COMPSCI 240: Reasoning Under Uncertainty

Nic Herndon and Andrew Lan

University of Massachusetts at Amherst

Spring 2019

Lecture 17: Correlation and Causation

Quantifying Dependence: Correlation

- The range of $\text{cov}(X, Y)$ values depends on the means of X, Y, and XY.
- The **correlation** ρ between X and Y is closely related to the covariance, but is *normalized* to the range $[-1, 1]$:

$$\rho(X, Y) = corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

- $\rho = 1$ indicates maximum positive covariance (e.g., $\rho(X, X)$) and $\rho = -1$ indicates maximum negative covariance (e.g., $\rho(X, -X)$).

Quantifying Dependence: Correlation

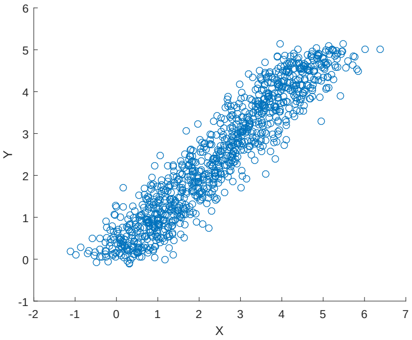
P(X,Y)		
X\Y	$Y = 0$	$Y = 1$
$X = 0$	0.4	0.1
$X = 1$	0.2	0.3

- We computed that $cov(X, Y) = 0.1$
- $P(X = 0) = 0.5, P(X = 1) = 0.5$ and so $var(X) = E[X^2] - E[X]^2 = 0.5 - 0.25 = 0.25$
- $P(Y = 0) = 0.6, P(Y = 1) = 0.4$ and so $var(Y) = E[Y^2] - E[Y]^2 = 0.4 - 0.16 = 0.24$
- Then,

$$\rho(X, Y) = \frac{0.1}{\sqrt{0.25 \times 0.24}} = 0.41$$

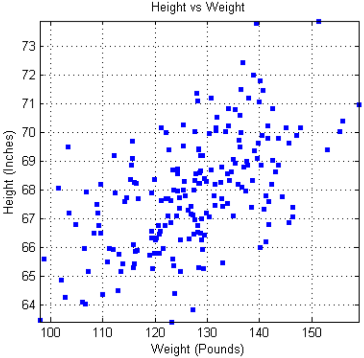
Quantifying Dependence: Correlation

- The computed (empirical) correlation was $\rho = 0.95$.

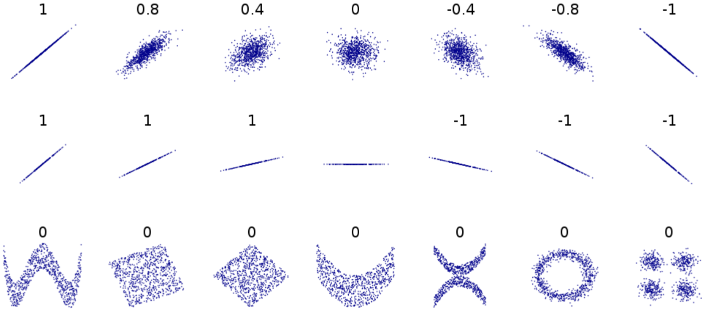


Visualizing Correlations: Height vs Weight

- The computed (empirical) correlation was $\rho = 0.56$.



Visualizing Correlations: Linear vs Non-Linear



Example

- Let X and Y be discrete random variables with the following joint PMF:

$$P_{X,Y}(x,y) = \frac{1}{4}, \text{for all } (x,y) \in \{(0,0),(1,1),(1,-1),(2,0)\}$$

- What is the covariance and correlation of X and Y?

$$Cov(X,Y) = E[X,Y] - E[X]E[Y] = 0 - (1 \times 0) = 0$$

$$\rho(X,Y) = 0$$

- Are X and Y independent?
- No, since $P_{X,Y}(X,Y) \neq P_X(x)P_Y(y)$.

Example

- Let X and Y be continuous random variables with the following joint PDF:

$$f_{X,Y}(x, y) = 3x, 0 \leq y \leq x \leq 1$$

- What is the covariance and correlation of X and Y?

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

- To compute $E[X]$ and $E[Y]$, we need to compute the marginal PDF of X and Y.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \int_0^x 3xdy = 3x^2, 0 \leq x \leq 1$$

Then,

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx = \int_0^1 3x^3 dx = \frac{3}{4}$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \int_y^1 3xdx = \frac{3}{2}(1 - y^2), 0 \leq y \leq 1$$

Then,

$$E[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_0^1 \frac{3}{2}(y - y^3)dy = \left(\frac{3}{4}y^2 - \frac{3}{8}y^4\right)\Big|_0^1 = \frac{3}{8}$$

Example

- $E[X, Y]$ can be computed as

$$
\begin{aligned}
E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dydx = \int_0^1\int_0^x 3x^2ydydx \\
&= \int_0^1 \frac{3}{2}x^2y^2\Big|_0^x dx = \int_0^1 \frac{3}{2}x^4dx \\
&= \frac{3}{10}y^5\Big|_0^1 = \frac{3}{10}
\end{aligned}
$$

- Then, $Cov(X, Y)$ is

$$
Cov(X, Y) = E[X, Y] - E[X]E[Y] = \frac{3}{10} - \frac{3}{4}\times\frac{3}{8} = \frac{3}{160}
$$

- The correlation $\rho(X, Y)$ is

$$
\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{\frac{3}{160}}{\frac{3}{80}\times\frac{19}{320}} = 0.397
$$

Causation

- **Question:** When two random variables are correlated does this mean one random variable causes the other?
- **Example:** In the height/weight example, height and weight were positively correlated. Does increasing your weight make you taller?
- **Example:** There are more fireman at the scene of larger fires. Do fireman cause an increase in the size of a fire?
- **Example:** More people drown on days where a lot of ice cream is sold. Does ice cream cause drowning?

Causation

Given two correlated random variables X and Y:

- X might cause Y (i.e., causation)
- Y might cause X (i.e., reverse causation)
- A third random variable Z might cause X and Y (i.e., common cause)
- A combination of all of these (e.g., self-reinforcement)
- The correlation might be spurious due to small sample size