Regression & Correlation

So far we have dealt with univariate data (e.g quiz marks).

But what if you have bivariate data ?

↑

2 variables

(son's height & father's height)

"Relationship" between 2 variables is captured by regression & correlation.

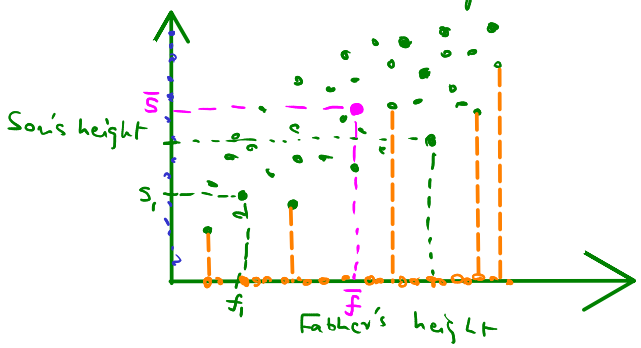
Father-Son pairs

f_1, s_1

f_2, s_2

$\vdots$

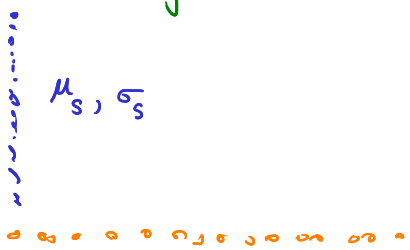
f_N, s_N

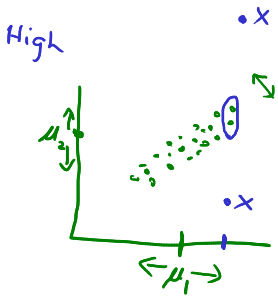


← Scatter Diagram

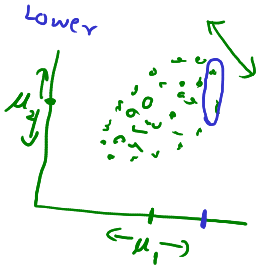
μ_f, μ_s

σ_f, σ_s

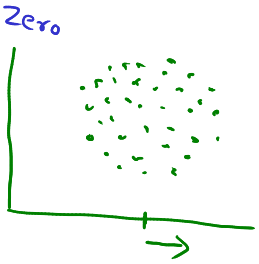




Thin



Thick



No association between x and y.

→ Correlation

y is more strongly associated with x

y is less strongly associated with x

↓

Measures the strength of associated.

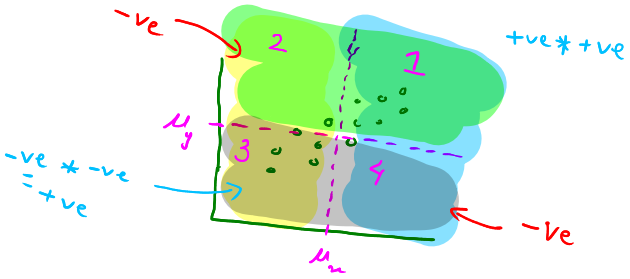
Step 1

x	y	$\hat{x} = \frac{x-\mu_x}{\sigma_x}$		$\hat{y} = \frac{y-\mu_y}{\sigma_y}$	$\hat{x} * \hat{y}$
–	–	–		–	–
–	–	–		–	–
⋮	⋮	⋮	*	⋮	⋮
–	–	–		–	–
			Step 2		avg.

Step 3

Correlation = avg (x in standard units * y in standard units)

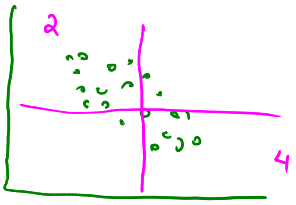
Correlation



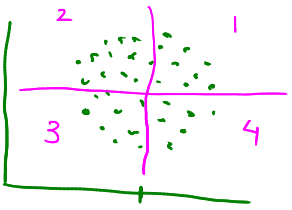
$\hat{x} * \hat{y}$

$\frac{x-\mu_x}{\sigma_x} \geq 0$ $\quad$ $\frac{y-\mu_y}{\sigma_y} \geq 0$

Correlation will be +ve bez. it is avg (mostly positive products).



Correlation will be -ve.



Correlation will be 0.

Correlation is usually denoted by the letter 'r'.

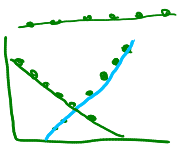
① r is always between -1 and 1

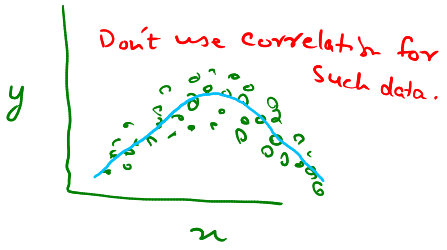
② $r = 0 \Rightarrow$ no association between x and y.

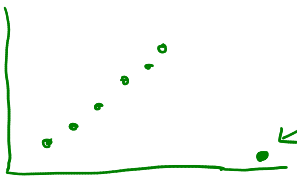
③ $|r| = 1 \Rightarrow$ perfect linear association btw. x & y

$r = 1 \Rightarrow$ _____ , +ve _____

$r = -1 \Rightarrow$ _____ , -ve _____







This outlier will bring value of r down to almost 0.

Try to remove outliers before using correlation.

Regression —— predict y from x.

$P(s \geq 72 \text{ inches} \mid f = 70 \text{ inches})$

Tests of Significance
Hypothesis testing