Probability Theory

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

Anne Dougherty

March 14, 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.

Motivating Examples: In statistics and data science, we frequently collect data from several random variables and we want to understand and quantify the strength of their interactions.

- The length of time a student studies and their score on an exam.
- The relationship between male and female life expectancy in a certain country.
- The relationship between the quantity of two different products purchased by a consumer.

Recall:

- $E(X) = \sum_k kP(X = k)$ if X is discrete
- $E(X) = \int_{-\infty}^{\infty} xf(x)\, dx$ if X is continuous.

What can we say about $E(g(X))$? We've already used this before:

$$E(g(X)) = \begin{cases} \sum_k g(k)\, P(X=k), & X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x)\, f(x)\, dx, & X \text{ continuous} \end{cases}$$

$$E(\underbrace{aX + b}_{g(X)}) = \sum_k (ak+b)\, P(X=k)$$

$$= \underbrace{a\sum_k kP(X=k)}_{E(X)} + b\underbrace{\sum_k P(X=k)}_{1}$$

$$E(aX+b) = aE(X) + b$$

Example: Suppose a university has 15,000 students and let X equal the number of courses for which a randomly selected student is registered. The pmf is

x	1	2	3	4	5	6	7
p(x)	.01	.03	.13	.25	.39	.17	.02

If a student pays \$500 per course plus a \$100 per-semester registration fee, what is the average amount a student pays each semester?

X = # of courses student takes

Want $E(500X+100) = 500E(X)+100$

$E(X) = \sum_{k=1}^{7} k\,P(X=k) = 1(.01)+2(.03)+\cdots+7(.02) = 4.57$

So, $E(500X+100) = 500(4.57)+100 = \$2385$

Recall: $\sigma^2 = V(X) = E[\overbrace{(X-\mu)^2}^{g(X)}]$ ~~and~~ $= E(X^2) - (E(X))^2$

- $V(X) = \sum_k \underbrace{(k-\mu)^2}_{g(k)} P(X=k)$ if X is discrete

- $V(X) = \int_{-\infty}^{\infty} \underbrace{(x-\mu)^2}_{g(x)} f(x)\, dx$ if X is continuous.

What about $V(g(X))$? Think of $g(X)$ as a new r.v.

$$V(g(X)) = \begin{cases} \sum_k \left(g(k) - E(g(X)) \right)^2 P(X=k) \\ \int_{-\infty}^{\infty} [g(x) - E(g(X))]^2 f(x)\, dx \end{cases} = E((g(X))^2) - E(g(X))^2$$

First compute $E(g(X))$

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

Intuition: $E(aX+b) = aE(X)+b$ scales the mean by a & shifts it by b.

For Variance, variance measures the spread of the data, so b has no effect.

Example: Suppose a university has 15,000 students and let X equal the number of courses for which a randomly selected student is registered. The pmf is

x	1	2	3	4	5	6	7
p(x)	.01	.03	.13	.25	.39	.17	.02

If a student pays \$500 per course plus a \$100 per-semester registration fee, what is the average amount a student pays each semester?

We found $E(X) = 4.57$ and $E(500X + 100) = \$2,385$.

$V(X) = E(X^2) - (E(X))^2 = \sum_{k=1}^{7} k^2 P(X=k) - (4.57)^2 = 22.15 - (4.57)^2 = 1.2651$

$V(500X + 100) = 500^2 V(X) = 316,275$

Now, we understand how to compute expected values & variances for fns of r.v.
In the next video, we'll look at what happens

when you have 2 random variables