Unit 3: Correlation

Definition: a correlation is a link between two separate distributions

# of hours studying ↑

nutrition ?

nutrition ↑

nutrition ↑

exercise ↑

work experience

plants/trees ↑

plants/trees ↑

housing development ↑

fox population ↑

amount of virgin plastic created ↑

marks on an exam ↑ positive correlation

grades ↑ · positive weak / moderate

health ↑ positive

weigth ↓ negative correlation

energy ↑

salary

happiness ↑

CO_2 ↓ negative

habitats ↓ negati

rabbit population ↓ negative

population ↑ → destruction of habitats ↑

2 ways to graph two distributions

Joint Distribution Table

p 3.3

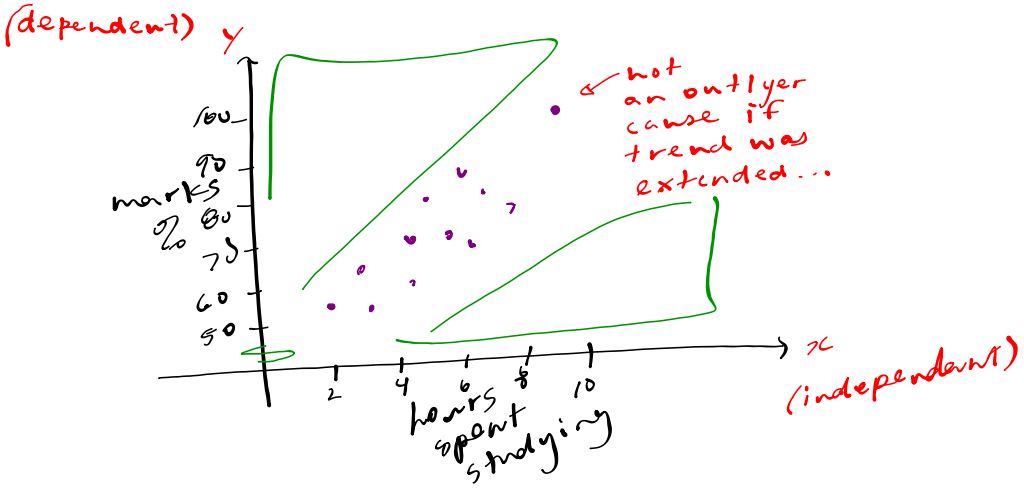
y marks \ x hours spent studying	2h	3h	4h	5h	6h
[40, 50[	1				
[50, 60[	2	3	1		
[60, 70[		4	5	4	
[70, 80[			1	2	
[80, 90[					

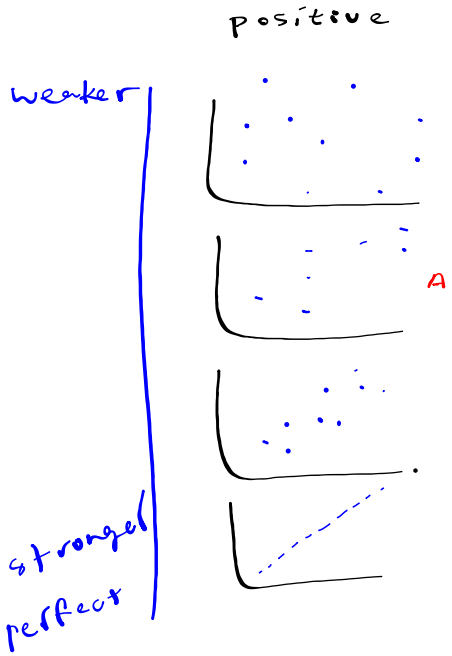
* if corners are more or less empty, then there's a correlation

* higher frequency around diagonal means stronger correlation.

outlyer - a data point that doesn't follow trend.

Scatter Plot: used to graph 2 distributions

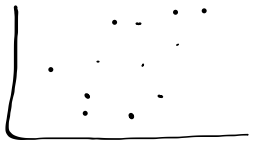



positive
weaker
stronger
perfect
A

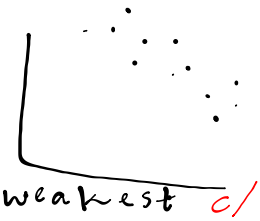

negative
weight
healthy food
B

Typical Exam Question

a) 

b) 

c) 

d) 

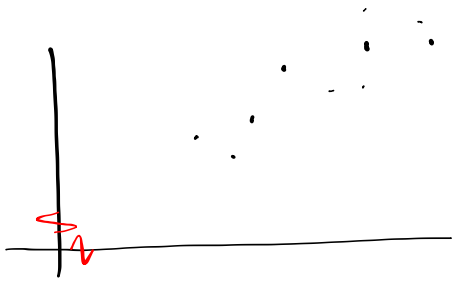
- Rank from strongest to weakest c/d/b/a/
- match scenario with graph
  - i) amount water consumed ↑ v.s. how hydrated you are ↑
  - ii) how many eggs you eat ↑ v.s. # of dates you have. no correlation

p 3.29

graph:

life exp men y

life exp women x



· label axis

p 3.36

Review

1-3

5

Unit 4: Correlation Coefficient (r)

↳ a # in between -1 and 1

↳ It describes the strength and direction/sign of the link between 2 distributions

$r = 1$ → perfect positive

$0.75 \leq r < 1$ → strong positive

$0.6 \leq r < 0.75$ → moderate positive

$0.4 \leq r < 0.6$ → weak positive

$0 \leq r < 0.4$ → no correlation

$-0.4 < r \leq 0$ → no correlation

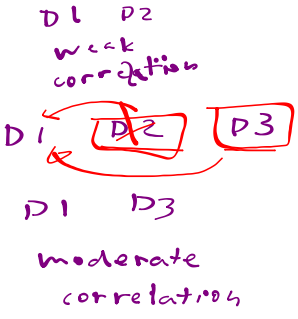
$-0.6 < r \leq -0.4$ → weak negative

$-0.75 < r \leq -0.6$ → moderate negative

$-1 < r \leq -0.75$ → strong negative

$r = -1$ → perfect negative

no action
no predictions



electric cars ↑

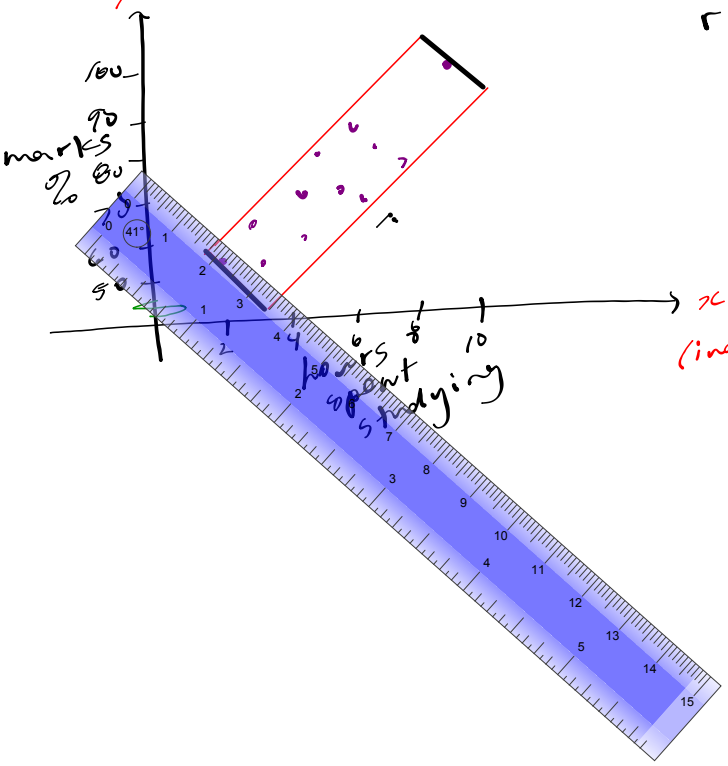
vs CO_2 ↓

3 ways to calculate r

steps pg 4.11

- graphically
- rectangle method

(dependent) y



(independent)

$$r = \pm\left(1 - \frac{\text{short}}{\text{long}}\right)$$

$$r = +\left(1 - \frac{1.7}{6}\right)$$

$$r = +(0.72)$$

$$r = 0.72$$

moderate positive.

calculate r. life
w men/women
exp. graph.

2nd way to calculate r algebraically

$$r = \frac{\sum_{i=1}^{n} z_{x_i} \times z_{y_i}}{n-1} = \frac{z_{x_1} \times z_{y_1} + z_{x_2} \times z_{y_2} + \cdots + z_{x_n} \times z_{y_n}}{n-1}$$

women's life exp x	men's life exp y	$\frac{x_i - \bar{x}}{s_x}$ " z_{x_i}	$\frac{y_i - \bar{y}}{s_y}$ " z_{y_i}	$z_{x_i} \times z_{y_i}$
~~x~~ 67.6		-1.52		
				sum

$$r = \frac{\text{sum}}{n-1}$$

3rd way to calculate r

→ calculator

p 4.24 - instructions

Stat 0 → 1 distrib

Stat 1 → 2 distrib

try
trial
&
error.

enter in distributions from p 3.29

Q1 • find r

Q2 • who's life expectancy varies more? men or women?

justify!

p 3.36
Review
#1-3
#5

p
4.31
-4.32
1-2
4.33
#3 not d)

#5 p 4.37