23.5

IQR = Q3 - Q1 = 23.5 - 20 = 3.5

```
[33] np.median(a)

     23.0
```

```
[34] np.percentile(a, 25)

     20.0
```

```
[35] np.percentile(a, 75)

     23.5
```

How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Step 2: Calculate 1.5 x IQR below the first quartile and check for low outliers.

Q1 - 1.5 x IQR = 20 - (1.5 x 3.5)

= 14.75

Low Outliers: 5 7 10

How to make a Box Plot (1.5xIQR Rule)

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

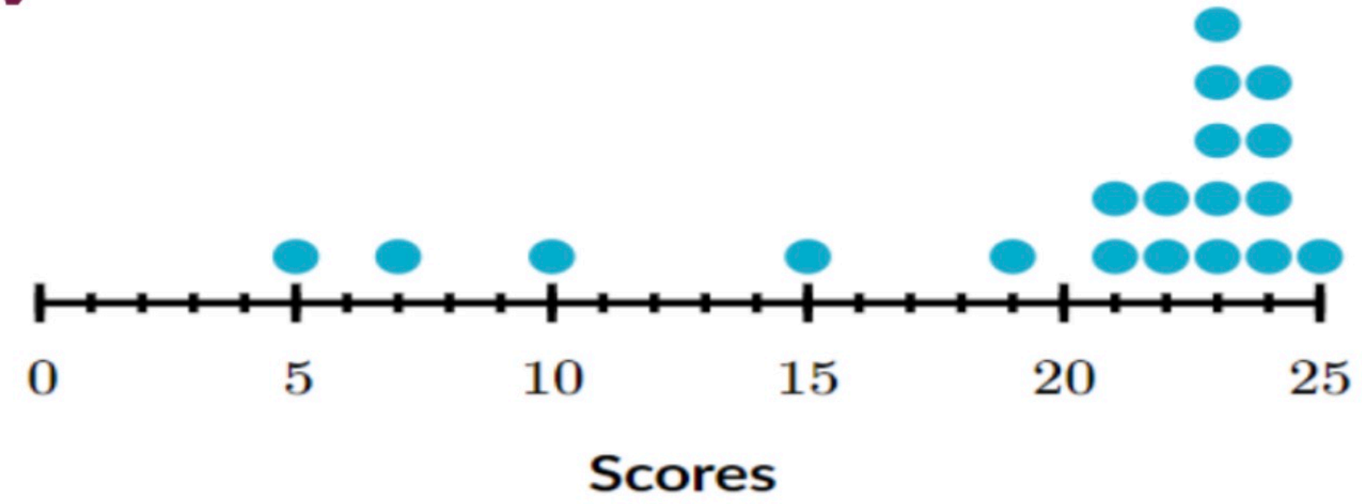
Step 3: Calculate 1.5 x IQR above the third quartile and check for high outliers.

$$Q3 + 1.5 \times IQR = 23.5 + (1.5 \times 5)$$
$$= 28.75$$

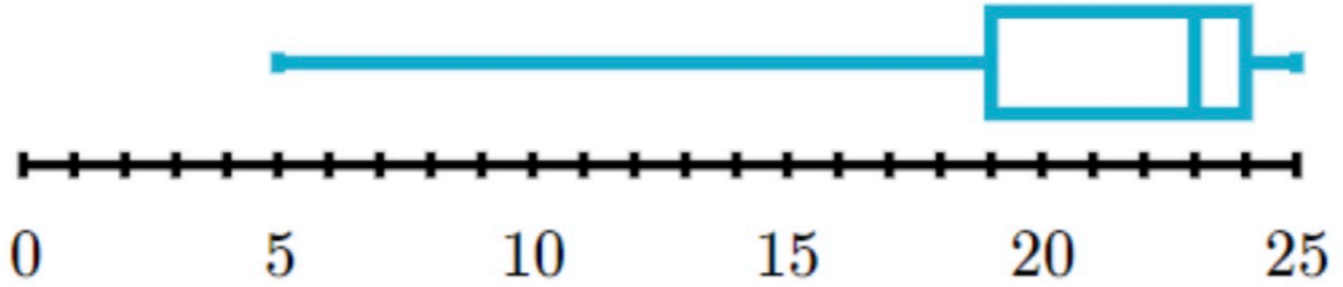
High Outliers: None

Box Plot *Overview*

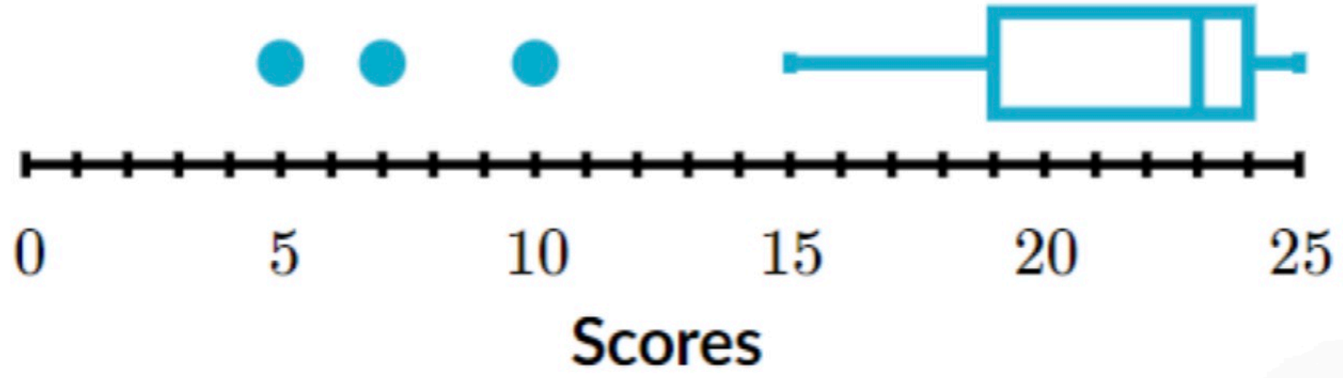
Data set



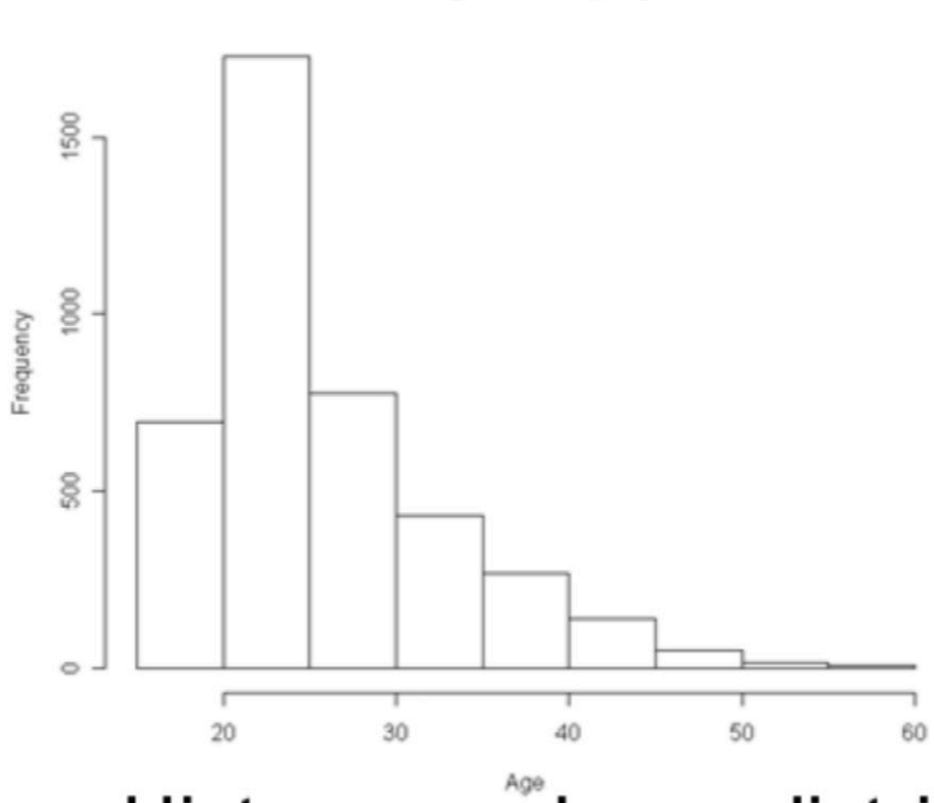
Min & Max

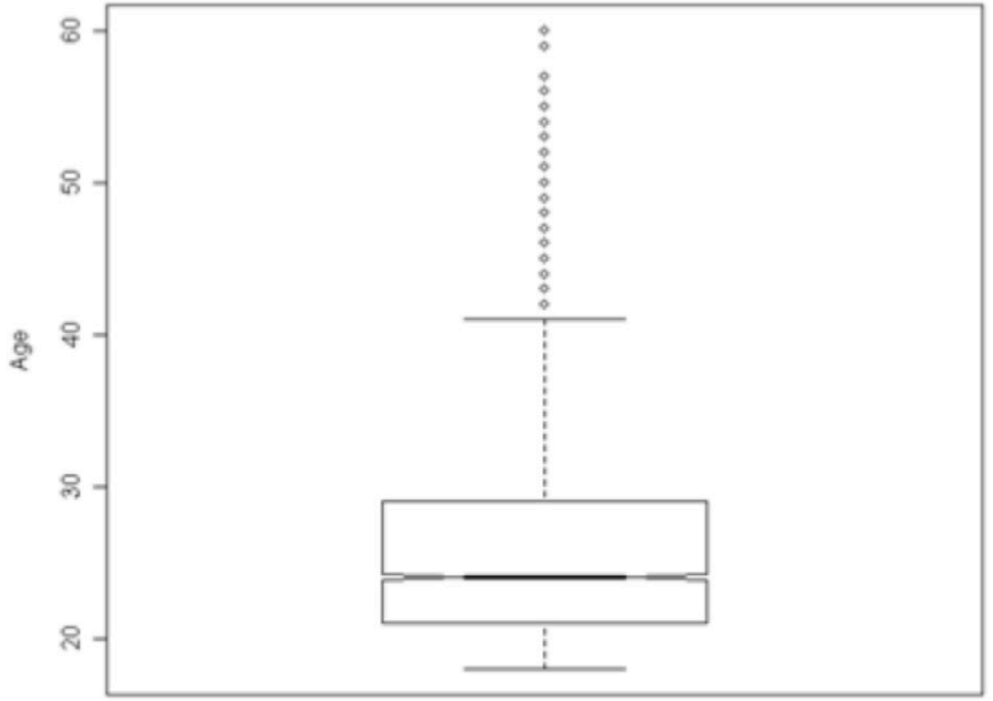


1.5 x IQR Rule



Box Plot vs. Histogram

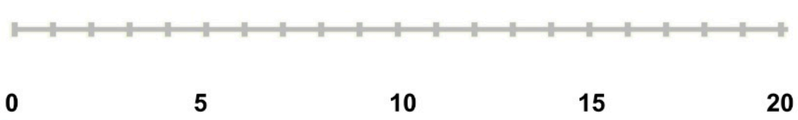




- Histogram shows distribution of the data in two dimensions – the boxplot is in one dimension
  - Histogram shows frequency of observations within ranges
  - Boxplot only shows summary statistics

Let's Practice *Draw Box Plot*

3	6	8	9	9	10	12	14	19



Students, draw anywhere on this slide!

No Drawing Response

You didn't answer this question

Let's Practice *Draw Box Plot*