DS-13/22 EU Session-2

DS-13 EU Statistics 2
Training Clarusway
Pear Deck - November 23, 2022 at 11:53AM

Part 1 - Summary

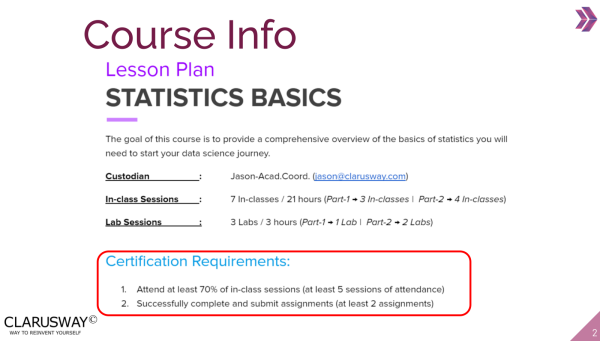
Use this space to summarize your thoughts on the lesson

Part 2 - Responses

Slide 1

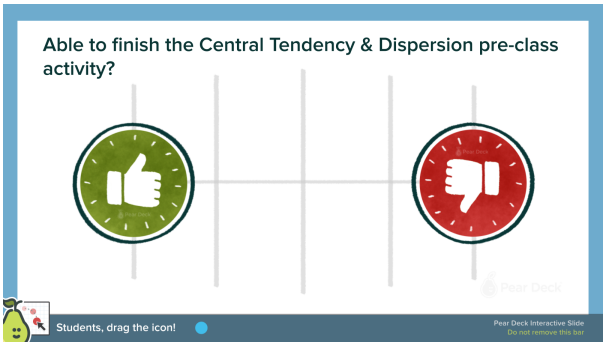
Use this space to take notes:

Slide 2



Use this space to take notes:

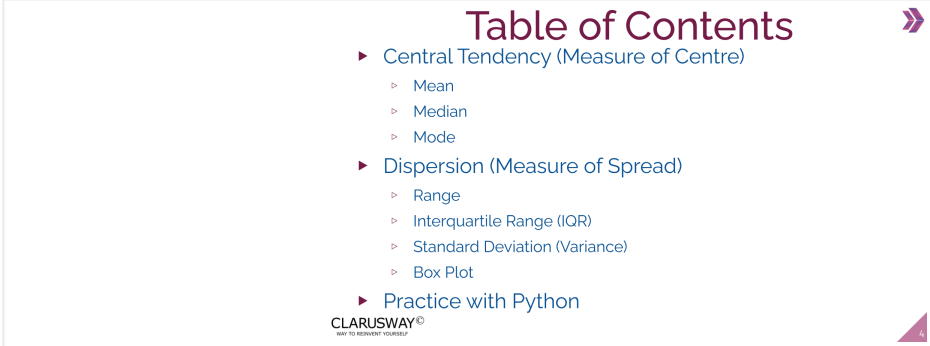
Slide 3



Your Response

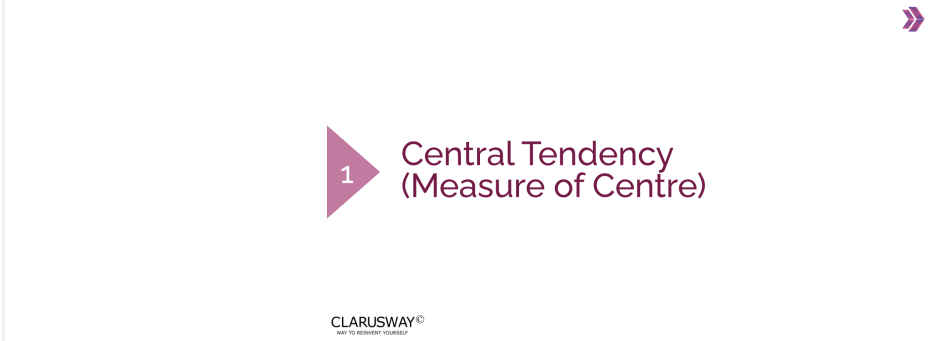
Use this space to take notes:

Slide 4



Use this space to take notes:

Slide 5

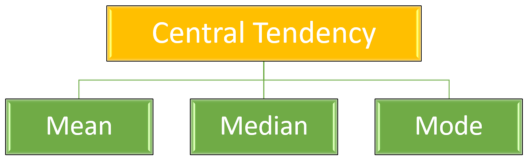


Use this space to take notes:

Slide 6

Central Tendency (Measure of Centre)

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



CLARUSWAY©

Use this space to take notes:

Slide 7

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.

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

CLARUSWAY©

Use this space to take notes:

Slide 8

Mean *Notation*

Population Mean

$\sum X = Sum\ of\ all\ X\ values$
$N = Number\ of\ X\ values$
$\mu = Population\ mean$

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 9

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

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

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

10

Use this space to take notes:

Slide 11

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©
WAY TO REINVENT YOURSELF

11

Use this space to take notes:

Slide 12

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©
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

CLARUSWAY©
WAY TO REINVENT YOURSELF

13

Use this space to take notes:

Slide 14

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

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

15

Use this space to take notes:

Slide 16

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

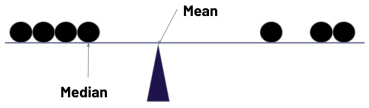
16

Use this space to take notes:

Slide 17

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

17

Use this space to take notes:

Slide 18

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

18

Use this space to take notes:

Slide 19

Selection of Imputation Method for Missing Values

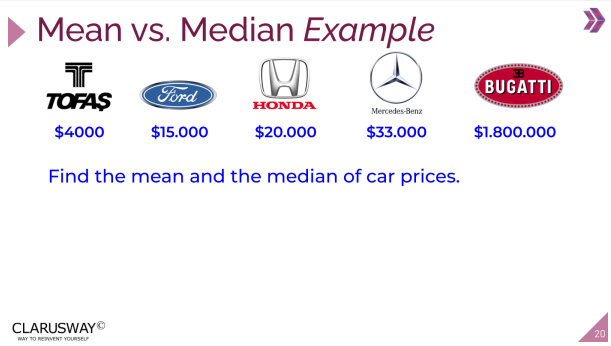
	First Name	Gender	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	97308.0	6.945	True	Marketing
1	Thomas	Male	61933.0	NaN	True	NaN
2	Jerry	Male	NaN	9.340	True	Finance
3	Dennis	n.a.	115163.0	10.125	False	Legal
4	NaN	Female	NaN	11.598	NaN	Finance
5	Angela	NaN	NaN	18.523	True	Engineering
6	Shawn	Male	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

CLARUSWAY©
WAY TO REINVENT YOURSELF

19

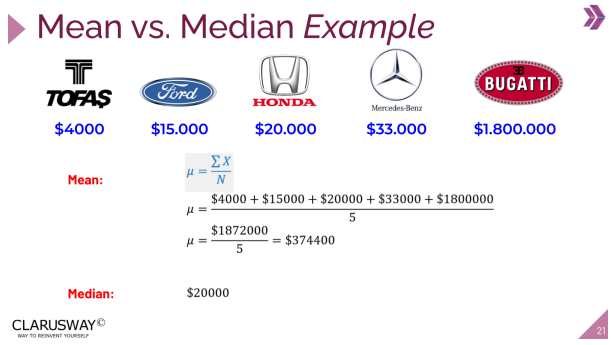
Use this space to take notes:

Slide 20



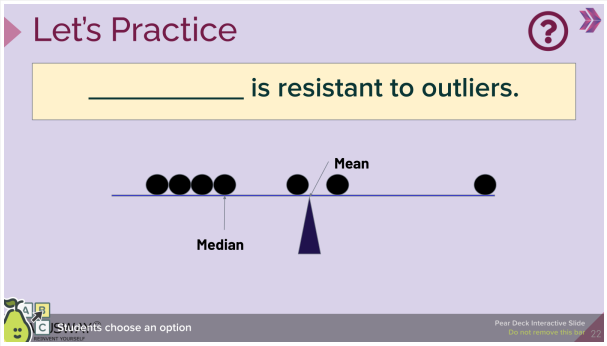
Use this space to take notes:

Slide 21

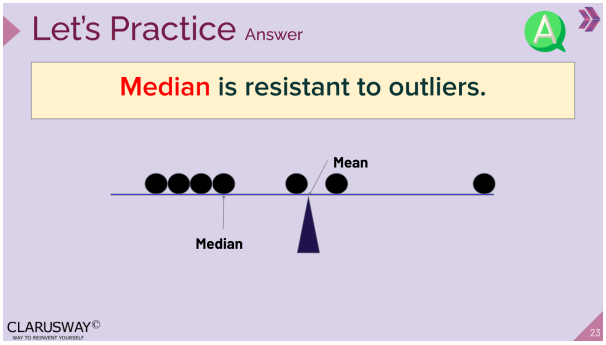


Use this space to take notes:

Slide 22	Your Response

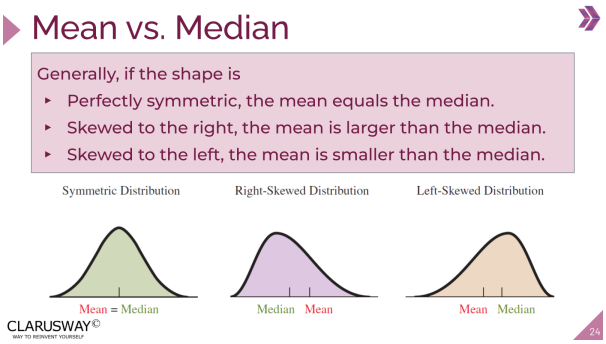
Slide 22	Your Response
	

Use this space to take notes:

Slide 23


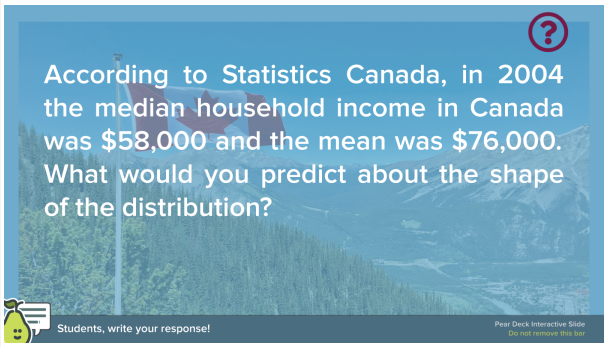
Use this space to take notes:

Slide 24



Use this space to take notes:

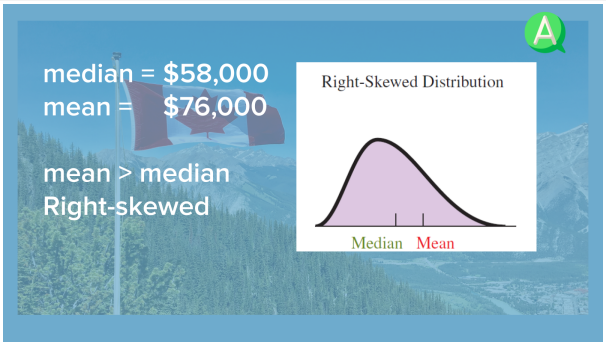
Slide 25



Your Response

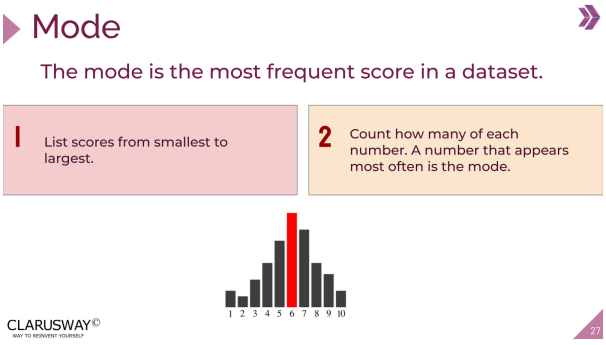
Use this space to take notes:

Slide 26



Use this space to take notes:

Slide 27



Use this space to take notes:

Slide 28

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©

28

Use this space to take notes:

Slide 29

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©

29

Use this space to take notes:

Slide 30

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.

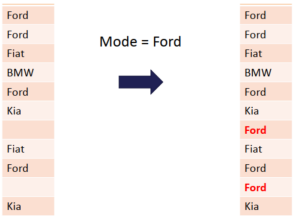
CLARUSWAY©
WAY TO REINVENT YOURSELF

30

Use this space to take notes:

Slide 31

Selection of Imputation Method for Missing Values



CLARUSWAY©
WAY TO REINVENT YOURSELF

31

Use this space to take notes:

Slide 32

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 33

2 Dispersion (Measure of Spread)

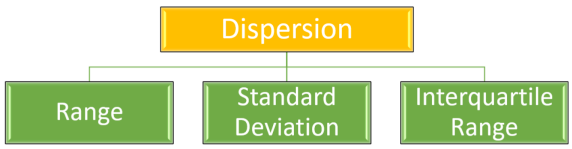
CLARUSWAY©
WAY TO REINVENT YOURSELF

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.



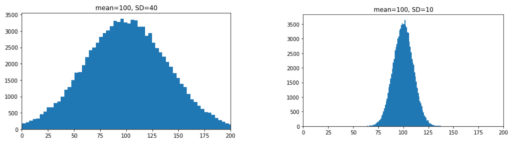
CLARUSWAY©
WAY TO REINVENT YOURSELF

34

Use this space to take notes:

Slide 35

Dispersion vs Central Tendency



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

CLARUSWAY©
WAY TO REINVENT YOURSELF

35

Use this space to take notes:

Slide 36

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

36

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.

- 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

37

Use this space to take notes:

Slide 38

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

38

Use this space to take notes:

Slide 39

IQR *Example*

Ordered data set.

0 1 2 3 4 5 6 7 8 9 10

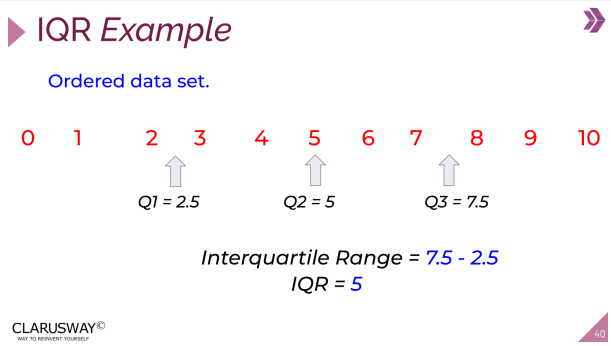
Q2 = 5

CLARUSWAY©
WAY TO REINVENT YOURSELF

39

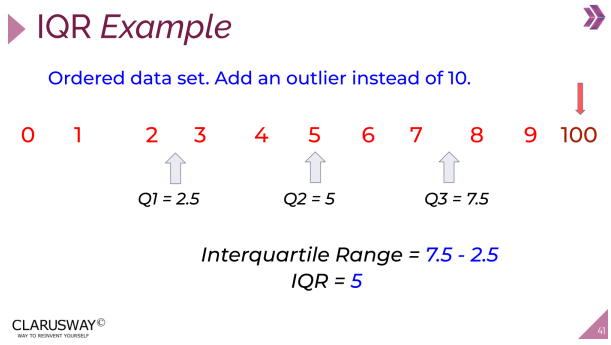
Use this space to take notes:

Slide 40



Use this space to take notes:

Slide 41



Use this space to take notes:

Slide 42	Your Response

Slide 42	Your Response
IQR *Practice* **Question**: What is the interquartile range of these numbers? 8 9 10 10 12 13 14 15 16 Students, enter a number!	

Use this space to take notes:

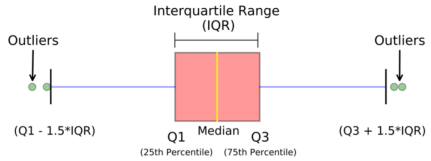
Slide 43
IQR Practice Answer **Question**: What is the interquartile range of these numbers? 8 9 (10) 10 12 13 (14) 15 16 IQR = Q3 - Q1 IQR = 14 - 10 IQR = 4 CLARUSWAY©

Use this space to take notes:

Slide 44

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.



CLARUSWAY©

44

Use this space to take notes:

Slide 45

Outlier Practice

Question:

Are there any outliers?

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

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

CLARUSWAY©

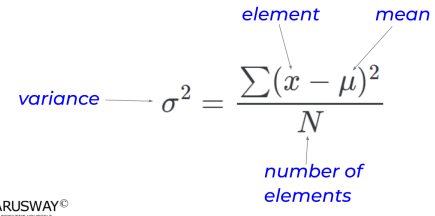
45

Use this space to take notes:

Slide 46

Variance (Population)

Variance is the average squared deviation from the mean.



CLARUSWAY©
WAY TO REINVENT YOURSELF

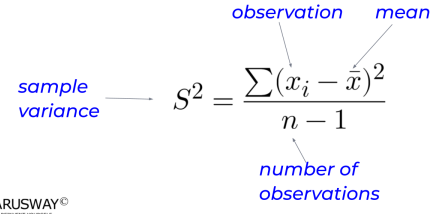
46

Use this space to take notes:

Slide 47

Variance (Sample)

Variance is the average squared deviation from the mean.



CLARUSWAY©
WAY TO REINVENT YOURSELF

47

Use this space to take notes:

Slide 48

Variance *Example*

Find the Variance.

0 1 5 6

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

CLARUSWAY©

48

Use this space to take notes:

Slide 49

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©

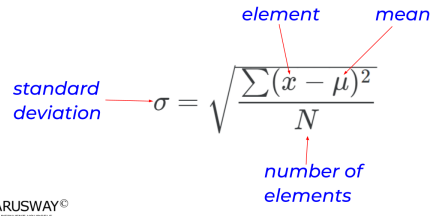
49

Use this space to take notes:

Slide 50

Standard Deviation

Standard deviation is the square root of the variance.



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 51

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?

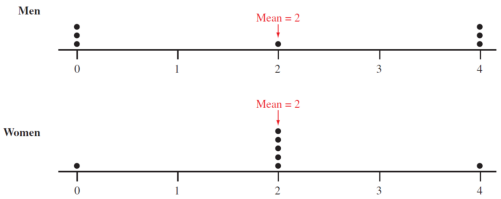
CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 52

Standard Deviation *Example*

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



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 53

Standard Deviation *Example*

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

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

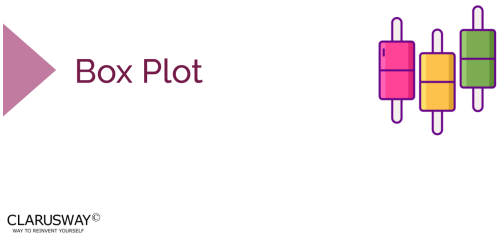
Women: $s = 1.2$

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.

CLARUSWAY©
WAY TO REINVENT YOURSELF

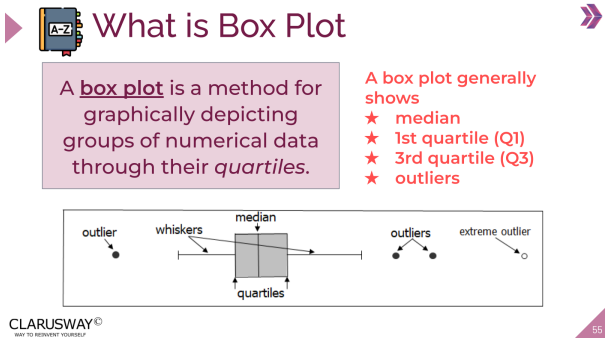
Use this space to take notes:

Slide 54



Use this space to take notes:

Slide 55



Use this space to take notes:

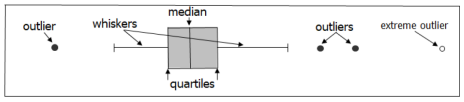
Slide 56

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**.



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 57

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 58

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©

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

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©

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

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©

60

Use this space to take notes:

Slide 61

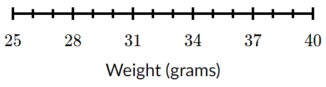
How to make a Box Plot (Min & Max)

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©

61

Use this space to take notes:

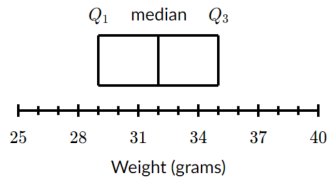
Slide 62

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©

62

Use this space to take notes:

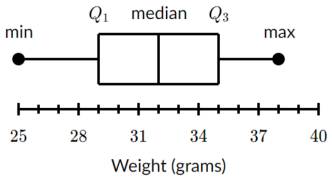
Slide 63

How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

Step 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©

63

Use this space to take notes:

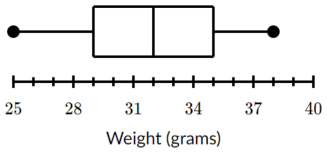
Slide 64

How to make a Box Plot (Min & Max)

Min	25
Q1	29
Median	32
Q3	35
Max	38

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

25 28 29 29 30 34 35 35 37 38



CLARUSWAY©

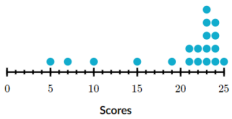
64

Use this space to take notes:

Slide 65

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

65

Your Response

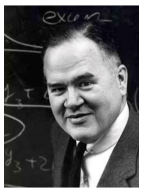
Use this space to take notes:

Slide 66

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©
WAY TO REINVENT YOURSELF

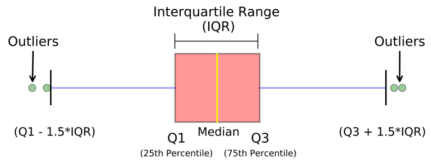
66

Use this space to take notes:

Slide 67

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©
WAY TO REINVENT YOURSELF

67

Use this space to take notes:

Slide 68

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

68

Use this space to take notes:

Slide 69

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

CLARUSWAY©
WAY TO REINVENT YOURSELF

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 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

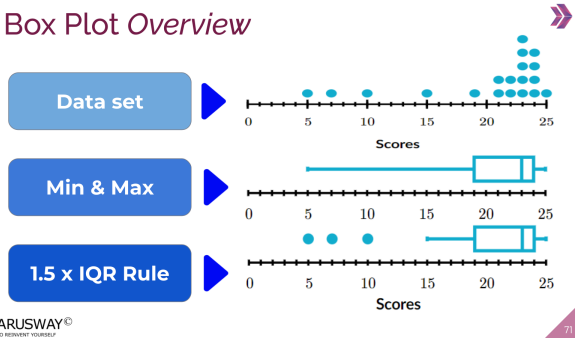
CLARUSWAY©

70

Use this space to take notes:

Slide 71

Box Plot *Overview*



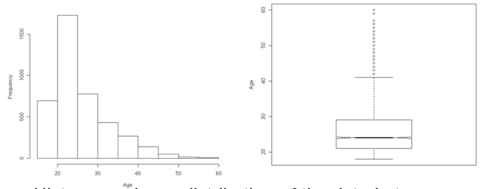
CLARUSWAY©

71

Use this space to take notes:

Slide 72

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

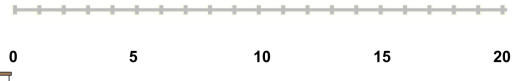
72

Use this space to take notes:

Slide 73

Let's Practice *Draw Box Plot*

3	6	8	9	9	10	12	14	19



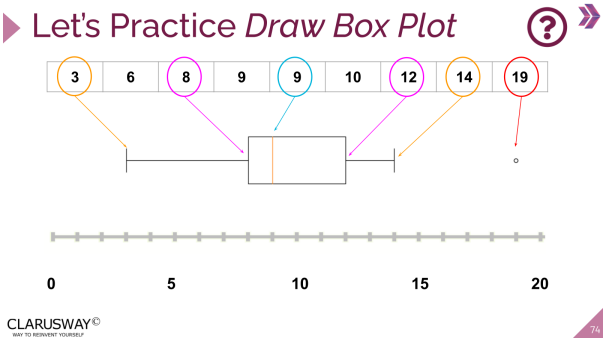
Students, draw anywhere on this slide!

Pear Deck Interactive Slide
Do not remove this bar 73

Your Response

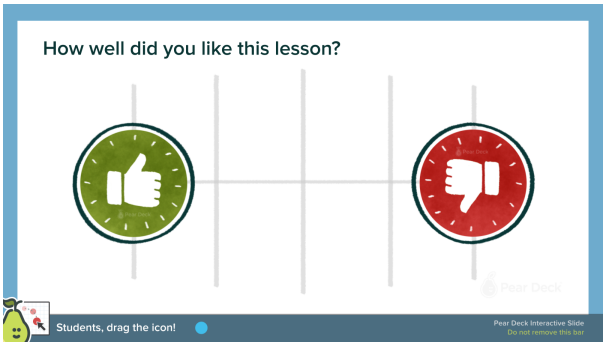
Use this space to take notes:

Slide 74



Use this space to take notes:

Slide 75



Your Response

Use this space to take notes:

Slide 76

THANKS!

Any questions?

You can find us at:

- richard@clarusway.com
- jason@clarusway.com

CLARUSWAY©

Use this space to take notes: