Covariance and Total Expectation Intro

Covariance: measure of the relationship between two RVs

$$\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

The sign of $\text{cov}(X,Y)$ illustrates how X and Y are related; a positive value means that X and Y tend to increase and decrease together, while a negative value means that X increases as Y decreases (and vice versa). A covariance of zero means that the two random variables are uncorrelated—there is no relationship between them.

Properties: for random variables X, Y, Z and constant a,

- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X,Y)$
- $\text{cov}(X,X) = \text{Var}(X)$
- $\text{cov}(X,Y) = \text{cov}(Y,X)$
- Bilinearity: $\text{cov}(X+Y,Z) = \text{cov}(X,Z) + \text{cov}(Y,Z)$ and $\text{cov}(aX,Y) = a\,\text{cov}(X,Y)$

Conditional Expectation: When we want to find the expectation of a random variable X conditioned on an event A, we use the following formula:

$$\mathbb{E}[X \mid A] = \sum_x x \cdot \mathbb{P}[(X = x) \mid A].$$

This is an application of the definition of expectation. We still consider all values of X but reweigh them based on their probability of occurring together with A.

Total Expectation: For any random variable X and events $A_1, A_2, \ldots, A_n$ that partition the sample space Ω,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i]\,\mathbb{P}[A_i].$$

We can think of this as splitting the sample space into partitions (events) and looking at the expectation of X in each partition, weighted by the probability of that event occurring.

1 Covariance

Note 16

(a) We have a bag of 5 red and 5 blue balls. We take two balls uniformly at random from the bag without replacement. Let X_1 and X_2 be indicator random variables for the events of the first and second ball being red, respectively. What is $\text{cov}(X_1, X_2)$? Recall that $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

(b) Now, we have two bags A and B, with 5 red and 5 blue balls each. Draw a ball uniformly at random from A, record its color, and then place it in B. Then draw a ball uniformly at random from B and record its color. Let X_1 and X_2 be indicator random variables for the events of the first and second draws being red, respectively. What is $\text{cov}(X_1, X_2)$?

Solution:

(a) We can use the formula $\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$.

$$\mathbb{E}[X_1] = \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2},$$
$$\mathbb{E}[X_2] = \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2},$$
$$\mathbb{E}[X_1 X_2] = \frac{5}{10} \cdot \frac{4}{9} \times 1 + \left(1 - \frac{5}{10} \cdot \frac{4}{9}\right) \times 0 = \frac{2}{9}.$$

Therefore,

$$\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] = \frac{2}{9} - \frac{1}{2} \times \frac{1}{2} = -\frac{1}{36}.$$

(b) Again, we use the formula $\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$.

$$\mathbb{E}[X_1] = \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2}$$
$$\mathbb{E}[X_2] = \left(\frac{5}{10} \times \frac{6}{11} + \frac{5}{10} \times \frac{5}{11}\right) \times 1 + \left(\frac{5}{10} \times \frac{5}{11} + \frac{5}{10} \times \frac{6}{11}\right) \times 0 = \frac{1}{2}$$
$$\mathbb{E}[X_1 X_2] = \frac{5}{10} \times \frac{6}{11} \times 1 = \frac{30}{110}.$$

Therefore,

$$\mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = \frac{30}{110} - \frac{1}{4} = \frac{1}{44}.$$

Note that in part (a), if one event happened, the other would be less likely to happen, and thus the covariance was negative. Similarly, in part (b), if one event happened, the other would be more likely to happen, and thus the covariance was positive.

Regression Intro

Note 20

Estimation: In estimation, we have an unknown random variable Y that we want to estimate. Y may also depend on another random variable X that we know. In the simplest case, we don't incorporate any information about X when creating our estimate $\hat{Y}$ and just estimate Y with a constant. Our choice of constant will minimize the **mean squared error**, $\mathbb{E}[(Y - \hat{Y})^2]$. This minimum occurs at

$$\hat{Y} = \mathbb{E}[Y].$$

If we want to incorporate X into our estimate, we can model $Y = g(X)$ and try to find the best $\hat{Y}$ such that the mean squared error $\mathbb{E}[(Y - \hat{Y})^2 \mid X]$ is again minimized. This occurs at

$$\hat{Y} = \mathbb{E}[Y \mid X].$$

We call this the **minimum mean squared estimate** (MMSE) of Y given X.

Since finding the conditional expectation is often very difficult, we compromise by estimating with a *linear function*: $\hat{Y} = aX + b$. Here, we want to minimize $\mathbb{E}[(Y - aX - b)^2 \mid X]$, which has a minimum at

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - \mathbb{E}[X]) \coloneqq \text{LLSE}[Y \mid X].$$

This is known as the **linear least squares estimate** (LLSE) of Y given X.

2 Number Game

Note 20

Sinho and Vrettos are playing a game where they each choose an integer uniformly at random from $[0, 100]$, then whoever has the larger number wins (in the event of a tie, they replay). However, Vrettos doesn't like losing, so he's rigged his random number generator such that it instead picks randomly from the integers between Sinho's number and 100. Let S be Sinho's number and V be Vrettos' number.

(a) What is $\mathbb{E}[S]$?

(b) What is $\mathbb{E}[V \mid S = s]$, where s is any constant such that $0 \leq s \leq 100$?

(c) What is $\mathbb{E}[V]$?

Alec sees Sinho and Vrettos playing this game, and wants to estimate Vrettos' number using an estimator $\hat{V}$, which may be a function of another random variable. The goal is to minimize the mean squared error (MSE) of the estimator, which is defined as $\text{MSE}(\hat{V}) = \mathbb{E}[(\hat{V} - V)^2]$.

(d) If Alec sees no information about either players' number, what is the optimal constant estimator $\hat{V}$ that minimizes the mean squared error?

(e) Now, assume that Alec sees Sinho's number and uses it to estimate Vrettos' number. What is the optimal estimator $\hat{V}(S)$ that minimizes the mean squared error (i.e. the MMSE)?

(f) Assuming the same conditions as the previous part, what is the optimal linear estimator $LLSE[V \mid S] = aS + b$ that minimizes the mean squared error?

(g) What is the expected value of the MMSE estimator $\hat{V}(S)$ from part (e)? (Hint: Use the law of total expectation.)

Solution:

(a) S is a (discrete) uniform random variable between 0 and 100, so its expectation is $\frac{0+100}{2} = 50$.

(b) If $S = s$, we know that V will be uniformly distributed between s and 100. Similar to the previous part, this gives us that $\mathbb{E}[V \mid S = s] = \frac{s+100}{2}$.

(c) With the law of total expectation, we have that

$$\begin{aligned}
\mathbb{E}[V] &= \sum_{s=0}^{100} \mathbb{E}[V \mid S = s] \cdot \mathbb{P}[S = s] \\
&= \sum_{s=0}^{100} \frac{s + 100}{2} \cdot \frac{1}{101} \\
&= \frac{1}{202} \left(\sum_{s=0}^{100} s + \sum_{s=0}^{100} 100 \right)
\end{aligned}$$

The first summation comes out to $\frac{100(100+1)}{2} = 50 \cdot 101$; the second summation is just adding 100 to itself 101 times, so it comes out to $100 \cdot 101$. Plugging these values in, we get $\mathbb{E}[V] = 75$.

Alternate Solution:
Using the previous part and the Law of Total Expectation, we get

$$\begin{aligned}\mathbb{E}[V] = \mathbb{E}[\mathbb{E}[V \mid S]] &= \mathbb{E}\left[\frac{S+100}{2}\right] \\ &= \frac{\mathbb{E}[S]+100}{2} \\ &= \frac{150}{2} = 75.\end{aligned}$$

(d) The optimal constant estimator is the mean of V, which is $\mathbb{E}[V] = 75$.

(e) The optimal estimator is $\hat{V} = \mathbb{E}[V \mid S] = \frac{S+100}{2}$.

(f) The optimal estimator is already linear in S, so the optimal linear estimator is the MMSE as computed in the previous part.

(g) The MMSE is $\hat{V}(S) = \mathbb{E}[V \mid S]$. Taking the outer expectation over S, we have $\mathbb{E}[\hat{V}(S)] = \mathbb{E}[\mathbb{E}[V \mid S]] = \mathbb{E}[V] = 75$.

3 LLSE

Note 20

We have two bags of balls. The fractions of red balls and blue balls in bag A are $2/3$ and $1/3$ respectively. The fractions of red balls and blue balls in bag B are $1/2$ and $1/2$ respectively. Someone gives you one of the bags (unmarked) uniformly at random. You then draw 6 balls from that same bag with replacement. Let X_i be the indicator random variable that ball i is red. Now, let us define $X = \sum_{1 \le i \le 3} X_i$ and $Y = \sum_{4 \le i \le 6} X_i$.

(a) Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(b) Compute $\text{Var}(X)$.

(c) Compute $\text{cov}(X,Y)$. (*Hint*: Recall that covariance is bilinear.)

(d) Now, we are going to try and predict Y from a value of X. Compute $L(Y \mid X)$, the best linear estimator of Y given X. Recall that

$$L(Y \mid X) = \mathbb{E}[Y] + \frac{\text{cov}(X,Y)}{\text{Var}(X)}\big(X - \mathbb{E}[X]\big).$$

Solution: Although the indicator random variables are not independent, we can still apply linearity of expectation. By symmetry, we also know that each indicator follows the same distribution.

(a)

$$\mathbb{E}[X] = \mathbb{E}[Y] = 3 \cdot \mathbb{E}[X_1] = 3 \cdot \mathbb{P}[X_1 = 1] = 3 \cdot \left(\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{2}\right) = \frac{7}{4}.$$

(b)

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{cov}\Big(\sum_{1\le i\le 3} X_i, \sum_{1\le j\le 3} X_j\Big)\\ &= 3\cdot \mathrm{Var}(X_1) + 6\cdot \mathrm{cov}(X_1, X_2)\\ &= 3\big(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2\big) + 6\cdot(\mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2])\\ &= 3\Big[\frac{7}{12} - \Big(\frac{7}{12}\Big)^2\Big] + 6\cdot(\frac{1}{2}\cdot\Big(\frac{2}{3}\Big)^2 + \frac{1}{2}\cdot\Big(\frac{1}{2}\Big)^2 - \Big(\frac{7}{12}\Big)^2)\\ &= \frac{111}{144}.\end{aligned}$$

(c)

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathrm{cov}\Big(\sum_{1\le i\le 3} X_i, \sum_{4\le j\le 6} X_j\Big)\\ &= 9\cdot \mathrm{cov}(X_1, X_4)\\ &= 9\cdot\big(\mathbb{E}[X_1X_4] - \mathbb{E}[X_1]\cdot\mathbb{E}[X_4]\big)\\ &= 9\cdot\Big(\mathbb{P}[X_1 = 1, X_4 = 1] - \mathbb{P}[X_1 = 1]^2\Big)\\ &= 9\cdot\left(\Big[\frac{1}{2}\cdot\Big(\frac{2}{3}\Big)^2 + \frac{1}{2}\cdot\Big(\frac{1}{2}\Big)^2\Big] - \Big[\frac{1}{2}\cdot\Big(\frac{2}{3}\Big) + \frac{1}{2}\cdot\Big(\frac{1}{2}\Big)\Big]^2\right) = \frac{9}{144}.\end{aligned}$$

(d)

$$L(Y \mid X) = \frac{7}{4} + \frac{9}{111}\Big(X - \frac{7}{4}\Big) = \frac{3}{37}X + \frac{119}{74}.$$