Probability Theory

Applications for Data Science
Module 5: Expectation, Variance, Covariance, and Correlation

Anne Dougherty

March 20, 2021

TABLE OF CONTENTS

Expectation, Variance, Covariance, and Correlation

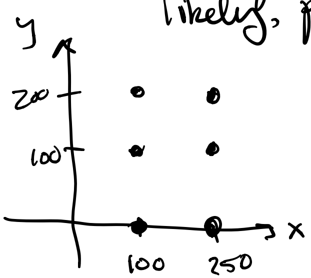
At the end of this module, students should be able to

- Compute the mean, variance, and standard deviation of a function of a random variable (i.e. $g(X)$).
- Explain the concept of jointly distributed random variables, for two random variables X and Y.
- **Define, compute, and interpret the covariance between two random variables X and Y.**
- **Define, compute, and interpret the correlation between two random variables X and Y.**

Example: An insurance agency services customers who have both a homeowner's policy and an automobile policy. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are \$100 or \$250 and for the homeowner's policy, the choices are \$0, \$100, or \$200.

Suppose the **joint probability table** is given by the insurance company as follows:

Note: each point is not equally likely, prob. are given by joint p.m.f.



		y (home)		
		0	100	200
x (auto)	100	.20	.10	.20
	250	.05	.15	.30

$P(X=100)=.5$

$P(X=250)=.5$

$P(Y=0)=.25$ $\quad P(Y=100)=.25$ $\quad P(Y=200)=.5$

$p(x,y) = P(X=x, Y=y)$ ← joint prob. mass function

When two random variables, X and Y, are not independent, it is frequently of interest to assess how strongly they are related to each other.

Definition: The **covariance** between two rv's, X and Y, is defined as:

$$\boxed{Cov(X,Y) = E[(X - E(X))(Y - E(Y))]}$$

$$= E[(X-\mu_X)(Y-\mu_Y)]$$

expectation need to sum over all possible x & y values

$$= \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) P(X=x, Y=y), & \text{(discrete)} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\,dx\,dy, & \text{(cont)} \end{cases}$$

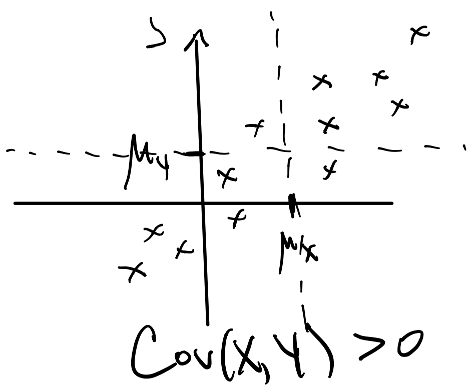
X & Y discrete

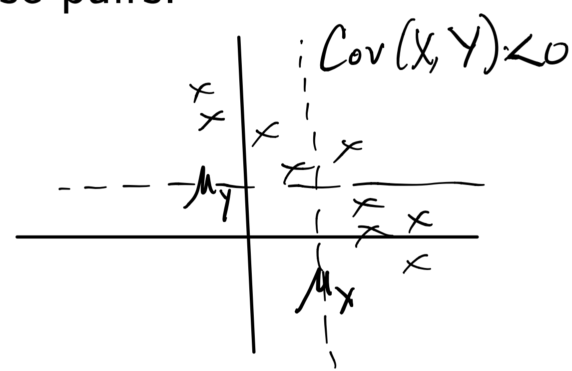
Definition: Covariance of X and Y is given by
$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$

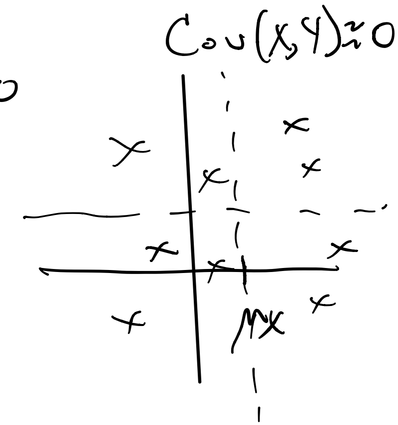
To calculate covariance:

$$Cov(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) P(X = x, Y = y) \\ \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\, dx\, dy \end{cases}$$

The covariance depends on both the set of possible pairs and the probabilities for those pairs.





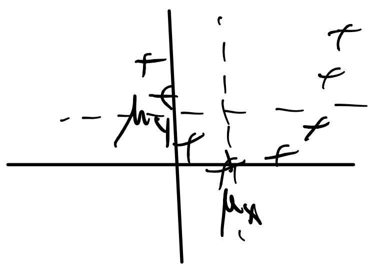


$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

- If both variables tend to deviate in the same direction (both go above their means or below their means at the same time), then the covariance will be positive.
- If the opposite is true, the covariance will be negative.
- If X and Y are not strongly (linearly) related, the covariance will be near 0.

Aside: It is possible to have a strong relationship between X + Y and still have $Cov(X, Y) \approx 0$

e.g.



$$Cov(X,Y) = \sum_x \sum_y (x-\mu_x)(y-\mu_y) P\{X=x, Y=y\}$$

Covariance example calculation:

	y (home)		
	0	100	200
x (auto) 100	.20	.10	.20
250	.05	.15	.30

$$\mu_X = \sum_x x P(X=x) = 100(.5) + 250(.5) = 175$$

$$\mu_Y = \sum y P(Y=y) = 0(.25) + 100(.25) + 200(.5) = 125$$

x	y	$x-\mu_X$	$y-\mu_Y$	$P(X=x, Y=y)$
100	0	-75	-125	.2
250	0	75	-125	.05
100	100	-75	-25	.1
250	100	75	-25	.15
100	200	-75	75	.2
250	200	75	75	.3

$Cov(X,Y) = 1875$

Is this a strong relationship between X & Y?
It seems like a "big" number, but it's hard to say. The correlation coef. will help.

But, before we get to correlation, there's a few more ideas related to covariance we need to discuss.

Computational formula for covariance:

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Recall: $V(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - \left(E(X)\right)^2$

$$Cov(X, Y) = E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)\right)$$

$$= E\left[XY - XE(Y) - YE(X) + E(X)E(Y)\right]$$

$$= E(XY) - E(X\underbrace{E(Y)}_{\text{constant}}) - E(Y\underbrace{E(X)}_{\text{constant}}) + E(\underbrace{E(X)E(Y)}_{\text{constant}})$$

$$= \underbrace{E(XY) - E(X)E(Y)}$$

What if X and Y are independent?

If X & Y are indep., $P(X=x, Y=y) = P(X=x)\,P(Y=y)$ for all x, y.
& discrete

$$Cor(X,Y) = \sum_x \sum_y (x-\mu_x)(y-\mu_y)\,P(X=x, Y=y)$$

X & Y indep

$$\overset{\checkmark}{=} \sum_x \sum_y (x-\mu_x)(y-\mu_y)\,P(X=x)\,P(Y=y)$$

$$= \left[\sum_x (x-\mu_x)\,P(X=x)\right]\left[\sum_y (y-\mu_y)\,P(Y=y)\right]$$

$$= \Big[\underbrace{\sum_x x\,P(X=x)}_{E(X)} - \mu_x \underbrace{\sum_x P(X=x)}_{1}\Big]\Big[\quad \text{similar} \quad\Big]$$

$$= 0$$

So, if X & Y are indep, $Cor(X,Y) = 0$

* It does not go the other way. If $Cor(X,Y)=0$ X & Y may still be dependent.

Useful formulas for random variables X and Y and real numbers a and b:

- $E(aX + bY) = aE(X) + bE(Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$

$$V(aX+bY) = E\left[\left(aX+bY - E(aX+bY)\right)^2\right]$$
$$= E\left[\left(a(X - E(X)) + b(Y - E(Y))\right)^2\right]$$
$$= a^2 E\left[(X - E(X))^2\right] + b^2 E\left[(Y - E(Y))^2\right]$$
$$+ 2ab\, E\left[(X - E(X))(Y - E(Y))\right]$$
$$= a^2 V(X) + b^2 V(Y) + 2ab\, Cov(X, Y)$$

Definition: The **correlation coefficient** of X and Y, denoted by $Cor(X, Y)$ or just $\rho_{X,Y}$, is defined by

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

It represents a "scaled" covariance. The correlation is always between -1 and 1.

Two ~~extreme examples~~ special cases:

- What if X and Y are independent?

$Cov(X,Y)=0$ so $\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0$

- What if $Y = aX + b$? $\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

① $Cov(X,Y) = Cov(X, aX+b)$

$= E[(X-E(X))(aX+\cancel{b}-E(aX+\cancel{b}))]$

$= a\,E[(X-E(X))^2] = a\,V(X) = a\sigma_X^2$

② Also $V(Y) = E[(Y-E(Y))^2]$

$= E[(aX+b-E(aX+b))] = a^2 \cdot V(X)$

$\sigma_Y = |a|\sigma_X$

③ $\rho_{X,Y} = \frac{a\,\sigma_X^2}{\sigma_X \cdot |a|\sigma_X} = \frac{a}{|a|} = \begin{cases} 1 & \text{if } a>0 \\ -1 & \text{if } a<0 \end{cases}$

	y (home)		
	0	100	200
x (auto) 100	.20 1/6	.10 1/6	.20 1/6	1/2
250	.05 1/6	.15 1/6	.30 1/6	1/2
	1/3	1/3	1/3	

Find ρ_{XY}

Earlier: $Cov(X,Y) = 1875$, $E(X) = 175$, $E(Y) = 125$

You should verify $V(X) = 75^2$ & $V(Y) = 6875 = \sigma_Y^2$

$$\rho_{X,Y} = \frac{1875}{75\sqrt{6875}} \approx .3$$

Conclusions: Correlation measures the strength of the linear relationship between X & Y. If X & Y are indep, $\rho_{X,Y} = 0$. But if you compute $\rho_{X,Y} = 0$ cannot conclude independence.

In the next module we'll transition to multiple random variables.