Covariance and Correlation

Probability Theory: Foundation for Data Science

with **Anne Dougherty**

Data Science
UNIVERSITY OF COLORADO **BOULDER**

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:

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

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:

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.

Covariance example calculation:

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

Computational formula for covariance:
$Cov(X, Y) = E(XY) - E(X)E(Y)$

What if X and Y are independent?

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

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

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

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

Two special cases:

- What if X and Y are independent?

- What if $Y = aX + b$?

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

Find $\rho_{X,Y}$