Question 1

(a) Prove that for any random variable X and any constant $a \in \mathbb{R}$,

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

(b) Prove that for any random variable X, Y and Z, and constants $a, b \in \mathbb{R}$,

$$\mathrm{Cov}\,(aX + bY, Z) = a\mathrm{Cov}\,(X, Z) + b\mathrm{Cov}\,(Y, Z)\,.$$

(c) For any random variables X and Y, and constants $a, b \in \mathbb{R}$, find an expression for

$$\mathrm{Cov}\,(aX + b, Y)$$

in terms of $\mathrm{Cov}\,(X, Y)$.

Solution to Question 1

Part (a):

Recall the definition of covariance:

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

where $\mu_X = \mathrm{E}[X]$ and $\mu_Y = \mathrm{E}[Y]$. Since a is a constant, $\mathrm{E}[a] = a$. Therefore,

$$\mathrm{Cov}\,(X, a) = \mathrm{E}[(X - \mu_X)\,(a - a)] = \mathrm{E}[0] = 0.$$

Part (b):

Using the linearity of expectation,

$$\mu_{aX+bY} = \mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y] = a\mu_X + b\mu_Y$$

Therefore, writing $\mu_Z = \mathrm{E}[Z]$, and using the linearity of expectation

$$\begin{aligned}
\mathrm{Cov}\,(aX + bY, Z) &= \mathrm{E}[(aX + bY - \mu_{aX+bY})\,(Z - \mu_Z)] \\
&= \mathrm{E}[(aX + bY - a\mu_X - b\mu_Y)\,(Z - \mu_Z)] \\
&= \mathrm{E}[(aX - a\mu_X + bY - b\mu_Y)\,(Z - \mu_Z)] \\
&= \mathrm{E}[(aX - a\mu_X)\,(Z - \mu_Z) + (bY - b\mu_Y)\,(Z - \mu_Z)] \\
&= \mathrm{E}[(aX - a\mu_X)\,(Z - \mu_Z)] + \mathrm{E}[(bY - b\mu_Y)\,(Z - \mu_Z)] \\
&= \mathrm{E}[a\,(X - \mu_X)\,(Z - \mu_Z)] + \mathrm{E}[b\,(Y - \mu_Y)\,(Z - \mu_Z)] \\
&= a\mathrm{E}[(X - \mu_X)\,(Z - \mu_Z)] + b\mathrm{E}[(Y - \mu_Y)\,(Z - \mu_Z)] \\
&= a\mathrm{Cov}\,(X, Z) + b\mathrm{Cov}\,(Y, Z)
\end{aligned}$$

Part (c):

Using Part (b),

$$\mathrm{Cov}\left(aX+b,Y\right)=a\mathrm{Cov}\left(X,Y\right)+\mathrm{Cov}\left(b,Y\right)$$

and using Part (a), $\mathrm{Cov}\left(b,Y\right)=0$, which implies

$$\mathrm{Cov}\left(aX+b,Y\right)=a\mathrm{Cov}\left(X,Y\right).$$

(To justify the first part, you could consider $b=bZ$, where Z is the random variable which is identically equal to 1, and then proceed from there.)

Question 2

Suppose you are tracking the value of two companies listed on the London Stock Exchange over the course of one week. Rather than record the actual values of the share prices, you record the increase or decrease in each share price value at the daily close to the nearest pound. You record the following table:

	Monday	Tuesday	Wednesday	Thursday	Friday
Company X	5	4	8	6	2
Company Y	3	2	7	4	-1

Table 1: Daily change in share price (£)

Do the following calculations (there is no need to use a calculator):

(a) Compute the sample covariance between the two sequences to two decimal places.

(b) Compute the sample correlation between the two sequences. You may leave your answer as a fraction.

(c) Compute the sample correlation between the two sequences to two decimal places.

(d) Are the two sequences significantly correlated?

Solution to Question 2

Part (a):

Let $\mathbf{x} = (5, 4, 8, 6, 2)$ and let $\mathbf{y} = (3, 2, 7, 4, -1)$. Then

$$\overline{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(25) = 5$$

$$\overline{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5}(15) = 3$$

Then,

$$\mathbf{x} - \overline{x} = (0, -1, 3, 1, -3)$$
$$\mathbf{y} - \overline{y} = (0, -1, 4, 1, -4)$$

Which implies that the sample covariance is

$$\frac{1}{5-1}\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{4}\left[(0)(0) + (-1)(-1) + (3)(4) + (1)(1) + (-3)(-4)\right]$$

$$= \frac{1}{4}\left[0 + 1 + 12 + 1 + 12\right]$$

$$= \frac{26}{4} = \frac{13}{2} = 6.5$$

Part (b):

We need to compute the sample variances of $\mathbf{x}$ and $\mathbf{y}$, or at least the sum of squared deviations:

$$\sum_{i=1}^{5}(x_i-\overline{x})^2=(0)^2+(1)^2+(3)^2+(1)^2+(-3)^2=0+1+9+1+9=20$$
$$\sum_{i=1}^{5}(y_i-\overline{y})^2=(0)^2+(-1)^2+(4)^2+(1)^2+(-4)^2=0+1+16+1+16=34$$

Then the sample correlation is

$$r_{XY}=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{5}(y_i-\overline{y})^2}}.$$

$$=\frac{26}{\sqrt{20}\sqrt{34}}=\frac{26}{2\sqrt{5}\sqrt{34}}=\frac{13}{\sqrt{5}\sqrt{34}}$$

Part (c): You can use a calculator for this question, but it is not necessary. Using the computation in Part (b):

$$r_{XY}=\frac{13}{\sqrt{5}\sqrt{34}}=\frac{\sqrt{169}}{\sqrt{170}}=\sqrt{\frac{169}{170}}=\sqrt{\frac{170-1}{170}}=\sqrt{1-\frac{1}{170}}$$

$$>\sqrt{1-\frac{1}{100}}=\sqrt{0.99}$$

$$>0.99$$

The first inequality follows since

$$\begin{aligned}&170>100\\ \Rightarrow&\frac{1}{170}<\frac{1}{100}\\ \Rightarrow&1-\frac{1}{170}>1-\frac{1}{100}\\ \Rightarrow&\sqrt{1-\frac{1}{170}}>\sqrt{1-\frac{1}{100}}\end{aligned}$$

The second inequality follows because the functions $f(x)=\sqrt{x}$ and $g(x)=x$ on the interval (0,1) have the property $f(x)>g(x)$ (to see this, plot the functions).

Finally, although we always have $r_{XY}\leq 1$, in this case $\frac{169}{170}<1\Rightarrow\sqrt{\frac{169}{170}}<1$, and therefore

$$0.99<r_{XY}<1.$$

Part (d):

Although we do not have the distribution for r_{XY} in this case, and so cannot make an inference with any degree of confidence, a value of $r_{XY}>0.99$ is likely to be significant.

Question 3

Suppose that X and Y are two normally distributed random variables that are neither independent nor identically distributed. In fact, suppose it is known that $X \sim \mathrm{N}(1, 8)$ and $Y \sim \mathrm{N}(5, 2)$, and their correlation is $\mathrm{Cor}(X, Y) = \frac{1}{9}$. Defining the new random variable $Z = 2X + Y$, compute the correlation $\mathrm{Cor}(Y, Z)$.

Solution to Question 3

By properties of the covariance function,

$$\begin{aligned}\mathrm{Cov}(Y, Z) &= \mathrm{Cov}(Y, 2X + Y) \\ &= \mathrm{Cov}(Y, 2X) + \mathrm{Cov}(Y, Y) \\ &= \mathrm{Cov}(2X, Y) + \mathrm{Var}(Y) \\ &= 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).\end{aligned}$$

Now, since the correlation is defined as $\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$, we have

$$\mathrm{Cov}(X, Y) = \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}\mathrm{Cor}(X, Y) = \sqrt{8}\sqrt{2}(\tfrac{1}{9}) = \frac{4}{9}.$$

Then,

$$\begin{aligned}\mathrm{Cov}(Y, Z) &= 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y) \\ &= 2(\tfrac{4}{9}) + 2 \\ &= \tfrac{8}{9} + 2 \\ &= \frac{26}{9}.\end{aligned}$$

Now, by the properties of the variance (or Exercise 6.1.4 in the notes),

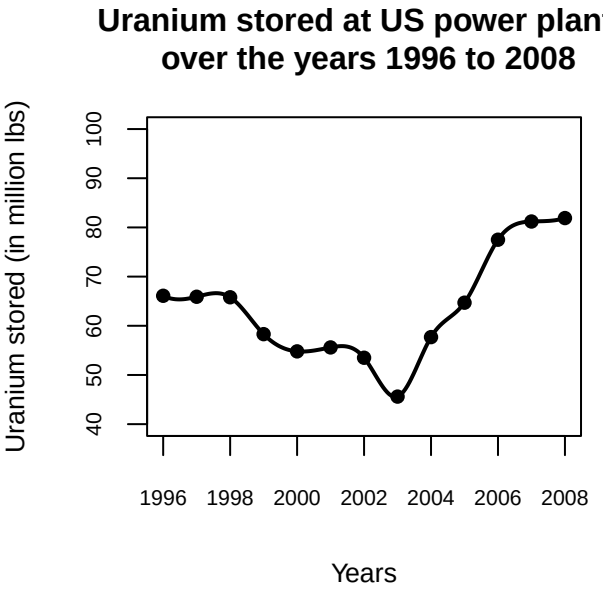
$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}(2X + Y) \\ &= \mathrm{Var}(2X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(2X, Y) \\ &= 4\mathrm{Var}(X) + \mathrm{Var}(Y) + 4\mathrm{Cov}(X, Y) \\ &= 4 \cdot 8 + 2 + 4(\tfrac{4}{9}) \\ &= 34 + \tfrac{16}{9} \\ &= \frac{322}{9}.\end{aligned}$$

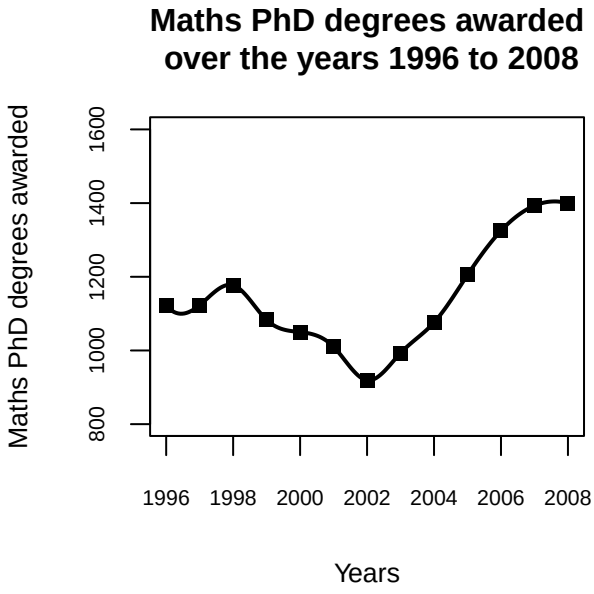
Finally, we can compute the correlation of Y and Z as

$$\begin{aligned}\mathrm{Cor}(Y, Z) &= \frac{\mathrm{Cov}(Y, Z)}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Z)}} = \frac{\frac{26}{9}}{\sqrt{2}\sqrt{\frac{322}{9}}} = \frac{\frac{26}{9}}{\frac{2}{3}\sqrt{161}} \\ &= \frac{13}{3\sqrt{161}} \approx 0.3415.\end{aligned}$$

Question 4

The figures below show X, the amount of uranium stored in US power plants and Y, the number of PhD degrees awarded in mathematics in the US for the years 1996 to 2008. The graphs appear to be very similar, and if one computes the sample correlation between the two data sets for X and Y, one obtains a value of $r = 0.952$. Taking this sample correlation value into account, can we conclude that the two quantities X and Y are related and have an influence on each other? Provide justification for your answer.





Solution to Question 4

We cannot make any conclusion regarding the dependence/influence of X and Y on each other, since correlation does not imply causation.