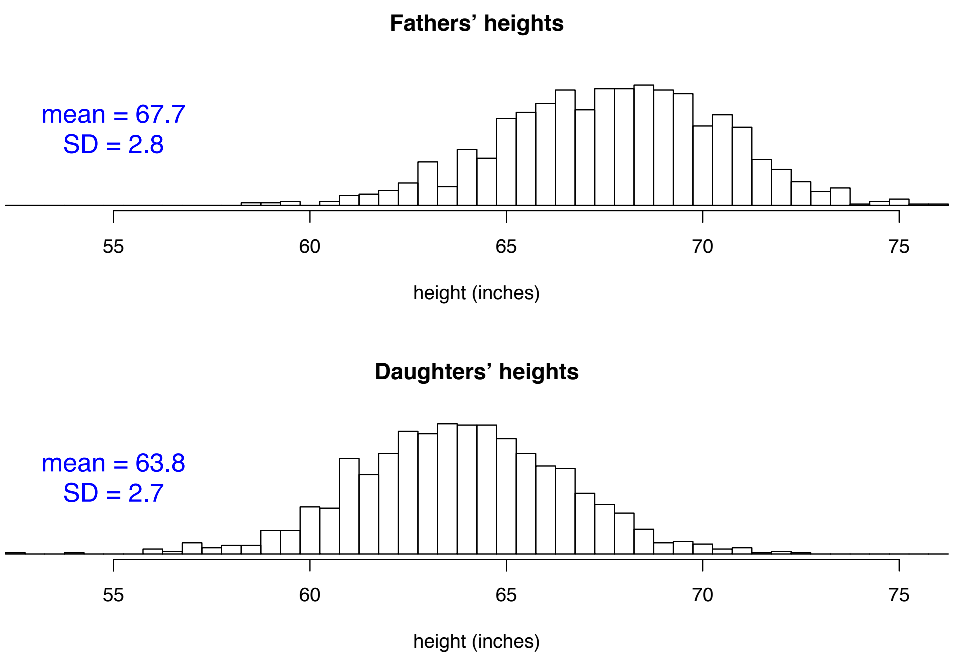
Correlation

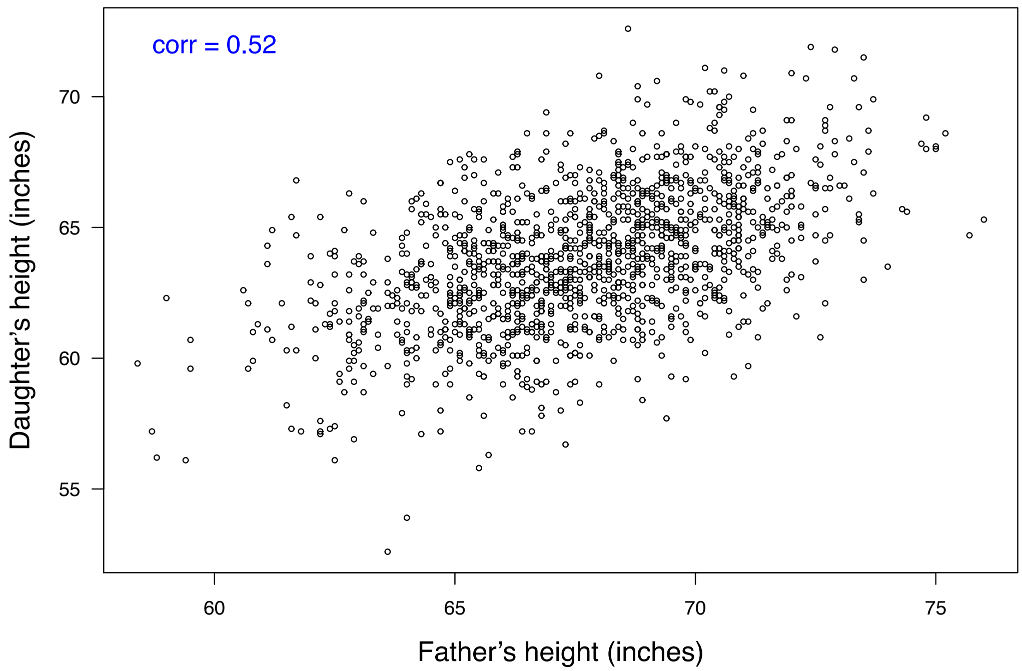
Fathers' and daughters' heights



Reference: Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

Fathers' and daughters' heights



Reference: Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

Covariance and correlation

Let X and Y be random variables with

μ_X = E(X), μ_Y = E(Y), σ_X = SD(X), σ_Y = SD(Y)

For example, sample a father/daughter pair and let

X = the father's height and Y = the daughter's height.

Covariance

$$\text{cov}(X,Y) = E\{(X - \mu_X)\,(Y - \mu_Y)\}$$

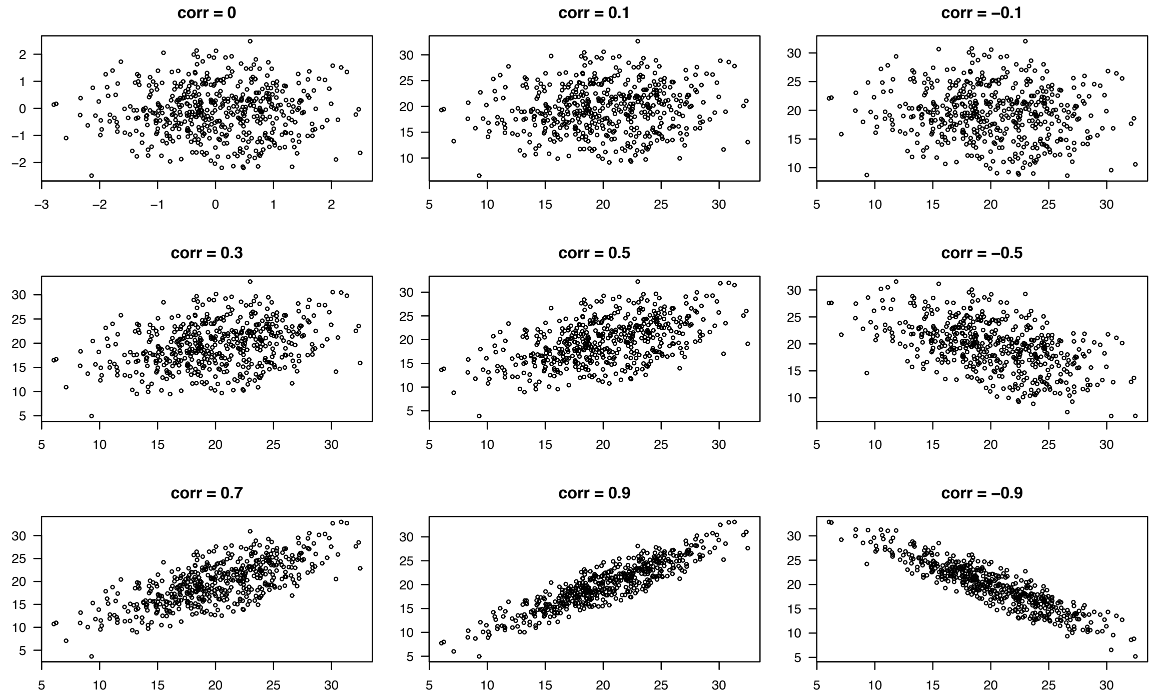
$\longrightarrow$ cov(X,Y) can be any real number

Correlation

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$\longrightarrow$ $-1 \leq \text{cor}(X,Y) \leq 1$

Examples



Estimated correlation

Consider n pairs of data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$

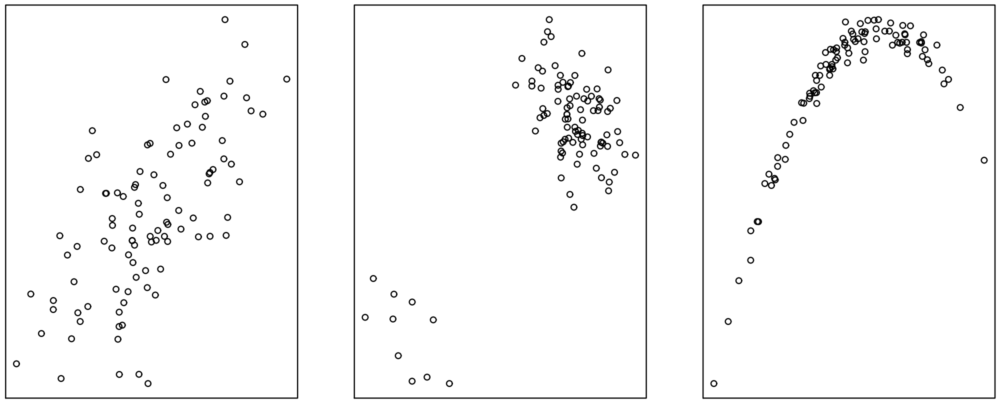
We consider these as independent draws from some bivariate distribution.

We estimate the correlation in the underlying distribution by:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

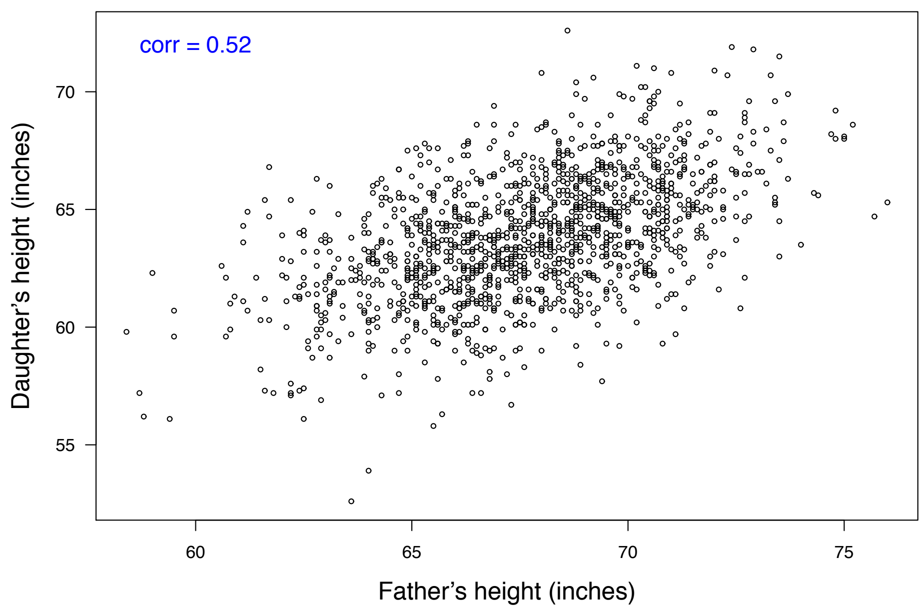
This is sometimes called the correlation coefficient.

Correlation measures linear association



$\longrightarrow$ All three plots have correlation ≈ 0.7!

Fathers' and daughters' heights



Pearson and Spearman