Statistics 1 Chapter 4
CORRELATION

⚠ this chapter is special: it has been dumbed down, and online questions may seem very hard.

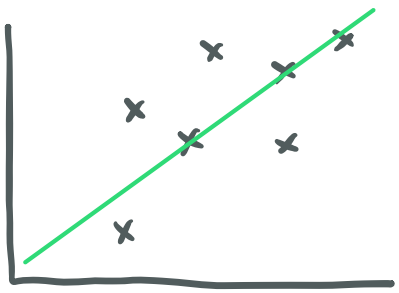
When introducing a second variable, we need to consider the relationship between them

CORRELATION is the strength of said relationship

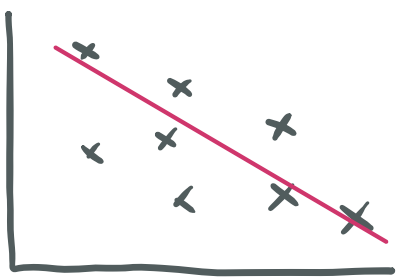
POSITIVE CORRELATIONS

NEGATIVE CORRELATIONS

WEAK CORRELATIONS

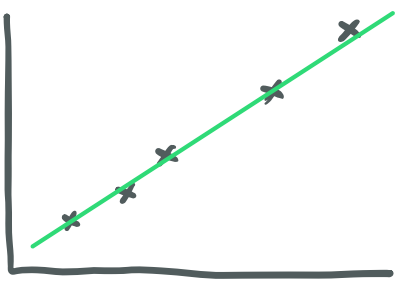


weak positive

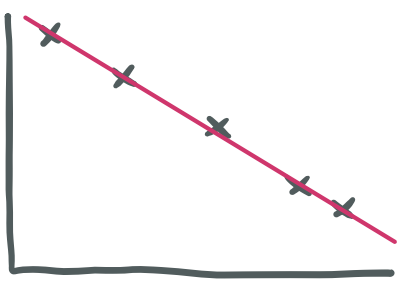


weak negative

STRONG CORRELATIONS



strong positive



strong negative

⚠ Correlation doesn't have to be linear

no correlation

⚠ CORRELATION ≠ CAUSATION

just because there is a correlation between x and y doesn't mean x causes y

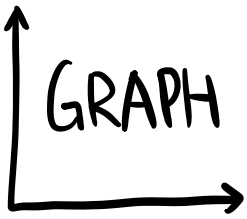
⚠ VARIABLES & AXIS

independent variable on x axis

dependent variable on y axis

EXAM SKILLS

① Interpret the correlation from a graph



⇨ positive / negative correlation
(strength usually not needed)

② Correlation or Causation?

Identify if the variable are causally related
(not casually)

EX 4A (p. 61)(Ans p.215)

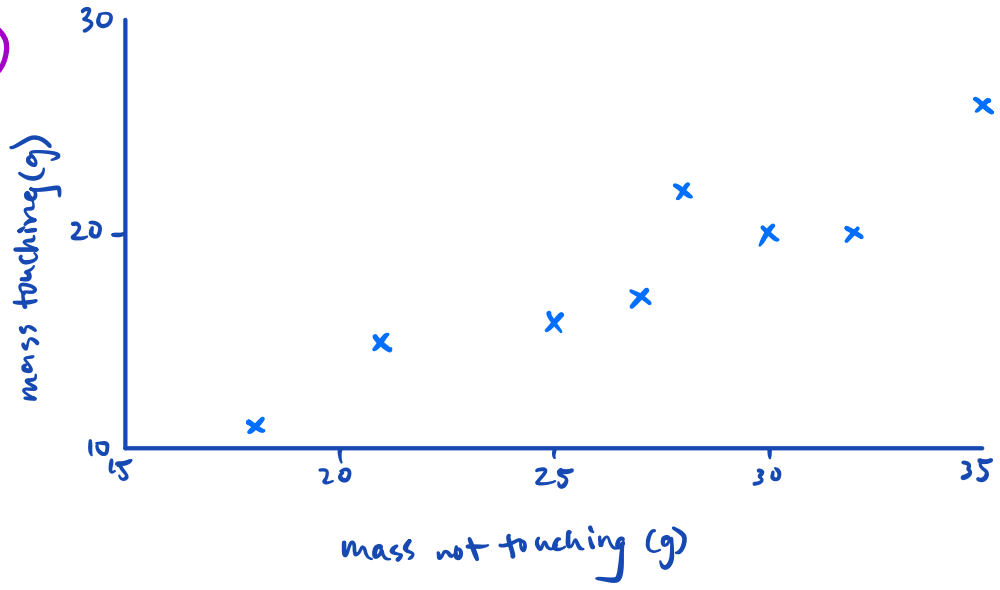
① a) Positive

b) The longer someone takes this drug, the larger the loss in weight

② a) No correlation

b) The scatter graph doesn't support this claim

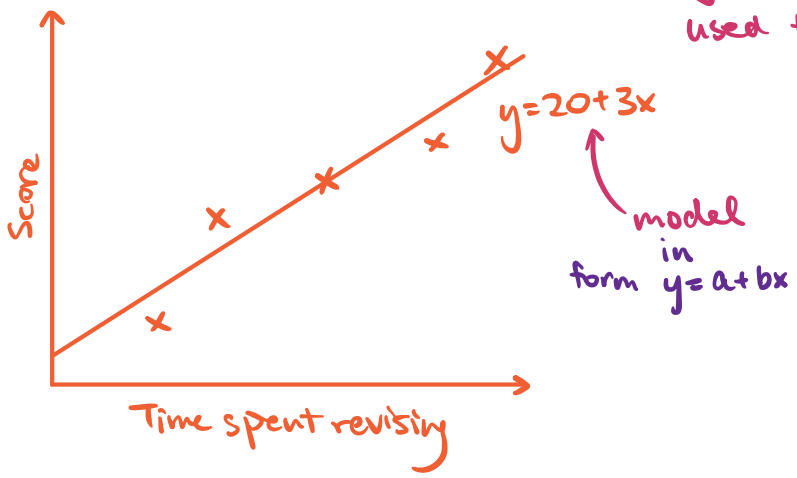
③ a)



b) Positive: when mass not touching increases, mass touching increases

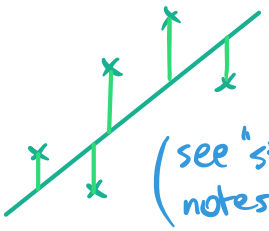
4.2: Linear Regression

used to predict a result



① least squares regression line

minimize $\sum(\hat{y}-y)^2$



No need to perform.

(see "straight line graphs" notes in pure math)

Non-linear regression:

They will NOT give you this!

$y = a+bx$

Interpreting a: y value when x is 0
(e.g. when you don't revise, you will get 20 marks)

Interpreting b: y increase when x increases 1 unit
(e.g. you get 3 marks for every hour you revise)

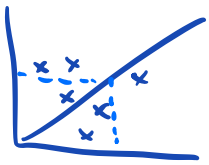
Justifying linear regression

"data suggests linear relationship"

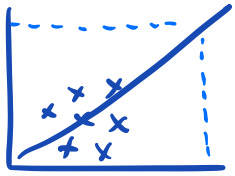
Interpolating vs Extrapolating

Interpolation is more reliable

Extrapolation is less reliable



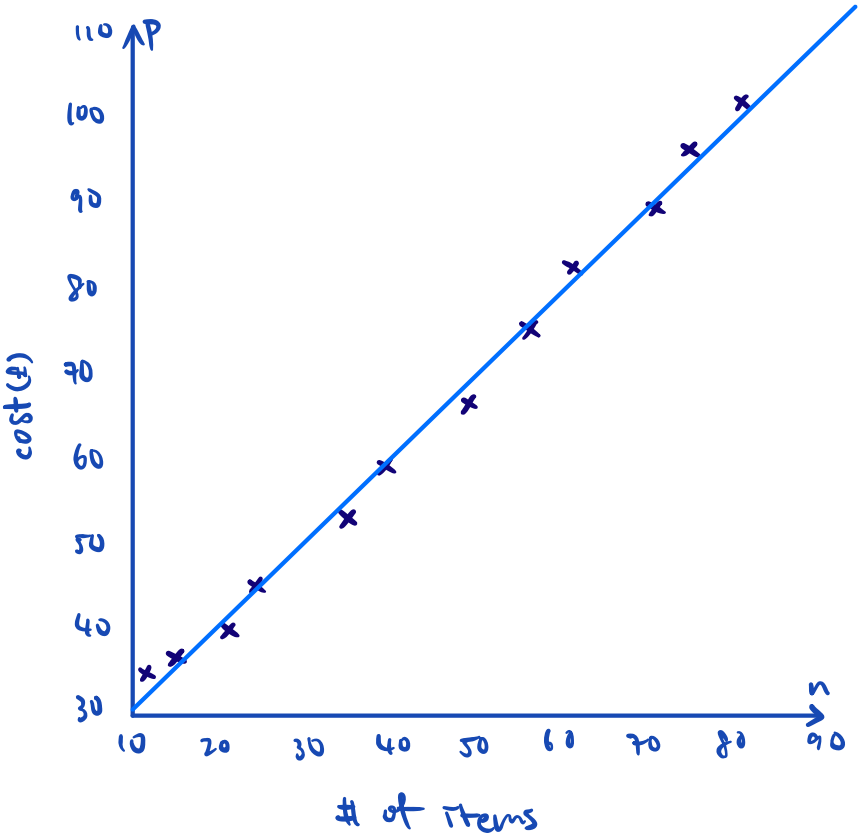
line of best fit ↓ estimate within range



line of best fit ↓ estimate out of range

EX 4B (p.65)

① a), b)



c) 21.0: it would cost £21 to produce 0 items

0.98: each item costs £0.98 to produce

d) it's suitable because the regression line is very close to the actual costs and the costs show a linear trend.