Probability and Statistics

Alvin Lin

Probability and Statistics: January 2017 - May 2017

Expected Values and Covariance

Let X and Y be discrete random variables with joint pmf $p(x, y)$. The expected value of $h(X, Y)$ is:

$$\begin{aligned} E\Big[h(X,Y)\Big] &= \sum_x \sum_y h(x,y)p(x,y) \\ &= \sum_x \sum_y h(x,y)P(X = x \text{ and } Y = y) \\ \mu_x &= \sum_x x p_x(x) \\ \mu_y &= \sum_y x p_y(y) \\ &= \sum_y y P(Y = y) \end{aligned}$$

$p_x(x)$ and $p_y(y)$ are marginal probability mass functions of X and Y, respectively. Recall the formal definitions:

$$\begin{aligned}
p_x(x) &= \sum_y p(x,y) \\
p_y(y) &= \sum_x p(x,y) \\
(\sigma_x)^2 &= \sum_x (x-\mu_x)^2 p_x(x) \\
&= \sum_x (x-\mu_x)^2 P(X=x) \\
(\sigma_y)^2 &= \sum_y (y-\mu_y)^2 p_y(y) \\
\sigma_x &= \sqrt{(\sigma_x)^2} \\
\sigma_y &= \sqrt{(\sigma_y)^2}
\end{aligned}$$

The **covariance** between the jointly distributed random variables X and Y is:

$$\begin{aligned}
Cov(X,Y) &= E\Big[(x-\mu_x)(y-\mu_y)\Big] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y) p(x,y)
\end{aligned}$$

Let X and Y be jointly distributed random variables.

$$Cov(X,Y) = E(XY) - \mu_x \mu_y$$

where:

$$E(XY) = \sum_x \sum_y xy \; p(x,y)$$

Correlation Coefficient

The correlation coefficient of X and Y, denoted by $Corr(X,Y)$, $\rho_{x,y}$, or just ρ, is:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

If the variables a and c are both positive or both negative, then:

$$Corr(aX+b, cY+d) = Corr(X,Y)$$

$$-1 \leq Corr(X, Y) \leq 1$$

If X and Y with joint pmf $p(x, y)$ and joint pdf $f(x, y)$ are independent, the $\rho = 0$. If $\rho = 0$, then the random variables X and Y may or may not be independent. $\rho = \pm 1$ if and only if $Y = aX + b$ for some numbers a and b with $a \neq 0$.

Uses of the Correlation Coefficient

Let X be the random variable for the height of a randomly selected person at RIT and let Y be the random variable for their weight. $Corr(X, Y)$ is not likely to be 1 or -1 because there is no strong correlation between height and weight.

Let X be the average temperature of a randomly selected city in degrees Centigrade, with Y being the average temperature expressed in degrees Fahrenheit. The correlation coefficient in this case is 1.

Example

Randomly select a card from a file containing three cards:

- (Name, Readiness Assessment Score, Common Core Score)
- (James, 58, 81)
- (Mike, 73, 78)
- (Jane, 76, 85)

Let X be the random variable for the readiness assessment score and Y be the random variable for the Common Core score. Recall that a random variable is a function. Find the correlation coefficient of X and Y.

x	y	$p(x, y)$	$p_x(x)$	$p_y(y)$	$p_x(x)p_y(y)$
58	81	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{9}$
73	78	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{9}$
76	85	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{9}$

Are X and Y independent?
No, $p(x,y) \neq p_x(x)p_y(y)$.

$$\begin{aligned}
\mu_x &= \sum_x x p_x(x) \\
&= 58 \cdot p_x(58) + 73 \cdot p_x(73) + 76 \cdot p_x(76) \\
&= 58 \cdot \frac{1}{3} + 73 \cdot \frac{1}{3} + 76 \cdot \frac{1}{3} = 69 \\
\mu_y &= \sum_y p_y(y) \\
&= 81 \cdot \frac{1}{3} + 78 \cdot \frac{1}{3} + 85 \cdot \frac{1}{3} = \frac{244}{3} \\
(\sigma_x)^2 &= \sum_x (x-\mu_x)^2 p_x(x) \\
&= (58-69)^2 \cdot \frac{1}{3} + (73-69)^2 \cdot \frac{1}{3} + (76-69)^2 \cdot \frac{1}{3} \\
&= \frac{1}{3}\left[11^2 + 4^2 + 7^2\right] \\
&= 62 \\
(\sigma_y)^2 &= \sum_y (y-\mu_y)^2 p_y(y) \\
&= (81 - \frac{244}{9})^2 \cdot \frac{1}{3} + (78 - \frac{244}{9})^2 \cdot \frac{1}{3} + (85 - \frac{244}{9})^2 \cdot \frac{1}{3} \\
&= \frac{1}{3}\left[(\frac{1}{3})^2 + (-\frac{10}{3})^2 + (\frac{11}{3})^2\right] \\
&= \frac{74}{9}
\end{aligned}$$

$$
\begin{aligned}
Cov(X,Y) &= E\Big[(X-\mu_x)(Y-\mu_y)\Big] \\
&= E(XY) - \mu_x\mu_y \\
&= \Big[\sum_x \sum_y xy\ p(x,y)\Big] - \mu_x\mu_y \\
&= 58 \cdot 81 \cdot \frac{1}{3} + 73 \cdot 78 \cdot \frac{1}{3} + 76 \cdot 85 \cdot \frac{1}{3} - 69 \cdot (81 + \frac{1}{3}) \\
&= 5.33 \\
Corr(X,Y) &= \frac{5.33}{\sqrt{62}\sqrt{8.22}} = 0.236
\end{aligned}
$$

The positive correlation indicates that we can reasonable expect there is a correlation between the Readiness Assessment score and the Common Core score. The correlation is not perfectly linear however.

Example

A surveyor wishes to lay out a square region with each side having length L. However, because of measurement error, he instead lays out a rectangle in which both the north-south sides are of length X and the east-west sides are both of length Y. Suppose X and Y are independent and that each is uniformly distributed on the interval $[L-A, L+A]$ where $0 < A < L$. What is the expected area of the resulting rectangle? Find the joint pdf of X and Y.

$$
f(x,y) = \begin{cases} h & , L-A \le x \le L+A \text{ and } L-A \le y \le L+A \\ 0 & , otherwise \end{cases}
$$

How do we find the value of h?

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{L-A}^{L+A}\int_{L-A}^{L+A} h\mathrm{d}x\mathrm{d}y \\
&= h\int_{L-A}^{L+A}\int_{L-A}^{L+A} \mathrm{d}x\mathrm{d}y \\
&= h(2A)(2A) \\
h &= \frac{1}{4A^2}
\end{aligned}
$$

The expected value of the area of the rectange is $E(XY)$.

$$\begin{aligned}
E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ p(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{L-A}^{L+A}\int_{L-A}^{L+A} xyh\mathrm{d}x\mathrm{d}y \\
&= h\int_{L-A}^{L+A} y\left[\int_{L-A}^{L+A} x\mathrm{d}x\right]\mathrm{d}y \\
&= h\left[\int_{L-A}^{L+A} x\mathrm{d}x\right]\left[\int_{L-A}^{L+A} y\mathrm{d}y\right] \\
&= h\left[\int_{L-A}^{L+A} x\mathrm{d}x\right]^2 \\
&= h\left[\frac{1}{2}(x^2)_{x=L-A}^{x=L+A}\right] \\
&= h\frac{1}{2}\left[(L+A)^2-(L-A)^2\right]^2 \\
&= h\frac{1}{2}\left[((L+A)+(L-A))((L+A)-(L-A))\right]^2 \\
&= h\frac{1}{2}\left[2L\cdot 2A\right]^2 \\
&= L^2
\end{aligned}$$

You can find all my notes at `http://omgimanerd.tech/notes`. If you have any questions, comments, or concerns, please contact me at alvin@omgimanerd.tech