DS-12/22 EU Session-3

DS 12/22 Statistics Session-3
Training Clarusway
Pear Deck - July 28, 2022 at 0:55PM

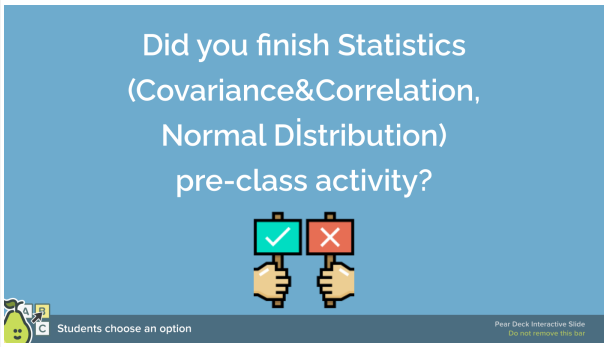
Part 1 - Summary

Use this space to summarize your thoughts on the lesson

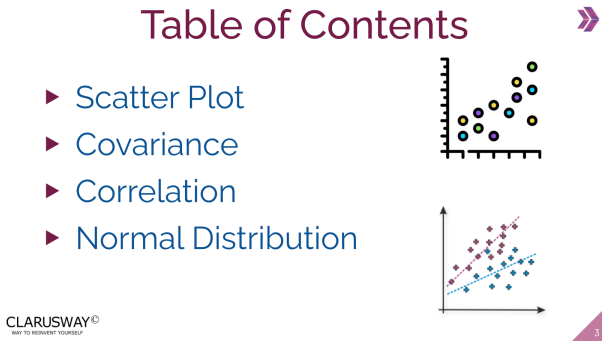
Part 2 - Responses

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 Scatter Plot

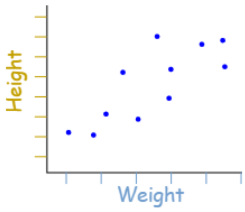
CLARUSWAY©

Use this space to take notes:

Slide 5

What is Scatter Plot

A **scatter plot** shows the relationship between two numerical variables.



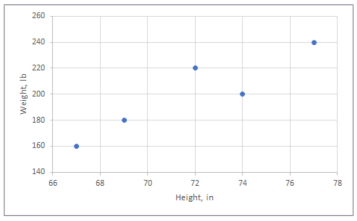
CLARUSWAY©

Use this space to take notes:

Slide 6

How to make a Scatter Plot

Height, in	Weight, lb
67	160
72	220
77	240
74	200
69	180



Scatter plots reveal patterns in bivariate data.

CLARUSWAY©
WAY TO REINVENT YOURSELF

6

Use this space to take notes:

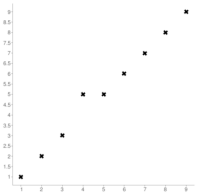
Slide 7

Patterns of Data in Scatter Plot

Linearity | Slope | Strength | Unusual Features

Linearity refers to whether a data pattern is linear (straight) or nonlinear (curved).

Linear

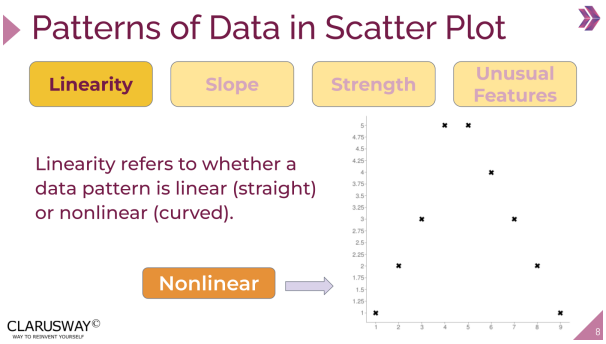


CLARUSWAY©
WAY TO REINVENT YOURSELF

7

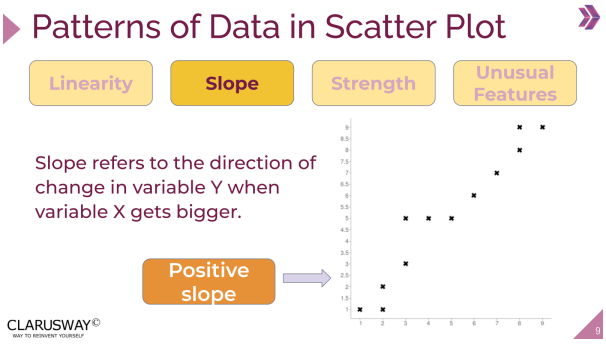
Use this space to take notes:

Slide 8



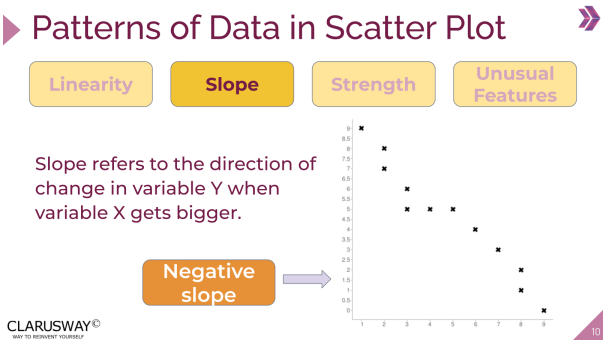
Use this space to take notes:

Slide 9



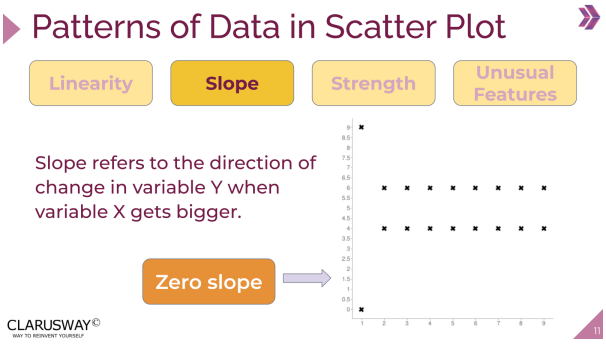
Use this space to take notes:

Slide 10



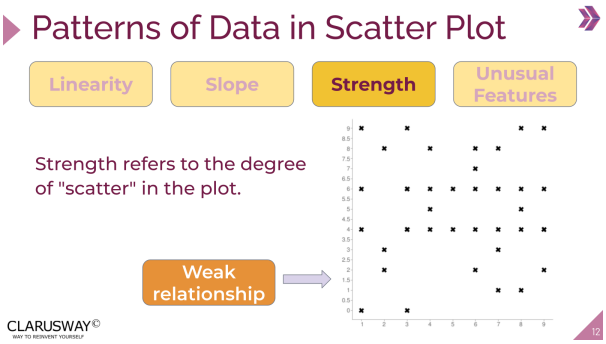
Use this space to take notes:

Slide 11



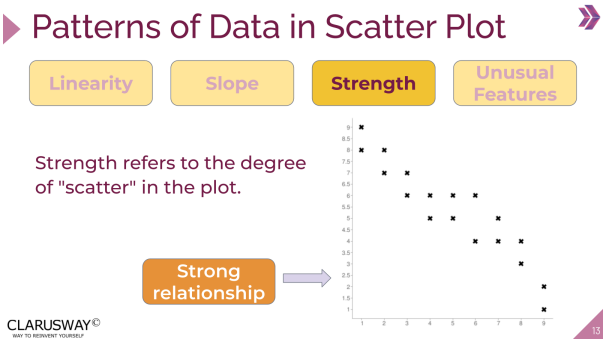
Use this space to take notes:

Slide 12



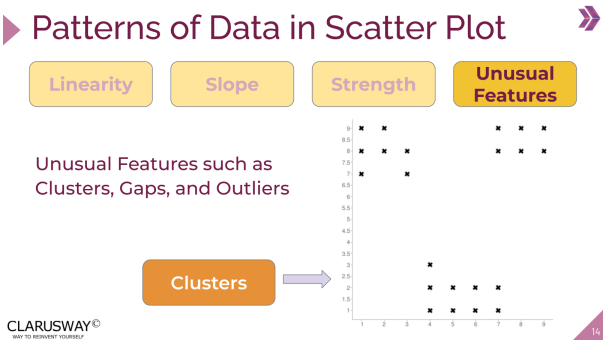
Use this space to take notes:

Slide 13



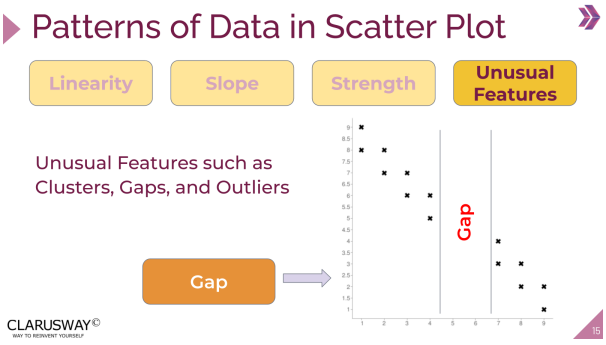
Use this space to take notes:

Slide 14



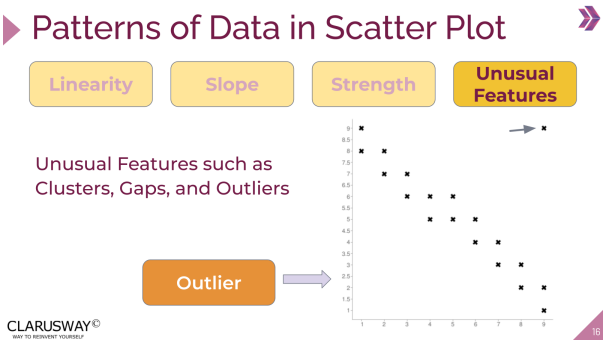
Use this space to take notes:

Slide 15



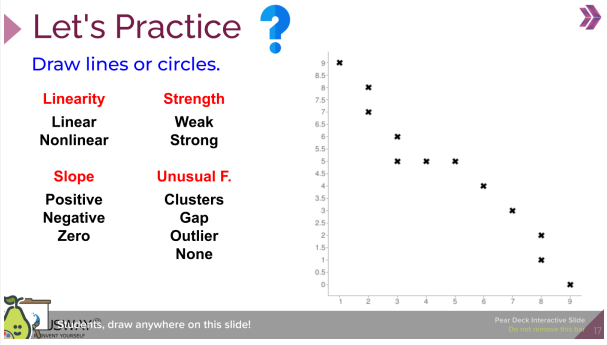
Use this space to take notes:

Slide 16



Use this space to take notes:

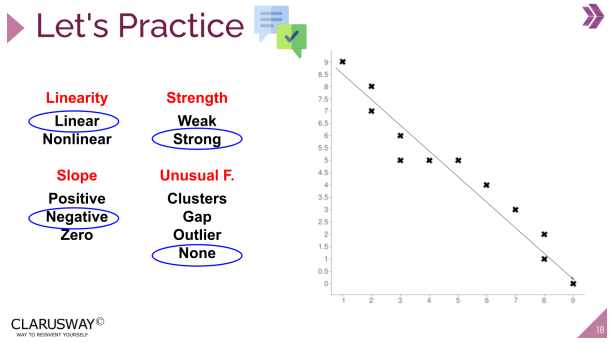
Slide 17



Your Response

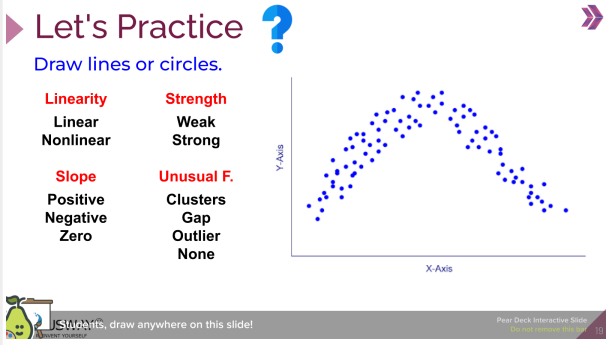
Use this space to take notes:

Slide 18



Use this space to take notes:

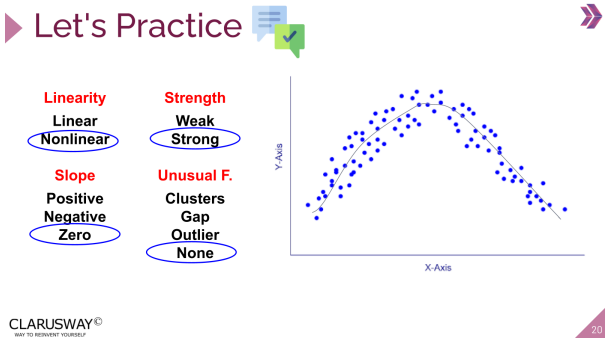
Slide 19



Your Response

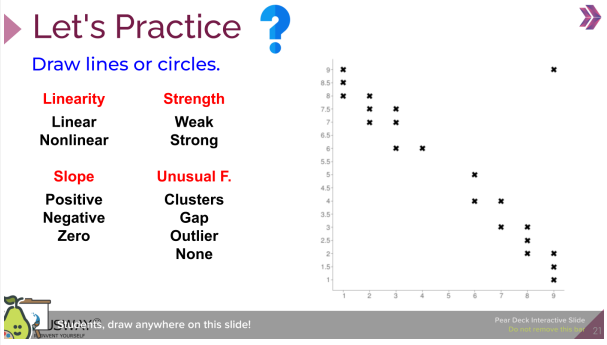
Use this space to take notes:

Slide 20



Use this space to take notes:

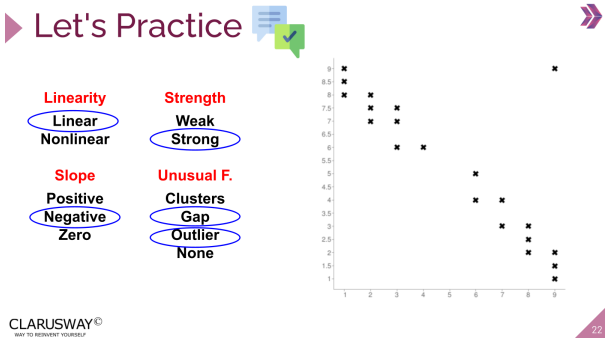
Slide 21



Your Response

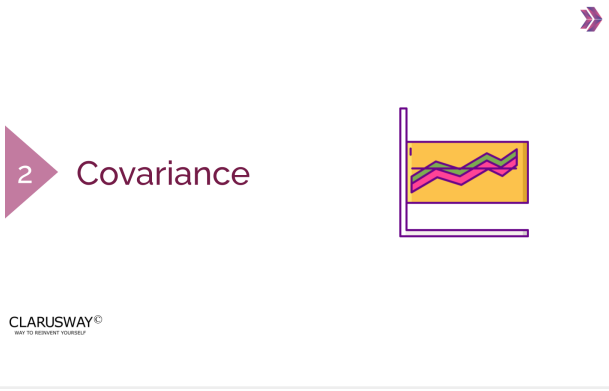
Use this space to take notes:

Slide 22



Use this space to take notes:

Slide 23



Use this space to take notes:

Slide 24

Intro

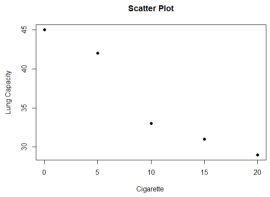
- ★ We are interested in how variables behave together.
- ★ For example, suppose we want to investigate the relationship between smoking and lung capacity.
- ★ We might ask a group of people about their smoking habits, and measure their lung capacities.

Use this space to take notes:

Slide 25

Scatter Plot

Cigarette (X)	Lung Capacity (Y)
0	45
5	42
10	33
15	31
20	29



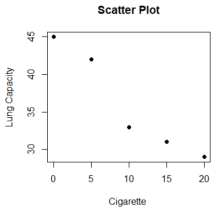
Scatter plots reveal patterns in bivariate data.

Use this space to take notes:

Slide 26

Scatter Plot

- We can see easily from the graph that as smoking goes up, lung capacity tends to go down.
- The two variables covary in opposite directions.



Use this space to take notes:

Slide 27

Covariance

- ★ **Covariance** is a measure of how much two random variables change together.
- ★ The covariance gives info on the strength of the linear relationship between two variables.
- ★ Since it is not a scaled metric, it is not useful to compare two different values of covariances.

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Use this space to take notes:

Slide 28

Covariance Interpretation

x and y tend to move in the same direction	Cov(x,y) > 0
x and y tend to move in opposite directions	Cov(x,y) < 0
x and y are independent	Cov(x,y) = 0

CLARUSWAY©
WAY TO REINVENT YOURSELF

28

Use this space to take notes:

Slide 29

Covariance *Example*

Table shows years of education (S) and hourly earnings in dollars (Y).

Observation	*S*	*Y*	*Observation*	*S*	*Y*
1	15	17.24	11	17	15.38
2	16	15.00	12	12	12.70
3	8	14.91	13	12	26.00
4	6	4.50	14	9	7.50
5	15	18.00	15	15	5.00
6	12	6.29	16	12	21.63
7	12	19.23	17	16	12.10
8	18	18.69	18	12	5.55
9	12	7.21	19	12	7.50
10	20	42.06	20	14	8.00

CLARUSWAY©
WAY TO REINVENT YOURSELF

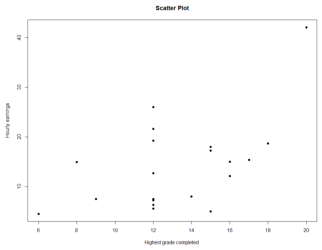
29

Use this space to take notes:

Slide 30

Covariance *Example*

- Figure shows the data plotted as a scatter plot.
- You can see that there is a weak positive association between the two variables.



CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 31

Covariance *Example*

Observation	S	Y	$(S - \bar{S})$	$(Y - \bar{Y})$	$(S - \bar{S})(Y - \bar{Y})$
1	15	17.24	1.75	3.016	5.277
2	16	15.00	2.75	0.775	2.133
3	8	14.91	-5.25	0.685	-3.599
4	6	4.5	-7.25	-9.725	70.503
5	15	18.00	1.75	3.776	6.607
6	12	6.29	-1.25	-7.935	9.918
7	12	19.23	-1.25	5.006	-6.257
8	18	18.69	4.75	4.466	21.211
9	12	7.21	-1.25	-7.015	8.768
10	20	42.06	6.75	27.836	187.890
11	17	15.38	3.75	1.156	4.333
12	12	12.70	-1.25	-1.525	1.906
13	12	26.00	-1.25	11.776	-14.719
14	9	7.50	-4.25	-6.725	28.579
15	15	5.00	1.75	-9.225	-16.143
16	12	21.63	-1.25	7.406	-9.257
17	16	12.10	2.75	-2.125	-5.842
18	12	5.55	-1.25	-8.675	10.843
19	12	7.50	-1.25	-6.725	8.406
20	14	8.00	0.75	-6.225	-4.668

$$\text{Cov}(x,y) = \frac{1}{n}[(x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Therefore, covariance is calculated as

305.888 / 19 = 16.0993

CLARUSWAY©
WAY TO REINVENT YOURSELF

Use this space to take notes:

Slide 32

3 Correlation

CLARUSWAY©

Use this space to take notes:

Slide 33

Correlation

The **correlation** summarizes the direction of the association between two quantitative variables and the strength of its linear (straight-line) trend.

Direction

- Positive : Move in same direction
- Negative : Move in opposite directions

Strength

- Weak : Widely spread
- Strong : Concentrated around a line

CLARUSWAY©

Use this space to take notes:

Slide 34

Correlation *Notation*

Sample Correlation	r
Population Correlation	ρ or R

CLARUSWAY©

Use this space to take notes:

Slide 35

Correlation *Formula*

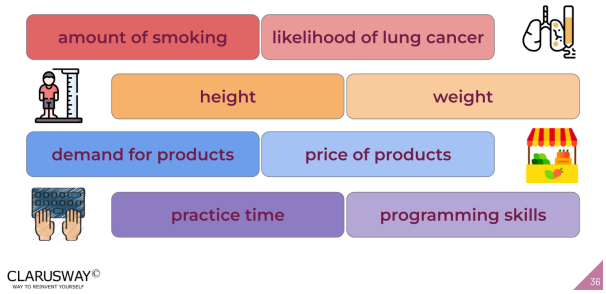
Sample Correlation	$r = \frac{Cov(x,y)}{s_X s_Y}$
Population Correlation	$\rho = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$

CLARUSWAY©

Use this space to take notes:

Slide 36

Examples



CLARUSWAY©

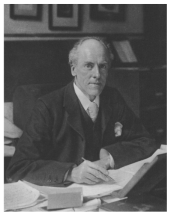
36

Use this space to take notes:

Slide 37

Correlation *Coefficient*

The correlation coefficient, ***r***, developed by *Karl Pearson* in the early 1900s.



CLARUSWAY©

37

Use this space to take notes:

Slide 38

Correlation *Coefficient*

Denoted by *r*, it takes values between *-1* and *+1*.

Strength

- Greater absolute value, stronger relationship
- Strongest relationship is correlation of -1 or 1
- Weakest relationship is correlation of zero.

Direction

- Sign of coefficient describes direction
- Positive sign: Variables move in same direction
- Negative sign: Move in opposite directions

CLARUSWAY©
WAY TO REINVENT YOURSELF

38

Use this space to take notes:

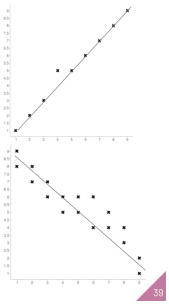
Slide 39

Linear Relationship

When the data points follow a roughly straight-line trend, the variables are said to have an approximately **linear relationship**.

On the line

Around the line



CLARUSWAY©
WAY TO REINVENT YOURSELF

39

Use this space to take notes:

Slide 40

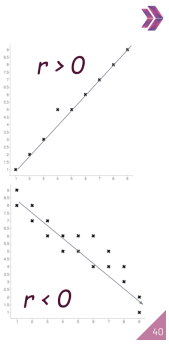
Direction

Slope determines the sign of the correlation coefficient.

When the slope of the line in the plot is negative, the correlation is negative; and vice versa.

Positive

Negative



CLARUSWAY©
WAY TO REINVENT YOURSELF

40

Use this space to take notes:

Slide 41

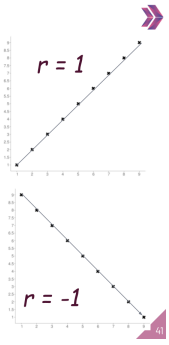
Strength

Strength is determined by scatter in data pattern.

The strongest correlations (r = 1.0 and r = -1.0) occur when data points fall exactly on a straight line.

Max positive correlation

Max negative correlation

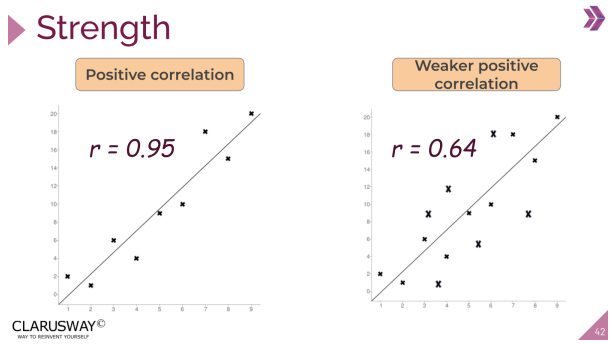


CLARUSWAY©
WAY TO REINVENT YOURSELF

41

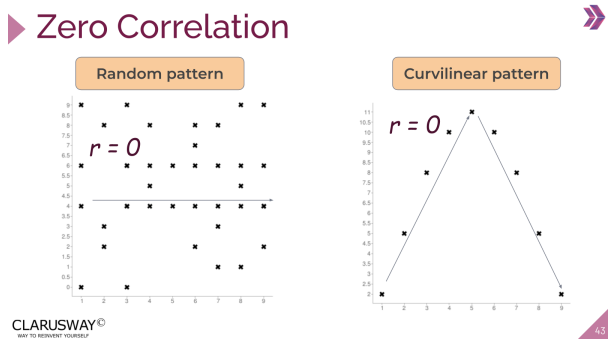
Use this space to take notes:

Slide 42



Use this space to take notes:

Slide 43



Use this space to take notes:

Slide 44	Your Response

Slide 44

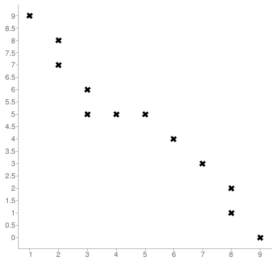
Patterns of Data in Scatter Plot

Draw lines or circles.

Linearity
Linear
Nonlinear

Strength
Weak
Strong

Slope
Positive
Negative
Zero



Students, draw anywhere on this slide!

Pear Deck Interactive Slide
Do not remove this bar 44

Your Response

Use this space to take notes:

Slide 45

Examples of Approximate r Values

Graph A (r = 1.0): perfect positive correlation between x and y

Graph B (r = -1.0): perfect negative correlation between x and y

Graph C (r = 0.6): a moderately positive relationship: y tends to increase as x increases, but not necessarily at the steady rate we observed in Graph A

Graph D (r = -0.4): a relatively weak negative relationship: the correlation coefficient is closer to zero, negative r value so y tends to decrease as x increases

Graph E (r = 0): no relationship between x and y

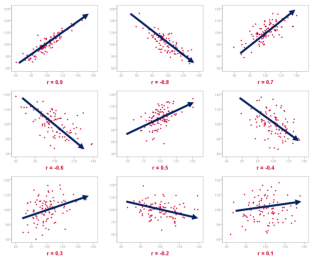
CLARUSWAY©

45

Use this space to take notes:

Slide 46

Examples of Approximate r Values



CLARUSWAY©
WAY TO REINVENT YOURSELF

46

Use this space to take notes:

Slide 47

Let's Practice

Which pair of variables exhibits the strongest linear relationship?
Which pair of variables exhibits the weakest linear relationship?

Pair	Variable 1	Variable 2	Correlation
A	Internet Users	Facebook Users	0.507
B	Internet Users	Broadband Subscribers	0.949
C	Internet Users	Population	0.744
D	Facebook Users	Broadband Subscribers	0.619
E	Facebook Users	Population	0.097
F	Broadband Subscribers	Population	0.533

Students choose an option

Pear Deck Interactive Slide

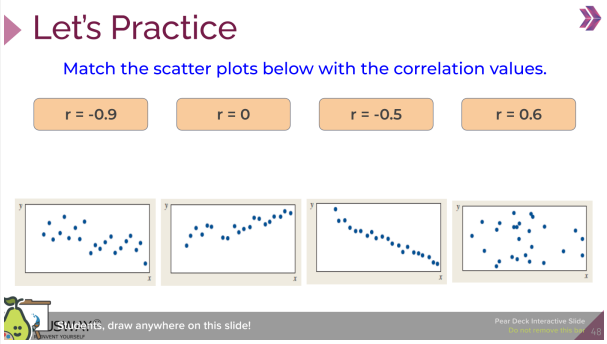
47

Your Response

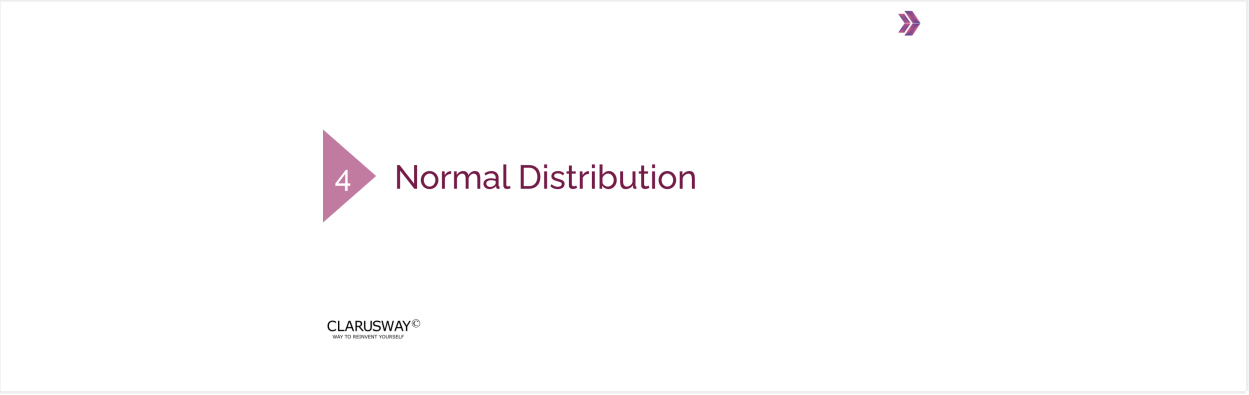
Use this space to take notes:

Slide 48

Your Response

Slide 48	Your Response
	

Use this space to take notes:

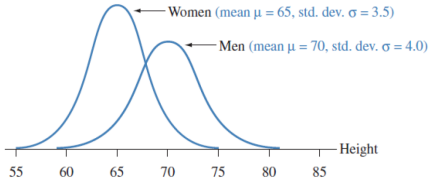
Slide 49


Use this space to take notes:

Slide 50

Normal Distribution

The normal distribution is symmetric, bell-shaped, and characterized by its mean **μ** and standard deviation **σ**.



Normal Distributions for Women's Height and Men's Height.

CLARUSWAY©

50

Use this space to take notes:

Slide 51

Normal Distribution

- ★ Symmetric around their mean.
- ★ The mean, median, and mode are equal.
- ★ The area under the normal curve is equal to 1.0.
- ★ Denser in the center and less dense in the tails.
- ★ Defined by two parameters: μ and σ. $\mathcal{N}(\mu, \sigma^2)$
- ★ Bell-shaped curve or Gaussian curve.

CLARUSWAY©

51

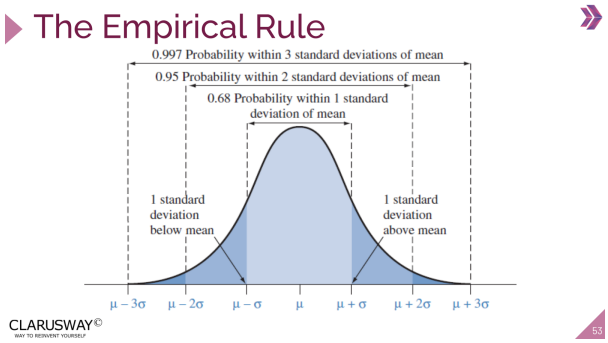
Use this space to take notes:

Slide 52	Your Response

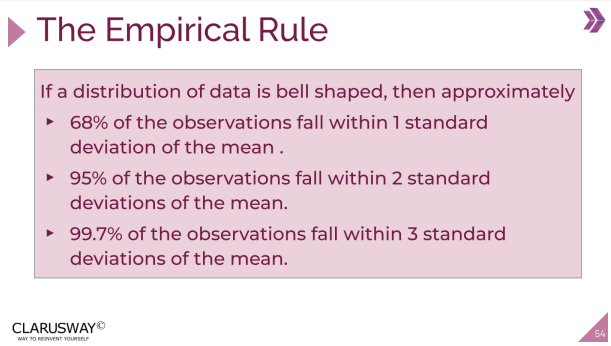
Slide 52	Your Response
	

Use this space to take notes:

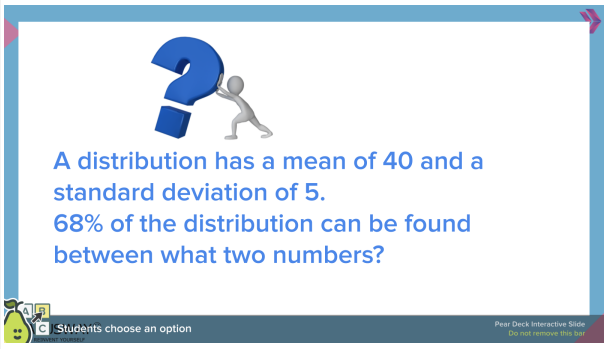
Slide 53



Use this space to take notes:

Slide 54


Use this space to take notes:

Slide 55	Your Response
	

Use this space to take notes:

Slide 56	Your Response

Slide 56

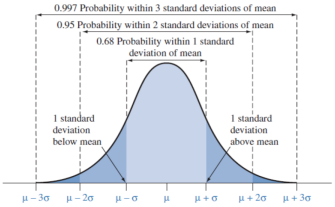
Your Response

The Empirical Rule

Question:

About what percentage (%) would fall more than 2 standard deviations from the mean?

Enter a number between 0-100



Students, enter a number!

Pear Deck Interactive Slide

55

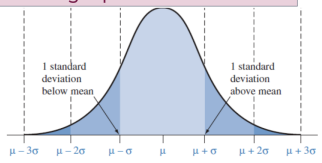
Use this space to take notes:

Slide 57

Finding Standard Score (z-score)

- The normal random variable of a standard normal distribution is called a standard score or a z-score.
- Every normal random variable x can be transformed into a z score via the following equation:

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$



CLARUSWAY©

57

Use this space to take notes:

Slide 58

Finding z-score

- SAT scores are approximately normally distributed with mean μ = 500 and standard deviation σ = 100.
- If your SAT score from one of the three components was x = 650, how many standard deviations from the mean was it?

$$z = \frac{x - \mu}{\sigma} = \frac{650 - 500}{100} = 1.50$$

CLARUSWAY©
WAY TO REINVENT YOURSELF

58

Use this space to take notes:

Slide 59

Finding z-score

- SAT scores are approximately normally distributed with mean μ = 500 and standard deviation σ = 100.
- What percentage of SAT scores was higher than yours?

Table 4 Normal Curve Areas
Standard normal probability in right-hand tail
(for negative values of z, areas are found by symmetry)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	Second decimal place of z									
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

CLARUSWAY©
WAY TO REINVENT YOURSELF

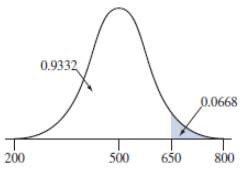
59

Use this space to take notes:

Slide 60

Finding z-score

- SAT scores are approximately normally distributed with mean μ = 500 and standard deviation σ = 100.
- What percentage of SAT scores was higher than yours?



CLARUSWAY©
WAY TO REINVENT YOURSELF

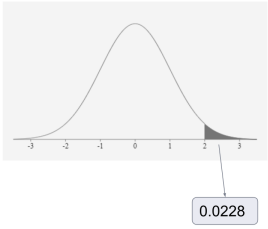
60

Use this space to take notes:

Slide 61

Normal Distribution - *Example*

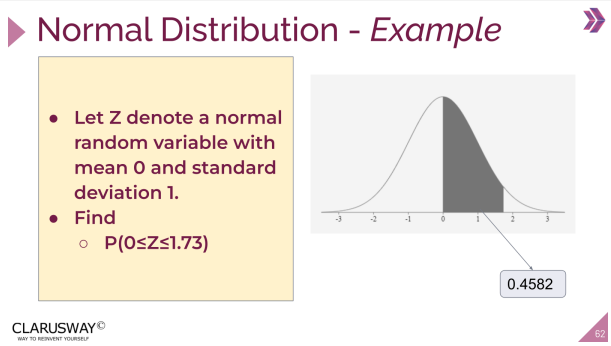
- Let Z denote a normal random variable with mean 0 and standard deviation 1.
- Find
  - P(Z>2)



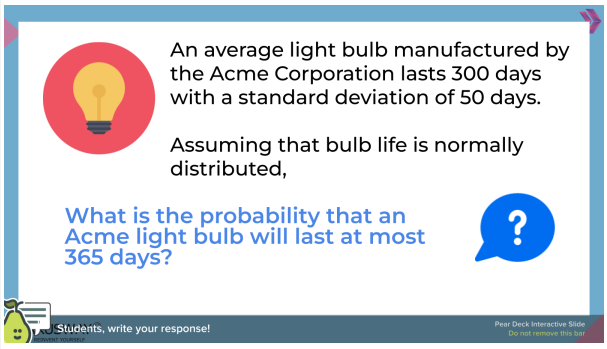
CLARUSWAY©
WAY TO REINVENT YOURSELF

61

Use this space to take notes:

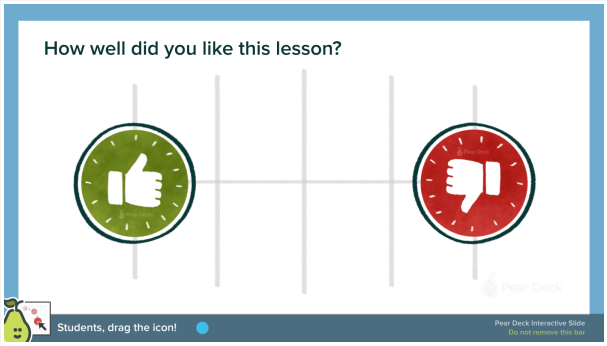
Slide 62


Use this space to take notes:

Slide 63	Your Response
	

Use this space to take notes:

Slide 64	Your Response

Slide 64	Your Response
	

Use this space to take notes:

Slide 65


Use this space to take notes: