工程概率统计

Probability and Statistics for Engineering

第二章 随机变量与分布

Chapter 2 Random Variables and Distributions

Chapter 2 Random Variables and Distributions

- 2.1 Introduction
  - 2.1.1 Definition of Random Variable
  - 2.1.2 Description of Probability Distribution
  - 2.1.3 Expectation and Variance
- 2.2 Common Discrete Distributions
- 2.3 Common Continuous Distributions
- 2.4 Transformation of Random Variables

2.1 Introduction

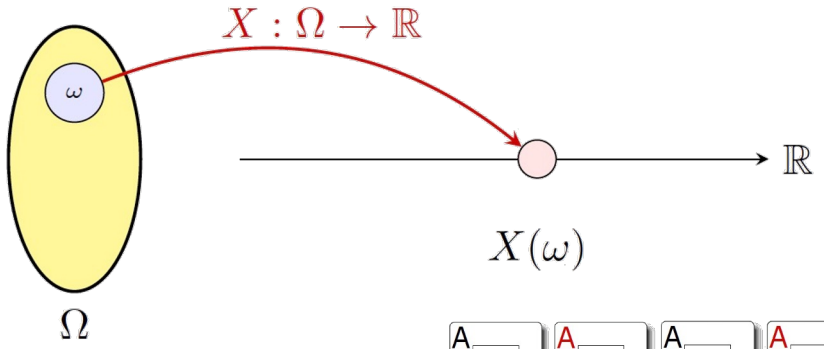
- In Chapter 1, we learned the basic concepts like random experiment, sample space, random event, classical model of probability, geometric model of probability, etc.
- Dealing with complex problems, we need a tool that allows for a unified study of random events.
- The goal is to simplify problems into functional operations, enabling the use of mathematical tools such as calculus.
- A **random variable** is such a tool: it connects the real world with the mathematical world.
  - Simply put, a random variable converts each possible outcome of a random experiment into a number.
  - For example, suppose an organization wants to survey public opinion on a proposal, asking whether people support or oppose it, by randomly sampling 50 individuals.
  - According to classical probability, there are 2^{50} possible outcomes.
  - However, if we use X to represent the number of people who support the proposal, then X is a random variable, and its range of values is limited to $\{0, 1, 2, \ldots, 50\}$.
  - Thus, random variables helps us to describe the problem in a precise and concise manner.

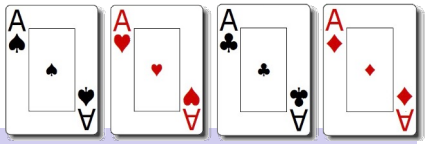
2.1.1 Definition of Random Variable

- Here we provide the mathematical definition of a **random variable (随机变量)**.

Random Variable

A random variable (or r.v. for short) is a **real-valued function** defined on the sample space, typically denoted by capital letters like X, Y, Z.





Example 3.1

- On Valentine's Day a restaurant offers a Lucky Lovers discount that could save couples money on their romantic dinners. How to define the Lucky Lovers discount as a random variable:
- When the waiter brings the check, he'll also bring the four aces from a deck of cards. He'll shuffle them and lay them out face down on the table. The couple will then get to turn one card over.
- If it's a black ace, they'll owe the full amount, but if it's the ace of hearts, the waiter will give them a $20 Lucky Lovers discount.
- If they first turn over the ace of diamonds (hey—at least it's red!), they'll then get to turn over one of the remaining cards, earning a $10 discount for finding the ace of hearts this time.

2.1.1 Definition of Random Variable

Solution

- It is not difficult to obtain the sample space of the game of Lucky Lovers:

$$\Omega = \{\text{A}\spadesuit, \text{A}\heartsuit, \text{A}\clubsuit, (\text{A}\diamondsuit, \text{A}\heartsuit), (\text{A}\diamondsuit, \text{A}\spadesuit), (\text{A}\diamondsuit, \text{A}\clubsuit)\}$$

- The Lucky Lovers discount can be defined as:

$$X(\omega) = \begin{cases} 20 & \text{if } \omega = \text{A}\heartsuit, \\ 10 & \text{if } \omega = (\text{A}\diamondsuit, \text{A}\heartsuit), \\ 0 & \text{otherwise.} \end{cases}$$

- It is a mapping from the outcomes in the sample space to numbers on the real line.

- With random variables defined, the study of random events can be simplified to the study of random variables.
- The study of r.v.s essentially involves examining all possible values that the r.v. can take and the probability associated with each value, which is known as the **probability distribution (概率分布)**.

2.1.1 Definition of Random Variable

- With the probability distribution, we can grasp the overall certainty of the random event, providing a foundation for further exploration of its underlying regularity.
- Based on the possible values a random variable can take, they can be classified into:
  - **Discrete random variable (离散型随机变量)**: take a finite or countable number of values, e.g., the number of customers that visit a restaurant.
  - **Continuous random variable (连续型随机变量)**: take continuous values (an uncountable number of values), e.g., the waiting time for a bus.
- The differences and similarities in how the probability distributions of these two types of r.v.s are described:
  - Discrete r.v. can be described using a **probability mass function (PMF, 概率质量函数)**.
  - Continuous r.v. can be described using a **probability density function (PDF, 概率密度函数)**.
  - Both types of r.v.s can be described using a **cumulative distribution function (CDF, 累积分布函数)**.

2.1.2 Description of Probability Distribution

Probability Mass Function

Let X be a discrete random variable, $S = \{a_1, a_2, \ldots\}$ be the **support (支撑集)** of X, i.e., the set of values that X can take. Then the **probability mass function (PMF, 概率质量函数)** of X is $p: S \to [0, 1]$ defined as

$$p(a_i) = p_i = P(X = a_i), i = 1, 2, \ldots,$$

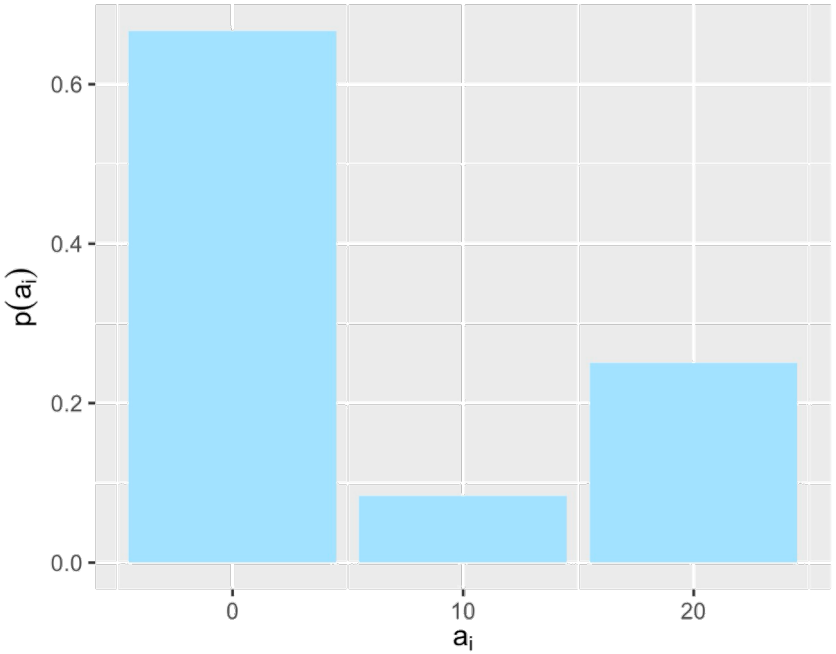
which satisfies:

- **Non-negativity (非负性)**: $p(a_i) \geq 0, i = 1, 2, \ldots;$
- **Normalization (规范性)**: $\sum_i p(a_i) = 1.$

- The PMF can also be displayed in a tabular format

Value	a_1	a_2	$\cdots$	a_i	$\cdots$
Probability	p_1	p_2	$\cdots$	p_i	$\cdots$

- The PMF can be displayed in a graphical format by plotting $p(a_i)$ on the y-axis against a_i on the x-axis.



2.1.2 Description of Probability Distribution

Example 3.1 (Continued)

Obtain the PMF of the Lucky Lovers discount.

Solution

- It is not difficult to get:

$$P(\text{A}\spadesuit) = P(\text{A}\heartsuit) = P(\text{A}\clubsuit) = \frac{1}{4},$$

$$P\{(\text{A}\diamondsuit, \text{A}\heartsuit)\} = P\{(\text{A}\diamondsuit, \text{A}\spadesuit)\} = P\{(\text{A}\diamondsuit, \text{A}\clubsuit)\} = \frac{1}{4}\times\frac{1}{3} = \frac{1}{12}.$$

- Then:

$$P(X = 20) = P(\text{A}\heartsuit) = \frac{1}{4}, P(X = 10) = P\{(\text{A}\diamondsuit, \text{A}\heartsuit)\} = \frac{1}{12}, P(X = 0) = \frac{1}{4} + \frac{1}{4} + \frac{1}{12} + \frac{1}{12} = \frac{2}{3}.$$

- Display the PMF in a tabular format:

Value	0	10	20
Probability	0.667	0.083	0.250

2.1.2 Description of Probability Distribution

Example 3.2

Suppose that the support of a r.v. X is $\{0, 1, 2, \ldots\}$ and the PMF of X is given by $P(X = k) = c\lambda^k/k!$, where λ is some positive constant. Please express the value of c in terms of λ.

Solution

- By the normalization property of the PMF, we have

$$1 = \sum_{k=0}^{\infty} P(X = k) = c \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

- This implies that

$$1 = ce^{\lambda} \Longrightarrow c = e^{-\lambda}.$$

2.1.2 Description of Probability Distribution

Continuous Random Variable and Probability Density Function

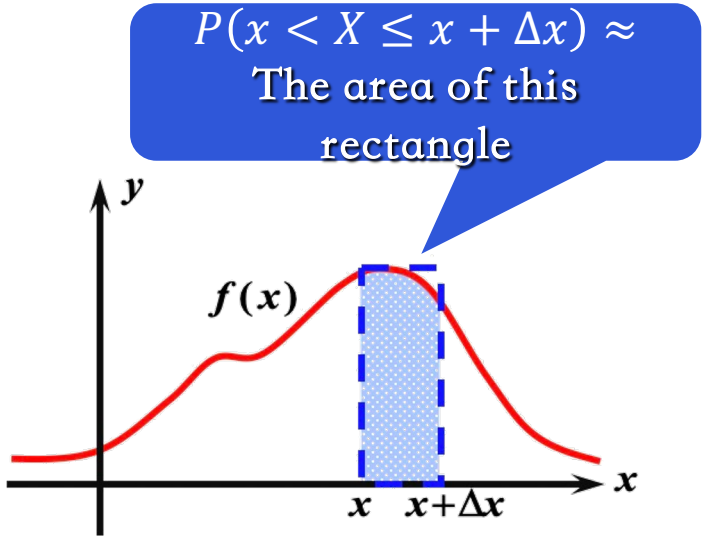
X is called a continuous random variable if there exists a non-negative function f, defined for all $x \in \mathbb{R}$, satisfies that

$$P(a \le X \le b) = \int_a^b f(x)dx$$

for any $-\infty < a \le b < \infty$. The function f is called the probability density function (PDF, 概率密度函数) of X.

It is not difficult to see that the PDF, like the PMF, has these properties:

- Non-negativity (非负性): $f(x) \ge 0, \forall x \in \mathbb{R}$;
- Normalization (规范性): $\int_{-\infty}^{\infty} f(x)dx = 1$.



- Note that for any $x \in \mathbb{R}$, we have $P(X = x) = \int_x^x f(u)du = 0$.
- Thus, different from the PMF, $f(x)$ does not reflect the probability of X taking the value x.

- The larger $f(x)$ is, the greater the probability that X takes a value near x.
- $f(x)$ reflects the degree to which the probability is concentrated around x.

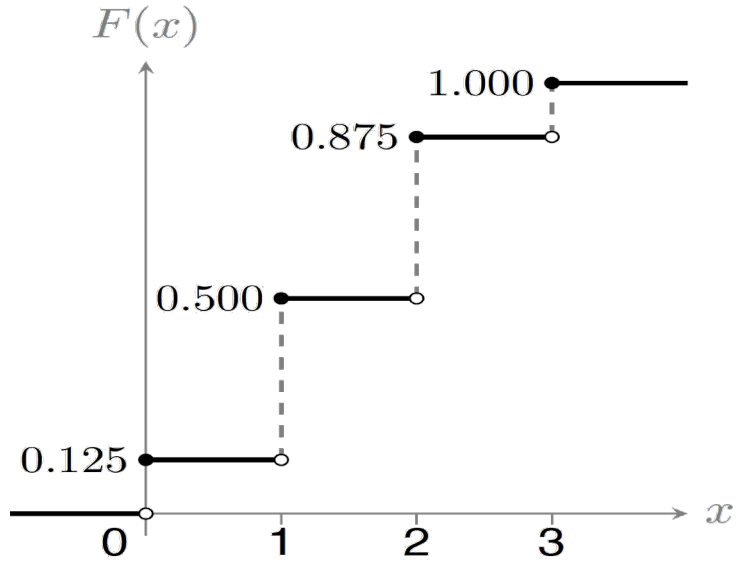
2.1.2 Description of Probability Distribution

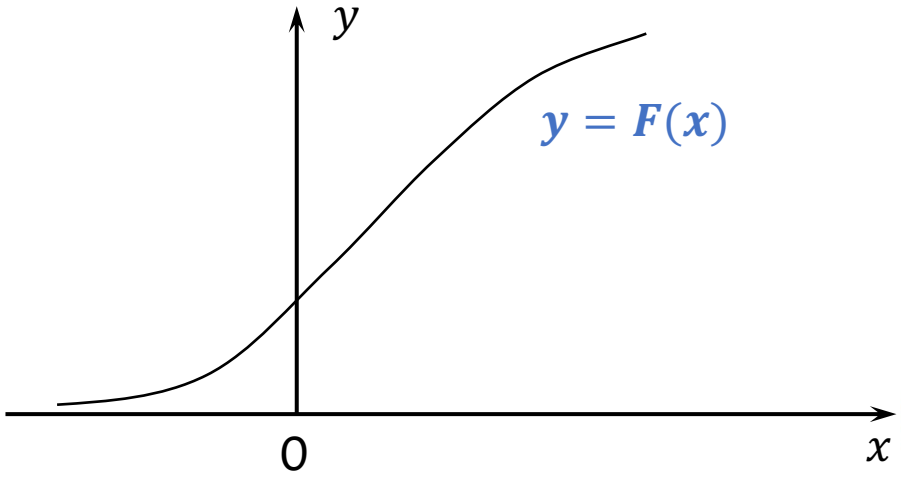
Cumulative Distribution Function

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

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

- For a **discrete r.v.**, its CDF is a step function (阶梯函数), which is $F(x) = \sum_{a_i \leq x} p(a_i)$.
- For a **continuous r.v.**, its CDF is a continuous function, which is $F(x) = \int_{-\infty}^{x} f(u)du$. Consequently, $f(x) = F'(x)$.
- The CDF is **non-decreasing** and **right-continuous**.
- The maximum of the CDF is $F(\infty) = \lim_{x \to \infty} F(x) = 1$.
- The minimum of the CDF is $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$.
- For any real numbers $a < b$, $P(a < X \leq b) = F(b) - F(a)$.





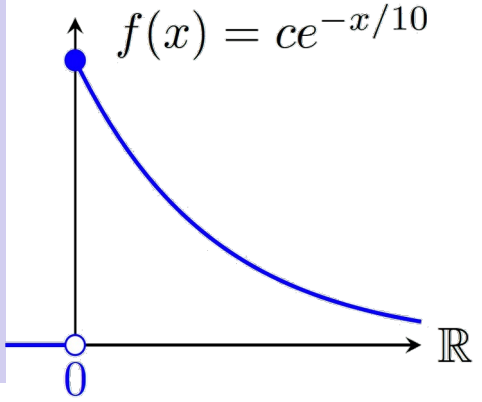
2.1.2 Description of Probability Distribution

Example 3.3

- Suppose that the lifespan in years of a certain household appliance is a r.v. with PDF given by

$$f(x) = \begin{cases} ce^{-x/10}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$

- What is the probability that an appliance will function between 1 and 2 years?



Solution

- By the normalization property of the PDF, we have

$$1 = \int_{-\infty}^{\infty} f(x)dx = c\int_{0}^{\infty} e^{-x/10}dx \Longrightarrow 1 = 10c \Longrightarrow c = \frac{1}{10}.$$

- Let X be the r.v. representing the lifespan (in years) of a computer, then

$$P(1 \leq X \leq 2) = \int_{1}^{2} f(x)dx = \frac{1}{10}\int_{1}^{2} e^{-x/10}dx = e^{-1/10} - e^{-1/5} \approx 0.086.$$

2.1.3 Expectation and Variance

- A probability distribution provides a comprehensive and complete description of a r.v. However, sometimes we may want to describe the r.v. from specific aspects with just a few numbers.
- These numbers should be simple, clear, distinctive, and intuitive.
- The most commonly used numerical characteristics are the **mathematical expectation (数学期望)** and **variance (方差)**.

Mathematical Expectation

The **mathematical expectation (数学期望)** of a random variable X is denoted as $E(X)$:

- If X is a **discrete r.v.** with PMF $p(x_k) = p_k = P(X = x_k), k = 1, 2, \ldots$, given $\sum_{k=1}^{\infty} |x_k| p_k < \infty$, then

$$\mathrm{E}(X) \triangleq \sum_{k=1}^{\infty} x_k P(X = x_k) = \sum_{k=1}^{\infty} x_k p_k .$$

- If X is a **continuous r.v.** with PDF $f(x)$, given $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, then

$$\mathrm{E}(X) \triangleq \int_{-\infty}^{\infty} x f(x) dx .$$

Also called the mean (均值), expectation (期望), expected value (期望值).

2.1.3 Expectation and Variance

- Simply put, the expectation is the weighted average of all possible values of a random variable.
- Expectation represents the average result or long-term value that we can anticipate from a series of random events, and it has wide applications in real life.
  - In the insurance industry, expectation can be used for setting the premium setting.
  - In the financial sector, investors use expectation to predict investment returns.

Example 3.3 (Continued)

- A store adopts a 'use first, pay later' approach for selling the household appliance, with the payment amount determined based on the lifespan X of the appliance.

Lifespan (year)	$X \leq 1$	$1 < X \leq 2$	$2 < X \leq 3$	$X > 3$
Payment	1500	2000	2500	3000

- What's the expected return for the store to sell a random appliance?

2.1.3 Expectation and Variance

Solution

- The CDF of X is

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du = \frac{1}{10}\int_{0}^{x} e^{-u/10} = 1 - e^{-x/10}.$$

- Let Y be the return for the store to sell a random appliance, then the PMF of Y is

$$P(Y = 1500) = P(X \le 1) = F_X(1) \approx 0.095,$$
$$P(Y = 2000) = P(1 < X \le 2) = F_X(2) - F_X(1) \approx 0.086,$$
$$P(Y = 2500) = P(2 < X \le 3) = F_X(3) - F_X(2) \approx 0.078,$$
$$P(Y = 3000) = P(X > 3) = 1 - F_X(3) \approx 0.741.$$

- The PMF of Y is summarized in the table below:

Value	1500	2000	2500	3000
Probability	0.095	0.086	0.078	0.741

- Thus, the expectation of Y is computed as

$$\mathrm{E}(Y) = \sum_{k=1}^{4} y_k P\{Y = y_k\} = 1500\times0.095 + 2000\times0.086 + 2500\times0.078 + 3000\times0.741 = 2732.5.$$

2.1.3 Expectation and Variance

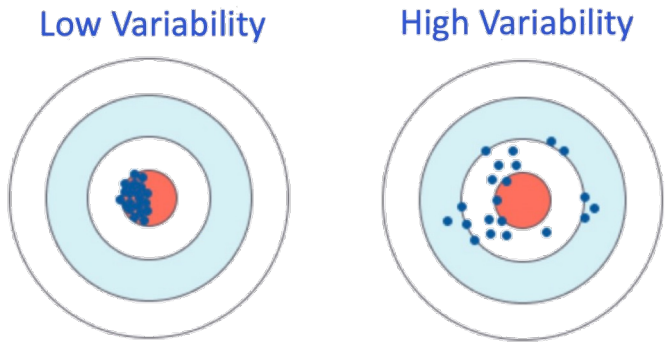
- While expectation represents the average result, we are also interested in the stability/variability (稳定性/波动性) of the result.
- Variance (方差) describes the variability of a random variable around its expectation.

Variance and Standard Deviation

If a r.v. X satisfies that $\mathrm{E}(X^2) < \infty$, then

$$\mathrm{Var}(X) \triangleq \mathrm{E}\big(X - \mathrm{E}(X)\big)^2$$

is called the **variance (方差)** of X, and $\mathrm{SD}(X) \triangleq \sqrt{\mathrm{Var}(X)}$ is called the **standard deviation (SD, 标准差)**.



Q: Why not use $\mathrm{E}|X - \mathrm{E}(X)|$ to describe the variability?

- Specifically, by the definition of expectation:
  - If X is a **discrete r.v.** with PMF $p(x_k) = p_k = P(X = x_k), k = 1, 2, \ldots$, then $\mathrm{Var}(X) = \sum_{k=1}^{\infty} [x_k - \mathrm{E}(X)]^2 \cdot p_k$.
  - If X is a **continuous r.v.** with PDF $f(x)$, then $\mathrm{Var}(X) = \int_{-\infty}^{\infty} [x - \mathrm{E}(X)]^2 f(x) dx$.
- The expectation of function of X: if X is a discrete r.v., $\mathrm{E}\big(g(X)\big) = \sum_{k=1}^{\infty} g(x_k) p_k$;

  if X is a continuous r.v., $\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f(x) dx$.

2.1.3 Expectation and Variance

Example 3.4

- Suppose that the time deviation (in seconds) of two brands of watches are summarized below

Deviation	-3	-2	-1	0	1	2	3
Prob. (Brand A)	0.10	0.15	0.15	0.20	0.15	0.15	0.10
Prob. (Brand B)	0.15	0.10	0.10	0.30	0.10	0.10	0.15

- Which brand has better quality?

Solution

- Let X and Y be the time deviation of the two brands of watches, it is not difficult to get

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

- Thus, from the perspective of average time deviation, the two brands are of equal quality.
- Then we consider the variance:

$$\mathrm{Var}(X) = \sum [x_k - 0]^2 \cdot p_k = 3^2{\times}0.1 + 2^2{\times}0.15 + \cdots + 3^2{\times}0.1 = 3.3, \quad \mathrm{Var}(Y) = 3.7.$$

- Therefore, from the perspective of stability, brand A is of better quality.

2.1.3 Expectation and Variance

- There are some basic properties of the expectation and variance:
  - For any constants a, b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.
  - For any constants a, b, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$, and thus $\mathrm{SD}(aX + b) = a\mathrm{SD}(X)$.
  - $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.

Proof: Without loss of generality, show the case for a discrete r.v. with PMF $p(x_k) = p_k = P(X = x_k)$.

Let $\mu \triangleq \mathrm{E}(X)$, then

$$\mathrm{Var}(X) = \mathrm{E}\big(X - \mathrm{E}(X)\big)^2 = \sum_{k=1}^{\infty}(x_k - \mu)^2 p_k$$

$$= \sum_{k=1}^{\infty}\left(x_k^2 - 2\mu x_k + \mu^2\right)p_k = \sum_{k=1}^{\infty} x_k^2 p_k - 2\mu\sum_{k=1}^{\infty} x_k p_k + \mu^2$$

$$= \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

- $\mathrm{E}[g_1(X) \pm g_2(X)] = \mathrm{E}[g_1(X)] \pm \mathrm{E}[g_2(X)]$, for any functions $g_1(x)$ and $g_2(x)$.

Chapter 2 Random Variables and Distributions

- 2.1 Introduction
- 2.2 Common Discrete Distributions
- 2.3 Common Continuous Distributions
- 2.4 Transformation of Random Variables

2.2 Common Discrete Distributions

- Bernoulli distribution (伯努利分布) is the simplest discrete distribution.

Bernoulli Trial and Bernoulli Distribution

- If a random experiment only have two possible outcomes, A and $\overline{A}$, then the experiment is called a **Bernoulli trial (伯努利试验)**.
- If a random variable X only takes values 0 and 1, and $P(X=1)=p$, $P(X=0)=1-p$, then X is called to follow a **Bernoulli distribution (伯努利分布)** with parameter p, denoted as
$$X \sim \text{Bernoulli}(p).$$
- The PMF of $\text{Bernoulli}(p)$ can be expressed as
$$p(x)=p^x(1-p)^{1-x}, x=0,1.$$
- The expectation and variance of $X \sim \text{Bernoulli}(p)$:
$$\text{E}(X)=p, \text{Var}(X)=p(1-p).$$

Jacob Bernoulli
(1655-1705)

- The Bernoulli distribution is the foundation of many classical probability distributions, such as the binomial distribution (二项分布), the geometric distribution (几何分布), etc.

2.2 Common Discrete Distributions

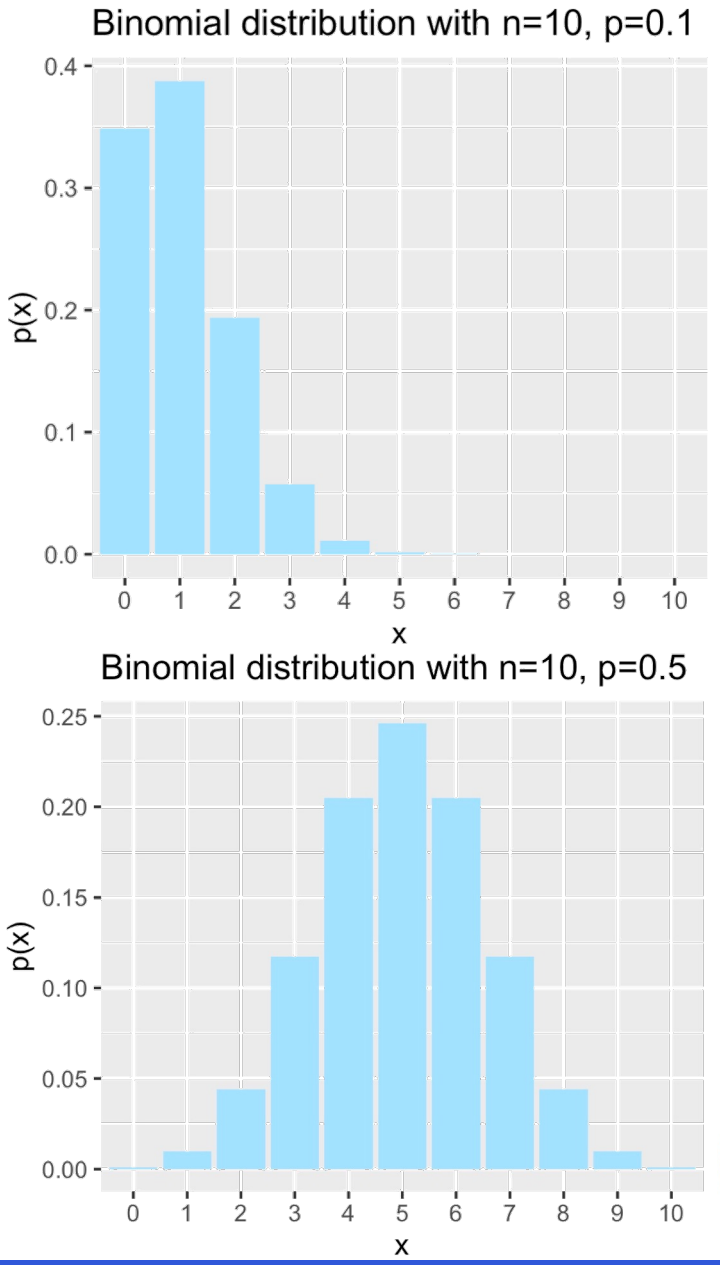
n-Fold Bernoulli Trial and Binomial Distribution

- An n-fold Bernoulli trial (n重伯努利试验) is an experiment that repeats a Bernoulli trial n times independently. Note:
  - "independently" suggests that the result of each Bernoulli trial would not affect each other.
  - "repeat" suggests that the probability of event A in each Bernoulli trial, i.e., $P(A) = p$, remains the same.
- Let X be a r.v. that records the number of times event A happens in an n-fold Bernoulli trial, then X is called to follow a binomial distribution (二项分布) with parameter n and p, denoted as

$$X \sim \text{Binomial}(n, p) \text{ (or } \text{Bin}(n, p), \text{B}(n, p)).$$

- The PMF of $\text{Binomial}(p)$ can be derived as

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = C_n^x p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots, n.$$



2.2 Common Discrete Distributions

Example 3.5

- A factory has 80 pieces of the same type of equipment, each operating independently, with a failure probability of 0.01.
- A single maintainer can only repair one piece of equipment at a time. The factory is considering two strategies for allocating maintainers:
  - Allocate 4 maintainers, with each responsible for maintaining 20 pieces of equipment.
  - Allocate 3 maintainer, with them jointly responsible for maintaining all 80 pieces of equipment.
- Please compare these two strategies in terms of the probability that a piece of equipment cannot be repaired in time when a failure occurs.

2.2 Common Discrete Distributions

Solution

- First consider the first strategy, let X_1, X_2, X_3, X_4 be the number of equipment that fail at the same time among the 20 pieces maintained by the four maintainers, respectively. Then

$$X_i \sim B(20, 0.01), i = 1, 2, 3, 4.$$

- Thus, the probability that a piece of equipment cannot be repaired in time is

$$P\left(\bigcup_{i=1}^{4}\{X_i \geq 2\}\right) \geq P\{X_1 \geq 2\} = 1 - P\{X_1 = 0\} - P\{X_1 = 1\}$$
$$= 1 - \binom{20}{0} \times 0.01^0 \times 0.99^{20} - \binom{20}{1} \times 0.01^1 \times 0.99^{19} \approx 0.0169.$$

- Then consider the second strategy, let X be the number of equipment that fail at the same time among the 80 pieces, then

$$X \sim B(80, 0.01).$$

- Thus, the probability that a piece of equipment cannot be repaired in time is

$$P\{X \geq 4\} = 1 - \sum_{i=0}^{3} P\{X = i\} = 1 - \sum_{i=0}^{3} \binom{80}{i} \times 0.01^i \times 0.99^{80-i} \approx 0.0087.$$

- Therefore, the second strategy is more efficient.

2.2 Common Discrete Distributions

- The expectation and variance of $X \sim \text{Binomial}(n, p)$ can be derived to be

$$\text{E}(X) = np, \text{Var}(X) = np(1-p).$$

Proof: By the definition of expectation, we have:

$$\text{E}(X) = \sum_{k=0}^{n} k \cdot P(X=k) = \sum_{k=1}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \cdot \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} p^{k-1} (1-p)^{n-k} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1}$$

binomial expansion (二项式展开)

$$= np[(1-p)+p]^{n-1} = np.$$

The variance can also be derived similarly, details are not shown here.

Easier derivations are available once we introduce the independence between random variables.

2.2 Common Discrete Distributions

Geometric Distribution

- Suppose that a Bernoulli trial is repeated independently until A occurs, let X be a r.v. that records the number of trials required, then X is called to follow a **geometric distribution (几何分布)** with parameter p, denoted as

$$X \sim \text{Geometric}(p).$$

- The PMF of $\text{Geometric}(p)$ can be derived as

$$p(x) = p(1-p)^{x-1}, x = 1, 2, \ldots.$$

- The expectation and variance of $X \sim \text{Geometric}(p)$:

$$\text{E}(X) = \frac{1}{p}, \text{Var}(X) = \frac{1-p}{p^2}.$$

- The geometric distribution is the only discrete distribution which has the **memoryless property (无记忆性)**, i.e., for $X \sim \text{Geometric}(p)$ and any positive integers m, n, we have

$$P(X > m+n | X > m) = P(X > n).$$

Example 3.6

- A gambler at a casino is betting on 'big' in a game of 'big or small,' but the table has shown 'small' for ten consecutive rounds.
- The gambler believes that since 'small' has come up so many times already, the probability of another 'small' is very low.
- He is considering betting all his money on 'big' to try to win back his losses. What do you think?

Gambler's Fallacy (赌徒谬误)

2.2 Common Discrete Distributions

- Some probability distributions can be directly derived from the data generation mechanism itself, such as the Bernoulli distribution, binomial distribution, geometric distribution.
- Some other probability distributions are first derived mathematically or summarized from data, and later found to describe many real-world patterns.
- The Poisson distribution is an example of the second type which was first introduced by Siméon-Denis Poisson when he study the number of wrongful convictions in a given country.

Poisson Distribution

- Let X be a discrete r.v. with support $\{0, 1, 2, \dots\}$, if its PMF is

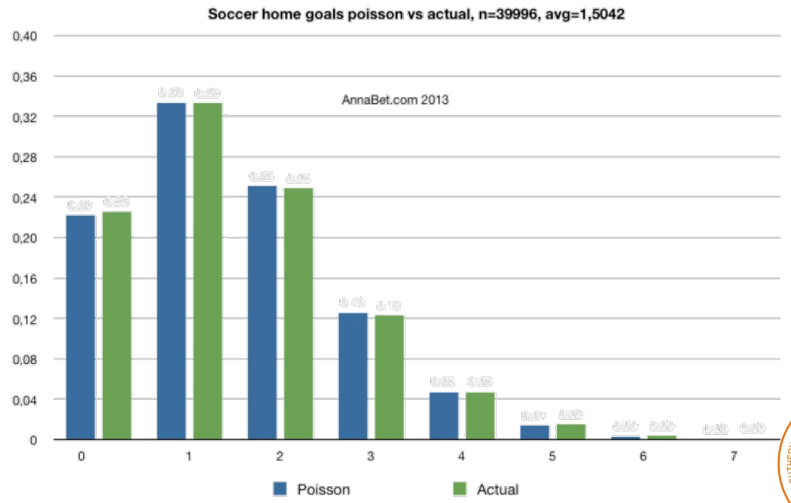
$$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots,$$

- where $\lambda > 0$ is a constant, then X is said to follow a **Poisson distribution (泊松分布)** with parameter λ, denoted by $X \sim \text{Poisson}(\lambda)$.

Siméon-Denis Poisson
西莫恩·德尼·泊松
(1781-1840)

2.2 Common Discrete Distributions

- While the Poisson distribution was initially introduced to solve a legal and criminal justice problem, it is widely used in various fields latter. Examples:
  - In 1860, an American astronomer fitted the Poisson distribution to the number of stars found in a space.
  - In 1898, a Russian economist showed that the frequency with which soldiers in the Prussian army accidentally killed by horse kicks could be well modeled by a Poisson distribution.
- Poisson distribution is used to describe the number of events occurring in a **fixed interval of time /space** if the events occur with a **constant rate** and **independently**. Examples:
  - The number of customers entering a store each day.
  - The number of traffic accidents at an intersection each year.
  - The number of typos on a page of a book.
  - The number of home goals in a World Cup soccer match.
  - The number of mutations on a strand of DNA per unit length.



2.2 Common Discrete Distributions

- The expectation and variance of $X \sim \text{Poisson}(\lambda)$ can be derived to be

$$\mathrm{E}(X) = \lambda, \mathrm{Var}(X) = \lambda.$$

Proof: By the definition of expectation, we have:

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k \cdot P(X = k) = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda.$$

$$\mathrm{E}(X^2) = \sum_{k=0}^{\infty} k^2 \cdot P(X = k) = \sum_{k=1}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \left[\sum_{k=1}^{\infty} \frac{(k-1)\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \right] = \lambda^2 + \lambda.$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \lambda.$$

- The parameter λ is also called the **intensity (泊松强度)** of the Poisson distribution.

2.2 Common Discrete Distributions

Example 3.7

- A Council is considering whether to base a recovery vehicle (救援车辆) on a stretch of road to help clear incidents as quickly as possible.
- Records show that, on average, the number of incidents during the morning rush hour is 5.
- The Council won't base a recovery vehicle on the road if the probability of having more than 5 incidents during the morning rush hour is less than 30%.
- Based on this information, should the Council provide a vehicle?

2.2 Common Discrete Distributions

Solution

- Let X be the number of incidents during the morning rush hour of a random day, then $X \sim \text{Poisson}(5)$.
- The target is to compute $P(X > 5) = 1 - P(X \leq 5)$. Since:

$$P(X=0) = \frac{5^0}{0!}e^{-5} = 0.00674, \quad P(X=1) = \frac{5^1}{1!}e^{-5} = 5 \times P(X=0) = 0.03369$$

$$P(X=2) = \frac{5}{2} \times P(X=1) = 0.08422, \quad P(X=3) = \frac{5}{3} \times P(X=2) = 0.14037,$$

$$P(X=4) = \frac{5}{4} \times P(X=3) = 0.17547, \quad P(X=5) = \frac{5}{5} \times P(X=4) = 0.17547.$$

- So, the probability of having more than 5 incidence on the road during the morning rush hour is

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.00674 - 0.03369 - \cdots - 0.17547 = 0.38403 > 30\%.$$

- Therefore, the Council should provide a recovery vehicle on the road.

2.2 Common Discrete Distributions

- The **Poisson distribution** can be obtained derived as a **limiting case of the binomial distribution.**

Proof: Consider the number of events within a unit time interval $[0, 1]$ and divide it into n subintervals:

$$\left[0, \frac{1}{n}\right], \left[\frac{1}{n}, \frac{2}{n}\right], \cdots, \left[\frac{i-1}{n}, \frac{i}{n}\right], \ldots, \left[\frac{n-1}{n}, 1\right].$$

Several assumptions are made:

1. Let n be large so that each subinterval is very short, making it impossible for two or more events to occur within any subinterval (i.e., either no event occurs or exactly one event occurs).
2. The probability of an event occurring is proportional to the length of the subinterval, i.e., λ/n.
3. Whether an event occurs in a subinterval is independent of the others.

Let X be the number of events within $[0, 1]$, by the assumptions above, we have $X \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right)$. So

$$P(X = x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

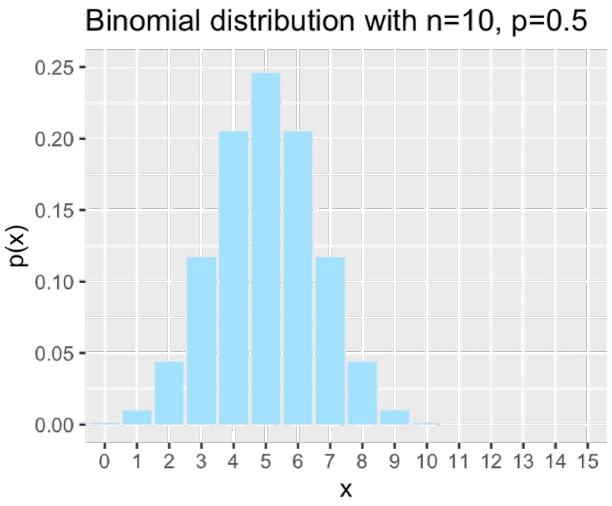
Let $n \to \infty$, it is not difficult to get

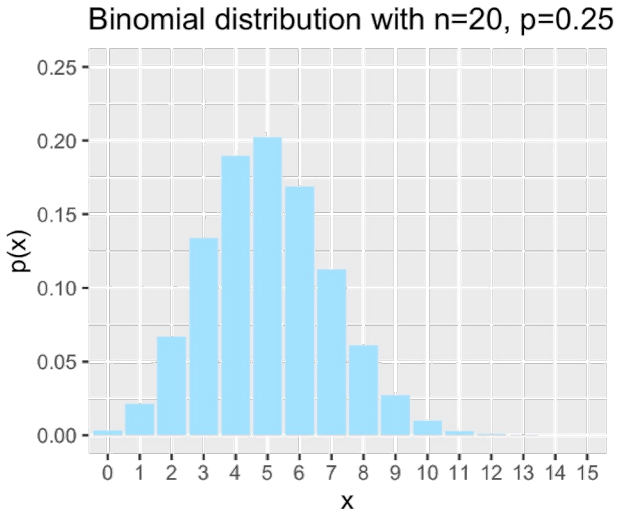
$$\lim_{n\to\infty} P(X = x) = \lim_{n\to\infty} \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!}e^{-\lambda}.$$

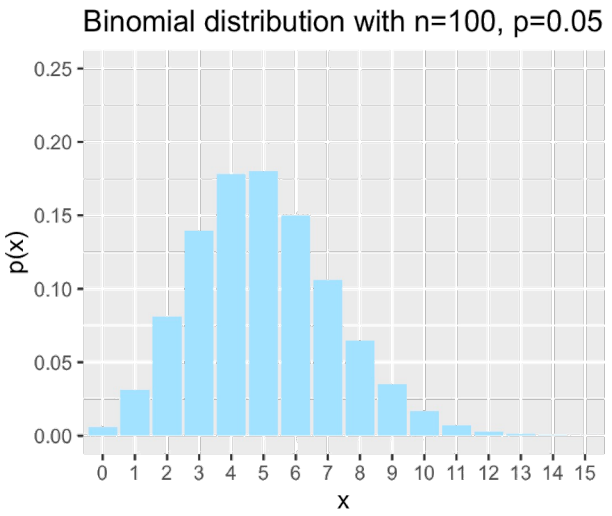
Poisson Theorem
（泊松定理）

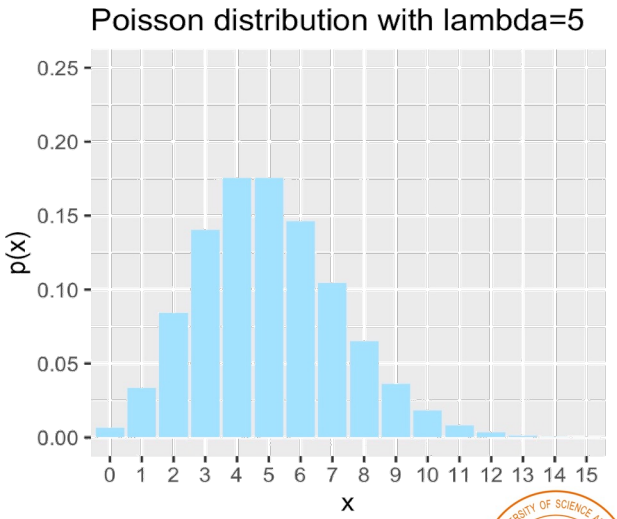
2.2 Common Discrete Distributions

- The theorem suggests that for an n-fold Bernoulli trial with large n and small p, the binomial distribution $\textbf{Binomial}(n, p)$ can be approximated by the Poisson distribution $\textbf{Poisson}(np)$.
- How large should n be and how small should p be?
- A rule of thumb is that when $n > 100$ and $p < 0.05$, the Poisson distribution provides a good approximation to the binomial distribution.









2.2 Common Discrete Distributions

Example 3.8

- An insurance company has launched a life insurance policy where each participant is required to pay a premium of $12 on January 1st.
- If a participant dies within the year, his/her family can receive a compensation of $2,000 from the insurance company.
- Suppose 2,500 people participate in this insurance, and the probability of death for each person within the year is 0.002.
- What is the probability that the insurance company's profit from this life insurance policy is no less than $20,000?

2.2 Common Discrete Distributions

Solution

- For the insurance company to make a profit of no less than \$20,000 from this life insurance policy, the maximum number of deaths is

$$\frac{2500\times 12-20000}{2000}=5.$$

- Let X be the number of deaths among the 2500 participants, then $X\sim\text{Binomial}(2500,0.002)$. We would like to compute

$$P(X\le 5)=\sum_{k=0}^{5}\binom{2500}{k}0.002^k 0.998^{2500-k}.$$

- Since $n=2500>100$ and $p=0.002<0.05$ in the binomial distribution, we apply the Poisson theorem, i.e., use $\text{Poisson}(2500*0.002)=\text{Poisson}(5)$ to approximate $\text{Binomial}(2500,0.002)$:

$$P(X\le 5)\approx\sum_{k=0}^{5}\frac{5^k}{k!}e^{-5}\approx 0.616.$$

2.2 Common Discrete Distributions

- In summary, we introduced the following discrete distributions:

Distribution	PMF	Expectation	Variance
Bernoulli(p)	$p(x) = p^x(1-p)^{1-x}, x = 0, 1$	p	$p(1-p)$
binomial(n, p)	$p(x) = \binom{n}{x} p^x(1-p)^{n-x}, x = 0, 1, 2, \dots, n$	np	$np(1-p)$
Geometric(p)	$p(x) = p(1-p)^{x-1}, x = 1, 2, \dots$	$1/p$	$(1-p)/p$
Poisson(λ)	$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots$	λ	λ

- There are many other discrete distributions that are not covered, for example:
  - Hypergeometric distribution (超几何分布): describes the number of successes in the first m trials of an n-fold Bernoulli trial.
  - Negative binomial distribution (负二项分布): a generalization of the geometric distribution, which describes the number of Bernoulli trials required until a predefined number r of successes occurs.

Chapter 2 Random Variables and Distributions

- 2.1 Introduction
- 2.2 Common Discrete Distributions
- 2.3 Common Continuous Distributions
- 2.4 Transformation of Random Variables

2.3 Common Continuous Distributions

- Uniform distribution (均匀分布) is the simplest continuous distribution.

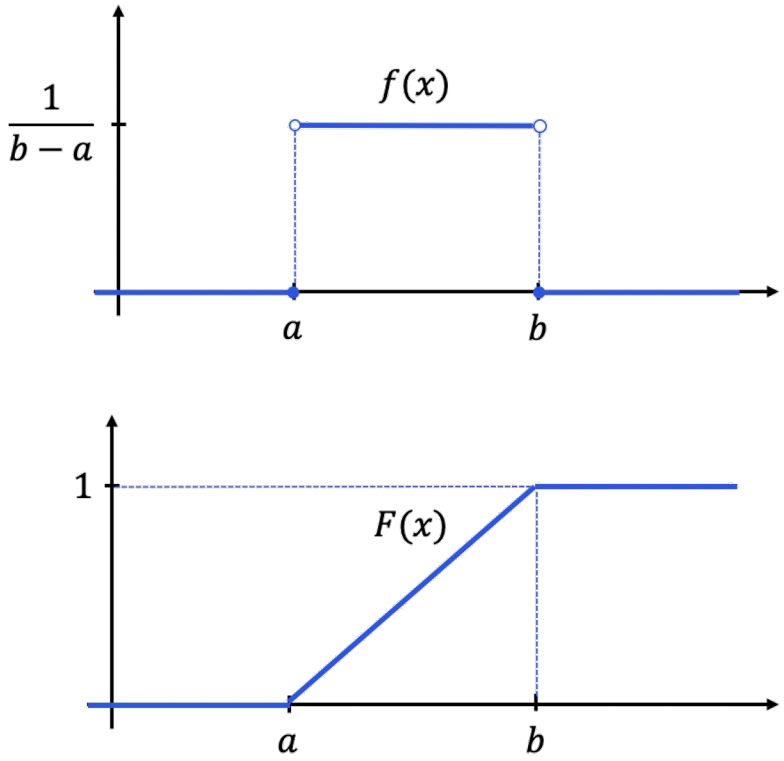
Uniform Distribution

- If the probability density function (PDF) of a random variable X is

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & \text{if } a<x<b\\ 0, & \text{otherwise}\end{cases},$$

- then X is said to follow a **uniform distribution (均匀分布)** on (a,b), denoted as $X\sim\text{Uniform}(a,b)$ or simply $X\sim\text{U}(a,b)$.
- The cumulative distribution function (CDF) of $X\sim\text{U}(a,b)$ is

$$F(x)=P(X\le x)=\int_{-\infty}^{x}f(u)du=\begin{cases}0, & \text{if } x\le a\\ \dfrac{x-a}{b-a}, & \text{if } a<x<b\\ 1, & if\ x\ge b\end{cases}.$$



2.3 Common Continuous Distributions

- For $\forall(c, c+L) \in (a, b)$, $P(c < X < c+L) = L/(b-a)$, which only depends on the length but not the position of the interval, suggesting kind of "equal likelihood (等可能性)".
- The expectation and variance of $X \sim \mathrm{U}(a, b)$ can be derived to be

$$\mathrm{E}(X) = \frac{a+b}{2}, \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

Proof:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2},$$

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{a^2+ab+b^2}{3},$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}.$$

2.3 Common Continuous Distributions

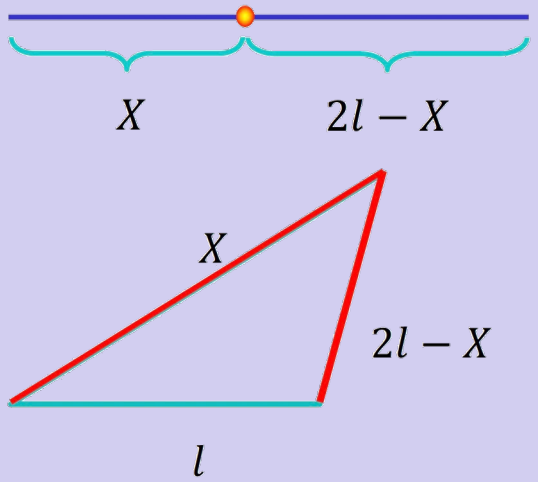
Example 3.9

- A wooden stick of length $2l$ is randomly cut into two pieces.
- What is the probability that these two pieces, together with another stick of length l, can form a triangle?

Solution

- Let the length of the two pieces be X and $2l - X$, then $X \sim \mathrm{U}(0, 2l)$.
- For the two pieces to form a triangle with another stick of length l, we need

$$\begin{cases} X + l > 2l - X \\ 2l - X + l > X \end{cases} \Leftrightarrow \frac{l}{2} < X < \frac{3l}{2}.$$



- Therefore, the probability is

$$P\left(\frac{l}{2} < X < \frac{3l}{2}\right) = \int_{l/2}^{3l/2} \frac{1}{2l}\, dx = \frac{1}{2}.$$

2.3 Common Continuous Distributions

- The next common continuous distribution is the exponential distribution (指数分布).

Exponential Distribution

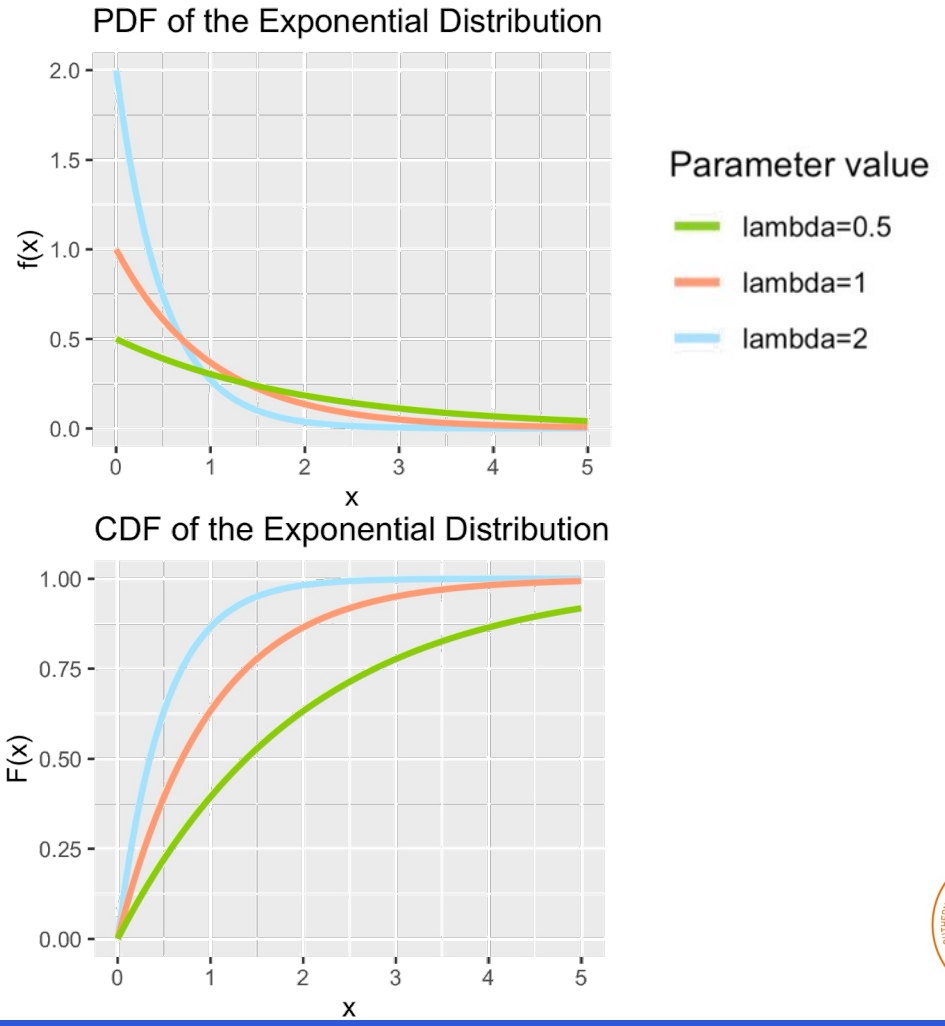
- If the PDF of a random variable X is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases},$$

- then X is said to follow an **exponential distribution (指数分布)** with parameter λ, denoted as $X \sim \text{Exp}(\lambda)$.
- The CDF of $X \sim \text{Exp}(\lambda)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

- In practice, the exponential distribution often arises as the distribution of the amount of time until some specific event occurs.

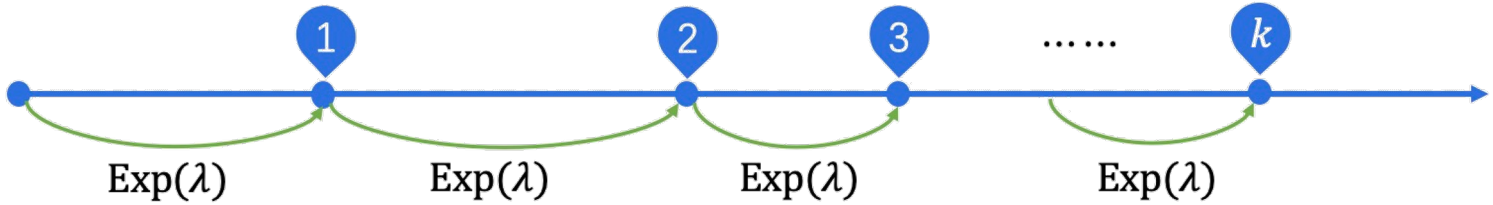


2.3 Common Continuous Distributions

- Have you noticed that the exponential distribution and the Poisson distribution use the same symbol λ for their parameter. Is there any relationship between them?
- In fact, the exponential distribution can be used to describe the distribution of the time intervals between events in a Poisson process (泊松过程).
  - A Poisson process can be simply understood as a process where random events occur independently and with a constant rate along the time axis.
  - The number of events occurring within a unit time interval follows $\text{Poisson}(\lambda)$, then the number of events occurring in $[0, t]$ follows $\text{Poisson}(\lambda t)$.
  - Let X be the time until the first event occurs, then

$$P(X \le t) = 1 - P(X > t) = 1 - P(\text{no event occured in } [0, t]) = 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} = 1 - e^{-\lambda t},$$

  - which means that $X \sim \text{Exp}(\lambda)$.
  - Similarly, we can show that the time intervals between events independently follow $\text{Exp}(\lambda)$.



2.3 Common Continuous Distributions

- The expectation and variance of $X \sim \text{Exp}(\lambda)$ can be derived to be

$$\text{E}(X) = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}.$$

Proof:

$$\text{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = -\int_0^{\infty} x d\left(e^{-\lambda x}\right) = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{\infty} = \frac{1}{\lambda},$$

$$\text{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = -\int_0^{\infty} x^2 d\left(e^{-\lambda x}\right) = 2\int_0^{\infty} x e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

$$\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

- For a Poisson process with intensity λ, i.e., the expected number of events in a unit time interval is λ, its natural that the expected time interval between events is $1/\lambda$.

2.3 Common Continuous Distributions

Example 3.10

- There was an extraordinary rainstorm in Hong Kong on September 7th, 2023.
- The news reported that it was a once-in-500-years torrential rain.
- 500 year ago, it was the second year of the Jiajing reign of the Ming Dynasty (明朝嘉靖二年).
- Does it mean that this is the heaviest rain since the Ming Dynasty?

News: https://www.hk01.com/article/939036?utm_source=01articlecopy&utm_medium=referral

2.3 Common Continuous Distributions

Answer

- In fact, the Observatory (天文台) describe the rain as a "once-in-100-years rain" and the Drainage Services Department (渠务署) referred to it as a "once-in-500-years rain".
- The Observatory and the Drainage Services Department have different computation method.
- The Drainage Services Department's computation is based on flood prevention standards.
- The Observatory's "once-in-XXX-years" is the event rate of the Poisson process describing the event of rainfall with the same scale.
- Therefore, the time interval between two "once-in-100-years rain" follows **Exp(1/100)**.
- The expected time interval between two "once-in-100-years rain" is 100 years, but is doesn't mean that we need to wait exactly 100 years to have one such rain.

其實新聞報道中常提及的「幾多年一遇」是表示概率。天文台與渠務署的計算公式亦不一樣。簡單而言，天文台的「多少年一遇」，是基於以往錄得的雨量及出現頻率作為統計基礎作出推算，是數據上的結論；而渠務署所指的「多少年一遇」，其實是防洪標準，一般市區排水幹渠系統足以應付重現期為200年一遇的暴雨，而今次「500年一遇」的暴雨侵襲，最後便導致多區出現水浸。

2.3 Common Continuous Distributions

- The exponential distribution is the only continuous distribution with the memoryless property (无记忆性), i.e., if $X \sim \text{Exp}(\lambda)$, then for any $s, t > 0$:

$$P(X > s + t | X > s) = P(X > t).$$

Proof: For any $s, t > 0$, since $\{X > s + t\} \subset \{X > s\}$, it follows that

$$P(\{X > s + t\} \cap \{X > s\}) = P(X > s + t) = 1 - F(s + t) = e^{-\lambda(s+t)}.$$

Then, by the definition of conditional probability:

$$P(X > s + t | X > s) = \frac{P(\{X > s + t\} \cap \{X > s\})}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).$$

- The memoryless property greatly simplifies analysis, however, it makes the exponential distribution inappropriate for many real world applications.
- **Think:** Recall that the geometric distribution is the only discrete distribution with the memoryless property, is there any relationship between the geometric and exponential distributions?
- The ceiling (上取整) of an exponential random variable follows geometric distribution. (HW2)

2.3 Common Continuous Distributions

- The normal distribution (正态分布) is the most important distribution, without exception.

Normal Distribution

- If the PDF of a random variable X is

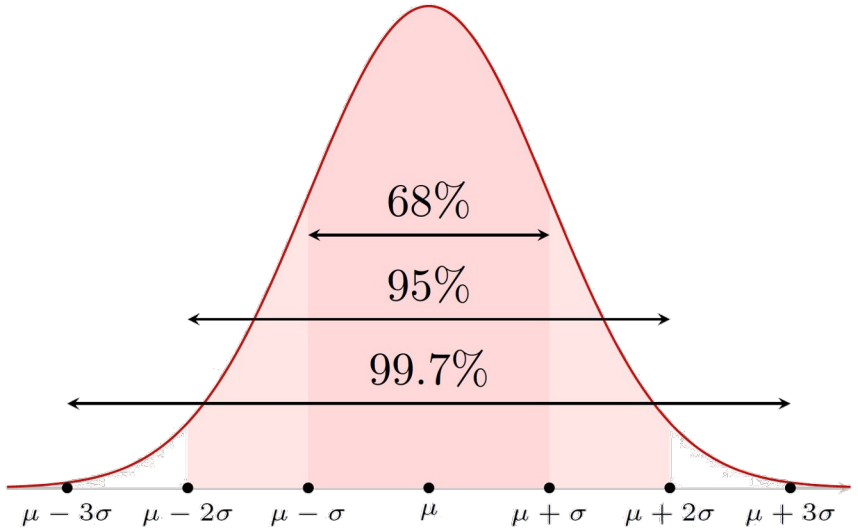
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty,$$

- then X is said to follow a **normal distribution (正态分布)** with parameter μ and σ^2 $(\sigma>0)$, denoted as $X\sim\mathrm{N}(\mu,\sigma^2)$.
- Specifically, $\mathrm{N}(0,1)$ is the **standard normal distribution (标准正态分布)**, with PDF

$$\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},-\infty<x<\infty.$$

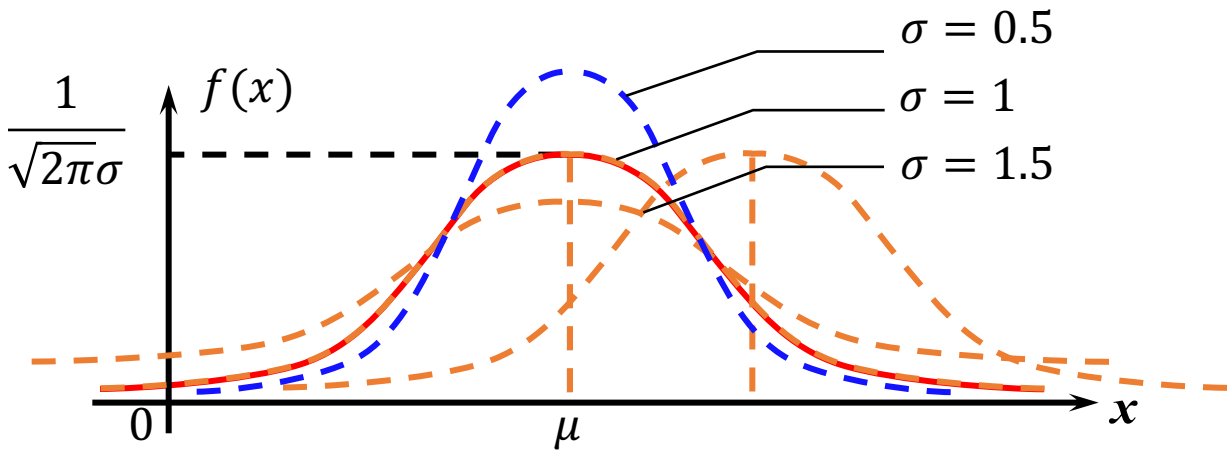
- The CDF of $X\sim\mathrm{N}(0,1)$ has no explicit expression, however, it is used very often and thus expressed as $\Phi(x)$:

$$\Phi(x)=P(X\le x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du.$$



- The PDF of $\mathrm{N}(\mu,\sigma^2)$ is an elegant bell-shaped curve, symmetric about the parameter μ.

2.3 Common Continuous Distributions



- $\mu \uparrow$: The PDF moves from the left to the right;
- $\mu \downarrow$: The PDF moves from the right to the left;
- $\sigma \uparrow$: The PDF becomes flatter;
- $\sigma \downarrow$: The PDF becomes sharper;

- If $X \sim \mathrm{N}(\mu, \sigma^2)$ and define r.v. $Z = \frac{X-\mu}{\sigma}$, then $Z \sim \mathrm{N}(0,1)$.

Standardize (标准化)

Proof: For $Z = \frac{X-\mu}{\sigma}$, consider its CDF:

$$F_Z(x) = P(Z \le x) = P\left(\frac{X-\mu}{\sigma} \le x\right)$$
$$= P(X \le \sigma x + \mu) = F_X(\sigma x + \mu).$$

By differentiation, the PDF of Z is given by

$$f_Z(x) = \frac{dF_X(\sigma x+\mu)}{dx} = \sigma f_X(\sigma x + \mu)$$
$$= \frac{\sigma}{\sqrt{2\pi}\sigma} e^{-\frac{(\sigma x+\mu-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

which shows that of $Z \sim \mathrm{N}(0,1)$.

2.3 Common Continuous Distributions

- The expectation and variance of $X \sim \mathrm{N}(\mu, \sigma^2)$ can be derived to be

$$\mathrm{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2.$$

Proof: Let $Z = \frac{X-\mu}{\sigma}$, it suffices to prove $\mathrm{E}(Z) = 0, \mathrm{Var}(Z) = 1$.

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} x\phi(x)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}}dx = -\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big|_{-\infty}^{\infty} = 0.$$

$$\mathrm{E}(Z^2) = \int_{-\infty}^{\infty} x^2\phi(x)\,dx = -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\,d\left(e^{-x^2/2}\right) = -\frac{1}{\sqrt{2\pi}}\left(xe^{-\frac{x^2}{2}}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx\right)$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx = \int_{-\infty}^{\infty} \phi(x)dx = 1.$$

By the normalization of a PDF

$$\mathrm{Var}(Z) = E(Z^2) - [E(Z)]^2 = 1.$$

2.3 Common Continuous Distributions

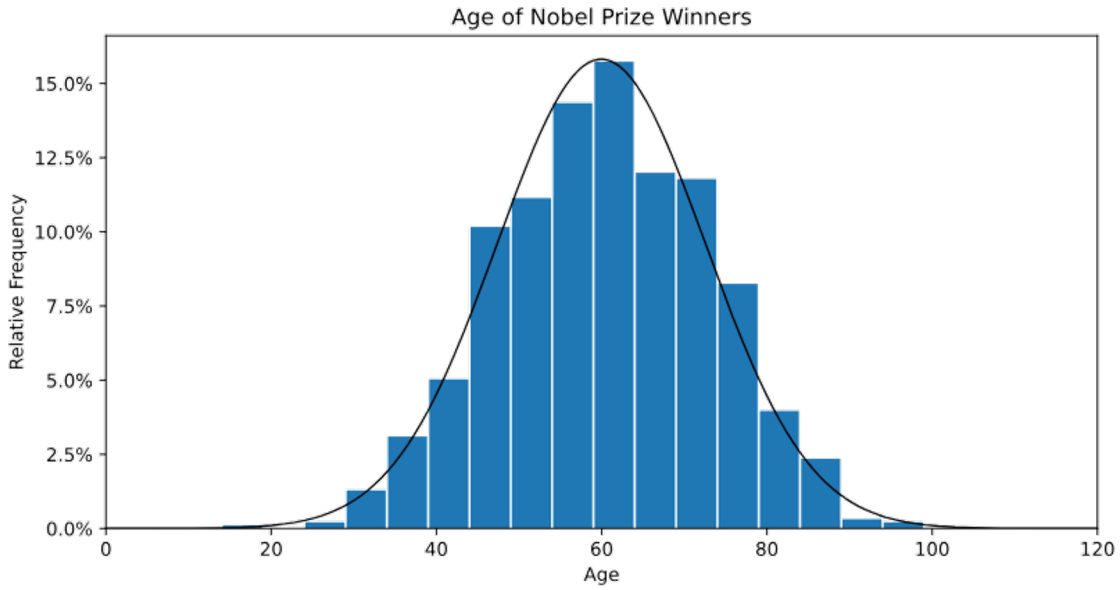
- The normal distribution is also called the **Gaussian distribution (高斯分布)**.

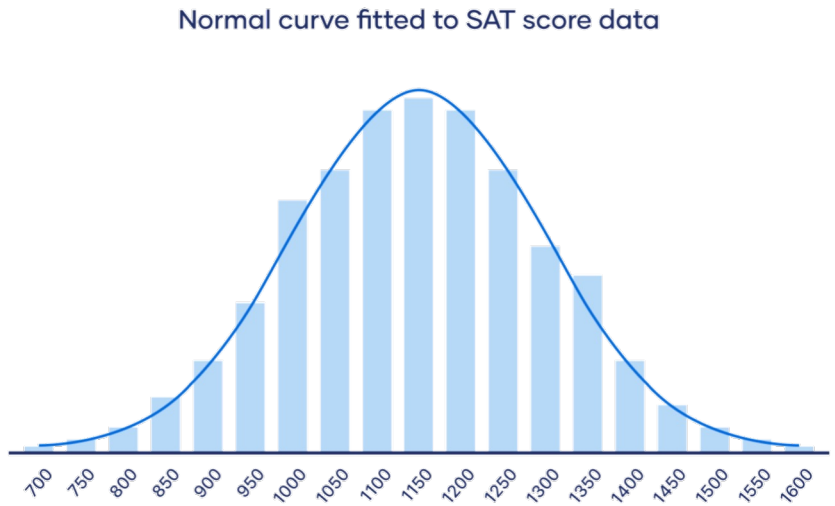
Carl Gauss, German mathematician, astronomer, and physicist, "greatest mathematicians of all time"

- Many people believe that it was Gauss who discovered the normal distribution.
- While Gauss did play a key role in establishing the significance of the normal distribution in history, he was not the first to propose the distribution.
- The French mathematician Poincaré suggested using the neutral term "normal distribution".

2.3 Common Continuous Distributions

- The normal distribution plays a vital role in Probability and Statistics, mostly because of the **Central Limit Theorem (CLT, 中心极限定理)**, which will be introduced in the next chapter.
- The CLT states that the sum/average of r.v.'s generally follows a normal distribution.
- Due to this fact, various fluctuations and measurement errors appear normally distributed.
- Moreover, normal distribution is often found to be a good model for weight, height, intelligence, temperature, pollution level, student grades, etc.





Galton board (高尔顿钉板): https://haokan.baidu.com/v?pd=wisenatural&vid=13179278539362444762

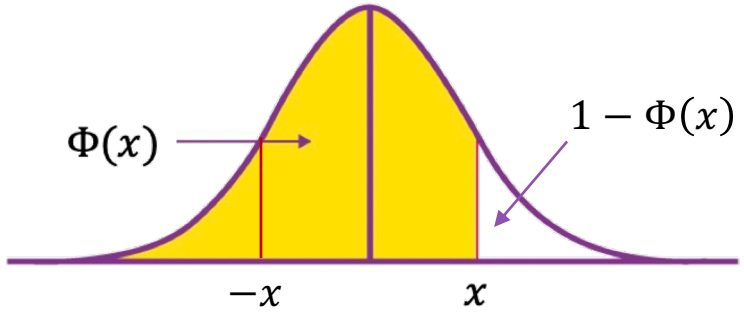
2.3 Common Continuous Distributions

- If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim \mathrm{N}(0,1)$ and consequently,

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right)$$
$$= P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- Moreover, by symmetry of the PDF of $\mathrm{N}(0,1)$:

$$\Phi(-x) = 1 - \Phi(x), -\infty < x < \infty.$$



Standard Normal Probabilities

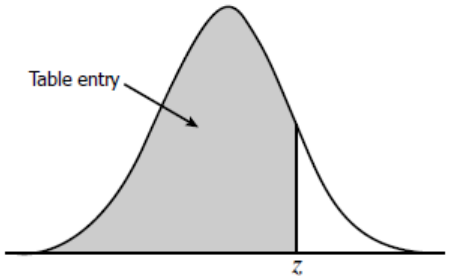


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

- Since $\Phi(x)$ does not have analytical expression, we would check its values in a probability table. E.g., $\Phi(1.96) = 0.975$.

2.3 Common Continuous Distributions

Example 3.11

- An expert witness (专家证人) in a paternity suit (亲子诉讼案) testifies that the length (in days) of human gestation (妊娠期) is approximately normally distributed with parameters $\mu = 270$ and $\sigma^2 = 100$.
- The defendant in the suit is able to prove that he was out of the country during a period between 290 days and 240 days before the child was born.
- If the defendant is, in fact, the father of the child, what is the probability that the mother could have had the very long or very short gestation indicated in her testimony(证词)?

Solution

- Let the mother's length of gestation be X, then $X \sim \mathrm{N}(270,100)$.
- The probability to be computed is $P(\{X > 290\} \cup \{X < 240\})$. Consider

$$P(X > 290) = P\left(\frac{X-270}{10} > \frac{290-270}{10}\right) = P(Z > 2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228.$$

$$P(X < 240) = P\left(\frac{X-270}{10} < \frac{240-270}{10}\right) = P(Z < -3) = 1 - \Phi(3) = 1 - 0.9987 = 0.0013.$$

- Therefore, the probability is $0.0228 + 0.0013 = 0.0241$.

2.3 Common Continuous Distributions

Example 3.12

- A bus manufacturer is designing a bus. When determining the door height, they must ensure that it is not too high but also allows 99% of male passengers to pass through without bending.
- Assuming the height of all males (in cm) follows a normal distribution $\mathrm{N}(170, 36)$, what should be the minimum door height to meet this requirement?

Solution

- Let X denote the height of a randomly selected male, h be the door height of the bus.
- Then the requirement can be expressed as $P(X \le h) \ge 0.99$.
- Checking the probability table inversely, we find that for $Z \sim \mathrm{N}(0,1)$, $P(Z \le 2.33) = 0.9901 > 0.99$. So

$$P(X \le h) = P\left(\frac{X-170}{6} \le \frac{h-170}{6}\right) = P\left(Z \le \frac{h-170}{6}\right) \ge 0.99.$$

$$\Rightarrow \frac{h-170}{6} \ge 2.33 \quad \Rightarrow h \ge 170 + 13.98 \approx 184 \text{ cm}.$$

2.3 Common Continuous Distributions

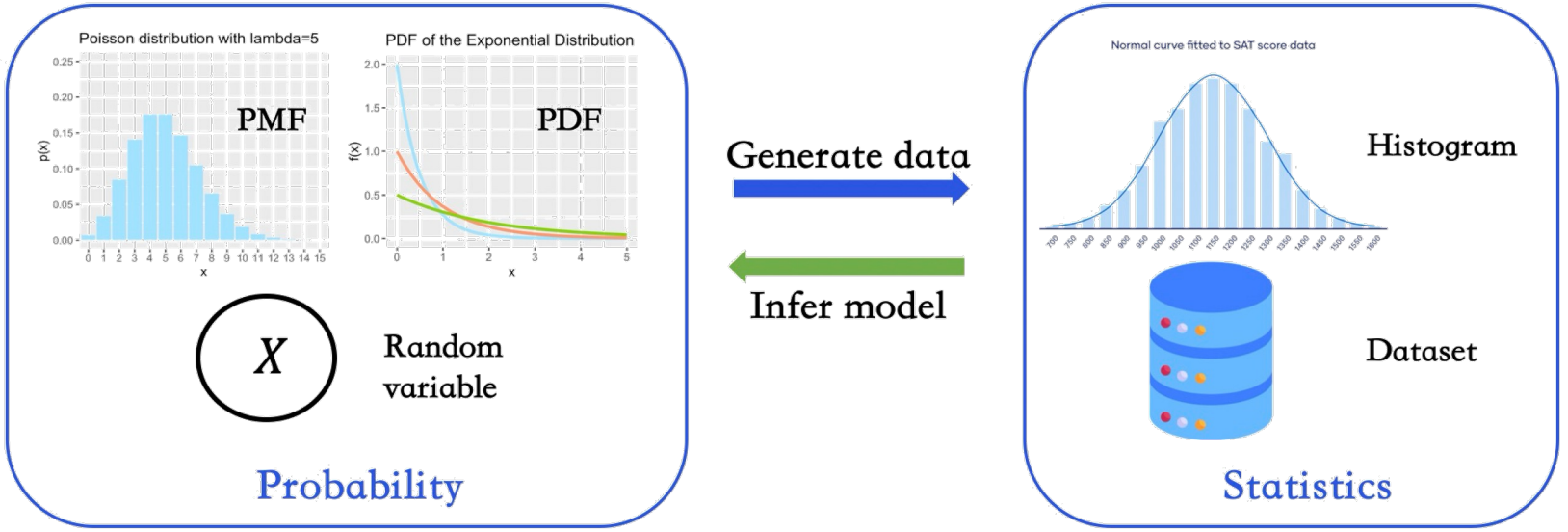
- In summary, we introduced the following continuous distributions:

Distribution	PDF	Expectation	Variance
Uniform(a, b)	$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
N(μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$	μ	σ^2

- There are many other continuous distributions that are not covered here, for example:
  - **Beta distribution (贝塔分布)**: typically used to describe the distribution of a random variable with support $[0, 1]$, widely used in Bayesian Statistics.
  - **Gamma distribution (伽马分布)**: a generalization of the exponential distribution, widely used for the total time of a multistage scheme.

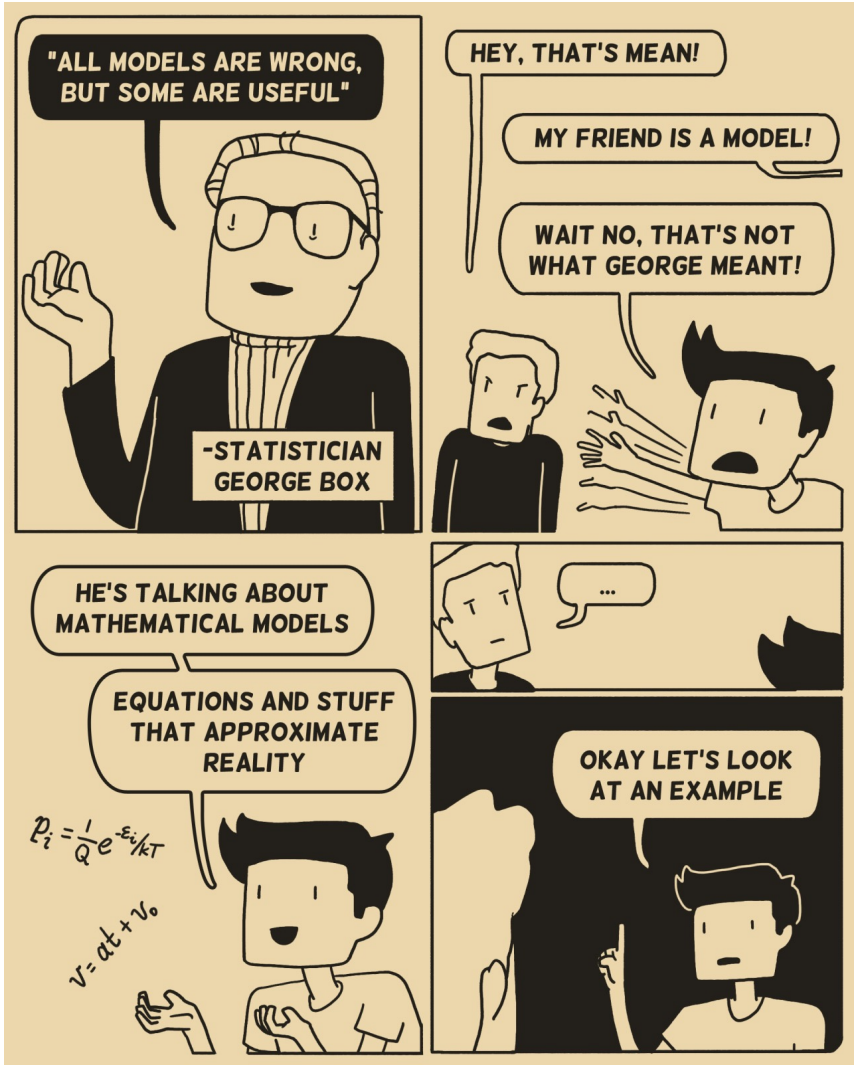
2.3 Common Continuous Distributions

- **Question:** What is the essence of a probability distribution?
- Typically, we can also plot a distribution based on the data we collect, e.g., using the histogram (直方图). How does this relate to the probability distribution mentioned?
  - The distribution plotted from collected data is just a form of data visualization.
  - A probability distribution does not correspond to a specific dataset, but is **a model**.
  - We can use this model to describe the mechanism of data generation or to summarize the patterns underlying the data.



2.3 Common Continuous Distributions

- Probability distributions are essentially models:
  - We usually assume a probability distribution model for the data, and then verify the assumption with the data collected.
  - It's hard to get a perfect match, but a certain range of deviation is acceptable.
  - "All models are wrong, but some are useful".
  - In fact, the model itself is neither right nor wrong; the mistake lies in choosing the wrong model for a specific problem.
  - Probability distributions is like a toolbox, and each distribution is a specific tool inside the toolbox.
  - When faced with a problem, we need to find an appropriate tool from the toolbox to solve it.



Chapter 2 Random Variables and Distributions

- 2.1 Introduction
- 2.2 Common Discrete Distributions
- 2.3 Common Continuous Distributions
- 2.4 Transformation of Random Variables

2.4 Transformation of Random Variables

- Sometimes, we may know the distribution of a r.v. X and would like to derive the distribution of some function of the r.v., i.e., $Y = g(X)$. For example:
  - Suppose that you invested ￥1000 in an account with continuous compounding interest rate R (连续复利)
  - R is a realization of a continuous r.v. with PDF $f(r)$.
  - How is the amount in the account after one year, i.e., $A = 1000e^R$, distributed?

  - A scientist measures the radius of a circle and the result is a random variable (denoted by R) due to the measurement error.
  - R is a realization of a continuous r.v. with PDF $f(r)$.
  - What is the distribution of the computed area of the circle, i.e., $A = \pi R^2$?

2.4 Transformation of Random Variables

- For a discrete r.v. X with PMF $p_X(x)$, it is not difficult to determine the PMF of $Y = g(X)$:

$$p_Y(y) = P(Y = y) = \sum_{x:g(x)=y} p_X(x),$$

Discrete to Discrete

- which consider both cases where g is one-to-one and not one-to-one.

Example 3.13

- The PMF of r.v. X is given below, obtain the PMF of $Y = (X-1)^2$.

x	-1	0	1	2
$p_X(x)$	0.2	0.3	0.1	0.4

Solution

- For each possible value x of X consider $y = (x-1)^2$:

y	4	1	0	1
$p_Y(y)$	0.2	0.3	0.1	0.4

- Combine the probability of the same y:

y	0	1	4
$p_Y(y)$	0.1	0.7	0.2

2.4 Transformation of Random Variables

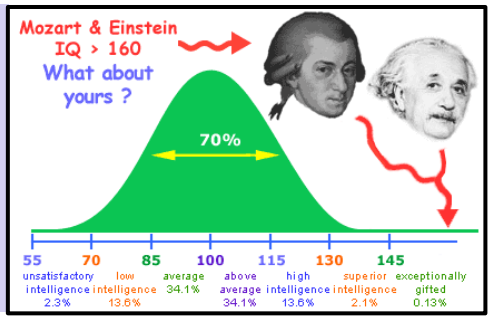
- For a continuous r.v. X with PDF $f_X(x)$, if $Y = g(X)$ is a discrete r.v., then the PMF of Y is:

$$p_Y(y) = P(Y = y) = \int_{x:g(x)=y} f_X(x)\,dx.$$

Continuous to Discrete

Example 3.14

- Suppose that the IQ test score of a randomly selected person is $X \sim N(100, 225)$.
- A random variable Y is defined to be $Y = \begin{cases} 1, & \text{if } X \le 85 \\ 2, & \text{if } 85 < X \le 115 \\ 3, & \text{if } X > 115 \end{cases}$.
- What is the PMF of Y?



Solution

$$P(Y = 1) = P(X \le 85) = P\left(\frac{X - 100}{15} \le \frac{85 - 100}{15}\right)$$

$$= P(Z \le -1) = 1 - P(Z \le 1) = 0.1587.$$

- Therefore, the PMF of Y is

y	1	2	3
$p_Y(y)$	0.1587	0.6826	0.1587

2.4 Transformation of Random Variables

- For a continuous r.v. X with PDF $f_X(x)$, if $Y = g(X)$ is also a continuous r.v., then deriving the PDF of Y is a bit more complicated.
- If $g(x)$ is a strictly monotonic function (严格单调函数) on the support of X, and it has a continuously-differentiable inverse function $h(y) = g^{-1}(y)$, then the PDF of Y is

$$f_Y(y) = \begin{cases} |h'(y)| \cdot f_X\big(h(y)\big), & \text{where } h(y) \text{ is defined;} \\ 0, & \text{otherwise.} \end{cases}$$

Continuous to Continuous

Proof: Consider the CDF of Y: $F_Y(y) = P(Y \le y) = P(g(X) \le y)$.

If $g(x)$ is a strictly increasing function, then $h'(y) > 0$ and:

$$F_Y(y) = P\big(X \le g^{-1}(y)\big) = F_X\big(h(y)\big) \Longrightarrow f_Y(y) = F_Y'(y) = h'(y) f_X\big(h(y)\big).$$

If $g(x)$ is a strictly decreasing function, then $h'(y) < 0$ and:

$$F_Y(y) = P\big(X \ge g^{-1}(y)\big) = 1 - F_X\big(h(y)\big) \Longrightarrow f_Y(y) = F_Y'(y) = -h'(y) f_X\big(h(y)\big).$$

Put these two cases together, the PDF of Y is the one given in the formula above in red.

2.4 Transformation of Random Variables

Example 3.15

- Consider the time it takes to transfer a file over a network depends on the network speed X, which vary due to traffic and other conditions and $X \sim \text{Uniform}[2, 4]$ (in Mbps).
- Let Y denote the time required to transfer a 100Mb file, please derive the PDF of Y.

Solution

- According to the scenario described, we have:

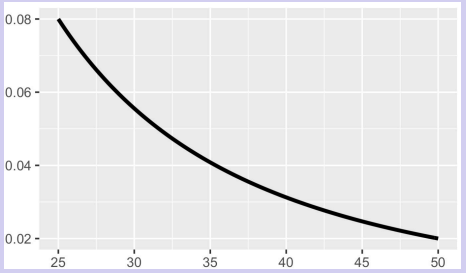
$$f_X(x) = \begin{cases} 1/2, & \text{if } 2 \le x \le 4 \\ 0, & \text{otherwise} \end{cases} \text{ and } Y = g(X) = \frac{100}{X}.$$

- $g(x) = 100/x$ is a strictly decreasing function on $[2, 4]$ and its inverse function is $h(y) = 100/y$. So

$$|h'(y)| \cdot f_X(h(y)) = \frac{100}{y^2} \cdot \frac{1}{2} = \frac{50}{y^2}.$$

- The PDF of Y is

$$f_Y(y) = \begin{cases} 50/y^2, & \text{if } 25 \le y \le 50 \\ 0, & \text{otherwise} \end{cases}.$$



2.4 Transformation of Random Variables

- A famous application of the transformation of r.v.s is based on the following results:
- If the CDF of a continuous r.v. X is $F(x)$ and its inverse function $F^{-1}(x)$ exists. Define a r.v. $Y = F(X)$, then $Y \sim \text{Uniform}(0,1)$.
- On the other hand, if $F(x)$ is the CDF of some r.v. and its inverse function $F^{-1}(x)$ exists, let $U \sim \text{Uniform}(0,1)$, then for $X = F^{-1}(U)$ we have $X \sim F(x)$, i.e., the CDF of X is $F(x)$.

Proof: Consider the CDF of Y: $F_Y(y) = P(Y \le y) = P(F(X) \le y)$.

Since $F(x)$ is a non-decreasing function and $F^{-1}(x)$ exists, then for any $y \in [0,1]$:

$$F_Y(y) = P\left(X \le F^{-1}(y)\right) = F\left(F^{-1}(y)\right) = y \Longrightarrow f_Y(y) = F_Y'(y) = 1.$$

This suggest that $Y \sim \text{Uniform}[0,1]$. The proof of the second result is similar and omitted here.

- The second result can be used in random number sampling, which is the called the inverse transform sampling (逆变换采样).
- It is a widely used technique for generating random samples from a complicated distribution.

2.4 Transformation of Random Variables

- Under the case when $g(x)$ is not a strictly monotonic function on the support of X, how to derive the PDF of $Y = g(X)$?

Example 3.16

- Assume that r.v. $X \sim \mathrm{N}(0,1)$, what is the PDF of $Y = X^2$?

chi-squared (χ^2) distribution (卡方分布)

Solution

- Since $X \sim \mathrm{N}(0,1)$, the PDF of is $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$.
- Although $g(x) = x^2$ is not a monotonic function, however, it is strictly increasing on $(0,\infty)$ and strictly decreasing on $(-\infty, 0)$. For any $y \in (0,\infty)$, we have

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$$

$$\Rightarrow F_Y'(y) = \phi(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} - \phi(-\sqrt{y}) \cdot \left(-\frac{1}{2\sqrt{y}}\right) = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} \Rightarrow f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}, y > 0 \\ 0, \quad \text{otherwise} \end{cases}.$$

2.4 Transformation of Random Variables

Example 3.15 (Continued)

- Compare the probabilities $P(25 \le Y \le 30)$ and $P(45 \le Y \le 50)$.
- Calculate $\mathrm{E}(Y)$, i.e., the expected time required to transfer a 100Mb file.

Solution

- Based on the PDF of Y, it is not difficult to obtain

$$P(25 \le Y \le 30) = \int_{25}^{30} \frac{50}{y^2}\,dy = -\frac{50}{y}\Big|_{25}^{30} = \frac{50}{25} - \frac{50}{30} \approx 0.333,\ P(45 \le Y \le 50) = \frac{50}{45} - \frac{50}{50} \approx 0.111.$$

- By the definition of expectation, we have

$$\mathrm{E}(Y) = \int_{25}^{50} y \cdot \frac{50}{y^2}\,dy = 50 \ln y\Big|_{25}^{50} = 50 \ln 50 - 50 \ln 25 \approx 34.657.$$

- What's the relationship between $\mathrm{E}\left(\frac{100}{X}\right)$ and $\frac{100}{\mathrm{E}(X)}$?
- In general, do we have a rule describing the relationship between $\mathrm{E}\big(g(X)\big)$ and $g\big(\mathrm{E}(X)\big)$?

2.4 Transformation of Random Variables

- Actually, to calculate the expectation of $Y = g(X)$, there is no need to derive the PMF/PDF of Y first, we can use the PMF/PDF of X directly:

  - If X is a **discrete r.v.** with PMF $P(X = x_k) = p_k, k = 1, 2, \ldots$, given $\sum_{k=1}^{\infty} |g(x_k)| p_k < \infty$, then

$$\mathrm{E}(Y) = \mathrm{E}\big(g(X)\big) = \sum_{k=1}^{\infty} g(x_k) p_k \,.$$

  - If X is a **continuous r.v.** with PDF $f(x)$, given $\int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty$, then

$$\mathrm{E}(Y) = \mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f(x) dx \,.$$

Solution of Example 3.15 (Continued)

- The expectation of Y can also be obtained by directly using $f_X(x)$:

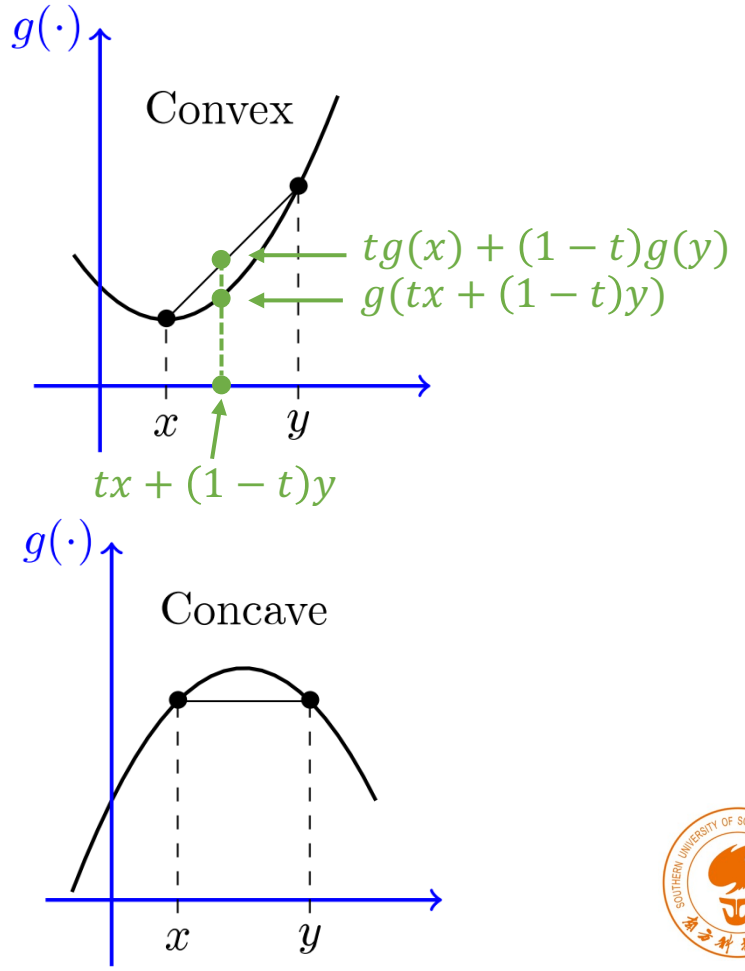
$$\mathrm{E}(Y) = \int_2^4 \frac{100}{x} \cdot \frac{1}{2} dx = 50 \ln x \Big|_2^4 = 50 \ln 4 - 50 \ln 2 \approx 34.657.$$

2.4 Transformation of Random Variables

- Then we come back to the question that do we have a rule describing the relationship between $\mathrm{E}\big(g(X)\big)$ and $g\big(\mathrm{E}(X)\big)$ in general?
- We do have such a rule for certain types of functions.
- E.g., if $g(x) = ax + b$, then we always have $\mathrm{E}\big(g(X)\big) = g\big(\mathrm{E}(X)\big)$.
- Any cases other than the linear function?

Convex and Concave Function

- A function $g: S \to \mathbb{R}$ is said to be a **convex function** (凸函数), if for any $t \in [0,1]$ and $x, y \in S$, we have
$$g(tx + (1-t)y) \leq t\,g(x) + (1-t)g(y).$$
- A function $g: S \to \mathbb{R}$ is said to be a **concave function** (凹函数), if for any $t \in [0,1]$ and $x, y \in S$, we have
$$g(tx + (1-t)y) \geq t\,g(x) + (1-t)g(y).$$



2.4 Transformation of Random Variables

Jensen's Inequality (琴生不等式)

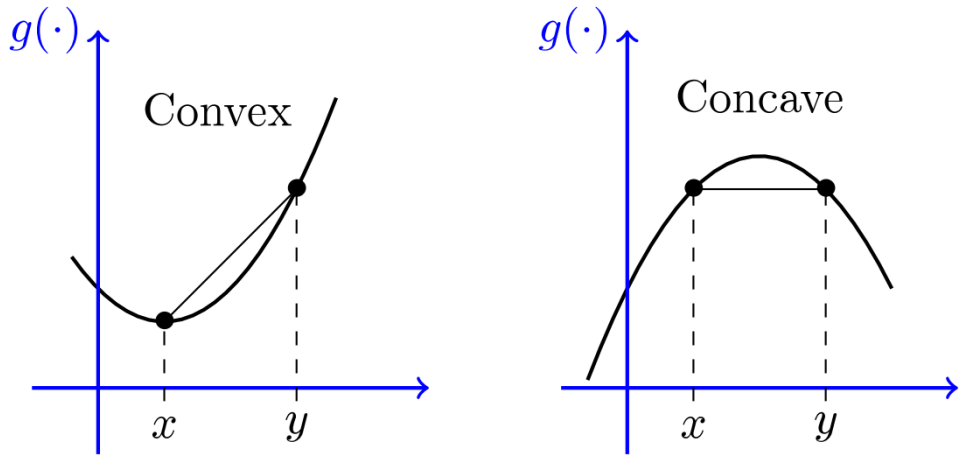
Let X be a random variable, then

- for any **convex** function g,

$$\mathrm{E}\big(g(X)\big) \geq g\big(\mathrm{E}(X)\big).$$

- for any **concave** function g,

$$\mathrm{E}\big(g(X)\big) \leq g\big(\mathrm{E}(X)\big).$$



- By the Jensen's inequality, we have the following results:
  - $\mathrm{E}(|X|) \geq |\mathrm{E}(X)|$ $(g(x) = |x|)$;
  - $\mathrm{E}(X^2) \geq \big(\mathrm{E}(X)\big)^2$ $(g(x) = x^2)$;
  - $\mathrm{E}(|X|^p) \geq |\mathrm{E}(X)|^p$ for $p \geq 1$ $(g(x) = |x|^p, p \geq 1)$;
  - $\mathrm{E}(e^{cX}) \geq e^{c\mathrm{E}(X)}$ $(g(x) = e^{cx})$.

The Jensen's inequality has many applications in information theory, machine learning, and optimization, etc.

2.4 Transformation of Random Variables

Example 3.17

- One of the application of Jensen's inequality is related to the **Kullback-Leibler divergence (KL divergence, KL散度)**.
- KL divergence is called **the information gain (信息增益)** in the context of decision trees and also called the **relative entropy (相对熵)**.
- You may get to know this concept latter when doing coursework in machine learning or information theory. The concept is actually pretty straightforward.
- Put it simply, if you have two probability distributions $p(x)$ and $q(x)$, the KL divergence measures the difference/distance between them:

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

- The KL divergence has the property that $D_{KL}(p||q) \geq 0$ and $D_{KL}(p||q) = 0$ if and only if $q(x) = p(x)$ almost every where.
- Can you show that $D_{KL}(p||q) \geq 0$ using the Jensen's inequality?

2.4 Transformation of Random Variables

Solution

- Let X be a r.v. with PDF $p(x)$, define another r.v. $Y = q(X)/p(X)$.
- Let function $g(x) = -\log(x)$, then $g(x)$ is a convex function, so that by the Jensen's inequality:

$$\mathrm{E}\big(g(Y)\big) \geq g\big(\mathrm{E}(Y)\big).$$

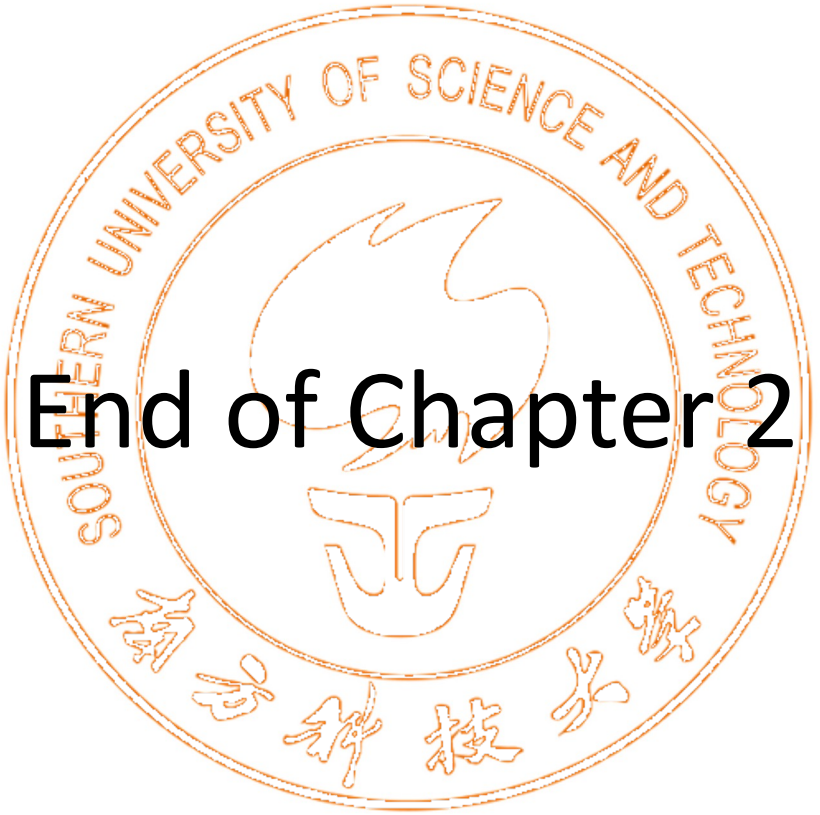
- Since

$$\mathrm{E}(g(Y)) = E\left(g\left(\frac{q(X)}{p(X)}\right)\right) = -\int_{-\infty}^{\infty} p(x)\log\left(\frac{q(x)}{p(x)}\right)dx = D_{KL}(p||q),$$

$$\mathrm{E}(Y) = E\left(\frac{q(X)}{p(X)}\right) = \int_{-\infty}^{\infty} p(x)\frac{q(x)}{p(x)}dx = \int_{-\infty}^{\infty} q(x)dx = 1 \Longrightarrow g\big(E(Y)\big) = -\log(1) = 0.$$

- Put these together, we have

$$D_{KL}(p||q) \geq 0.$$



End of Chapter 2