Probability Theory

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

Anne Dougherty

March 15, 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 an individual, let's say Bob, is selected at random from the agency's files. Let X be the deductible amount on the auto policy and let Y be the deductible amount on the homeowner's policy.

We want to understand the relationship between X and Y.

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

regular mass fcn for X

$P(X=100)=.5$

$P(X=250)=.5$

1

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

regular mass fcn for Y

$P(X=100, Y=0)=.20$

$P(\underbrace{X=250, Y=100}_{\text{intersection of 2 events } X=250 \text{ and } Y=100})=.15$

-This table gives interaction for X + Y.

In the next video we'll discuss how X + Y are correlated.

But for now, just get used to the probabilities

Definition: Given two discrete random variables, X and Y, $p(x, y) = P(X = x, Y = y)$ is the **joint probability mass function** for X and Y.

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

$P(X = 100) = .5$

Are X + Y indep in this example? $P(Y=100) = .25$

$P(X=100, Y=100) = .1$

$P(X=100)P(Y=100) = (.5)(.25) = .125$

X and Y are not indep

Important property: X and Y are **independent random variables** if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all possible values of x and y.

Similar definition holds for X and Y continuous r.v.

Definition: If X and Y are continuous random variables, then $f(x, y)$ is the **joint probability density function** for X and Y if $P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\, dx\, dy$

Importantant property: X and Y are **independent random variables** if $f(x, y) = f(x)f(y)$ for all possible values of x and y.

Example: Suppose a room is lit with two light bulbs. Let X_1 be the lifetime of the first bulb and X_2 be the lifetime of the second bulb. Suppose $X_1 \sim Exp(\lambda_1 = 1/2000)$ and $X_2 \sim Exp(\lambda_2 = 1/3000)$. If we assume the lifetimes of the light bulbs are independent of each other, find the probability that the room is dark after 4000 hours.

$E(X_1) = \frac{1}{\lambda_1} = 2000$ hrs. and $E(X_2) = \frac{1}{\lambda_2} = 3000$ hrs

Light bulbs fen independently so

$$P(X_1 \leq 4000,\ X_2 \leq 4000) = P(X_1 \leq 4000)\, P(X_2 \leq 4000)$$

$$= \int_0^{4000} \lambda_1 e^{-\lambda_1 x_1} dx_1 \int_0^{4000} \lambda_2 e^{-\lambda x_2} dx_2$$

$$= \left(-e^{-\lambda_1 x_1}\right)\Big|_0^{4000} \cdot \left(-e^{-\lambda x_2}\right)\Big|_0^{4000}$$

$$= \left(1 - e^{-4000/2000}\right)\left(1 - e^{-4000/3000}\right)$$

$$= \left(1 - e^{-2}\right)\left(1 - e^{-4/3}\right) \cong .6368$$

Recap - discussed joint distribution of 2 r.v.
& also what it means for them to be indep.
(just like we had indep events earlier)

Statistical Inference: Soon, we will be focusing on making "statistical inferences" about the true mean and true variance of a population by using sample datasets. We'll return to this in subsequent modules, but for now, $X_1, X_2, \ldots, X_n$ are said to form a **random sample** of size n if

- $X_1, X_2, \ldots, X_n$ are independent
- each random variable has the same distribution

We say that these X_i's are *iid*, independent and identically distributed. ← You'll be hearing more about this in coming lessons.