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 $\text{cov}(X_1, X_2)$?

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

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

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.)

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

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