DS-11/22 Statistics Session-2

DS-11/22 EU Statistics Session-2
Training Clarusway
Pear Deck - April 20, 2022 at 2:38PM

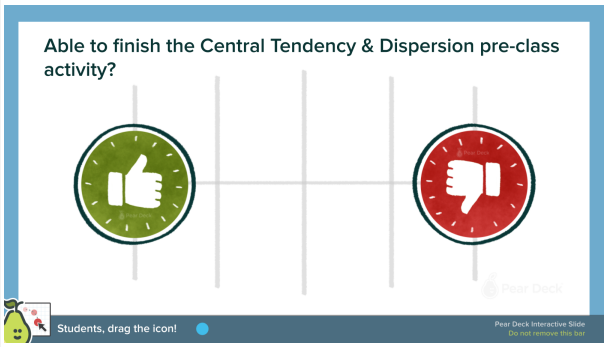
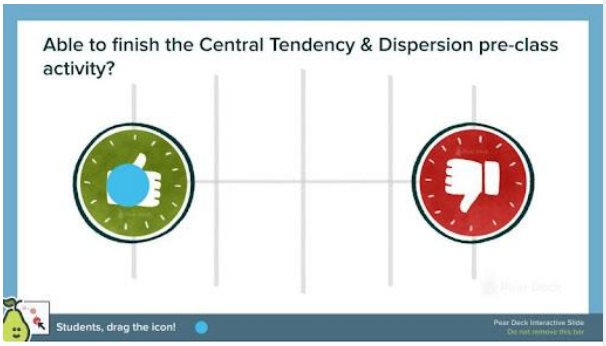
Part 1 - Summary

Use this space to summarize your thoughts on the lesson

Part 2 - Responses

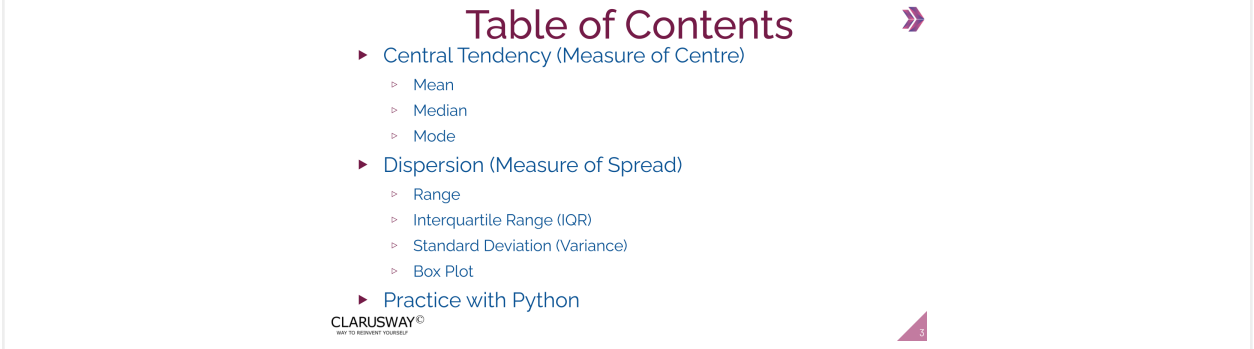
Slide 1

Use this space to take notes:

Slide 2	Your Response
	

Use this space to take notes:

Slide 3



Use this space to take notes:

Slide 4

1 Central Tendency (Measure of Centre)

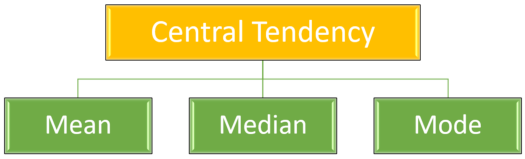
CLARUSWAY©

Use this space to take notes:

Slide 5

Central Tendency (Measure of Centre)

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

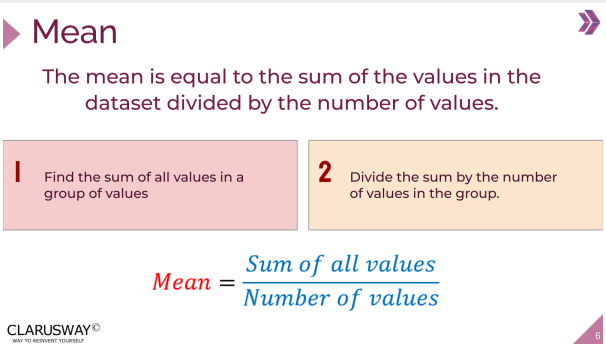


CLARUSWAY©

5

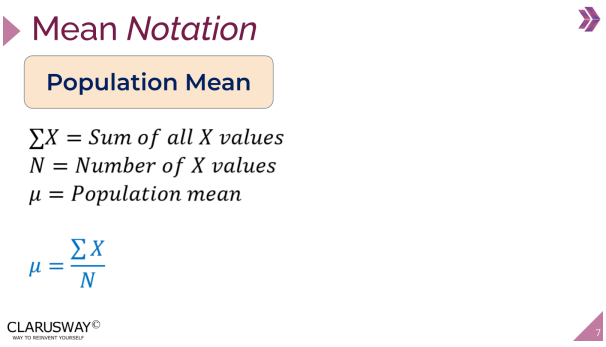
Use this space to take notes:

Slide 6



Use this space to take notes:

Slide 7



Use this space to take notes:

Slide 8

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

$\sum x = Sum\ of\ all\ x\ values$
$n = Number\ of\ x\ values$
$\bar{x} = Sample\ mean$

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

8

Use this space to take notes:

Slide 9

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

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

9

Use this space to take notes:

Slide 10

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

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

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

$$N = 5$$

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

CLARUSWAY©

Use this space to take notes:

Slide 11

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.

CLARUSWAY©

Use this space to take notes:

Slide 12

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

12

Use this space to take notes:

Slide 13

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

13

Use this space to take notes:

Slide 14

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

14

Use this space to take notes:

Slide 15

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.

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

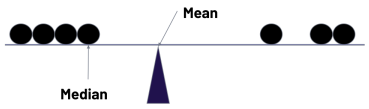
15

Use this space to take notes:

Slide 16

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.



CLARUSWAY©
WAY TO REINVENT YOURSELF

16

Use this space to take notes:

Slide 17

Mean vs. Median *Example*

$4000 $15.000 $20.000 $33.000 $1.800.000

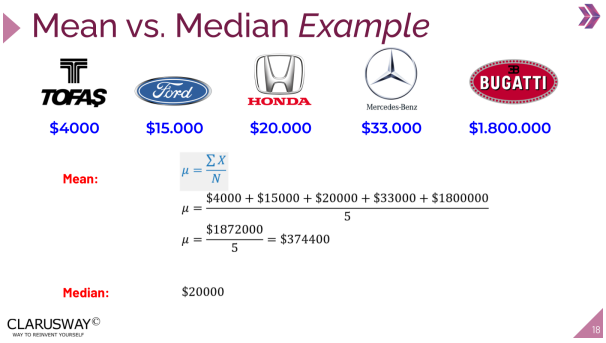
Find the mean and the median of car prices.

CLARUSWAY©
WAY TO REINVENT YOURSELF

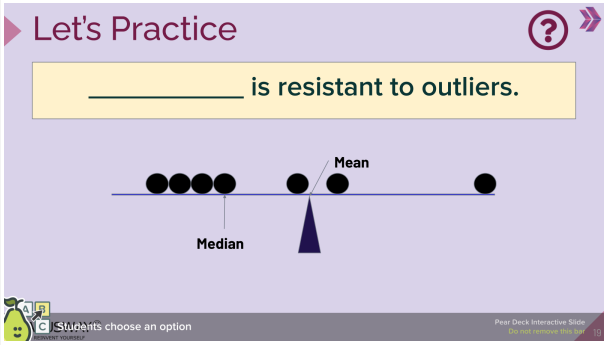
17

Use this space to take notes:

Slide 18

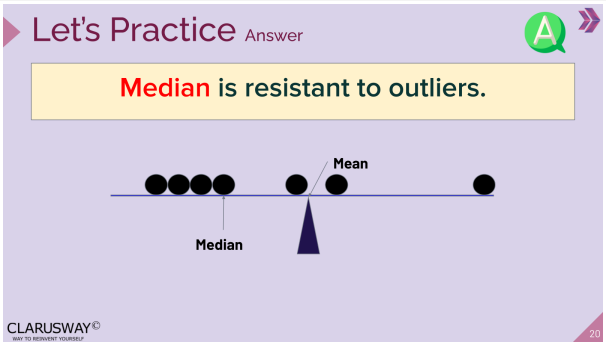


Use this space to take notes:

Slide 19	Your Response
	

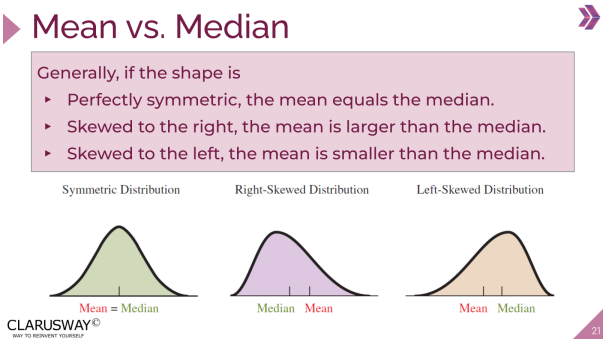
Use this space to take notes:

Slide 20



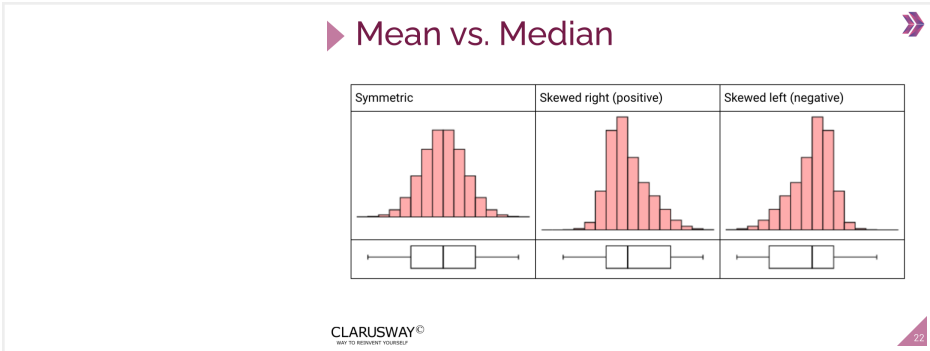
Use this space to take notes:

Slide 21

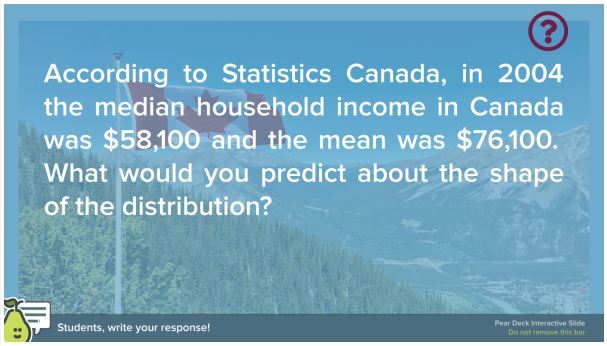


Use this space to take notes:

Slide 22

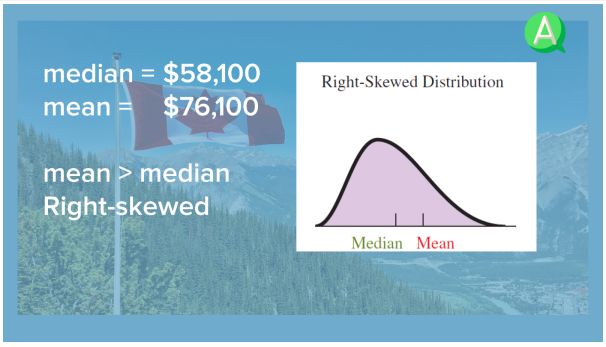


Use this space to take notes:

Slide 23	Your Response
	**Answer 1:** + skewed

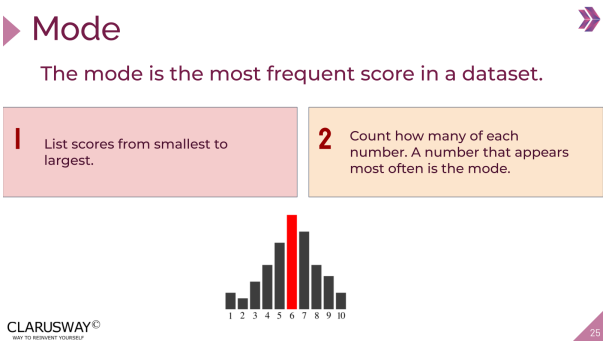
Use this space to take notes:

Slide 24



Use this space to take notes:

Slide 25



Use this space to take notes:

Slide 26

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

CLARUSWAY©

Use this space to take notes:

Slide 27

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

CLARUSWAY©

Use this space to take notes:

Slide 28

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 qualitative 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.

CLARUSWAY©
WAY TO REINVENT YOURSELF

28

Use this space to take notes:

Slide 29	Your Response

Add a constant

Question:

What happens to the mean, median, and mode when you add a constant to every value in x column?

	x	Add 10
Data	0	10
	20	30
	40	50
	40	50
	50	60
Mean		
Median		
Mode		

Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

29

You Chose

- **The mean, median, and mode increase by the same constant.**

Other Choices

- The mean and median increase by the same constant, the mode not.
- The mode and median increase by the same constant, the mean not.

Use this space to take notes:

Slide 30

Add a constant

Question:

What happens to the mean, median, and mode when you add a constant to every value in x column?

Answer:

The mean, median, and mode increase by the same constant.

	x	Add 10
Data	0	10
	20	30
	40	50
	40	50
	50	60
Mean	30	40
Median	40	50
Mode	40	50

CLARUSWAY©

30

Use this space to take notes:

Slide 31	Your Response
## Multiply by a constant **Question**: What happens to the mean, median, and mode when you multiply every value by a constant in x column? Students choose an option Pear Deck Interactive Slide Do not remove this bar 31	You Chose ● **The mean, median, and mode are multiplied by the same constant.** Other Choices ● The mean and median are multiplied by the same constant, the mode not. ● The mode and median are multiplied by the same constant, the mean not.

	x	Multiply by 10
Data	0	0
	20	200
	40	400
	40	400
	50	500
Mean		
Median		
Mode		

Use this space to take notes:

Slide 32

Multiply by a constant

Question:

What happens to the mean, median, and mode when you multiply every value by a constant in x column?

Answer:

The mean, median, and mode are also multiplied by the same constant.

	x	Multiply by 10
Data	0	0
	20	200
	40	400
	40	400
	50	500
Mean	30	300
Median	40	400
Mode	40	400

CLARUSWAY©
WAY TO REINVENT YOURSELF

32

Use this space to take notes:

Slide 33

2 Dispersion (Measure of Spread)

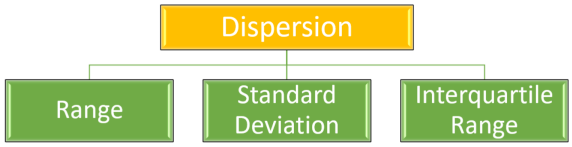
CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 34

Dispersion (Measure of Spread)

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



CLARUSWAY©
WAY TO REINVENT YOURSELF

34

Use this space to take notes:

Slide 35

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

35

Use this space to take notes:

Slide 36

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.

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 37

The Interquartile Range (IQR)

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

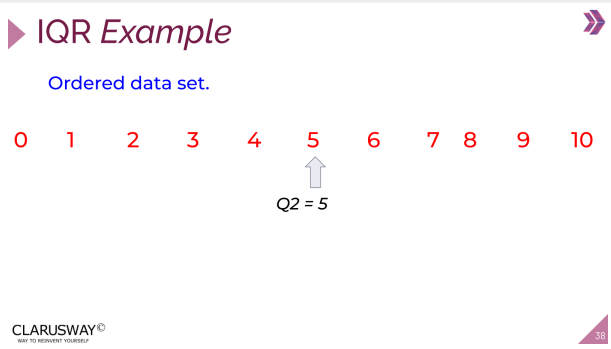
- Q2 is the median of the entire data set.
- Q1 is the median of the data below Q2.
- Q3 is the median of the data above Q2.

Interquartile Range = Q3 - Q1

CLARUSWAY©
WAY TO REINVENT YOURSELF

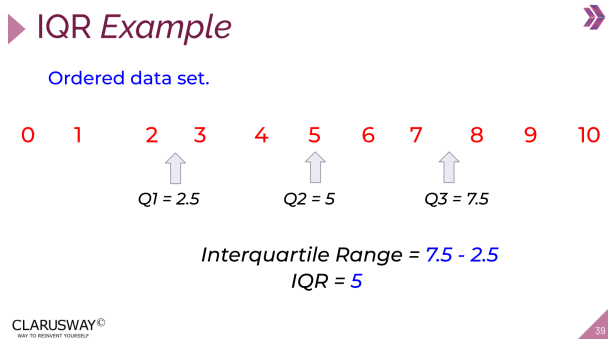
Use this space to take notes:

Slide 38



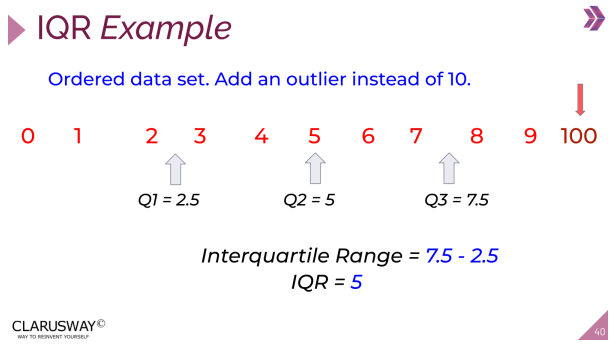
Use this space to take notes:

Slide 39

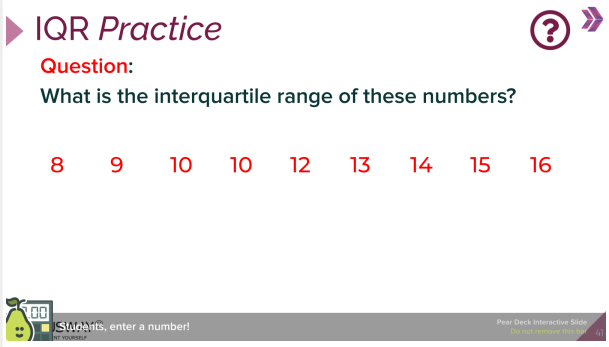


Use this space to take notes:

Slide 40



Use this space to take notes:

Slide 41	Your Response
	

Use this space to take notes:

Slide 42

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

CLARUSWAY©

42

Use this space to take notes:

Slide 43

Variance (Population)

Variance is the average squared deviation from the mean.

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

element, mean, number of elements

CLARUSWAY©

43

Use this space to take notes:

Slide 44

Variance (Sample)

Variance is the average squared deviation from the mean.

observation mean

sample variance → $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

number of observations

CLARUSWAY©

Use this space to take notes:

Slide 45

Variance *Example*

Find the Variance.

0 1 5 6

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

CLARUSWAY©

Use this space to take notes:

Slide 46

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

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

CLARUSWAY

Use this space to take notes:

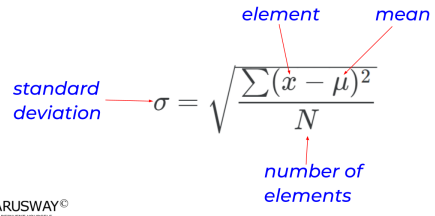
Slide 47	Your Response
## Variance **Question:** Would the variance of 10, 12, 17, 20, 25, 27, 42, and 45 be larger if the numbers represented a population or a sample? $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ $S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ Students choose an option Pear Deck Interactive Slide Do not remove this bar 	You Chose • **Sample** Other Choices • Population

Use this space to take notes:

Slide 48

Standard Deviation

Standard deviation is the square root of the variance.



CLARUSWAY©

Use this space to take notes:

Slide 49

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?

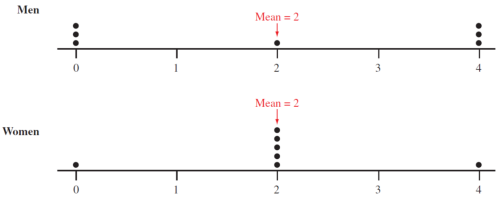
CLARUSWAY©

Use this space to take notes:

Slide 50

Standard Deviation *Example*

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



CLARUSWAY©
WAY TO REINVENT YOURSELF

50

Use this space to take notes:

Slide 51

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 tended 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 varied more around the mean.

CLARUSWAY©
WAY TO REINVENT YOURSELF

51

Use this space to take notes:

Slide 52	Your Response
	You Chose • **Some of them change.** Other Choices

Add a constant

Question:

What happens to measures of variability when you add a constant to each value in x column?

	x	Add 10
Data	0	10
	20	30
	40	50
	40	50
	50	60
Range		
IQR		
Std Dev		

Students choose an option

Pear Deck Interactive Slide

52

- They change.
- They stay the same.

Use this space to take notes:

Slide 53

Add a constant

Question:

What happens to measures of variability when you add a constant to each value in x column?

Answer:

All measures of variability stay the same.

CLARUSWAY©

	x	Add 10
Data	0	10
	20	30
	40	50
	40	50
	50	60
Range	50	50
IQR	20	20
Std Dev	17.9	17.9

53

Use this space to take notes:

Slide 54	Your Response

Multiply by a constant

Question:

What happens to measures of variability when you multiply each value by a constant in x column?

	x	Multiply by 10
Data	0	0
	20	200
	40	400
	40	400
	50	500
Range		
IQR		
Std Dev		

Students choose an option

Pear Deck Interactive Slide

54

Use this space to take notes:

Slide 55

Multiply by a constant

Question:

What happens to measures of variability when you multiply each value by a constant in x column?

Answer:

The range, IQR, and standard deviation are multiplied by the constant.

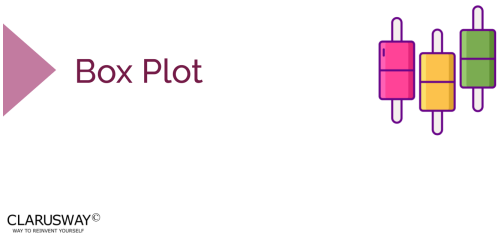
CLARUSWAY©

	x	Multiply by 10
Data	0	0
	20	200
	40	400
	40	400
	50	500
Range	50	500
IQR	20	200
Std Dev	17.9	179

55

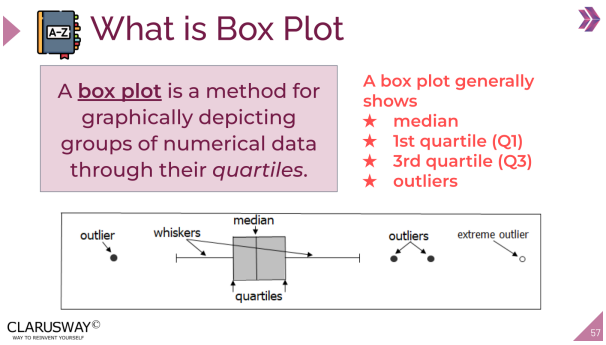
Use this space to take notes:

Slide 56



Use this space to take notes:

Slide 57



Use this space to take notes:

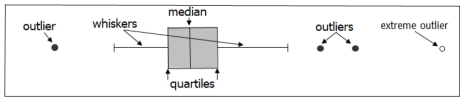
Slide 58

Box Plot

Boxplots show distribution in one dimension

- ★ Only useful for continuous variables
- ★ Good for comparing distributions of a continuous variable between categorical groups
- ★ Will not show multiple modes

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



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 59

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 60

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

60

Use this space to take notes:

Slide 61

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*

CLARUSWAY©
WAY TO REINVENT YOURSELF

61

Use this space to take notes:

Slide 62

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*

CLARUSWAY©

62

Use this space to take notes:

Slide 63

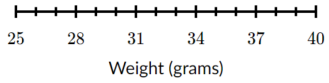
How to make a Box Plot (Min & Max)

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*



CLARUSWAY©

63

Use this space to take notes:

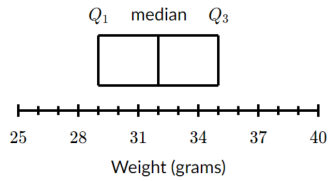
Slide 64

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



CLARUSWAY©

64

Use this space to take notes:

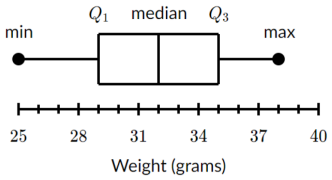
Slide 65

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



CLARUSWAY©

65

Use this space to take notes:

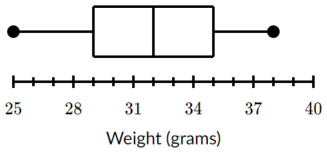
Slide 66

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



CLARUSWAY©

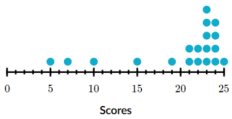
66

Use this space to take notes:

Slide 67	Your Response

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!

Pear Deck Interactive Slide
Do not remove this bar

67

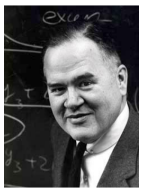
Use this space to take notes:

Slide 68

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.



CLARUSWAY©

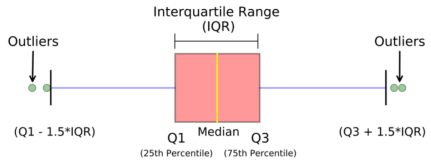
68

Use this space to take notes:

Slide 69

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.



CLARUSWAY©

69

Use this space to take notes:

Slide 70

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
```

CLARUSWAY©
WAY TO REINVENT YOURSELF

70

Use this space to take notes:

Slide 71

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.

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

71

Use this space to take notes:

Slide 72

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.

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

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

Q3 + 1.5 x IQR *= 23.5 + 1.5 x 5*

= 28.75

High Outliers: None

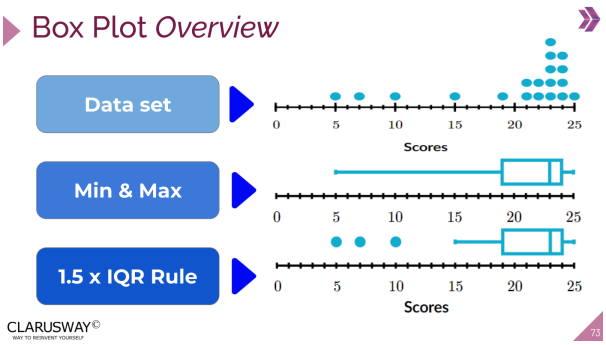
CLARUSWAY©

72

Use this space to take notes:

Slide 73

Box Plot *Overview*



CLARUSWAY©

73

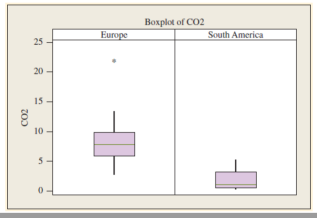
Use this space to take notes:

Slide 74	Your Response

Box Plot *Comparison*

Box plots are particularly useful when comparing many categories or populations.

Compare two continents in terms of CO2 emissions.



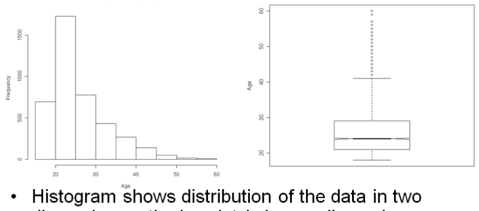
Students, write your response!

Pear Deck Interactive Slide
Do not remove this bar
74

Use this space to take notes:

Slide 75

Box Plot vs. Histogram



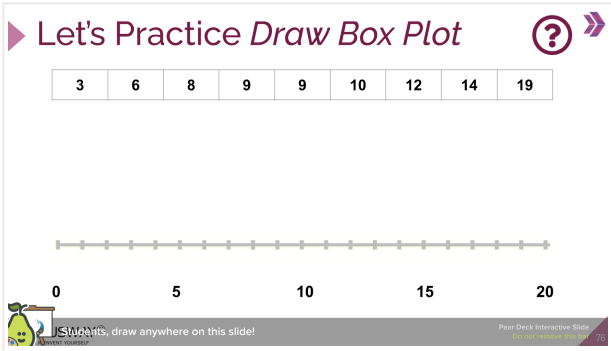
- Histogram shows distribution of the data in two dimensions – the boxplot is in one dimension
  - Histogram shows frequency of observations within ranges
  - Boxplot only shows summary statistics

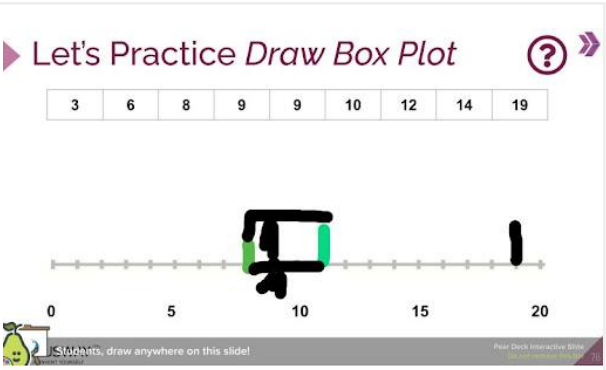
CLARUSWAY©
WAY TO REINVENT YOURSELF

75

Use this space to take notes:

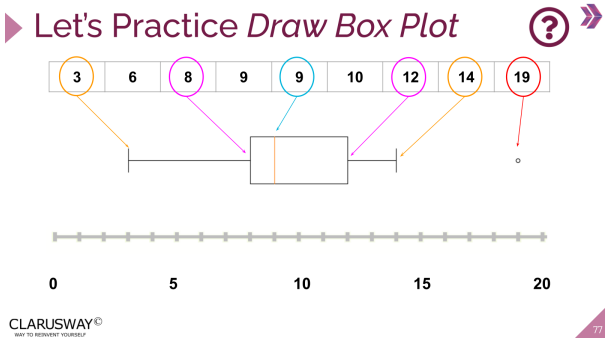
Slide 76	Your Response





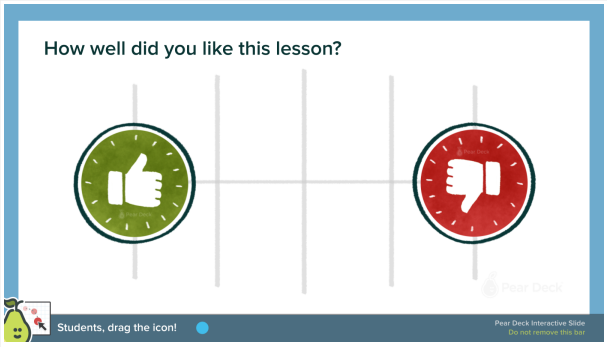
Use this space to take notes:

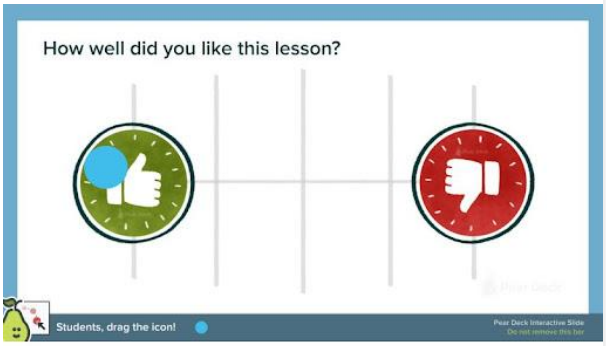
Slide 77



Use this space to take notes:

Slide 78	Your Response





Use this space to take notes:

Slide 79

Use this space to take notes: