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.

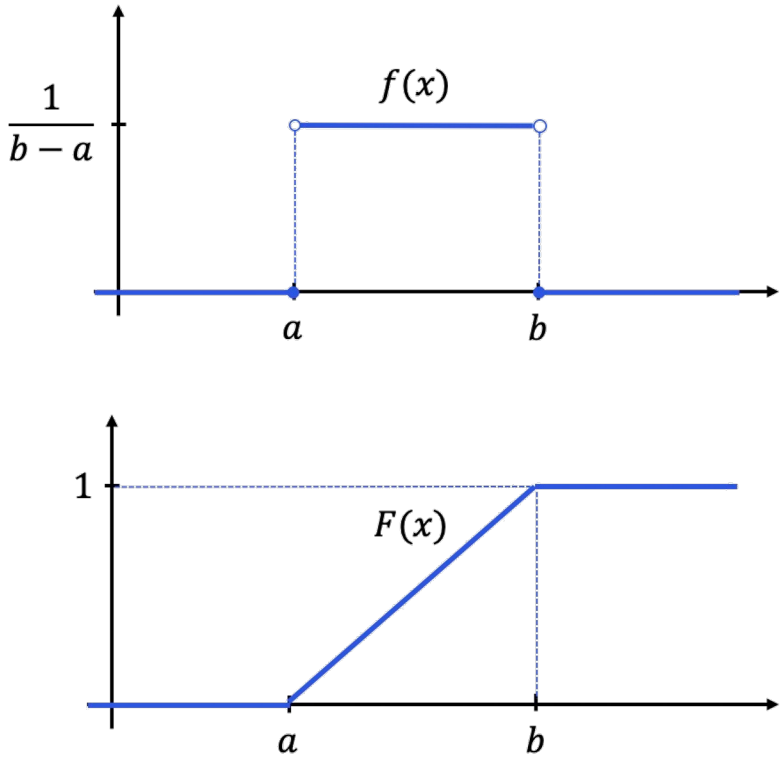
Uniform Distribution

- 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

2.3 Common Continuous Distributions

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

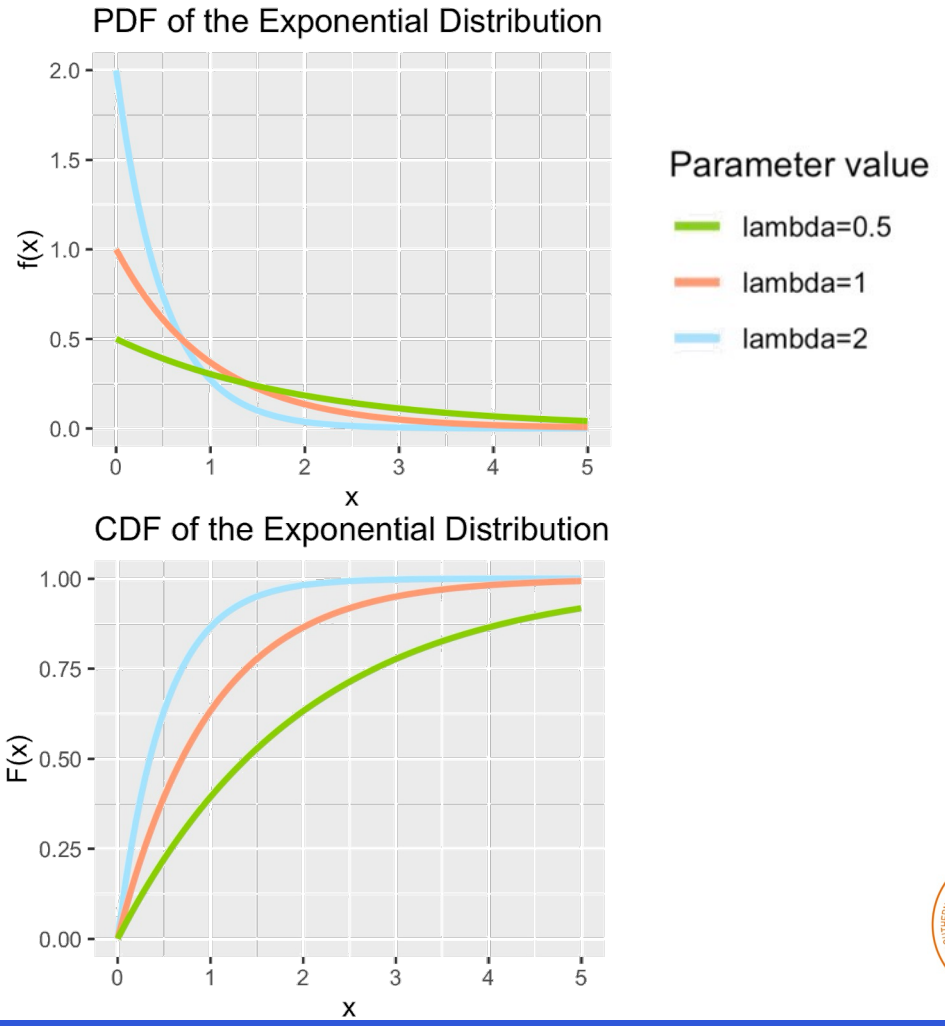
Exponential Distribution

- 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

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



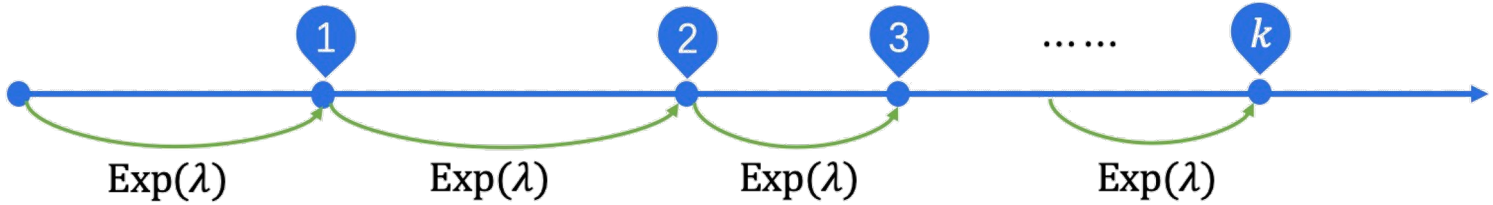
- In practice, the exponential distribution often arises as the distribution of the amount of time until some specific event occurs.

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

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?

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

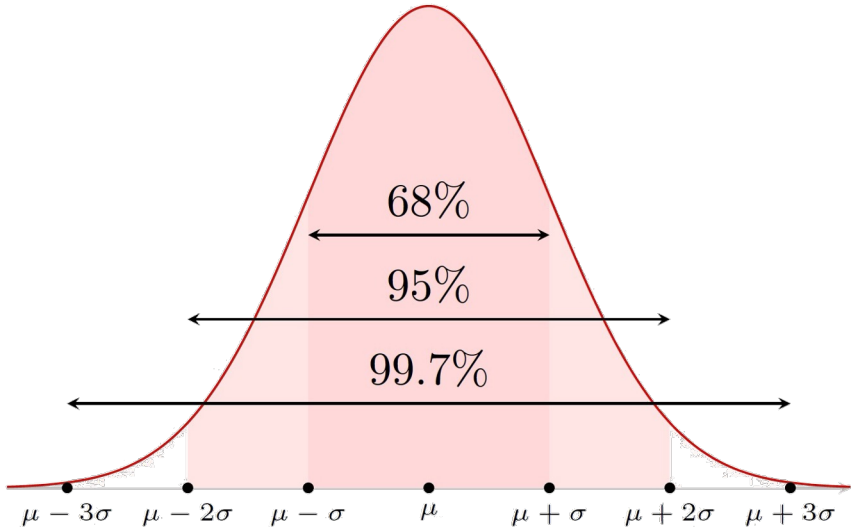
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty,$$

- 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

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

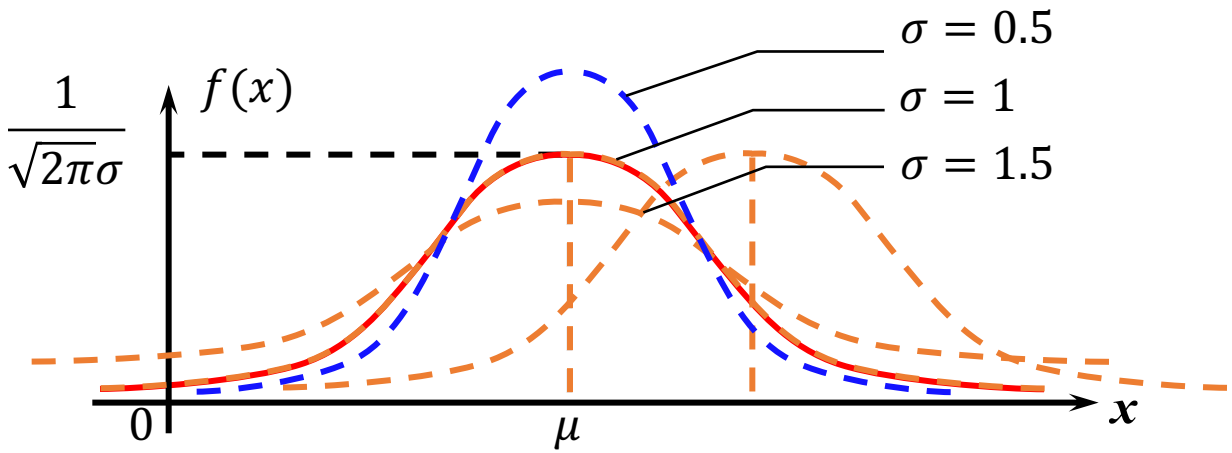
- The CDF of $X\sim\mathrm{N}(0,1)$ has no explicit expression, however, it is used very often and thus expressed as $\Phi(x)$:

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



- 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

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

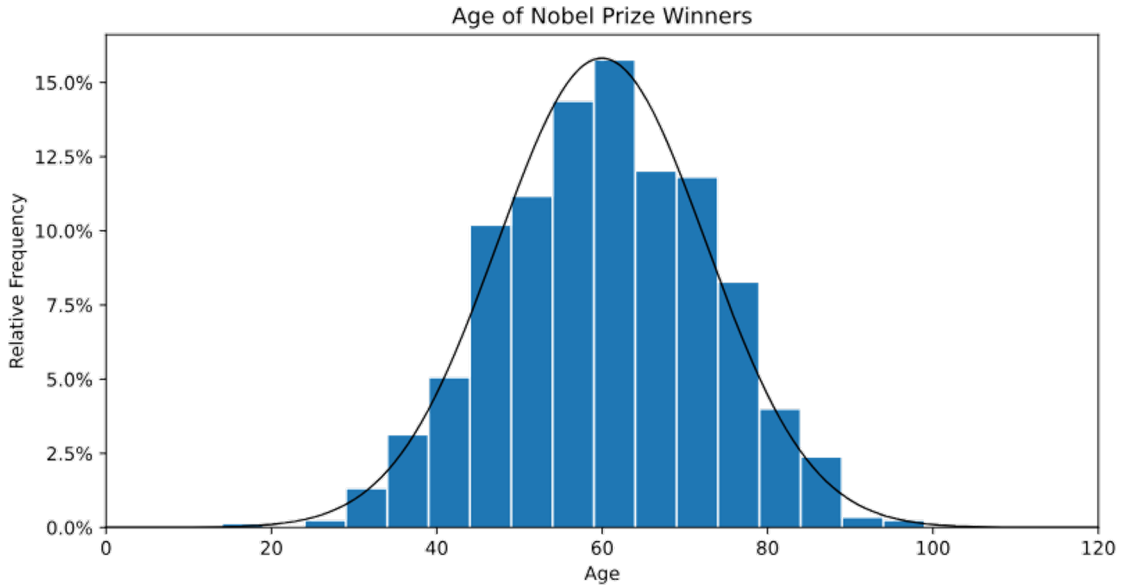
2.3 Common Continuous Distributions

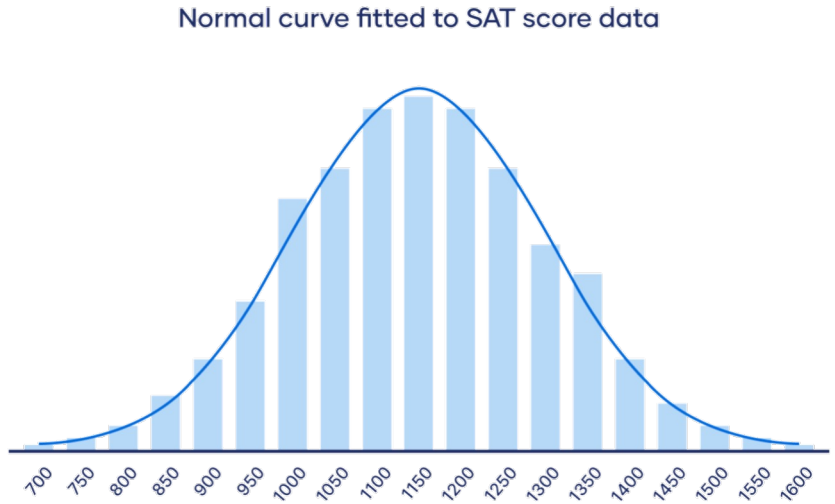
- The normal distribution is also called the **Gaussian distribution (高斯分布)**.

- 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.





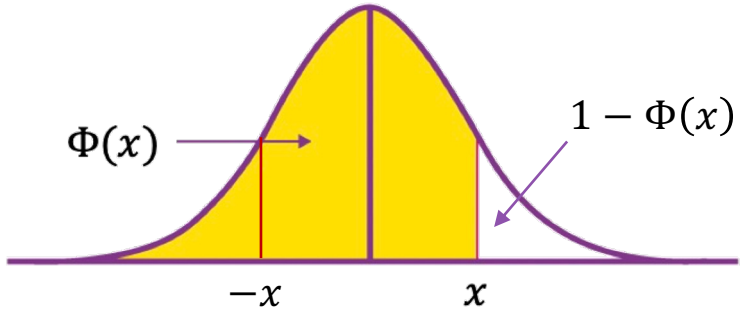
2.3 Common Continuous Distributions

- 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

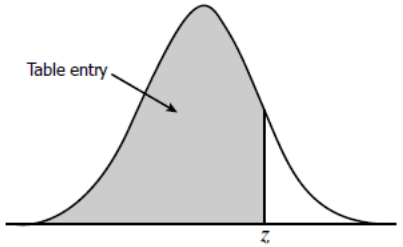


Table entry for z is the area under the standard normal curve to the left of z.

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

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

2.3 Common Continuous Distributions

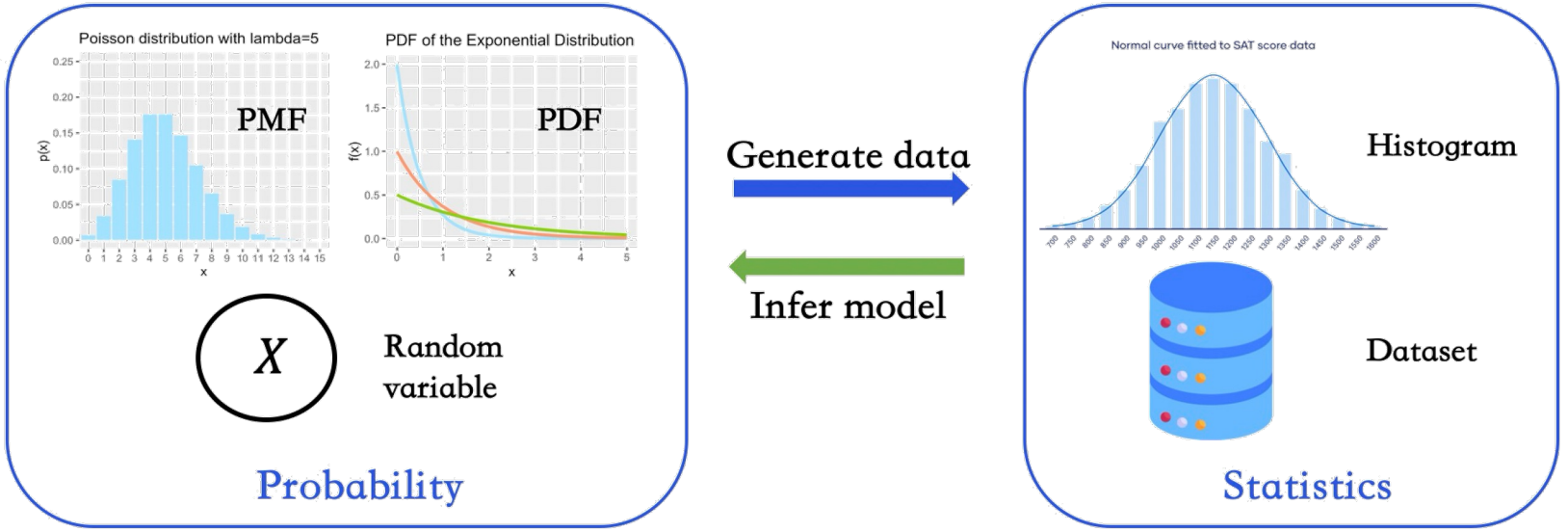
- In summary, we introduced the following continuous distributions:

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 discrete 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.

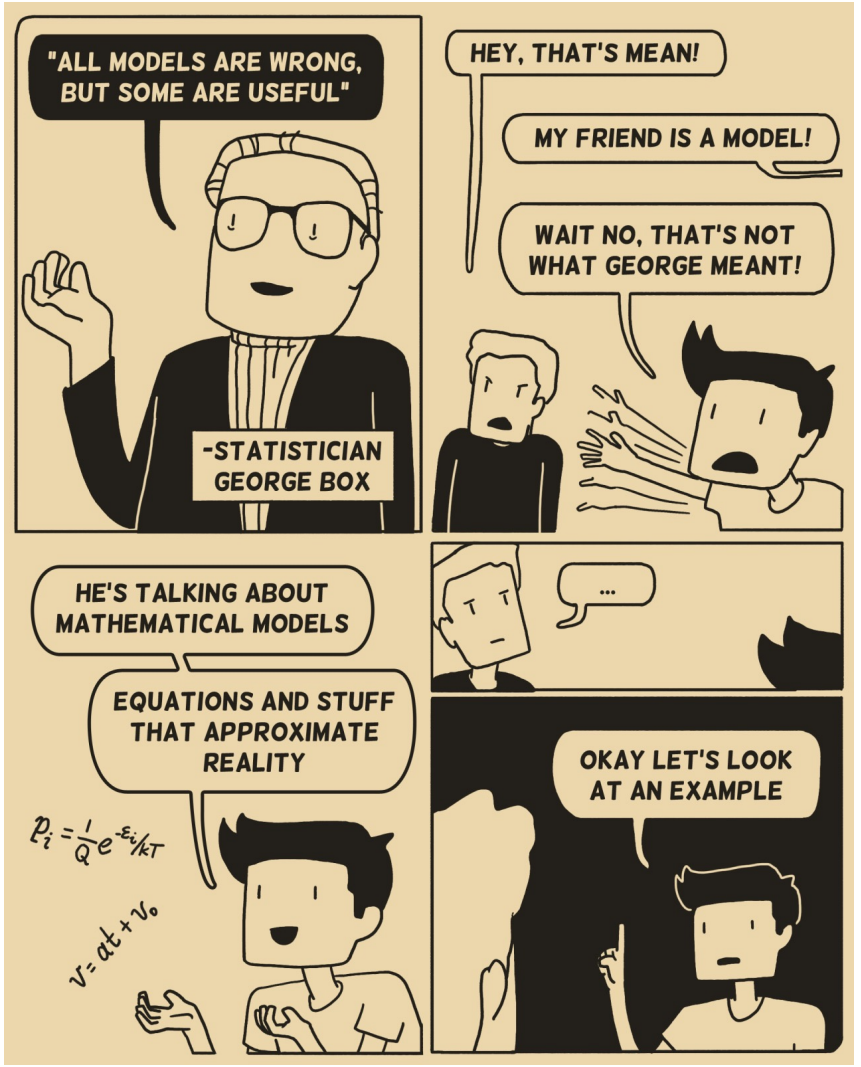
2.3 Common Continuous Distributions

- **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
  - 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

2.4 Transformation of Random Variables

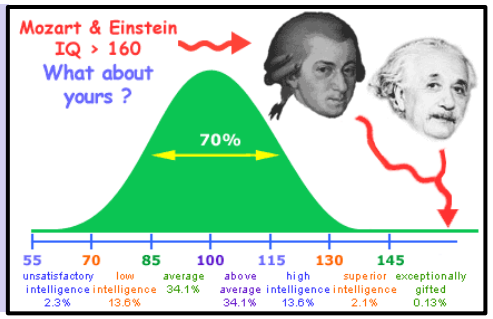
- For a continuous r.v. X with PDF $f_X(x)$, if $Y = g(X)$ is a discrete r.v., then the PMF of Y is:

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

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

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:

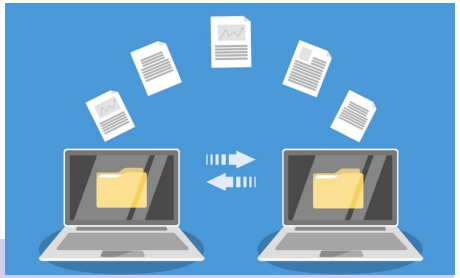
$$F_Y(y) = P\big(X \le g^{-1}(y)\big) = F_X\big(h(y)\big) \Longrightarrow f_Y(y) = F_Y'(y) = h'(y) f_X\big(h(y)\big).$$

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

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 $Y = X^2$?

Solution

2.4 Transformation of Random Variables

Example 3.15 (Continued)

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

Solution

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

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

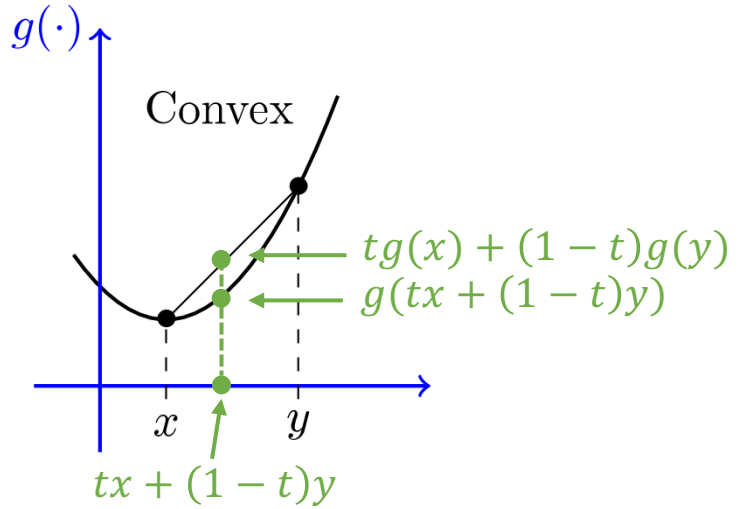
  - If X is a **continuous r.v.** with PDF $f(x)$, given $\int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty$, then

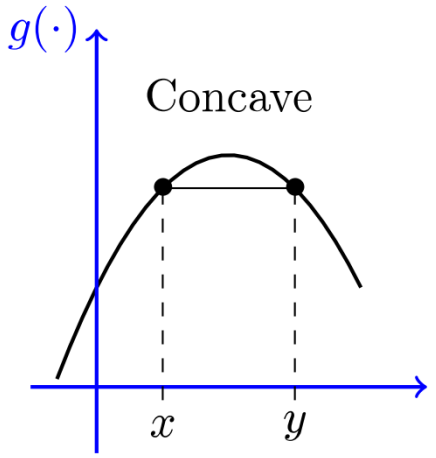
$$\mathrm{E}(Y) = \mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f(x) dx \,.$$

Solution of Example 3.15 (Continued)

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)

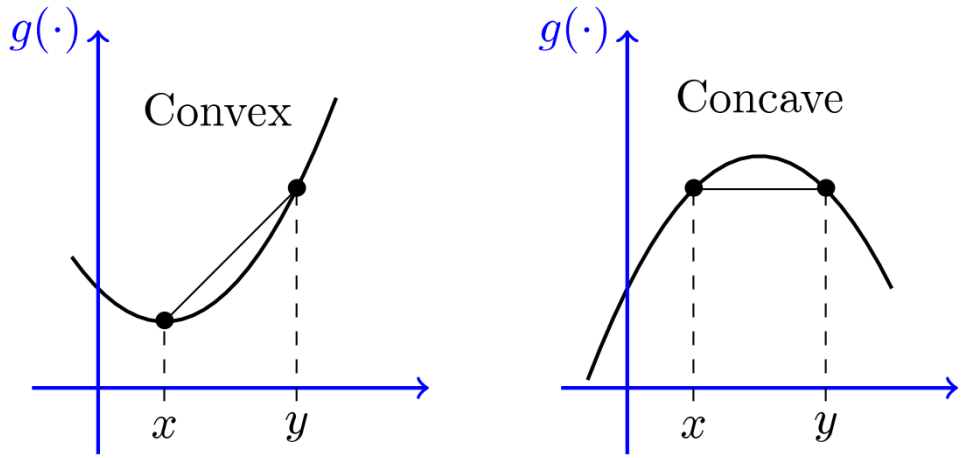
Let X be a random variable, then

- 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

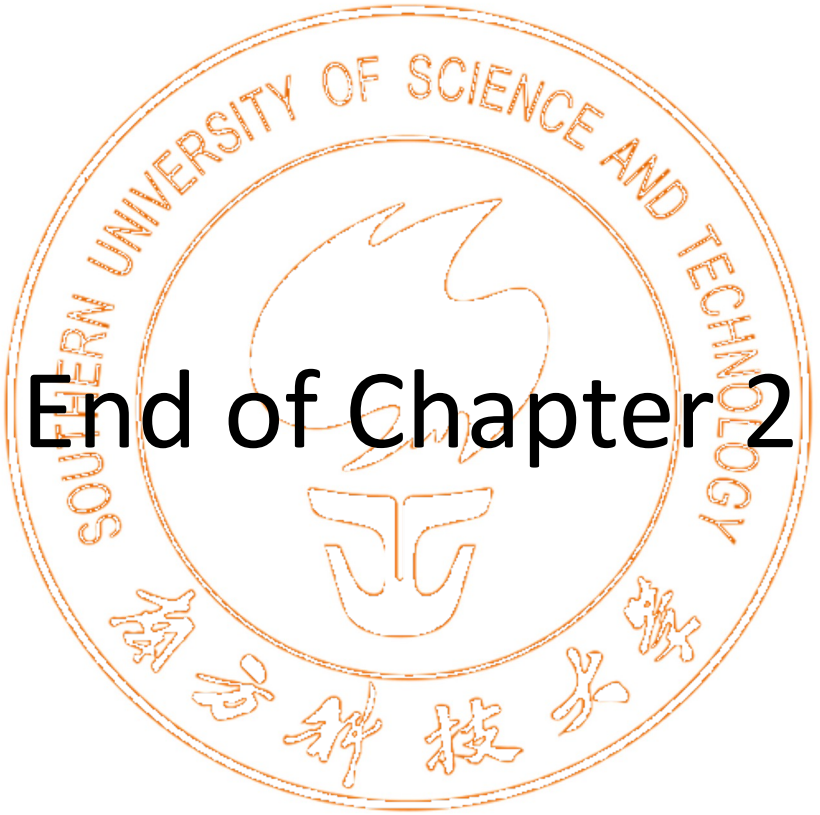
- One of the application of Jensen's inequality is related to the **Kullback-Leibler divergence (KL divergence, KL散度)**.
- KL divergence is called **the information gain (信息增益)** in the context of decision trees and also called the **relative entropy (相对熵)**.
- You may get to know this concept latter 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:

$$D_{KL}(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

- 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



End of Chapter 2