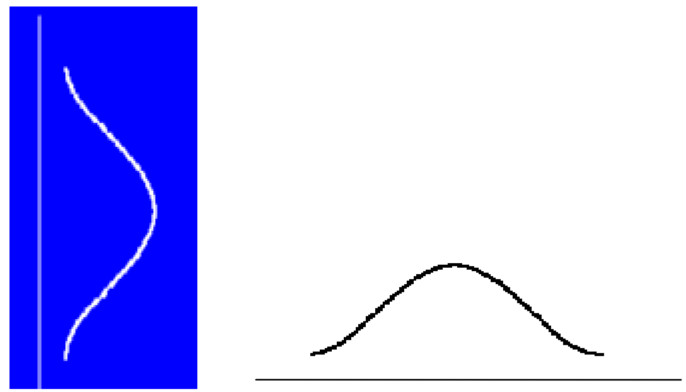
KORRELÁCIÓ-ANALÍZIS

több (két) valószínűségi változó közti kapcsolat vizsgálata

Két valószínűségi változó közti kapcsolat

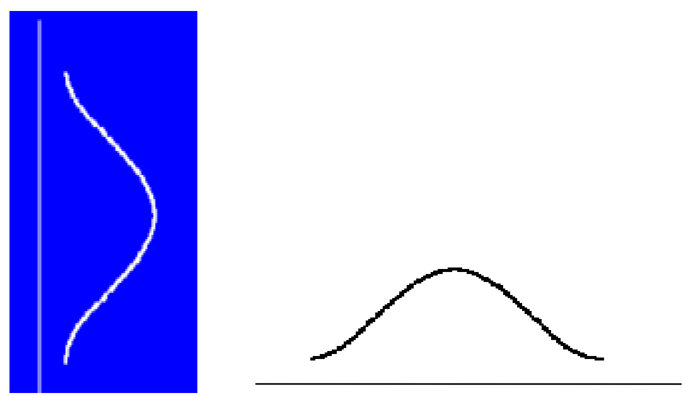
fiúk
magassága



apák magassága

összefüggő

mennyi idő
alatt tanul meg
az unokaöcs
spanyolul



apák magassága

független

Két valószínűségi változó együttes eloszlása

A kétváltozós eloszlás sűrűségfüggvénye:

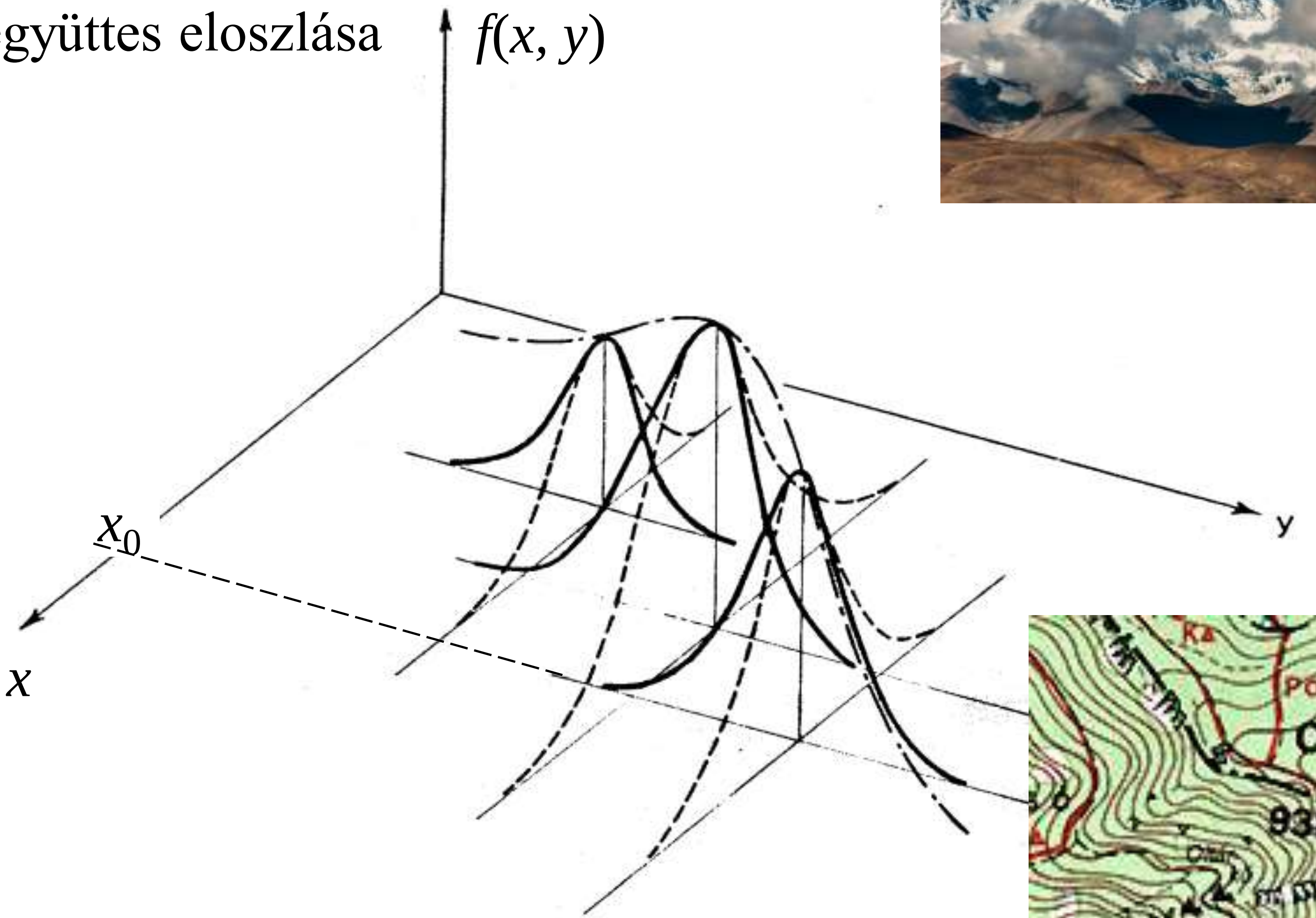
$$P\left(x_0 < x \le x_0 + \Delta x\,;\, y_0 < y \le y_0 + \Delta y\right) =$$

$$= \int_{x_0}^{x_0+\Delta x} \int_{y_0}^{y_0+\Delta y} f\left(x, y\right) dx\,dy \approx f\left(x_0, y_0\right) \Delta x \Delta y$$

És eloszlásfüggvénye:

$$P\left(x \le x_0\,;\, y \le y_0\right) = F\left(x_0, y_0\right) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f\left(x, y\right) dx\,dy$$

Két normális eloszlású valószínűségi változó együttes eloszlása

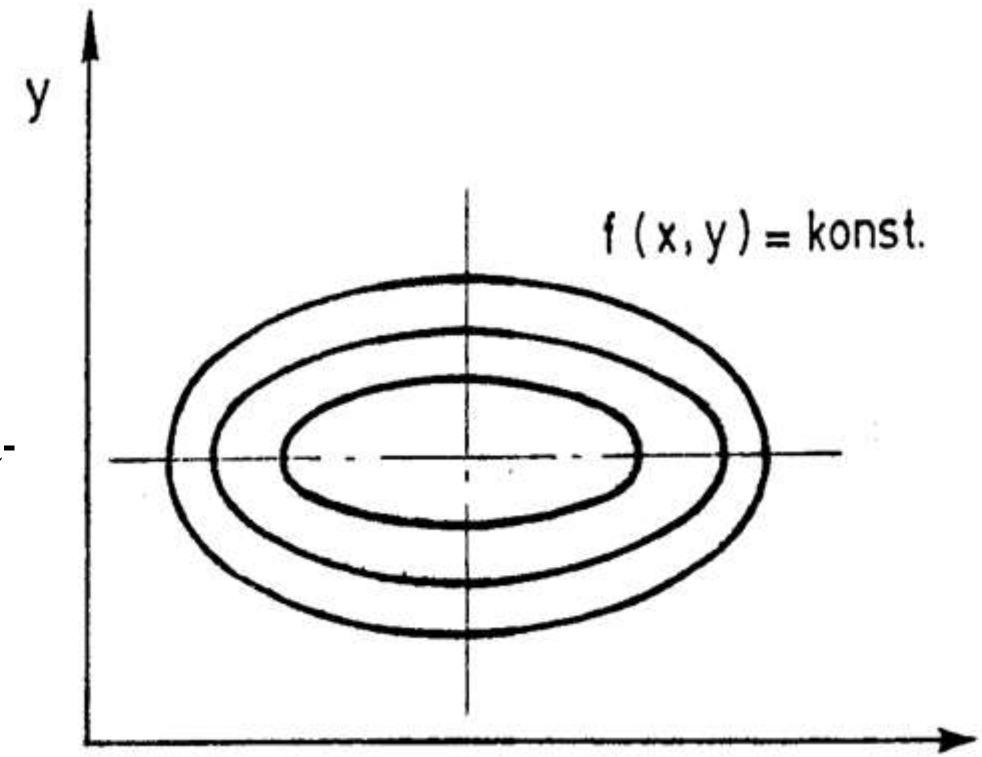


Valószínűségi változók függetlensége

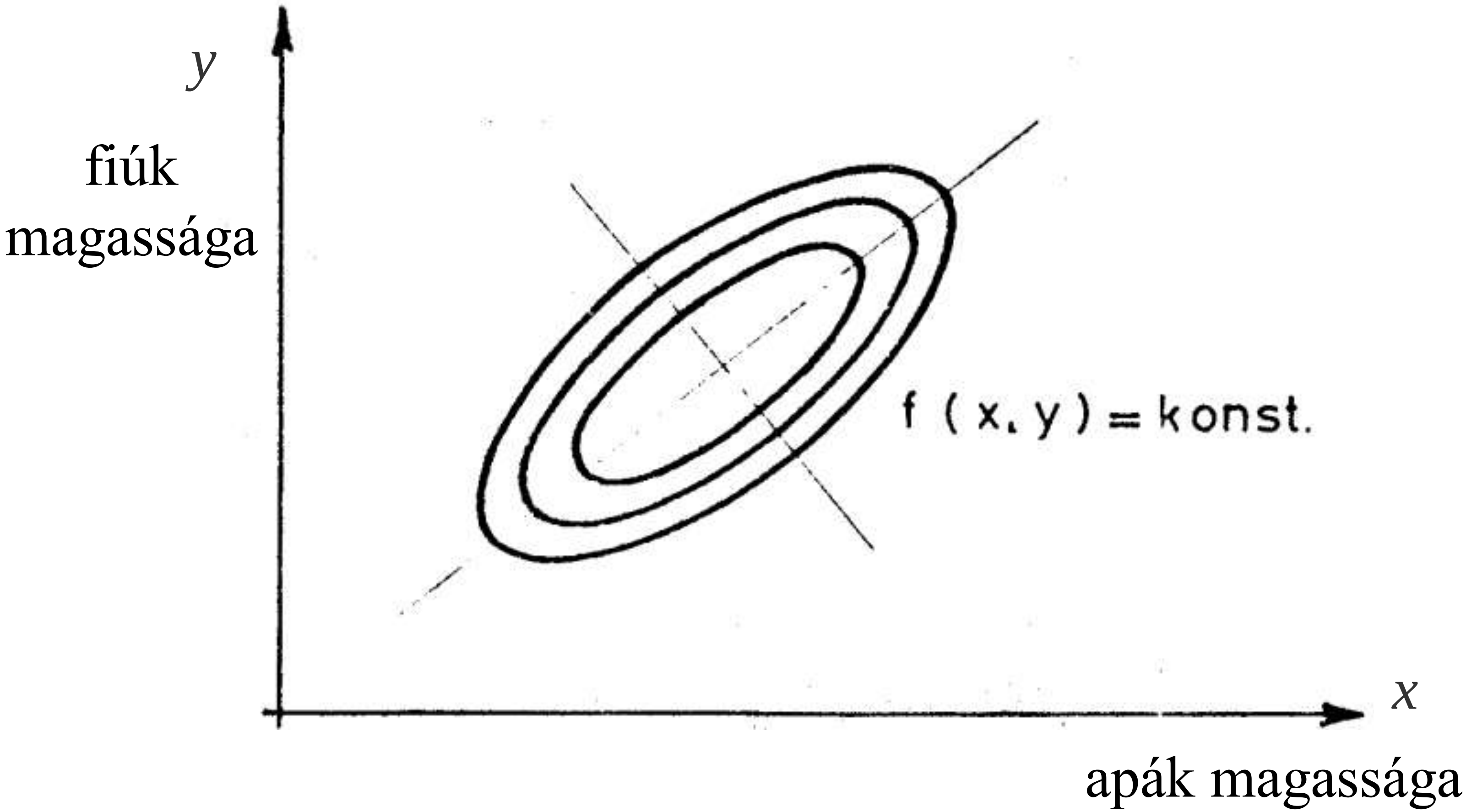
Az x és y valószínűségi változók függetlenek, ha

$$f(x, y) = f(x) f(y)$$

A szintvonalas ábrán az ellipszisek főtengelyei párhuzamosak a koordináta-tengelyekkel



Szintvonalas ábra **nem független valószínűségi változókra**



Valószínűségi változók közti kapcsolat jellemzése

Két mutatószám:

- kovariancia
- (Pearson –féle) korrelációs együttható

Kovariancia

Két valószínűségi változó összegének varianciája:

$$Var(x+y)=E\left\{\left[(x+y)-E(x+y)\right]^2\right\}=Var(x)+Var(y)+2Cov(x,y)$$

mert:

$$E\left\langle\left\{\left[x-E(x)\right]+\left[y-E(y)\right]\right\}^2\right\rangle=E\left\{\left[x-E(x)\right]^2\right\}+E\left\{\left[y-E(y)\right]^2\right\}+$$
$$+2E\left\{\left[x-E(x)\right]\left[y-E(y)\right]\right\}=Var(x)+Var(y)+2Cov(x,y)$$

$$Var(x)=\sigma_x^2=E[(x-\mu)^2]$$

$$Cov(x,y)=\sigma_{xy}=E\{[x-E(x)][y-E(y)]\}$$

Ha x és y függetlenek: $\sigma_{xy}=0$ $\quad Var(x+y)=Var(x)+Var(y)$

Kovariancia mintabeli becslése

$$Cov(x, y) = \sigma_{xy} = E\{[x - E(x)][y - E(y)]\}$$

σ_{xy} becslése: $$\hat{\sigma}_{xy} = s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

(emlékeztetőül) $$\hat{\sigma}^2 = s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

Korrelációs együttható

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E\left[\left(x - E(x)\right)\left(y - E(y)\right)\right]}{\sigma_x \sigma_y}$$

Tulajdonságai:

- léptékfüggetlen
- előjele van
- értéke: $-1 \leq \rho \leq 1$
- a **lineáris** kapcsolat szorosságát méri

Ha *x* és *y* függetlenek: $\sigma_{xy} = 0$ $\rho_{xy} = 0$

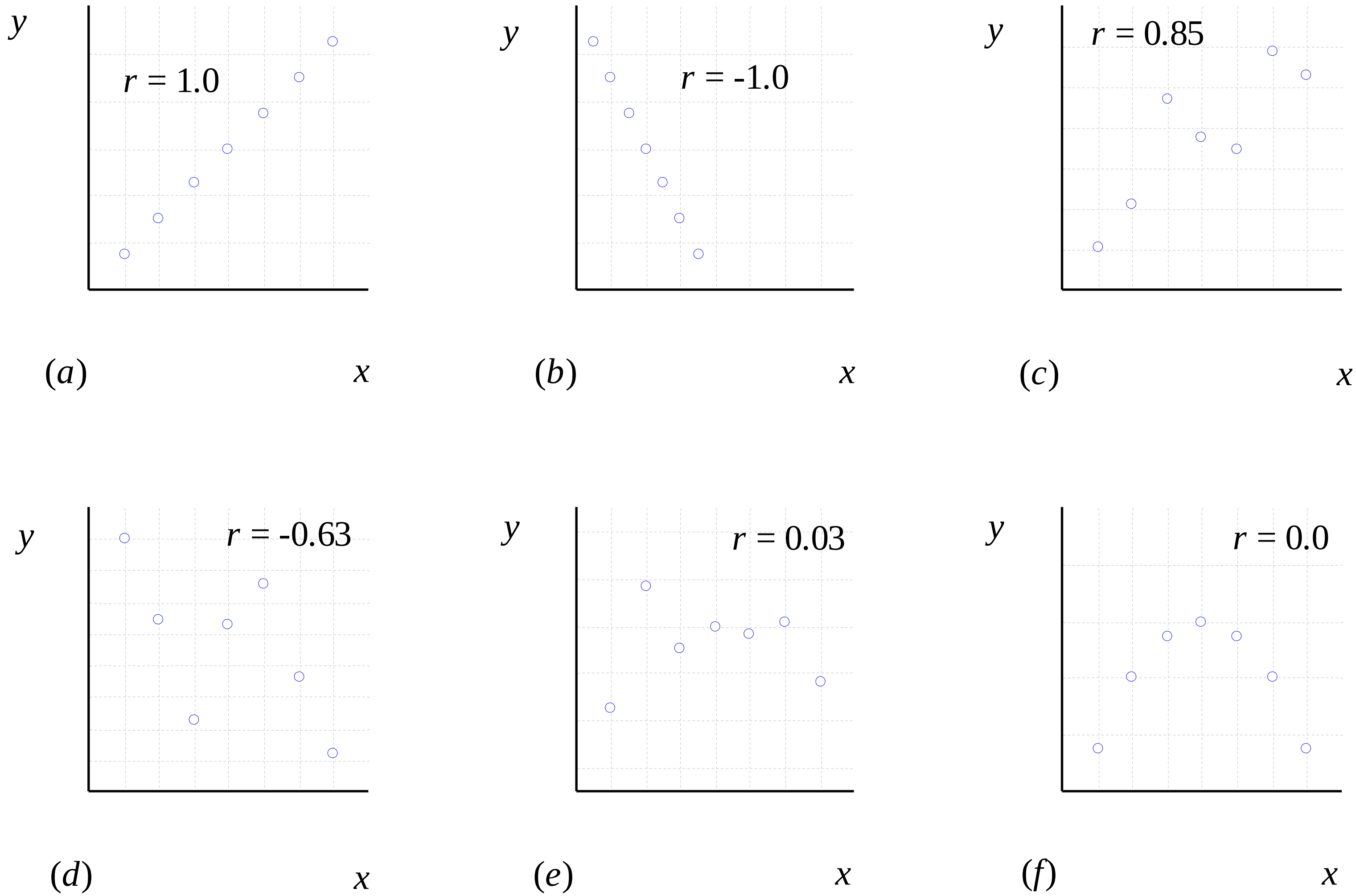
Mintabeli (becsült) korrelációs együttható

$$\hat{\rho} \equiv r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_i (x_i - \bar{x})^2\right]\left[\sum_i (y_i - \bar{y})^2\right]}}$$

(emlékeztetőül) $\quad \hat{\sigma}^2 = s^2 = \dfrac{\sum_i (x_i - \bar{x})^2}{n-1}$

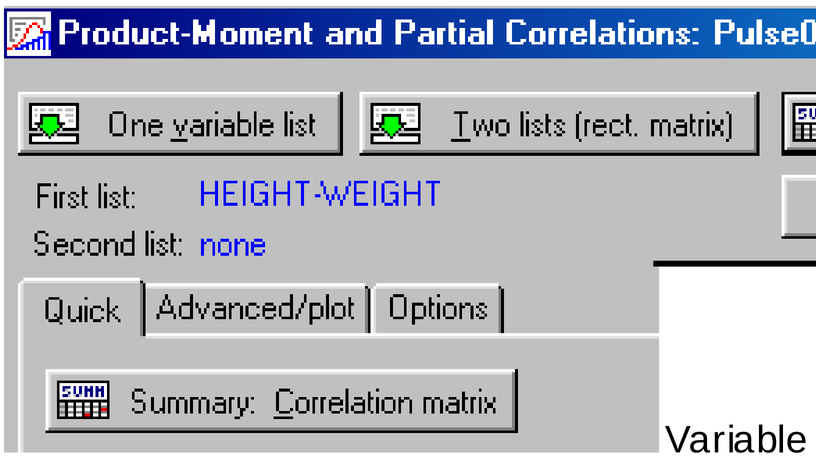
σ_{xy} becslése $\quad \hat{\sigma}_{xy} = s_{xy} = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Korreláció két valószínűségi változó között



Példa:
Számítsuk ki a pulzus-példában a testmagasság és a testsúly közötti korrelációs együtthatót!

(Statistics > Basic Statistics and Tables > Correlation matrices)



Correlations (Pulse00)
Marked correlations are significant at $p < .0500$
N=92 (Casewise deletion of missing data)

Variable	HEIGHT	WEIGHT
HEIGHT	1.00	0.78
WEIGHT	0.78	1.00

A korrelációs együttható szignifikanciájának vizsgálata

Ha $\rho=0$ $$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$ $\nu=n-2$

DE: A szignifikáns korreláció nem jelent kauzális összefüggést!

Pl. a tűz okozta kár nagysága és az oltásban részt vevő tűzoltók száma

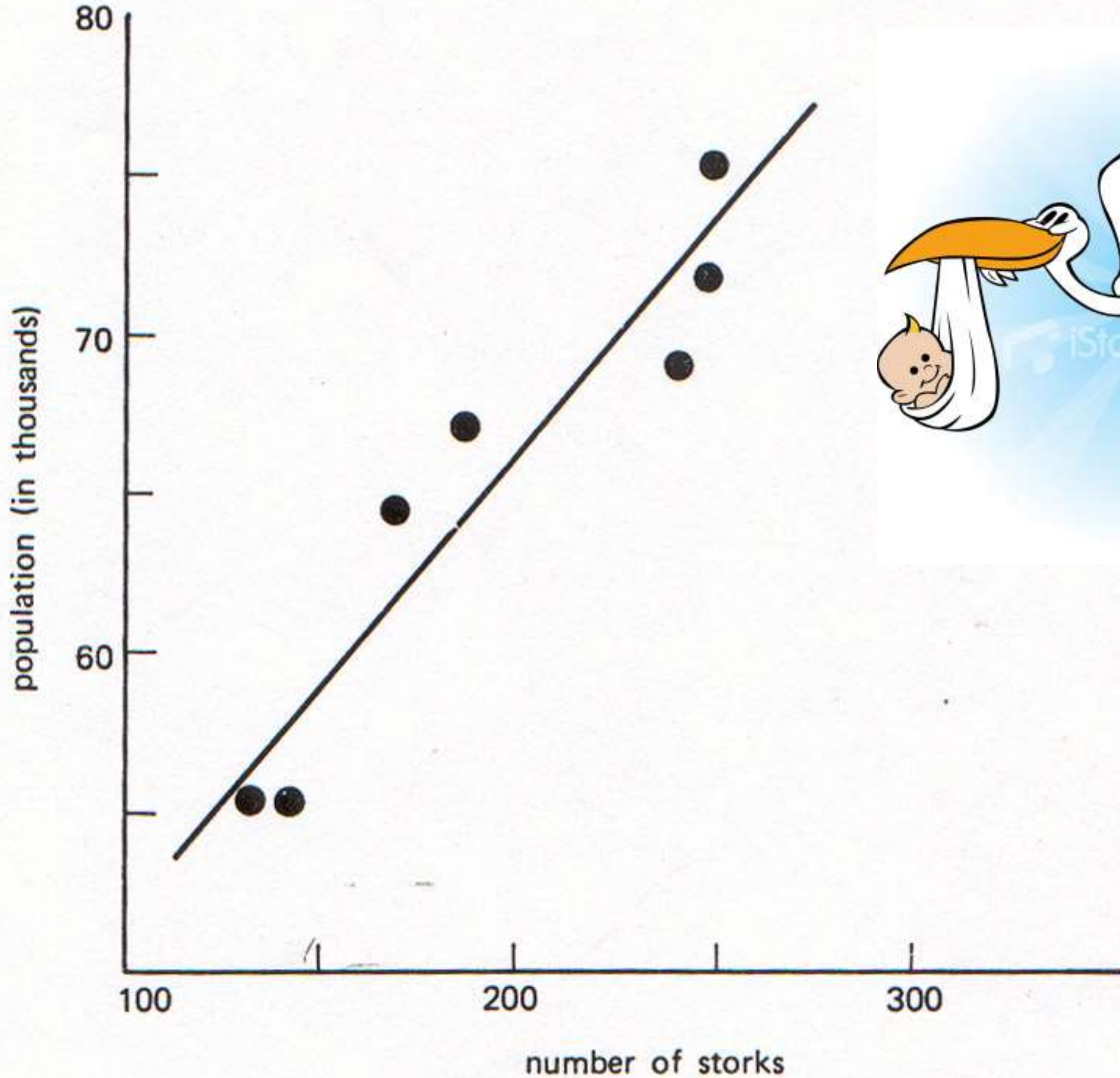


FIGURE 1.4. A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year, 1930–1936.