

Section 8.4

Covariance and correlation

So far, we have independent r.v.s, and dependent r.v.s.

Q) What can we say about the *level of dependence of r.v.s*?

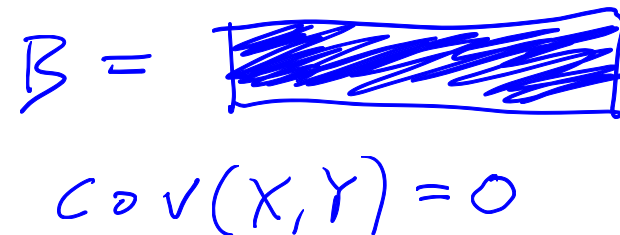
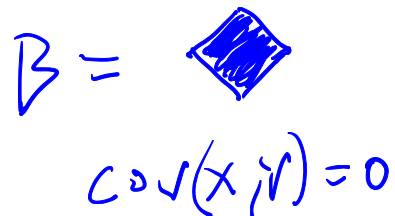
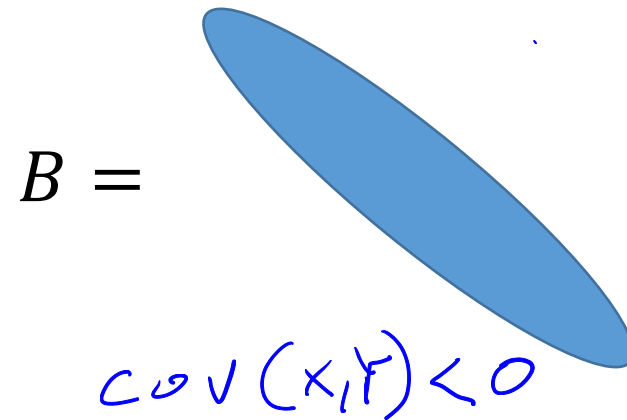
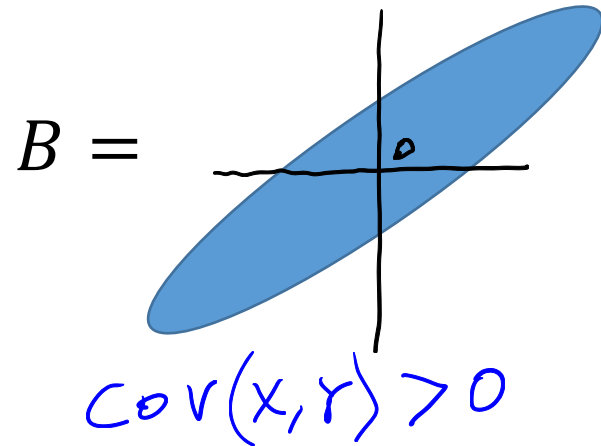
Covariance (and correlation) give a partial answer.

Def: Let X, Y be r.v.s with means μ_X, μ_Y . The *covariance* of X and Y is defined as

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

When is covariance positive?

Let $(X, Y) \sim \text{Unif}[B]$. Is covariance positive or negative?



Covariance and dependence

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

If X, Y are independent, then $\text{Cov}(X, Y) = \frac{\mathbb{E}[(X - \mu_X) \cdot (Y - \mu_Y)]}{\text{indep}}$
 $= \mathbb{E}[X - \mu_X] \cdot \mathbb{E}[Y - \mu_Y] = (\mathbb{E}[X] - \mu_X) \cdot (\mathbb{E}[Y] - \mu_Y)$
 $= \underbrace{\mu_X}_{\text{mean}} \cdot \underbrace{\mu_Y}_{\text{mean}} = 0 \cdot 0 = 0$

But, if $\text{Cov}(X, Y) = 0$, then X & Y are not
necessarily indep. Counterexample: X is any $\forall$ r.v.

$Y = Z \cdot X$ where Z is indep. of X & $Z = \begin{cases} 1 & \text{wp. } 1/2 \\ -1 & \text{wp. } 1/2 \end{cases}$

$$\text{Cov}(X, Y) = \mathbb{E}[X \cdot Y] = \mathbb{E}[X^2 \cdot Z] = \mathbb{E}[X^2] \cdot \mathbb{E}[Z] = 0$$

X, Y are dependent in general since $|X| = |Y|$.

Properties of covariance for r.v.s X, Y, Z

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

1. $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \cdot \text{Var}(Y)}$
2. Symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. (Bi-)linear: $\text{Cov}(aX + bZ, Y) = \underline{a \text{Cov}(X, Y) + b \text{Cov}(Z, Y)}$
4. $\text{Cov}(X, X) = \underline{\text{Var}(X)}$, $\text{Cov}(-X, X) = \underline{-\text{Var}(X)}$
5. $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \cdot \mu_Y$.

Alternative formula for cov. Exercise: prove it!

Examples

Let A, B be events and $1_A, 1_B$ be the indicators of those events. What is $\text{Cov}(1_A, 1_B)$?

$$\begin{aligned}\text{Cov}(1_A, 1_B) &= E[1_A 1_B] - E[1_A] \cdot E[1_B] \\ &= E[1_{AB}] - P(A) \cdot P(B) \\ &= P(AB) - P(A) \cdot P(B)\end{aligned}$$

Thus, A & B are indep $\Leftrightarrow \text{Cov}(1_A, 1_B) = 0$.

Ex) Let $(X, Y) \sim \text{Unif}(B)$, where $B = \{(x, y): x^2 + y^2 \leq 1\}$.

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

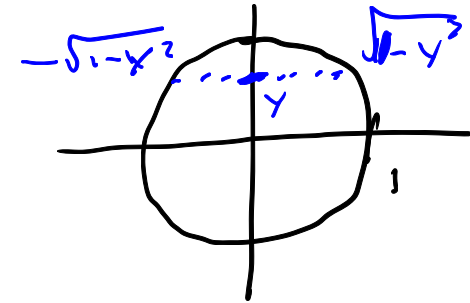
$$\mu_X = \mu_Y = 0 \quad \text{by symm}$$

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

$$= \iint xy f(x, y) dx dy$$

$$= \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} xy \frac{1}{\pi} dx dy$$

$$= \frac{1}{\pi} \int_{-1}^1 y \cdot \underbrace{\frac{x^2}{2} \Big|_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}}}_0 dy = 0$$



$$f(x, y) = \begin{cases} 1/\pi & (x, y) \in B \\ 0 & (x, y) \notin B \end{cases}$$

Variance of an arbitrary sum of r.v.s

Recall: For *independent* r.v.s $\text{Var}(\sum_i X_i) = \underline{\sum_i \text{Var}(X_i)}$

Fact: For r.v.s $X_1, X_2, \dots, X_n$,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \underline{\sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)}$$

summed over $\binom{n}{2}$ pairs

Note: $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i,j} \mathbb{E}(X_i - \mu_{X_i})(X_j - \mu_{X_j})$

Special case, $\text{Var}(X + Y)$

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

Proof:
$$\begin{aligned}\text{Var}(X + Y) &= \mathbb{E}(X + Y - \mu_X - \mu_Y)^2 \\ &= \mathbb{E}[(X - \mu_X) + (Y - \mu_Y)]^2 \\ &= \mathbb{E}[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\ &= \text{Var}(X) + \text{Var}(Y) + 2 \text{Cor}(X, Y)\end{aligned}$$

Ex) Let X be the number of aces in a 5-card poker hand. What is $\text{Var}(X)$?

$$X = \sum_{i=1}^5 1_{A_i}$$

$A_i = \{i^{\text{th}} \text{ card is an ace}\}$

$$\text{Var}(X) = \sum_{i=1}^5 \text{Var}(1_{A_i}) + 2 \sum_{i < j} \text{cov}(1_{A_i}, 1_{A_j})$$

- $\text{Var}(1_{A_i}) = \text{Var}(\text{Bern}(P(A_i))) = P(A_i) \cdot (1 - P(A_i)) = \frac{4}{52} \cdot (1 - \frac{4}{52})$

- $\text{cov}(1_{A_i}, 1_{A_j}) = \underbrace{P(A_i A_j)}_{\frac{4}{52} \cdot \frac{3}{51}} - \underbrace{P(A_i) \cdot P(A_j)}_{(\frac{4}{52})^2} \leftarrow \text{see slide above}$

$$\Rightarrow \boxed{\text{Var}(X) = 5 \cdot \frac{4}{52} \left(1 - \frac{4}{52}\right) + 2 \cdot \binom{5}{2} \cdot \left(\frac{4}{52} \cdot \frac{3}{51} - \left(\frac{4}{52}\right)^2\right)}$$

Correlation: normalized version of covariance

Recall: $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \cdot \text{Var}(Y)}$

Def: $\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$

Properties:

- $-1 \leq \text{Corr}(X, Y) \leq 1$
- $Y = aX + b$ if and only if $\text{Corr}(X, Y) = \underline{1}$ $a > 0$
- $\text{Sign}(\text{Cov}(X, Y)) = \text{sign}(\text{Corr}(X, Y))$

Terminology: correlated

- Uncorrelated: $\text{Corr}(X, Y) = \underline{0}$
- Positively correlated: $\text{Corr}(X, Y) \underline{> 0}$
- Negatively correlated: $\text{Corr}(X, Y) \underline{< 0}$

Note: Since $\text{sign}(\text{Cov}(X, Y)) = \text{sign}(\text{Corr}(X, Y))$, we can replace $\text{Corr}(X, Y)$ with $\text{Cov}(X, Y)$ above.

Variance of uncorrelated sums

If $X_1, X_2, \dots, X_n$ are pairwise uncorrelated, i.e., $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, then

$$\text{Corr}(X_i, X_j) = 0$$

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$