Statistics 301: Probability and Statistics

Joint Distributions

Module 6

2018

Two Discrete Random Variables

The probability mass function (pmf) of a single discrete rv X specifies how much probability mass is placed on each possible value of X. The joint pmf of two discrete RVs X and Y describes how much probability mass is placed on each possible pair of values (x, y).

Definition

Let X and Y be two discrete RVs defined on the sample space $\mathcal{S}$ of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \textit{ and } Y = y)$$

It must be the case that $p(x, y) \geq 0$ and $\Sigma_x \Sigma_y p(x, y) = 1$

Discrete Distribution Example

A large insurance agency services a large number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible is specified; the auto poilcy has deductibles of \$100 or \$250, whereas a homeowner's policy has deductibles of \$0, \$100 or \$200. Let $X =$ the deductible amount on the auto policy and Let $Y =$ the deductible amount on the homeowner's policy. The next slide contains the table distribution.

Find: $P(X = 100 \textit{ and } Y = 100) = p(100, 100)$
$P(Y \geq 100)$

Discrete Distribution Example Data

			y	
	$p(x, y)$	0	100	200
x	100	0.20	0.10	0.20
	250	0.05	0.15	0.30

Discrete Example: Probabilities

$P(X = 100 \textit{ and } Y = 100) = p(100, 100) = 0.10$

$P(Y \geq 100) = p(100, 100) + p(100, 200) + p(250, 100) + p(250, 200) = 0.1 + 0.2 + 0.15 + 0.3 = 0.75$
OR (complement rule)
$1 - P(Y < 100) = 1 - P(Y = 0) = 1 - [p(100, 0) + p(250, 0)] = 1 - (0.2 + 0.05) = 1 - 0.25 = 0.75$

Discrete Marginal Distributions (marginal pmfs)

The **marginal probability mass function** of X, denoted by $p_X(x)$, is given by

$$p_X(x) = \sum_y p(x,y) \ \forall \ x$$

Similarly, the **marginal probability mass function** of Y, denoted by $p_Y(y)$, is given by

$$p_Y(y) = \sum_x p(x,y) \ \forall \ y$$

Discrete Example: Marginal Distributions of X and Y

$p_X(100) = \sum_y p(x,y) = p(100,0) + p(100,100) + p(100,200) = 0.5$

$p_X(250) = \sum_y p(x,y) = p(250,0) + p(250,100) + p(250,200) = 0.5$

$p_Y(0) = \sum_x p(x,y) = p(100,0) + p(250,0) = 0.25$

$p_Y(100) = \sum_x p(x,y) = p(100,100) + p(250,100) = 0.25$

$p_Y(200) = \sum_x p(x,y) = p(100,200) + p(250,200) = 0.5$

Discrete Marginal Distributions

$$p_X(x) = \begin{cases} 0.5 & x = 100 \\ 0.5 & x = 250 \\ 0 & otherwise \end{cases}$$

$$p_Y(y) = \begin{cases} 0.25 & y = 0 \\ 0.25 & y = 100 \\ 0.5 & y = 200 \\ 0 & otherwise \end{cases}$$

Independence of X and Y

Two random variables X and Y are **independent** if for every pair of x and y values,

$$p(x,y) = p_X(x) \cdot p_Y(y) \quad when\ X\ and\ Y\ are\ discrete$$

Or

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad when\ X\ and\ Y\ are\ continuous$$

If the above are not satisfied for all (x,y)*, then all* X *and* Y *are said to be dependent*

Discrete Example: Independence

Are X and Y independent?

$$? \ p(100, 100) = p_X(100) \cdot p_Y(100) \ ?$$

$$\Rightarrow 0.1 \neq (0.5)(0.25)$$

No, they are not independent

Continuous Distribution Example: Independence

Are X and Y independent?

$$? \ f(x, y) = f_X(x) \cdot f_Y(y) \ ?$$

Try $f(1, 1)$

$$? \ f(1, 1) = f_X(1) \cdot f_Y(1) \ ?$$

$$f(1, 1) = 0.35; \ f_X(1) = 0.35; \ f_Y(1) = 1.00$$

$$\Rightarrow 0.35 = (0.35)(1)$$

So, X and Y are independent

Joint Conditional Probabilities

Recall the formula for conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The same follows for discrete and continuous distributions:

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}$$

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Discrete Example: Conditional Probabilities

$$p_{Y|X}(Y = 200|X = 100) = \frac{p(100, 200)}{p_X(100)}$$

$$= \frac{0.2}{0.5} = \frac{2}{5}$$

Continuous Example: Conditional Probabilities

$$f_{Y|X}(Y > 0.5|X < 1) = \frac{f(X < 1, Y > 0.5)}{f_X(X < 1)}$$

$$f(X < 1, Y > 0.5) = \int_0^1 \int_{0.5}^1 f(x, y)\ dydx$$

$= \int_0^1 \int_{0.5}^1 \left(\frac{9xy^2}{10} + \frac{1}{5}\right)\ dydx = \cdots = 0.23125$

$$f_X(x < 1) = \int_0^1 f_X(x)\ dx$$

$= \int_0^1 \left(\frac{3x}{10} + \frac{1}{5}\right)\ dx = \cdots = 0.35$

$$f_{Y|X}(Y > 0.5|X < 1) = \frac{0.23125}{0.35} = 0.08094$$

Expected Values, Covariance, Correlation

Expected values (EX, EY), variances (VX, VY), and standard deviations (SDX, SDY) are calculated as learned previously, in addition to updated rules of expectation.

Covariance Definition

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another. The **covariance** between two RVs X and Y is:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = EXY - (EX)(EY)$$

For discrete RVs:

$$\sum_x \sum_y (x - EX)(y - EY)p(x, y)$$

For continuous RVs:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - EX)(y - EY)f(x, y)\ dxdy$$

Covariance Properties

Covariance

Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated. However, it is also often used informally as a general measure of how monotonically related two variables are.

The major defect in covariance is that although it is a measure of linear dependence, its computed value depends critically on the units of measurement. However, if we standardize the covariance (by dividing it by standard deviations), we get a better measure of linear dependence, called correlation.

If X and Y are independent, the covariance of X and Y ($Cov(X, Y) = 0$), but it does not hold in reverse. Just because the covariance is 0 does not mean independence; it could mean they are not *linearally* related.

Covariance Formulas

$Cov(X,Y) = EXY - (EX)(EY)$ where

For discrete RVs:

$$EXY = \sum (xyp(x,y))$$

For continuous RVs:

$$EXY = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\ dxdy$$

Discrete Example: Covariance

All products that equal 0 will not be shown in calculation

$$EXY = (100)(100)(.1) + (100)(200)(.2) + (250)(100)(.15)$$

$$+(250)(200)(.3) = 23750$$

$$Cov(X,Y) = 23750 - (175)(125) = 1875$$

Correlation

This is the standardized version of covariance. Correlation refers to the extent to which two variables have a linear relationship with each other. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.

Properties of Correlation

- describes the *linear* relationship between two *quantitative* variables X and Y
- $-1 \leq \rho \leq 1$
- There are no units of measurement associated with ρ (and will not change if units of measurement are changed)
- Makes no distinction between X and Y

Warning!

Correlation is often used in misleading and incorrect ways. The main thing to remember with correlation is that it implies only that there is an association; it does *not* mean that X causes Y. The only way to determine causation is with experimentation.

Formulas

For both continuous and discrete RVs:

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{(SDX)(SDY)}$$

The sample correlation is usually referred to as r

Discrete Example: Covariance

$Cov(X,Y) = 1875$, $SDX = 75$, $SDY = 82.9156$

$$\rho_{XY} = \frac{Cov(X,Y)}{(SDX)(SDY)} = \frac{1875}{(75)(82.9156)} = 0.301511$$

$\rho_{XY} = 0.3015$, which is close to 0 and positive, indicating that there is a weak, positive linear relationship between X (auto insurance) and Y (home insurance). Generally, more people that have auto insurance will also have home insurance through the same company (or at least in this company).

Continuous Joint Distributions

Let X and Y be continuous RVs. A **joint probability density function** $f(x,y)$ for these two variables is a function satisfying $f(x,y) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ dxdy = 1$$

Then for any two dimensional set A

$$P[(X,Y) \in A] = \iint_A f(x,y)\ dxdy$$

Continuous Joint Distributions (con't)

In particular, if A is the two-dimensional rectangle $(x,y) : a \leq x \leq b, c \leq y \leq d$, then

$$P[(X,Y) \in A] =$$

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ dxdy$$

Continuous Distribution Example

A college professor wants to learn if there is a relationship between time spend on homework and the percent of the homeowrk that is completed. Let X = the number of weeks after being distributed that an assignment is turned in and Y = percent of completed assignment. Suppose X, Y have the following joint pdf:

$$f(x,y) = \begin{cases} \frac{9}{10}xy^2 + \frac{1}{5} & 0 \leq x \leq 2,\ 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

Continuous Example: Probabilities

The probability that a randomly selected student will turn in an assignment in less than one week with more than half of the assignment completed. That is, find $P(X < 1, y > 0.5)$

$P(X < 1, Y > 0.5) = \int_0^1 \int_{0.5}^1 \frac{9}{10}xy^2 + \frac{1}{5}\ dydx$

$= \int_0^1 \left[\frac{3}{10}xy^3 + \frac{1}{5}y\right]_{0.5}^1 \; dx$

$= \int_0^1 \left(\frac{21}{80}x + \frac{1}{10}x\right) \; dx$

$= \left(\frac{21}{160}x^2 + \frac{1}{10}x\right)|_0^1 = 0.23125$

Continuous Marginal Distributions (marginal pdfs)

The **marginal probability density function** of X, denoted by $f_X(x)$, is given by

$$f_X(x) = \int_y f(x,y) \; dy$$

Similarly, the **marginal probability density function** of Y, denoted by $f_Y(y)$, is given by

$$f_Y(y) = \int_x f(x,y) \; dx$$

Continuous Example: Marginal Distribution of X

$f_X(x) = \int_y f(x,y) \; dy = f_X(x) = \int_0^1 \frac{9}{10}xy^2 + \frac{1}{5} \; dy$

$= \left[\frac{3xy^3}{10} + \frac{y}{5}\right]_0^1 = \frac{3x}{10} + \frac{1}{5}$

$$f_X(x) = \begin{cases} \frac{3x}{10} + \frac{1}{5} & 0 \leq x \leq 2 \\ 0 & otherwise \end{cases}$$

Continuous Marginal Distribution of Y

$f_Y(y) = \int_x f(x,y) \; dx = f_Y(y) = \int_0^2 \frac{9}{10}xy^2 + \frac{1}{5} \; dx$

$= \left[\frac{9x^2y^2}{20} + \frac{x}{5}\right]_0^2 = \frac{9y^2}{5} + \frac{2}{5}$

$$f_Y(y) = \begin{cases} \frac{9y^2}{5} + \frac{2}{5} & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

Continuous Covariance

$$EXY = \int_0^2 \int_0^1 xy\left(\frac{9}{10}xy^2 + \frac{1}{5}\right) \; dydx$$

$= \int_0^2 \left[\frac{9x^2y^3}{40} + \frac{xy^2}{10}\right]_0^1 \; dx$

$= \int_0^2 \left(\frac{9x^2}{40} + \frac{x}{10}\right) \; dx$

$= \left[\frac{9x^3}{120} + \frac{x^2}{20}\right]_0^2 = \frac{4}{5}$

$Cov(X,Y) = \frac{4}{5} - (\frac{6}{5})(\frac{13}{20}) = 0.02$

Continuous Correlation

$Cov(X,Y) = 0.02$, $SDX = 0.5416$, $SDY = 0.2661$

$$\rho_{XY} = \frac{Cov(X,Y)}{(SDX)(SDY)} = \frac{0.02}{(0.5416)(0.2661)} = 0.138773$$

$\rho_{XY} = 0.1388$, which is close to 0 and positive, indicating that there is a weak, positive linear relationship between X and Y. Generally, papers will be more complete as the time spent on them increases.

Rules of Expectation

Adding, subtracting, or multiplying RV by a constant

$$E(X \pm a) = E(X) \pm E(a) = E(X) \pm a$$
$$V(X \pm a) = V(X) + 0$$

$$E(aX) = aE(X)$$
$$V(aX) = a^2 V(X)$$

Rules of Expectation: Independent RVs

When X and Y are independent RVs:

$$E(X \pm Y) = E(X) \pm E(Y)$$
$$V(X \pm Y) = V(X) + V(Y)$$
$$SD(X \pm Y) = \sqrt{V(X) + V(Y)}$$

Rules involving constants still hold and can be applied

Rules of Expectation: Dependent RVs

When X and Y are dependent RVs:

$$E(X + Y) = E(X) + E(Y)$$
$$V(X + Y) = V(X) + V(Y) + 2Cov(X,Y)$$
$$SD(X + Y) = \sqrt{V(X) + V(Y) + 2Cov(X,Y)}$$

$$E(X - Y) = E(X) - E(Y)$$
$$V(X - Y) = V(X) + V(Y) - 2Cov(X,Y)$$
$$SD(X - Y) = \sqrt{V(X) + V(Y) - 2Cov(X,Y)}$$

Rules involving constants still hold and can be applied