

**证明** 根据题意可得  $E[X_i] = 0$ , 以及  $\text{Var}(X_i) = E[X_i^2] = i^{1/2}$ , 根据 Chebysheve 不等式和独立性有

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i \right| \geq \epsilon \right] \leq \frac{1}{n^2 \epsilon^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n^2 \epsilon^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{\epsilon^2} \frac{1}{n^2} \sum_{i=1}^n i^{1/2} \leq \frac{1}{\epsilon^2 \sqrt{n}}$$

再根据

$$\sum_{i=1}^n i^{1/2} \leq \sum_{i=1}^n \int_i^{i+1} i^{1/2} dx \leq \sum_{i=1}^n \int_i^{i+1} x^{1/2} dx = \int_1^{n+1} x^{1/2} dx = 2((n+1)^{3/2} - 1)/3$$

由此可得当  $n \rightarrow +\infty$  时有

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i \right| \geq \epsilon \right] \leq \frac{2((n+1)^{3/2} - 1)/3}{\epsilon^2 n^2} \rightarrow 0$$

大数定律小结:

- Markov 大数定律: 若随机变量序列  $\{X_i\}$  满足  $\text{Var}(\sum_{i=1}^n X_i)/n^2 \rightarrow 0$ , 则满足大数定律;
- Chebyshev 大数定律: 若独立随机变量序列  $\{X_i\}$  满足  $\text{Var}(X_i) \leq c$ , 则满足大数定律;
- Khintchine 大数定律: 若独立同分布随机变量序列  $\{X_i\}$  期望存在, 则满足大数定律;
- Bernoulli 大数定律: 对二项分布  $X_n \sim B(n, p)$ , 有  $X_n/n \xrightarrow{P} p$ .

## 7.2 中心极限定理

对独立的随机变量序列  $X_1, X_2, \dots, X_n, \dots$ , 我们考虑标准化后随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{\text{Var}(\sum_{i=1}^n X_i)}}$$

的极限分布是否为服从正态分布. 首先介绍依分布收敛.

**定义 7.2** 设随机变量  $Y$  的分布函数为  $F_Y(y) = \Pr(Y \leq y)$ , 以及随机变量序列  $Y_1, Y_2, \dots, Y_n, \dots$  的分布函数分别为  $F_{Y_n}(y) = \Pr(Y_n \leq y)$ , 如果

$$\lim_{n \rightarrow \infty} \Pr[Y_n \leq y] = \Pr[Y \leq y], \quad \text{即} \quad \lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y),$$

则称随机变量序列  $Y_1, Y_2, \dots, Y_n, \dots$  依分布收敛于  $Y$ , 记  $Y_n \xrightarrow{d} Y$ .

下面介绍独立同分布中心极限定理, 又被称为林德贝格-勒维 (Lindeberg-Lévy) 中心极限定理”:

**定理 7.6** 设独立同分布的随机变量  $X_1, X_2, \dots, X_n, \dots$  的期望  $E(X_1) = \mu$  和方差  $\text{Var}(X_1) = \sigma^2$ , 则

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

前面介绍标准正态分布的分布函数为  $\Phi(x)$ , 则上述中心极限定理等价于

$$\lim_{n \rightarrow \infty} \Pr[Y_n \leq y] = \Phi(y).$$

随机变量  $Y_n$  是随机变量  $X_1, X_2, \dots, X_n$  的标准化, 其极限服从标准正态分布. 当  $n$  足够大时近似有  $Y_n \sim \mathcal{N}(0, 1)$ , 中心极限定理的变形公式为

$$\sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2), \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

大数定律给出了当  $n \rightarrow \infty$  时随机变量平均值  $\frac{1}{n} \sum_{i=1}^n X_i$  的趋势, 而中心极限定理给出了  $\frac{1}{n} \sum_{i=1}^n X_i$  的具体分布.

**例 7.2** 设一电压接收器同时接收到 20 个独立同分布的信号电压  $V_k$  ( $k \in [20]$ ), 且  $V_k \sim U(0, 10)$ , 求电压和大于 105 的概率.

**解** 根据题意可知独立同分布的随机变量  $V_1, V_2, \dots, V_{20}$  服从均匀分布  $U(0, 10)$ , 于是有  $E(V_k) = 5$  和  $\text{Var}(V_k) = 100/12 = 25/3$ . 设  $V = \sum_{k=1}^{20} V_k$ , 则有

$$E(V) = 100 \quad \text{Var}(V) = 500/3.$$

根据中心极限定理近似有

$$\frac{V - E(V)}{\sqrt{\text{Var}(V)}} = \frac{V - 100}{\sqrt{500/3}} \sim \mathcal{N}(0, 1).$$

根据标准正态分布的分布函数  $\Phi(x)$  有

$$\Pr(V \geq 105) = \Pr\left(\frac{V - 100}{\sqrt{500/3}} \geq \frac{105 - 100}{\sqrt{500/3}}\right) = \Pr\left(\frac{V - 100}{\sqrt{500/3}} \geq 0.387\right) = 1 - \Phi(0.387).$$

查表完成证明.

**例 7.3** 某产品装箱, 每箱重量是随机的, 假设其期望是 50 公斤, 标准差为 5 公斤. 若最大载重量为 5 吨, 问每车最多可装多少箱能以 0.997 以上的概率保证不超载?

**解** 假设最多可装  $n$  箱不超重, 用  $X_i$  表示第  $i$  箱重量 ( $i \in [n]$ ), 有  $E(X_i) = 50$  和  $\text{Var}(X_i) = 25$ . 设总重量  $X = \sum_{i=1}^n X_i$ , 则有  $E(X) = 50n$  和  $\text{Var}(X) = 25n$ . 由中心极限定理近似有

$$(X - 50n)/\sqrt{25n} \sim \mathcal{N}(0, 1).$$

根据标准正态分布的分布函数  $\Phi(x)$  有

$$\Pr(X \leq 5000) = \Pr\left(\frac{X - 50n}{\sqrt{25n}} \leq \frac{5000 - 50n}{\sqrt{25n}}\right) = \Phi\left(\frac{5000 - 50n}{\sqrt{25n}}\right) > 0.977 = \Phi(2).$$

根据分布函数的单调性有

$$\frac{1000 - 10n}{\sqrt{n}} > 2 \implies 1000n^2 - 2000n + 1000^2 > 4n.$$

求解可得  $n > 102.02$  或  $n < 98.02$ , 根据由题意可知  $n = 98$ .

下面介绍另一个中心极限定理: 棣莫弗-拉普拉斯 (De Moivre-Laplace) 中心极限定理:

**推论 7.1** 设随机变量  $X_n \sim B(n, p)$ , 则

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

由此中心极限定理可知: 当  $n$  非常大时随机变量  $X_n \sim B(n, p)$  满足  $X_n \overset{\text{近似}}{\sim} \mathcal{N}(np, np(1-p))$ , 从而有如下近似估计:

$$\Pr[X_n \leq y] = \Pr\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leq \frac{y - np}{\sqrt{np(1-p)}}\right] \approx \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right).$$

针对上式, 可以考虑三种问题: i) 已知  $n$  和  $\Pr[X_n \leq y]$ , 求  $y$ ; ii) 已知  $n$  和  $y$ , 求  $\Pr[X_n \leq y]$ ; iii) 已知  $y$  和  $\Pr[X_n \leq y]$ , 求  $n$ . 下面看三个例子:

**例 7.4** 车间有 200 台独立工作的车床, 每台工作的概率为 0.6, 工作时每台耗电 1 千瓦, 至少供电多少千瓦才能以 99.9% 的概率保证正常生产.

**解** 设工作的车床数为  $X$ , 则  $X \sim B(200, 0.6)$ . 设至少供电  $y$  千瓦. 根据棣莫弗-拉普拉斯中心定理近似有  $X \sim \mathcal{N}(120, 48)$ , 进一步有

$$\Pr(X \leq y) \geq 0.999 \implies \Pr\left(\frac{X - 120}{\sqrt{48}} \leq \frac{y - 120}{\sqrt{48}}\right) \approx \Phi\left(\frac{y - 120}{\sqrt{48}}\right) \geq 0.999 = \Phi(3.1).$$

所以有  $\frac{y-120}{\sqrt{48}} \geq 3.1$ , 求解可得  $y \geq 141$ .

**例 7.5** 系统由 100 个相互独立的部件组成, 每部件损坏率为 0.1, 至少 85 个部件正常工作系统才能运行, 求系统运行的概率.

**解** 设  $X$  是损坏的部件数, 则  $X \sim B(100, 0.1)$ , 有  $E(X) = 10$  和  $\text{Var}(X) = 9$ . 根据棣莫弗-拉普拉斯中心定理近似有  $X \sim \mathcal{N}(10, 9)$ , 求系统运行的概率为

$$\Pr(X \leq 15) = \Pr\left(\frac{X - 10}{\sqrt{9}} \leq \frac{15 - 10}{\sqrt{9}}\right) \approx \Phi(5/3).$$

**例 7.6** 一次电视节目调查中调查  $n$  人, 其中  $k$  人观看了电视节目, 因此收看比例  $k/n$  作为电视节目收视率  $p$  的估计, 要以 90% 的概率有  $|k/n - p| \leq 0.05$  成立, 需要调查多少对象?

**解** 用  $X_n$  表示  $n$  个调查对象中收看节目的人数, 则有  $X_n \sim B(n, p)$ . 根据棣莫弗-拉普拉斯中心定理近似有  $(X_n - np)/\sqrt{np(1-p)} \sim \mathcal{N}(0, 1)$ , 进一步有

$$\begin{aligned} \Pr\left[\left|\frac{X_n}{n} - p\right| \leq 0.05\right] &= \Pr\left[\frac{|X_n - np|}{n} \leq 0.05\right] = \Pr\left[\frac{|X_n - np|}{\sqrt{np(1-p)}} \leq \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right] \\ &= \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \end{aligned}$$

对于标准正太分布函数有  $\Phi(-\alpha) = 1 - \Phi(\alpha)$  以及  $p(1-p) \leq 1/4$ , 于是有

$$\Pr\left[\left|\frac{X_n}{n} - p\right| \leq 0.05\right] = 2\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 > 2\Phi(\sqrt{n}/10) - 1 > 0.9.$$

所以  $\Phi(\sqrt{n}/10) \geq 0.95$ , 查表解得  $n \geq 271$ .

对独立不同分布的随机变量序列, 有李雅普诺夫 (Lyapunov) 中心极限定理:

**定理 7.7** 设独立随机变量  $X_1, X_2, \dots, X_n, \dots$  的期望  $E[X_n] = \mu_n$  和方差  $\text{Var}(X_n) = \sigma_n^2 > 0$ . 记  $B_n^2 = \sum_{k=1}^n \sigma_k^2$ , 若存在  $\delta > 0$ , 当  $n \rightarrow \infty$  时有

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E[|X_k - \mu_k|^{2+\delta}] \rightarrow 0$$

成立, 则有

$$Y_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E(X_k)}{\sqrt{\text{Var}(\sum_{k=1}^n X_k)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

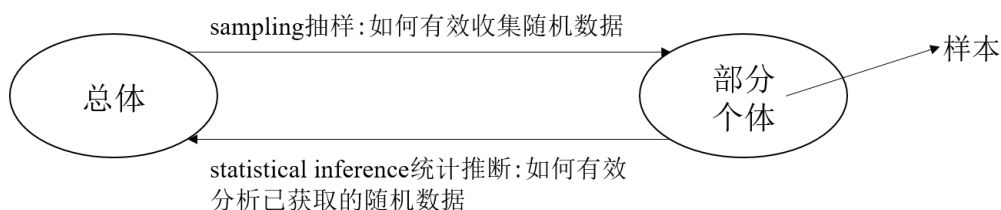
中心极限定理小结:

- 独立同分布中心极限定理: 若  $E[X_k] = \mu$  和  $\text{Var}(X_k) = \sigma^2$ , 则  $\sum_{k=1}^n X_k \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2)$ ;
- 棣莫弗-拉普拉斯中心极限定理: 若  $X_k \sim B(k, p)$ , 则  $X_k \xrightarrow{d} \mathcal{N}(np, np(1-p))$ ;
- 独立不同分布中心极限定理: 李雅普诺夫定理.

## 第8章 统计的基本概念

到 19 世纪末 20 世纪初, 随着近代数学和概率论的发展, 诞生了统计学.

统计学: 以概率论为基础, 研究如何有效收集研究对象的随机数据, 以及如何运用所获得的数据揭示统计规律的一门学科. 统计学的研究内容具体包括: 抽样、参数估计、假设检验等.



### 8.1 总体 (population) 与样本 (sample)

‘总体’是研究问题所涉及的对象全体; 总体中每个元素称为‘个体’. 总体分为有限或无限总体. 例如: 全国人民的收入是总体, 一个人的收入是个体.

在研究总体时, 通常关心总体的某项或某些数量指标, 总体中的每个个体是随机试验的一个观察值, 即随机变量  $X$  的值. 对总体的研究可转化为对随机变量  $X$  的分布或数字特征的研究, 后面总体与随机变量  $X$  的分布不再区分, 简称总体  $X$ .

总体: 研究对象的全体  $\Rightarrow$  数据  $\Rightarrow$  随机变量 (分布未知).

样本: 从总体中随机抽取一些个体, 一般表示为  $X_1, X_2, \dots, X_n$ , 称  $X_1, X_2, \dots, X_n$  为取自总体  $X$  的随机样本, 其样本容量为  $n$ .

抽样: 抽取样本的过程.

样本值: 观察样本得到的数值, 例如:  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  为样本观察值或样本值.

样本的二重性: i) 就一次具体观察而言, 样本值是确定的数; ii) 不同的抽样下, 样本值会发生变化, 可看作随机变量.

**定义 8.1 (简单随机样本)** 称样本  $X_1, X_2, \dots, X_n$  是总体  $X$  的简单随机样本, 简称样本, 是指样本满足: 1) 代表性, 即  $X_i$  与  $X$  同分布; 2) 独立性, 即  $X_1, X_2, \dots, X_n$  之间相互独立.

本书后面所考虑的样本均为简单随机样本.

设总体  $X$  的联合分布函数为  $F(x)$ , 则  $X_1, X_2, \dots, X_n$  的联合分布函数为

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i);$$

若总体  $X$  的概率密度为  $f(x)$ , 则样本  $X_1, X_2, \dots, X_n$  的联合概率密度为

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i).$$

若总体  $X$  的分布列  $\Pr(X = x_i)$ , 则样本  $X_1, X_2, \dots, X_n$  的联合分布列为

$$\Pr(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n \Pr(X_i = x_i).$$

## 8.2 常用统计量

为研究样本的特性, 我们引入统计量:

**定义 8.2** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本,  $g(X_1, X_2, \dots, X_n)$  是关于  $X_1, X_2, \dots, X_n$  的一个连续、且不含任意参数的函数, 称  $g(X_1, X_2, \dots, X_n)$  是一个 **统计量**.

由于  $X_1, X_2, \dots, X_n$  是随机变量, 因此统计量  $g(X_1, X_2, \dots, X_n)$  是一个随机变量. 而  $g(x_1, x_2, \dots, x_n)$  为  $g(X_1, X_2, \dots, X_n)$  的一次观察值. 下面研究一些常用统计量.

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **样本均值** 为

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

根据样本的独立同分布性质有

**引理 8.1** 设总体  $X$  的期望为  $E[X] = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 则有

$$E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n, \quad \bar{X} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **样本方差** 为

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

**引理 8.2** 设总体  $X$  的期望为  $E[X] = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 则有

$$E[S_0^2] = \frac{n-1}{n} \sigma^2.$$

**证明** 根据  $E[X_i^2] = \sigma^2 + \mu^2$  有

$$E(\bar{X}^2) = E \left[ \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 \right] = \frac{1}{n^2} E \left[ \left( \sum_{i=1}^n X_i \right)^2 \right] = \frac{1}{n^2} E \left[ \sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j \right] = \frac{\sigma^2}{n} + \mu^2,$$