Covariance and Correlation

Statistics 110

Summer 2006

Joint Distributions and Expectation

Let X and Y have joint density $f(x, y)$. Then the expectation of $g(X, Y)$, $E[g(X, Y)]$ is

$$E[g(X,Y)] = \int\int g(x,y)f(x,y)dxdy$$

For example, if X and Y are independent $U(0, 1)$,

$$\begin{aligned} E[XY] &= \int_0^1 \int_0^1 xydxdy \\ &= \int_0^1 \frac{1}{2}ydy \\ &= \frac{1}{4} \end{aligned}$$

If the function g is only a function of a single variable, such as $g(x, y) = h(x)$, then the expectation reduces to the marginal expectation as

$$
\begin{aligned}
E[g(X,Y)] &= \int \int h(x) f(x,y) dy dx \\
&= \int h(x) \left[\int f(x,y) dy \right] dx \\
&= \int h(x) f(x) dx \\
&= E[h(X)]
\end{aligned}
$$

Note that the same results hold for discrete RVs. Also the obvious extensions hold for 3 or more RVs.

Covariance

Lets consider the case where X and Y are both $Bern(p)$ marginally.

1. If X and Y are independent

$$\mathrm{Var}(X+Y) = \mathrm{Var}(Bin(2,p)) = 2p(1-p)$$

2. If $X = Y$ (positive dependence), then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(2X) = \mathrm{Var}(2Bin(1,p)) = 4p(1-p)$$

3. If $X = -Y$ (negative dependence), then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(0) = 0$$

To quantify the amount of covariation, consider for any X and Y, the difference

$$\mathrm{Var}(X+Y) - [\mathrm{Var}(X) + \mathrm{Var}(Y)]$$

Lemma. *This difference is equal to*

$$2E\left[(X - E[X])(Y - E[Y])\right]$$

Proof.

$$\begin{aligned} \mathrm{Var}(X+Y) &= E[(X - EX + Y - EY)^2] \\ &= E[(X-EX)^2 + (Y-EY)^2 + 2(X-EX)(Y-EY)] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2E[(X-E[X])(Y-E[Y])] \end{aligned}$$

□

Definition. *The* **Covariance** *of X and Y is defined as*

$$\mathrm{Cov}(X,Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

Properties of Covariance:

1. $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$

2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

3. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

4. $\mathrm{Cov}(aX + bY + c, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$ where X, Y, Z are RVs and a, b, c are constants.

5. $\mathrm{Cov}(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{Cov}(X_i, Y_j)$

   Proof. By 4), $LHS = \sum_{i=1}^{m} \mathrm{Cov}(X_i, \sum_{j=1}^{n} Y_j)$. By applying 4) to each term in the sum gives the result. $\square$

Theorem.

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

Proof. Use 3) and 5) with $Y_i = X_i, i = 1, \ldots, n$ $\square$

A useful extension to this theorem is

Theorem.

$$\mathrm{Var}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} b_i b_j \mathrm{Cov}(X_i, X_j)$$

Theorem. *If X and Y are independent, then*

$$\mathrm{Cov}(X, Y) = 0$$

Proof. If X and Y are independent (and X and Y are continuous RVs)

$$\begin{aligned}
E[XY] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} xy f_{X,Y}(x,y) dy dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} xy f_X(x) f_Y(y) dy dx \\
&= \left(\int_{\mathcal{X}} x f(x) dx \right) \left(\int_{\mathcal{Y}} y f(y) dy \right) = E[X]E[Y]
\end{aligned}$$

(Note that a similar result holds for discrete RVs). Then

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$$

□

A direct consequence of this result is that if $X_1, X_2, \ldots, X_n$ are independent RVs, then

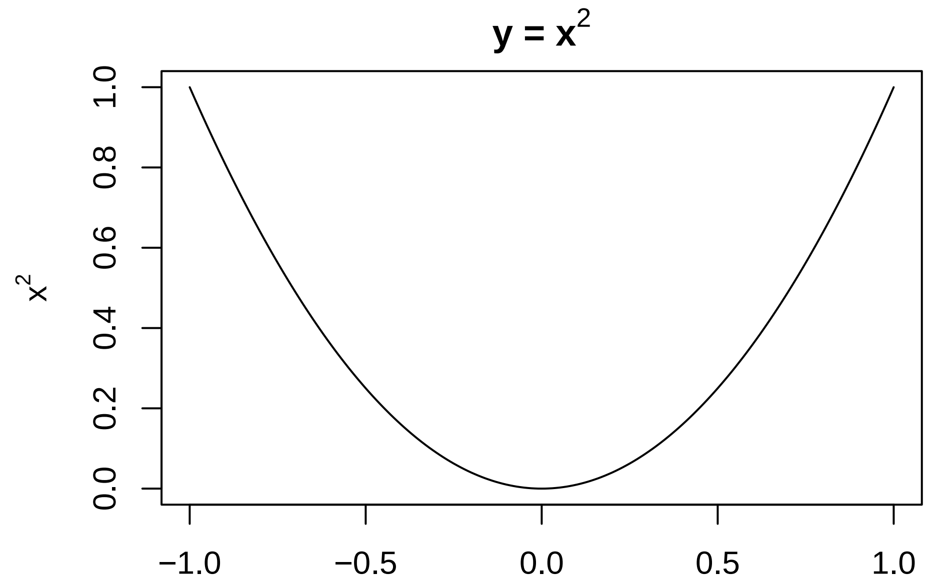
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Note that the converse of this theorem is not true. $\mathrm{Cov}(X, Y) = 0$ does not imply that X and Y are independent. For example, let $X \sim U(-1, 1)$ and $Y = X^2$. $\mathrm{Cov}(X, Y) = 0$, but the variables are highly dependent.

$$E[X] = \int_{-1}^{1} \frac{x}{2} dx = 0$$

$$E[X^2] = \int_{-1}^{1} \frac{x^2}{2} dx = \frac{1}{3} = E[Y]$$

$$E[X^3] = \int_{-1}^{1} \frac{x^3}{2} dx = 0 = E[XY]$$



$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0 - 0\frac{1}{3} = 0$$

Correlation

Definition. *If X and Y are jointly distributed random variables and the variances and covariances all exist with the variances non-zero, then the* **Correlation** *of X and Y, denoted by ρ, is*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

Properties of correlation:

1. The correlation is dimensionless

$$\rho_{aX+b,cY+d} = \rho_{X,Y}$$

   So for example, it doesn't matter whether you measure height in inches or meters or weight in pounds or kilograms.

$$
\begin{aligned}
\rho_{aX+b,cY+d} &= \frac{\mathrm{Cov}(aX+b, cY+d)}{\sqrt{\mathrm{Var}(aX+b)\mathrm{Var}(cY+d)}} \\
&= \frac{ac\mathrm{Cov}(X,Y)}{\sqrt{a^2\mathrm{Var}(X)c^2\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \rho_{X,Y}
\end{aligned}
$$

2. $|\rho| \leq 1$

Proof.

$$
\begin{aligned}
0 &\leq \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\
&= \mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= \frac{\mathrm{Var}(X)}{\sigma_X^2} + \frac{\mathrm{Var}(Y)}{\sigma_Y^2} + \frac{2\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&= 2(1+\rho)
\end{aligned}
$$

which implies $\rho \geq -1$. Similarly

$$0 \leq \operatorname{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho)$$

which implies that $\rho \leq 1$. $\square$

3. If $|\rho| = 1$, then $Y = a + bX$ with probability 1.

Proof. Assume that $\rho = 1$. Then

$$\operatorname{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0$$

This implies that

$$P\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right] = 1$$

for some constant c or that

$$P\left[Y = \frac{\sigma_Y}{\sigma_X} X + c\sigma_Y\right] = 1$$

A similar argument holds when $\rho = -1$. $\square$

The correlation ρ is the 5th parameter of the bivariate normal distribution with density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

As shown in the text

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_Y)f(x, y)dxdy = \rho\sigma_X\sigma_Y$$

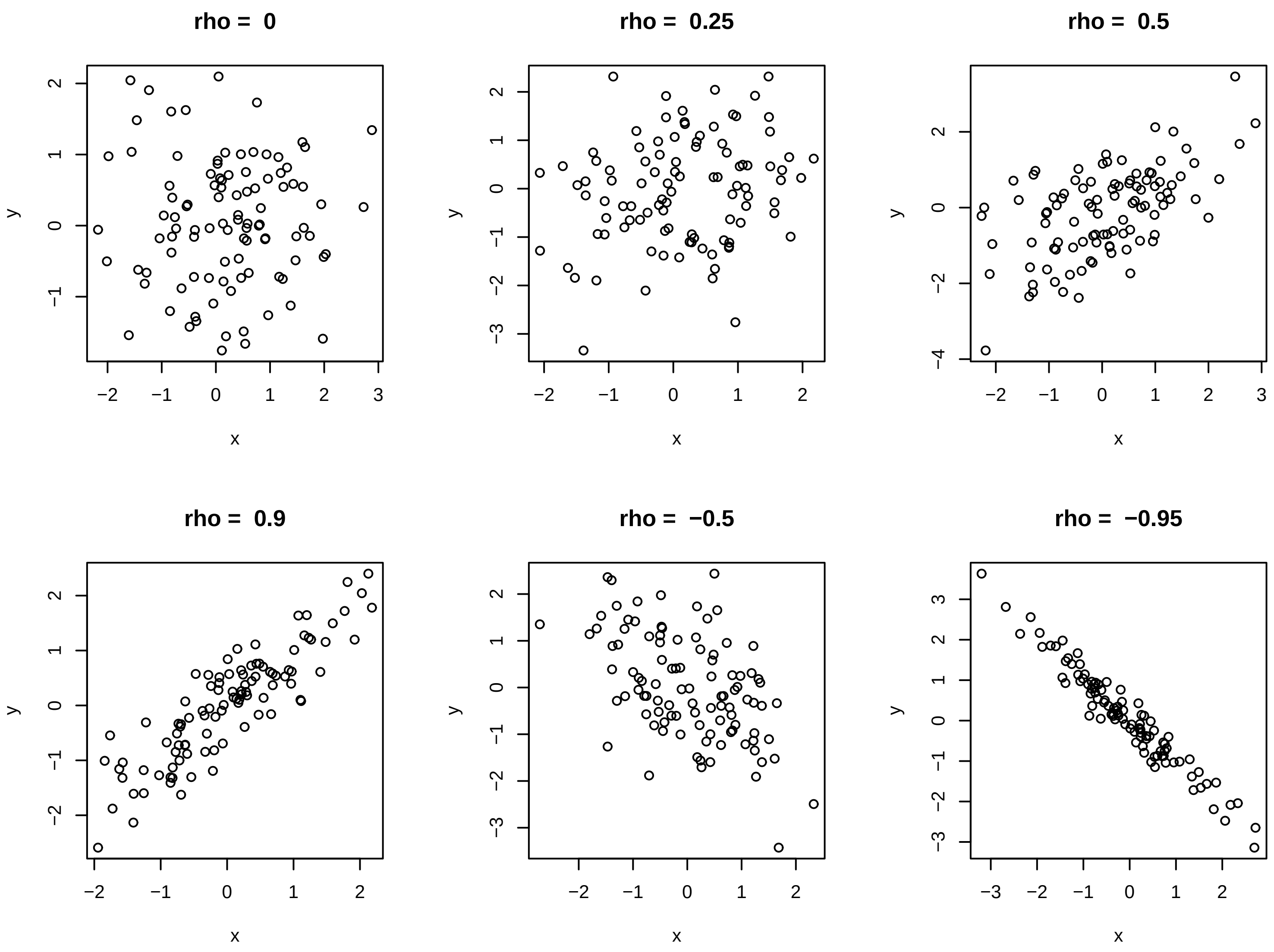
which implies $\mathrm{Corr}(X, Y) = \rho$. Also we've seen that

$$Y|X = x \sim N(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2)$$

so $E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ is a linear relationship with the slope proportional to ρ.

Also $\mathrm{Var}(Y|X = x) = (1 - \rho^2)\sigma_Y^2$ is constant for all x and this variance decreases as $|\rho|$ increases.

In general, the correlation coefficient ρ measures the strength of a linear relationship between two variables, not just for bivariate normal ones. In the case of bivariate normal, data under different ρ look like


rho = 0
rho = 0.25
rho = 0.5
rho = 0.9
rho = −0.5
rho = −0.95
x
y