工程概率统计

Probability and Statistics for Engineering

第三章 联合分布

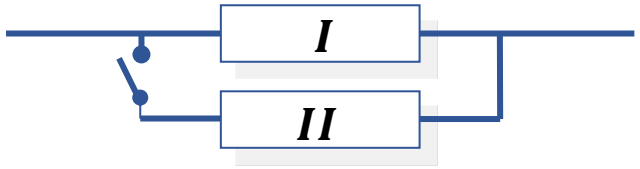
Chapter 3 Joint Distributions

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

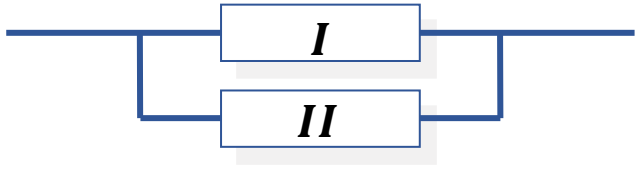
3.3 Function of Multiple Random Variables

- 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

System lifespan = $X + Y$



I and II are connected parallelly

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

I and II are connected in series

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

3.3 Function of Multiple Random Variables

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

The PDF of Z

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

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:

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

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

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

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

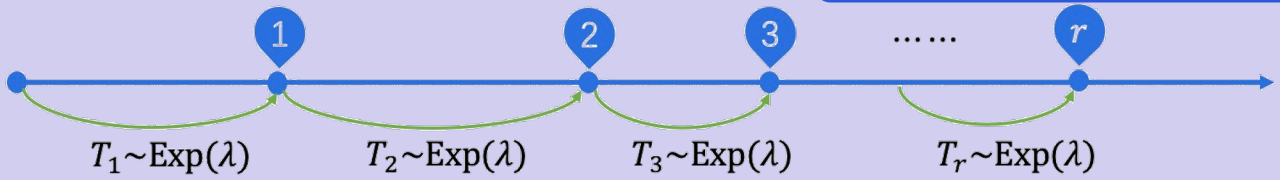
3.3 Function of Multiple Random Variables

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 $T_1 \sim \text{Exp}(\lambda), \ldots, T_r \sim \text{Exp}(\lambda)$, and $T_1, T_2, \ldots, T_r$ are independent.



We say that $T_1, T_2, \ldots, T_r \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, 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 $z > 0$:

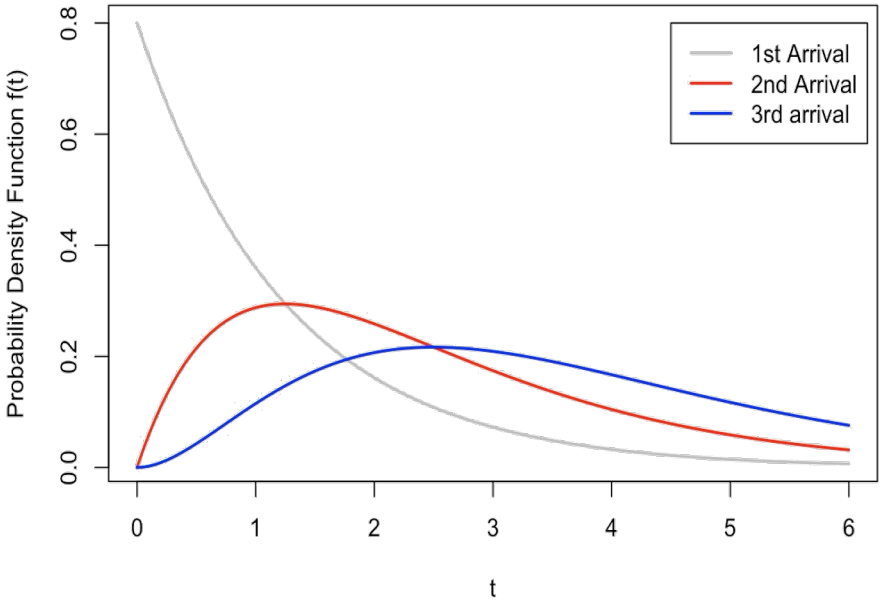
$$f_Z(z) = \int_{-\infty}^{\infty} f_{T_1+T_2}(t) f_{T_3}(z-t)dt = \int_0^z \lambda^2 t e^{-\lambda t} \cdot \lambda e^{-\lambda(z-t)} dt$$

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

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

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

- 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

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

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

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

- 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 $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ $(i = 1, 2, \ldots, n)$. Then for constants $a_1, a_2, \ldots, a_n$ (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:

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

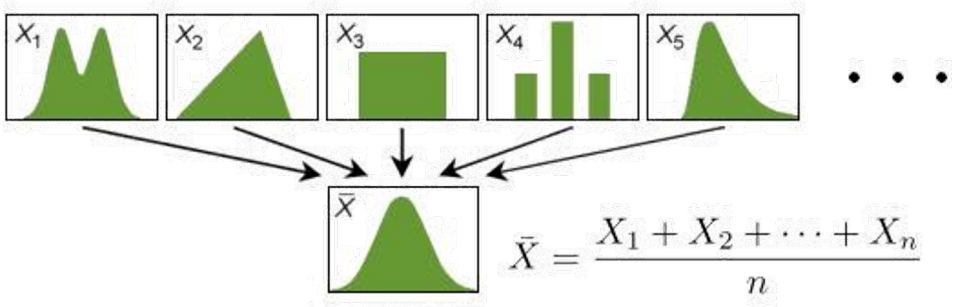
- As $n \to \infty$, Z_n converges in distribution (依分布收敛) to a standard normal random variable, that is:

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

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

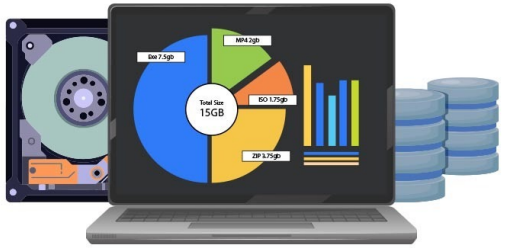
3.3 Function of Multiple Random Variables

- The CLT does not require $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots, X_n$ 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:

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

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

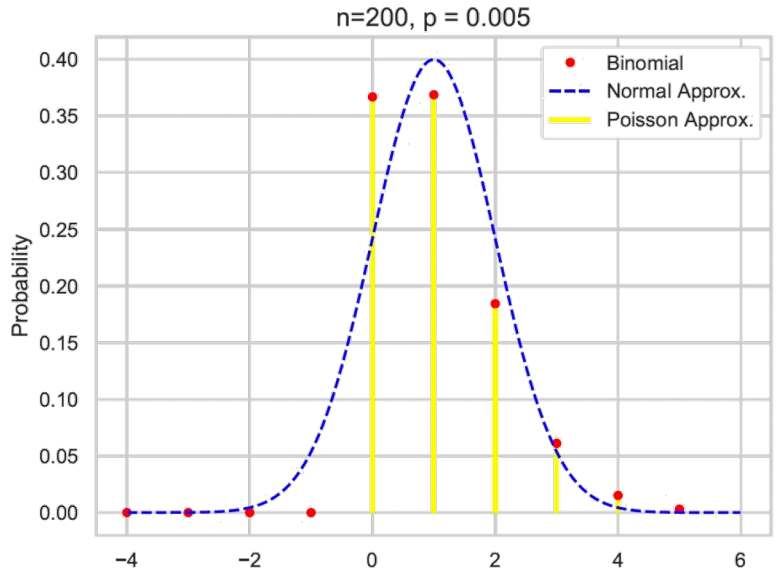
- $\text{Binomial}(n, p)$ can be approximated by $\text{Poisson}(np)$.
- The approximation works well when n is large and p is small, e.g., $n > 100$ and $p < 0.05$.

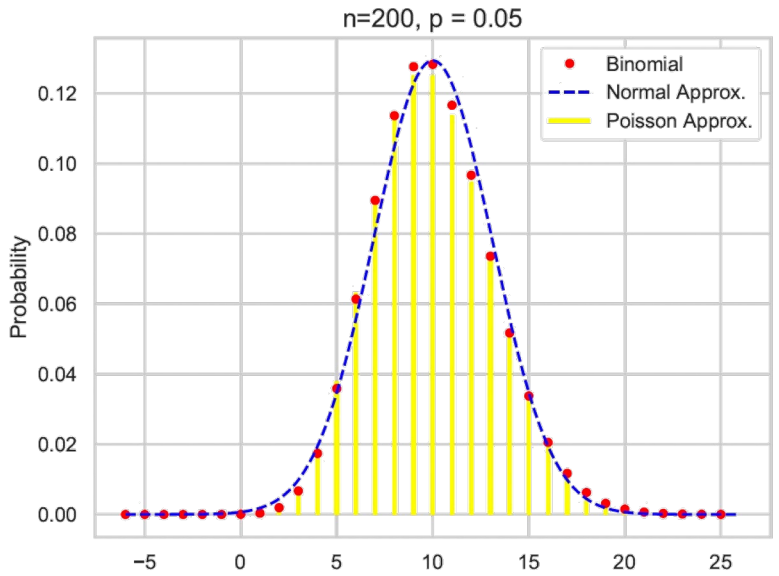
Normal approximation

- $\text{Binomial}(n, p)$ can be approximated by $\text{N}\big(np, np(1-p)\big)$.
- The approximation works well when $np \geq 5$ and $n(1-p) \geq 5$.

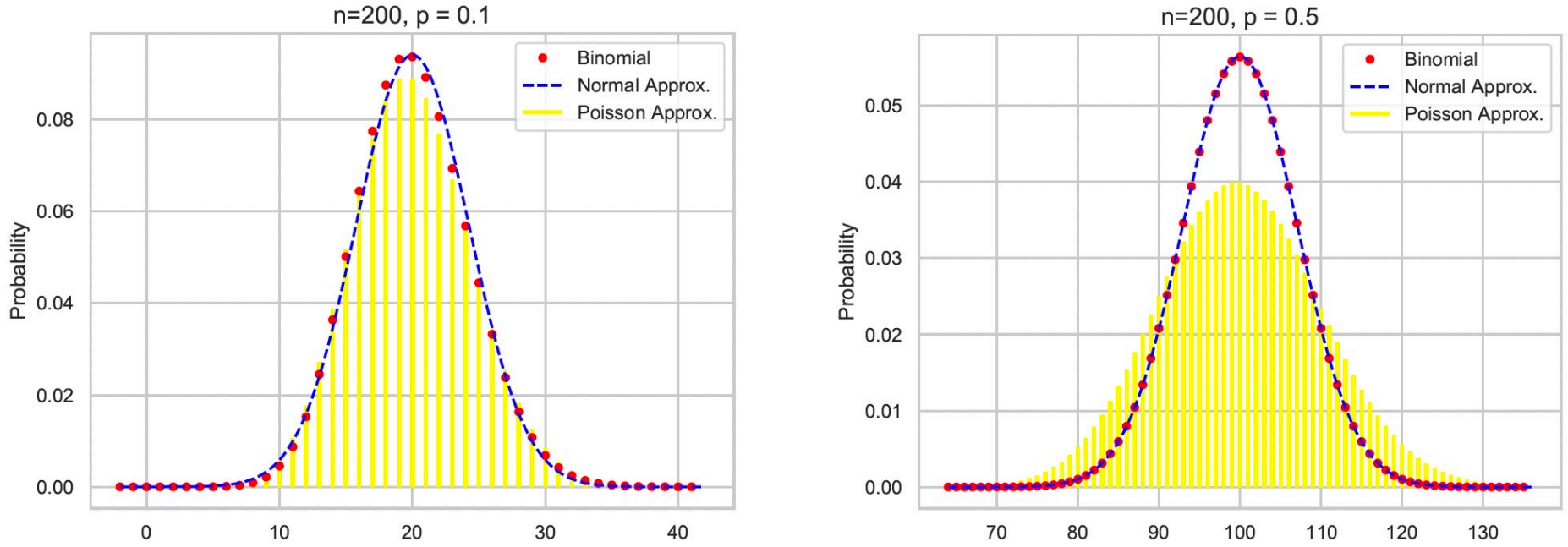
- The best way to understand this is to visualize the three distributions in Python.

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

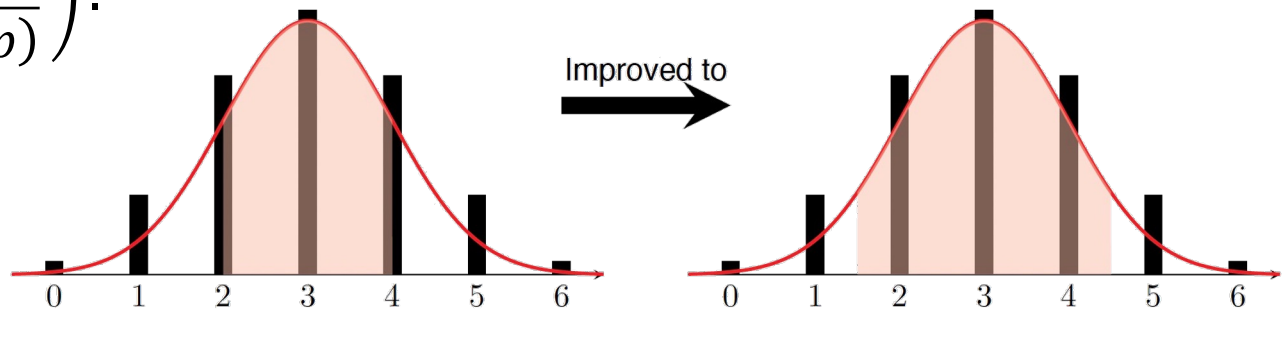
$np(1-p) \approx np$ only when p is small

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,

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

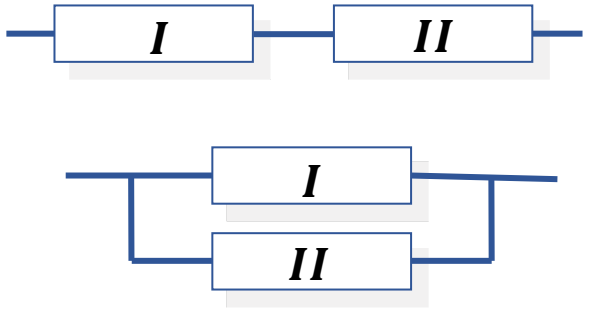
$$F_{\min}(z) = 1 - P(X > z) \cdot P(Y > z) = 1 - [1 - F_X(z)][1 - F_Y(z)].$$

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

3.3 Function of Multiple Random Variables

Example 3.15

- A system is made up of two independent components I and II, with lifespan $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$, 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 $Z = \min(X, Y)$, the CDF of Z is

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

- This suggest that $Z \sim \text{Exp}(\lambda_1 + \lambda_2)$, so that $\text{E}(Z) = 1/(\lambda_1 + \lambda_2)$.
- It's not difficult to find that $\text{E}(Z) < \text{E}(X_1)$ and $\text{E}(Z) < \text{E}(X_2)$, 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 $Z = \max(X, Y)$, the CDF and PDF of Z is

$$F_Z(z) = F_X(z)F_Y(z) = \begin{cases} (1 - e^{-\lambda_1 z})(1 - e^{-\lambda_2 z}), z > 0 \\ 0, \quad \text{otherwise} \end{cases},$$

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

- 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 $\mathbf{Exp(1)}$.
- 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:

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

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

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,

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

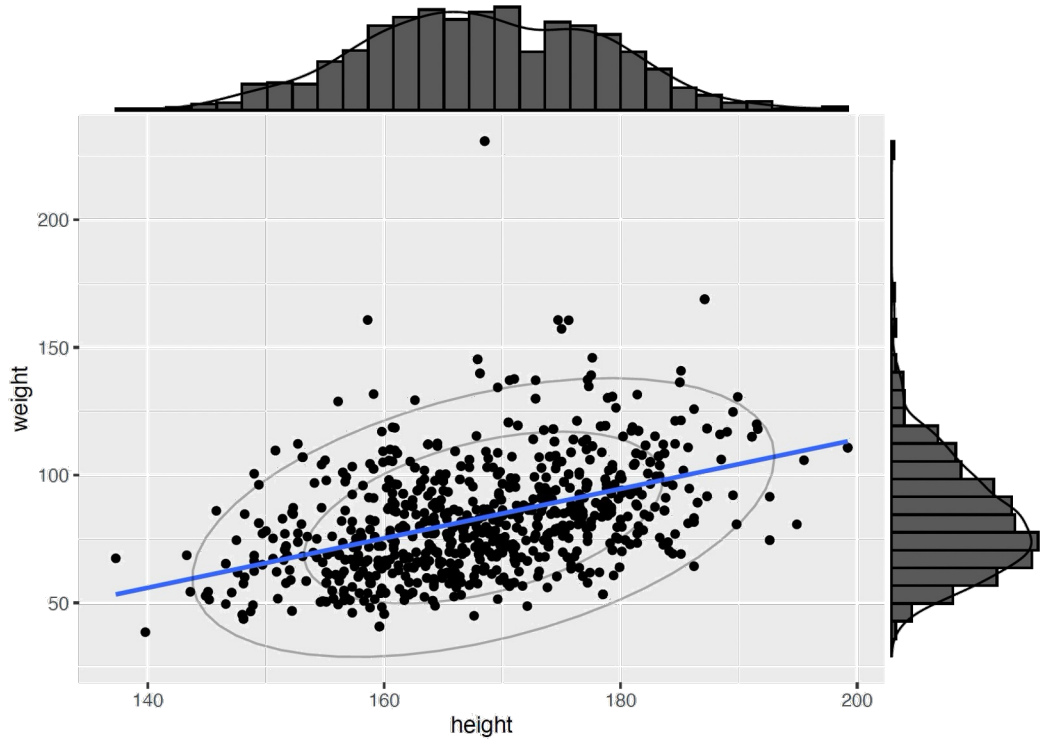
$$=10\int_{10}^{20}\left(\int_{y}^{20}ydx\right)dy+5\int_{10}^{20}\left(\int_{10}^{y}(x+y)dx\right)dy=\frac{20000}{3}+5\times1500\approx14166.67.$$

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

3.4 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 $(-\infty < x, y < \infty)$

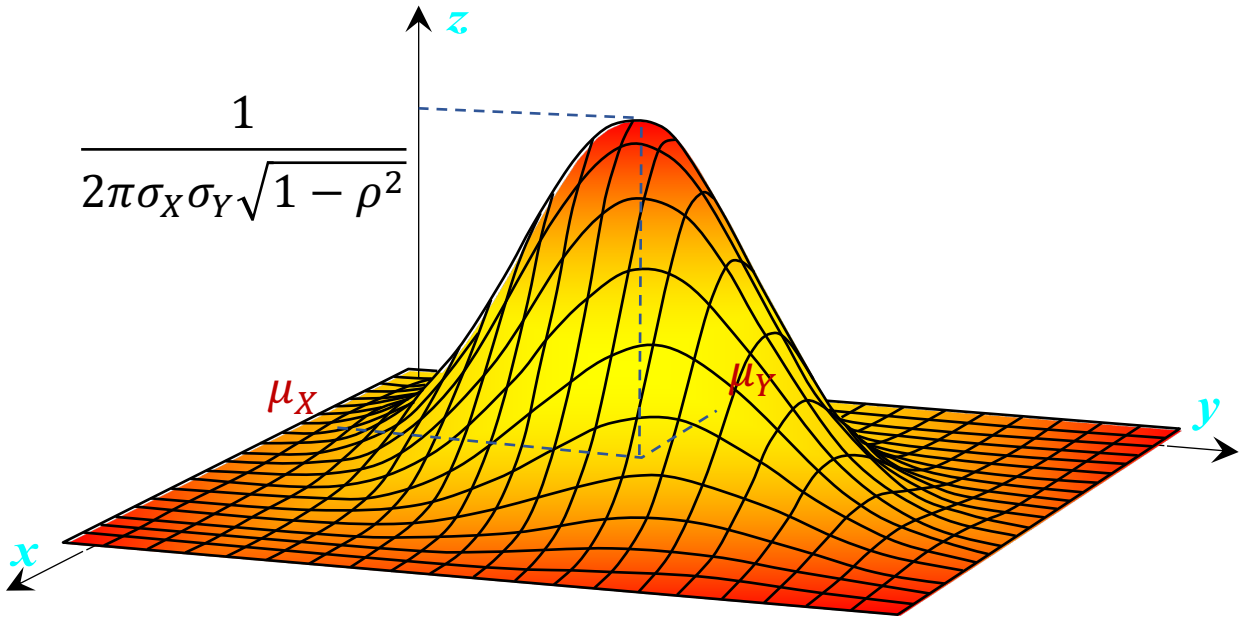
$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right).$$

- It can be expressed as

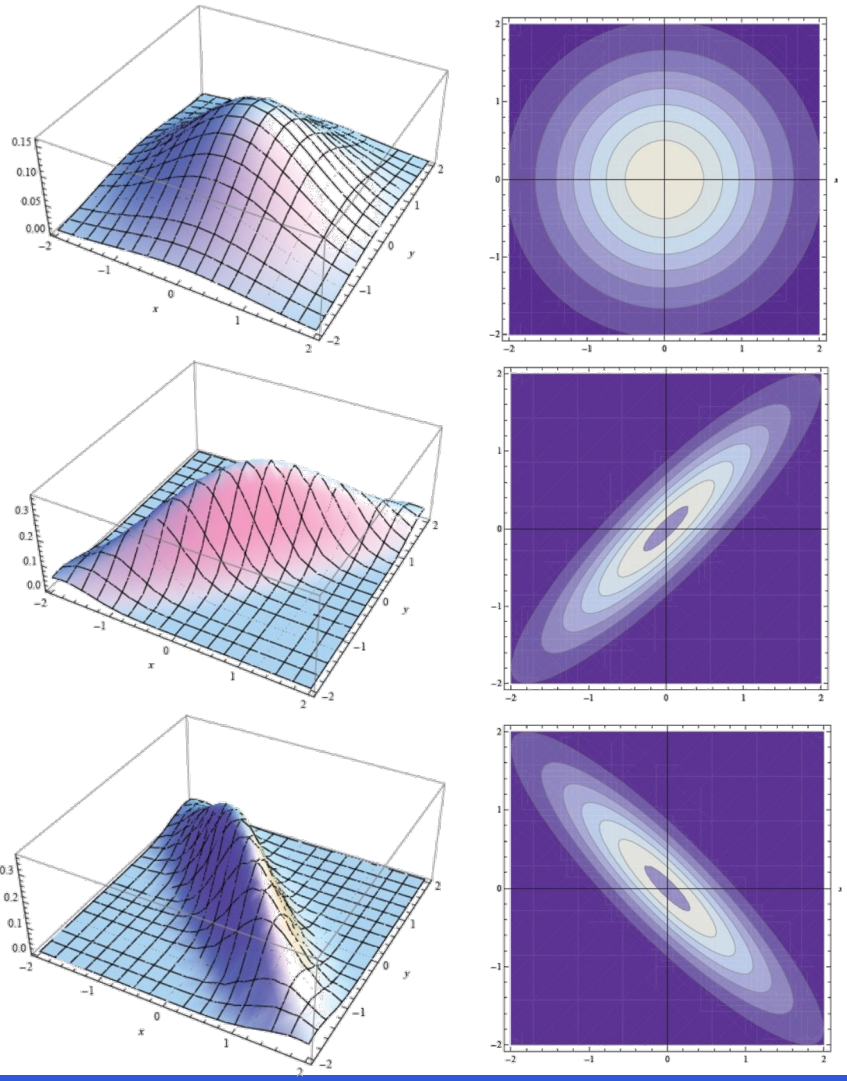
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

- 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., $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

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.

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

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

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

By variable substitution, details omitted.

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

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

3.4 Multivariate Normal Distribution

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

X and Y are independent ✓ → ✗ 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

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

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

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

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:

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

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

- Therefore, $Y \sim \mathrm{N}(0,1)$.
- (2) To show that X and Y are uncorrelated, calcul… $\mathrm{Cov}(X,Y) = \mathrm{E}\big[\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\big] = \mathrm{E}(XY)$:

X and Z are independent

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

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

- 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

- 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

Conditional distributions are normal distributions.

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

Proof:

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

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

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

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

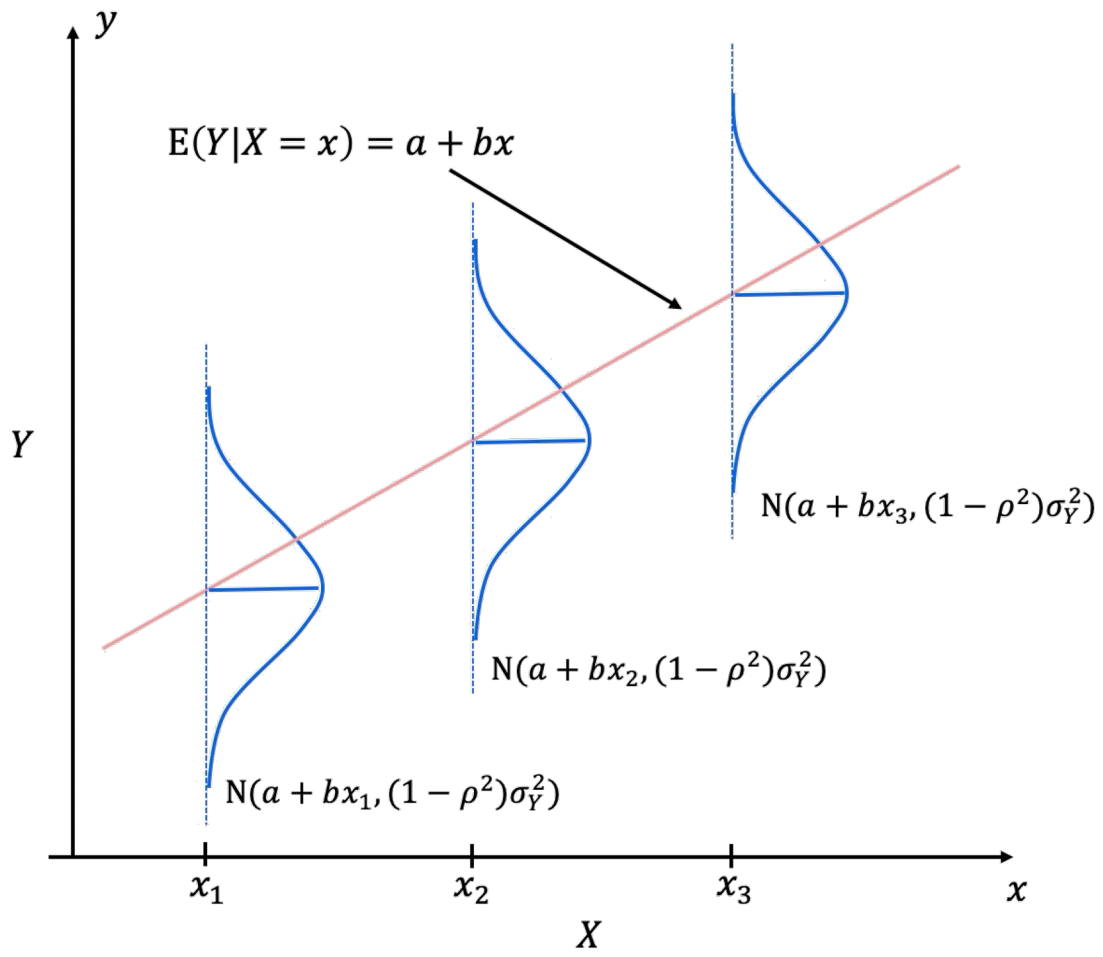
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_X} \exp\left(-\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x - \left(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y)\right)\right]^2\right)$$

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

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

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

- Furthermore, 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 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

$$\mathrm{P}(X > Y | X + Y = 50) = \mathrm{P}(X - Y > 0 | X + Y = 50).$$

- 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 $\mathrm{Var}(W_1), \mathrm{Var}(W_2), \mathrm{Cov}(W_1, W_2)$

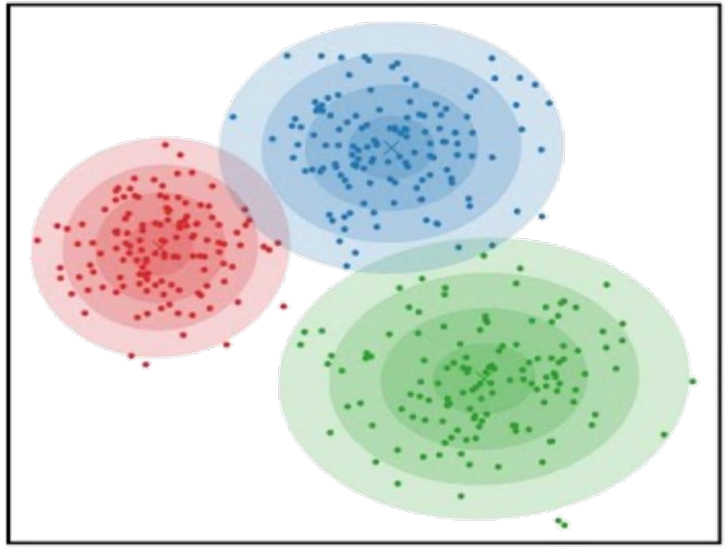
- The correlation coefficient between W_1 and W_2 is $\rho_{12} = 5.39/\sqrt{6.958 \times 52.822} \approx 0.281$.
- Then the conditional distribution of W_1 given $W_2 = w_2 = 50$ is

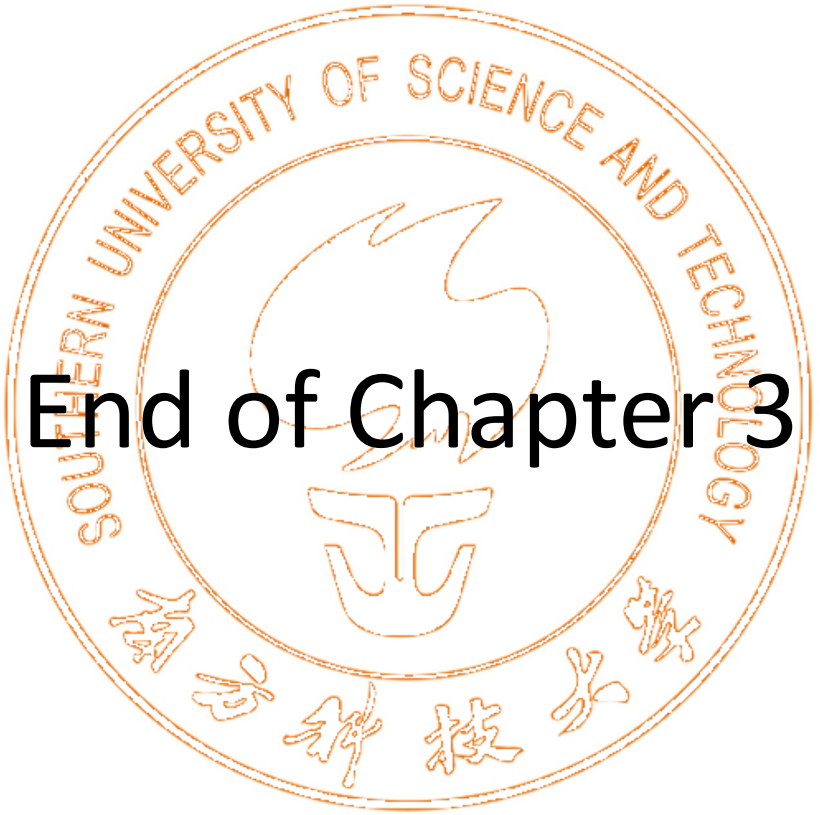
$$\mathrm{N}\left(\mu_1 + \frac{\rho_{12}\sigma_1}{\sigma_2}(w_2 - \mu_2), (1 - \rho_{12}^2)\sigma_1^2 \right) = \mathrm{N}\left(0 + \frac{0.281\sqrt{6.958}}{\sqrt{52.822}}(50 - 45.4), (1 - 0.281^2)6.958 \right) \approx \mathrm{N}(0.47, 6.41)$$

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

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 $\{1, 2, \dots, K\}$, $\boldsymbol{X}$ 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