Lecture 21: Correlation and Covariance

Satyajit Thakor
IIT Mandi

Normal distribution

- Recall: For $X \sim N(0,1)$, $\qquad \phi_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

  and
  $$\Phi_X(x) = \int_{-\infty}^{x} \phi_X(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$
- Evaluation of this integral is not straightforward.
- Hence, numerical approximations of $\Phi_X(x)$ are used in practice.

- To find $P(X \leq x), P(X \geq x)$ or $P(x_1 \leq X \leq x_2)$ for given values x, x_1, x_2, use the table of CDF for $X \sim N(0,1)$:

  `https://www.mathsisfun.com/data/standard-normal-distribution-table.html`
  OR
  `https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf`

Recall: $\Phi_X(-x) = 1 - \Phi_X(x)$ $\Rightarrow$ knowing Φ_X for $x \geq 0$ is sufficient.

Normal distribution

Symmetric about mean $\Rightarrow \Phi_X(\mu) = \Phi_X(0)$
$= .5$

$\Rightarrow \mu$ is the median.

- Example: For $X \sim N(0,1)$, find

$$P(X \leq 0.31), \quad P(X \geq 1.05), \quad P(-1.5 \leq X \leq 1.18).$$

– Using the table for the PDF of an r.v. with standard normal distribution:

$$P(X \leq 0.31) = \Phi_X(0.31) \approx 0.6217$$

$$P(X \geq 1.05) = 1 - \Phi_X(1.05) \approx 1 - .8531$$
$$= .1469$$

$$P(-1.5 \leq X \leq 1.18) = \Phi_X(1.18) - \overbrace{\Phi_X(-1.5)}$$

$1 - \Phi_X(1.5)$
$\approx 1 - .9331$
$= .0668$

$$\approx .8810 - .0668 = 0.8142.$$

Covariance and correlation

- When we consider the joint distribution of two random variables, the means, the medians, and the variances of the variables provide useful information about their marginal distributions.
- However, these values do not provide any information about the dependence between the two variables.
- The strength of the dependence of two random variables on each other is indicated by their covariance, which is defined as

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

- Covariance is a generalization of variance.
- In other words, variance is a special case of covariance:

$$\operatorname{Cov}(X, X) = \operatorname{Var}(X)$$

Covariance and correlation

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

$$\text{Cov}(X,Y) = E[(X-E(X))(Y-E(Y))]$$
$$= E[XY - XE(Y) - YE(X) + E(X)E(Y)]$$
$$\Downarrow \ (\because \text{ linearity of } E(\cdot))$$
$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$
$$= E(XY) - E(X)E(Y)$$

$\left(\because \text{ independence } \Rightarrow E(XY) = E(X)E(Y)\right)$

- Independent r.v.s have a covariance of zero.
- The positive covariance indicates a tendency for high values of one random variable to be associated with high values of the other random variable (we shall see this by example).
- Similarly, the negative covariance indicates a tendency for high values of one random variable to be associated with low values of the other random variable.

Covariance and correlation

- The correlation between two r.v.s X and Y is defined as

$$\boxed{\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}}$$

- $\mathrm{Corr}(X, Y)$ is also denoted as $\rho(X, Y)$ or $\rho_{X,Y}$.
- Independent random variables have a correlation of 0.
- The correlation takes values between -1 and 1. (Why?: proof involves Cauchy–Schwarz inequality - advance topic)
- It is said that:
  X and Y are positively correlated if $\mathrm{Corr}(X, Y) > 0$,
  X and Y are negatively correlated if $\mathrm{Corr}(X, Y) < 0$, and
  X and Y are uncorrelated if $\mathrm{Corr}(X, Y) = 0$.

Covariance and correlation

Example :

- A company services air conditioner units in residences/offices.
- If the random variable X is the service time in hours taken at a particular location, and the random variable Y is the number of air conditioner units at the location, then these two r.v.s can be thought of as jointly distributed.

Y = number of air conditioner units		X = service time (hrs): 1	2	3	4
	1	0.12	0.08	0.07	0.05
	2	0.08	0.15	0.21	0.13
	3	0.01	0.01	0.02	0.07

Covariance and correlation

Homework

- Find the $\underbrace{\text{correlation}}$ between X and Y. → discussion at the end.
- Think: Will it be positive or negative? Why?
- We need to find marginals to compute expectations:

$$P(Y=1) = \sum_{x=1}^{4} P(X=x, Y=1)$$

X = service time (hrs)

Y = number of air conditioner units	1	2	3	4	Marginal distribution of Y
1	0.12	0.08	0.07	0.05	0.32
2	0.08	0.15	0.21	0.13	0.57
3	0.01	0.01	0.02	0.07	0.11
Marginal distribution of X	0.21	0.24	0.30	0.25	

Similarly, compute marginal PMF values from the joint PMF.

$$P(X=1) = \sum_{y=1}^{3} P(X=1, Y=y)$$

Covariance and correlation

- Now we find expectations:

$$E(X) = \sum_{x=1}^{4} x\, P(X=x) = 1(0.21) + 2(0.24) + 3(0.3) + 4(0.25)$$
$$= 2.59 \text{ hours}$$

$$E(Y) = \sum_{y=1}^{3} y\, P(Y=y) = 1(0.32) + 2(0.57) + 3(0.11)$$
$$= 1.79 \text{ units (of AC)}$$

$$E(XY) = \sum_{x=1}^{4} \sum_{y=1}^{3} xy\, P(X=x, Y=y)$$
$$= (1 \cdot 1 \cdot 0.12) + (1 \cdot 2 \cdot 0.08) + \cdots + (4 \cdot 3 \cdot 0.07)$$
$$= 4.86.$$

Covariance and correlation

- Now we find the covariance:

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$
$$= 4.86 - (2.59 \cdot 1.79)$$
$$= 0.224.$$

- The covariance is positive since there is a tendency for locations with a large number of air conditioner units to require relatively long service times. This can also be observed in the joint PMF table.

For example, $\underbrace{P(X=1, Y=3)}_{0.01} < \underbrace{P(X=4, Y=3)}_{0.07}$