

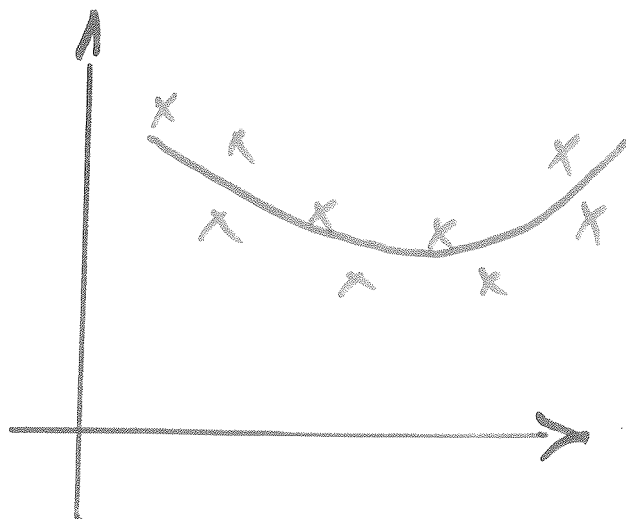
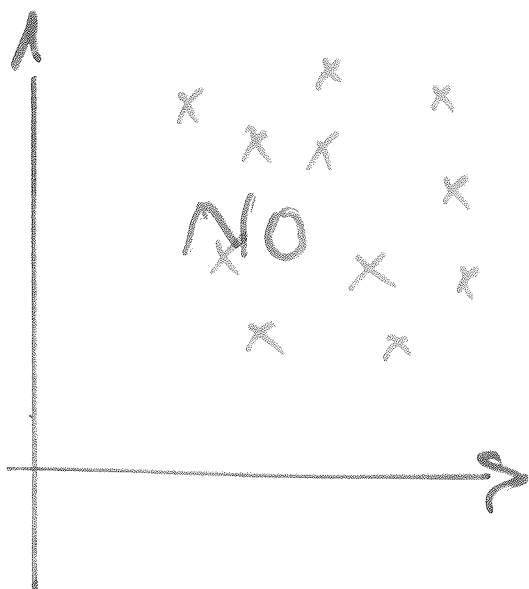
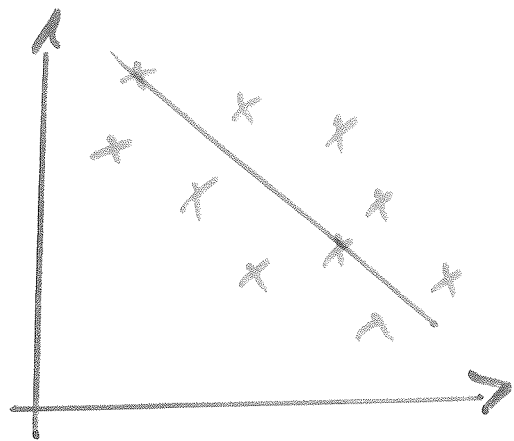
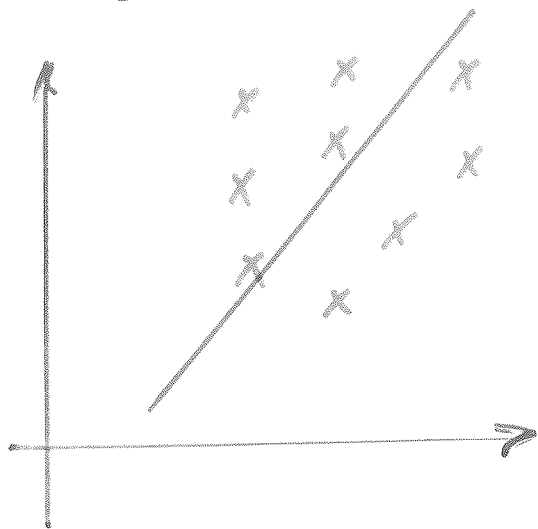
CORRELATION

P42-44

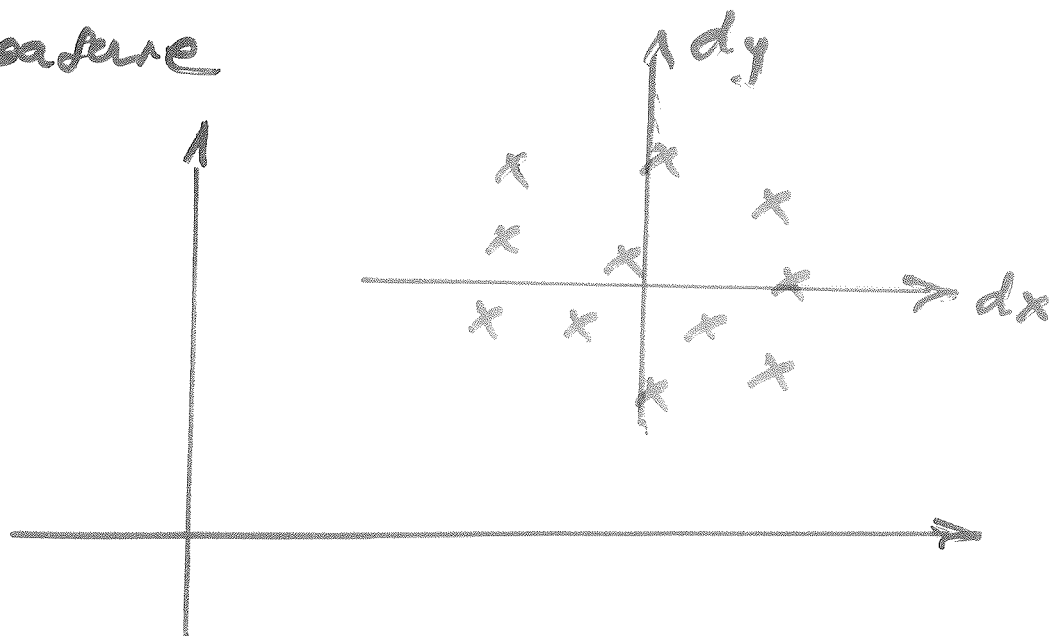
Two random variables, e.g. height H , and weight W are not functionally related, but they can be correlated.

This means that a tendency in one is marked by a tendency in the other.

We plot sets of random observations of two random variables X, Y on a scatter diagram. Regression curves are shown.



To establish the amount of correlation we measure



where $dx = x_i - \bar{x}$

$dy = y_i - \bar{y}$

$\sum dx dy$ or sum of products of deviations from means gives a measure. We normalise by dividing by $\sqrt{\sum dx^2 \sum dy^2}$ and define

$$r = \frac{\sum dx dy}{(\sum dx^2 \sum dy^2)^{1/2}}$$

It can be proved that $-1 \leq r \leq 1$ with r taking the values ± 1 when there is an exact straight line relationship. When $r = 0$ X and Y are uncorrelated.

r is obtained from the data and is called the sample correlation coefficient

Note that r , like \bar{x} and s^2 is a sample based statistical parameter.

If X and Y are the random variables of the parent population, e.g. height and weight then we can define ρ to be the theoretical correlation coefficient,

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\left\{ E[X - \mu_x]^2 E[Y - \mu_y]^2 \right\}^{1/2}}$$
$$= \frac{\text{Covariance}[X, Y]}{\left\{ \text{Var}[X] \text{Var}[Y] \right\}^{1/2}} = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

$\rho = 0$ when X, Y are independent

$$Z = \frac{\sqrt{n-3}}{2} \ln \left(\frac{1+r}{1-r} \right)$$

$$\text{So } \ln \left(\frac{1+r}{1-r} \right) = \frac{2Z}{\sqrt{n-3}}$$

$$\begin{aligned} \therefore \frac{1+r}{1-r} &= \exp \left(\frac{2Z}{\sqrt{n-3}} \right) \\ &= e^{2\alpha Z}, \text{ say.} \end{aligned}$$

$$\therefore 1+r = (1-r) e^{2\alpha Z}$$

$$\text{or } r(1 + e^{2\alpha Z}) = e^{2\alpha Z} - 1$$

$$\text{i.e. } r = \frac{e^{2\alpha Z} - 1}{e^{2\alpha Z} + 1}$$

$$= \frac{e^{\alpha Z} - e^{-\alpha Z}}{e^{\alpha Z} + e^{-\alpha Z}}$$

$$= \frac{\sinh \alpha Z}{\cosh \alpha Z} = \tanh \alpha Z.$$

As with \bar{x} and s^2 vis-à-vis μ and σ^2 , we can validate as to whether the sample estimate r is a good estimate of ρ_0 .

Set up the null hypothesis

$H_0: \rho = \rho_0$ versus $H_1: \rho \neq \rho_0$ by computing $\frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$ because

$$Z = \frac{\sqrt{n-3}}{2} \left[\ln \left[\frac{1+r}{1-r} \right] - \ln \left[\frac{1+\rho_0}{1-\rho_0} \right] \right]$$

Can be shown to be distributed $N(0,1)$.

If we wish we can choose either a one-tail or two-tail test.

Significance of the Correlation Coefficient

$H_0: \rho_0 = 0$, say

$H_1: \rho_0 \neq 0$

So $Z = \frac{\sqrt{n-3}}{2} \ln \left(\frac{1+r}{1-r} \right)$

$$\ln \frac{1+\rho_0}{1-\rho_0} = 0 \text{ if } \rho_0 = 0$$

So $r = \tanh \left(\frac{Z}{\sqrt{n-3}} \right)$. So if $|Z_{crit}| = 1.96$
for $n=10$, $r_{crit} = 0.63$
 $n=20$, $r_{crit} = 0.44$

Example (theoretical correlation)

If $Z = X + Y$ Q. book p 21

$$\text{then } E[Z] = E[X] + E[Y]$$

$$\text{Var}[Z] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$

Let us call $\text{Cov}[X, Y]$, σ_{XY}

$$\begin{aligned} \text{so } \sigma_Z^2 &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \\ &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \end{aligned}$$

In the example, if $\sigma_X = 1 \text{ min}$, $\sigma_Y = 2 \text{ min}$
then $\sigma_Z^2 = 1^2 + 2^2 + 2 \times 0.5 \times 1 \times 2 = 7$ $\rho = 0.5$

$$\text{i.e. } \sigma_Z = \sqrt{7}$$

Given $E[X] = 12$ and $E[Y] = 20$,

$$\begin{aligned} P(\text{Running time} \geq 33 \text{ min}) &= P(Z \geq 1/\sqrt{7}) \\ &= 0.35 \end{aligned}$$

where Z is $N(0, 1)$.

Daily rainfall x (cm)	Particles removed y (microgm. m ⁻³)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

$$\sum (x_i - \bar{x})^2 = 19.26, \quad \sum (y_i - \bar{y})^2 = 804.2, \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = -121.8$$

$$\text{So } r = \frac{-121.8}{\sqrt{19.26 \times 804.2}} = -0.9786$$

We assume no correlation between rainfall volume and particulate deposit, i.e.

$$H_0: \rho = 0, \quad H_1: \rho \neq 0, \quad \text{with } |Z_{\text{crit}}| = 1.96$$

$$\text{Computing: } Z = \frac{\sqrt{n-3} \ln\left(\frac{1+r}{1-r}\right)}{2} = \frac{\sqrt{6} \ln\left(\frac{0.0214}{1.9786}\right)}{2} = -5.54.$$

So accept H_1 (by a marked margin)

For revision, repeat with $\rho_0 = -0.95, -0.995$.