Covariance - key facts and exercises

In the following capital letters will denote random variables, unless otherwise specified, and \( \mu \) will denote a mean, often with a subscript to clarify to which random variable it applies, so. e.g., \( \mu_X = E(X) \). Lower case letter will denote constants.

**Background:** In the study of a single random variable \( X \) its mean and variance are important. When studying a higher dimensional variable, such as a *bivariate* random variable \((X, Y)\), then these quantities remain important but they do not address a new question which arises, namely how the variables vary together. For example, do higher values of \( X \) tend to occur with higher values of \( Y \)? An important quantity in addressing this question is the covariance and its scaled variant, the correlation.

**Definition:** \( \text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \)

**Key facts:**

1. If \( X \) and \( Y \) are independent then \( \text{cov}(X, Y) = 0 \) (but *not* conversely)
2. \( \text{cov}(X, Y) \) can be negative
3. \( \text{cov}(X, X) = \text{var}(X) \)
4. \( \text{cov}(X, Y) = \text{cov}(Y, X) \)
5. \( \text{cov}(aX, bY) = ab\text{cov}(X, Y) \)
6. \( \text{cov}(X, Y) = E(XY) - E(X)E(Y) \)
7. \( \text{cov}(X, Y) = E(X'Y') \)

The final item is just a restatement of the definition but serves to introduce the dash notation, where a dash denotes *centering*, i.e. subtracting its mean from a random variable, so \( X' = X - \mu_X \). This is useful in deriving more complex formulae for covariances

**Important formulae involving covariances:**

1. \( \text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2\text{cov}(X, Y) \)
2. The correlation between \( X \) and \( Y \), \( \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \)
3. \( \text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z) \)
4. \( \text{cov}\left(\sum_{i=1}^{n} a_iX_i, \sum_{j=1}^{m} b_jY_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{cov}(X_i, Y_j) \)
5. If \( X_1, \ldots, X_n \) are independent then \( \text{cov}(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j X_j) = \sum_{i=1}^{n} a_i b_i \text{var}(X_i) \)

6. A special case of 4 is \( \text{var}(\sum_{i=1}^{n} a_i X_i) = \sum_i a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j) \)

**Sample covariances and simulation:** If \((X, Y)\) denotes a bivariate random variable with distribution \(D\) and \((X_1, Y_1), \ldots, (X_n, Y_n)\) are \(n\) independent pairs with distribution \(D\) then:

\[
\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})
\]

is an estimator of \(\text{cov}(X, Y)\) (where as usual \(\bar{X} = n^{-1} \sum_{i=1}^{n} X_i\) etc.). If we assume that each of \(X\) and \(Y\) have zero mean then, by the Strong Law of Large Numbers:

\[
\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \xrightarrow{a.s.} \text{cov}(X, Y) \quad \text{as} \quad n \to \infty
\]

n.b. the restriction to zero means is inessential but convenient.

It follows that if you can generate random variables from \(D\), e.g. in \(\mathbb{R}\), then we can generate many realisations, e.g., \((x_1, y_1), \ldots, (x_N, y_N)\) with, say \(N = 1000\) and compute \(C_i = x_i y_i\) and then the mean of the \(C_i\) is an approximation to \(\text{cov}(X, Y)\) (remember this assumes \(X\) and \(Y\) have zero mean).

---

**Worked example 1:**

Suppose that \(X_1\) and \(X_2\) are independent variables with mean 0 and variance \(\sigma^2\). What is the covariance of \(X_1 + 2X_2\) and \(4X_1 - 3X_2\)? What is the correlation between these variables?

**Solution 1:** Writing \(U = X_1 + 2X_2\) and \(V = 4X_1 - 3X_2\) then item 3 in the list of important formulae and property 5 gives:

\[
\text{cov}(U, V) = \text{cov}(X_1 + 2X_2, V) = \text{cov}(X_1, V) + 2\text{cov}(X_2, V) = \text{cov}(X_1, V) + 2[4\text{cov}(X_1, X_1) - 3\text{cov}(X_1, X_2)]
\]

The independence of \(X_1\) and \(X_2\) implies that \(\text{cov}(X_1, X_2) = 0\) and \(\text{cov}(X_1, X_1) = \text{var}(X_1) = \sigma^2\), so:

\[
\text{cov}(U, V) = 4\sigma^2 - 6\sigma^2 = -2\sigma^2
\]

The variance of \(U\) is \(\text{var}(X_1 + 2X_2) = \text{var}(X_1) + 4\text{var}(X_2) = 5\sigma^2\). The variance of \(V\) is \(\text{var}(4X_1 - 3X_2) = 16\sigma^2 + 9\sigma^2 = 25\sigma^2\). Therefore the correlation between \(U\) and \(V\) is:

\[
\text{corr}(U, V) = \frac{\text{cov}(U, V)}{\sqrt{\text{var}(U)\text{var}(V)}} = \frac{-2\sigma^2}{\sqrt{5\sigma^2 \cdot 25\sigma^2}} = -\frac{2}{5\sqrt{5}} \approx -0.179
\]

This can be illustrated by taking \(X_1\) and \(X_2\) to be independent standard Normal variables. A million realisations of each can be generated in \(\mathbb{R}\) and realisations of \(U\) and \(V\) can then be found. The covariance and correlation can then be found.
These are only sample values and we have had to assume a value for $\sigma$ but the sample sizes are large enough to illustrate the theoretically calculated values. The use of the Normal distribution is not necessary. Try carrying out the calculations using another distribution with mean 0 and variance 1 and see that the covariances and correlations remain very close to the theoretical values.

**Solution 2:** This uses a more formulaic approach to finding $\text{cov}(U, V)$ but is otherwise the same as Solution 1. As $X_1$ and $X_2$ are independent then formula 5 applies. Hence:

$$U = X_1 + 2X_2 \quad V = 4X_1 - 3X_2 \quad \Rightarrow a_1 = 1, \quad a_2 = 2 \quad b_1 = 4; \quad b_2 = -3$$

So

$$\text{cov}(U, V) = \left( \sum_{i=1}^{2} a_i b_i \right) \sigma^2 = (1 \times 4 + 2 \times (-3))\sigma^2 = -2\sigma^2$$

---

**Worked example 2:**

$X_1, \ldots, X_4$ are independent random variables each with variance $\sigma^2$. What is the covariance and correlation between $X_1 + X_2 + X_3 + X_4$ and $2X_1 - 3X_2 + 6X_3$.

As the random variables are independent, formula 5 can again be used. The covariance is therefore:

$$(1 \times 2 + 1 \times (-3) + 1 \times 6 + 1 \times 0)\sigma^2 = 5\sigma^2$$

To get the correlation we need the variance of $X_1 + X_2 + X_3 + X_4$, which is $[1^2 + 1^2 + 1^2 + 1^2]\sigma^2 = 4\sigma^2$ and the variance of $2X_1 - 3X_2 + 6X_3$, which is $[2^2 + (-3)^2 + 6^2 + 0^2]\sigma^2 = [4 + 9 + 36]\sigma^2 = 49\sigma^2$. Consequently the correlation is:

$$\frac{5}{\sqrt{4 \times 49}} = \frac{5}{14} = 0.357$$

Again this can be illustrated by simulation as follows:
Examples

In Q1-Q6 assume that $X_1, \ldots, X_n$ are independent random variables with variance $\sigma^2$.

**Q1** Find the covariance and correlation between $U = X_1 + 2X_2 + 3X_4$ and $V = 3X_1 + 2X_2 + X_4$?

**Q2** Find the covariance and correlation between $U = 2X_1 + X_2 - 7X_4$ and $V = 8X_1 + 4X_3 + 3X_4$?

**Q3** For any angles $\theta$ and $\phi$ what is the correlation between $X_1 \sin \theta + X_2 \cos \theta$ and $X_1 \sin \phi + X_2 \cos \phi$?

**Q4** Find the value of $a$ so that $U = X_1 + 2X_2 + 2X_3 + X_4$ and $V = X_1 + aX_2 + X_4$ are uncorrelated.

**Q5** If $S = 2X_1 + X_2 + X_3$ and $T = X_1 + aX_2 - 2X_3$ between what limits must $\text{corr}(S, T)$ lie?

**Q6** If $U = \sum_{i=1}^{n} X_i$ and $V = \sum_{j=1}^{n} b_j X_j$ what condition must the $b_i$ satisfy if $\text{corr}(U, V) = 0$?

**Q7** What is the covariance between $X_1 - 3X_4$ and $X_2 + 16X_3$? What is the correlation between these random variables?

In the following assume $\text{cov}(X_i, X_j) = \rho \sigma^2$ ($i \neq j$) and $\sigma^2$ ($i = j$)

**Q8** What is $\text{cov}(X_1 + X_2, X_1 - X_2)$?

**Q9** What is $\text{cov}(X_1 + X_3, X_1 - 2X_2 + 4X_3)$? Also find $\text{corr}(X_1 + X_3, X_1 - 2X_2 + 4X_3)$.

**Q10** What is $\text{corr}(X_1, X_1 + X_2)$?

**Q11** What is the covariance between $X_1 - 3X_4$ and $X_2 + 16X_3$? What is the correlation between these random variables?