Covariance and Correlation

CS 70, Summer 2019

Lecture 22, 7/31/19

Last Time...

- Variance measures **deviation from mean**
- Variance is additive for **independent RVs**
- Use **linearity of expectation** and **indicator variables**

Today:

- Proof that var. is additive for ind. RVs
- Talk about **covariance and correlation**
- Some RV practice (if time)

Product of RVs

Let X be a RV with values in A.
Let Y be a RV with values in B.

What does the distribution of XY look like?

$XY = \left\{ \text{all } (ab) \text{ where } a \in A, b \in B \right.$ $\quad \mathbb{P}[XY = ab] = \mathbb{P}[X = a, Y = b]$

What if X, Y are **independent**?

$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\,\mathbb{P}[Y = b]$ $\quad \forall a \in A, b \in B$

will treat things like $X = 1, Y = 2$ vs $Y = 1, X = 2$ as 2 separate cases.

For Independent RVs...

If X and Y are independent, we can show that:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Definition ×2

RHS: $\mathbb{E}[X]\mathbb{E}[Y] = \left(\sum_{a \in A} a \cdot \mathbb{P}[X = a]\right)\left(\sum_{b \in B} b \cdot \mathbb{P}[Y = b]\right)$

Distribute sum: $= \sum_{a \in A, b \in B} (ab)\underbrace{\left[\mathbb{P}[X = a]\,\mathbb{P}[Y = b]\right]}_{= \mathbb{P}[X = a, Y = b] \text{ by indep.}}$

$= \sum_{a \in A, b \in B} (ab) \cdot \mathbb{P}[X = a, Y = b]$

$= \mathbb{E}[XY]$ ← see prev slide, when XY's dist. defined.

Variance is Additive for Ind. RVs

If X and Y are independent, we can show that:

$$\text{Var}(X + Y) = \text{Var}[X] + \text{Var}[Y]$$

Def.

$\mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2$ $\quad \mu_1 := \mathbb{E}[X]$, $\mu_2 := \mathbb{E}[Y]$

lin / lin

$\mathbb{E}[X^2 + 2XY + Y^2] - [\mu_1 + \mu_2]^2$

lin

$\mathbb{E}[X^2] + \cancel{2\mathbb{E}[XY]} + \mathbb{E}[Y^2] - \mu_1^2 - \cancel{2\mu_1\mu_2} - \mu_2^2$

cancel b/c $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

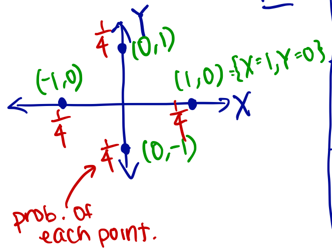
$\text{Var}(X) + \text{Var}(Y)$ $\quad$ * For multiple RVs: induction!!

Can You Multiply?

If X and Y are **independent**, then:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Is the **converse** true? No.



$\mathbb{E}[X] = 0$
$= 0 \cdot \frac{1}{2} + (-1)\frac{1}{4} + 1\left(\frac{1}{4}\right)$

$\mathbb{E}[Y] = 0$

$XY = 0$ always.

$\mathbb{E}[XY] = 0$

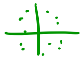
X, Y dep.

$\mathbb{P}[X = 0, Y = 0] \neq \mathbb{P}[X = 0]\,\mathbb{P}[Y = 0]$

Covariance

if cov(X,Y) > 0    cov(X,Y) < 0 

Measures "how independent" two RVs are. Let $\mathbb{E}[X] = \mu_1$, and let $\mathbb{E}[Y] = \mu_2$.

cov(X,Y) = 0 

$$\text{Cov}(X, Y) = \mathbb{E}[(X-\mu_1)(Y-\mu_2)]$$

Alternate Form:

$\mathbb{E}[\mu_1 Y] = \mu_1 \mathbb{E}[Y] = \mu_1\mu_2$

$$\text{Cov}(X, Y) = \boxed{\mathbb{E}[XY] - \mu_1\mu_2}$$

$\mathbb{E}[(X-\mu_1)(Y-\mu_2)] = \mathbb{E}[XY - \mu_1 Y - \mu_2 X + \mu_1\mu_2]$

lin. $\rightarrow$ $= \mathbb{E}[XY] - \cancel{\mu_1\mu_2} - \mu_2\mu_1 + \cancel{\mu_1\mu_2}$

$= \mathbb{E}[XY] - \mu_1\mu_2$

Example: Coin Flips I

I flip two **fair** coins.
Let X count the number of heads, and let Y be an indicator for the first coin being a head.

First, what is XY?

$X = \begin{cases} 0 & \text{wp } \frac{1}{4} \quad TT \\ 1 & \frac{1}{2} \quad TH, HT \\ 2 & \frac{1}{4} \quad HH \end{cases}$

$X \sim \text{Bin}(2, \frac{1}{2})$

$\mathbb{E}[X] = np = 1$

$Y = \begin{cases} 0 & \text{wp } \frac{1}{2} \quad TT, TH \\ 1 & \text{wp } \frac{1}{2} \quad HH, HT \end{cases}$

$\mathbb{E}[Y] = \frac{1}{2}$

$XY = \begin{cases} 0 & \text{wp } \frac{1}{2} \quad TT, TH \\ 1 & \frac{1}{4} \quad HT \\ 2 & \frac{1}{4} \quad HH \end{cases}$

$\mathbb{E}[XY] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4}$
$= \frac{3}{4}$

$\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$
$\Rightarrow X, Y$ not ind!!
(contrapositive)

Example: Coin Flips I

What is $\text{Cov}(X, Y)$?

$\underbrace{\mathbb{E}[XY]} - \mathbb{E}[X]\mathbb{E}[Y]$

$\frac{3}{4} - 1\left(\frac{1}{2}\right)$

$\frac{1}{4}$

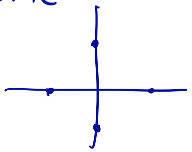
Properties of Covariance I

If X and Y are independent, then:

$\text{Cov}(X, Y) = 0$ because $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Is the **converse** true? NO.

Same counterexample as before.



Properties of Covariance II

What happens when we take $\text{Cov}(X, X)$?

$\mu = \mathbb{E}[X]$

Can use **either definition** of covariance!

① $\text{cov}(X,X) = \mathbb{E}[(X-\mu)(X-\mu)]$
$= \mathbb{E}[(X-\mu)^2] = \text{Var}(X)$

② $\text{cov}(X,X) = \mathbb{E}[X(X)] - \mathbb{E}[X]\mathbb{E}[X]$
$= \text{Var}(X)$

Properties of Covariance III

Covariance is **bilinear.**

$\text{Cov}(a_1X_1 + a_2X_2, Y) =$
$a_1 \text{Cov}(X_1, Y) + a_2 \text{Cov}(X_2, Y)$

$\text{Cov}(X, b_1Y_1 + b_2Y_2) =$
$b_1 \text{Cov}(X, Y_1) + b_2 \text{cov}(X, Y_2)$

⊛ $\text{Var}(cX) = \text{cov}(cX, cX) = c\,\text{Cov}(X, cX)$
$= c^2 \text{Cov}(X, X).$

Practice: Bilinearity!

Simplify $\text{Cov}(3X + 4Y, 5X - 2Y)$.

first coord → $= 3\,\text{Cov}(X, 5X-2Y) + 4\,\text{Cov}(Y, 5X-2Y)$

$= 15\,\text{Cov}(X,X) - 6\,\text{Cov}(X,Y) + 20\,\text{Cov}(Y,X) - 8\,\text{Cov}(Y,Y)$

$= 15\,\text{Var}(X) + 14\,\text{Cov}(X,Y) - 8\,\text{Var}(Y)$

Example: Coin Flips II

Same setup: Two fair flips. X is number of heads, and Y is indicator for the first coin heads.

Recall: $\text{Cov}(X, Y) = \frac{1}{4}$

Now, let Y' be an indicator for the first coin being a tail. How does the covariance change?

$\text{Cov}(X, Y') = \text{Cov}(X, 1-Y)$

Observe: $Y' + Y = 1$
$Y' = 1 - Y$

$= \underbrace{\text{Cov}(X, 1)} - \text{Cov}(X,Y)$

$= \cancel{\mathbb{E}[X \cdot 1] - \mathbb{E}[X]\,\mathbb{E}[1]} - \frac{1}{4}$

$= -\frac{1}{4}$

Properties of Covariance IV

For **any** two RVs X, Y:

$\text{Var}(X + Y) = \text{Cov}(X+Y, X+Y)$

$= \text{Cov}(X, X+Y) + \text{Cov}(Y, X+Y)$

$= \text{Cov}(X,X) + \text{Cov}(X,Y) + \text{Cov}(Y,X) + \text{Cov}(Y,Y)$

$= \text{Var}(X) + 2\,\text{Cov}(X,Y) + \text{Var}(Y)$

Break

If you could eliminate one food so that **no one would eat it ever again**, what would you pick to destroy?

Correlation

so the denom. is $\neq 0$

For **any** two RVs X, Y that are **not constant**:

$\text{Corr}(X, Y) = \dfrac{\text{Cov}(X,Y)}{\sigma(X)\,\sigma(Y)}$

std dev $\sqrt{\text{Var}(Y)}$

Sanity Check!

What is $\text{Corr}(X, X)$? $\dfrac{\text{Var}(X)}{(\sqrt{\text{Var}(X)})^2} = 1$

What is $\text{Corr}(X, -X)$? $-1.$

What is $\text{Corr}(X, Y)$ for X, Y independent? 0

Example: Coin Flips III

I flip two **fair** coins.
Let X count the number of heads, and let Y be an indicator for the first coin being a head.

What is $\text{Corr}(X, Y)$?

$= \dfrac{\text{Cov}(X,Y)}{\sigma(X)\,\sigma(Y)}$ ← $\frac{1}{4}$

$X \sim \text{Bin}(2, \frac{1}{2})$
$\text{Var}(X) = np(1-p)$
$= \frac{1}{2}$
$\sigma(X) = \frac{\sqrt{2}}{2}$

$Y \sim \text{Ber}(\frac{1}{2})$
$\text{Var}(Y) = p(1-p)$
$= \frac{1}{4}$
$\sigma(Y) = \frac{1}{2}$

$\text{Corr}(X,Y) = \dfrac{\frac{1}{4}}{\frac{\sqrt{2}}{2} \cdot \frac{1}{2}} = \frac{\sqrt{2}}{2}$

Size of Correlation?

For any RVs X and Y that are **not constants**:

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

Proof: Define new RVs using X and Y:

$\tilde{X} =$

$\tilde{Y} =$

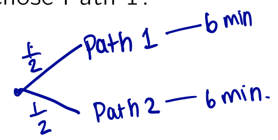
Size of Correlation?

(Continued:)

RV Practice: Two Roads

There are two paths from Soda to VLSB.
I usually choose a path uniformly at random.
minutes spent on Path 1 is a Geometric(p_1) RV.
minutes spent on Path 2 is a Geometric(p_2) RV.

Today, it took me 6 minutes to walk from Soda to VLSB. Given this, what is the probability that I chose Path 1?



T_1 = time path 1 $\sim \text{Geom}(p_1)$

T_2 = time path 2 $\sim \text{Geom}(p_2)$

W_1 = event I chose path 1

RV Practice: Two Roads

Continued:

Goal $\rightarrow$ $\mathbb{P}[W_1 \mid 6 \text{ min}] = \dfrac{\mathbb{P}[W_1 \cap (T_1 = 6)]}{\mathbb{P}[6 \text{ min}]}$ conditional prob. def.

$$= \frac{(\frac{1}{2})\underbrace{(1-p_1)^5 p_1}}{\mathbb{P}[W_1 \cap (T_1 = 6)] + \mathbb{P}[\overline{W_1} \cap (T_2 = 6)]}$$

from Geo(p_1) distribution.

$\mathbb{P}[T_1 = i] = (1-p_1)^{i-1} p_1$

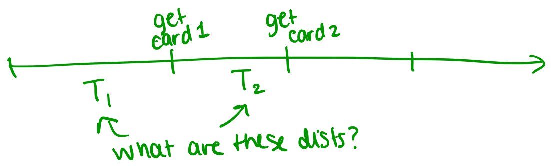
Total Prob Rule

$$= \frac{\cancel{\frac{1}{2}}(1-p_1)^5 p_1}{\cancel{\frac{1}{2}}(1-p_1)^5 p_1 + \cancel{\frac{1}{2}}(1-p_2)^5 p_2}$$

RV Practice: RandomSort

I have cards labeled $1, 2, \ldots, n$. They are shuffled.

I want them in order. I sort them in a naive way. I start with all cards in an "unsorted" pile. I draw cards from the unsorted pile uniformly at random until I get card 1. I place card 1 in a "sorted" pile, and continue, this time looking for card 2.



RV Practice: RandomSort

What is the expected number of draws I need?

What is the variance of the number of draws?

RV Practice: Packets

Packets arrive from sources A and B.
I fix a time interval. Over this interval, the number of packets from A and B have $\text{Poisson}(\lambda_A)$ and $\text{Poisson}(\lambda_B)$ distributions, and are **independent**.

What is the distribution of the total number of packets I receive in this time interval?

RV Practice: Packets

What is the probability that over this interval, I receive **exactly** 2 packets?

What is the expected number of packets I receive over this interval?

What is the variance of this number?

Summary

- Covariance and correlation measure **how independent** two RVs are.
- Variance can be **expressed** and **manipulated** in terms of covariance.
- Independent RVs have **zero covariance** and **zero correlation**. However, **the converse is not true!**