Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

IPDA1005 INTRODUCTION TO PROBABILITY AND DATA ANALYSIS

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Outline

Functions of Random Variables

- Consider a continuous RV X with CDF F. Define a new RV U as $U = F(X)$. Then U has $U(0,1)$ distribution.
- **Proof:** Note that since $0 \leq F(x) \leq 1$, U takes values between 0 and 1. For $u \in (0,1)$,

$$\begin{aligned} P(U \leq u) &= P\left(F(X) \leq u\right) \\ &= P\left(X \leq F^{-1}(u)\right) \\ &= F\left(F^{-1}(u)\right) \\ &= u \end{aligned}$$

- Thus $F_U(u) = u$ for $u \in (0,1)$, showing that U has $U(0,1)$ distribution.

Random Number Generation - Continuous distributions

- We can go in the opposite direction of this idea. If we start from $U \sim U(0,1)$ and calculate $X = F^{-1}(U)$ then we obtain random observations of X.
- Thus we can generate random observations from a continuous distribution with CDF F if:
  - we can generate random numbers from $U(0,1)$, and
  - we can calculate the inverse distribution function F^{-1}.
- **Example:** If $u_1, u_2, \ldots$ is a random sample from $U(0,1)$, use it to generate a random sample from a distribution whose CDF is given by

$$F(x) = \frac{\sqrt{x}}{2}, 0 \le x \le 4$$

- $g(u) = F^{-1}(u) = 4u^2$, so if we apply the function $g(u)$ to the $U(0,1)$ sample to get $g(u_1), g(u_2), \ldots$ we have a random sample from the required distribution.

Random Number Generation - Discrete distributions

- To generate random numbers from a discrete distribution we treat the CDF as a "look-up" table.
- Suppose X is a discrete RV with CDF F, and possible values $x_1, x_2, \ldots x_N$.
- We generate random observations of X by:
  1. Generate random number u from $U(0,1)$
  2. Find x_j where $F(x_{j-1}) < u \leq F(x_j)$, for j in 1, 2, ..., N
- Repeating this n times, we will have a sample of size n from a discrete distribution with CDF F.

Jointly Distributed Discrete Random Variables

- In many situations we need to consider two or more random variables defined together on a single sample space. This gives us insight into -
  - Probability of joint occurrence
  - Dependence structure, covariance, correlation
  - Mean and variance of functions related to the joint distribution.

  Examples:
  - Reliability: failure-time distribution of two or more parts of a machine
  - Financial modelling: distribution of two or more stocks traded in a stock exchange
  - Weather forecasting: joint distribution of climatological variables, e.g., temperature and humidity

Jointly Distributed Discrete Random Variables

- In this example, X and Y are dependent while X and Z are independent.
- Similarly, Y and Z are independent.
- When two random variables are independent, we can calculate the joint probabilities by multiplying their individual probabilities.

$$P(X = 2, Z = 3) = \binom{10}{2} p_1^2 (1 - p_1)^8 \binom{20}{3} p_2^3 (1 - p_2)^{17}.$$

- On the other hand, for X and Y,

  $P(X = 3, Y = 4) = 0$ because if $X = 3$, Y has to be 7.

  $P(X = 3, Y = 7) = P(X = 3) = \binom{10}{3} p_1^3 (1 - p_1)^7.$

Jointly Distributed Discrete Random Variables

- Joint distribution of X and Y can be written in the form of a table.
- But first, we make a preliminary table of possible values the dice come up with and the corresponding X and Y. The ordered pair in each cell is (X, Y).

Table: Outcomes for the Dice

	1	2	3	4	5	6
1	$(2, 0)$	$(3, 1)$	$(4, 2)$	$(5, 3)$	$(6, 4)$	$(7, 5)$
2	$(3, 1)$	$(4, 0)$	$(5, 1)$	$(6, 2)$	$(7, 3)$	$(8, 4)$
3	$(4, 2)$	$(5, 1)$	$(6, 0)$	$(7, 1)$	$(8, 2)$	$(9, 3)$
4	$(5, 3)$	$(6, 2)$	$(7, 1)$	$(8, 0)$	$(9, 1)$	$(10, 2)$
5	$(6, 4)$	$(7, 3)$	$(8, 2)$	$(9, 1)$	$(10, 0)$	$(11, 1)$
6	$(7, 5)$	$(8, 4)$	$(9, 3)$	$(10, 2)$	$(11, 1)$	$(12, 0)$

Joint and Marginal PMF

- The *joint pmf* of two discrete RVs X and Y is denoted by p_{XY} where $p_{XY}(x, y) = P(X = x, Y = y)$.
- $p_{XY}(x, y)$ is sometimes written as $p(x, y)$ for short.
- The joint PMF satisfies the property $\sum_x \sum_y p(x, y) = 1$.
- If A is a set of pairs of (x, y) values, then the probability that the RV pair (X, Y) lies in A is the sum of the joint PMF over all pairs in A.

$$P\left[(X, Y) \in A\right] = \sum_{(x,y)\in A} \sum p(x, y)$$

- The *marginal pmfs* of X and Y are given by

$$p_X(x) = \sum_y p_{XY}(x, y), \qquad p_Y(y) = \sum_x p_{XY}(x, y)$$

Joint CDF

The *joint cumulative distribution function* of the discrete random variables X and Y is defined by

$$F(x, y) = P(X \le x, Y \le y) = \sum_{u \le x, v \le y} p(u, v)$$

Example: Consider the dice example.

1. Construct the joint CDF table.
2. Find $P(X \le 4, Y \le 2)$
3. Find $P(X + Y \le 6)$

Conditional Distribution

- If $P(Y = y) > 0$, the *conditional distribution* of X given $Y = y$ is

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

- For example, given $Y = 2$,

$$p_{X|Y}(x|2) = P(X = x|Y = 2) = \frac{P(X = x, Y = 2)}{P(Y = 2)}$$

$$= \frac{P(X = x, Y = 2)}{\frac{2}{9}}$$

- The conditional distribution of x at $Y = 2$ is then -

x	4	6	8	10
$p_{X\|Y}(x\|2)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Jointly Distributed Discrete Random Variables

- **Solution:** The number of different ways of selecting 2 books from the 8 books is $\binom{8}{2}$.
- So

$$P(X = x, Y = y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}}$$

for $0 \le x \le 2, 0 \le y \le 2, x + y \le 2$.

- From this we construct the joint probability distribution table:

		X		
		0	1	2
Y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$
	1	$\frac{6}{28}$	$\frac{6}{28}$	0
	2	$\frac{1}{28}$	0	0

$P(X = 2, Y = 2) = 0$
$P(X = 2)P(Y = 2) = \frac{3}{784} \neq 0.$
So X and Y are dependent.

Covariance and Correlation

The covariance between X and Y is given by

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

The correlation between X and Y is given by

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Example: Suppose X and Y have the following joint distribution:

		X		
		-1	0	1
Y	-1	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{9}$
	0	0	$\frac{1}{9}$	0
	1	$\frac{2}{9}$	0	$\frac{1}{9}$

Find the covariance and correlation between X and Y.

Covariance and Correlation

- The marginal distribution of Y is given by

x	-1	0	1
$p_Y(y)$	$\frac{5}{9}$	$\frac{1}{9}$	$\frac{1}{3}$

  from which it follows that

$$E(Y) = \frac{-2}{9}, E(Y^2) = \frac{8}{9}, Var(Y) = \frac{68}{81}, \mathrm{SD}(Y) = \frac{\sqrt{68}}{9}.$$

- The covariance between X and Y is given by

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = -\frac{1}{9} - \left(-\frac{1}{9}\right)\left(-\frac{2}{9}\right) = \frac{-11}{81}$$

  and hence

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \frac{\frac{-11}{81}}{\frac{\sqrt{44}}{9}\frac{\sqrt{68}}{9}} = \frac{-11}{\sqrt{44}\sqrt{68}} = -0.201.$$

Covariance and Correlation

- From the marginal distributions, $E(X) = 7$ and $E(Y) = \frac{35}{18}$.
- So $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{245}{18} - 7\left(\frac{35}{18}\right) = 0$.
- Hence the correlation between X and Y is also zero.

- Independent RVs will always have zero covariance and correlation.
- This example shows that the converse is not true.

Properties of Covariance and Correlation

8. $-1 \leq \text{Corr}(X, Y) \leq 1$
9. $$\begin{aligned} \text{Corr}(aX + b, cY + d) &= \text{Corr}(X, Y) \text{ if } ac > 0 \\ &= -\text{Corr}(X, Y) \text{ if } ac < 0 \\ &= 0 \text{ if } ac = 0 \end{aligned}$$
10. If X and Y are independent,
    a. $\text{Cov}(X, Y) = 0$
    b. $\text{Corr}(X, Y) = 0$
    c. $Var(X + Y) = Var(X) + Var(Y)$

Note: The converse of 10(a) above is false. Independence implies that covariance is zero, but two dependent random variables can have zero covariance, as in the dice example.

Properties of Covariance and Correlation

Example: X and Y are two RVs with $Var(X) = 4$, $Var(Y) = 9$, $\mathrm{Cov}(X, Y) = -2$. Find -

1. $\mathrm{Cov}(2X + 3, -4Y + 2)$
2. $\mathrm{Cov}(2X - 1, 4 - X)$
3. $Var(2X - Y)$
4. $\mathrm{Corr}(X, Y)$
5. $\mathrm{Corr}(3X + 1, -2Y + 2)$

Solution:

1. $\mathrm{Cov}(2X + 3, -4Y + 2) = 2(-4)\mathrm{Cov}(X, Y) = 16$
2. $\mathrm{Cov}(2X - 1, 4 - X) = 2(-1)\mathrm{Cov}(X, X) = (-2)Var(X) = -8$
3. $Var(2X - Y) = 4Var(X) + Var(Y) - 4\mathrm{Cov}(X, Y) = 16 + 9 + 8 = 33$
4. $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \frac{-2}{(2)(3)} = \frac{-1}{3}$
5. $\mathrm{Corr}(3X + 1, -2Y + 2) = (-1)\frac{-1}{3} = \frac{1}{3}$

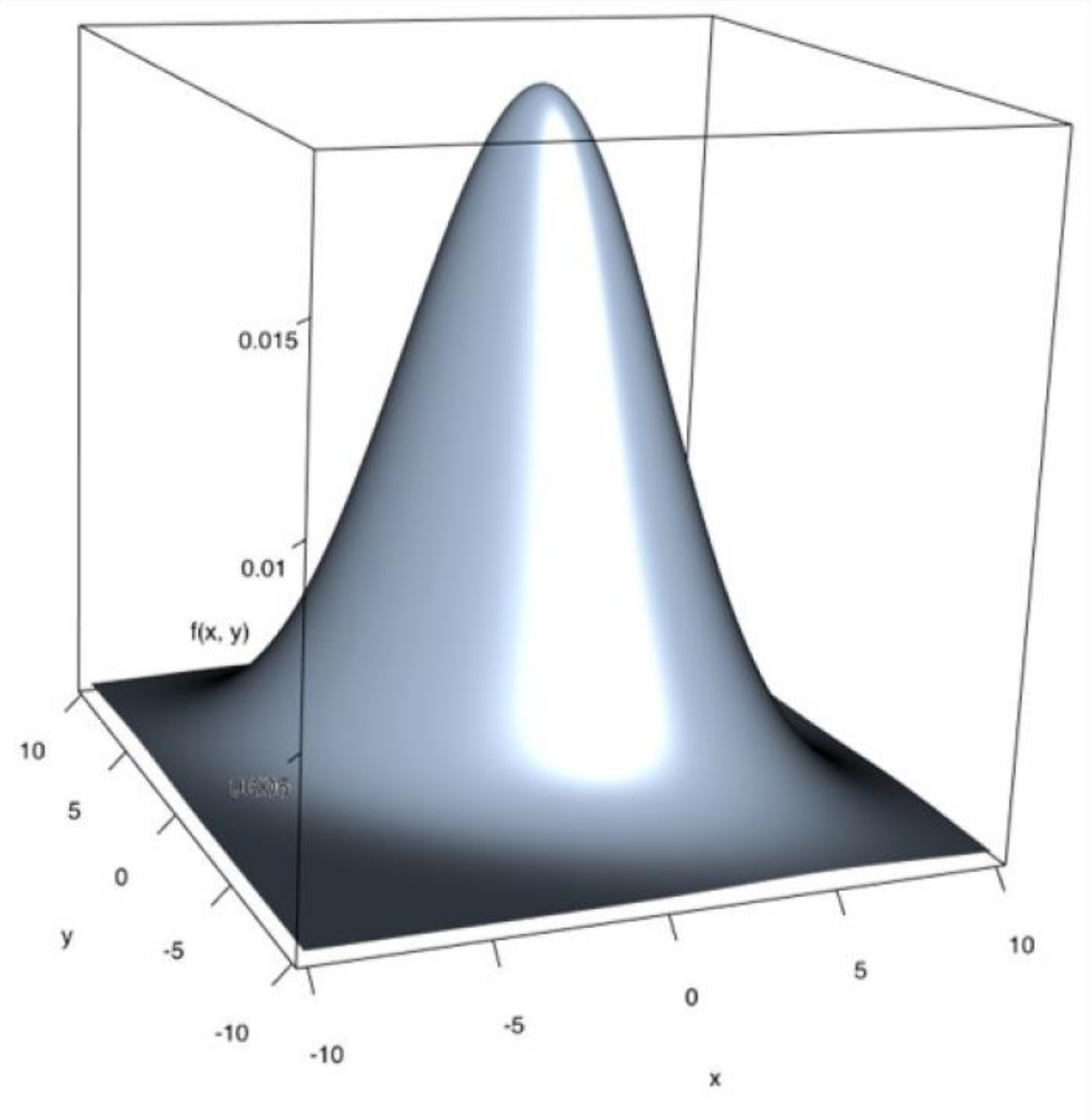
Jointly Distributed Continuous RVs

If X and Y are two jointly distributed continuous RVs, their *joint density function*, defined on the two-dimensional plane, is denoted by f_{XY} (or just f) and has the following properties:

1. $f(x, y) \geq 0$ for all x and y.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$
3. $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \left(\int_{-\infty}^{x} f_{XY}(u, v) du \right) dv$
4. $P\left[(X, Y) \in A\right] = \iint_{A} f(x, y) dx dy$

The function F above is the *joint CDF* of X and Y.

Bivariate Normal Density



Jointly Distributed Continuous RVs

Example: Let the joint density of X and Y be given by

$$f(x, y) = ke^{-(x+2y)}, 0 \leq x < \infty, 0 \leq y < \infty.$$

1. Find k.
2. Find $P(X \leq 3, Y \leq 2)$.
3. Find F.

Examples

Example: Let (X, Y) be uniformly distributed on the unit disc.

- Let A be the region on the unit disc bounded by the line $y = x$ and the x-axis and let B be the disc of radius $\frac{1}{2}$ centred at the origin.
  1. Find the joint density function of (X, Y).
  2. Find $P((X, Y) \in A)$.
  3. Find $P((X, Y) \in B)$.

Solution: 'Uniformly distributed on the unit disc' means the joint density is constant inside the circle of radius 1 centred at the origin.

1. The joint density function is given by $f(x, y) = \frac{1}{\pi}$ when $x^2 + y^2 \leq 1$.

   The probabilities are equal to the areas multiplied by $\frac{1}{\pi}$.

   The areas of A and B are $\frac{\pi}{8}$ and $\frac{\pi}{4}$ respectively.
2. $P((X, Y) \in A) = \frac{1}{\pi} \cdot \frac{\pi}{8} = \frac{1}{8}$.
3. $P((X, Y) \in B) = \frac{1}{\pi} \cdot \frac{\pi}{4} = \frac{1}{4}$.

Examples

- **Example:** Find the marginal densities of X and Y if

$$f_{XY}(x,y) = \frac{1}{4}(2x+y), 0 \leq x \leq 1, 0 \leq y \leq 2.$$

- **Solution:** For $0 \leq x \leq 1$,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\ &= \frac{1}{4}\int_0^2 (2x+y)dy \\ &= \frac{1}{4}\left[2xy + \frac{y^2}{2}\right]_{y=0}^{y=2} \\ &= \frac{1}{2}(2x+1). \end{aligned}$$

Examples

Example: A bank operates both a drive-up facility and a walk-up window. Let X and Y be the proportions of time that the drive-up facility and the walk-up window are, respectively, in use on a randomly selected day. Then the set of possible values for (X, Y) is the rectangle $[0, 1] \times [0, 1]$. Suppose the joint pdf of (X, Y) is given by

$$f(x, y) = \frac{6}{5}\left(x + y^2\right), 0 \leq x \leq 1, 0 \leq y \leq 1.$$

1. Verify that this is a legitimate PDF.
2. What is the probability that neither facility is busy more than one-quarter of the time?
3. Find the marginal density function for Y.
4. Find the marginal density function for X. (Left as an exercise.)

Solution

Solution:

1. $$\int_0^1 \int_0^1 \frac{6}{5}\left(x+y^2\right) dx dy = \int_0^1 \left(\int_0^1 \frac{6}{5}\left(x+y^2\right) dx\right) dy$$
$$= \int_0^1 \left(\frac{6}{5}\left[\frac{x^2}{2} + xy^2\right]_{x=0}^{1}\right) dy$$
$$= \int_0^1 \left(\frac{6}{10} + \frac{6}{5}y^2\right) dy$$
$$= \left[\frac{6}{10}y + \frac{6}{15}y^3\right]_{y=0}^{1} = \frac{6}{10} + \frac{6}{15} = 1$$
2. $P(X \leq 0.25, Y \leq 0.25) = \int_0^{0.25} \int_0^{0.25} \frac{6}{5}\left(x+y^2\right) dx dy = \frac{7}{640}$.
3. $f_Y(y) = \frac{6}{5}\left(y^2 + \frac{1}{2}\right)$ from the inner integral calculation above.
4. $f_X(x) = \frac{6}{5}\left(x + \frac{1}{3}\right)$ (check this for yourself)

Examples

Solution:

- For this example, we found that the marginal density of X is $f_X(x) = \frac{1}{2}(2x+1)$ for $0 \leq x \leq 1$.
- For $x \in [0,1]$ and $y \in [0,2]$, the conditional density of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{\frac{1}{4}(2x+y)}{\frac{1}{2}(2x+1)} = \frac{2x+y}{4x+2}.$$

Thus $f_{Y|X}(y|x) = \frac{2x+y}{4x+2}$ when $0 \leq x \leq 1, 0 \leq y \leq 2$

- Conditional density of Y given $X = \frac{1}{4}$ is given by

$$f_{Y|X}\left(y \middle| \frac{1}{4}\right) = \frac{f_{XY}(\frac{1}{4},y)}{f_X(\frac{1}{4})} = \frac{2\left(\frac{1}{4}\right)+y}{4\left(\frac{1}{4}\right)+2} = \frac{2y+1}{6} \text{ for } y \in [0,2]$$

Independence

Two continuous RVs are said to be *independent* if for all x and y, $f_{XY}(x, y) = f_X(x) f_Y(y)$.

This is equivalent to $F_{XY}(x, y) = F_X(x) F_Y(y)$

and $P(X \leq x, Y \leq y) = P(X \leq x) P(Y \leq y)$

- In other words, X and Y are independent when the joint PDF factorizes as product of the marginal PDFs.
- **Caution:** Independence of jointly distributed continuous RVs also requires that the support of f (the region where $f(x, y) > 0$) must be a rectangular set.

Expectation of Functions of Jointly Distributed RVs

Let X and Y be jointly distributed continuous RVs with PDF $f(x, y)$ and let h be a function defined on the range of (X, Y). Then the expectation of $h(X, Y)$ is defined as

$$E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy.$$

Examples include $h(x, y) = xy$, $h(x, y) = x + y$, $h(x, y) = x^2 y$, $h(x, y) = x$, etc.

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy.$$

Expectation of Functions of Random Variables

Result:

Let h_1 and h_2 be bivariate functions and let a_2, a_2, b be constants. Then

$$E\left[a_1 h_1(X,Y) + a_2 h_2(X,Y) + b\right] = a_1 E\left[h_1(X,Y)\right] + a_2 E\left[h_2(X,Y)\right] + b$$

Result:

If X and Y are independent, then for univariate functions g_1 and g_2,

$$E\left[g_1(X) g_2(Y)\right] = E\left[g_1(X)\right] E\left[g_2(Y)\right]$$

Example:

Let X and Y have joint density $f(x, y) = k(x + y)$ for $0 \leq x \leq 1, 0 \leq y \leq 1$. Find $\text{Corr}(X, Y)$.

Solution:

- Integrating the joint PDF, we find that $k = 1$.
- Integrating the joint PDF with respect to y and with respect to x respectively, we find that for $0 \leq x \leq 1$ and for $0 \leq y \leq 1$,

$$f_X(x) = x + \frac{1}{2}, \quad f_Y(y) = y + \frac{1}{2}.$$

$$E(X^2) = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \left[\frac{x^4}{4}+\frac{x^3}{6}\right]_0^1 = \frac{5}{12}$$

$$V(X) = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}.$$

Similarly, $E(Y) = \dfrac{7}{12}$ and $V(Y) = \dfrac{11}{144}$.

Hence

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \left(\frac{7}{12}\right)^2 = \frac{-1}{144}$$

and

$$\mathrm{Corr}(X,Y) = \frac{\frac{-1}{144}}{\sqrt{\left(\frac{11}{144}\right)\left(\frac{11}{144}\right)}} = \frac{-1}{11}$$