Covariance and Correlation

Definition 6.2. The **correlation** of the two random variables X and Y, with variances σ_X^2 and σ_Y^2, respectively, is the number defined as

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{6.6}$$

Corollary 6.2.3. For any random variables X and Y,

$$-1 \leq \rho_{XY} \leq 1.$$

♦

Lemma 6.2.4. Suppose Z is a non-negative random variable. Then $\mathrm{E}[Z] = 0$ implies that $\mathrm{P}(Z = 0) = 1$. ♦

$Z \geq 0$ and $E[Z] = 0 \Rightarrow P(Z=0) = 1$

$\Longleftarrow$

$Z \geq 0$ and $P(Z=0) = 1 \Rightarrow E[Z] = 0$

Corollary 6.2.6. For any two random variables X and Y, $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $\mathrm{P}(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$. ♦

$Y = aX + b$

$a > 0 \Rightarrow \rho_{XY} = 1$

$a < 0 \Rightarrow \rho_{XY} = -1$

Definition 6.3. Suppose the random variables $X_1, X_2, \ldots, X_n$ are observed as $x_1, x_2, \ldots, x_n$, respectively, and the random variables $Y_1, Y_2, \ldots, Y_n$ are observed as $y_1, y_2, \ldots, y_n$, respectively. Then the observed sample correlation is defined as

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Proposition 6.2.8

Suppose we have pairs of measurements $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Define pairs $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)$

where $u_i = ax_i + b$ $\quad i = 1, 2, \ldots, n$

$v_i = cy_i + d$

for some $a, b, c, d \in \mathbb{R}$ and $a > 0, c > 0$.

Then $r_{XY} = r_{UV}$

Proof: Substitute in transformations; and $r_{XY} = r_{UV}$.

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b)$$

$$= a\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + b\left(\frac{1}{n}\sum_{i=1}^{n} 1\right)$$

$$= a\bar{x} + b$$

$$\Rightarrow \quad u_i = a x_i + b$$

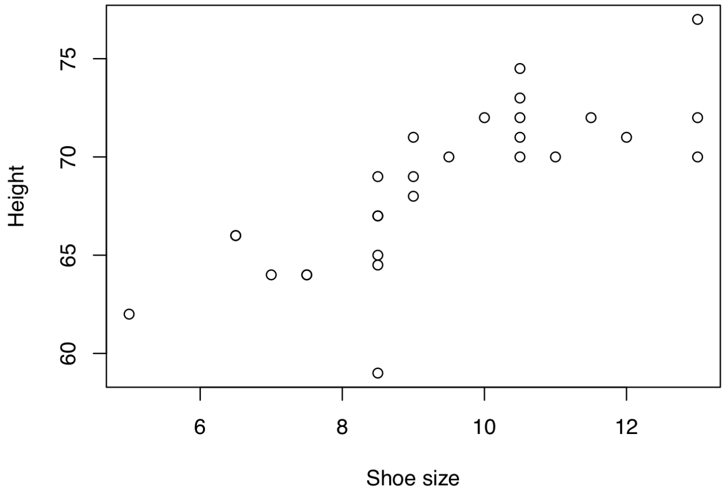
$$\bar{u} - u_i = a\bar{x} + b - (a x_i + b)$$

$$= a(\bar{x} - x_i) \quad \text{, etc}$$

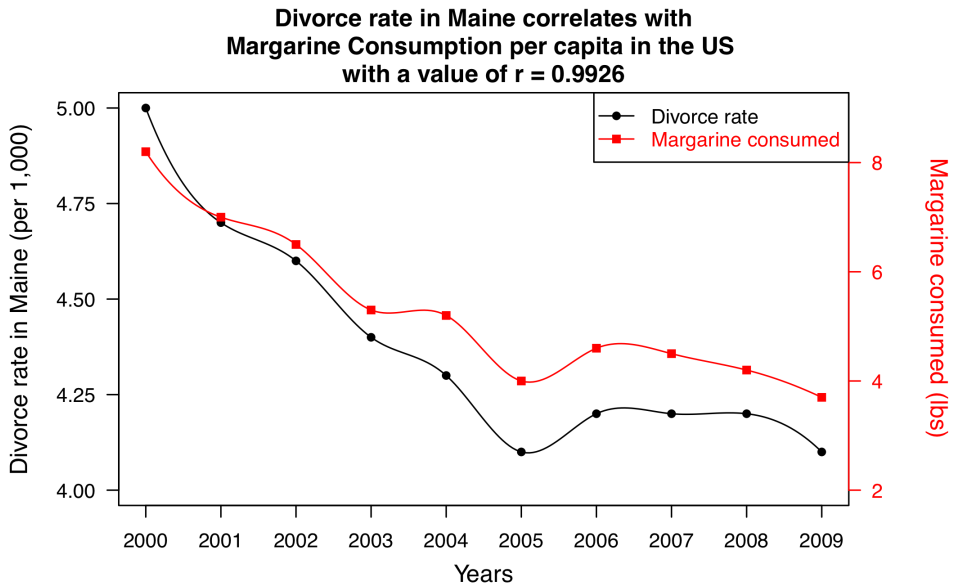
6.2.2 A real data example: height and shoe size

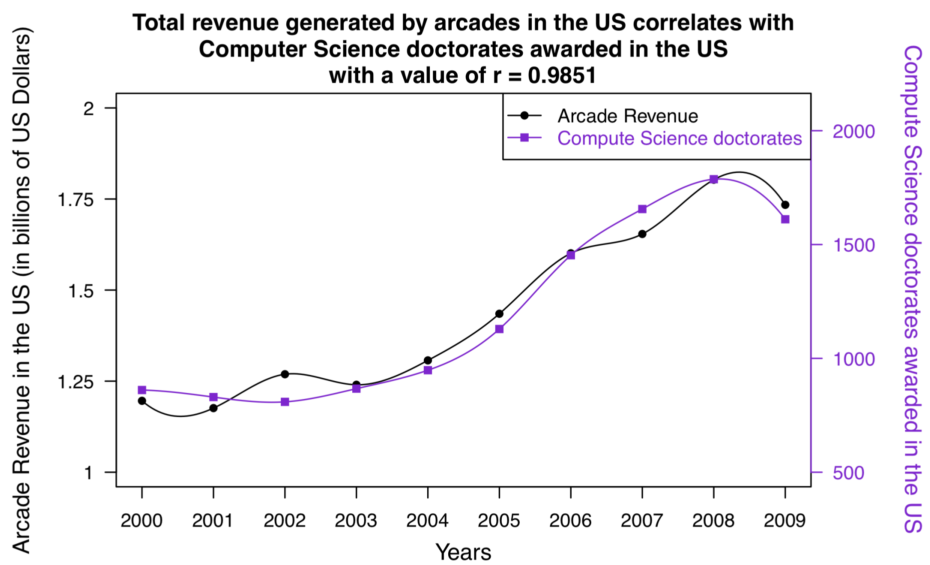
Shoe size	Height	Gender	Shoe Size	Height	Gender
6.5	66.0	F	13.0	77.0	M
9.0	68.0	F	11.5	72.0	M
8.5	64.5	F	8.5	59.0	F
8.5	65.0	F	5.0	62.0	F
10.5	70.0	M	10.0	72.0	M
7.0	64.0	F	6.5	66.0	F
9.5	70.0	F	7.5	64.0	F
9.0	70.0	F	8.5	67.0	M
13.0	72.0	M	10.5	73.0	M
7.5	64.0	F	8.5	69.0	F
10.5	74.0	M	10.5	72.0	M
8.5	67.0	F	11.0	70.0	M
12.0	71.0	M	9.0	69.0	M
10.5	71.0	M	13.0	70.0	M

Table 6.1: Shoes sizes (US scale) and heights (inches) of 28 students from Arizona State University.



$r_{xy} = 0.78$





CORRELATION DOES NOT IMPLY
CAUSATION

7.2 Inference using a probability model

We have random variable X, and know (assume we know) the distribution. And then we observe x.

(a) Compute an estimate of a plausible value for x, e.g. using the expected value of x following our probability model.

(b) Construct a subset that has a high probability of containing the true value of x.

(c) Assess whether or not an observed value of x is an implausible value, given the known probability model.

Example 7.2.1. Suppose it is known that the lifespan X in years for a particular smartphone follows the distribution $X \sim \text{Exp}(\lambda)$ with $\lambda = 1$; see Figure 7.1 for a plot of this distribution.

a) $E[X] = 1$ year

b) 95% interval $(0, c)$

$$f(x) = e^{-x} \quad ; x \geq 0$$

$$0.95 = \int_0^c e^{-x}\,dx = 1 - e^{-c}$$

$$c = -\log(0.05) = 2.996 \approx 3$$

c) Suppose we hope $x = 5$

$$P(X > 5) = \int_5^\infty e^{-x}\,dx = e^{-5}$$

$$= 0.0067$$

0.7%

7.3 Statistical models

Definition 7.6

The space of all possible values of a parameter θ is called the parameter space and denoted by Θ; ie. $\theta \in \Theta$.

A statistical model for data x is a set of probability measures $\{P_\theta \mid \theta \in \Theta\}$

Example 7.3.2. Suppose five friends all purchased the same smartphone when it was released. The manufacturer of the smartphones claims that the lifespan of the phones (in years) follows an Exp(0.5) distribution, while another source claims the lifespan of the phones follows an Exp(1) distribution. Therefore, in this example the statistical model for the lifespan of the smartphones is $\{\mathrm{P}_{0.5}, \mathrm{P}_1\}$, where $\mathrm{P}_{0.5}$ is the Exp(0.5) probability measure and P_1 is the Exp(1) probability measure, i.e. our indexing parameter is $\theta \in \Theta = \{0.5, 1\}$. Suppose that the friends record the lifespans of their phones, i.e. they use the phones until they break, and obtain the sample $(0.76, 1.18, 0.15, 0.14, 0.44)$ number of years. Comparing the p.d.f.'s of the Exp(0.5) and Exp(1) distributions in Figure 7.2, which model would you be inclined to say is the correct one? What if the observed data had been $(1.91, 2.46, 1.08, 5.79, 0.29)$? △

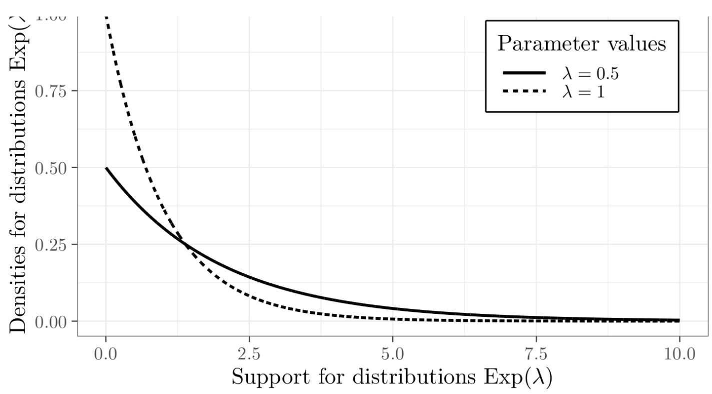


Figure 7.2: The density $f(x|\lambda) = \lambda e^{-\lambda}$ of the Exp(λ) distribution, for $\lambda \in \{0.5, 1\}$

Likelihood

Definition 8.1

Suppose we have a statistical model $\{P_\theta : \theta \in \Theta\}$ for the random variables X and where each P_θ is specified by a probability density (mass) function f_θ.
Having observed data x the likelihood function $L(\cdot|x) : \Theta \to \mathbb{R}$ is defined by $L(\theta|x) = f_\theta(x)$ for any $\theta \in \Theta$

Def. 8.2

For any $\theta \in \Theta$, $L(\theta|x)$ is called the likelihood of θ given data x.

Example 8.1.4

Suppose $x = (x_1, x_2, \dots, x_n)$ are <u>independently</u> sampled from a $N(\theta, 1)$.

Then the likelihood of θ given x, $L(\theta|x)$

$$L(\theta|x) = f_\theta(x) = f(x|\theta)$$

$$= \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\theta)^2}{2}\right)$$

$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$