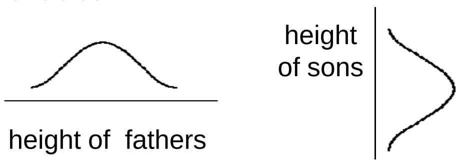
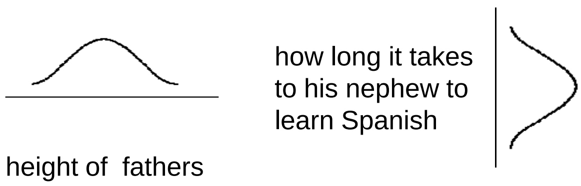
CORRELATION

JOINT DISTRIBUTION OF TWO RANDOM VARIABLES

Not independent variables:

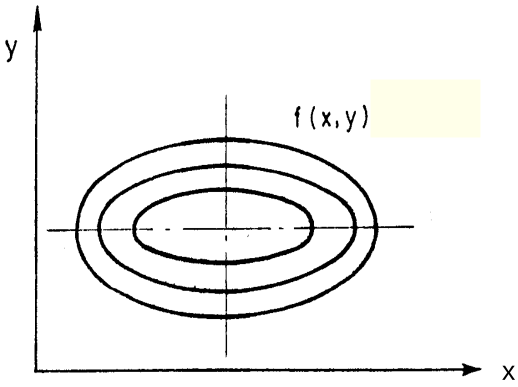


Independent variables:



CONTOUR PLOT (INDEPENDENT RANDOM VARIABLES)

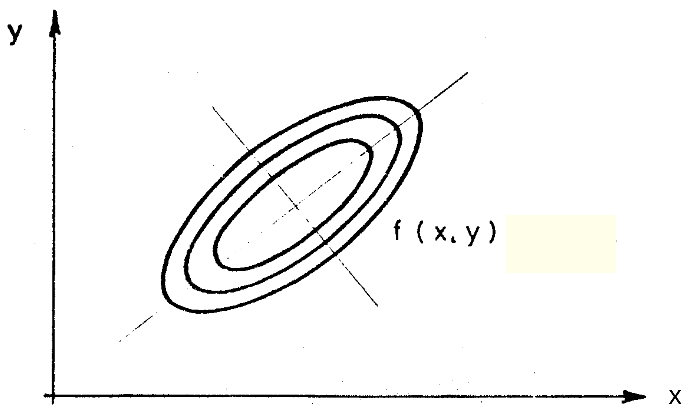
how long it takes
to his nephew to
learn Spanish



height of father

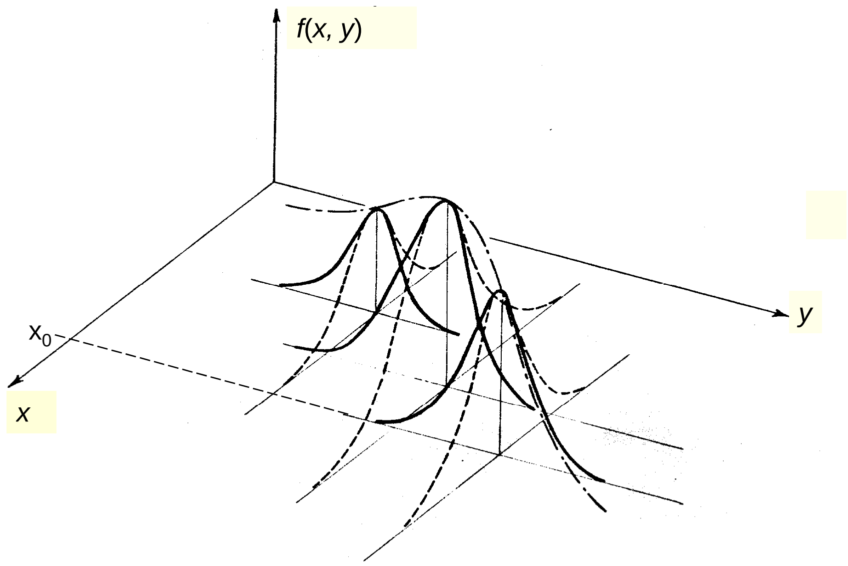
CONTOUR PLOT (INDEPENDENT RANDOM VARIABLES)

height
of son



height of father

JOINT DENSITY FUNCTION



5

CONFIDENCE INTERVAL AND HYPOTHESIS TEST

The probability density function:

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

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

The distribution function:

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

6

INDEPENDENCE OF RANDOM VARIABLES

The x and y random variables are independent if

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

The main axes of the ellipses are parallel to the co-ordinate axes

CHARACTERISATION OF THE INTERDEPENDENCE

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

bivariate Normal distribution,
ρ correlation coefficient

$$-1 \le \rho \le 1$$

When ρ=0 $\quad f(x,y) = f(x)f(y)$

COVARIANCE

Variance of the sum of two random variables:

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

$$Var(x) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x) dx = E\left[(x-\mu)^2\right]$$

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

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

COVARIANCE

Variance of the sum of two random variables:

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

$$\sigma_x^2 = Var(x) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x) dx = E\left[(x-\mu)^2\right]$$

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

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y) f(x,y) dx\, dy$$

If *x* and *y* are independent: $\sigma_{xy} = 0$ $\quad Var(x+y) = Var(x) + Var(y)$

COVARIANCE AND CORRELATION COEFFICIENT

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If *x* and *y* are independent: $\sigma_{xy} = 0$ $\rho_{xy} = 0$

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

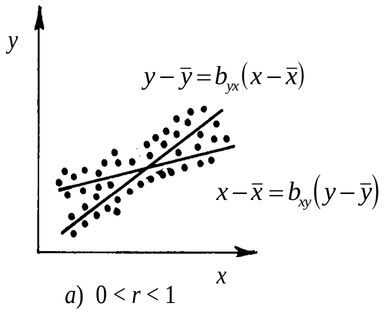
(to remind) $$\hat{\sigma}^2 = s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

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

CORRELATION COEFFICIENT

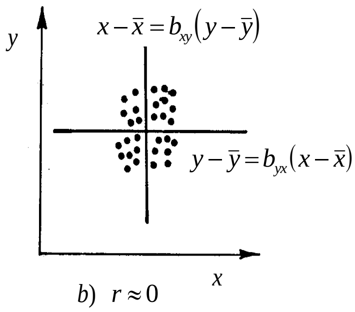
The correlation coefficient measures the strength of **linear** interdependence

$$r = \sqrt{b_{yx} b_{xy}}$$



$a)$ $0 < r < 1$

correlated variables



$b)$ $r \approx 0$

uncorrelated variables

SIGNIFICANCE TEST FOR THE CORRELATION COEFFICIENT

If ρ=0 $$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$ with $\nu = n-2$

Even if it is significant, causality is not proved!

E.g. damage in fire vs. number of firemen participating

8 SCIENCE AND STATISTICS

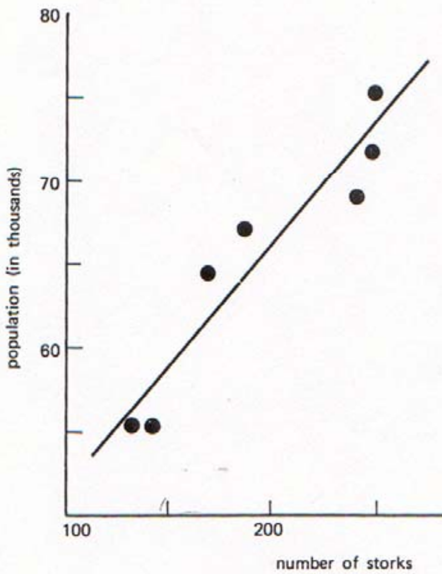


FIGURE 1.4. A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year, 1930–1936.