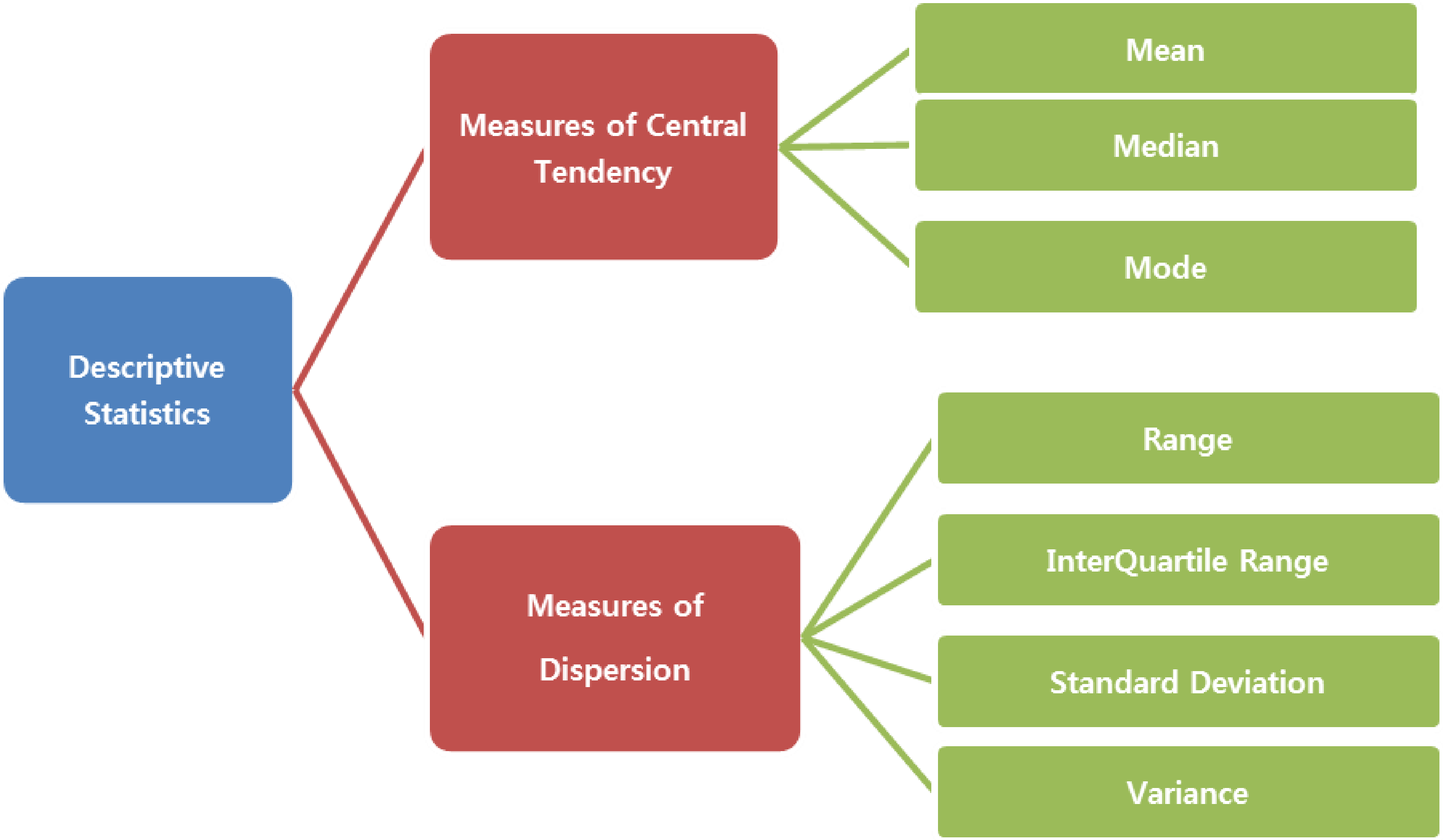
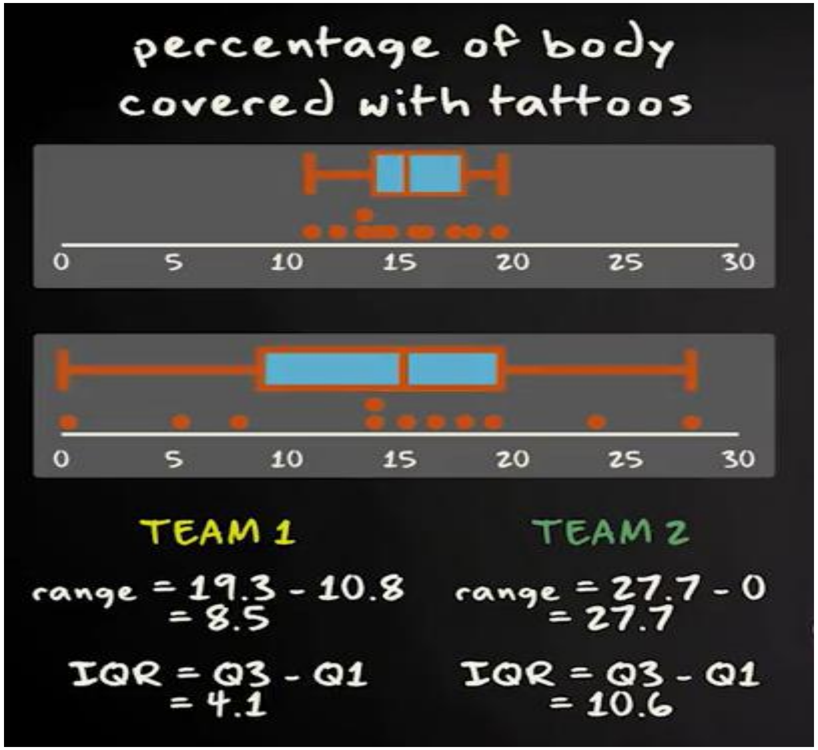
STATISTICAL ANALYSIS - LECTURE 3

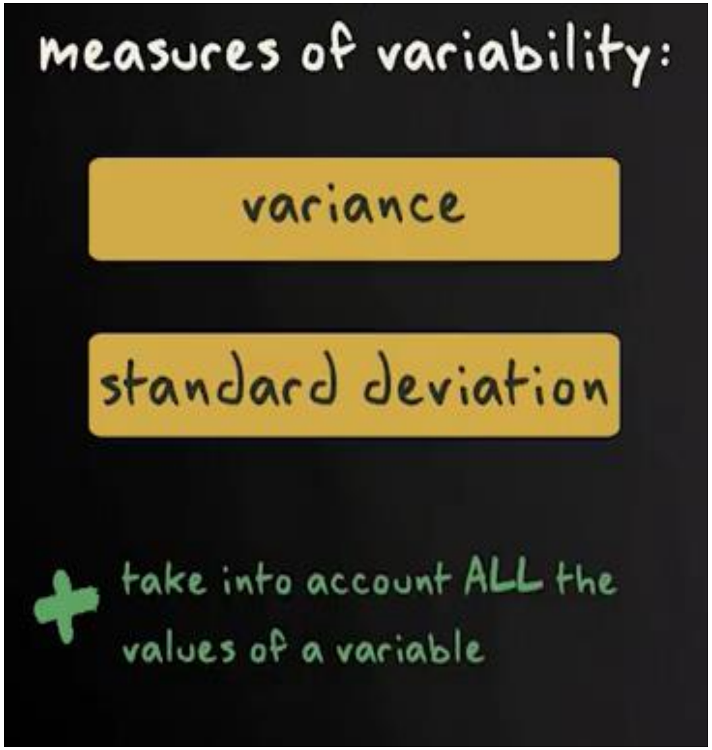
Dr. Mahmoud Mounir

mahmoud.mounir@cis.asu.edu.eg


Descriptive Statistics
Measures of Central Tendency
Mean
Median
Mode
Measures of Dispersion
Range
InterQuartile Range
Standard Deviation
Variance

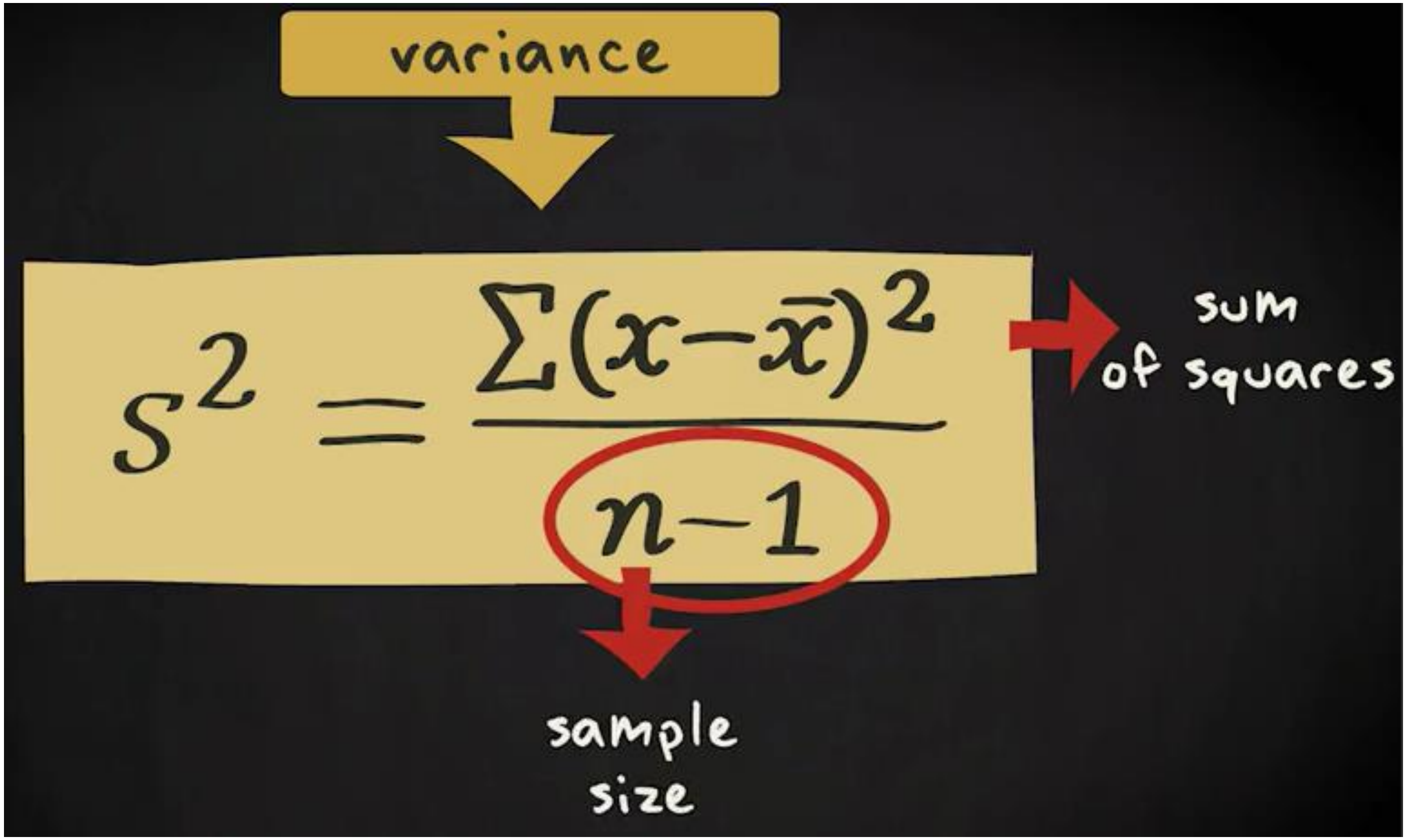
VARIANCE AND STANDARD DEVIATION





VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)

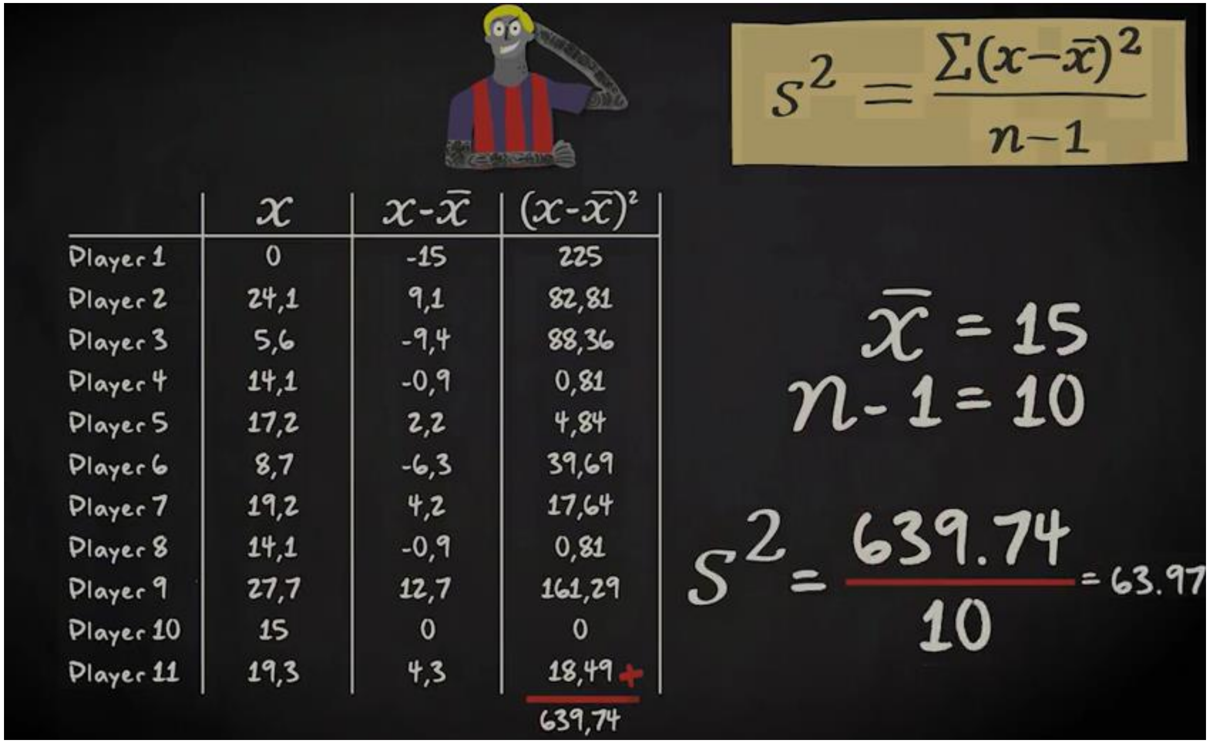


VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)

- **Mean is the point of balance, so we have positive and negative deviations from the mean.**

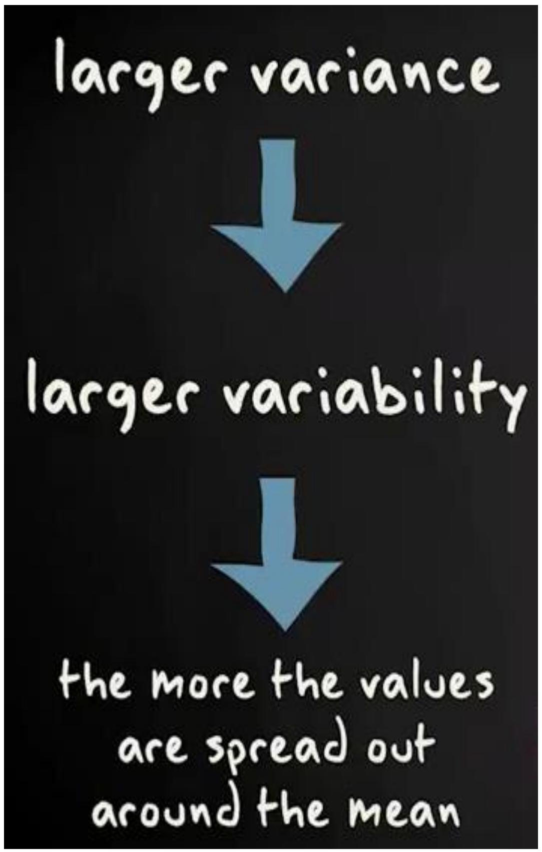
- **The sum of deviation sum to zero. That's why we don't use the original deviations, but the squared deviations.**

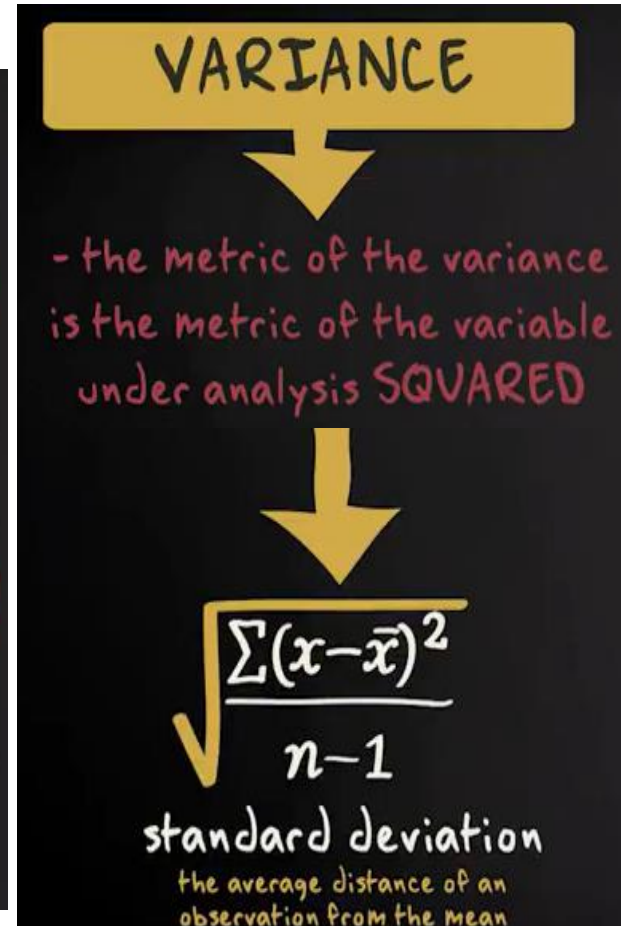


	x	$x-\bar{x}$	$(x-\bar{x})^2$
Player 1	0	-15	225
Player 2	24,1	9,1	82,81
Player 3	5,6	-9,4	88,36
Player 4	14,1	-0,9	0,81
Player 5	17,2	2,2	4,84
Player 6	8,7	-6,3	39,69
Player 7	19,2	4,2	17,64
Player 8	14,1	-0,9	0,81
Player 9	27,7	12,7	161,29
Player 10	15	0	0
Player 11	19,3	4,3	18,49 +
			639,74

VARIANCE AND STANDARD DEVIATION

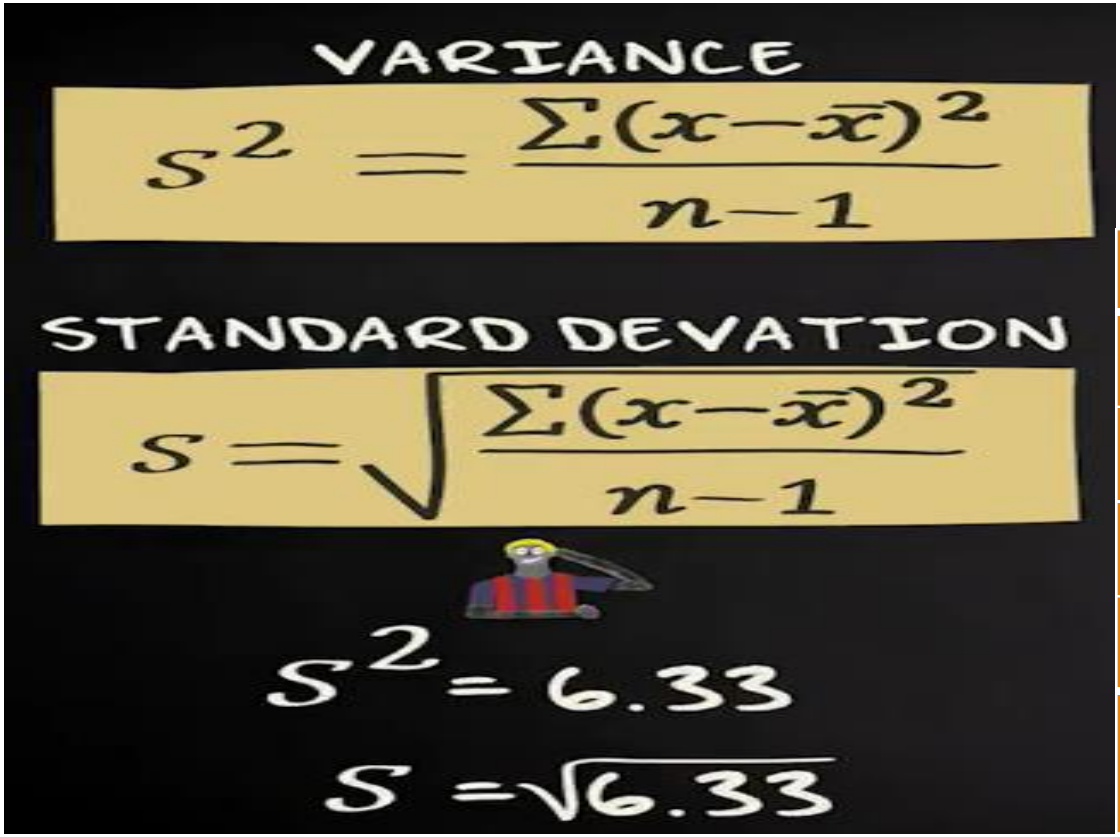
❑ VARIANCE (UNGROUPED DATA)





VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)

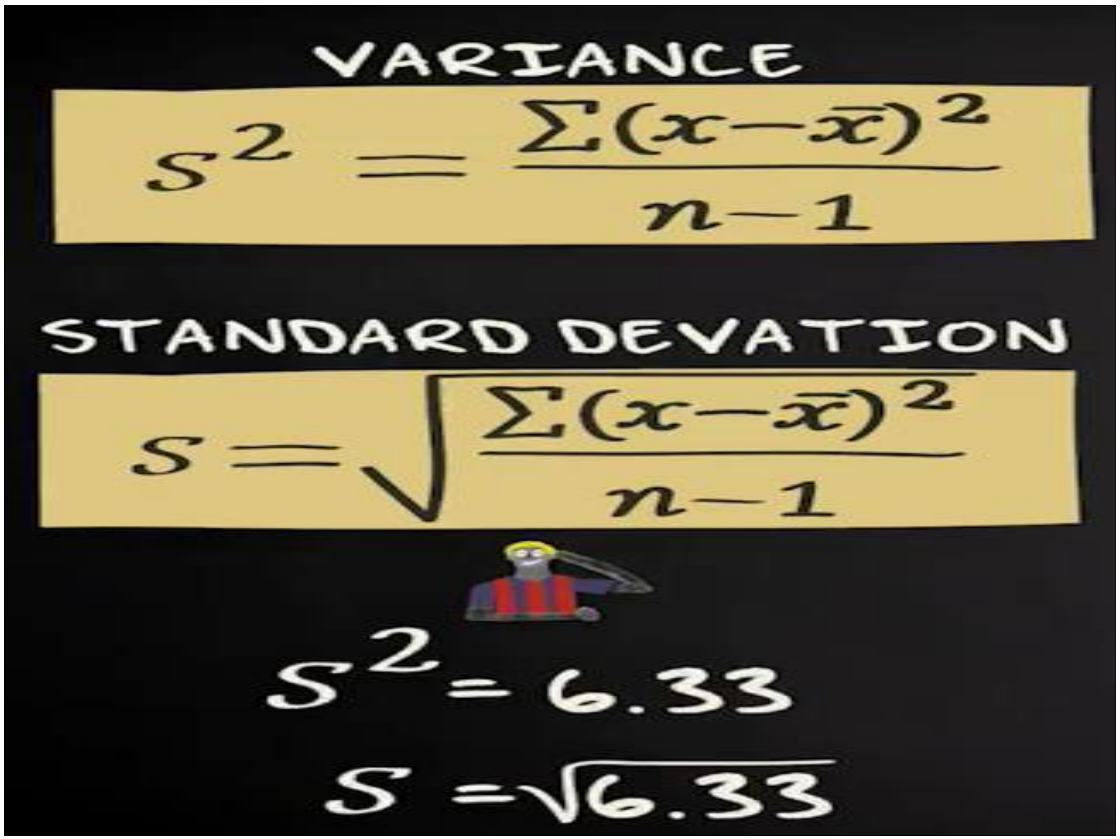


$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

(X)	(X - $\bar{x}$)	(X - $\bar{x}$)^2
50	-112.5	12656.25
100	-62.5	3906.25
200	37.5	1406.25
300	137.5	18906.25
$\bar{x}$ = 162.5		$\sum$ (X - $\bar{x}$)^2 = 36875

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)



$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

(X)	(X^2)
50	2500
100	10000
200	40000
300	90000
$\sum$ X = 650	$\sum$ X^2 = 142500

VARIANCE AND STANDARD DEVIATION

VARIANCE (GROUPED DATA)

$$S = \sqrt{\frac{\sum(x - \bar{x})^2 f}{n-1}}$$

$$\mathbf{x = class\ midpoint}$$

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (GROUPED DATA)

$$s = \sqrt{\frac{\sum x^2 f - \frac{(\sum xf)^2}{n}}{n-1}}$$

x = class midpoint

(X)	(X^2)	f	$X^2 * f$
50	2500	5	12500
100	10000	3	30000
200	40000	6	240000
300	90000	2	180000
$\sum$ X = 650	$\sum X^2$ = 142500	n = 16	$\sum X^2 * f$ = 462500

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (GROUPED DATA)

Age	*Frrquency (f)*	*Midpoint (x)*	*X-Mean*	*(X-Mean)^2*	*(X-Mean)^2 f*
30-34	*4*	*32*	*-9*	*81*	*324*
35-39	*5*	*37*	*-4*	*16*	*80*
40-44	*2*	*42*	*1*	*1*	*2*
45-49	*9*	*47*	*6*	*36*	*324*
Total	*20*				*730*

∑f = n = 20

Mean = 820/20 = 41

∑(X-Mean)^2 f = 730

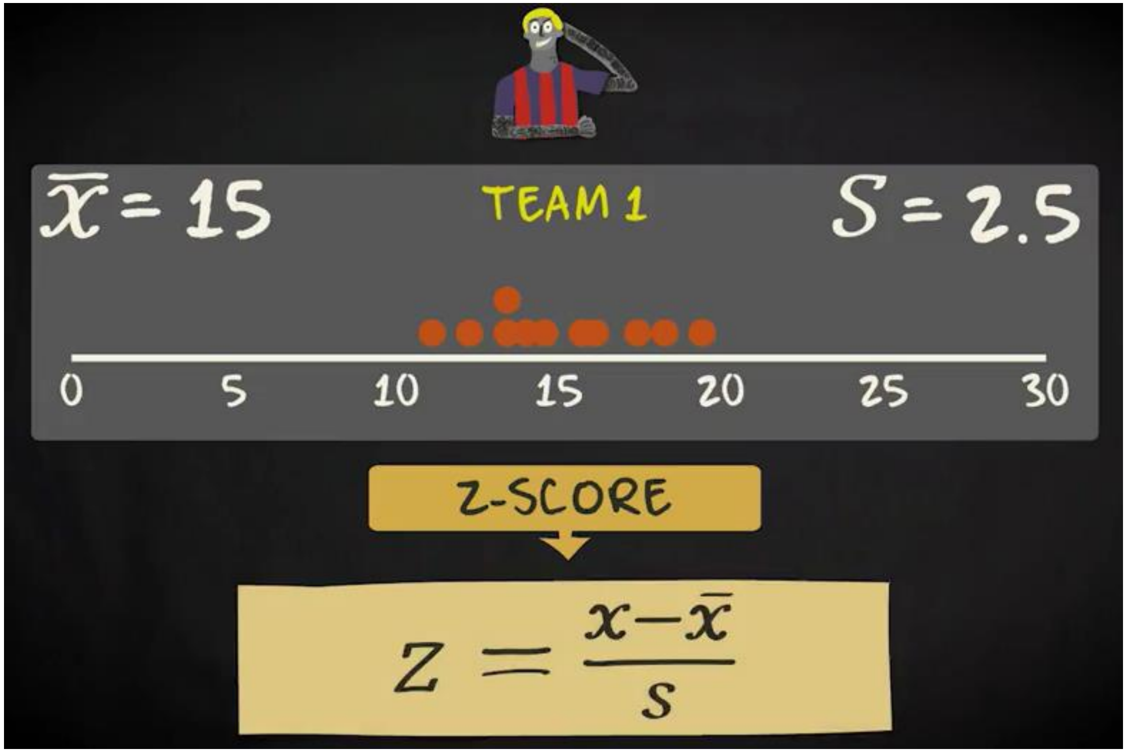
$$S = \sqrt{\frac{730}{20 - 1}}$$
$$= \sqrt{38.42} \approx 6.20$$

Z-SCORE

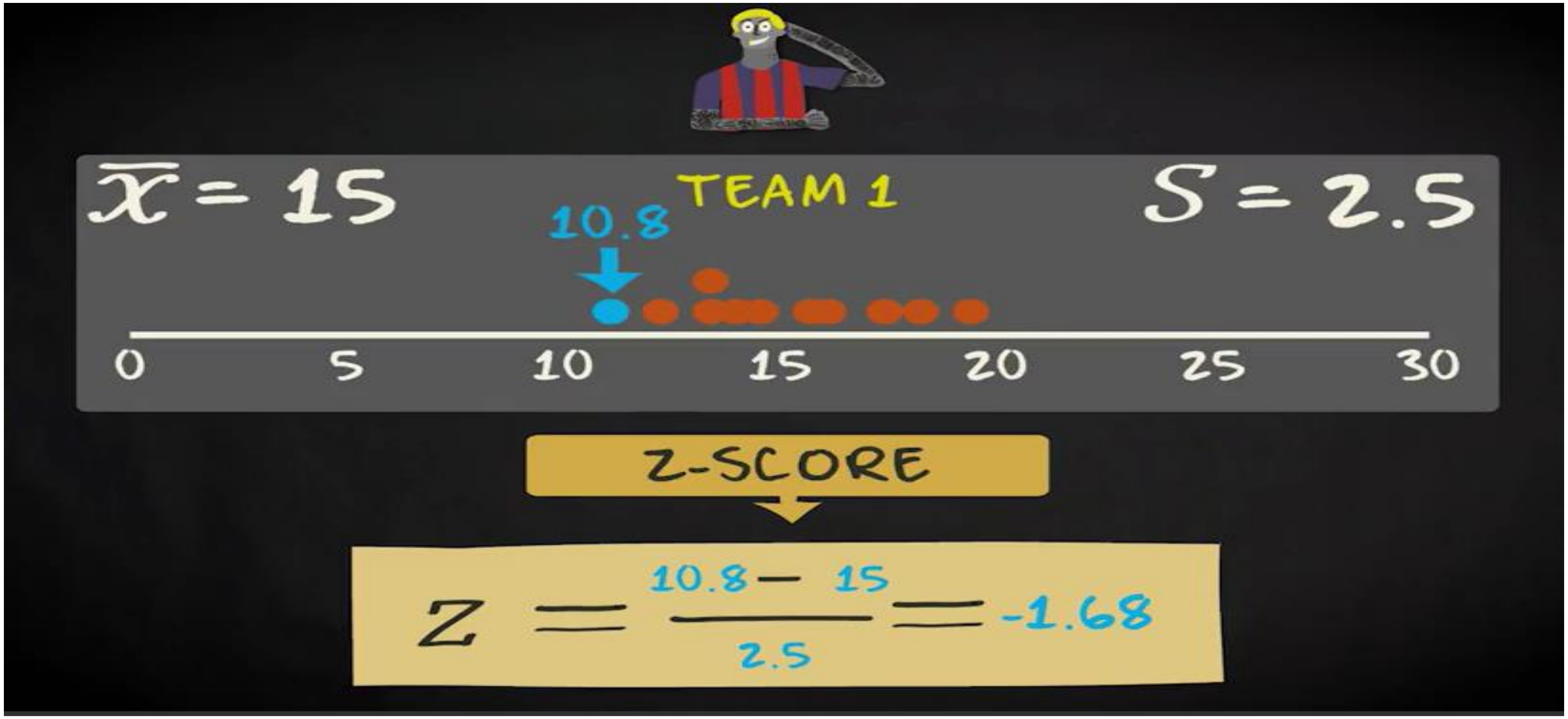
- Sometimes researchers want to know if a specific observation is **common** or **exceptional**.
- To answer that question, they express a score in terms of **how many standard deviations below or above the population mean a raw score is.**
- This number is what we call a **z-score**.
- If we recode original scores into z-scores, we say that we **standardize** a variable.

Z-SCORE


TEAM 1
S = 2.5
0 5 10 15 20 25 30
TEAM 2
S = 8.0
0 5 10 15 20 25 30
percentage of body covered with tattoos

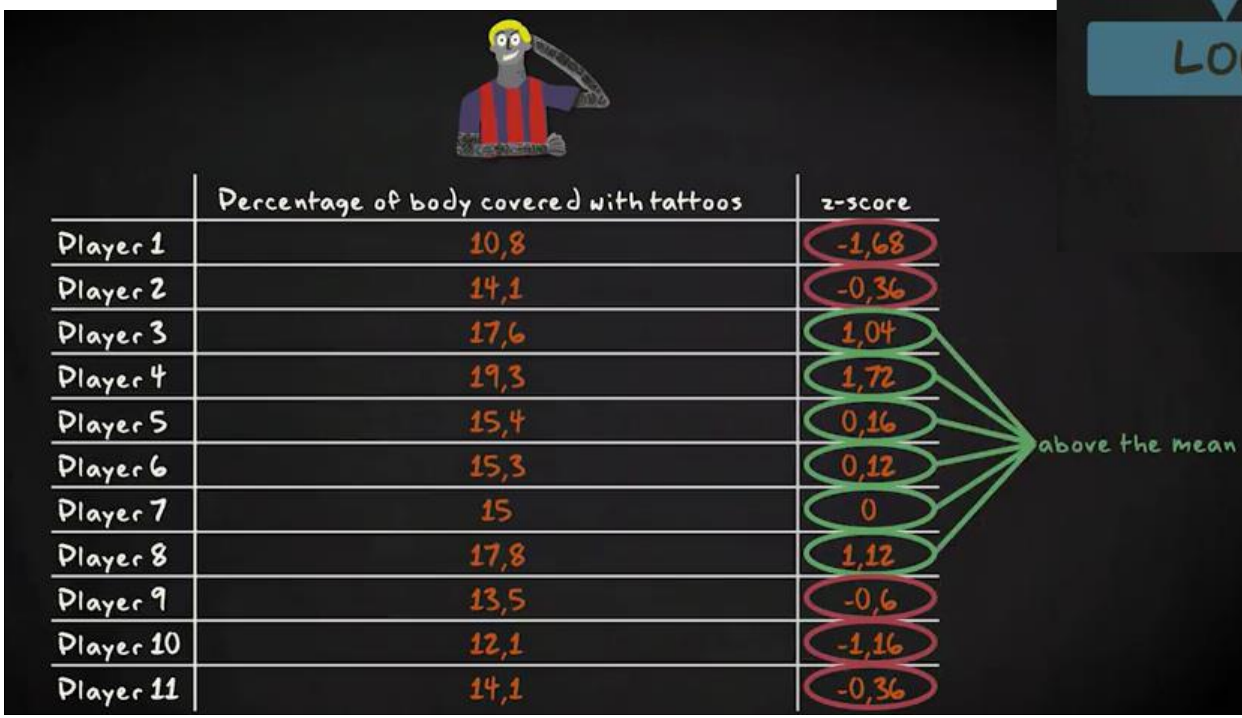

$\bar{x} = 15$
TEAM 1
$S = 2.5$
0 5 10 15 20 25 30
Z-SCORE
$Z = \frac{x - \bar{x}}{s}$

Z-SCORE



$\bar{x} = 15$
TEAM 1
$S = 2.5$
10.8
0
5
10
15
20
25
30
Z-SCORE
$Z = \frac{10.8 - 15}{2.5} = -1.68$

Z-SCORE

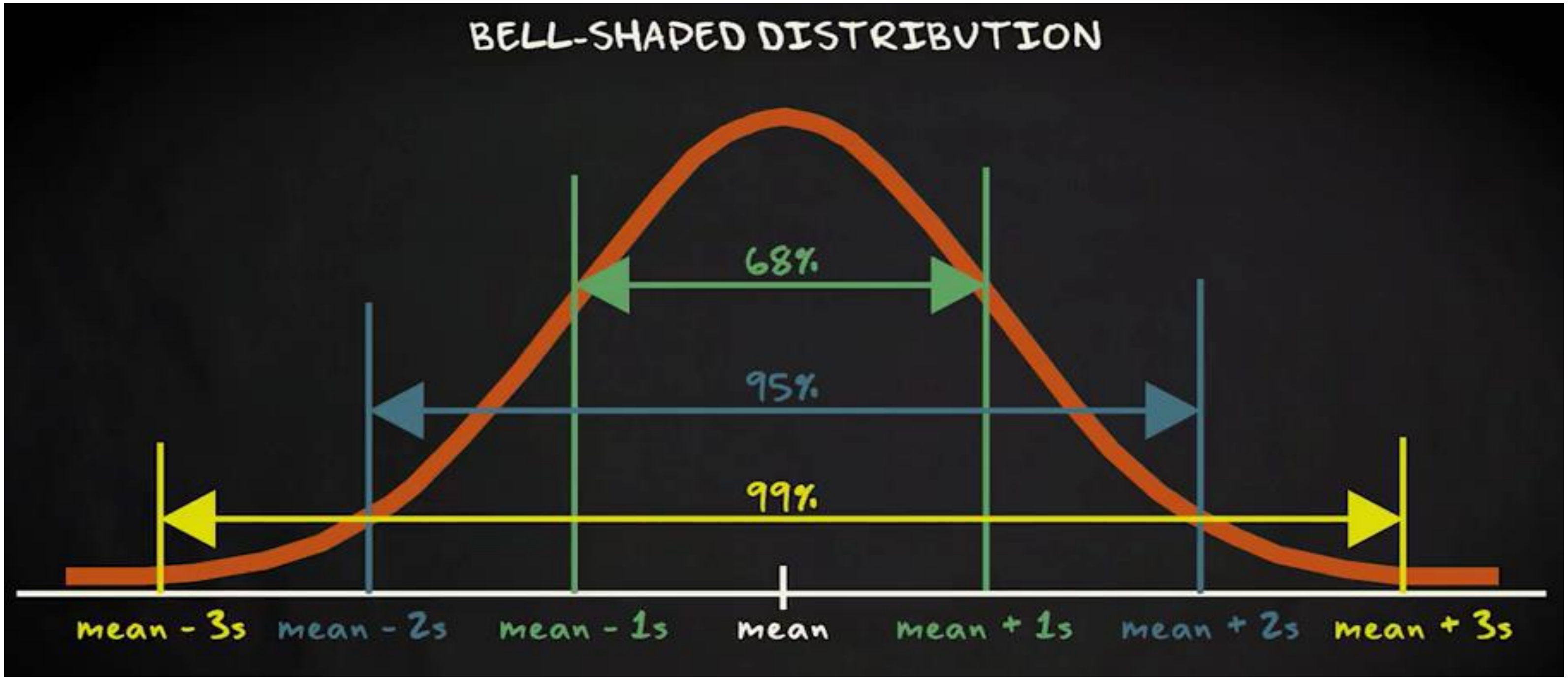


	Percentage of body covered with tattoos	z-score
Player 1	10,8	-1,68
Player 2	14,1	-0,36
Player 3	17,6	1,04
Player 4	19,3	1,72
Player 5	15,4	0,16
Player 6	15,3	0,12
Player 7	15	0
Player 8	17,8	1,12
Player 9	13,5	-0,6
Player 10	12,1	-1,16
Player 11	14,1	-0,36

Z-SCORE

❑ **EMPIRICAL RULE**

NORMAL DISTRIBUION (BELL SHAPED)



Z-SCORE

EMPIRICAL RULE “APPROXIMATION”

NORMAL DISTRIBUION (BELL SHAPED)

- ***Approximately 68%***

of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations.

- ***Approximately 95%***

of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations.

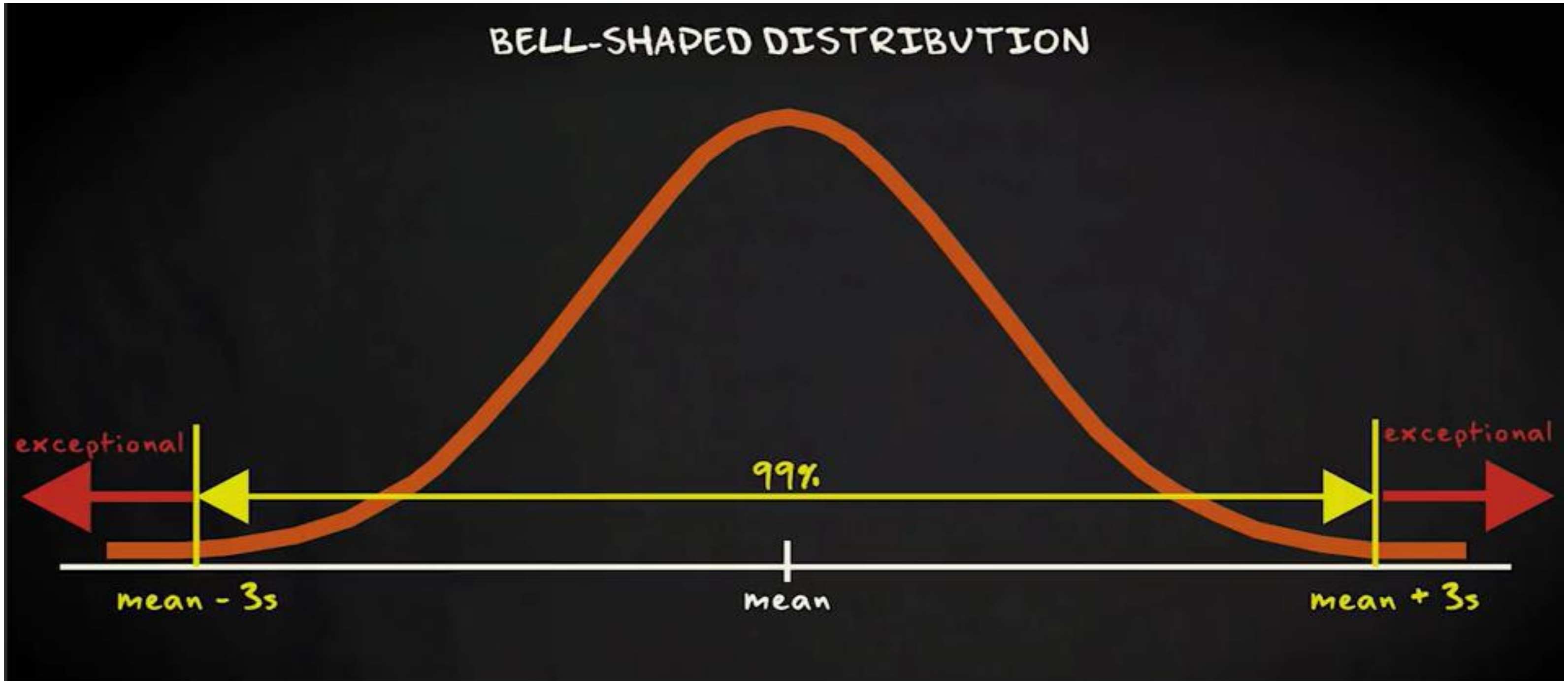
- ***Approximately 99.7%***

of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

Z-SCORE

- **EMPIRICAL RULE**

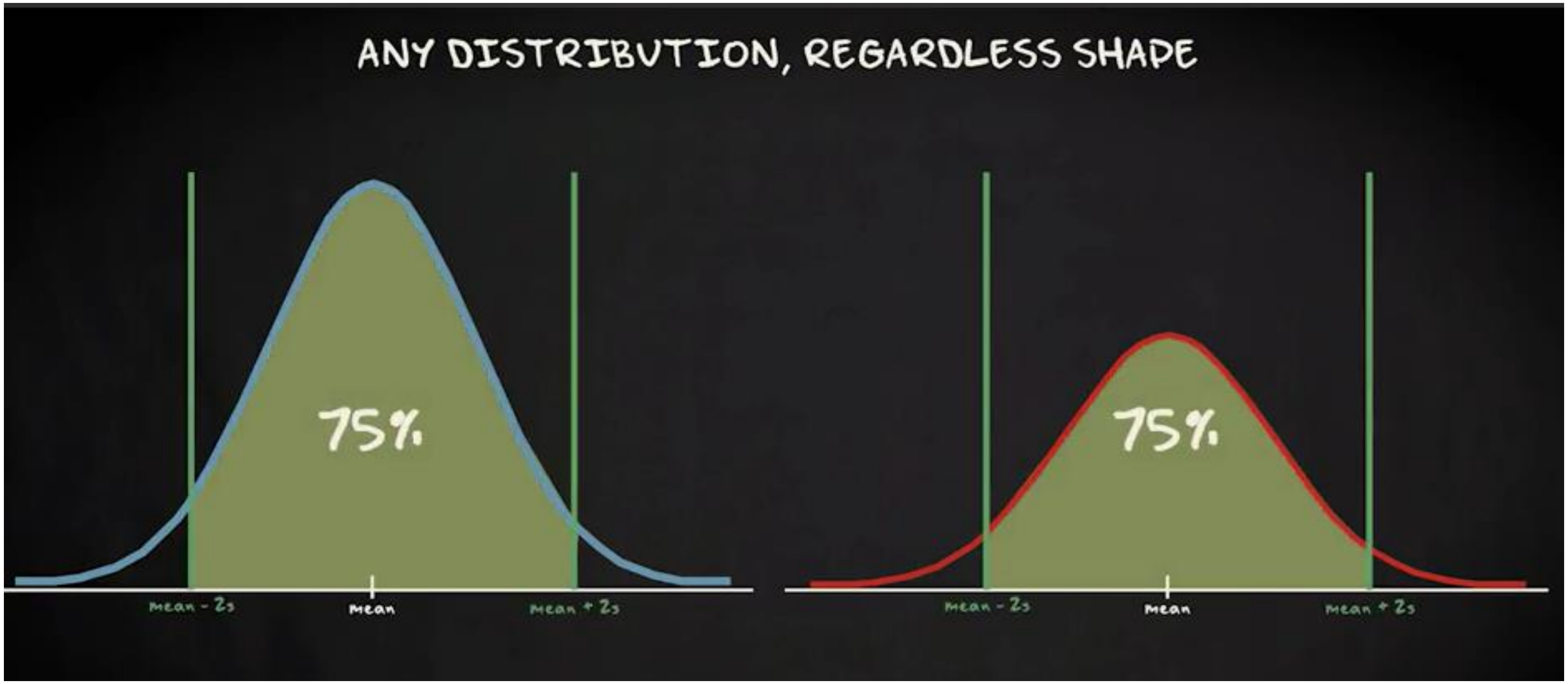
  NORMAL DISTRIBUION (BELL SHAPED)



Z-SCORE

❑ CHEBYSHEV'S RULE

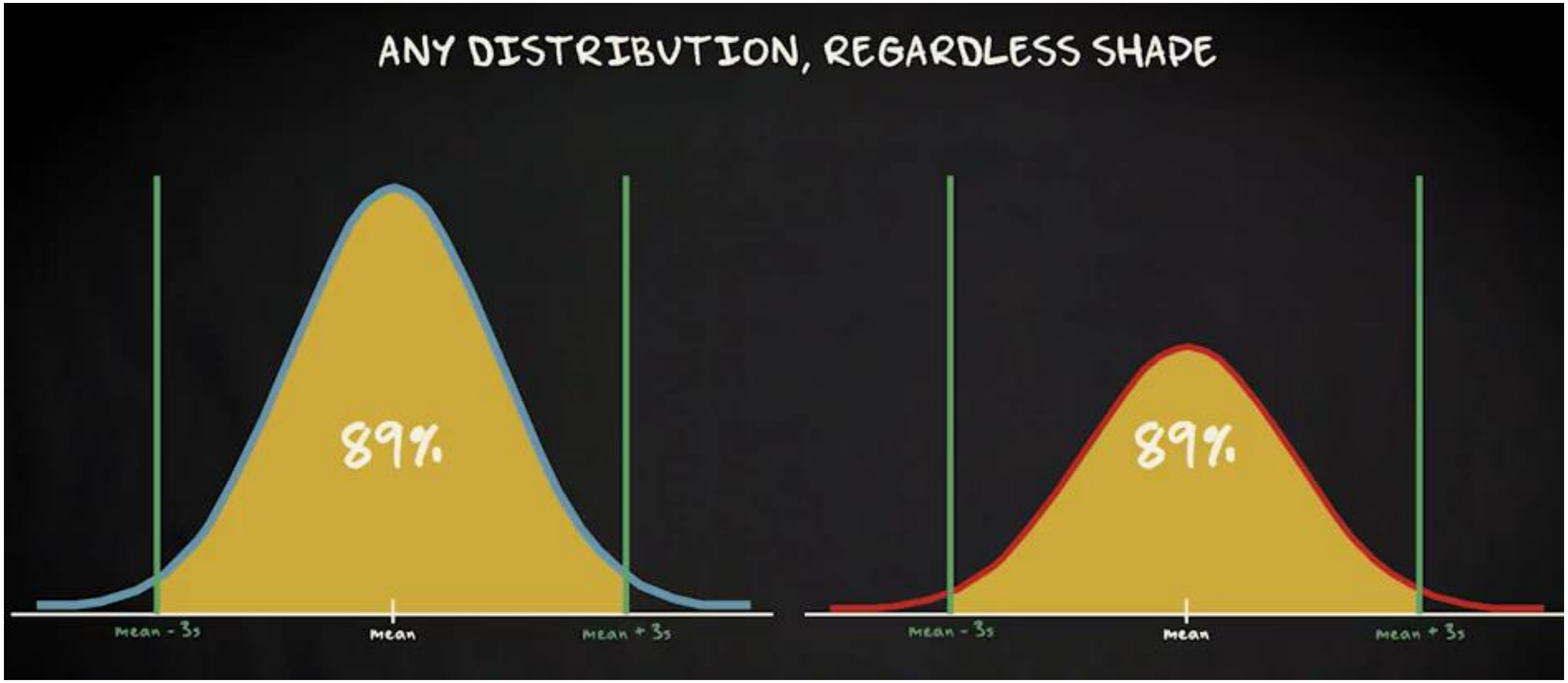
ANY DISTRIBUION



Z-SCORE

❑ **CHEBYSHEV'S RULE**

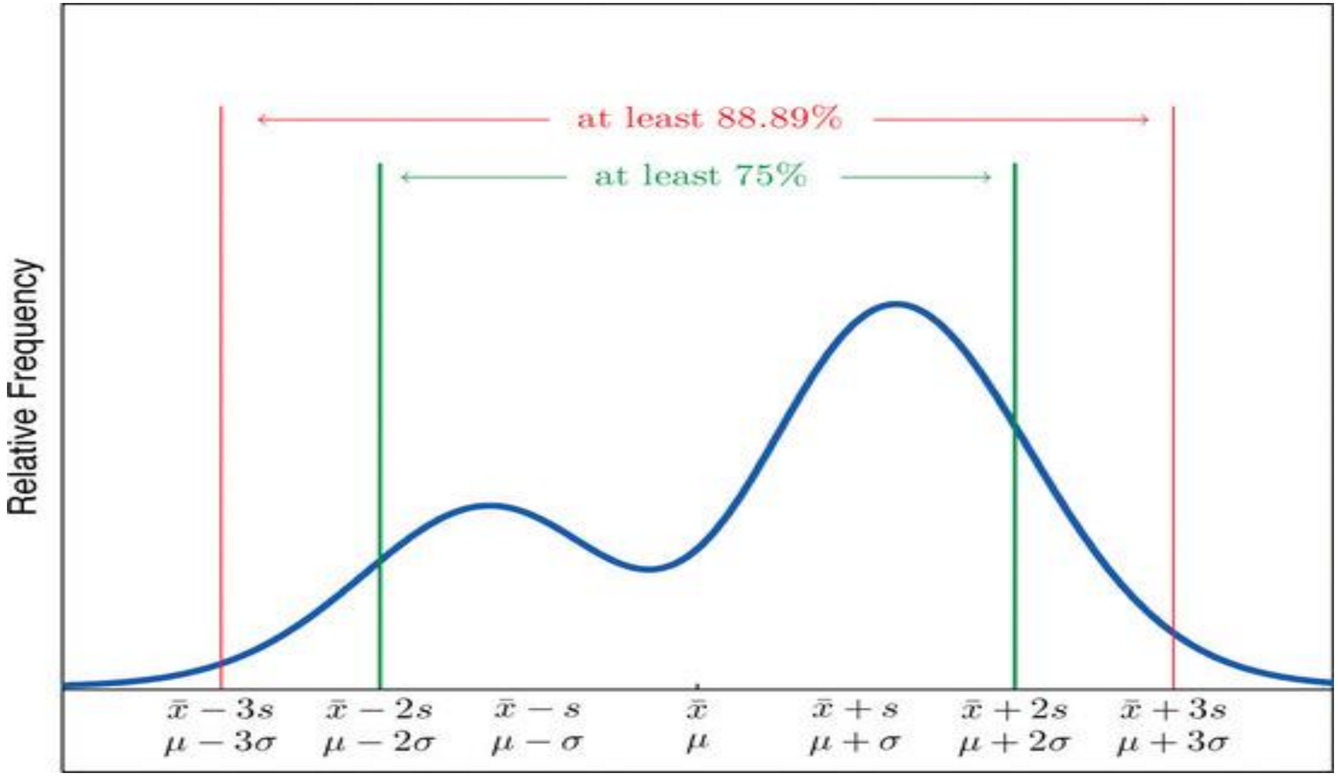
ANY DISTRIBUION



Z-SCORE

❑ CHEBYSHEV'S RULE

ANY DISTRIBUION



Z-SCORE

❑ CHEBYSHEV'S RULE "FACT"

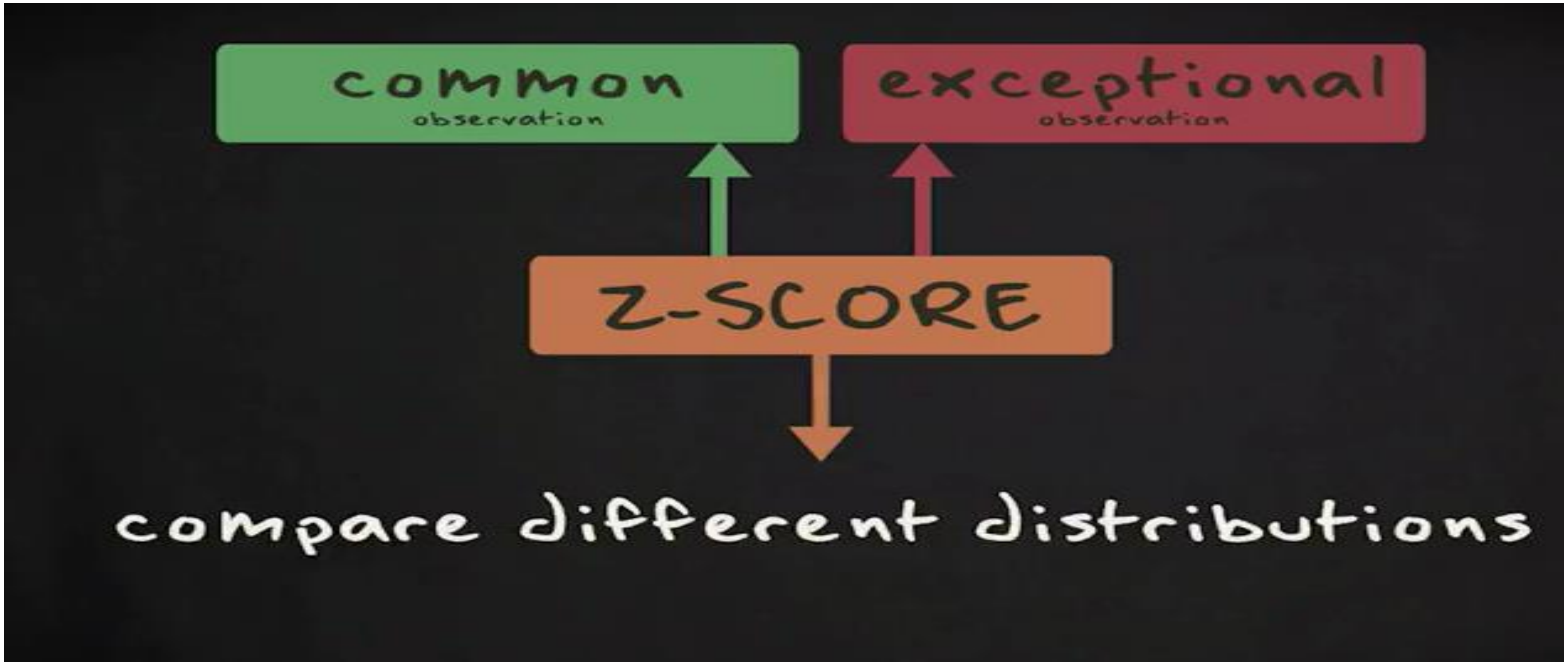
ANY DISTRIBUION

- ***At Least 75%***

  of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x}$ ±2s for samples and with endpoints μ±2σ for populations.
- ***At Least 89%***

  of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x}$ ±3s for samples and with endpoints μ±3σ for populations.

Z-SCORE


common
observation
exceptional
observation
Z-SCORE
compare different distributions

EXERCISE (1)

- **What does the distribution of the variable look like?**
- **What is the center of the distribution?**
- **Study the variability of the distribution.**
- **Construct a box plot.**
- **What is the z-score of school #3?**

	Average grade chemistry
School 1	7,4
School 2	7,9
School 3	4,1
School 4	8,1
School 5	6,2
School 6	7,1
School 7	7,4
School 8	6,7

EXERCISE (2)

- **The following table shows the heights in inches of 100 randomly selected adult men measured in inches.**

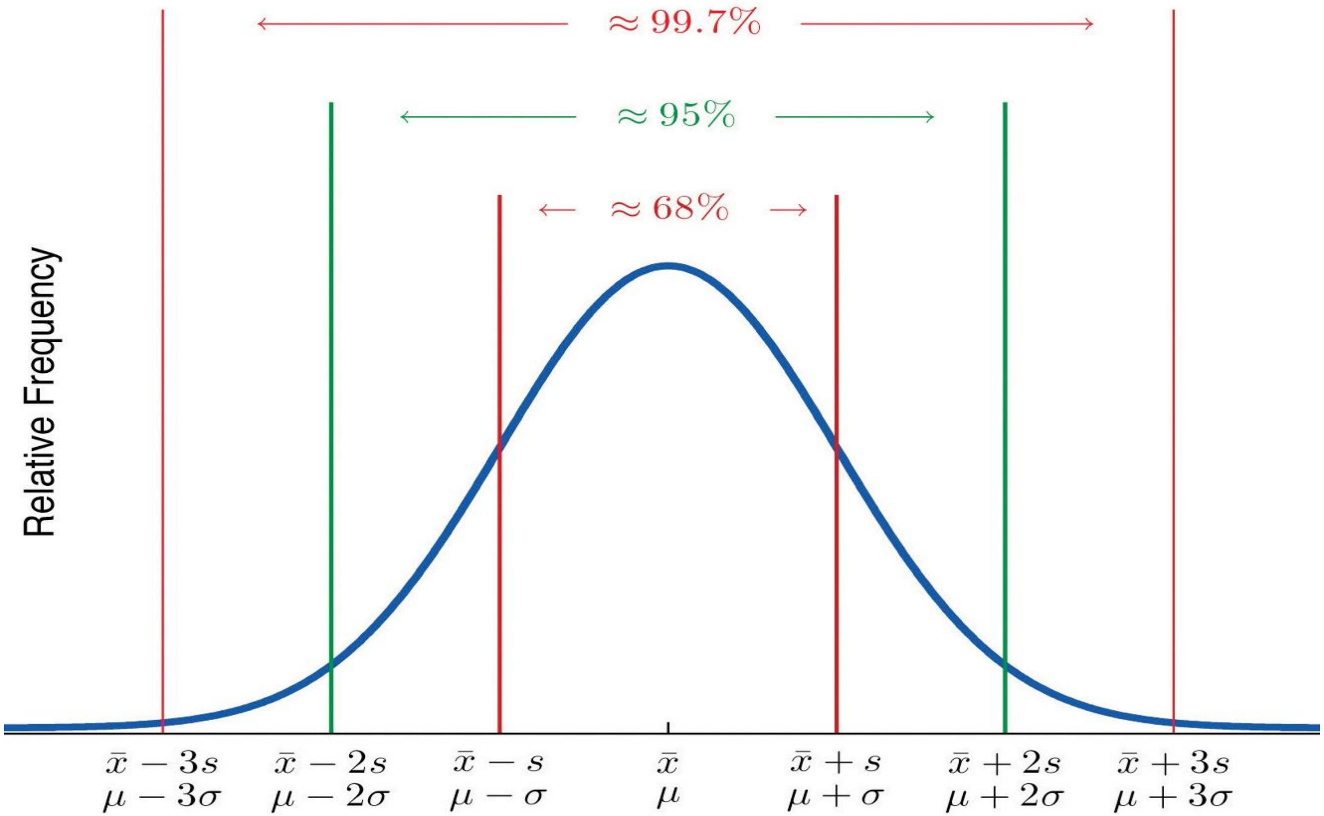
68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

Mean $\bar{x}$ = 69.92 inches
Standard Deviation S = 1.70 inches

MIN = 65.6 inches
MAX = 74.8 inches

EXERCISE (2)

- **A relative frequency histogram for the data**



EXERCISE (2)

- The number of observations that are within ONE standard deviation of the mean

$\bar{x} - S$ $\bar{x} + S$

69.92 – 1.70 = 68.22 inches and 69.92 + 1.70 = 71.62 inches

69

- The number of observations that are within TWO standard deviation of the mean

$\bar{x} - 2S$ $\bar{x} + 2S$

69.92 – 2(1.70) = 66.52 inches and 69.92 + 2(1.70) = 73.32 inches

95

- The number of observations that are within THREE standard deviation of the mean

$\bar{x} - 3S$ $\bar{x} + 3S$

69.92 – 3(1.70) = 64.822 inches and 69.92 + 3(1.70) = 75.02 inches

ALL

EXERCISE (3)

- **Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.**
  1. **About what proportion of all such men are between 68.2 and 71 inches tall?**
  2. **What interval centered on the mean should contain about 95% of all such men?**

EXERCISE (3) SOLUTION

- **The observations that are within ONE standard deviation of the mean**

$\bar{x} - S$ $\bar{x} + S$

69.6 – 1.40 = 68.2 inches and 69.6 + 1.40 = 71.71 inches

68 %

- **The observations that are within TWO standard deviation of the mean**

$\bar{x} - 2S$ $\bar{x} + 2S$

69.6 – 2(1.40) = 66.80 inches and 69.6 + 2(1.40) = 72.40 inches

95 %

- **The observations that are within THREE standard deviation of the mean**

$\bar{x} - 3S$ $\bar{x} + 3S$

69.6 - 3(1.40) = 65.40 inches and 69.6 + 3(1.40) = 73.80 inches

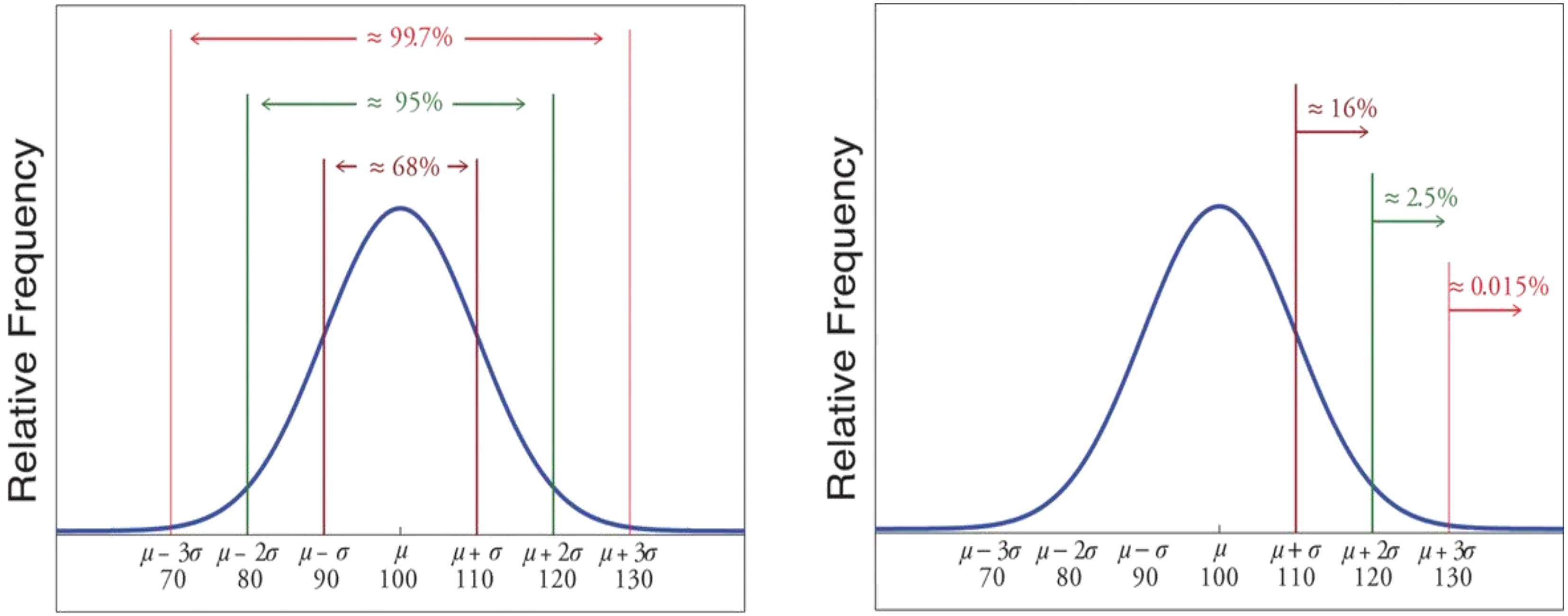
ALL

EXERCISE (4)

- Scores on IQ tests have a bell-shaped distribution with **mean $\mu=100$** and **standard deviation $\sigma=10$**. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

  - **Approximately 68% of the IQ scores in the population lie between 90 and 110,**
  - **Approximately 95% of the IQ scores in the population lie between 80 and 120, and**
  - **Approximately 99.7% of the IQ scores in the population lie between 70 and 130.**

EXERCISE (4)



(a) Whole Spectrum (b) Higher End

EXERCISE (5)

- **A sample of size n=50 has mean $\bar{x}$ =28 and standard deviation s=3. Without knowing anything else about the sample,**
- **what can be said about the number of observations that lie in the interval (22,34)?**
- **What can be said about the number of observations that lie outside that interval?**

EXERCISE (5) SOLUTION

By Chebyshev's Theorem:

- **The observations that are within TWO standard deviation of the mean**

  $\bar{x} - 2S$ $\qquad\qquad$ $\bar{x} + 2S$

  28 – 2(3) = 22 $\qquad$ **and** $\qquad$ **28 + 2(3) = 34**

  75 %

- **The observations that are within THREE standard deviation of the mean**

  $\bar{x} - 3S$ $\qquad\qquad$ $\bar{x} + 3S$

  28 – 3(3) = 19 $\qquad$ **and** $\qquad$ **28 + 3(3) = 37**

  89 %

EXERCISE (5) SOLUTION

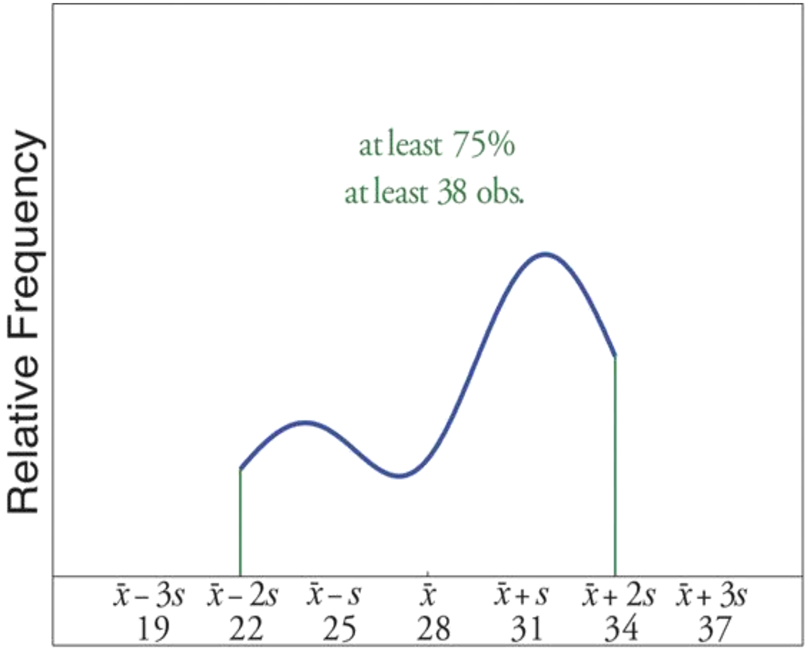
- **The interval (22,34) is the one that is formed by adding and subtracting two standard deviations from the mean.**

  By Chebyshev's Theorem,

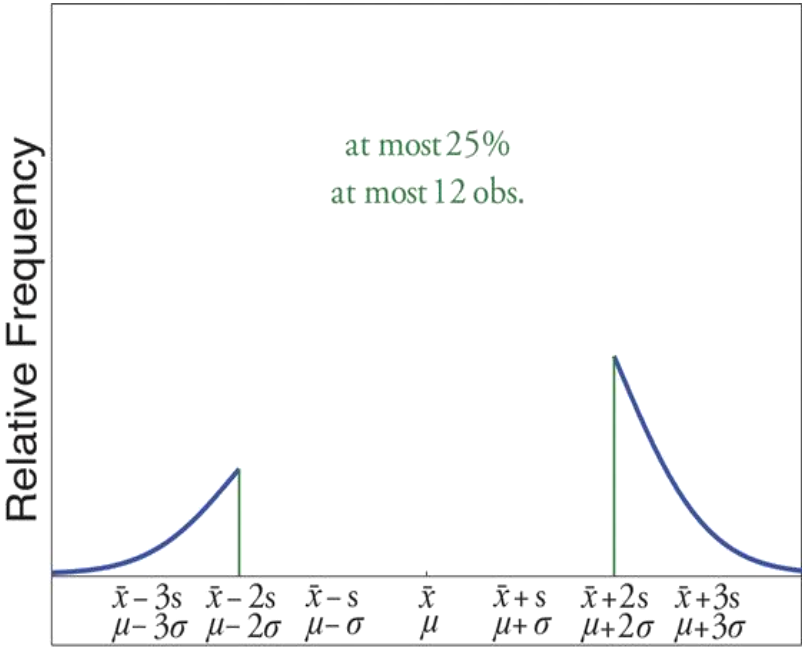
  - **At least 75 % of the data are within this interval.**
  - **Since 75 % of 50 is 37.5, this means that at least 37.5 observations are in this interval or at least 38 observations.**

  - **If at least 75 % of the observations are in the interval, then at most 25 % of them are outside it.**
  - **Since 1/4 of 50 is 12.5, at most 12.5 observations are outside the interval or 38 observations.**

EXERCISE (5) SOLUTION



(a) Within $\bar{x} \pm 2s$



(b) Outside $\bar{x} \pm 2$

EXERCISE (6)

1. (26 points total) Suppose that in 2004, the verbal portion of the Scholastic Aptitude Test (SAT) had a mean score of $\mu = 500$ and a standard deviation of $\sigma = 100$, while in the same year, the verbal exam from the American College Testing Program (known as ACT) had a mean of $\mu = 21.0$ and a standard deviation of $\sigma = 4.7$. Assume that the scores from both exams are approximately normally distributed in any given year.

a. (9 points) Two friends applying for college took the tests, the first of the two scoring 650 on the SAT and the second scoring 30 on the ACT. Which of these students scored higher among the population of students taking the relevant test? Exhibit clearly all the calculations that justify your answer.

EXERCISE (6)

$Z_{SAT} = (650-500)/100 = 1.5$ (93.32 Percentile)

$Z_{ACT} = (30-21)/4.7 = 1.91$ (97.19 Percentile)

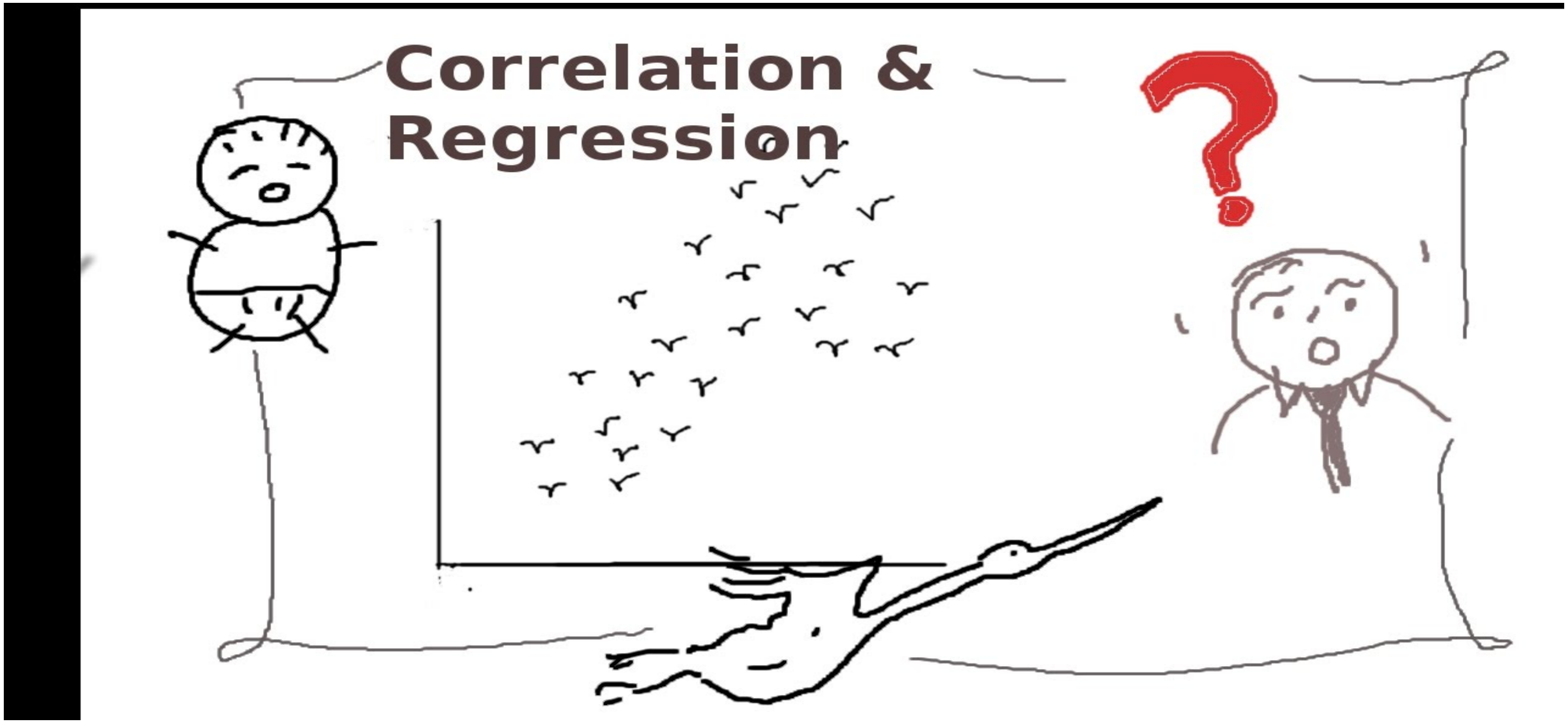
The student taking the ACT test performed better because his/her test score has a higher Z-score (or equivalently, higher percentile).

EXERCISE (7)

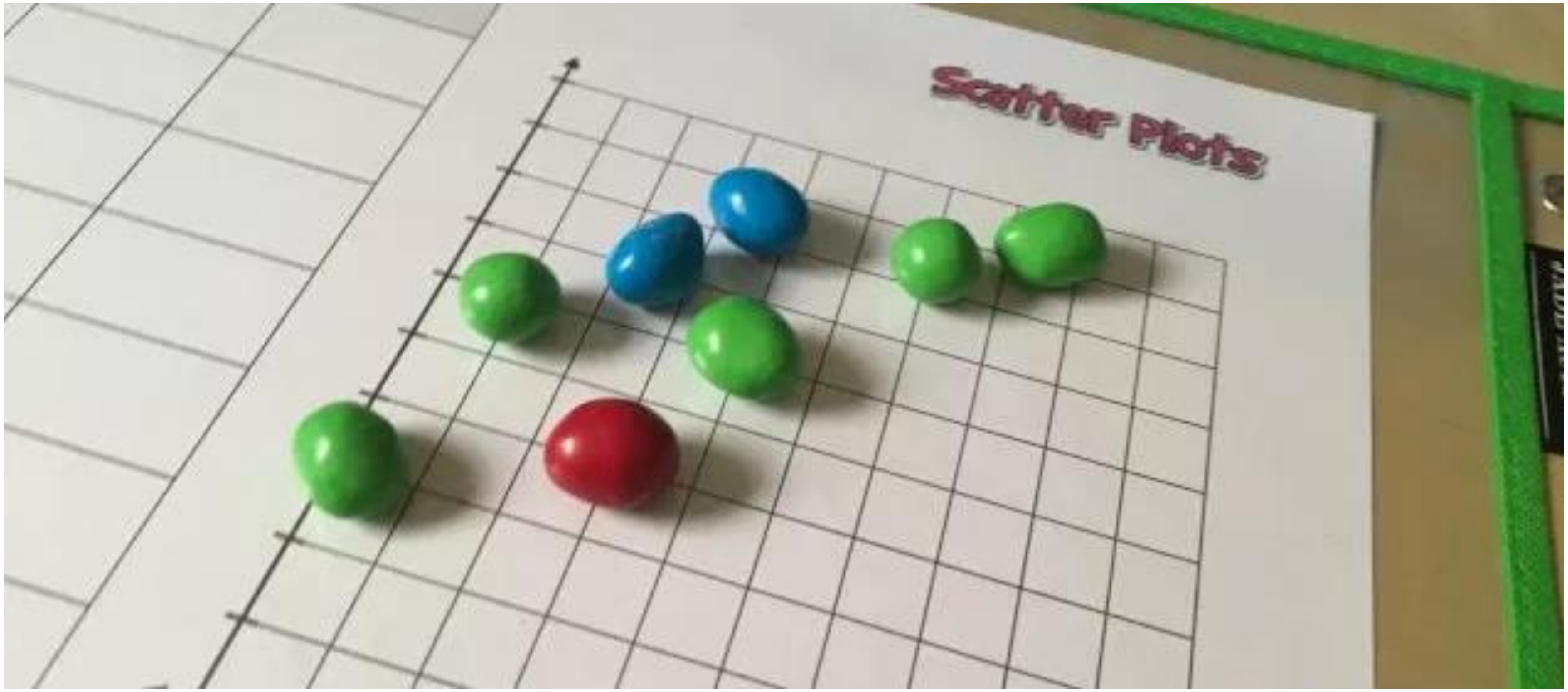
(2 Marks) Here are some summary statistics for the numbers of acres of soybeans فول الصويا and peanuts الفول السوداني harvested per county in Alabama in 2009, for counties that planted those crops.

In one southern county, there were 9 thousand acres of soybeans harvested and 3 thousand acres of peanuts harvested. **Relative to its crop, which plant had a better harvest?**

Crop	Mean harvest (thousands of acres)	Standard deviation (thousands of acres)
Soybeans	$\mu = 12$	$\sigma = 14$
Peanuts	$\mu = 10$	$\sigma = 8$

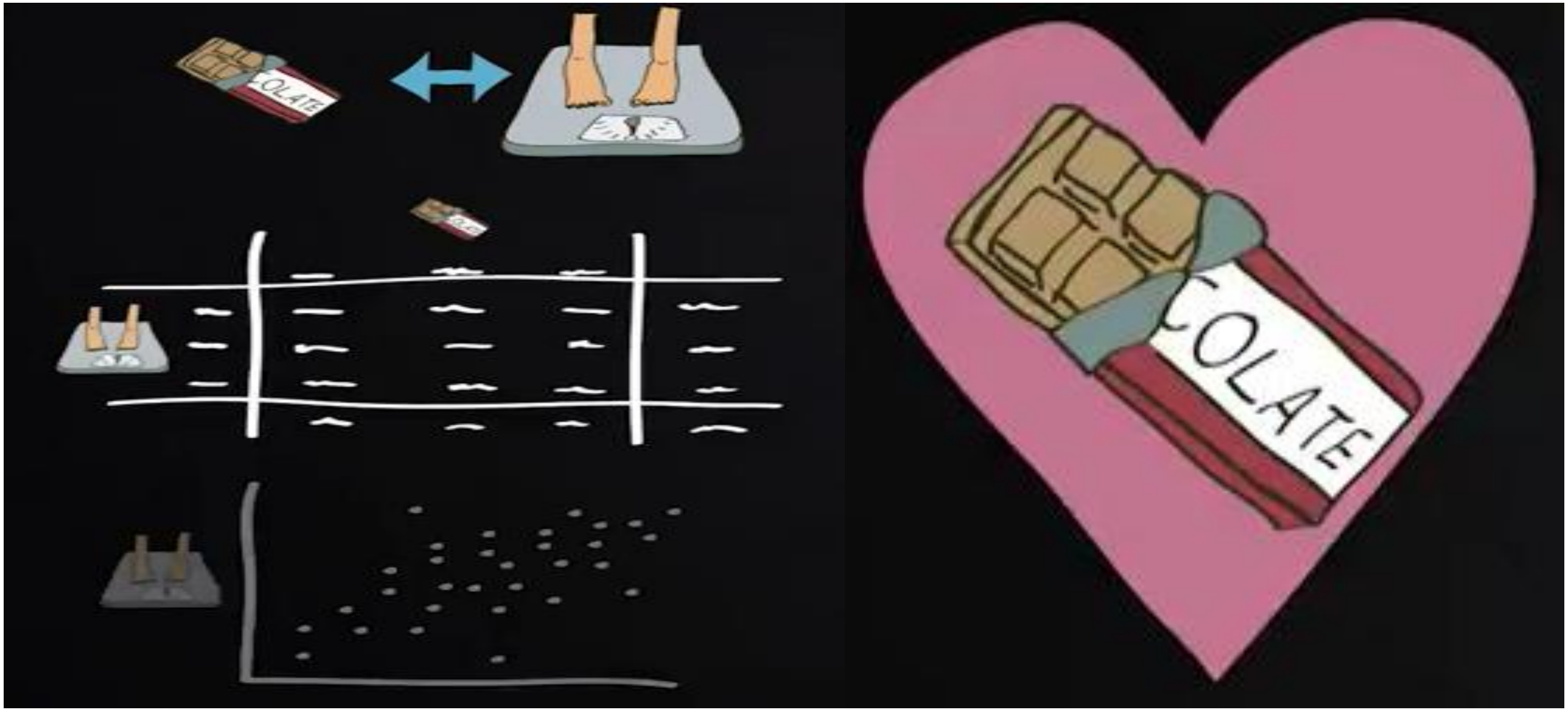


CORRELATION AND REGRESSION

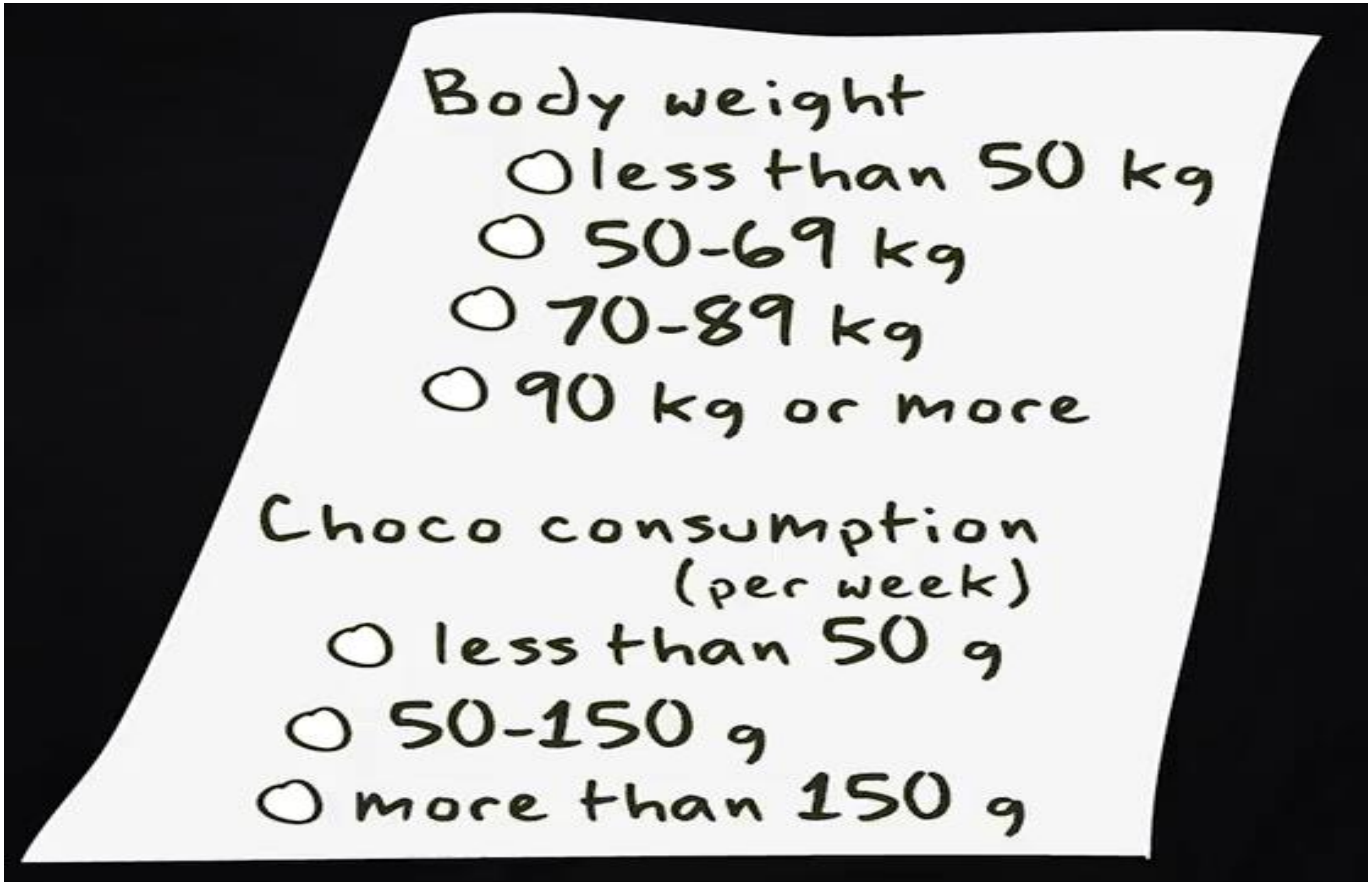


CORRELATION:
CROSSTABS AND SCATTER PLOTS

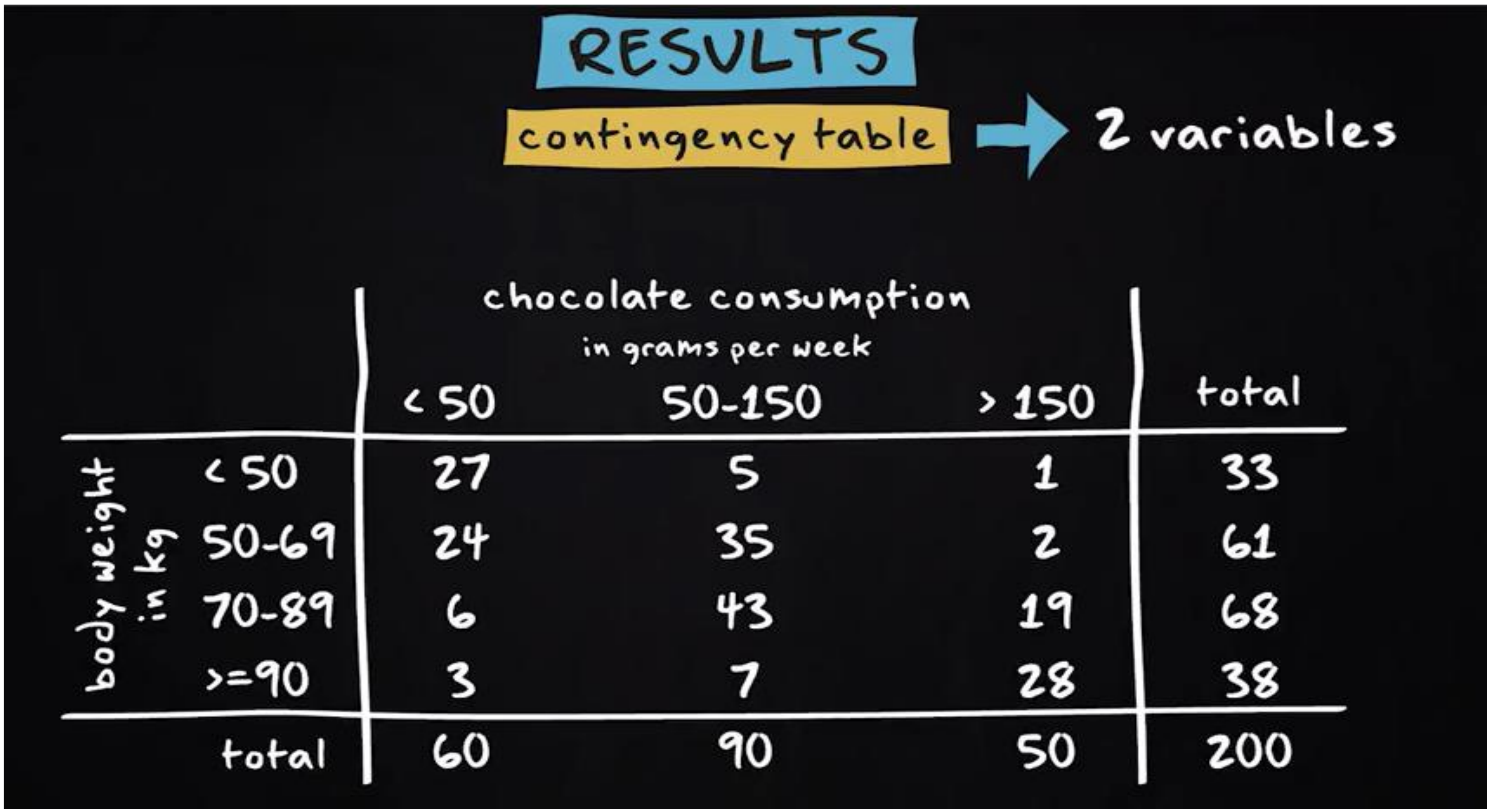
CORRELATION: CROSSTABS AND SCATTER PLOTS



CORRELATION: CROSSTABS AND SCATTER PLOTS

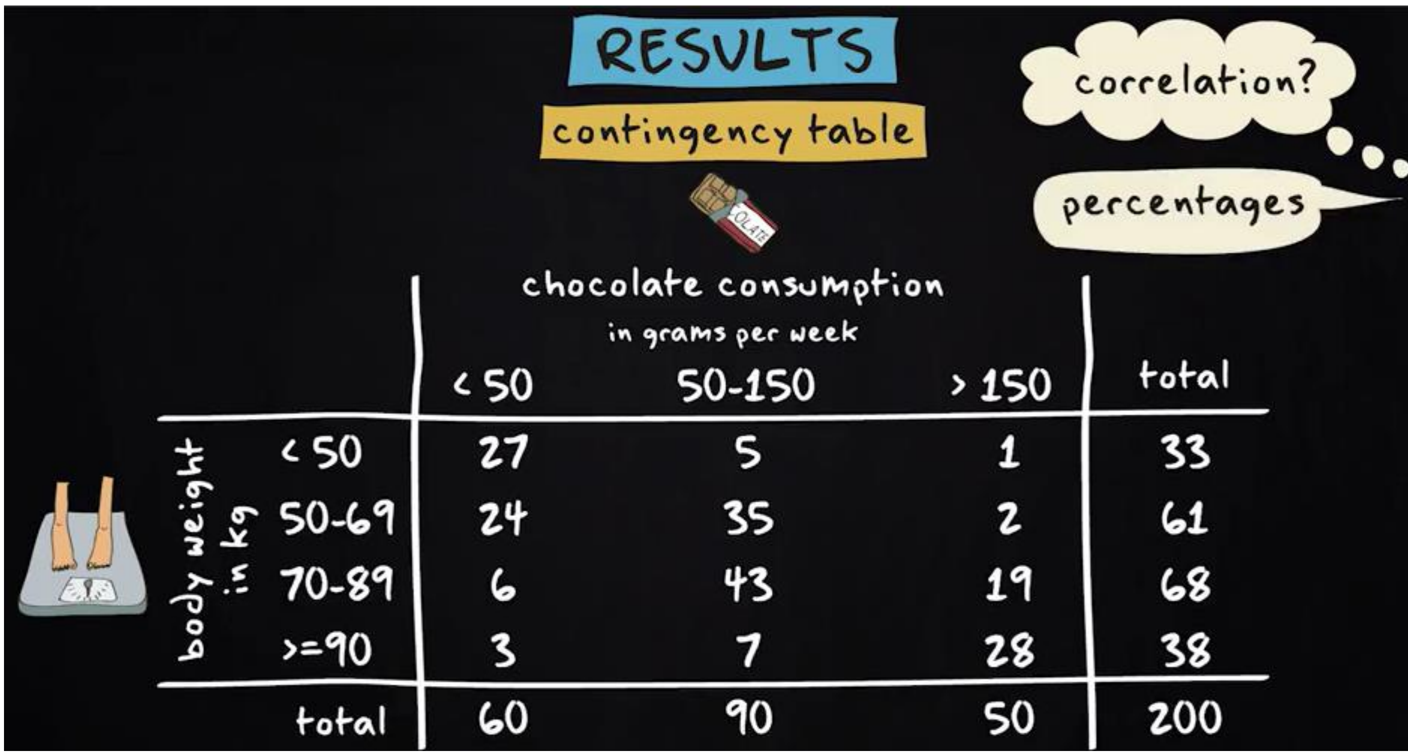


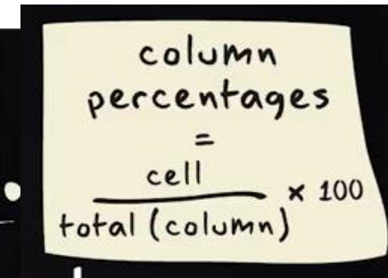
CROSSTABS (CONTINGENCY TABLES)



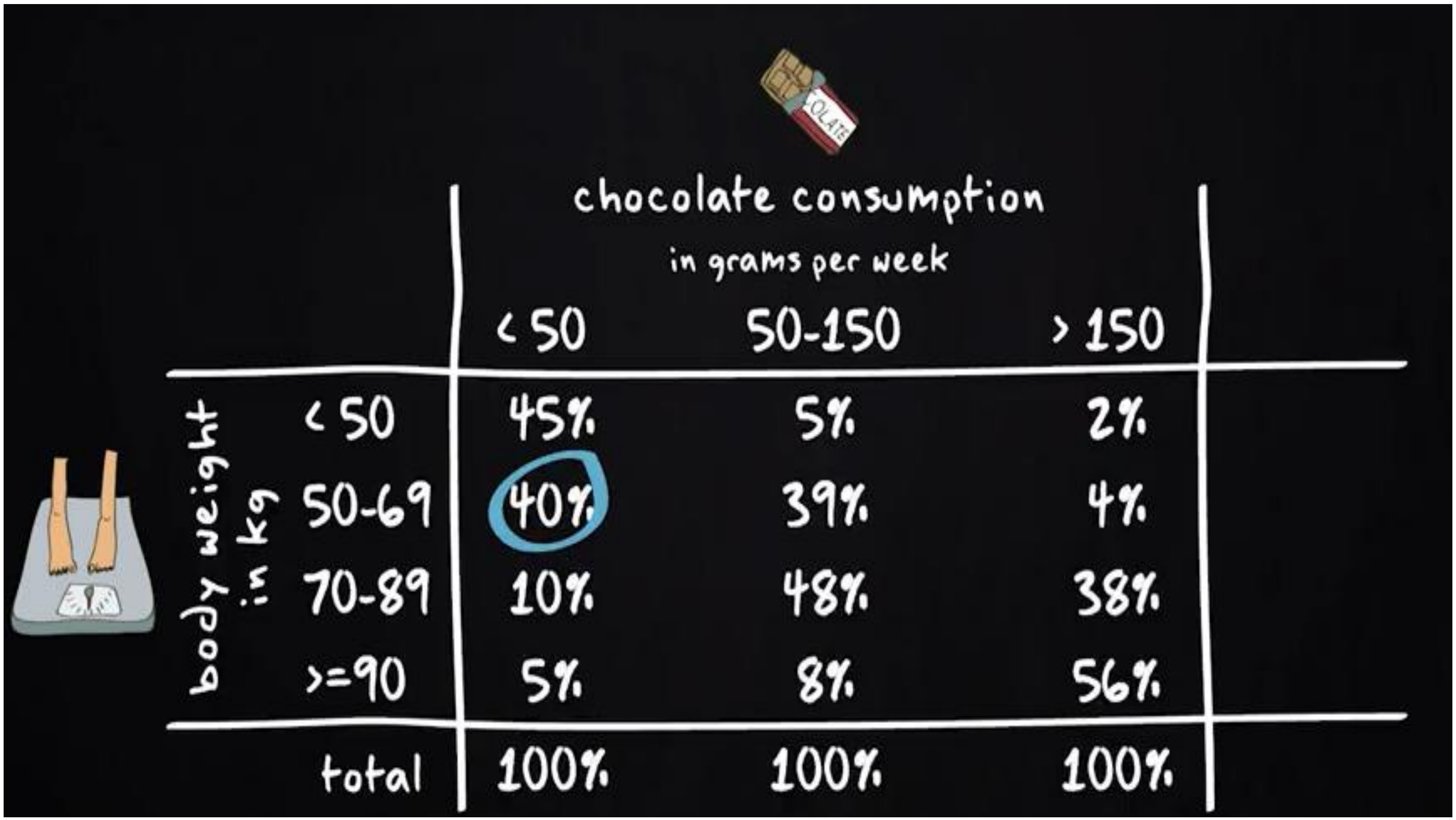
	< 50	50-150	> 150	total
< 50	27	5	1	33
50-69	24	35	2	61
70-89	6	43	19	68
>=90	3	7	28	38
total	60	90	50	200

CROSSTABS (CONTINGENCY TABLES)



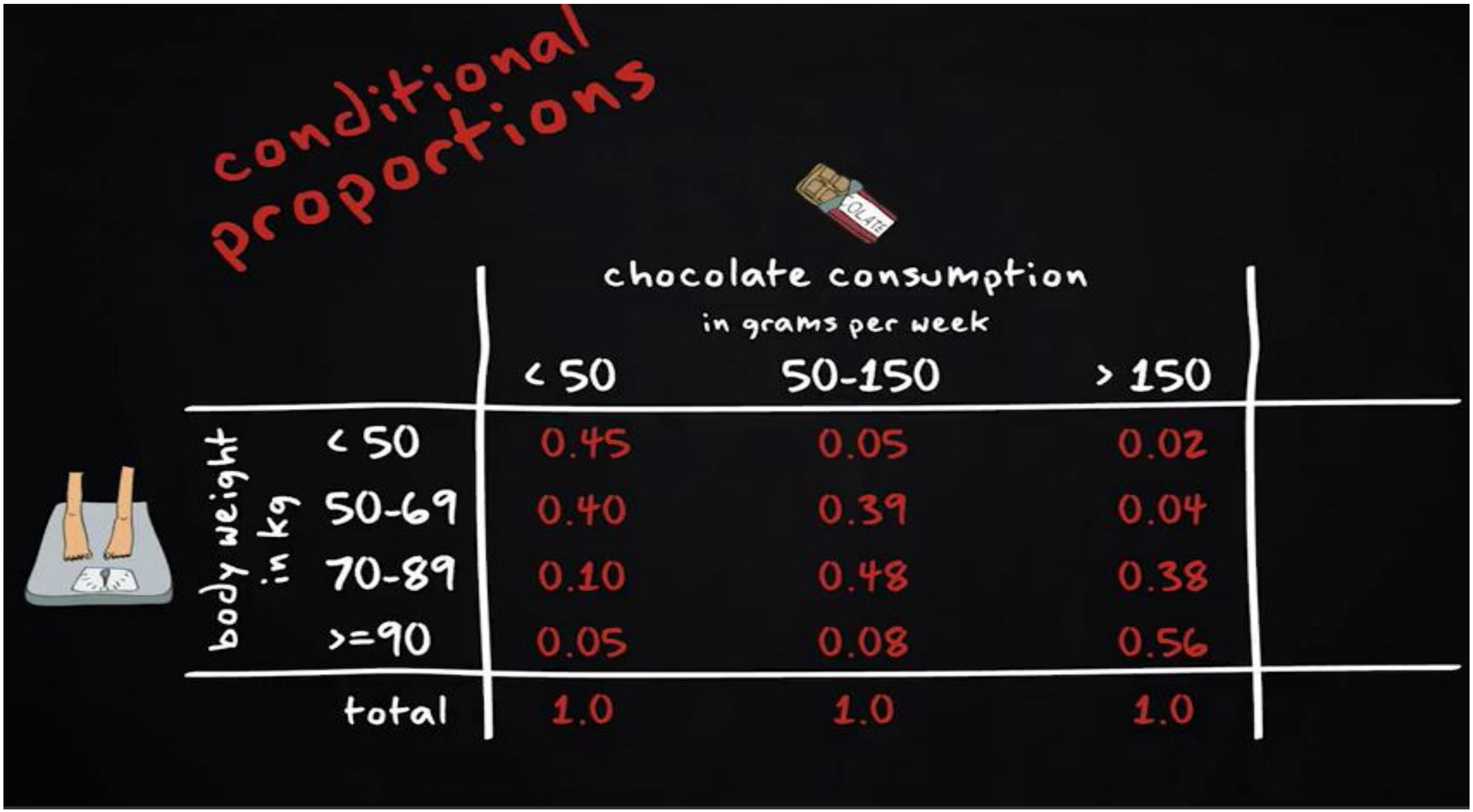


CROSSTABS (CONTINGENCY TABLES)



		chocolate consumption in grams per week		
		< 50	50-150	> 150
body weight in kg	< 50	45%	5%	2%
	50-69	40%	39%	4%
	70-89	10%	48%	38%
	>=90	5%	8%	56%
	total	100%	100%	100%

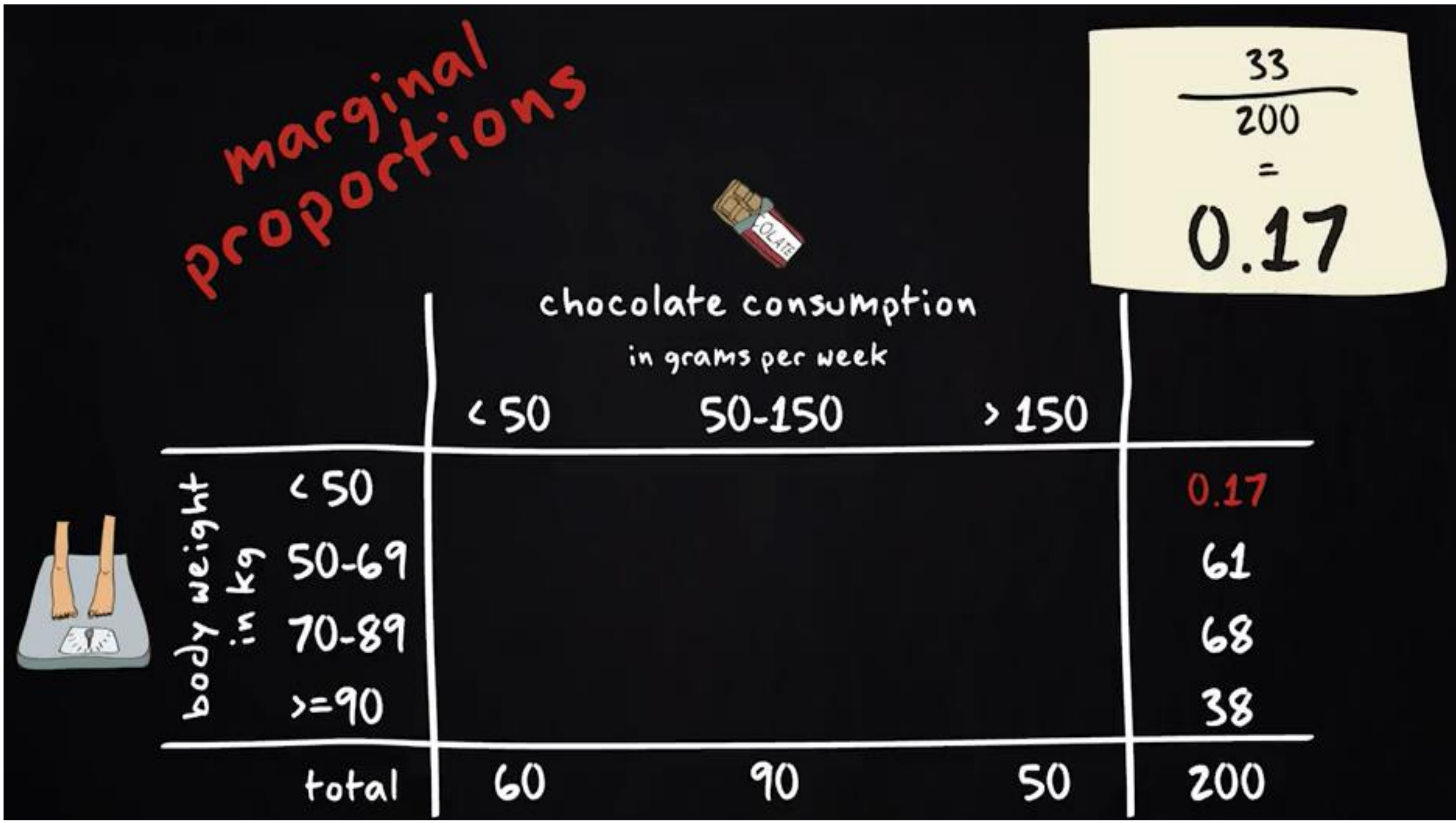
CROSSTABS (CONTINGENCY TABLES)



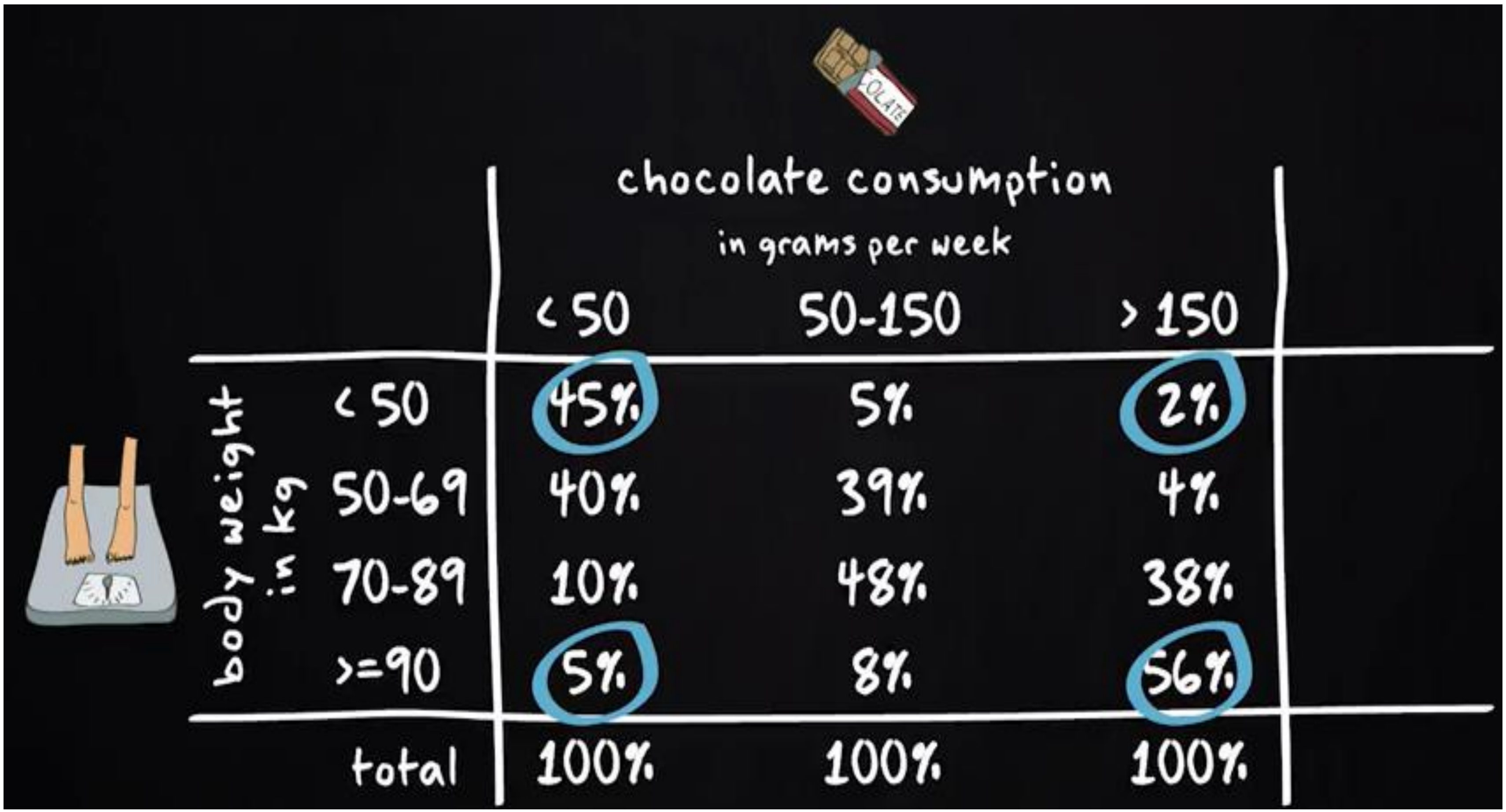
body weight in kg	chocolate consumption in grams per week: < 50	50-150	> 150
< 50	0.45	0.05	0.02
50-69	0.40	0.39	0.04
70-89	0.10	0.48	0.38
>=90	0.05	0.08	0.56
total	1.0	1.0	1.0

CROSSTABS (CONTINGENCY TABLES)

CROSSTABS (CONTINGENCY TABLES)

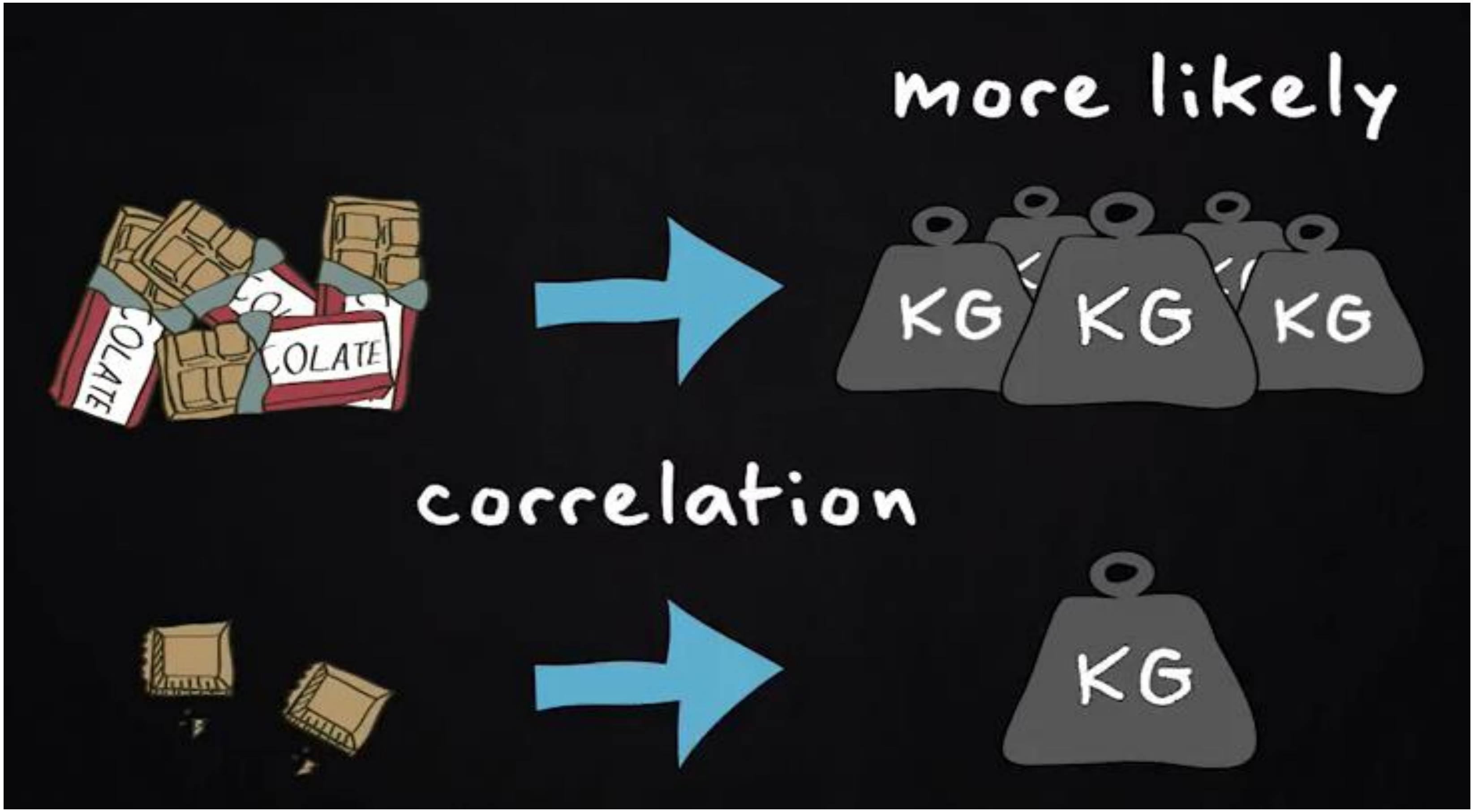


CROSSTABS (CONTINGENCY TABLES)

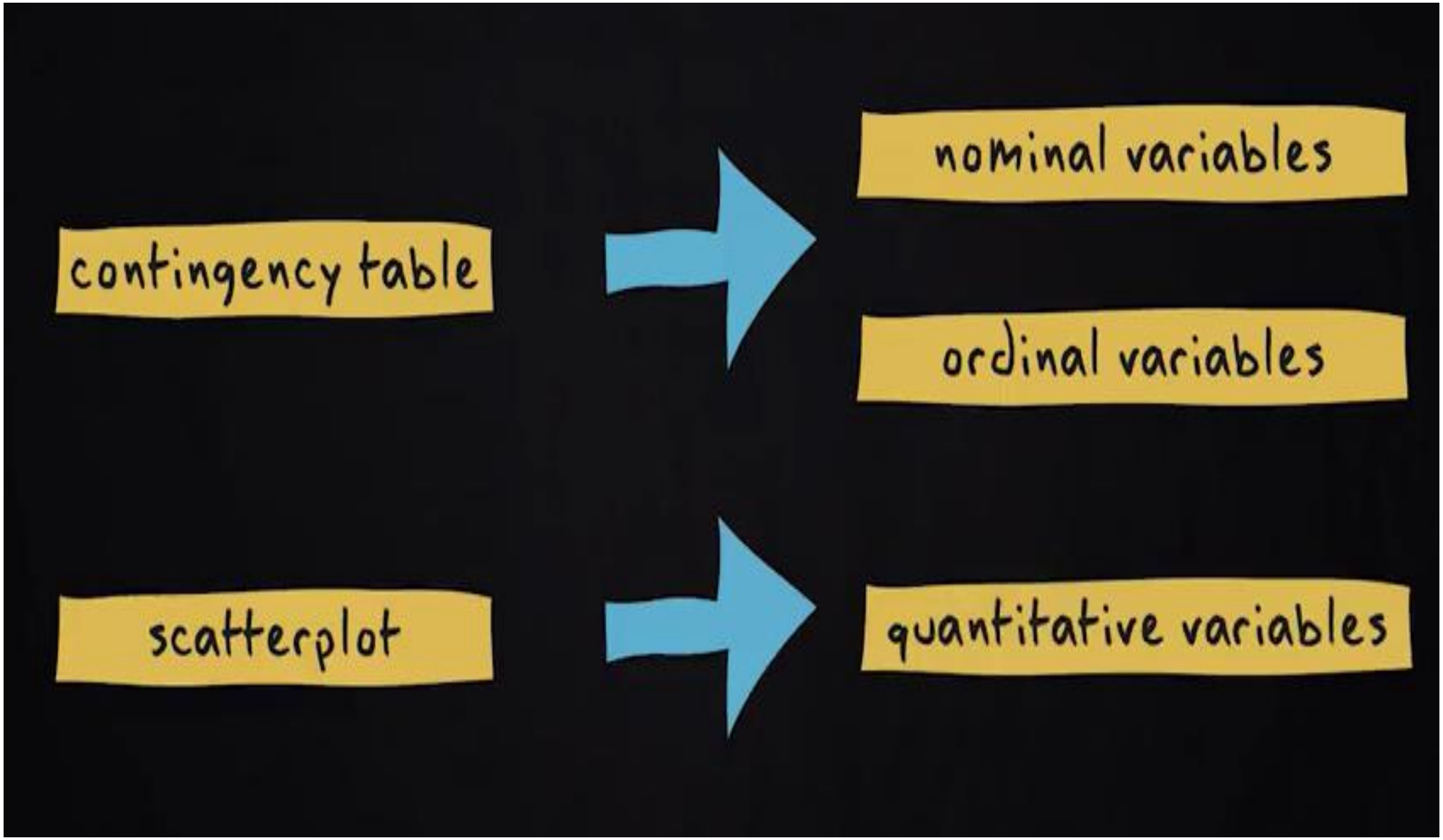


body weight in kg	chocolate consumption in grams per week < 50	50-150	> 150
< 50	45%	5%	2%
50-69	40%	39%	4%
70-89	10%	48%	38%
>=90	5%	8%	56%
total	100%	100%	100%

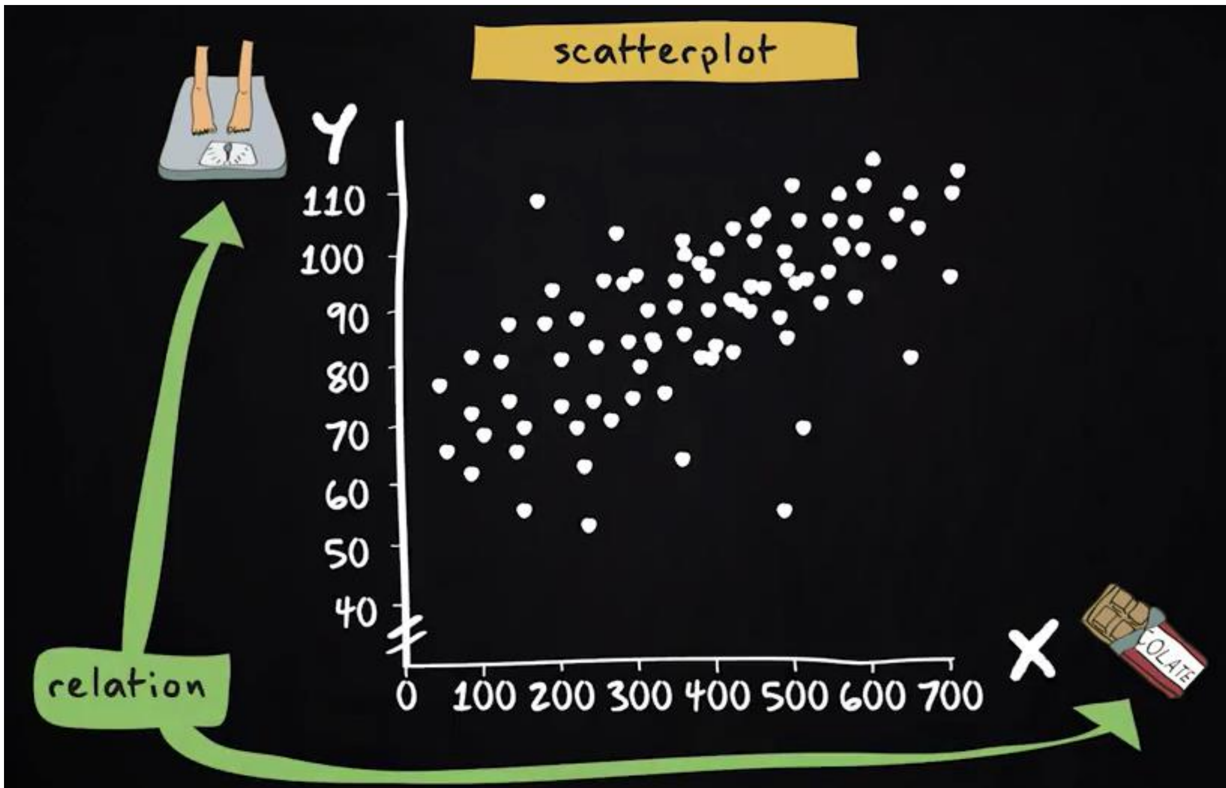
CROSSTABS (CONTINGENCY TABLES)

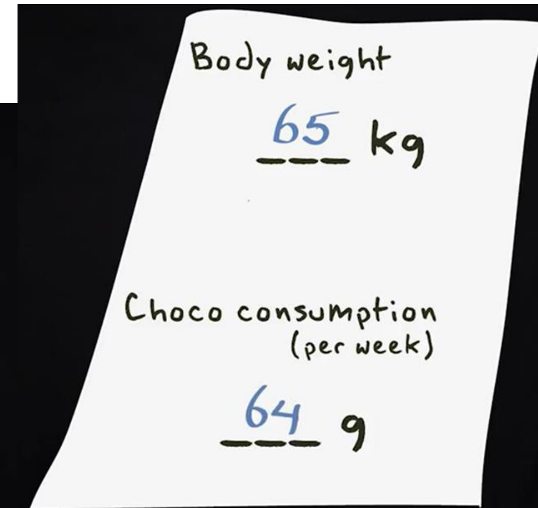


SCATTER PLOTS



SCATTER PLOTS

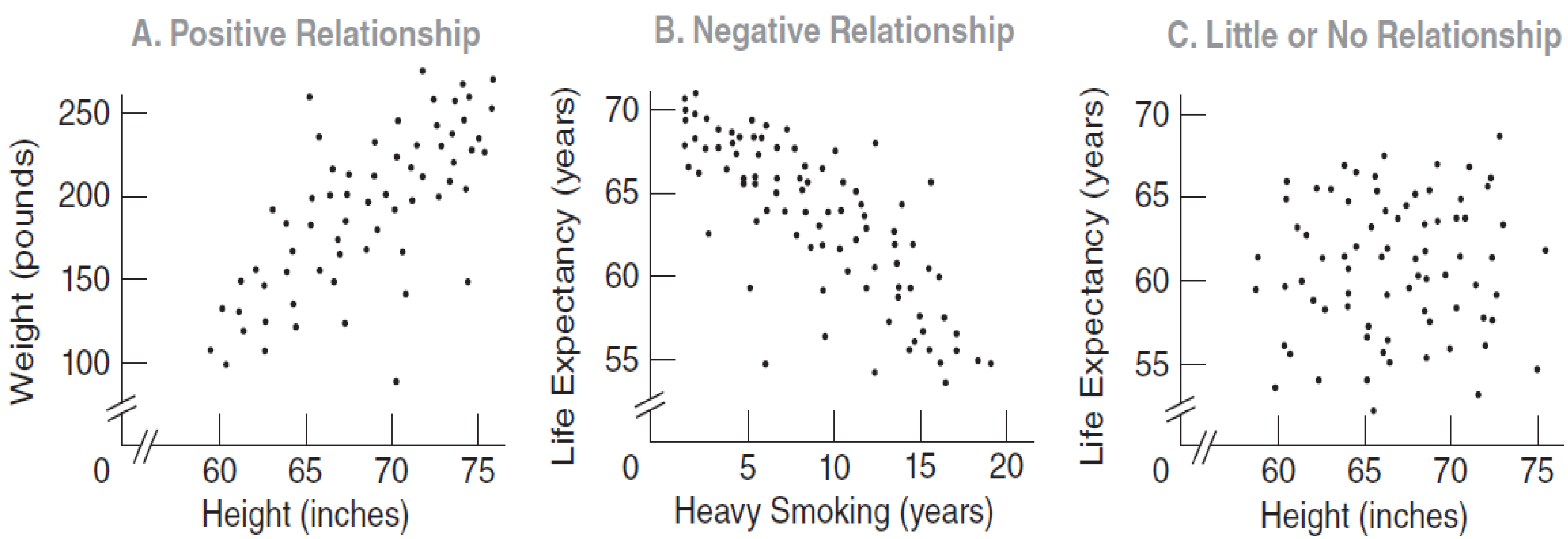




SCATTER PLOTS

TYPE OF RELATIONSHIP "DIRECTION"

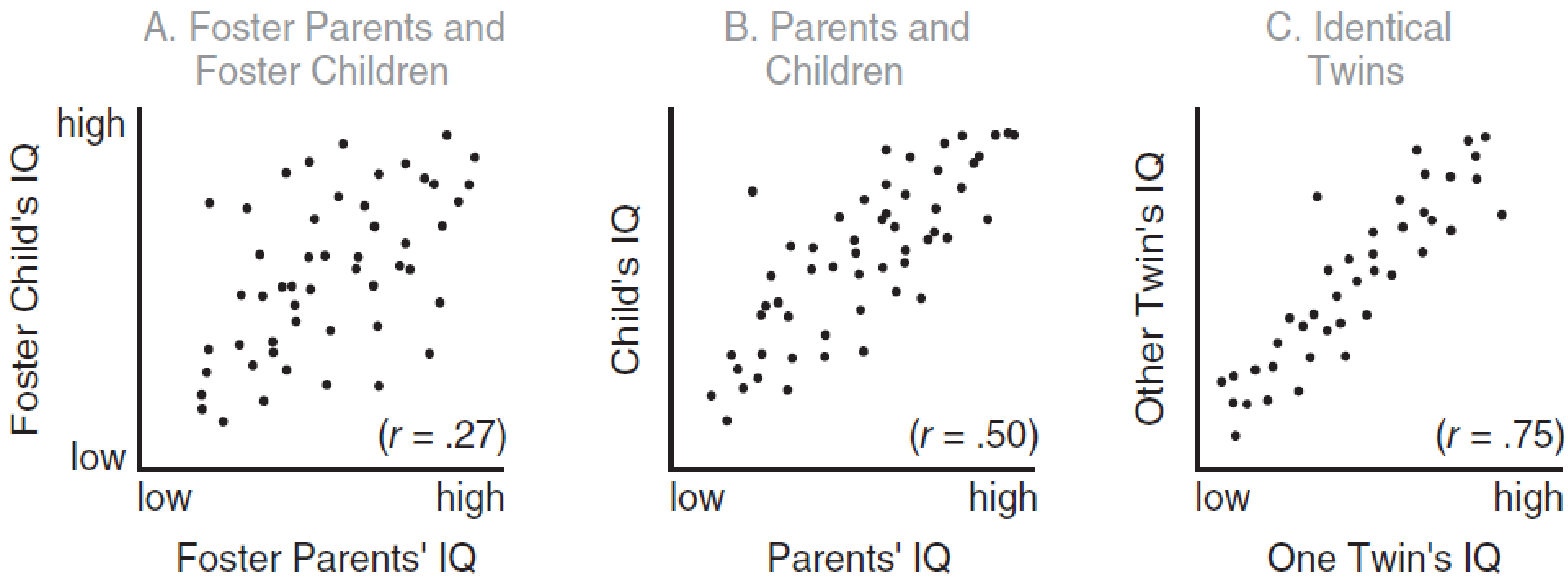
POSITIVE, NEGATIVE, OR NO RELATIONSHIP



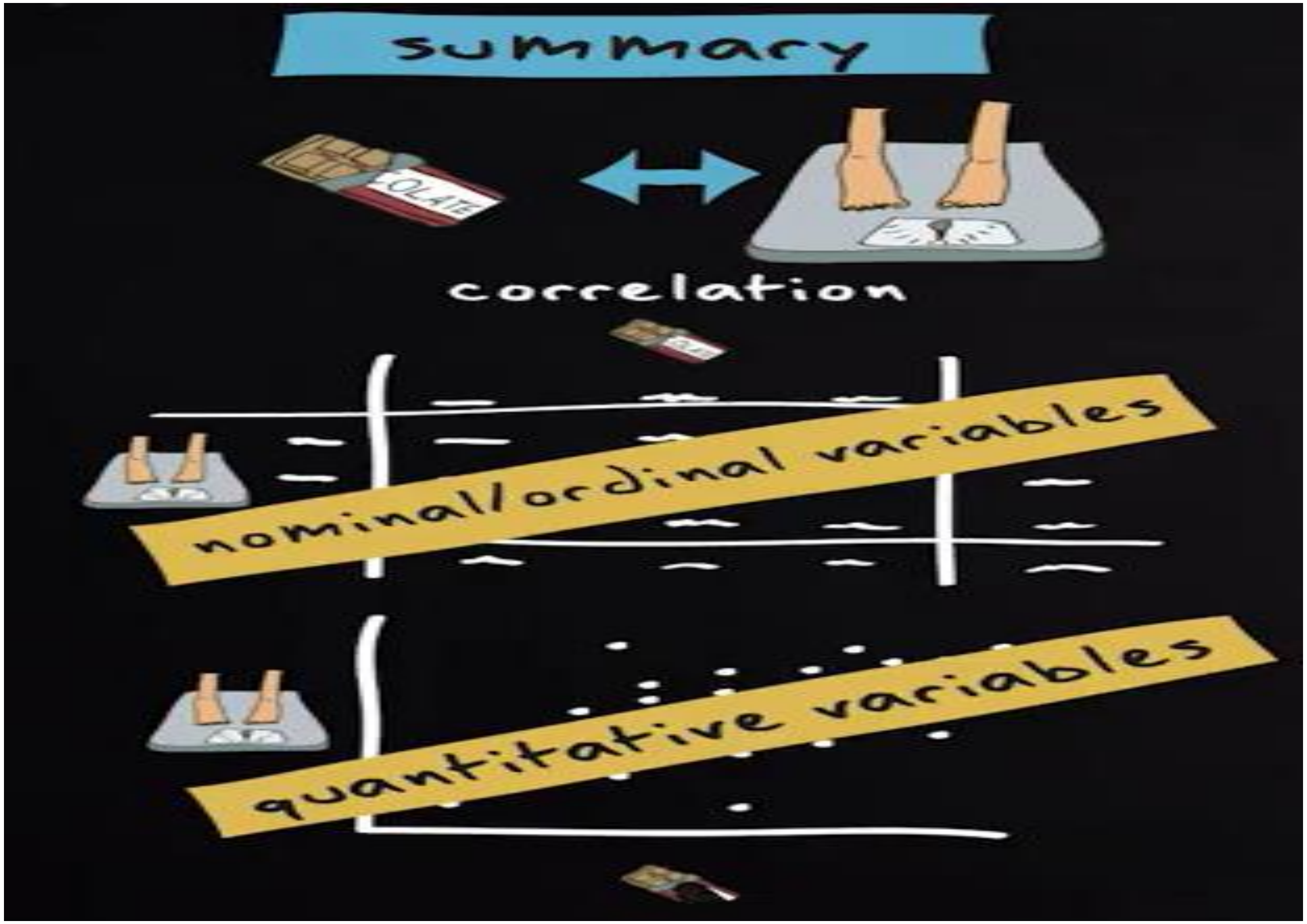
SCATTER PLOTS

STRENGTH OF RELATIONSHIP

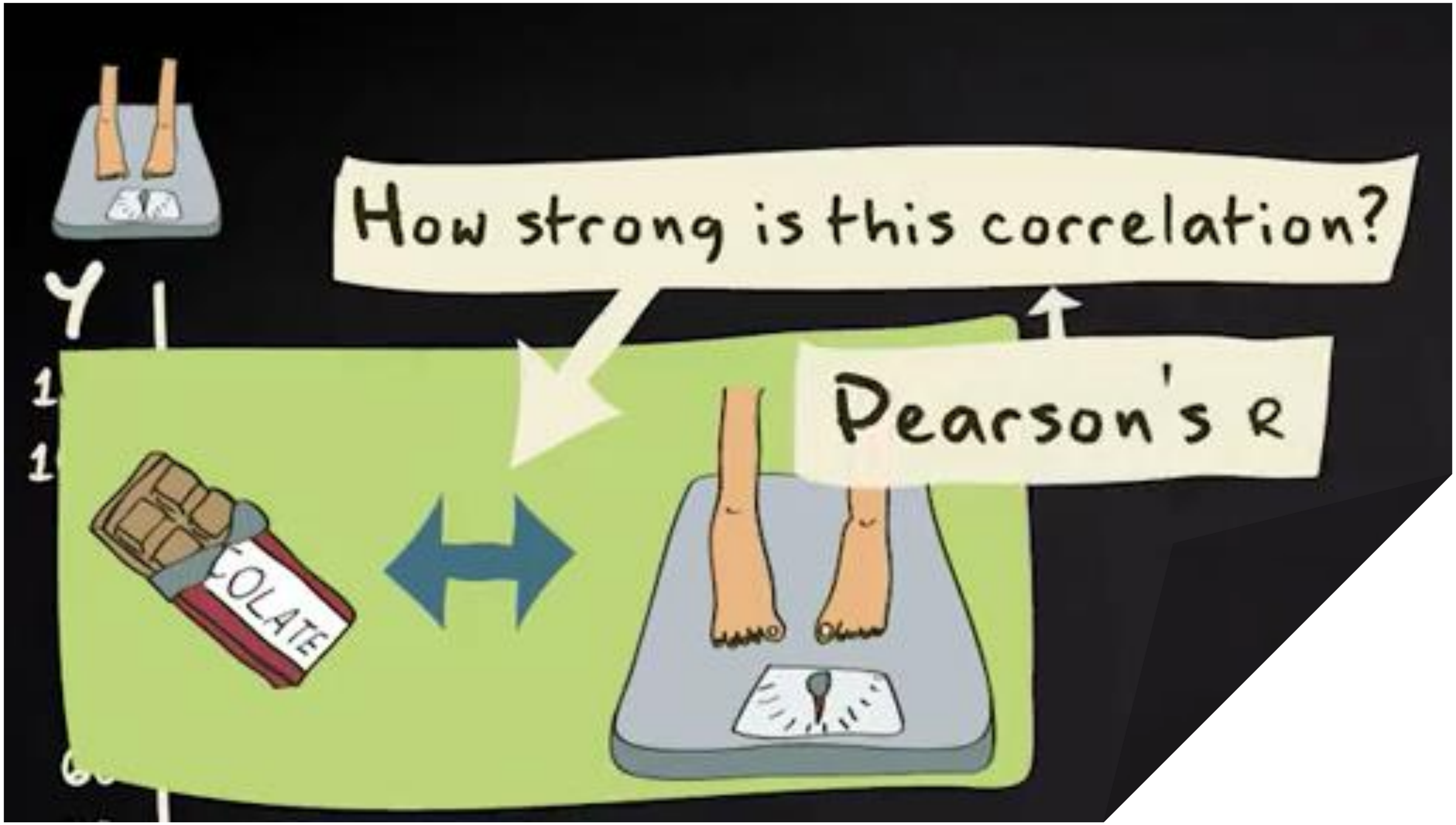
STRONG OR WEAK RELAIONSHIP



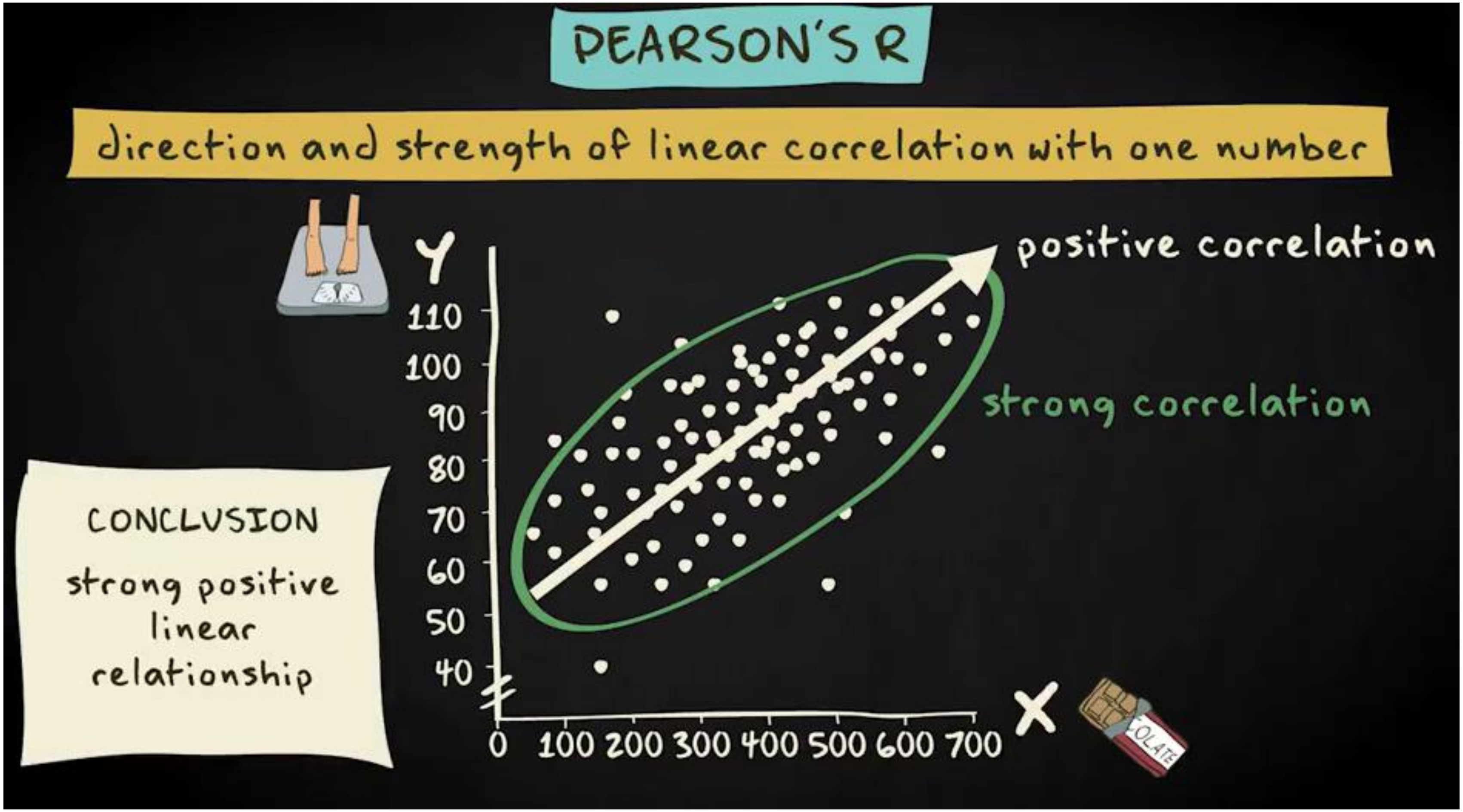
SCATTER PLOTS



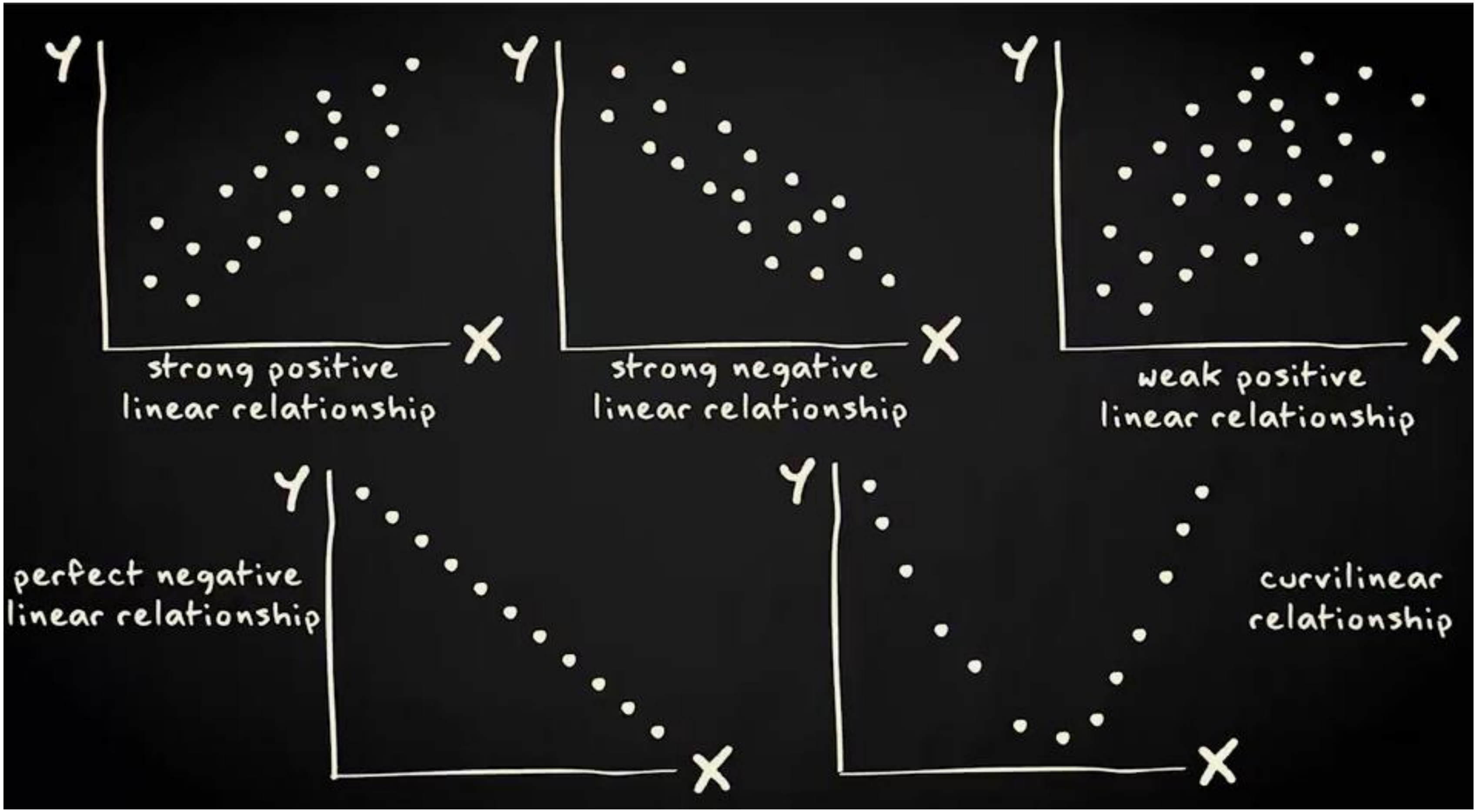
CORRELATION



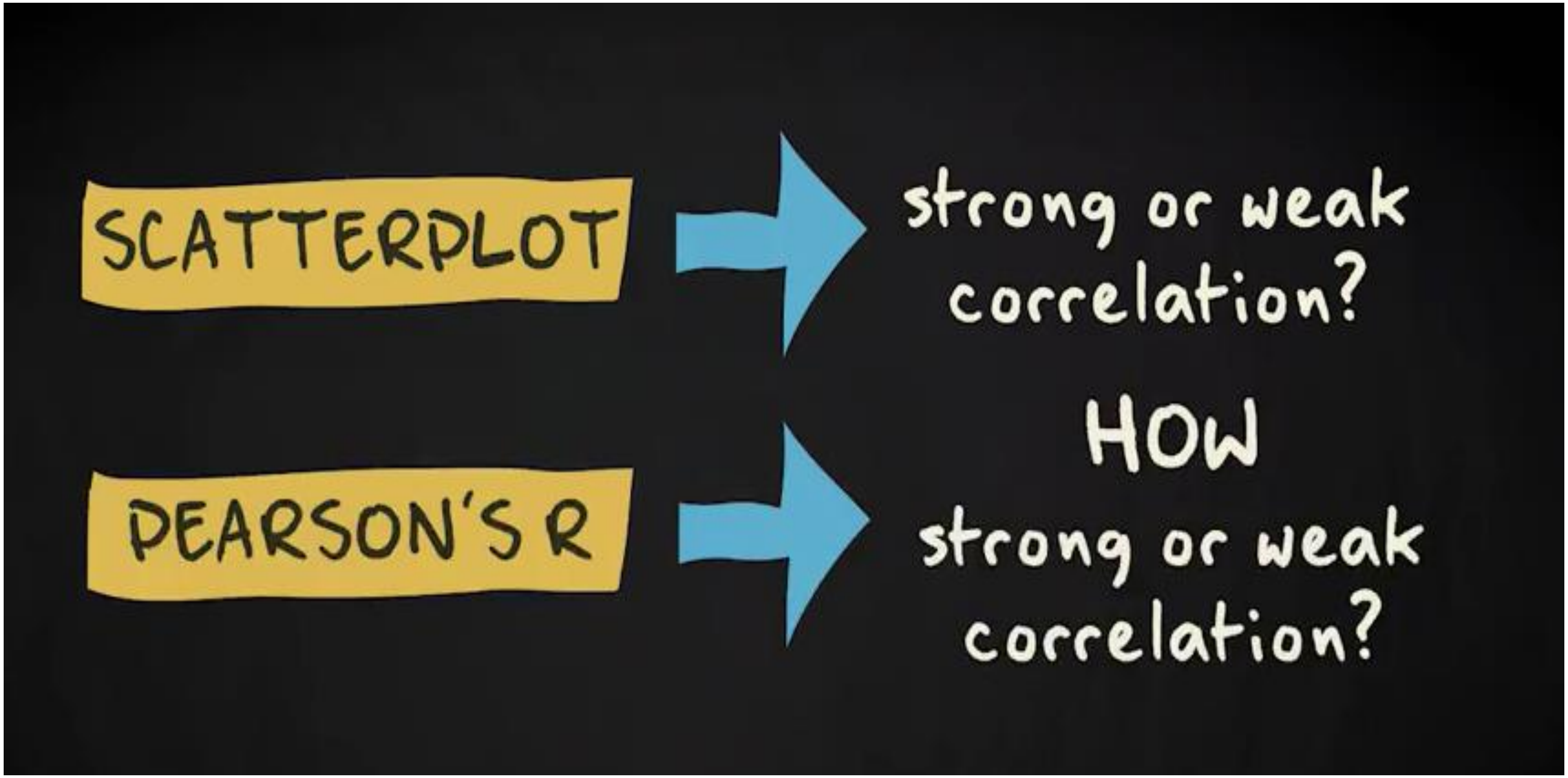
CORRELATION



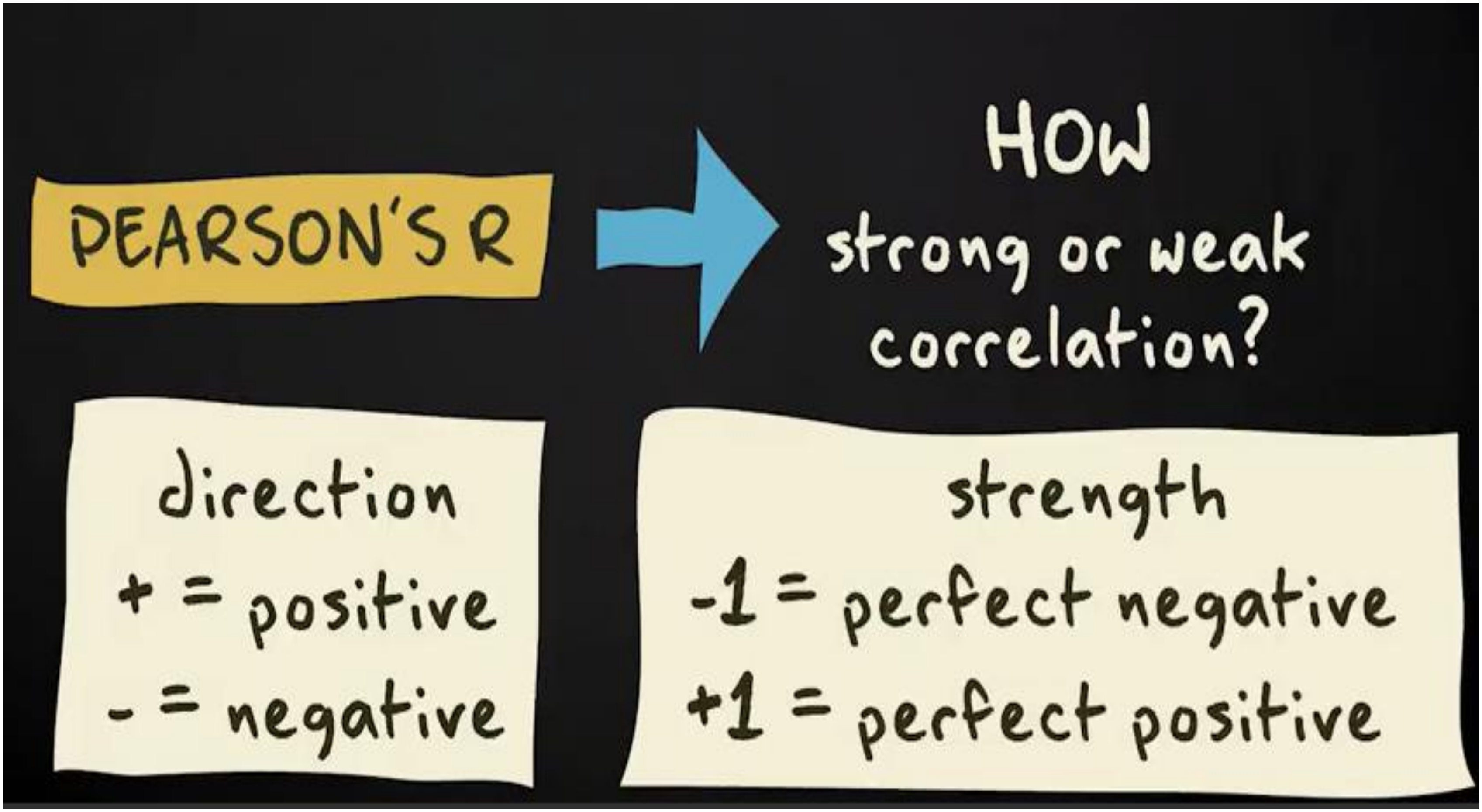
CORRELATION



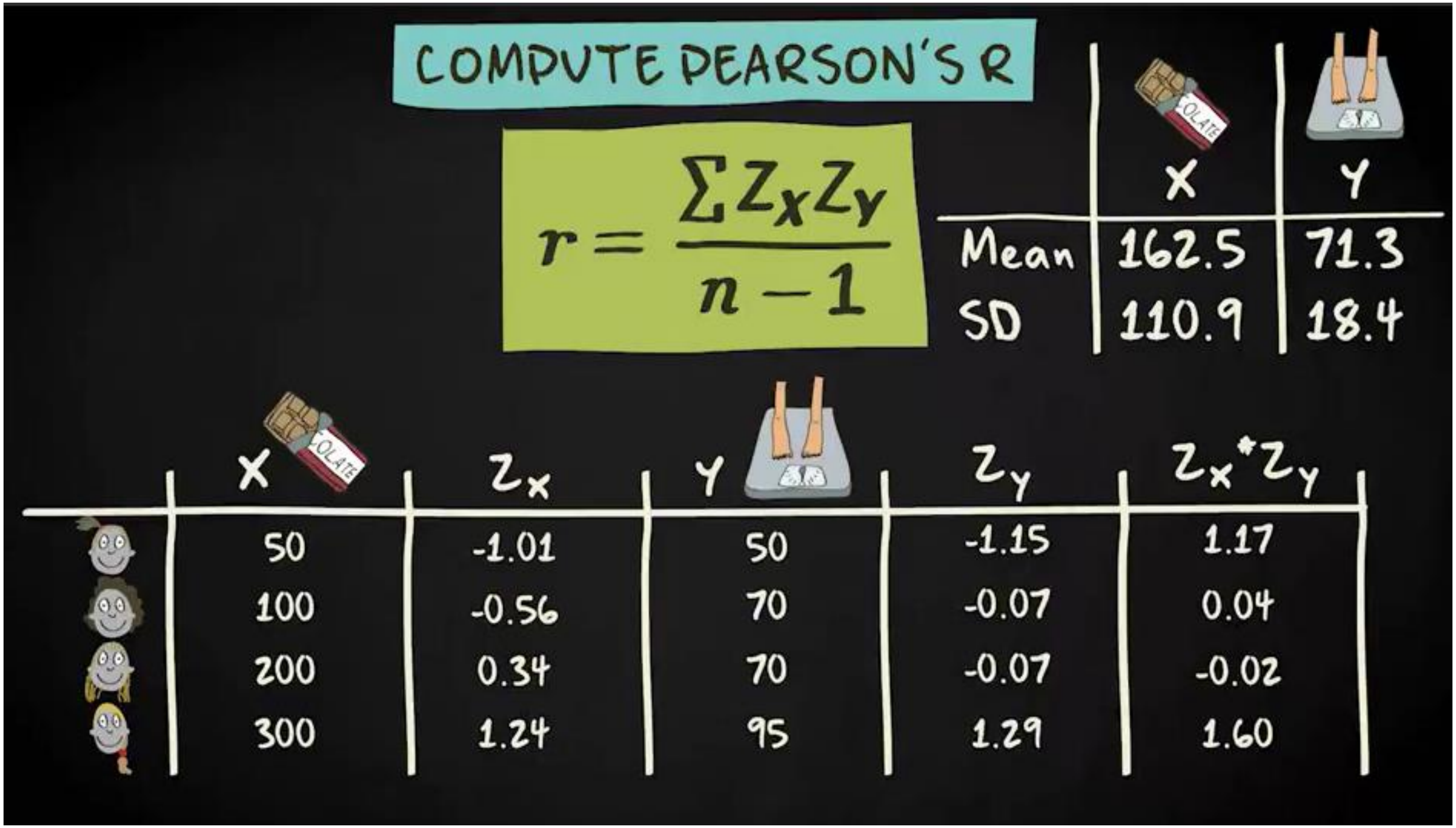
CORRELATION



CORRELATION



CORRELATION



CORRELATION

COMPUTE PEARSON'S R

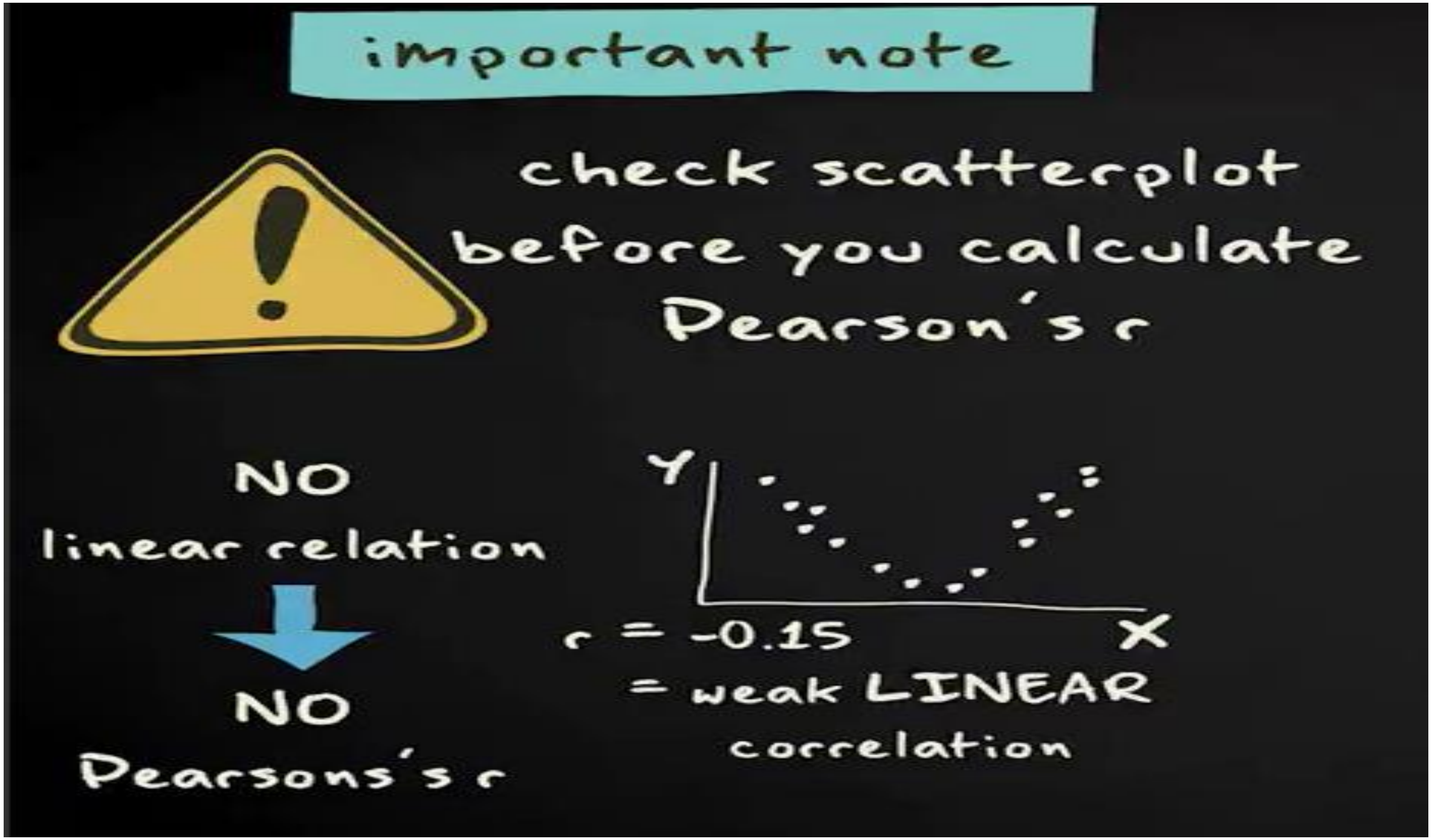
$$r = \frac{2.78}{4-1}$$

0.93

	X	Y
Mean	162.5	71.3
SD	110.9	18.4

X	Z_x	Y	Z_y	$Z_x * Z_y$
50	-1.01	50	-1.15	1.17
100	-0.56	70	-0.07	0.04
200	0.34	70	-0.07	-0.02
300	1.24	95	1.29	1.60
			+	2.78

CORRELATION



CORRELATION

The coefficient of determination r^2

$$0 \le r^2 \le +1$$

Example :

If $r^2 = 0.86$

This means that 86% of the variation in y can be described by x.

LINEAR REGRESSION

REGRESSION ANALYSIS

- Deals with finding the best relationship between Y and X, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the X.

LINEAR REGRESSION

SIMPLE REGRESSION

$$Y = \beta_0 + \beta_1 x$$

LINEAR REGRESSION

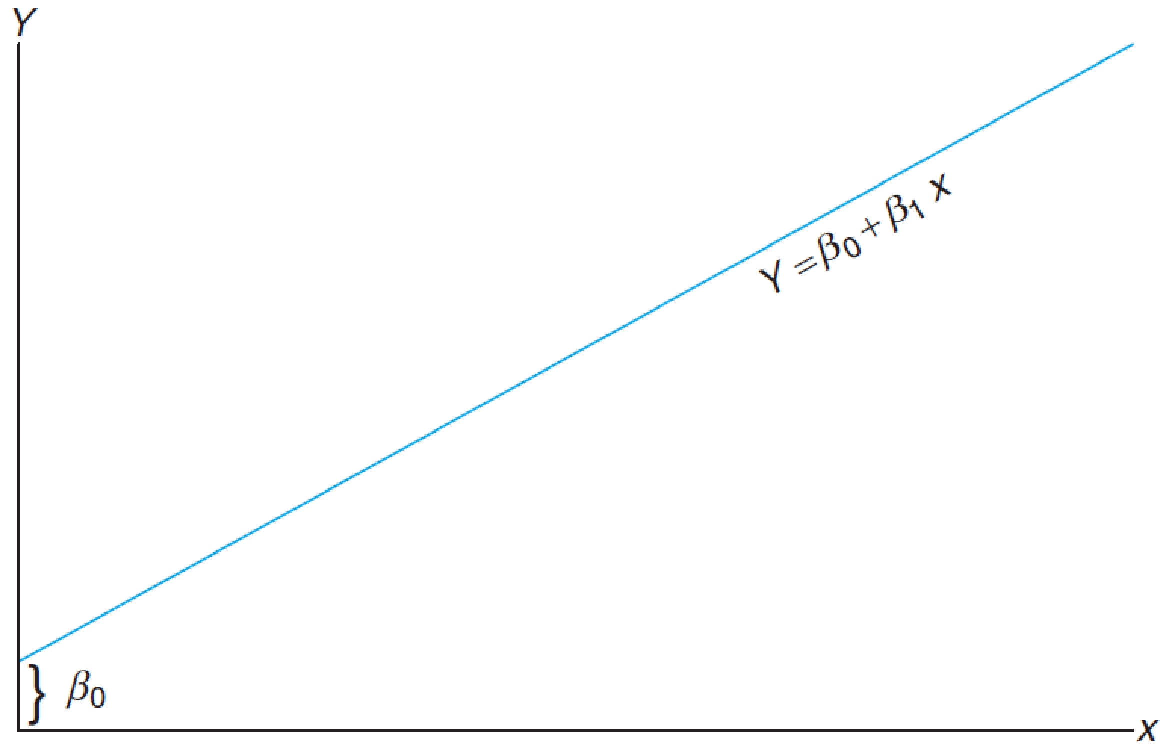
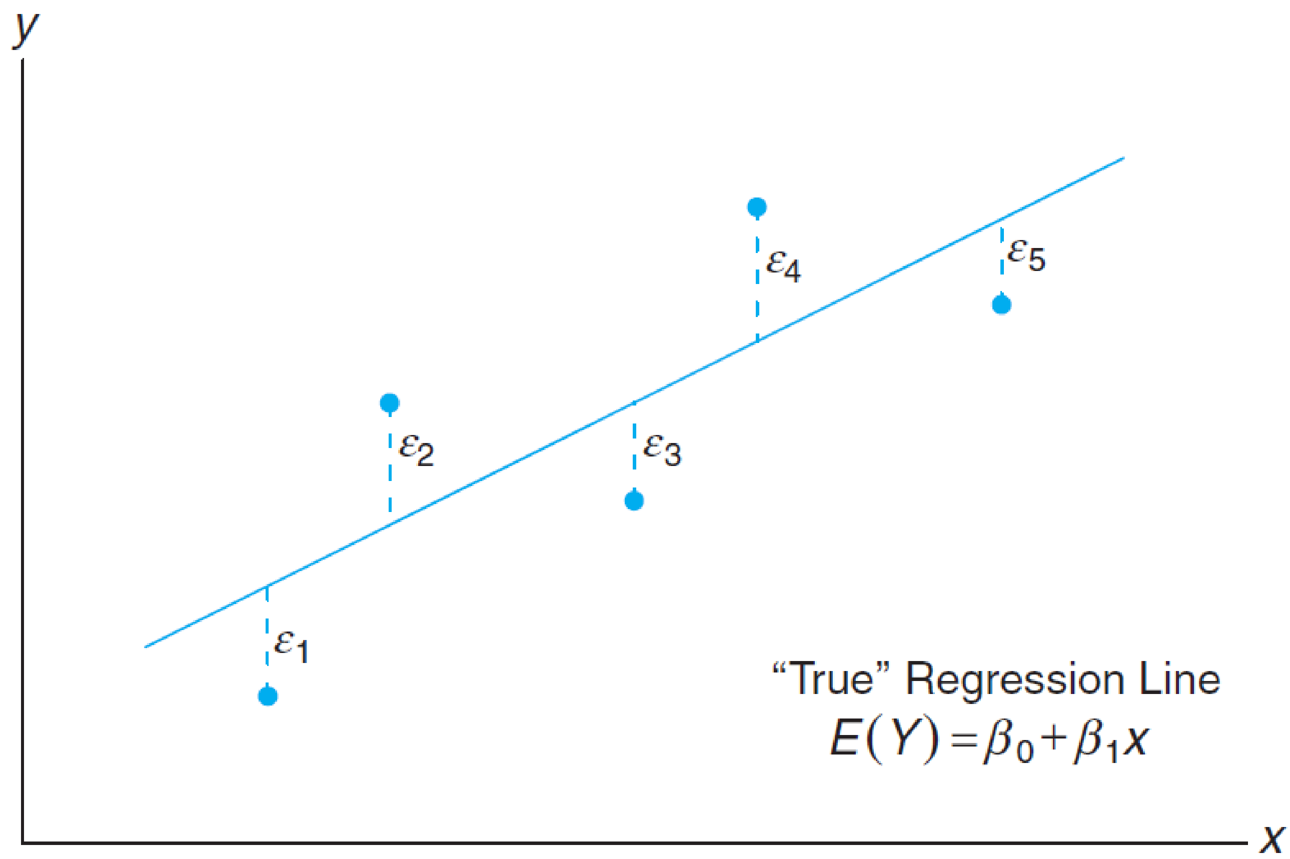
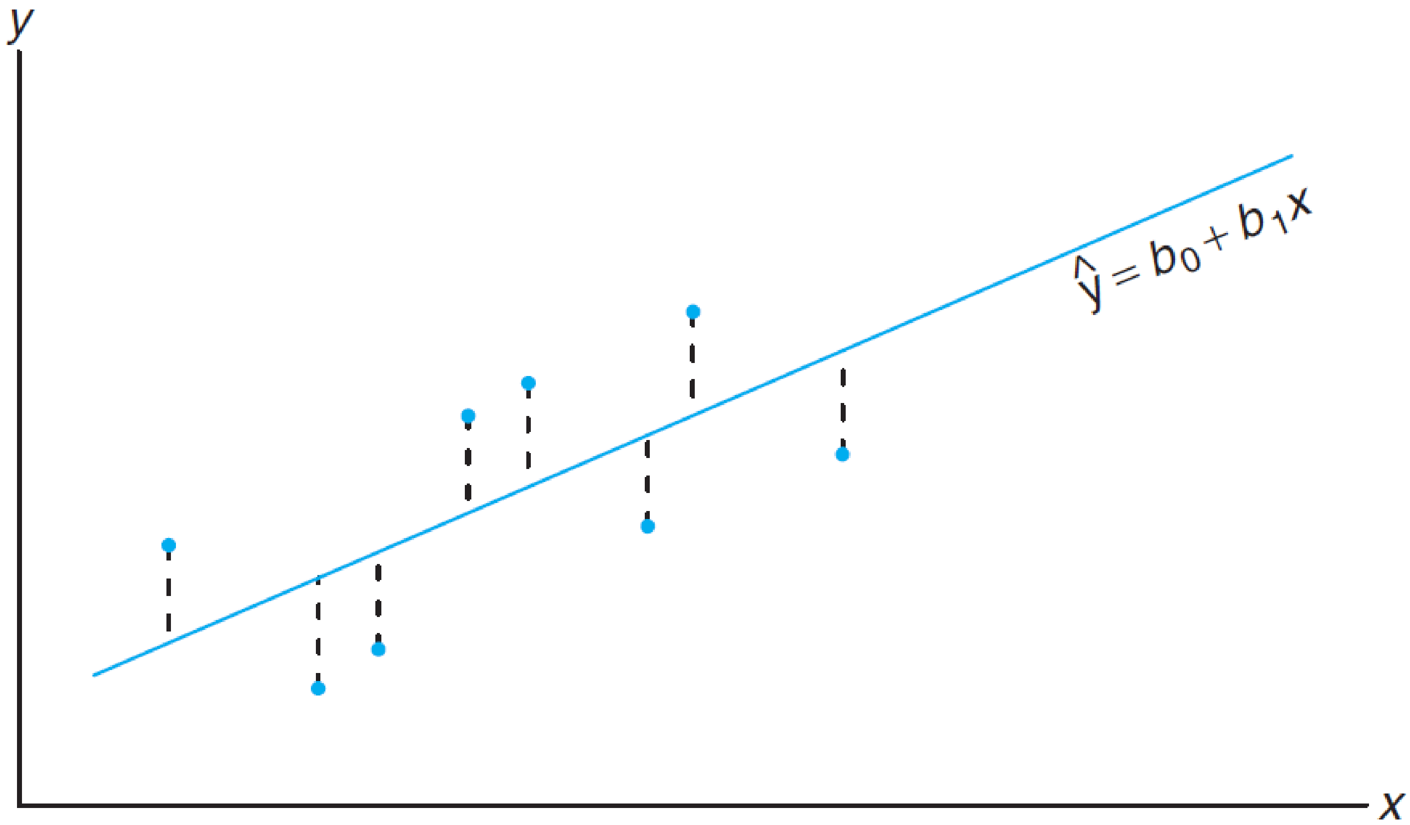


Figure 11.1: A linear relationship; β_0: intercept; β_1: slope.

LINEAR REGRESSION



LINEAR REGRESSION



LINEAR REGRESSION

There are different forms of these formulas
Don't get confused please ☺

LINEAR REGRESSION EQUATION

$$\hat{y} = b_1 x + b_0$$

$$b_1 = r\left(\frac{S_y}{S_x}\right)$$

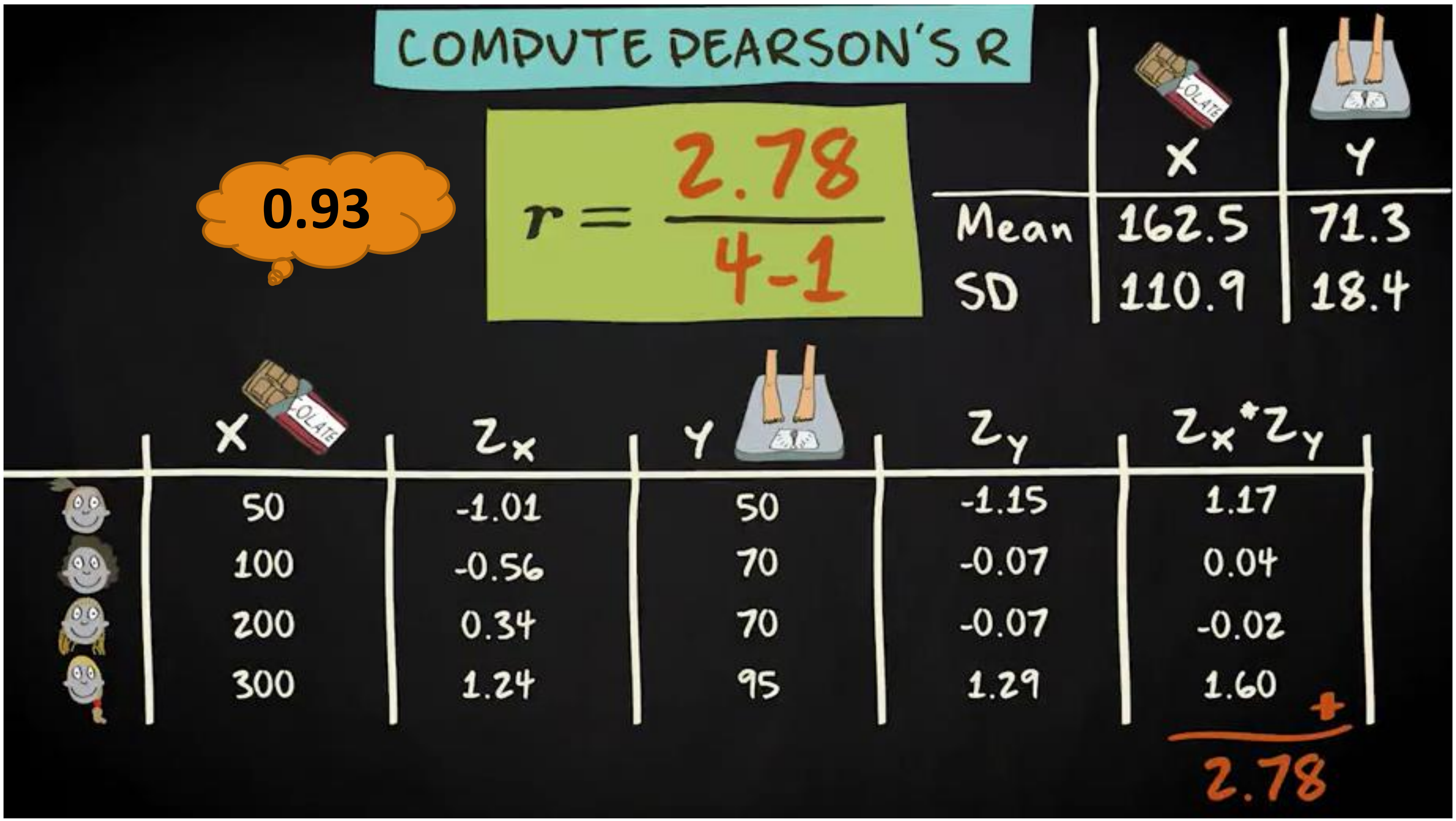
$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b_0 = \bar{y} - b_1(\bar{x})$$

$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

EXAMPLE (1) ON CORRELATION



	X	Y
Mean	162.5	71.3
SD	110.9	18.4

X	Z_X	Y	Z_Y	$Z_X * Z_Y$
50	-1.01	50	-1.15	1.17
100	-0.56	70	-0.07	0.04
200	0.34	70	-0.07	-0.02
300	1.24	95	1.29	1.60
			+	2.78

EXAMPLE (1) ON CORRELATION

(X)	50	100	200	300	
Z_X	-1.01	-0.56	0.34	1.24	
(Y)	50	70	70	95	
Z_Y	-.1.15	-0.07	-0.07	1.29	
$Z_X * Z_Y$	1.17	0.04	-0.02	1.60	$\sum Z_X * Z_Y = 2.78$

	X	Y
Mean	162.5	71.3
SD	110.9	18.4

$$Z_X = \frac{X - \overline{X}}{s_X}$$

$$Z_Y = \frac{Y - \overline{Y}}{s_Y}$$

$$r = \frac{\sum ZX * ZY}{n-1} = \frac{2.78}{3} = \underline{0.93}$$

Strong Positive or Direct Relationship

EXAMPLE (1) ON CORRELATION

ii. What would be the values of Y at X = 400 and 500?

$$\hat{y} = bo + b_1 X$$

b_1 = r $\frac{Sy}{Sx}$ = 0.93 $\frac{18.4}{110.9}$ = **0.154**

b_o = $\bar{y}$ – b_1 $\bar{x}$ = 71.3- (0.154)(162.5) = **46.275**

$\hat{y} = bo + b_1 X$ = $\boldsymbol{\hat{y} = 0.154\, X + 46.275}$

At X = 400 $\boldsymbol{y\hat{} = 0.154\,(400) + 46.275}$ **= 107.875**

At X = 500 $\boldsymbol{y\hat{} = 0.154\,(500) + 46.275}$ **= 123.275**

EXAMPLE (1) ON CORRELATION

iii. What is the error in the predicted value of Y at X = 200 and 300?

$$\hat{y} = 0.154\,X + 46.275$$

At X = 200 $\quad \hat{y} = 0.154\,(200) + 46.275 = 77.075$

Error = |y^ - y| = |77.075 – 70|= 7.075

At X = 300 $\quad \hat{y} = 0.154\,(300) + 46.275 = 92.475$

Error = |y^ - y| = |92.475 – 95|= 2.525