工程概率统计

Probability and Statistics for Engineering

第三章 联合分布

Chapter 3 Joint Distributions

Chapter 3 Joint Distributions

- 3.1 Random Vector and Joint Distribution
- 3.2 Relationship between Two Random Variables
- 3.3 Function of Multiple Random Variables
- 3.4 Multivariate Normal Distribution

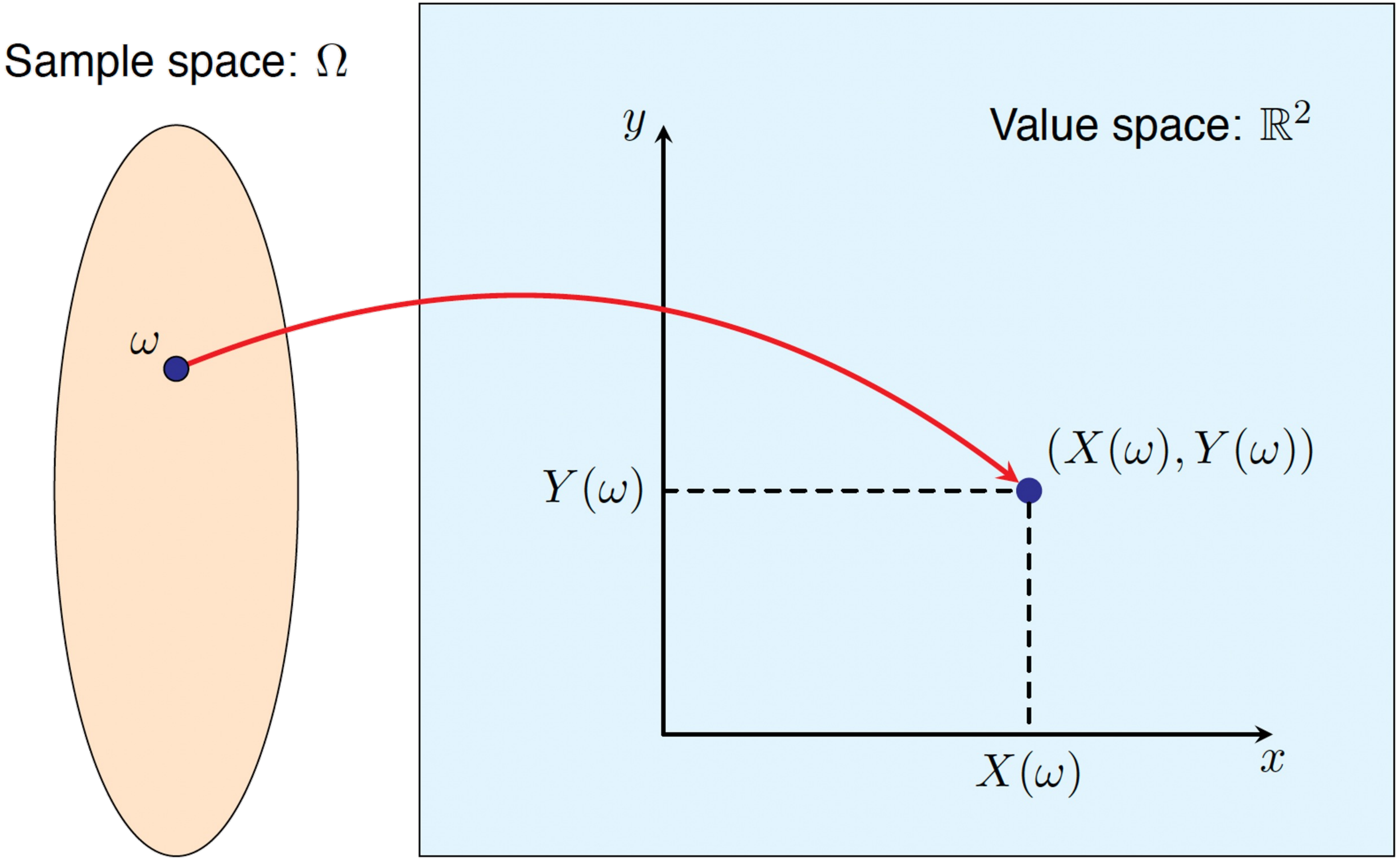
3.1 Random Vector and Joint Distribution

- We have been considering a single random variable each time, however, often we need to deal with several random variables. E.g.,
  - The height and weight of a randomly selected person.
  - The temperature, humidity, wind, precipitation of Singapore on a randomly selected day.
- Since different metrics of the same object are typically associated with each other, they should be considered simultaneously.
- Here we will talk about how to study two random variables X and Y simultaneously, all the concepts can be extended to n random variables $X_1, X_2, \dots, X_n$.

Random Vector

We say that (X, Y) is a **random vector (随机向量)** if $\omega \in \Omega \longmapsto (X(\omega), Y(\omega))$ is a function valued on $\mathbb{R}^2$. (X, Y) can also be called a two-dimensional random variable (二维随机变量).

3.1 Random Vector and Joint Distribution



3.1 Random Vector and Joint Distribution

- The major concepts that will be introduced for random vectors are:
  - (Joint) CDF/PMF/PDF;
  - Marginal CDF/PMF/PDF;
  - Conditional PMF/PDF.
- First, for a discrete/continuous random vector, we can describe its distribution using the (joint) CDF (cumulative distribution function).

Joint Cumulative Distribution Function

For a random vector (X, Y), either discrete or continuous, its cumulative distribution function (CDF, 累积分布函数) is defined as

$$F(x, y) = P(X \leq x, Y \leq y), \forall x, y \in \mathbb{R}.$$

$F(x, y)$ is also called the joint CDF (联合累积分布函数) of X and Y.

- You can treat $\{X \leq x\}$ as an event A, and $\{Y \leq y\}$ as an event B, then

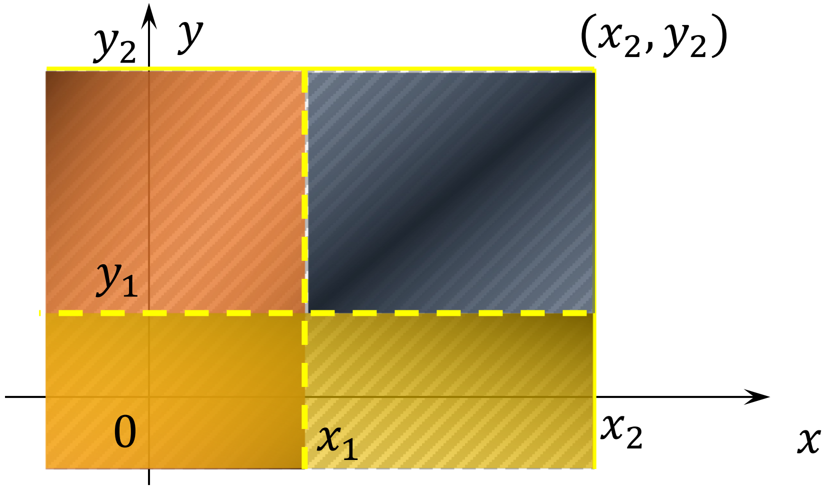
$$F(x, y) = P(A \cap B).$$

- It follows that for $\forall x, y \in \mathbb{R}$, we have $0 \leq F(x, y) \leq 1$ and

$$F(+\infty, +\infty) = 1, F(-\infty, -\infty) = 0,$$

$$F(-\infty, y) = 0, F(x, -\infty) = 0.$$

3.1 Random Vector and Joint Distribution



- The joint CDF can be used to compute $P(x_1 < X \le x_2, y_1 < Y \le y_2)$ for any $-\infty < x_1 < x_2 < \infty, -\infty < y_1 < y_2 < \infty$.

$$P\{x_1 < X \le x_2,\ \ y_1 < Y \le y_2\}$$
$$= F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$
$$\ge 0.$$

- With the joint CDF, it is straight forward to define the marginal CDF.

Marginal CDF

Let $F(x, y)$ be the joint CDF of (X, Y), $F_X(x)$ be the CDF of X without considering Y, and $F_Y(y)$ be the CDF of Y without considering X, then for $\forall x, y \in \mathbb{R}$:

$$F_X(x) = F(x, \infty), F_Y(y) = F(\infty, y).$$

F_X (F_Y) is also called the marginal CDF (边缘累积分布函数) of X (Y).

Note: While the joint CDF uniquely determines the marginal CDFs, the reverse is not true.

3.1 Random Vector and Joint Distribution

Example 3.1

- Suppose that the joint CDF of random vector (X,Y) is (a,b,c are constants)

$$F(x,y)=a\left(b+\arctan\frac{x}{2}\right)\left(c+\arctan\frac{y}{2}\right),-\infty<x,y<\infty.$$

- 1. Determine the value of a,b,c. 2. Calculate $P(-2<X\leq 2,-2<Y\leq 2)$.
- 3. Obtain the marginal CDFs of X and Y.

Solution

- 1. By the basic properties of the joint CDF, we have:

$$F(+\infty,+\infty)=a\left(b+\frac{\pi}{2}\right)\left(c+\frac{\pi}{2}\right)=1,\ F(-\infty,+\infty)=a\left(b-\frac{\pi}{2}\right)\left(c+\frac{\pi}{2}\right)=0,$$

$$F(+\infty,-\infty)=a\left(b+\frac{\pi}{2}\right)\left(c-\frac{\pi}{2}\right)=0.\quad\Rightarrow b=\frac{\pi}{2},c=\frac{\pi}{2},a=\frac{1}{\pi^2}.$$

- Therefore, the joint CDF is

$$F(x,y)=\frac{1}{\pi^2}\left(\frac{\pi}{2}+\arctan\frac{x}{2}\right)\left(\frac{\pi}{2}+\arctan\frac{y}{2}\right).$$

Solution

- 2. By the definition and property of the joint CDF, we have:

$$P(-2 < X \le 2, -2 < Y \le 2) = F(2,2) - F(2,-2) - F(-2,2) + F(-2,-2)$$
$$= \frac{1}{\pi^2}\left[\left(\frac{\pi}{2} + \arctan(1)\right)^2 - 2\left(\frac{\pi}{2} + \arctan(1)\right)\left(\frac{\pi}{2} + \arctan(-1)\right) + \left(\frac{\pi}{2} + \arctan(-1)\right)^2\right]$$
$$= \frac{1}{\pi^2}\left[\left(\frac{\pi}{2} + \arctan(1)\right) - \left(\frac{\pi}{2} + \arctan(-1)\right)\right]^2 = \frac{1}{\pi^2}\left[\frac{\pi}{4} - \left(-\frac{\pi}{4}\right)\right]^2 = \frac{1}{4}.$$

- 3. By the definition of the marginal distribution, we have:

$$F_X(x) = F(x, +\infty) = \frac{1}{\pi^2}\left(\frac{\pi}{2} + \arctan\frac{x}{2}\right)\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = \frac{1}{2} + \frac{1}{\pi}\arctan\frac{x}{2},$$
$$F_Y(y) = F(+\infty, y) = \frac{1}{2} + \frac{1}{\pi}\arctan\frac{y}{2}.$$

3.1 Random Vector and Joint Distribution

- Then for discrete random vectors, we introduce the joint and marginal PMF.

Joint and Marginal PMF for Discrete Random Vector

- For a random vector (X, Y), let $S_X = \{x_1, x_2, \dots\}$ and $S_Y = \{y_1, y_2, \dots\}$ be the support of X and Y, respectively. Then the joint PMF (联合概率质量函数) of (X, Y) is defined as
$$p(x_i, y_j) = P(X = x_i, Y = y_j) \triangleq p_{ij}, i, j = 1, 2, \dots.$$
- The marginal PMF (边缘概率质量函数) of X is the PMF of X without considering Y:
$$p_X(x_i) = P(X = x_i) = \sum_{j=1}^{\infty} p_{ij} \triangleq p_{i\cdot}, i = 1, 2, \cdots.$$
- Similarly, the marginal PMF of Y is
$$p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij} \triangleq p_{\cdot j}, j = 1, 2, \dots.$$

- The joint PMF satisfies:
- Non-negativity: $p_{ij} \geq 0, i, j = 1, 2, \dots;$
- Normalization: $\sum_i \sum_j p_{ij} = 1.$

- The joint PMF is typically displayed in a tabular format:

$X\backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$

3.1 Random Vector and Joint Distribution

Example 3.2

- Two dice are tossed independently. Let X be the smaller number of points and Y be the larger number of points. If both dice show the same number, say, z points, then $X = Y = z$.
- 1. Find the joint PMF of (X, Y); 2. Find the marginal PMF of X.

Solution

- 1. Since the two dice are tossed independently, it is not difficult to obtain the joint PMF of (X, Y):

$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/18	1/18	1/18	1/18	1/18
2	0	1/36	1/18	1/18	1/18	1/18
3	0	0	1/36	1/18	1/18	1/18
4	0	0	0	1/36	1/18	1/18
5	0	0	0	0	1/36	1/18
6	0	0	0	0	0	1/36

- 2. With the joint PMF of (X, Y), the marginal PMF of X can be directly obtained as:

Value	1	2	3	4	5	6
Prob.	$\frac{11}{36}$	$\frac{1}{4}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{1}{12}$	$\frac{1}{36}$

3.1 Random Vector and Joint Distribution

- Then for continuous random vectors, we introduce the joint PDF.

Joint PDF for Continuous Random Vector

- (X,Y) is said to be a continuous random vector if there exists a non-negative function $f(x,y)$, defined for all $(x,y) \in \mathbb{R}^2$, satisfies that for any $D \subset \mathbb{R}^2$,

$$P\big((X,Y) \in D\big) = \iint_{(x,y)\in D} f(x,y)dxdy.$$

  $f(x,y)$ is called the joint PDF (联合概率密度函数) of (X,Y).

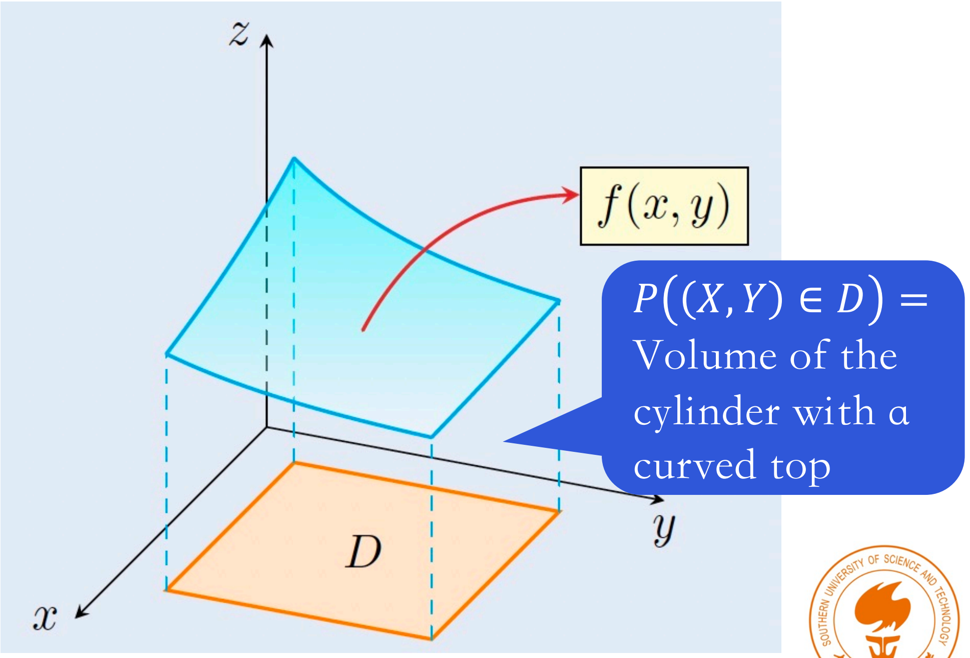
- Particularly, we have the joint CDF of (X,Y) to be

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v)dudv.$$

- It follows that

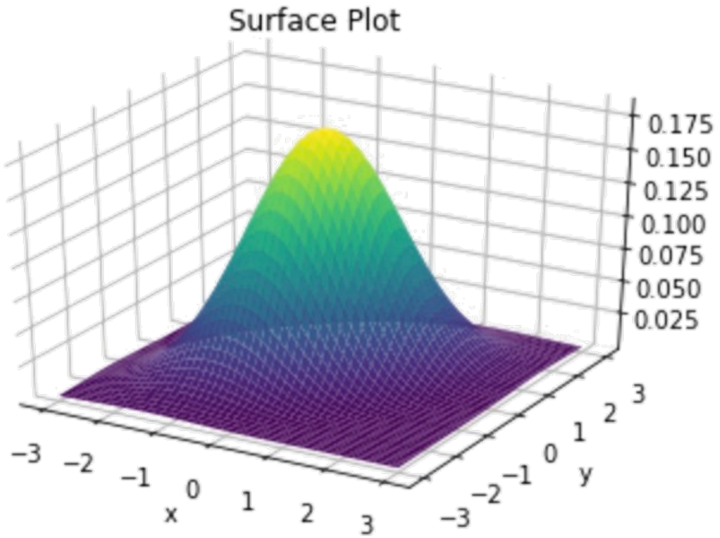
$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y).$$

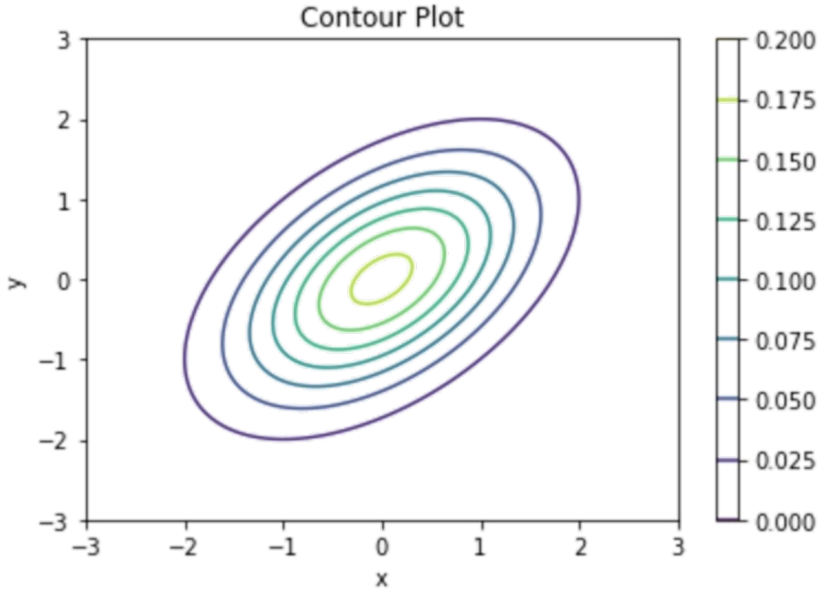
- The joint PDF satisfies:
- Non-negativity: $f(x,y) \geq 0, \forall (x,y) \in \mathbb{R}$;
- Normalization: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1$.



3.1 Random Vector and Joint Distribution

- Similar to the PDF of a single r.v., the joint PDF $f(x,y) \neq P(X=x, Y=y)$. Instead, it reflects the degree to which the probability is concentrated around (x,y).
- The joint PDF are typically visualized with the **surface plot (曲面图)** or the **contour plot (等高线图)** which help in intuitively understanding the distribution and relationship between X and Y.
  - **Surface plot**: a 3D plot where the height represents the value of the joint PDF at each point (x,y).
  - **Contour plot**: a 2D plot showing level curves where the joint PDF has constant values.





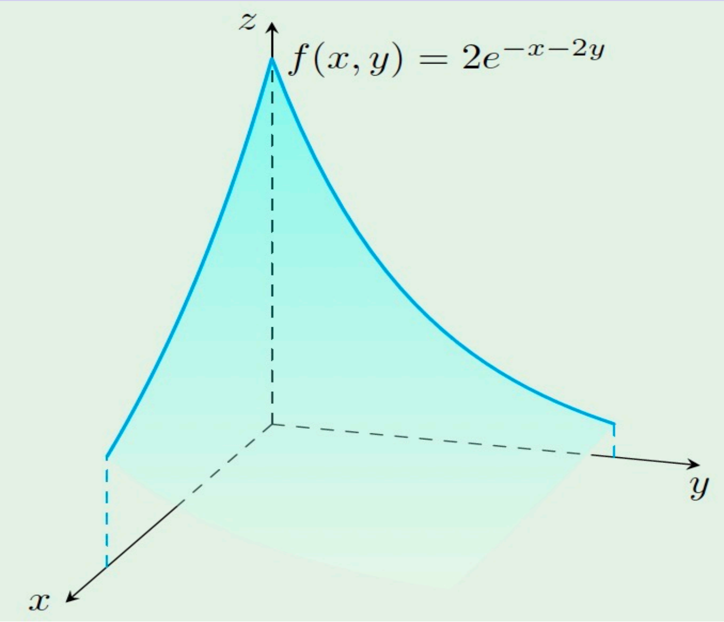
3.1 Random Vector and Joint Distribution

Example 3.3

- The lifetime (in years) of two electronic components of a randomly selected machine is denoted by r.v.s X and Y, which has a joint PDF

$$f(x,y)=\begin{cases}2e^{-x-2y}, & 0<x,y<\infty\\ 0, & \text{otherwise}\end{cases}.$$

- Compute: 1. $P(X<1, Y<1)$; 2. $P(X<Y)$.



3.1 Random Vector and Joint Distribution

Solution

- 1. By the definition of the joint PDF of (X,Y), we have:

$$P(X<1,Y<1)=\int_{-\infty}^{1}\left(\int_{-\infty}^{1}f(x,y)\,dx\right)dy\ =\int_{0}^{1}\left(\int_{0}^{1}e^{-x}dx\right)2e^{-2y}dy$$

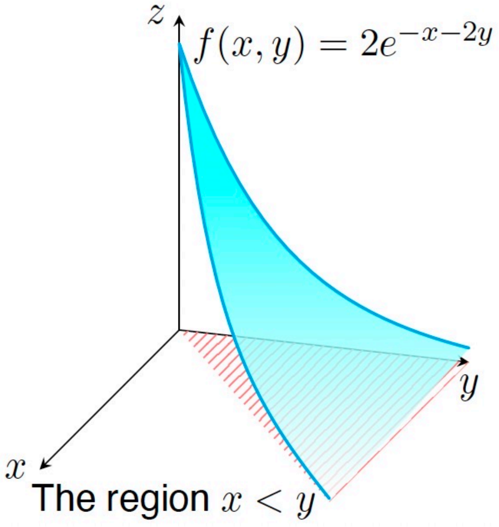
$$=(1-e^{-1})\int_{0}^{1}2e^{-2y}dy\ =(1-e^{-1})(1-e^{-2})\approx 0.5466.$$

- 2. Again, by the joint PMF of (X,Y):

$$P(X<Y)=\iint_{(x,y):x<y}f(x,y)dxdy\ =\int_{0}^{\infty}\left(\int_{0}^{y}2e^{-x-2y}dx\right)dy$$

$$=\int_{0}^{\infty}2e^{-2y}(1-e^{-y})dy\ =\int_{0}^{\infty}2e^{-2y}dy-\int_{0}^{\infty}2e^{-3y}dy$$

$$=1-\frac{2}{3}=\frac{1}{3}.$$



3.1 Random Vector and Joint Distribution

- With the joint PDF, the marginal PDF can be determined.
- The marginal PDF mirrors the definition of the marginal PMF for the discrete case, except with sums replaced by integrals and the joint PMF replaced by the joint PDF.

Marginal PDF for Continuous Random Vector

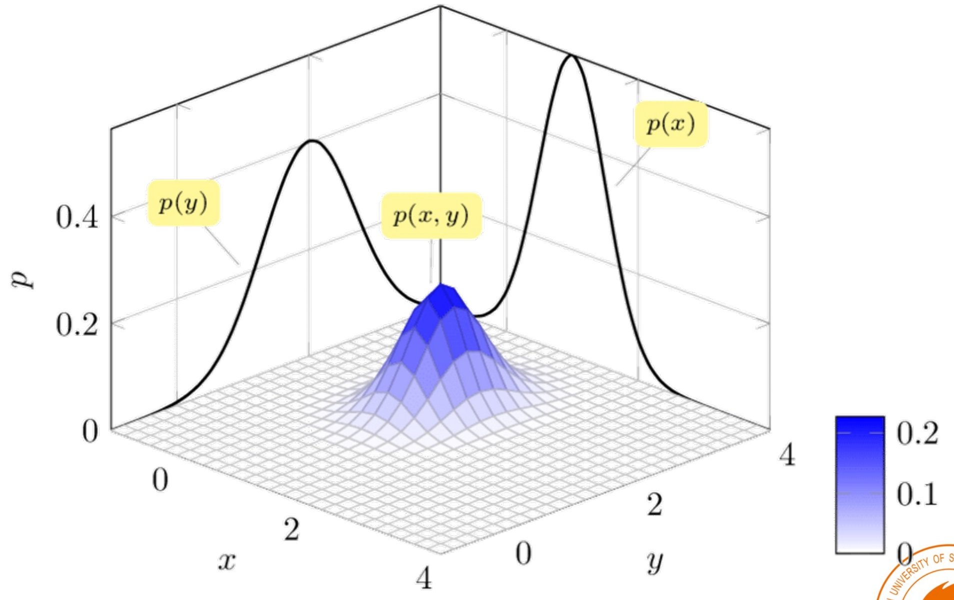
- (X, Y) a continuous random vector with joint PDF $f(x, y)$, then the **marginal PDF (边缘概率密度函数)** of X, i.e., the PDF of X without considering Y is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy .$$

- Likewise, the marginal PDF of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

- Note that a joint PDF uniquely defines the marginal PDFs, however, the reverse is not true.



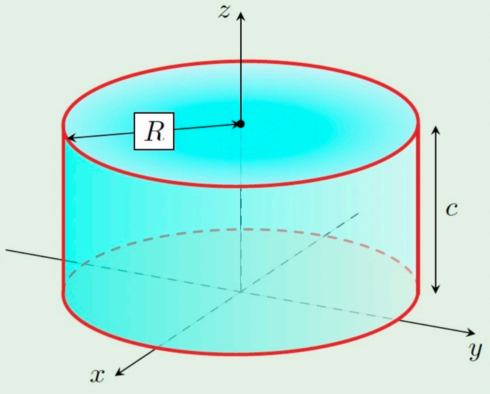
3.1 Random Vector and Joint Distribution

Example 3.4

- Suppose that the joint PDF of a random vector (X, Y) is given by

$$f(x,y)=\begin{cases} c, \text{if } x^2+y^2\le R^2 \\ 0, \quad\quad \text{otherwise} \end{cases}$$

- 1. Determine the constant c;
- 2. Find the marginal PDF of X.



Solution

- 1. By the normalization property of the joint PDF:

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

- it follows that

$$1=\iint_{(x,y):x^2+y^2\le R^2} c\,dxdy \Longrightarrow c=\frac{1}{\pi R^2}.$$

- 2. With the joint PDF of (X, Y), the marginal PDF of X can be directly obtained as:

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy=\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} c dy=\frac{2\sqrt{R^2-x^2}}{\pi R^2},$$

- if $|x|\le R$; and $f_X(x)=0$, otherwise.

3.1 Random Vector and Joint Distribution

- Finally, we talk about the conditional PMF/PDF.

Conditional PMF/PDF

- For a **discrete** random vector (X, Y) with joint PMF $p(x, y)$, the conditional PMF (条件概率质量函数) of X given $Y = y$ is defined as
$$p_{X|Y}(x|y) \triangleq P(X = x|Y = y) = \frac{p(x, y)}{p_Y(y)},$$
for all values of y such that $p_Y(y) > 0$.
- For a **continuous** random vector (X, Y) with joint PDF $f(x, y)$, the conditional PDF (条件概率密度函数) of X given $Y = y$ is defined as
$$f_{X|Y}(x|y) \triangleq \frac{f(x, y)}{f_Y(y)},$$
for all values of y such that $f_Y(y) > 0$.

- The conditional PDF mirrors the definition of the conditional PMF for the discrete case, except with the joint/marginal PMF replaced by the joint/marginal PDF.
- The conditional PMF/PDFs also satisfy the non-negativity and normalization properties.
- Question: for the continuous case, $P(Y = y) = 0$, so that $P(X = x|Y = y)$ or $P(X \leq x|Y = y)$ is not defined. Then how to understand the conditional PDF of X given $Y = y$?

3.1 Random Vector and Joint Distribution

- Here we talk about how to understand the conditional PDF of X given $Y = y$.
- Conditioning on $Y = y$ can be understood as conditioning on $\{y \le Y \le y + \varepsilon\}$ where $\varepsilon \to 0$.
- Consider the conditional CDF:

$$P\{X \le x | y < Y \le y + \varepsilon\} = \frac{P\{X \le x, y < Y \le y + \varepsilon\}}{P\{y < Y \le y + \varepsilon\}} = \frac{\int_{-\infty}^{x} \int_{y}^{y+\varepsilon} f(u, v) dv du}{\int_{y}^{y+\varepsilon} f_Y(y) dy}$$

By the mean value theorem of integrals (积分中值定理)

$$= \frac{\varepsilon \int_{-\infty}^{x} f(u, y_\varepsilon) du}{\varepsilon f_Y(\tilde{y}_\varepsilon)} \to \int_{-\infty}^{x} \frac{f(u, y)}{f_Y(y)} du \quad (\varepsilon \to 0).$$

The conditional PDF

- By the definition of the conditional PDF, we have

$$f(x, y) = f_{X|Y}(x|y) f_Y(y).$$

- Take the integration w.r.t. y on both sides, the marginal distribution of X can be expressed as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$$

The law of total probability under the continuous case

3.1 Random Vector and Joint Distribution

Example 3.5

- For a randomly selected person in an automobile accident, let X be his/her extent of injury and Y be the type of safety restraint he/she was wearing at the time of the accident. The joint PMF of X and Y is

$X \backslash Y$	1 (None)	2 (Belt Only)	3 (Belt and Harness)	$p_X(x)$
1 (None)	0.065	0.075	0.060	0.20
2 (Minor)	0.165	0.160	0.125	0.45
3 (Major)	0.145	0.10	0.055	0.30
4 (Death)	0.025	0.015	0.010	0.05
$p_Y(y)$	0.40	0.35	0.25	1.00

- 1. What is the PMF of extent of injury for a randomly selected person with no restraint?
- 2. What is the PMF of extent of injury for a randomly selected person with belt and harness?

3.1 Random Vector and Joint Distribution

Solution

- 1. By the definition of conditional PMF, we have the conditional distribution of X given $Y = 1$ is

Value	1 (None)	2 (Minor)	3 (Major)	4 (Death)
Prob.	$\frac{0.065}{0.4} = 0.1625$	$\frac{0.165}{0.4} = 0.4125$	$\frac{0.145}{0.4} = 0.3625$	$\frac{0.025}{0.4} = 0.0625$

- 2. Similarly, the conditional distribution of X given $Y = 3$ is

Value	1 (None)	2 (Minor)	3 (Major)	4 (Death)
Prob.	$\frac{0.06}{0.25} = 0.24$	$\frac{0.125}{0.25} = 0.50$	$\frac{0.055}{0.25} = 0.22$	$\frac{0.01}{0.25} = 0.04$

- With belt and harness, the probability of a major injury or death is 16.5% lower than the case without any safety restraint.

3.1 Random Vector and Joint Distribution

Example 3.4 (Continued)

- Determine the conditional PDF of X given $Y = y$ (where $|y| \le R$).

Solution

- Similar with 2, we obtain the marginal PDF of Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \begin{cases} \frac{2\sqrt{R^2 - y^2}}{\pi R^2}, & \text{if } |y| \le R \\ 0, & \text{otherwise} \end{cases}.$$

- Then, by the definition, the conditional PDF of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{2\sqrt{R^2 - y^2}}, & \text{if } |x| \le \sqrt{R^2 - y^2} \\ 0, & \text{otherwise} \end{cases}.$$

- This suggest that given $Y = y$, X follows $\text{Uniform}[-\sqrt{R^2 - y^2}, \sqrt{R^2 - y^2}]$.
- Since $f_{X|Y}(x|y) \ne f_X(x)$, we say that X is not independent of Y.

Chapter 3 Joint Distributions

3.2 Relationship Between Two Random Variables

- In the end of Section 3.1, we mentioned the concept of independence between random variables.
- Recall the independence between random events, the independence between random variables can be defined similarly: the value of Y does not affect the distribution of X or vice versa.
- For example, for the continuous case:

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

or

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

By $f(x,y) = f_X(x)f_Y(y)$, we have

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)du\,dv = \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv = F_X(x)F_Y(y).$$

3.2 Relationship Between Two Random Variables

Independence of Random Variables

- Let $F(x_1, x_2, \dots x_n)$ be the joint CDF of $(X_1, X_2, \dots, X_n)$, $F_{X_i}(x_i)$ be the marginal CDF of X_i, then if for $\forall x_1, x_2, \dots, x_n \in \mathbb{R}$ we have
$$F(x_1, x_2, \dots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n),$$
then we say that random variables $X_1, X_2, \cdots, X_n$ are **(mutually) independent (相互独立)**.
- For discrete random variables $X_1, X_2, \cdots, X_n$, if they are independent, then the PMF satisfies
$$p(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n).$$
- For continuous random variables $X_1, X_2, \cdots, X_n$, if they are independent, then the PDF satisfies
$$f(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

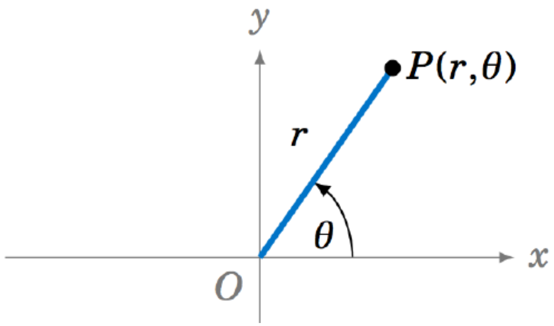
- **Question:** the definition of independence between three or more random events require multiple equations, why the independence between three or more random variables only require one?

3.2 Relationship Between Two Random Variables

Example 3.6

- The PDF of a standard normal random variable Z is $f(z) = ce^{-z^2/2}$, $-\infty < z < \infty$.
- We already know that $c = 1/\sqrt{2\pi}$, however, how is this value obtained?
- Surprisingly, the easiest way to determine c is to define two independent standard normal random variables and use the fact that their joint PDF must integrate to 1.

3.2 Relationship Between Two Random Variables



Solution

- Let random variables $X \sim \mathrm{N}(0,1)$, $Y \sim \mathrm{N}(0,1)$, X and Y are independent. Then the joint PDF of X and Y is

$$f(x,y) = ce^{-\frac{x^2}{2}} \cdot ce^{-\frac{y^2}{2}} = c^2 e^{-\frac{x^2+y^2}{2}}.$$

- Since the joint PDF must integrate to 1, we have

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c^2 e^{-\frac{x^2+y^2}{2}} dx\, dy\,.$$

- Surprisingly, this double integral can be evaluated even though the single integral could not.
- To evaluate the double integral, we convert to polar coordinates (极坐标), using the substitutions $x = r\sin\theta$, $y = r\cos\theta$ and $dxdy = rdrd\theta$:

$$c^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx\, dy = c^2 \int_{0}^{2\pi} \int_{0}^{\infty} e^{-\frac{r^2}{2}} rdr\, d\theta = c^2 \int_{0}^{2\pi} 1 d\theta = 2\pi c^2 \Longrightarrow c = \frac{1}{\sqrt{2\pi}}.$$

3.2 Relationship Between Two Random Variables

Example 3.7

- Suppose that the number of people who enter a shopping mall on a randomly selected weekday follows a Poisson distribution with parameter λ.
- If each person who enters the shopping mall is a male with probability 0.2 and a female with probability 0.8.
- Show that the number of males and females entering the shopping mall are independent Poisson random variables with parameters 0.2λ and 0.8λ, respectively.

Solution

- Let random variables X_1 and X_2 be the number of males and females that enter the shopping mall. By the definition of independence, we need to show that for $\forall i_1, i_2 = 0, 1, \dots,$

$$P(X_1 = i_1, X_2 = i_2) = P(X_1 = i_1)P(X_2 = i_2).$$

3.2 Relationship Between Two Random Variables

Solution

- Let $Y=X_1+X_2$ be the total number of people that enter the shopping mall, then $Y \sim \text{Poisson}(\lambda)$, i.e.,

$$P(Y=i_1+i_2)=e^{-\lambda}\frac{\lambda^{i_1+i_2}}{(i_1+i_2)!}.$$

- Given that $Y=i_1+i_2$, it follows that $X_1 \sim \text{Binomial}(i_1+i_2, 0.2)$, so that

$$P(X_1=i_1|Y=i_1+i_2)=\binom{i_1+i_2}{i_1}0.2^{i_1}0.8^{i_2}=\frac{(i_1+i_2)!}{i_1!\,i_2!}0.2^{i_1}0.8^{i_2}.$$

- Therefore,

$$\begin{aligned}P(X_1=i_1,X_2=i_2)&=P(X_1=i_1,X_2=i_2|Y=i_1+i_2)P(Y=i_1+i_2)\\&=\frac{(i_1+i_2)!}{i_1!\,i_2!}0.2^{i_1}0.8^{i_2}\times e^{-\lambda}\frac{\lambda^{i_1+i_2}}{(i_1+i_2)!}=\left(e^{-0.2\lambda}\frac{(0.2\lambda)^{i_1}}{i_1!}\right)\left(e^{-0.8\lambda}\frac{(0.8\lambda)^{i_2}}{i_2!}\right).\end{aligned}$$

- With the joint PMF, it is not difficult to determine the marginal PMFs, i.e., $X_1 \sim \text{Poisson}(0.2\lambda)$ and $X_2 \sim \text{Poisson}(0.8\lambda)$, and the independence between X_1 and X_2 is proved.

3.2 Relationship Between Two Random Variables

- If X and Y are independent random variables, then, for any functions g and h, we have

$$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)],$$

$$\mathrm{Var}[g(X) \pm h(Y)] = \mathrm{Var}[g(X)] + \mathrm{Var}[h(Y)].$$

Proof: Without loss of generality, show the case for the continuous case.

Suppose that X and Y have joint density $f(x,y)$, then

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

$$= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy = \mathrm{E}[g(X)]\mathrm{E}[h(Y)].$$

Let $\mathrm{E}[g(X)] = a$ and $\mathrm{E}[h(Y)] = b$, then

$$= \mathrm{E}[g(X) - a]\mathrm{E}[h(Y) - b] = 0$$

$$\begin{aligned}\mathrm{Var}[g(X) + h(Y)] &= \mathrm{E}[(g(X) + h(Y) - a - b)^2] \\ &= \mathrm{E}[(g(X) - a)^2] + \mathrm{E}[(h(Y) - b)^2] + 2\mathrm{E}[(g(X) - a)(h(Y) - b)] \\ &= \mathrm{Var}[g(X)] + \mathrm{Var}[h(Y)].\end{aligned}$$

3.2 Relationship Between Two Random Variables

- Special case: $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent.
- What if X and Y are not independent?
- Think about the difference between $\mathrm{Var}(X+Y)$ and $\mathrm{Var}(X) + \mathrm{Var}(Y)$:

$$\mathrm{Var}(X+Y) - \mathrm{Var}(X) - \mathrm{Var}(Y) = 2\mathrm{E}\big[\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\big] = 2[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)].$$

- So, if $\mathrm{E}\big[\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\big] \neq 0$, X and Y cannot be independent.
- Therefore, $\mathrm{E}\big[\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\big]$ can be used to measure the relationship between X and Y.

Covariance

- The **covariance (协方差)** between X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

$$\mathrm{Cov}(X,Y) \triangleq \mathrm{E}\big[\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\big] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

- If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. However, if $\mathrm{Cov}(X,Y) = 0$, X and Y may not be independent, we can only say that X and Y are **uncorrelated (不相关的)**.

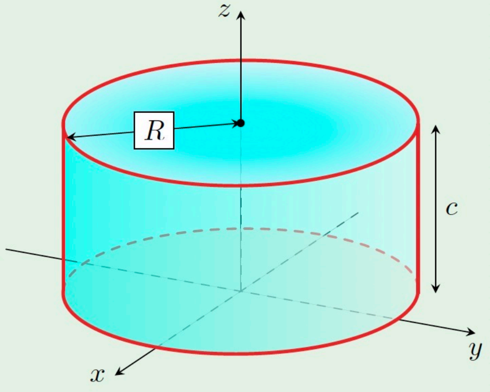
3.2 Relationship Between Two Random Variables

Example 3.4 (Continued)

- We already know that X and Y are not independent, however, show that $\mathrm{Cov}(X,Y)=0$.

Proof

- Since the marginal PDFs of X and Y are both even functions (偶函数), it follows directly that
$$\mathrm{E}(X)=\mathrm{E}(Y)=0.$$
- Then, we compute $\mathrm{E}(XY)$:
$$\mathrm{E}(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dx\,dy=\iint_{\{x^2+y^2\le R^2\}} cxydxdy$$
$$=c\int_{-R}^{R}\left(\int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} xdx\right)ydy=0.$$



- Therefore, $\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)=0$.

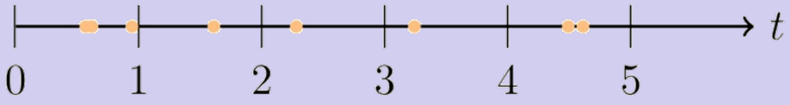
3.2 Relationship Between Two Random Variables

- Note that $\mathrm{Cov}(X,Y)$ is positive when X and Y tend to vary in the same direction and negative when they tend to vary in the opposite direction.
- The covariance has the following properties: (a, b, c are constants)
  - Covariance-variance relationship: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
  $$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y).$$
  - Symmetry: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
  - Constants cannot covary: $\mathrm{Cov}(X,c) = 0$.
  - Pulling out constants: $\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$.
  - Distributive property: $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y)$.
  - Bilinear property:
  $$\mathrm{Cov}(a_1X_1 + a_2X_2 + \cdots a_nX_n, b_1Y_1 + b_2Y_2 + \cdots + b_mY_m) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_j\mathrm{Cov}(X_i,Y_j).$$

3.2 Relationship Between Two Random Variables

Example 3.8

- A Geiger counter (盖革计数器) is a device used for detecting and measuring ionizing radiation (电离辐射).
- Each time it detects a radioactive particle(放射性粒子), it makes a clicking sound. E.g., the orange points below indicates the times at which the Geiger counter detect a particle.



- Suppose that in a city, radioactive particles reach a Geiger counter according to a Poisson process at a rate of $\lambda = 0.8$ particles per second.
- The time that the first particle is detected and the time that the second particle is detected are denoted by X and Y, respectively.
- Calculate the covariance between X and Y.

3.2 Relationship Between Two Random Variables

Solution

- Let $Z = Y - X$, then Z represents the time interval between the arrival of the first and the second particle.
- By our previous knowledge, we know that $X \sim \text{Exp}(\lambda)$, $Z \sim \text{Exp}(\lambda)$, and X and Z are independent.
- Therefore, we have

$$\mathrm{E}(X) = \mathrm{E}(Z) = \frac{1}{\lambda}, \ \mathrm{Var}(X) = \mathrm{Var}(Z) = \frac{1}{\lambda^2}, \text{ and } \mathrm{Cov}(X, Z) = 0.$$

- Then,

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, X + Z) = \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Z) = \mathrm{Var}(X) = \frac{1}{\lambda^2} = 1.5625.$$

- $\mathrm{Cov}(X, Y) > 0$ is consistent with our intuition: the longer it takes for the first arrival to happen, the longer we will have to wait for the second arrival, since the second arrival has to happen after the first.

3.2 Relationship Between Two Random Variables

- While the covariance measures the relationship between two random variables, its value depends on the unit/scale on which we measure the random variables.
  - E.g., let X (in m) and Y be the height and weight (in kg) of a randomly selected person, and $\tilde{X} = 100X$ (i..e., $\tilde{X}$ is the height measured in cm) then $\mathrm{Cov}\big(\tilde{X}, Y\big) = 100\mathrm{Cov}(X, Y)$.
  - Therefore, a larger covariance does not necessarily suggest a stronger relationship.
- To make the measure comparable, we need to remove the impact of unit/scale.

Correlation Coefficient

- The **correlation coefficient (相关系数)** between X and Y, denoted by $\mathrm{Cor}(X, Y)$ or ρ_{XY}, is defined by

$$\rho_{XY} = \mathrm{Cor}(X, Y) \triangleq \mathrm{E}\left[\frac{\big(X - \mathrm{E}(X)\big)}{\mathrm{SD}(X)} \cdot \frac{\big(Y - \mathrm{E}(Y)\big)}{\mathrm{SD}(Y)}\right] = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- ρ_{XY} is a normalized version of the covariance, which is a dimensionless quantity (无量纲数值).
- **Question**: what kind of relationship is ρ_{XY} measuring?

Example 3.9

- Calculate ρ_{XY} if the joint PDF of X and Y is $(0 < c \le 1)$

$$f(x,y) = \frac{1}{2\pi c}\exp\left\{-\frac{x^2 - 2\sqrt{1-c^2}xy + y^2}{2c^2}\right\}, -\infty < x, y < \infty.$$

Solution

Consider the normal PDF

- Consider the marginal PDF of X:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{2\pi c}\int_{-\infty}^{\infty}\exp\left\{-\frac{\left(y-\sqrt{1-c^2}x\right)^2}{2c^2} - \frac{c^2x^2}{2c^2}\right\}dy = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

- So $X \sim \mathrm{N}(0,1)$ and similarly, we obtain $Y \sim \mathrm{N}(0,1)$. It follows that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dx\,dy = \frac{1}{2\pi c}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} cv\left(cu + \sqrt{1-c^2}v\right)e^{-(u^2+v^2)/2}du\,dv$$

$$u = \frac{y - \sqrt{1-c^2}x}{c}, v = x$$

3.2 Relationship Between Two Random Variables

Solution

- Continued with the previous derivation

$$\mathrm{Cov}(X,Y)=\mathrm{E}(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dx\,dy=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(cuv+\sqrt{1-c^2}v^2\right)e^{-(u^2+v^2)/2}du\,dv$$

$$=\frac{c}{2\pi}\int_{-\infty}^{\infty}ue^{-u^2/2}du\int_{-\infty}^{\infty}ve^{-v^2/2}\,dv+\frac{\sqrt{1-c^2}}{2\pi}\int_{-\infty}^{\infty}e^{-u^2/2}du\int_{-\infty}^{\infty}v^2e^{-v^2/2}\,dv$$

$$=0+\frac{\sqrt{1-c^2}}{2\pi}\cdot\sqrt{2\pi}\cdot\sqrt{2\pi}=\sqrt{1-c^2}.$$

- Therefore, $\rho_{XY}=\mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}=\sqrt{1-c^2}$.
- This is an example showing that the marginal PDFs cannot uniquely determine the joint PDF.

3.2 Relationship Between Two Random Variables

- ρ_{XY} actually measure the direction and strength of **the linear relationship** between X and Y.

Proof: Consider to use a linear function of X to approximate Y, i.e., $\hat{Y} = a + bX$.

Then, the mean squared error (MSE, 均方误差) of the approximation is

$$\begin{aligned}\mathrm{MSE} &= \mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right] = \mathrm{E}[(Y - a - bX)^2] \\ &= \mathrm{E}(Y^2) + b^2\mathrm{E}(X^2) + a^2 - 2b\mathrm{E}(XY) + 2ab\mathrm{E}(X) - 2a\mathrm{E}(Y).\end{aligned}$$

Next, we would like to minimize the MSE w.r.t. a and b.

$$\begin{cases} \dfrac{\partial \mathrm{MSE}}{\partial a} = 2a + 2b\mathrm{E}(X) - 2\mathrm{E}(Y) = 0 \\ \dfrac{\partial \mathrm{MSE}}{\partial b} = 2b\mathrm{E}(X^2) - 2\mathrm{E}(XY) + 2a\mathrm{E}(X) = 0 \end{cases} \Rightarrow \begin{cases} b_0 = \dfrac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\mathrm{E}(X^2) - [\mathrm{E}(X)]^2} = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \\ a_0 = \mathrm{E}(Y) - b_0\mathrm{E}(X) = \mathrm{E}(Y) - \mathrm{E}(X)\dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \end{cases}$$

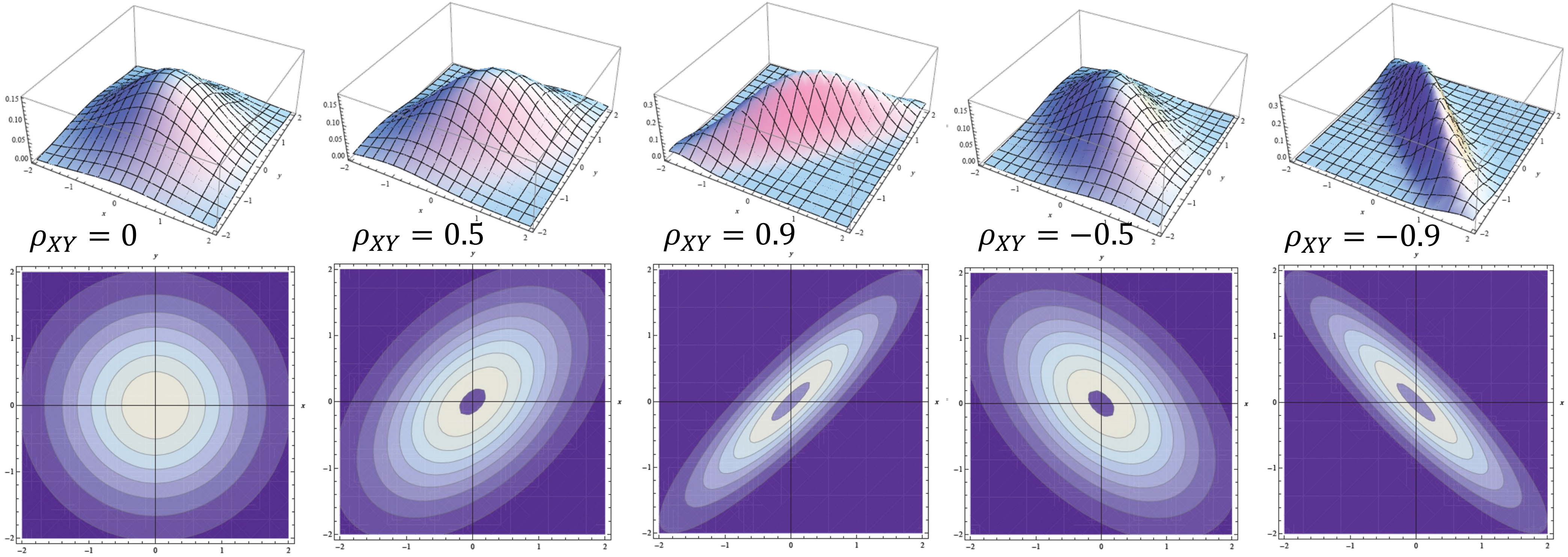
Therefore,

$$\begin{aligned}\min_{a,b} \mathrm{MSE} &= \mathrm{E}[(Y - a_0 - b_0X)^2] = \mathrm{E}[(Y - \mathrm{E}(Y) + b_0\mathrm{E}(X) - b_0X)^2] \\ &= \mathrm{Var}(Y) + b_0^2\mathrm{Var}(X) - 2b_0\mathrm{Cov}(X,Y) = \mathrm{Var}(Y)\left[1 - \frac{[\mathrm{Cov}(X,Y)]^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}\right] = \mathrm{Var}(Y)(1 - \rho_{XY}^2).\end{aligned}$$

3.2 Relationship Between Two Random Variables

- Since $\min_{a,b} \text{MSE} = \text{Var}(Y)(1 - \rho_{XY}^2) \geq 0$, so $\rho_{XY}^2 \leq 1 \Rightarrow -1 \leq \rho_{XY} \leq 1$.
  - $0 < \rho_{XY} \leq 1$: positively correlated; $-1 \leq \rho_{XY} < 0$: negatively correlated; $\rho_{XY} = 0$: uncorrelated.
  - When $|\rho_{XY}|$ is closer to 1, the mean squared error is smaller, i.e., the relationship between X and Y is closer to linear. Specifically, if $\rho_{XY} = \pm 1$, X and Y have **an almost perfect linear relationship**.

Not necessarily independent!



3.2 Relationship Between Two Random Variables

Example 3.10

- We would like to invest \$10,000 into shares of companies XX and YY.
- Shares of XX cost \$20 per share and the market analysis shows that the expected return is \$1 per share, with a standard deviation of \$0.5.
- Shares of YY cost \$50 per share, with an expected return of \$2.5 and a SD of \$1.
- What is the optimal portfolio (资产组合) consisting of shares of XX and YY, given their correlation coefficient ρ? (Note: number of shares can be any non-negative real value)

Solution

- Suppose that c dollars are invested into XX and $(10{,}000 - c)$ dollars into YY, the resulting return is R_c.
- Let r.v.s X, Y denote the return per share of XX and YY, respectively. First, consider the expected return:

$$\mathrm{E}(R_c) = \mathrm{E}\left(X \times \frac{c}{20} + Y \times \frac{10000 - c}{50}\right) = 1 \times \frac{c}{20} + 2.5 \times \frac{10000 - c}{50} = \$500.$$

- Therefore, the expected return does not vary with c.

3.2 Relationship Between Two Random Variables

Solution

- Next, consider the variance of the return

$$\begin{aligned}\mathrm{Var}(R_c) &= \mathrm{Var}\left(X\times\frac{c}{20}+Y\times\frac{10000-c}{50}\right)\\ &= \left(\frac{c}{20}\right)^2\mathrm{Var}(X)+\left(\frac{10000-c}{50}\right)^2\mathrm{Var}(Y)+2\left(\frac{c}{20}\right)\left(\frac{10000-c}{50}\right)\mathrm{Cov}(X,Y)\\ &= \left(\frac{c}{20}\right)^2 0.5^2+\left(\frac{10000-c}{50}\right)^2 1^2+2\left(\frac{c}{20}\right)\left(\frac{10000-c}{50}\right)\rho\times 0.5\times 1\\ &= \left(\frac{41-40\rho}{40000}\right)c^2-(8-10\rho)c+40000.\end{aligned}$$

- Minimizing $\mathrm{Var}(R_c)$ w.r.t. c, we have:
  - If $\rho \geq 0.8$, $c=0$, i.e., all \$10000 are invested into YY. In this case, $\mathrm{Var}(R_c)=40000$.
  - If $\rho < 0.8$, $c=40000(4-5\rho)/(41-40\rho)$. In this case

$$\mathrm{Var}(R_c)=40000-40000\times\frac{(4-5\rho)^2}{41-40\rho},$$ specifically, when $\rho=-1$, $\mathrm{Var}(R_c)=0$.

Perfect risk hedging!

3.2 Relationship Between Two Random Variables

- The conditional expectation (条件期望) of a random variable X given another random variable Y is the expected value of X when Y is known, denoted as $E(X|Y)$.

Conditional Expectation

- For a **discrete** random vector (X, Y), the conditional expectation of X given $Y = y$ is

$$E(X \mid Y = y) \triangleq \sum_{k=1}^{\infty} x_k \cdot P(X = x_k | Y = y).$$

- For a **continuous** random vector (X, Y), the conditional expectation of X given $Y = y$ is

$$E(X \mid Y = y) \triangleq \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx,$$

- where $f_{X|Y}(x|y)$ is the conditional PDF of X given $Y = y$.

- $E(X|Y = y)$ is a function of y, and we can denote $g(y) \triangleq E(X|Y = y)$.
- Then $E(X|Y) = g(Y)$, which is a function of Y, not a constant. It is itself a **random variable**.
- If X and Y are **independent**, then $E(X|Y) = E(X)$.

3.2 Relationship Between Two Random Variables

- The Law of Total Expectation (全期望公式) ***connects*** the **unconditional expectation** of X to its **conditional expectation** given Y.

Law of Total Expectation

- If (X, Y) is a random vector, and $E(X)$ exists, then
$$E(X) = E(E(X \mid Y)).$$
- Specifically, if (X, Y) are **discrete** random vector, then
$$E(X) = \sum_j E(X|Y = y_j) \cdot P(Y = y_j).$$
- If (X, Y) are **continuous** random vector, then
$$E(X) = \int_{-\infty}^{\infty} E(X|Y = y) \cdot f_Y(y)\, dy.$$

- $E(X)$ can be computed by first finding the conditional expectation $E(X|Y = y)$, and then taking the expectation of this conditional expectation over all possible values of Y.
- Useful when it is easier to compute $E(X|Y = y)$ than $E(X)$ directly. The whole sample space is split into several subspaces $\{Y = y\}$.

Example 3.11

- A miner (矿工) is trapped in a mine (矿井) with three doors.
- Door 1 leads to a tunnel (隧道) that takes 3 hours to reach a safe area.
- Door 2 leads to a tunnel that takes 5 hours to return to the original location.
- Door 3 leads to a tunnel that takes 7 hours to also return to the original location.
- Assume the miner always chooses one of the three doors with equal probability. What is the average time it will take for him to reach the safe area?

Solution

- Let X be the time in hours it takes for the miner to reach the safe area. The possible values of X are:

$$3, 5+3, 7+3, 5+5+3, 5+7+3, 7+7+3, \cdots$$

- It is difficult to write out the probability distribution of X, so we cannot directly compute $E(X)$.
- Instead, let Y represent the first door chosen, where $\{Y=j\}$ means the miner chooses the door j. Then,

$$P(Y=1)=P(Y=2)=P(Y=3)=\frac{1}{3}.$$

3.2 Relationship Between Two Random Variables

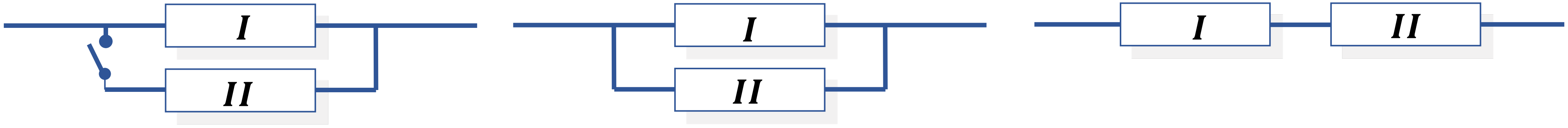
Solution

- If door 1 is chosen, it takes 3 hours to reach the safe area, so
$$E(X \mid Y = 1) = 3.$$
- If door 2 is chosen, it takes 5 hours to return to the original location, so
$$E(X \mid Y = 2) = 5 + E(X).$$
- If door 3 is chosen, it takes 7 hours to return to the original location, so
$$E(X \mid Y = 3) = 7 + E(X).$$
- Thus, by the Law of Total Expectation,
$$E(X) = \sum_j E(X|Y = y_j) \cdot P(Y = y_j)$$
$$= \frac{1}{3}[3 + 5 + E(X) + 7 + E(X)] = 5 + \frac{2}{3}E(X),$$
- Solving the equation, we get $E(X) = 15$.
- Therefore, the miner will take an average of 15 hours to reach the safe area.

Chapter 3 Joint Distributions

3.3 Function of Multiple Random Variables

- In Section 2.4, we talked about how to determine the distribution of some function of a random variable.
- Similarly, we may sometimes know the joint distribution of a random vector, e.g., (X, Y), and would like to derive the distribution of some function of it, e.g., $Z = g(X, Y)$.
- E.g., we want to derive the distribution of the lifespan of a system consists of two components.



Switch to II if I is broken

I and II are connected parallelly

I and II are connected in series

System lifespan = $X + Y$

System lifespan = $\max\{X, Y\}$

System lifespan = $\min\{X, Y\}$

3.3 Function of Multiple Random Variables

The PDF of Z

- Consider the continuous first.
- The most general solution is to derive the CDF of $Z = g(X,Y)$ starting from the definition of CDF:

$$F_Z(z) = P(Z \leq z) = P(g(X,Y) \leq z) \;=\; \iint_{g(x,y)\leq z} f(x,y)dxdy = \cdots = \int_{-\infty}^{z} f_Z(u)du\,.$$

The PDF of $Z = X + Y$ – Continuous Case

- Let $f(x,y)$ be the PDF of random vector (X,Y), $f_X(x)$ and $f_Y(y)$ be the marginal PDF of X and Y, respectively. Then, the PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f(z-y,y)dy = \int_{-\infty}^{\infty} f(x,z-x)dx\,.$$

- Specifically, if X and Y are independent, then

$$f_Z(z) = f_X * f_Y \triangleq \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx\,,$$

- where the two integrals are called the convolution（卷积）of f_X and f_Y, denoted as $f_X * f_Y$.

3.3 Function of Multiple Random Variables

- Here we provide the derivation of the PDF of $Z = X + Y$.

Proof:

$$F_Z(z) = P(Z \le z) = P(X + Y \le z) = \iint\limits_{x+y\le z} f(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f(x,y)dx\right)dy \xlongequal{\text{Let } x = u - y} \int_{-\infty}^{\infty}\int_{-\infty}^{z} f(u-y,y)dudy$$

$$= \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f(u-y,y)dy\right]du \Longrightarrow f_Z(z) = \int_{-\infty}^{\infty} f(z-y,y)dy\,.$$

Similarly, we can show $f_Z(z) = \int_{-\infty}^{\infty} f(x,z-x)dx\,.$

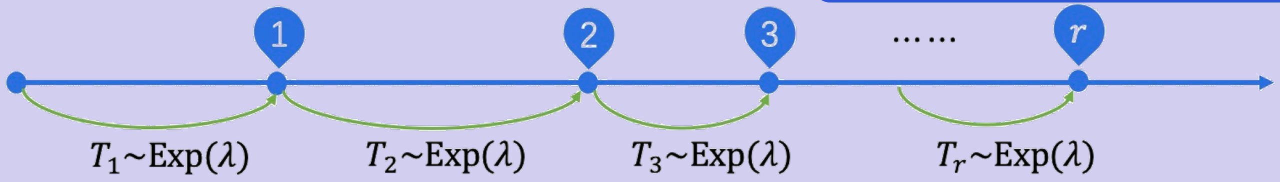
3.3 Function of Multiple Random Variables

Example 3.8 (Continued)

- Let T_1 be the duration from 0 to the arrival of the first particle, T_2 be the duration from the arrival of the first particle to the arrival of the second particle, ...
- Derive the distribution of time until the rth particle arrives.

Solution

- By our previous knowledge, we know that $T_1 \sim \text{Exp}(\lambda), \dots, T_r \sim \text{Exp}(\lambda)$, and $T_1, T_2, \dots, T_r$ are independent.



We say that $T_1, T_2, \dots, T_r \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, independent and identically distributed, abbreviated as i.i.d..

- First consider the case when $r = 2$, i.e., $Z = T_1 + T_2$, then for $z > 0$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{T_1}(t) f_{T_2}(z-t)dt = \int_0^z \lambda e^{-\lambda t} \cdot \lambda e^{-\lambda(z-t)} dt = \lambda^2 e^{-\lambda z} \int_0^z 1 dt = \lambda^2 z e^{-\lambda z}.$$

3.3 Function of Multiple Random Variables

Solution

- Then consider $r = 3$, i.e., $Z = T_1 + T_2 + T_3$, then for $z > 0$:

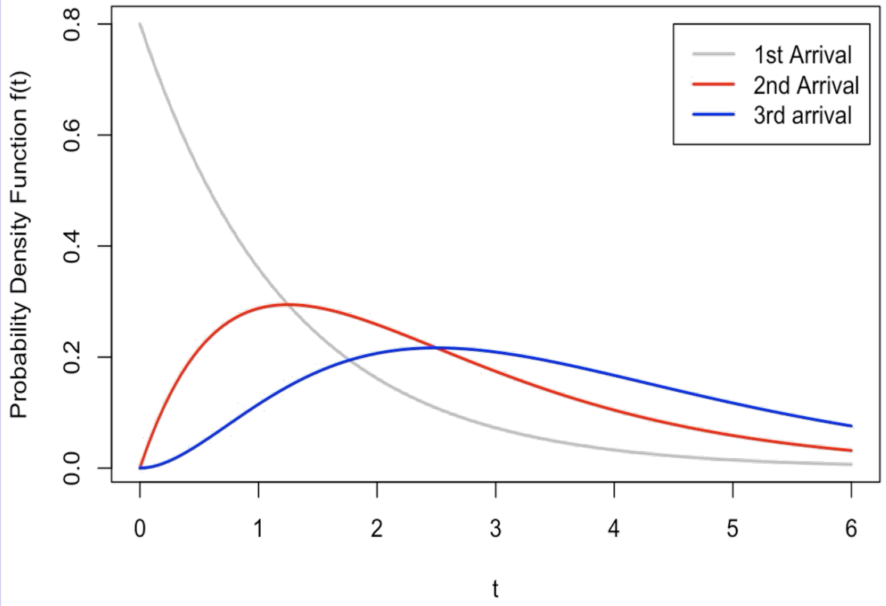
$$f_Z(z) = \int_{-\infty}^{\infty} f_{T_1+T_2}(t) f_{T_3}(z-t)dt = \int_{0}^{z} \lambda^2 t e^{-\lambda t} \cdot \lambda e^{-\lambda(z-t)} dt$$

$$= \lambda^3 e^{-\lambda z} \int_0^z t dt = \frac{\lambda^3 z^2 e^{-\lambda z}}{2}.$$

- Perform the computation recursively, it is not difficult to obtain that the PDF of $Z = T_1 + T_2 + \cdots + T_r$ is

$$f_Z(z) = \begin{cases} \frac{\lambda^r}{(r-1)!} z^{r-1} e^{-\lambda z}, z > 0 \\ 0, \quad \text{otherwise} \end{cases}.$$

- This distribution is known as the Gamma distribution (伽马分布), with parameters r and λ, denoted by $\text{Gamma}(r, \lambda)$.



3.3 Function of Multiple Random Variables

Example 3.12

- Let X and Y be independent standard normal random variables, $T = X + Y$.
- You should quickly be able to determine $\mathrm{E}(T)$ and $\mathrm{Var}(T)$, but what's the distribution of T?

Solution

- Since X and Y are independent, by the convolution formula, we have

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-x)^2}{2}} dx$$

$$= \frac{1}{2\pi} e^{-\frac{t^2}{4}} \int_{-\infty}^{\infty} e^{-\left(x-\frac{t}{2}\right)^2} dx \overset{Let\ u = x - t/2}{=\!=\!=\!=\!=} \frac{1}{2\pi} e^{-\frac{t^2}{4}} \int_{-\infty}^{\infty} e^{-u^2} du$$

$$= \frac{1}{2\pi} e^{-\frac{t^2}{4}} \sqrt{\pi} = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{t^2}{2\left(\sqrt{2}\right)^2}}$$

- This suggest that $T = X + Y \sim \mathrm{N}(0,2)$.

This result can be extended to more general cases

3.3 Function of Multiple Random Variables

General Results about the Sum of Independent Normal Random Variables

- Let X and Y be two independent random variables, $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Then
$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

- More generally, if random variables $X_1, X_2, \dots, X_n$ are independent and $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ $(i = 1, 2, \dots, n)$. Then for constants $a_1, a_2, \dots, a_n$ (not all zero),
$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

- In summary, a linear combination of independent normal random variables still follows a normal distribution.

3.3 Function of Multiple Random Variables

- The discrete case is similar.

The PMF of $Z = X + Y$ – Discrete Case

- Let X and Y be two discrete random variables, for simplicity, assume that the support of X and Y are both $\{0, 1, 2, \dots\}$, then the PMF of $Z = X + Y$ is: $(k = 0,1,2, \dots)$

$$P(Z = k) = \sum_{i=0}^{k} P(X = i, Y = k - i) = \sum_{j=0}^{k} P(X = k - j, Y = j).$$

- Specifically, if X and Y are **independent**, then

$$P(Z = k) = \sum_{i=0}^{k} P(X = i) \cdot P(Y = k - i) = \sum_{j=0}^{k} P(X = k - j) \cdot P(Y = j),$$

- which is the **discrete convolution formula (离散卷积公式)**.

- For $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$, and X, Y are independent, then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$ can be shown accordingly. Actually, Example 3.7 serves as an example.
- Therefore, sum of independent Poisson random variables still follows a Poisson distribution. (Linear combination of independent Poisson r.v.? No. Sum of Uniform distribution? No.)

3.3 Function of Multiple Random Variables

- The sum of random variables $S_n = X_1 + X_2 + \cdots + X_n$ appear in many real-life problems, however, determining the exact distribution of S_n is not an easy task generally.
- Each time we add one more random variable, we have to calculate a convolution. What if we work with the sum of hundreds of random variables? Calculating many convolutions is impractical.
- It would be great if there is an approximated distribution of S_n that is accurate and easy to use.
- The **Central Limit Theorem (CLT, 中心极限定理)** provides such an approximation.

Central Limit Theorem for i.i.d. Random Variables

- $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with $\mu \triangleq \mathrm{E}(X_i)$ and $\sigma^2 \triangleq \mathrm{Var}(X_i)$. Let $S_n = X_1 + X_2 + \cdots + X_n$ and $\bar{X}_n = S_n/n$, consider the standardized version of S_n:

$$Z_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

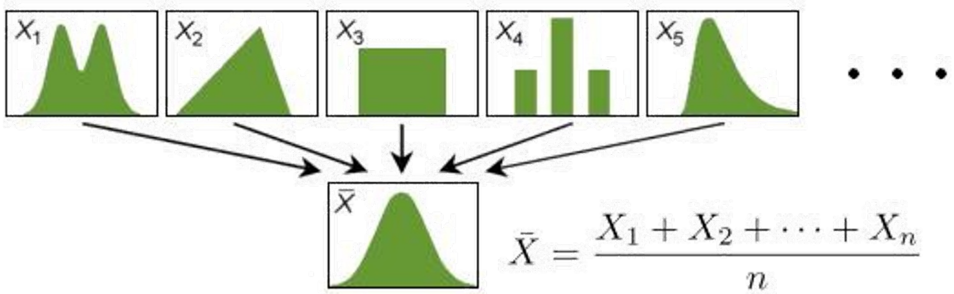
$\xrightarrow{n\to\infty}$ would be "$=$" if $X_1, \ldots, X_n$ are normal r.v.s, even for small n.

- As $n \to \infty$, Z_n **converges in distribution (依分布收敛)** to a **standard normal** random variable, that is:

$$F_{Z_n}(z) = P(Z_n \le z) \xrightarrow{n\to\infty} \Phi(z) \quad \text{for all } z.$$

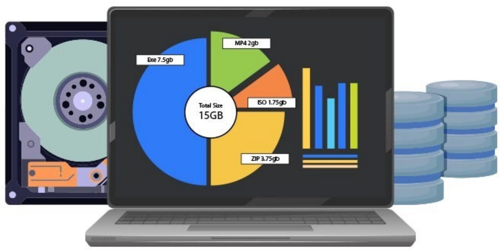
3.3 Function of Multiple Random Variables

- The CLT does not require $X_1, X_2, \dots, X_n$ to follow any specific distribution, so it is a universal behavior across different probability distributions with finite expectation and variance.
- The further research findings are surprising: even if $X_1, X_2, \dots, X_n$ are not independent and do not follow the same distribution, the CLT still holds. Details are omitted here.
- All the complexity and chaos are dissolved under the mysterious curve of the normal distribution.
- Initially, mathematicians refer to this theorem as the Limit Theorem. However, due to its importance in probability theory, the word "central" was added.



- The CLT explains why many measures in reality are normally distributed: they are typically the combined effect of multiple factors.
- How large n should be to apply the CLT? The rule of thumb (经验法则) is $n \geq 30$.

3.3 Function of Multiple Random Variables



Example 3.13

- A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1Mb with a standard deviation of 0.5Mb?

Solution

- We have $n = 300$, $\mu = 1$, $\sigma = 0.5$. As n is large, so the CLT applies to their total size S_n.
- Therefore, the probability of sufficient space is

$$P(S_n \leq 330) = P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq \frac{330 - 300\times 1}{0.5\sqrt{300}}\right) \approx \Phi(3.46) = 0.9997.$$

- This probability is very high, hence, the available disk space is very likely to be sufficient.

3.3 Function of Multiple Random Variables

- The binomial variable represent a special case of $S_n = X_1 + \cdots + X_n$, where $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\,(p)$.
- In this case, the exact distribution of S_n is $\text{Binomial}(n,p)$, and consider the approximated distribution of S_n applying the CLT:

$$\frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}} \overset{\text{approx.}}{\sim} \mathrm{N}(0,1) \Longrightarrow S_n \overset{\text{approx.}}{\sim} \mathrm{N}\big(np, np(1-p)\big).$$

- This suggests that the binomial distribution $\text{Binomial}(n,p)$ can be approximated by the normal distribution $\mathrm{N}(\mu,\sigma^2)$, where $\mu = np, \sigma = \sqrt{np(1-p)}$.

Galton board(高尔顿钉板)

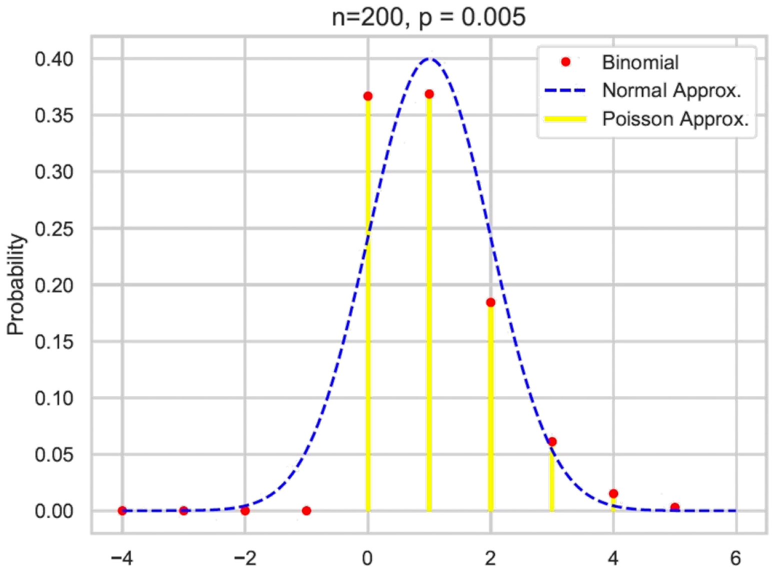
- This is called the **normal approximation to binomial distribution (二项分布的正态近似)**.
- Recall that we talked about the Poisson theorem, which is about the Poisson approximation to binomial distribution (see the PPT of Chapter 2, Page 28-29).
- **Question:** what's the relationship between these two approximations?

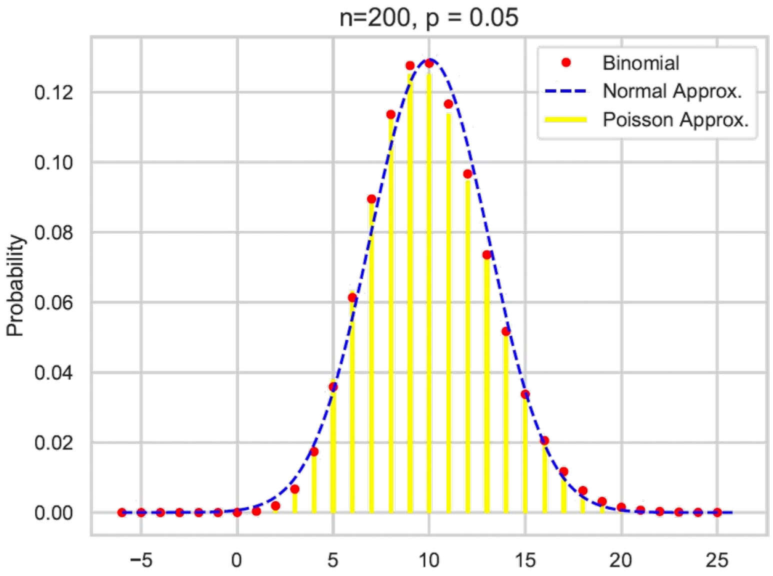
3.3 Function of Multiple Random Variables

Poisson approximation	Normal approximation
▪ Binomial(n, p) can be approximated by Poisson(np). ▪ The approximation works well when n is large and p is small, e.g., $n > 100$ and $p < 0.05$.	▪ Binomial(n, p) can be approximated by $\mathrm{N}\big(np, np(1-p)\big)$. ▪ The approximation works well when $np \geq 5$ and $n(1-p) \geq 5$.

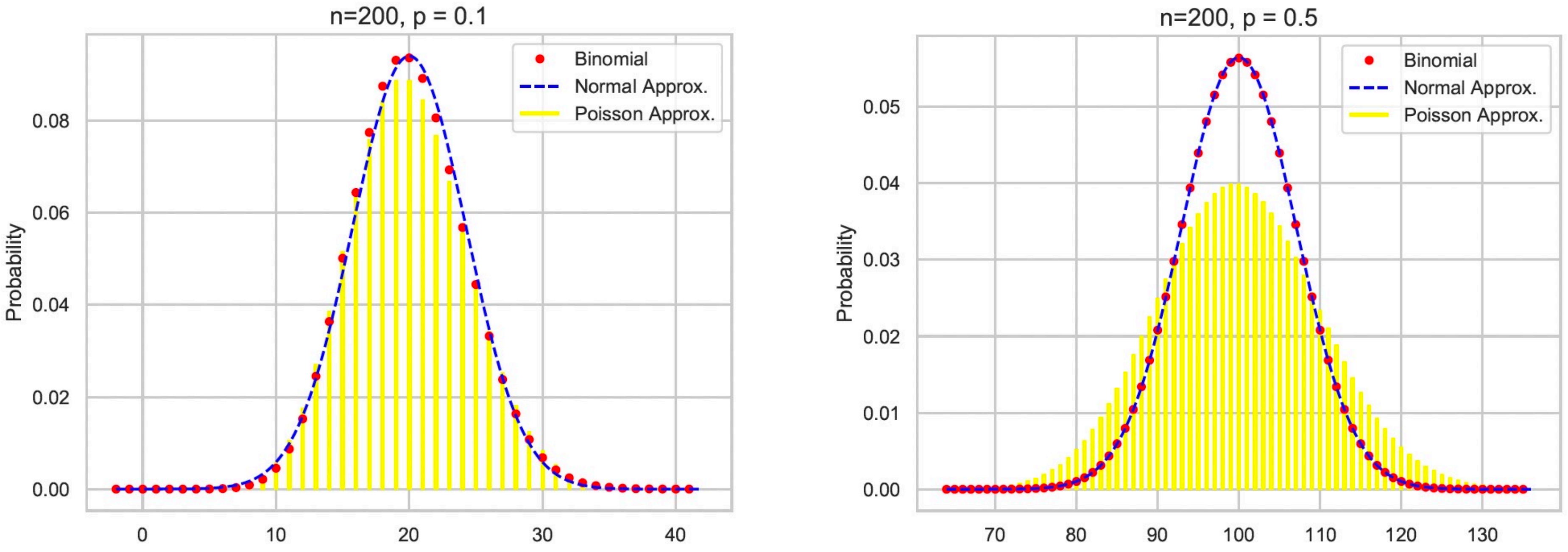
- The best way to understand this is to visualize the three distributions in Python.

https://jupyter.org/try-jupyter/lab/





3.3 Function of Multiple Random Variables



- We see that when p is small, the Poisson approximation is better (for a given small value of p, we need larger n for the normal approximation), while the normal approximation is better for large p.
- It is not surprising that the Poisson approximation works poorly for large p if we consider the variance of $\text{Binomial}(n, p)$ and $\text{Poisson}(np)$.

Variance $np(1-p)$

Variance np

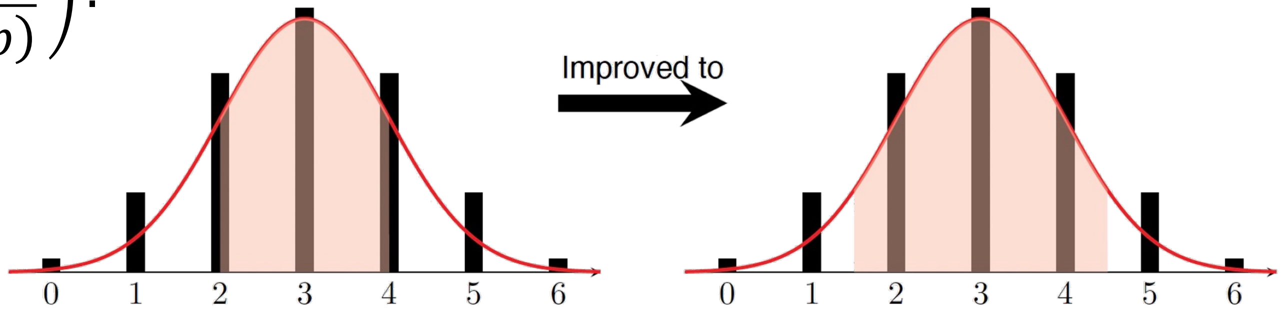
$np(1-p) \approx np$
only when p is small

3.3 Function of Multiple Random Variables

- If $X \sim \text{Binomial}(n, p)$ and we want to calculate $P(k \le X \le l)$ (k, l are integers) with the normal approximation, note that a **continuity correction (连续性修正)** needs to be applied.
- This correction is needed when we approximate a discrete distribution by a continuous one.
- The essential reason why a correction is needed is that $P(X = x)$ may be positive if X is discrete, whereas it is always 0 for continuous X.
- The continuity correction is to expand the interval by 0.5 in each direction:

$$P(k \le X \le l) = P(k - 0.5 \le X \le l + 0.5) = P\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}} \le \frac{X - np}{\sqrt{np(1-p)}} \le \frac{l + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{l + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right).$$



3.3 Function of Multiple Random Variables

Example 3.14

- A new computer virus attacks a folder consisting of 200 files.
- Each file gets damaged with probability 0.2 independently of other files.
- What is the probability that fewer than 50 files get damaged?

Solution

- Let X denote the number of files get damaged, then $X \sim \text{Binomial}(200, 0.2)$.
- Since $p = 0.2 > 0.05$, we would apply the normal approximation with the continuity correction:

$$P(X < 50) = P(X \leq 49) = P(X \leq 49.5) = P\left(\frac{X - 200\times0.2}{\sqrt{200\times0.2\times0.8}} \leq \frac{49.5 - 200\times0.2}{\sqrt{200\times0.2\times0.8}}\right)$$

$$\approx \Phi\left(\frac{49.5 - 40}{5.657}\right) \approx \Phi(1.68) = 0.9535.$$

- Notice that the properly applied continuity correction is $P(X \leq 49.5)$ instead of $P(X \leq 50.5)$, because the problem is asking for "the probability that fewer than 50 files get damaged".

3.3 Function of Multiple Random Variables

- Up to this point, we have been talking about the sum of multiple random variables.
- In the following, we consider how to determine the distribution of the maximum/minimum of two random variables, the result can be generalized to multiple random variables.

The CDF of $\max(X,Y)$ and $\min(X,Y)$ - Continuous Case

- Let $f(x,y)$ be the PDF of random vector (X,Y). Then, the CDFs of $\max(X,Y)$ and $\min(X,Y)$ are

$$F_{\max}(z) = \int_{-\infty}^{z}\int_{-\infty}^{z} f(x,y)dx\,dy, \quad F_{\min}(z) = 1 - \int_{z}^{\infty}\int_{z}^{\infty} f(x,y)dx\,dy.$$

- Let $F_X(x)$ and $F_Y(y)$ be the marginal CDF of X and Y, then if X and Y are independent, we have

$$F_{\max}(z) = F_X(z)F_Y(z), \quad F_{\min}(z) = 1 - [1 - F_X(z)][1 - F_Y(z)].$$

- Specifically, if $X_1, X_2, \dots, X_n \overset{\text{i.i.d.}}{\sim} F(x)$ with PDF $f(x)$, then the CDFs and PDFs of $\max(X_1, X_2, \dots, X_n)$ and $\min(X_1, X_2, \dots, X_n)$ are

$$F_{\max}(z) = [F(z)]^n, f_{\max}(z) = nf(z)[F(z)]^{n-1},$$

$$F_{\min}(z) = 1 - [1 - F(z)]^n, f_{\min}(z) = nf(z)[1 - F(z)]^{n-1}.$$

3.3 Function of Multiple Random Variables

- Here we provide the derivation of the CDF of $\max(X,Y)$ and $\min(X,Y)$.

Proof:

$$F_{\max}(z) = P(\max(X,Y) \le z) = P(X \le z, Y \le z) = \int_{-\infty}^{z}\int_{-\infty}^{z} f(x,y)dx\,dy.$$

$$F_{\min}(z) = P(\min(X,Y) \le z) = 1 - P(\min(X,Y) > z)$$

$$= 1 - P(X > z, Y > z) = 1 - \int_{z}^{\infty}\int_{z}^{\infty} f(x,y)dx\,dy$$

When X and Y are independent,

$$F_{\max}(z) = P(X \le z) \cdot P(Y \le z) = F_X(z)F_Y(z),$$

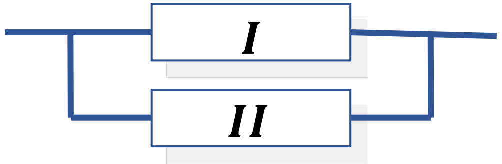
$$F_{\min}(z) = 1 - P(X > z) \cdot P(Y > z) = 1 - [1 - F_X(z)][1 - F_Y(z)].$$

- **Suggestion**: don't just memorize the resulting formulas, try to understand the process of derivation.

Example 3.15

- A system is made up of two independent components I and II, with lifespan $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$, respectively.
- Calculate the expected lifespan of the system in these two scenarios:

(1) I and II are connected in series; (2) I and II are connected parallelly.



Solution

- (1) In this case, the lifespan of the system is $Z = \min(X_1, X_2)$, the CDF of Z is

$$F_Z(z) = 1 - [1 - F_{X_1}(z)][1 - F_{X_2}(z)] = \begin{cases} 1 - e^{-(\lambda_1+\lambda_2)z}, z > 0 \\ 0, \quad \text{otherwise} \end{cases}.$$

- This suggest that $Z \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$, so that $\mathrm{E}(Z) = 1/(\lambda_1 + \lambda_2)$.
- It's not difficult to find that $\mathrm{E}(Z) < \mathrm{E}(X_1)$ and $\mathrm{E}(Z) < \mathrm{E}(X_2)$, so the expected lifespan of the system is shorter than that of any single component.

Solution

- (2) In this case, the lifespan of the system is $Z = \max(X_1, X_2)$, the CDF and PDF of Z is

$$F_Z(z) = F_{X_1}(z)F_{X_2}(z) = \begin{cases} \left(1 - e^{-\lambda_1 z}\right)\left(1 - e^{-\lambda_2 z}\right), z > 0 \\ 0, \qquad \text{otherwise} \end{cases},$$

$$\Rightarrow f_Z(z) = F_Z'(z) = \begin{cases} \lambda_1 e^{-\lambda_1 z} + \lambda_2 e^{-\lambda_2 z} - (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)z}, z > 0 \\ 0, \qquad \text{otherwise} \end{cases}.$$

- Then, we can obtain $\mathrm{E}(Z)$ by definition:

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} z f_Z(z) dz = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}.$$

- It is not difficult to find that $\mathrm{E}(Z) > \mathrm{E}(X_1)$, $\mathrm{E}(Z) > \mathrm{E}(X_2)$ and $\mathrm{E}(Z) < \mathrm{E}(X_1) + \mathrm{E}(X_2)$, so the expected lifespan of the system is longer than that of any single component but shorter than their sum.

3.3 Function of Multiple Random Variables

- Besides the sum, maximum, minimum functions, the distribution of other functions of multiple random variables can also be derived starting from the definition of CDF and do the integration.

Example 3.16

- X and Y are independent random variables and both follow the distribution $\mathrm{Exp}(1)$.
- Derive the PDF of $Z = X/Y$.

Solution

- The joint PDF of X and Y is $f(x,y) = \begin{cases} e^{-(x+y)}, x,y > 0 \\ 0,\ \text{otherwise} \end{cases}$.
- For any $z > 0$, consider the CDF $F_Z(z)$ of Z:

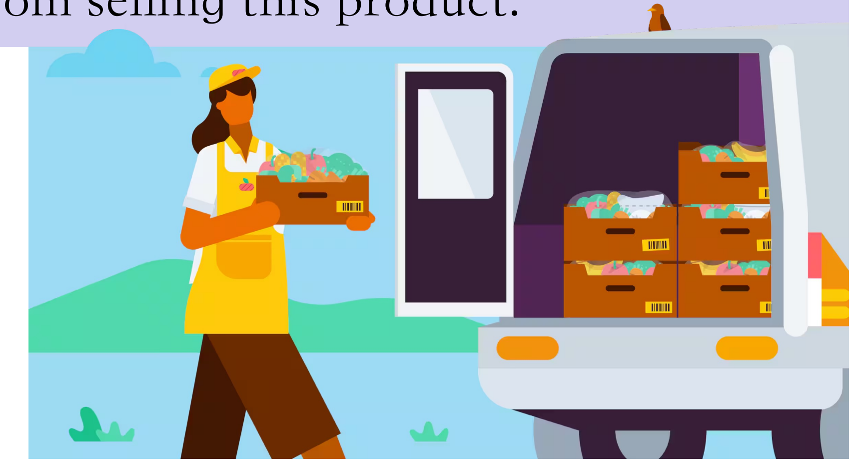
$$F_Z(z) = P\left(\frac{X}{Y} \le z\right) = \underset{\{x,y>0,\,x/y\le z\}}{\iint} e^{-(x+y)} dxdy = \int_0^\infty \left(\int_0^{yz} e^{-(x+y)} dx\right) dy = \int_0^\infty e^{-y}(1 - e^{-yz}) dy = 1 - \frac{1}{1+z}.$$

$$\Rightarrow f_Z(z) = F_Z'(z) = \begin{cases} (1+z)^{-2}, z > 0 \\ 0, \quad\quad \text{otherwise} \end{cases}.$$

3.3 Function of Multiple Random Variables

Example 3.17

- A store sells a certain product, where the weekly stock (进货量) and customer demand are independent random variables, both uniformly distributed over the interval $(10, 20)$.
- The store earns a profit of $1,000 for each unit of the product sold.
- However, if the demand exceeds the stock, the store can order the product from other stores, earning a profit of $500 per unit in such cases.
- Please calculate the store's expected weekly profit from selling this product.



3.3 Function of Multiple Random Variables

Solution

- Let X be the weekly stock and Y be the weekly customer demand. Then the joint PDF of X and Y is

$$f(x,y)=\begin{cases}1/100, & 10<x,y<20\\ 0, & \text{otherwise}\end{cases}.$$

- Let Z be the weekly profit of the store from selling this product, then Z must be a function of X and Y, i.e., $Z=g(X,Y)$. By the description of the problem, we have

$$g(x,y)=\begin{cases}1000y, & \text{if } y\le x\\ 1000x+500(y-x), & \text{if } y>x\end{cases}=\begin{cases}1000y, & \text{if } y\le x\\ 500(x+y), & \text{if } y>x\end{cases}.$$

- Therefore,

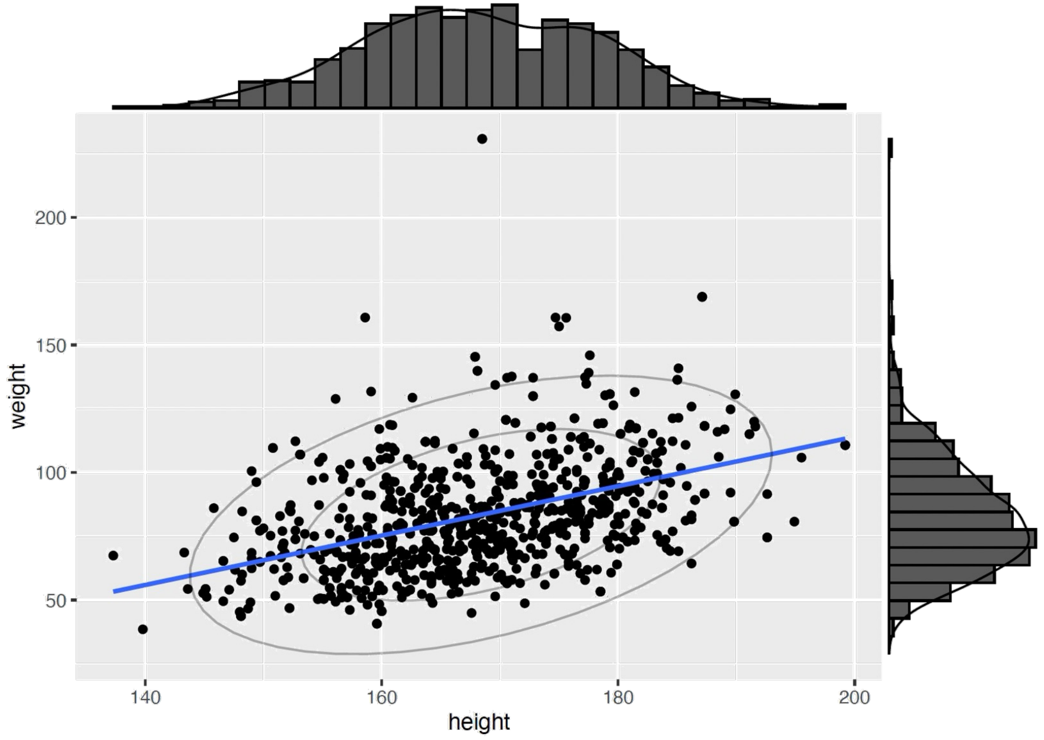
$$\mathrm{E}(Z)=\mathrm{E}[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)f(x,y)dx\,dy=\iint_{y\le x}1000yf(x,y)dxdy+\iint_{y>x}500(x+y)f(x,y)dxdy$$

$$=10\int_{10}^{20}\left(\int_{y}^{20}ydx\right)dy+5\int_{10}^{20}\left(\int_{10}^{y}(x+y)dx\right)dy=\frac{20000}{3}+5\times1500\approx 14166.67.$$

Chapter 3 Joint Distributions

3.4 Multivariate Normal Distribution

- In this section, we will talk about the **bivariate normal distribution (二元正态分布)**, and the results can be generalized to **multivariate normal distribution (多元正态分布)**.
- The bivariate normal/Gaussian distribution is commonly used to model the joint distribution of two normal random variables, particularly when they have some degree of linear relationship.
- Real-world examples:
  - Height and weight of adults
  - Father and son's heights
  - Test scores in two courses



3.4 Multivariate Normal Distribution

The Bivariate Normal Distribution

- Random vector (X, Y) is said to be bivariate normally distributed with means μ_X, μ_Y and variances σ_X^2 and σ_Y^2, and with correlation coefficient ρ, if the joint PDF of (X, Y) is given by $(-\infty < x, y < \infty)$

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right).$$

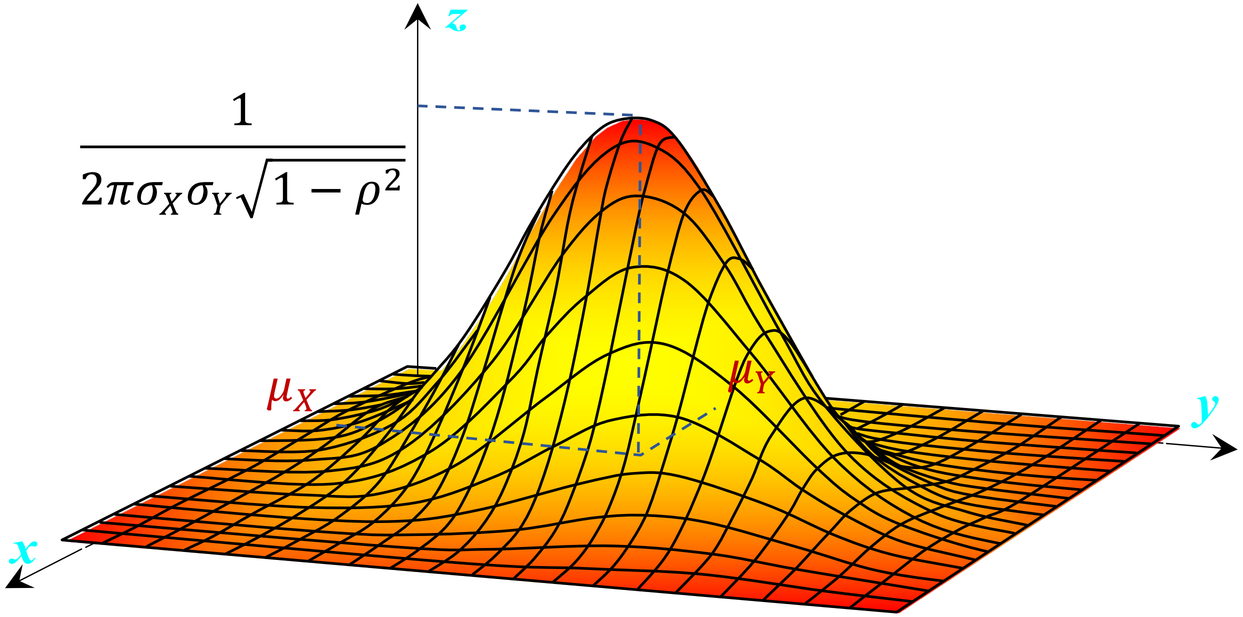
- It can be expressed as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

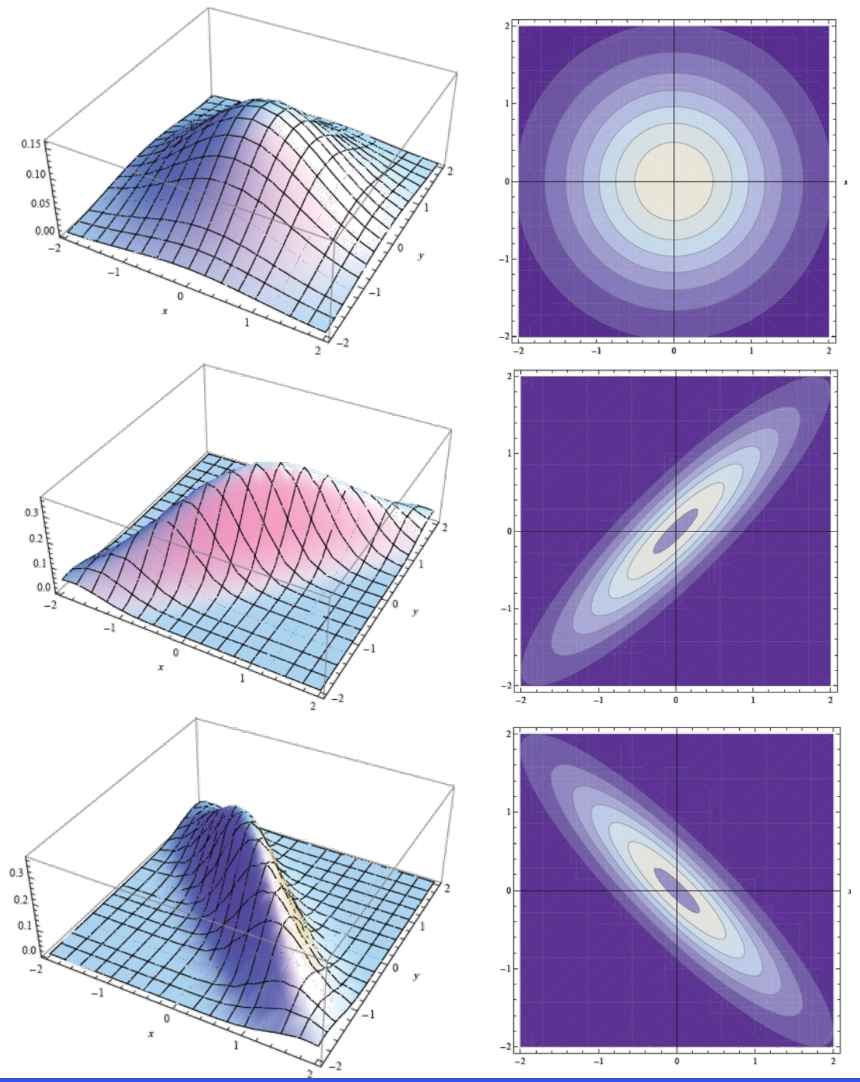
- where $\boldsymbol{\mu}$ is the **mean vector (均值向量)** and $\boldsymbol{\Sigma}$ is the **variance-covariance matrix (方差-协方差矩阵)**.
- Specifically, if $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\rho = 0$, then it is said to be a standard bivariate normal distribution, i.e., $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

- For n-dimensional random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$, mean vector $\mathrm{E}(\boldsymbol{X}) = (E(X_1), E(X_2), \ldots, E(X_n))'$,
- Covariance matrix $\mathrm{Cov}(\boldsymbol{X}) \triangleq \mathrm{E}\left[\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)'\right]$, e.g. in 2-d $= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{pmatrix}$.

3.4 Multivariate Normal Distribution



$\mu_X = \mu_Y = 0, \sigma_X = \sigma_Y = 1$



$\rho = 0$

$\rho = 0.9$

$\rho = -0.9$

3.4 Multivariate Normal Distribution

Marginal distributions are normal distributions.

- Property 1: It can be shown that if $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, then $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, and $\rho_{XY} = \rho$.

Proof:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sqrt{1-\rho^2}}\right)^2\left(\frac{y-\mu_Y}{\sigma_Y} - \rho\frac{x-\mu_X}{\sigma_X}\right)^2} dy$$

Let $t = \frac{1}{\left(\sqrt{1-\rho^2}\right)}\left(\frac{y-\mu_Y}{\sigma_Y} - \rho\frac{x-\mu_X}{\sigma_X}\right) \Rightarrow dy = \sigma_Y\sqrt{1-\rho^2}\,dt$

$$f_X(x) = \frac{1}{2\pi\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{2\pi\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \cdot \sqrt{2\pi} = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

By variable substitution similar to Example 3.9, details omitted.

$\therefore X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$ follows similarly.

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)dxdy = \rho\sigma_X\sigma_Y.$$

$$\Rightarrow \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = \rho.$$

3.4 Multivariate Normal Distribution

- Property 2: Generally, if random variables X and Y are uncorrelated, then we not necessarily have X and Y are independent.

X and Y are independent $\;\Rightarrow\;$ X and Y are uncorrelated

X and Y are independent $\;\not\Leftarrow\;$ X and Y are uncorrelated

- However, uncorrelated does imply independent if X and Y jointly follow a bivariate normal distribution.

Proof: If (X, Y) follow a bivariate normal distribution and they are uncorrelated, i.e., $\rho = \rho_{XY} = 0$, then the joint PDF is

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right).$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left(-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X}e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \times \frac{1}{\sqrt{2\pi}\sigma_Y}e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} = f_X(x)f_Y(y). \qquad \therefore X \text{ and } Y \text{ are independent.}$$

3.4 Multivariate Normal Distribution

Marginal distributions cannot uniquely determine the joint distribution!

Example 3.18

- Let r.v. $X \sim \mathrm{N}(0,1)$ and Z be a r.v. independent of X with PMF $\mathrm{P}(Z=1)=\mathrm{P}(Z=-1)=0.5$.
- Define $Y = ZX$, (1) show that $Y \sim \mathrm{N}(0,1)$; (2) show that X and Y are uncorrelated.

Solution

- (1) Consider the CDF of Y:

$$\begin{aligned} F_Y(y) = \mathrm{P}(Y \le y) &= \mathrm{P}(ZX \le y) \\ &= \mathrm{P}(X \le y \mid Z=1)\mathrm{P}(Z=1) + \mathrm{P}(X \ge -y \mid Z=-1)\mathrm{P}(Z=-1) \\ &= 0.5\Phi(y) + 0.5[1-\Phi(-y)] = \Phi(y). \end{aligned}$$

$\Phi(-y) = 1 - \Phi(y)$

- Therefore, $Y \sim \mathrm{N}(0,1)$.

X and Z are independent

$\mathrm{E}(X) = 0, \mathrm{E}(Y) = 0$

- (2) To show that X and Y are uncorrelated, calculate $\mathrm{Cov}(X,Y) = \mathrm{E}\big[\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\big] = \mathrm{E}(XY)$:

$$\mathrm{E}(XY) = \mathrm{E}(ZX^2) = \mathrm{E}(Z)\mathrm{E}(X^2) = 0.$$

- Therefore, X and Y are uncorrelated. However, it is obvious that X and Y are not independent.
- **Question:** doesn't zero correlation imply independence under the case of normal random variables?

3.4 Multivariate Normal Distribution

Conditional distributions are normal distributions.

- Property 3: Moreover, if $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, then

$$X|Y = y \sim \mathrm{N}\left(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y), (1-\rho^2)\sigma_X^2\right), \ \ Y|X = x \sim \mathrm{N}\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X), (1-\rho^2)\sigma_Y^2\right).$$

Proof:

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

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \Big/ \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{\rho^2(y-\mu_Y)^2}{\sigma_Y^2}\right]\right)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} \exp\left(-\frac{1}{2(1-\rho^2)\sigma_X^2}\left[(x-\mu_X) - \frac{\rho\sigma_X(y-\mu_Y)}{\sigma_Y}\right]^2\right)$$

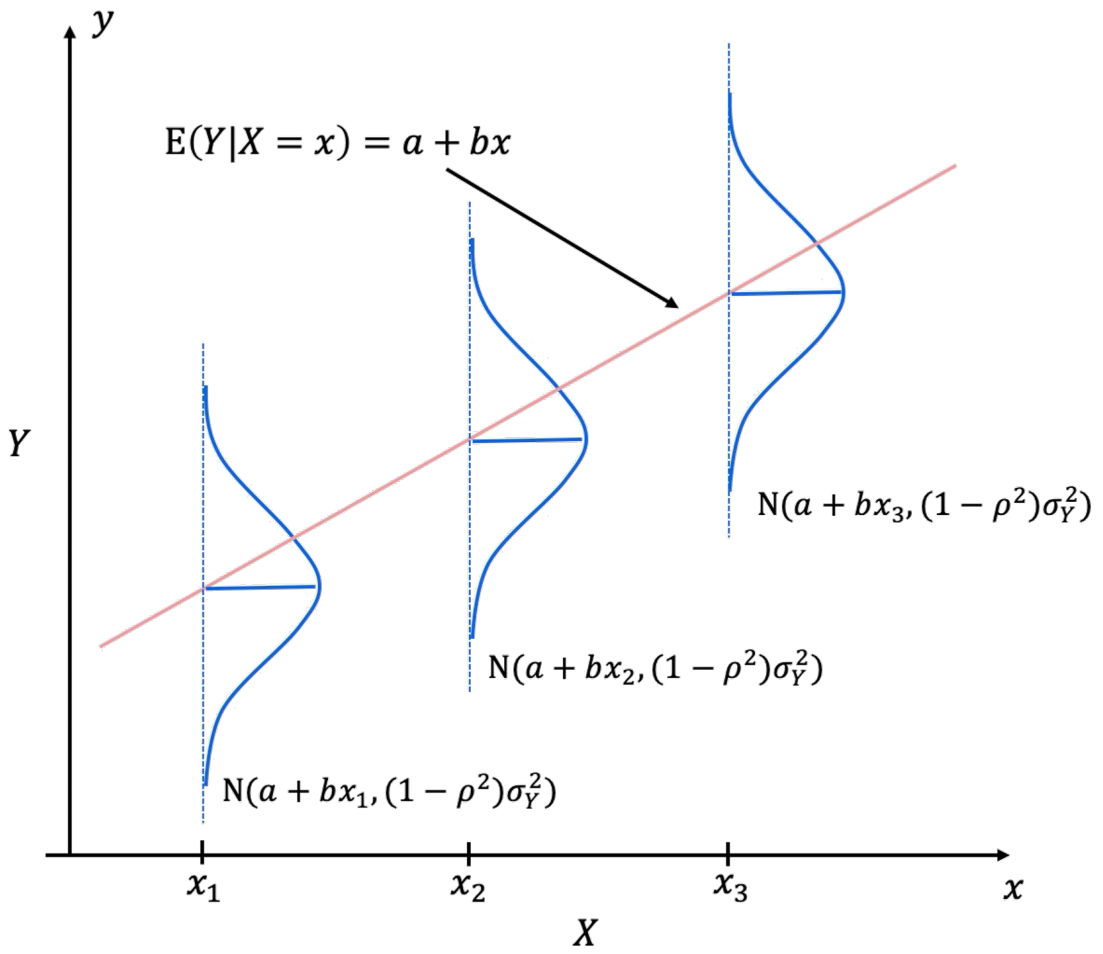
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} \exp\left(-\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x - \left(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y)\right)\right]^2\right)$$

$$\mathrm{E}(Y|X=x) = a + bx$$

$$b = \rho\frac{\sigma_Y}{\sigma_X}, a = \mu_Y - b\mu_X$$

3.4 Multivariate Normal Distribution

- Graphical illustration of the conditional distribution of Y given $X = x$.



- Y follows a normal distribution given any value of X.
- The mean of the normal distribution is a linear function of the value of X.
- These normal distributions have different means but **the same variance**.

3.4 Multivariate Normal Distribution

- Property 4: If $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$), then for any constants $c_1, c_2 \in \mathbb{R}$:

$$c_1X + c_2Y \sim \mathrm{N}(c_1\mu_X + c_2\mu_Y, c_1^2\sigma_X^2 + 2c_1c_2\rho\sigma_X\sigma_Y + c_2^2\sigma_Y^2)$$

Linear combinations still follow normal distributions.

Note: On Page 45, we provide the conclusion that a linear combination of independent normal random variables still follow a normal distribution.

The statement above provide a more general conclusion which does not require independence between the normal random variables, but require that their joint distribution is a bivariate (or. multivariate) normal distribution.

Proof of the statement is omitted here.

- More generally, for any real matrix $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, we have:

$$\boldsymbol{A}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} a_{11}X + a_{12}Y \\ a_{21}X + a_{22}Y \end{pmatrix} \sim \mathrm{N}(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^\top).$$

3.4 Multivariate Normal Distribution

Example 3.19

- Let X and Y denote the math score and the verbal score on the ACT college entrance exam of a randomly selected student.
- Previous history suggest that X and Y are bivariate normally distributed with means $\mu_X = \mu_Y = 22.7$, variances $\sigma_X^2 = 17.64$ and $\sigma_Y^2 = 12.25$, and correlation coefficient $\rho = 0.78$.
- Calculate:
- (1) The probability that a randomly selected student's math score is greater than 25?
- (2) The probability that a randomly selected student's math score is greater than 25 given that his/her verbal score is 25?
- (3) The probability that a randomly selected student has combined math and verbal score greater than 50?
- (4) The probability that a randomly selected student's math score is higher than his/her verbal score given that he/she has combined math and verbal score 50.

3.4 Multivariate Normal Distribution

Solution

- (1) According to the description, the math score $X \sim \mathrm{N}(22.7, 17.64)$, so

$$\mathrm{P}(X > 25) = \mathrm{P}\left(\frac{X - 22.7}{\sqrt{17.64}} > \frac{25 - 22.7}{\sqrt{17.64}}\right) \approx 1 - \Phi(0.55) = 0.2912.$$

- (2) By the property of bivariate normal distribution, the conditional distribution of $X|Y = 25$ is

$$\mathrm{N}\left(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y), (1 - \rho^2)\sigma_X^2\right) = \mathrm{N}\left(22.7 + \frac{0.78\sqrt{17.64}}{\sqrt{12.25}}(25 - 22.7), (1 - 0.78^2)17.64\right) \approx \mathrm{N}(24.85, 6.91)$$

$$\Longrightarrow \mathrm{P}(X > 25|Y = 25) = \mathrm{P}\left(Z > \frac{25 - 24.85}{\sqrt{6.91}}\right) \approx 1 - \Phi(0.06) = 0.4761.$$

- (3) Since X and Y are jointly bivariate normally distributed, linear combinations of X and Y still follow normal distribution, so

$$X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2) = \mathrm{N}(45.4, 52.822).$$

$$\Longrightarrow \mathrm{P}(X + Y > 50) = \mathrm{P}\left(Z > \frac{50 - 45.4}{\sqrt{52.822}}\right) \approx 1 - \Phi(0.63) = 0.2643.$$

3.4 Multivariate Normal Distribution

Solution

- (4) According to the description, we would like to calculate

$$P(X > Y | X + Y = 50) = P(X - Y > 0 | X + Y = 50).$$

- Therefore, we first determine the joint distribution of $W_1 = X - Y$ and $W_2 = X + Y$:

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} = \begin{pmatrix} X - Y \\ X + Y \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 22.7 \\ 22.7 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 17.64 & 11.466 \\ 11.466 & 12.25 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{\mathsf{T}} \right)$$

$$= \mathrm{N}\left(\begin{pmatrix} 0 \\ 45.4 \end{pmatrix}, \begin{pmatrix} 6.958 & 5.39 \\ 5.39 & 52.822 \end{pmatrix} \right)$$

This matrix can also be calculated with $\mathrm{Var}(W_1), \mathrm{Var}(W_2), \mathrm{Cov}(W_1, W_2)$

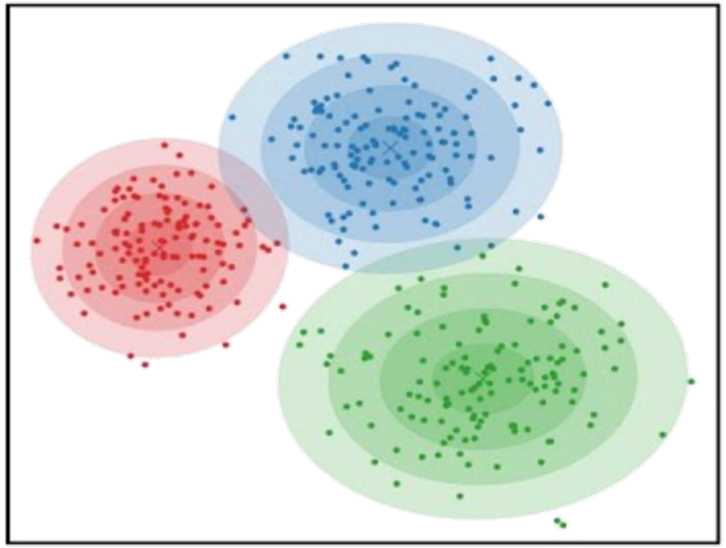
- The correlation coefficient between W_1 and W_2 is $\rho_{12} = 5.39/\sqrt{6.958 \times 52.822} \approx 0.281$.
- Then the conditional distribution of W_1 given $W_2 = w_2 = 50$ is

$$\mathrm{N}\left(\mu_1 + \frac{\rho_{12}\sigma_1}{\sigma_2}(w_2 - \mu_2), (1 - \rho_{12}^2)\sigma_1^2 \right) = \mathrm{N}\left(0 + \frac{0.281\sqrt{6.958}}{\sqrt{52.822}}(50 - 45.4), (1 - 0.281^2)6.958 \right) \approx \mathrm{N}(0.47, 6.41)$$

$$\Longrightarrow P(W_1 > 0 | W_2 = 50) = P\left(Z > \frac{0 - 0.47}{\sqrt{6.41}} \right) \approx P(Z < 0.185) = \frac{0.5714 + 0.5753}{2} \approx 0.5734.$$

3.4 Multivariate Normal Distribution

- Finally, we talk about an application of the multivariate normal distribution in machine learning, the **Gaussian Mixture Model (GMM, 高斯混合模型)**.
- GMM is a machine learning method used to determine the probability each data point belongs to a given cluster. It is a clustering method (聚类方法) used in unsupervised learning (非监督学习).
- Under GMM, the dataset is modeled as a mixture of several multivariate Gaussian distributions, assuming that individuals of different clusters come from different Gaussian distributions.
  - For a randomly selected individual, let Y be its cluster ID, which takes value in $\{1, 2, \ldots, K\}$, $\boldsymbol{X}$ be its feature vector, then $\boldsymbol{X}|Y = k \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
  - The goal is to infer $\mathrm{P}(Y = k|\boldsymbol{X} = \boldsymbol{x})$.
- The multivariate normal distribution has wide applications in pattern recognition, computer vision, natural language processing, signal processing, finance, and economics, etc.



End of Chapter 3