关于大数定律与中心极限定理

第一节 大数定律

背景

1. 为何能以某事件发生的频率作为该事件的概率的估计？
2. 为何能以样本均值作为总体期望的估计？
3. 为何正态分布在概率论中占有极其重要的地位？
4. 大样本统计推断的理论基础是什么？

大数定律

中心极限定理

1. 切比雪夫不等式

设随机变量X的数学期望$E(X)=\mu$, 方差$D(X)=\sigma^2$, 则对任意的正数ε，不等式

$$P\{|X-\mu|\geq\varepsilon\}\leq\frac{D(X)}{\varepsilon^2}$$

或

$$P\{|X-\mu|<\varepsilon\}>1-\frac{D(X)}{\varepsilon^2}$$

成立.

利用切比雪夫不等式可以估计一些随机事件的概率。

例1 设电站供电网有10000盏灯，夜晚每一盏灯开灯的概率是0.7，假定开、关时间彼此独立，估计夜晚同时开着的灯数在6800与7200之间的概率

解 设X表示在夜晚同时开着的灯的数目，它服从参数为n=10000, p=0.7的二项分布，则有

$$P(6800 < X < 7200) = \sum_{k=6801}^{7199} C_{10000}^{k} 0.7^{k} 0.3^{10000-k}$$

而用切比雪夫不等式估计

E(X)=np=7000, D(x)=np(1-p)=2100

P(6800<X<7200)=P(|X-7000|<200)>0.95

使用切比雪夫不等式只能得到事件的大致概率，能否得到其较精确的概率呢？这就要用到中心极限定理

2.大数定律

定义1　设Y_1，Y_2，…，Y_n，…，是一随机变量序列，a为一常数．若对任意给定正数ε>0, 有

$$\lim_{n\to+\infty} P\{|Y_n - a| < \varepsilon\} = 1$$

则称随机变量序列$Y_1,Y_2,\ldots,Y_n,\ldots$, 依概率收敛于$a$.

定义2　设X_1，X_2，…，X_n，　… 是一随机变量序列

. 若存在常数列$\{a_n\}$使对任意给定的正数ε，恒有

$$Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

，则称随机变量序列$\{Y_n\}$

$$\lim_{n\to+\infty} P\{|Y_n - a_n| < \varepsilon\} = 1$$

服从大数定律.

注意：

$\{X_n\}$依概率收敛于a，意味着对任意给定的$\varepsilon > 0$，当n充分大时，事件$|X_n - a| < \varepsilon$的概率很大，接近于1；并不排除事件$|X_n - a| \geq \varepsilon$的发生，而只是说它发生的可能性很小.

切比雪夫大数定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$ $D(X_k)=\sigma^2$ (k=1，2，⋯)，则对任意的正数$\varepsilon>0$，有

$$\lim_{n \to +\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| < \varepsilon\} = 1$$

或

$$\lim_{n \to +\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| \geq \varepsilon\} = 0$$

注意

1、定理中$\{|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu|<\varepsilon\}$是指一个随机事件，当$n\to\infty$时，这个事件的概率趋于1.

2、通俗地说，对于独立同分布且具有均值μ的随机变量$X_1,\cdots,X_n$，当n很大时，它们的算术平均$\frac{1}{n}\sum_{i=1}^{n}X_i$很可能接近于$\mu$。

证明:（利用切比雪夫不等式）

根据已知条件

$$E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n}E\left(\sum_{k=1}^{n}X_k\right)=\frac{1}{n}\sum_{k=1}^{n}E(X_k)=\mu$$

由切比雪夫不等式，有

$$D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}D\left(\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{\sigma^2}{n}$$

又

$$1-\frac{\sigma^2}{n\varepsilon^2}\le P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon)\le 1$$

所以

$$\lim_{n\to+\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}=1$$

伯努利大数定理

设n_A为是n次独立重复试验中事件A发生的次数，p 是事件A在每次试验中发生的概率, 则对任意的正数$\varepsilon>0$, 有

$$\lim_{n\to+\infty} P\{|\frac{n_A}{n}-p|<\varepsilon\}=1$$

或

$$\lim_{n\to+\infty} P\{|\frac{n_A}{n}-p|\geq\varepsilon\}=0$$

证：设

$$X_k=\begin{cases}0, & A\text{在第}k\text{次试验中发生,}\\ 1, & A\text{在第}k\text{次试验中未发生}\end{cases}$$

那么 $X_1,X_2,\cdots,X_n$ 相互独立，且服从参数为p的0—1分布，$E(X_k)=p$，$D(X_k)=p(1-p)$.

由切比雪夫大数定理, 有

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

所以

$$\lim_{n\to+\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-p|<\varepsilon\}=1$$

即

$$\lim_{n\to+\infty}P\{|\frac{n_A}{n}-p|<\varepsilon\}=1$$

辛钦大数定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$（k=1，2，…），则对任意的正数$\varepsilon>0$, 有

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k-\mu|<\varepsilon\}=1$$

或

$$\lim_{n\to+\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k-\mu|\geq\varepsilon\}=0$$

第二节中心极限定理

设 $\{X_n\}$ 为独立随机变量序列，记其和为

$$Y_n = \sum_{i=1}^{n} X_i$$

问这个和的**极限分布**是什么？

1. 独立同分布中心极限定理

若X_1，X_2，…，X_n，…，为独立同分布随机变量序列，$E(X_k)=\mu$ $D(X_k)=\sigma^2$ (k=1，2, ⋯)，则随机变量标准化量

$$Y_n=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}$$

的分布函数$F_n(x)$对于任意x满足

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

$$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \underset{\text{近似}}{\sim} N(0,1) \quad \Rightarrow \sum_{k=1}^{n} X_k \underset{\text{近似}}{\sim} N(n\mu, n\sigma^2)$$

$$\Downarrow$$

$$\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\sigma/\sqrt{n}} \underset{\text{近似}}{\sim} N(0,1) \Rightarrow \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \underset{\text{近似}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

$$\lim_{n\to+\infty} F_n(x) = \lim_{n\to+\infty} P\{\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

例2 每袋味精的净重为随机变量，平均重量为 100 克，标准差为10克. 一箱内装200袋味精，求一箱味精的净重大于20500克的概率?

解：设箱中第 i 袋味精的净重为 X_i, 则X_i 独立同分布，且 $E(X_i)=100$，$Var(X_i)=100$，

由中心极限定理得，所求概率为：

$$P\left(\sum_{i=1}^{200} X_i > 20500\right) \approx 1-\Phi\left(\frac{20500-200\times 100}{\sqrt{200\times 100}}\right)$$

$$=1-\Phi(3.54)$$

故一箱味精的净重大于20500克的概率为0.0002.

2. 李雅普诺夫中心极限定理

若X_1，X_2，…，X_n，…，为独立随机变量序列,

$$E(X_k)=\mu_k,\ D(X_k)=\sigma_k^2>0,\ B_n^2=\sum_{k=1}^{n}X_k^2$$

，若存在正数，使当 $n\to\infty$ 时

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

则随机变量标准化量Z_n的分布函数$F_n(x)$对于任意x满足

$$\lim_{n\to+\infty}F_n(x)=P\{Z_n\le x\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

$$Z_n=\frac{\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu_k}{\sqrt{\sum_{k=1}^{n}\sigma_k^2}}$$