Chapter 5: Law of Large Numbers and Limit Theorems

1. The Law of Large Numbers (大数定律)
2. The Central Limit Theorem (中心极限定理)

Chapter 5: Law of Large Numbers and Limit Theorems

1. The Law of Large Numbers (大数定律)
2. **The Central Limit Theorem (中心极限定理)**

Background of the Central Limit Theorem (中心极限定理)

There are many quantitative indices in real life that follow the normal distribution. Why?

$$Z \sim_{\text{close to}} N(\cdot,\cdot)$$

Researches show that these quantitative indices have the combined effects of a great number of independent random elements, i.e.:

$$Z = X_1 + X_2 + \cdots X_n$$

The research focus of the Central Limit Theorem (中心极限定理):

If $n \to \infty$, under what circumstances does the limiting distribution (极限分布) of $Z_n = \sum_{i=1}^{n} X_i$ is the normal distribution $N(\cdot\,,\ \cdot)$?

⇕

The limiting distribution of $\dfrac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i)}{\sqrt{D\left(\sum_{i=1}^{n} X_i\right)}} = \dfrac{Z_n - E(Z_n)}{\sqrt{D(Z_n)}}$ is $N(0,1)$?

If $\{X_n\}$ is an independent r.v. list, and both mean and variance exist as follows:

$$E(X_n) = \mu_n, D(X_n) = \sigma_n^2 \ (n = 1, 2, \cdots)$$

Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{D\left(\sum_{i=1}^n X_i\right)}} = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \ (n = 1,2, \cdots)$$

Thus $E(Z_n) = 0, D(Z_n) = 1 \ (n = 1, 2, \cdots)$.

Is the limiting distribution of $\{Z_n\}$ certainly $N(0,1)$?

Generally, the answer is NO!

Example: Let $X_i = 0$ for $i = 2, 3, \cdots$, then

$$Z_n = \frac{X_1 - \mu_1}{\sigma_1} \ (n = 1, 2, \cdots)$$

In this case, unless X_1 follows a normal distribution, the conclusion is not true.

If $\{X_n\}$ is an independent r.v. list, and both mean and variance exist as follows:

$$E(X_n) = \mu_n, D(X_n) = \sigma_n^2 \ (n = 1, 2, \cdots)$$

Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{D\left(\sum_{i=1}^n X_i\right)}} = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \ (n = 1,2, \cdots)$$

Thus $E(Z_n) = 0, D(Z_n) = 1 \ (n = 1, 2, \cdots)$.

Definition: If the distribution function $F_n(x)$ of Z_n for any x satisfies

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \le x\right\}$$

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

Then we say that $\{X_n\}$ **follows the Central Limit Theorem (中心极限定理)**

What are the **conditions** that $\{X_n\}$ follows the Central Limit Theorem?

① $X_1, X_2, \cdots, X_n, \cdots$ are independent

② $E(X_i)$ and $D(X_i)$ both exist $(i = 1, 2, \cdots)$

③ ……?

Theorem 1: The Central Limit Theorem under i.i.d. condition (独立同分布中心极限定理)

Assume that $\{X_n\}$is a list of independent and identically distributed r.v.s. The mathematical expectation and variance are

$$E(X_i)=\mu, D(X_i)=\sigma^2>0\ (i=1,2,\cdots)$$

Then $\{X_n\}$ follows the central limit theorem. Which means that the distribution function $F_n(x)$ of the standardized r.v. :

$$Z_n=\frac{\sum_{i=1}^n X_i-\sum_{i=1}^n E(X_i)}{\sqrt{D\left(\sum_{i=1}^n X_i\right)}}=\frac{\sum_{i=1}^n X_i-n\mu}{\sqrt{n}\sigma}\quad (n=1,2,\cdots)$$

satisfies the following condition for any real value x:

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^n X_i-n\mu}{\sqrt{n}\sigma}\le x\right\}$$

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

The practical meaning of the Central Limit Theorem

For a list of independent and identically distributed r.v.s $X_1, X_2, \cdots, X_n, \cdots$ with mean value μ and variance $\sigma^2 > 0$, we have that

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim_{\text{close to}} N(0,1)$$

Or equivalently

$$X_1 + X_2 + \cdots + X_n \sim_{\text{close to}} N(n\mu, n\sigma^2)$$

Since $X_1, X_2, \ldots, X_n$ are independent and identically distributed, every r.v. is only a small part of the whole list. **None of them plays the main role.**

In practical problems, if a quantitative index (数量指标) satisfies

1. The index is the summation of a great number of independent random elements.
2. The influence of the elements are tiny, i.e., none of the elements plays a main role.

then the quantitative index can be considered to follow a **normal distribution.**

Theorem 2: De Moivre-Laplace Central Limit Theorem
(棣莫弗-拉普拉斯中心极限定理)

Assume that $\{\eta_n\}$ is a list of random variables following the binomial distribution with parameter n and p $(0 < p < 1)$. Then for any real value x, we have:

$$\lim_{n\to\infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

Proof: Since binomial distributions are from Bernoulli processes, thus η_n can be decomposed to

$$\eta_n = X_1 + X_2 + \cdots + X_n,$$

where X_i $(i = 1, 2, \cdots)$ are i.i.d. r.v.s following the (0-1) distribution and:

$$E(X_i) = p, D(X_i) = p(1-p)\ (i = 1, 2, \cdots)$$

By the Central Limit Theorem under i.i.d. condition, we have

$$\lim_{n\to\infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\right\} = \lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \le x\right\}$$
$$= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

Application of the De Moivre-Laplace theorem

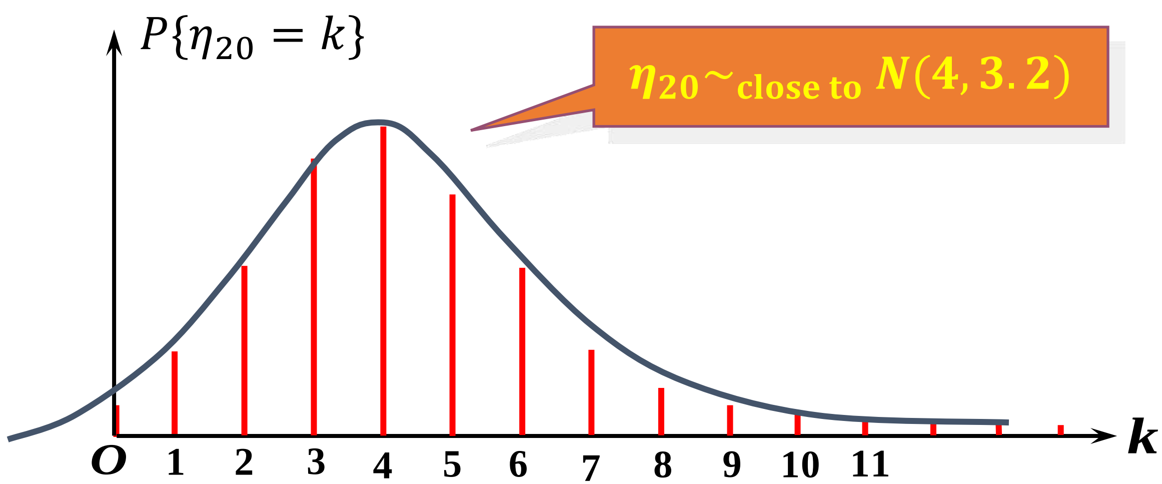
For a list of binomial distribution r.v.s $\eta_n \sim b(n,p)$ $(n=1,2,\cdots)$, we have

$$\frac{\eta_n - np}{\sqrt{np(1-p)}} \sim_{\text{close to}} N(0,1).$$

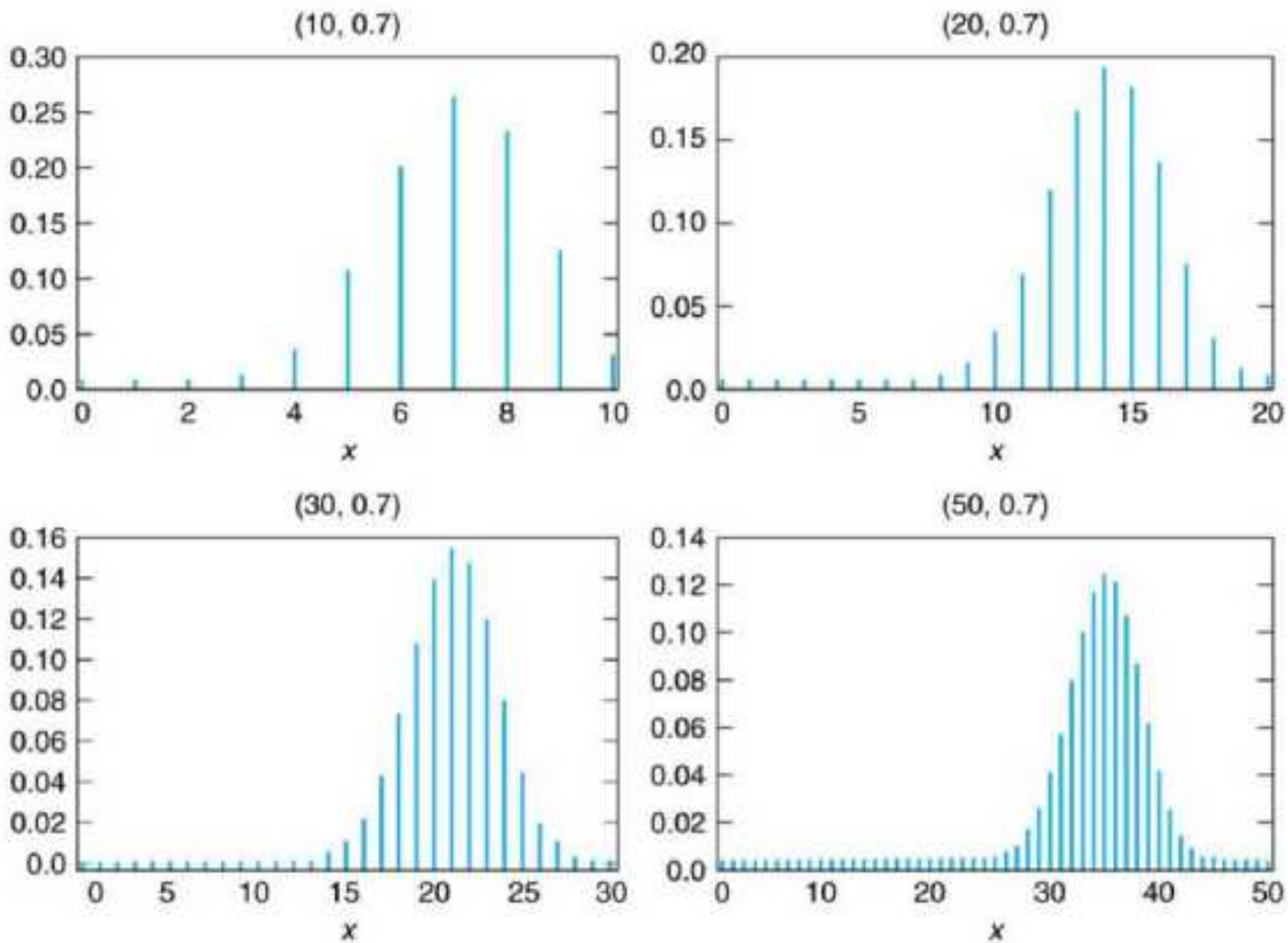
Therefore, when n is large enough, we have

$$\eta_n \sim_{\text{close to}} N\big(np, np(1-p)\big)$$

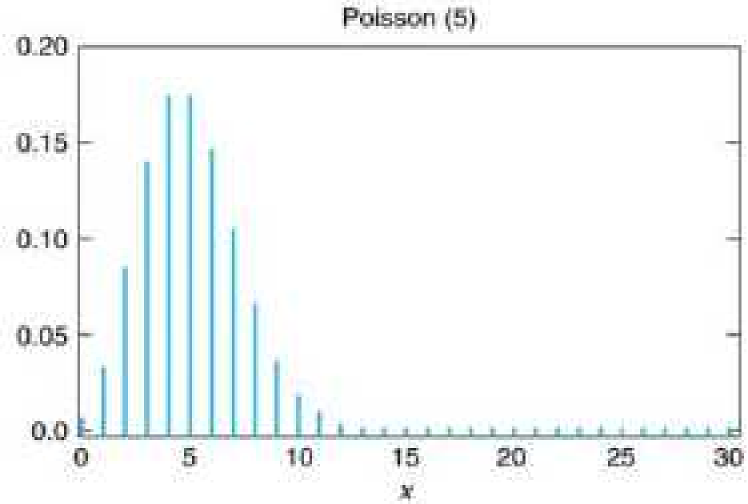
Example: The figure of $\eta_n \sim b(20,\ 0.2)$ is

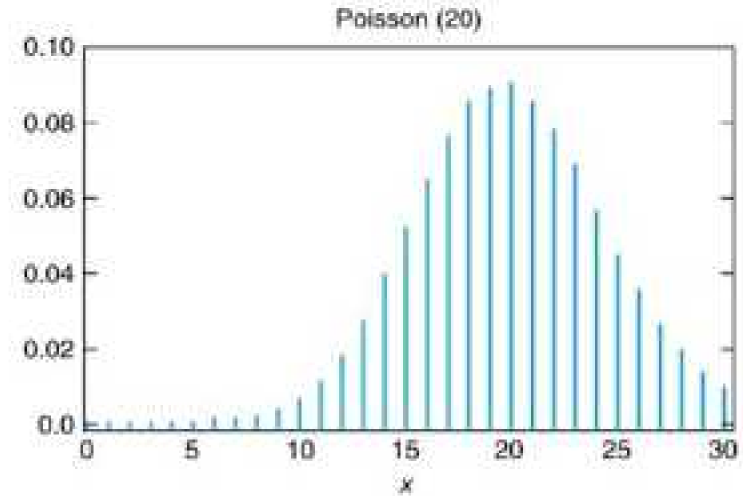


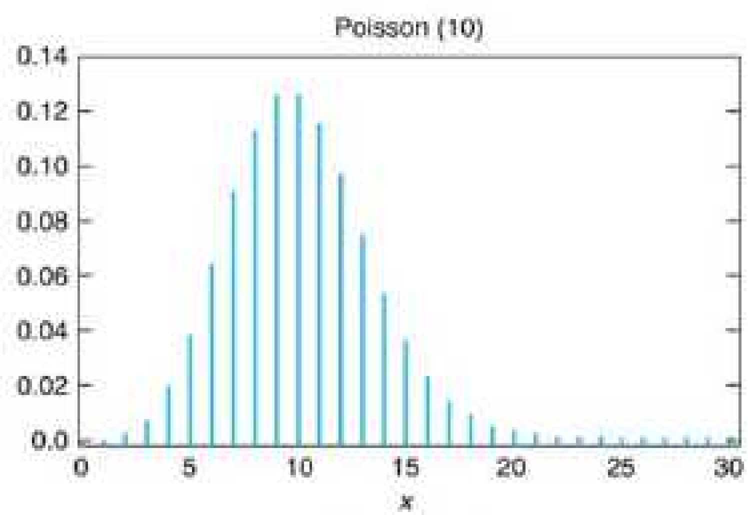
Binomial probability mass functions converging to the normal density



Furthermore: Poisson probability mass functions converging to the normal density



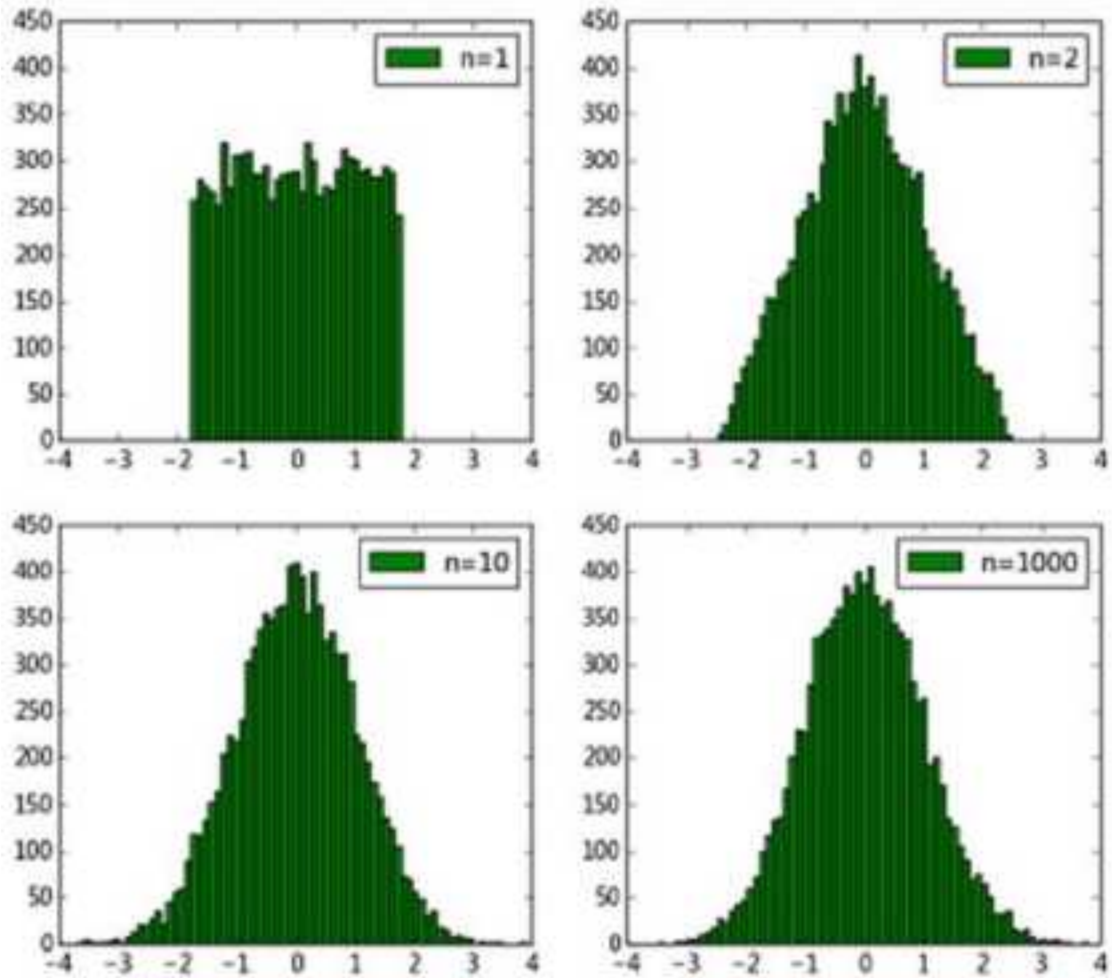




Sums of independent Poisson r.v.s are still Poisson r.v.s.

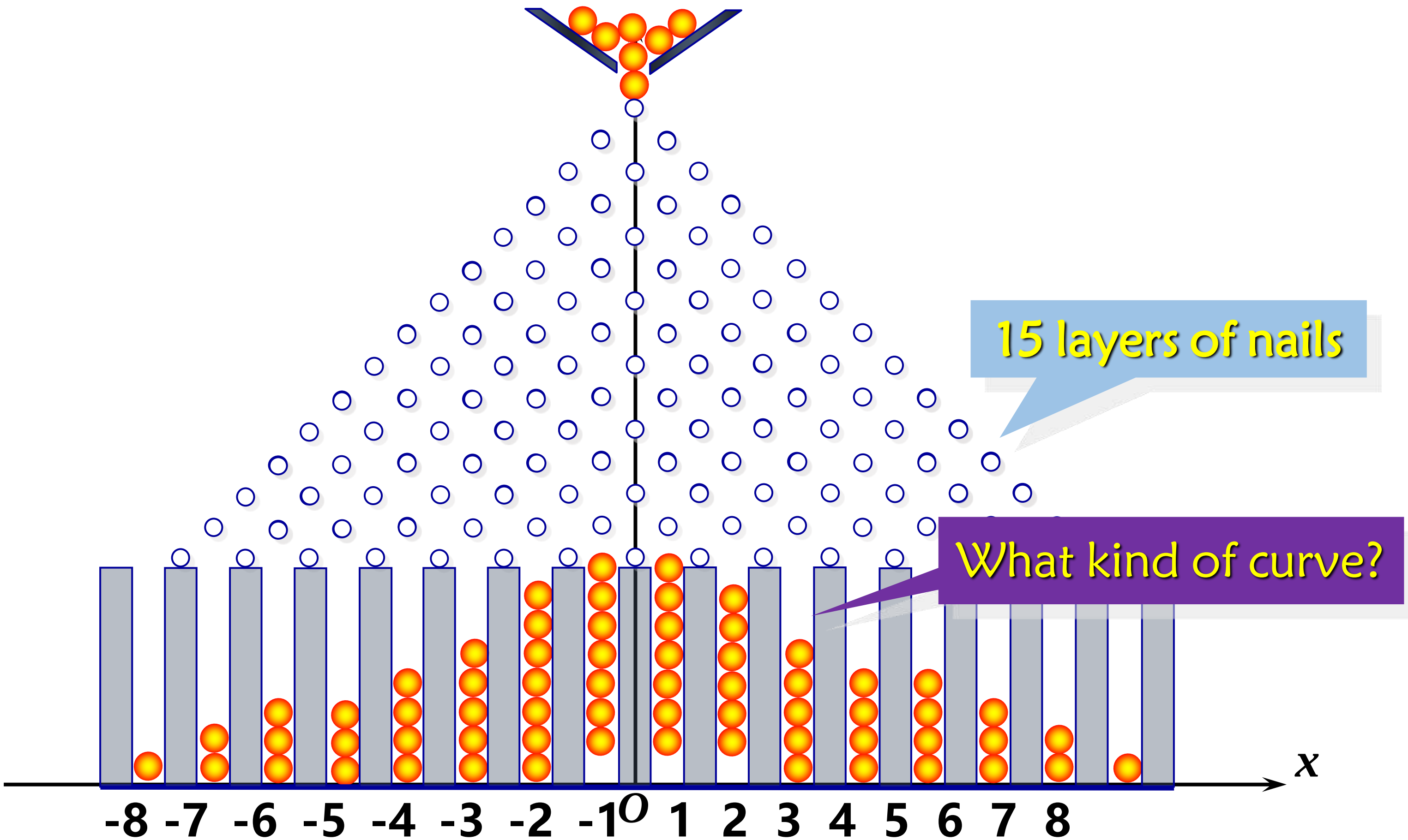
Furthermore: density functions of $\frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{n/12}}$ converging to the normal density

where $X_i \sim_{\text{i.i.d.}} U[0,1]$



When n=1, it is the uniform distribution. As n increases, the histogram (直方图) gradually becomes close to the shape of the standard normal distribution.

Galton's quincunx (梅花机) experiment



Let $X_i = \begin{cases} 1, \text{if the ball goes to the right on the } i\text{th layer} \\ -1, \text{if the ball goes to the left on the } i\text{th layer} \end{cases}, i = 1, 2, \cdots 15$

Then $\sum_{i=1}^{15} X_i \sim_{\text{close to}} N(0, 15).$

Exercise: A company has 500 phone extensions (分机) connected to a switchboard (总机), in which 96% calls are made within the company through internal lines. Assume that each phone extension calling external lines are independent. How many external lines are need to guarantee that people (phones) do not need to wait for external lines 95% of the time?

Answer: At any time, denote the random variable X_k as:

$$X_k = \begin{cases} 1, \text{the } k\text{th extension is calling external lines} \\ 0, \quad \text{otherwise} \end{cases}, k = 1, 2, \cdots, 500$$

Then $X_1, X_2, \cdots, X_{500}$ are i.i.d, and

$$P\{X_k = 1\} = p = 0.04, \quad P\{X_k = 0\} = 0.96$$

By the Central limit theorem under i.i.d condition, we have:

$$\sum_{k=1}^{500} X_k \sim_{\text{close to}} N(20, 19.2)$$

Suppose that M external lines are needed in order to meet the requirement, then:

$$P\left\{\sum_{k=1}^{500} X_k \leq M\right\} \geq 0.95$$

Exercise: A company has 500 phone extensions (分机) connected to a switchboard (总机), in which 96% calls are made within the company through internal lines. Assume that each phone extension calling external lines are independent. How many external lines are need to guarantee that people (phones) do not need to wait for external lines 95% of the time?

Answer: Since:

$$P\left\{\sum_{k=1}^{500} X_k \le M\right\} = P\left\{\frac{\sum_{k=1}^{500} X_k - 20}{\sqrt{19.2}} \le \frac{M-20}{\sqrt{19.2}}\right\} \approx \Phi\left(\frac{M-20}{\sqrt{19.2}}\right) \ge 0.95$$

According to the normal distribution table $\Phi(1.65) = 0.95$, therefore:

$$\frac{M-20}{\sqrt{19.2}} \ge 1.65 \Rightarrow M \ge 27.23$$

So, at least 28 external lines are needed to meet the requirement.

Why it called the Central Limit Theorem (中心极限定理)?

De Moivre-Laplace central limit theorem (棣莫弗-拉普拉斯极限定理) was the first limit theorem (极限定理) in the history of Probability theory proposed by De Moivre in 1730. Afterwards, discussion on the limiting distribution (极限分布) of the sum of independent random variables had been the focus (or center) of probability research for about 200 years. Therefore, it was named as the Central Limit Theorem.

Homework

1. Assume that the lifespan of an electronic component follows the exponential distribution with mean value being 100 hours. There are 16 randomly picked components and suppose that their lifespans are independent. What is the probability that the sum of their lifespans is greater than 1920 hours?

2. There are 10000 elder people participating in a type of insurance. The premium is 200 yuan per year. If the insured person passes away within the insured year, the beneficiary will receive 10000 yuan. Assume that the probability of death is 0.017, what is the probability that the insurance company will suffer a deficit in a year?

Supplementary questions

1. Let $\{X_k\}$ be a list of independent random variables, and its probability distribution satisfies:

$$P\left(X_k = \pm\sqrt{\ln k}\right) = \frac{1}{2}, \qquad k = 1,2,\ldots$$

Prove: $\{X_k\}$ follows the law of large numbers.

2. Let $\{X_n\}$ be a list of independent random variables, and its probability distribution satisfies:

$$P(X_n = 1) = p_n, \qquad P(X_n = 0) = 1 - p_n \quad n = 1,2,\ldots$$

Prove: $\{X_n\}$ follows the law of large numbers.

谢谢大家