2.3 大数定律与中心极限定理

本节要解决的问题

答复

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？

大数定律

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？

中心极限定理

§5.1 大数定律

一、契比雪夫(chebyshev)不等式

定理1 设 r.v. X 的数学期望为 $E(X)=\mu$，方差为 $D(X)=\sigma^2$，则对于任意实数 $\varepsilon > 0$,有

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

证
$$P(|X-\mu| \geq \varepsilon) = \int_{|X-\mu|\geq\varepsilon} f(x)dx$$

$$\leq \int_{|X-\mu|\geq\varepsilon} \frac{|x-\mu|^2}{\varepsilon^2} f(x)dx$$

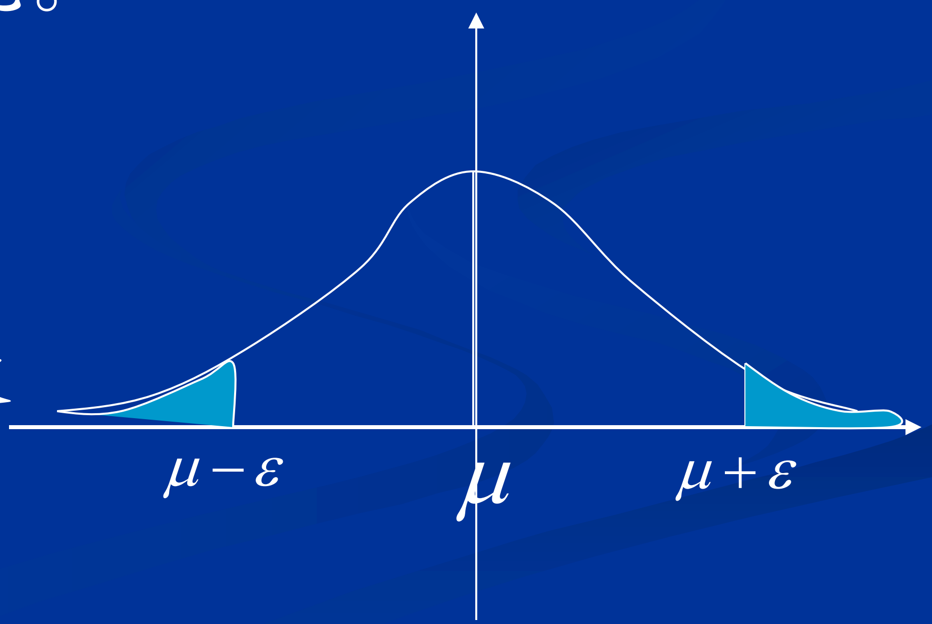
$$\leq \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \frac{D(X)}{\varepsilon^2}$$

或 $\boxed{P(|X-E(X)|<\varepsilon)\geq 1-\dfrac{D(X)}{\varepsilon^2}}$

当 $\varepsilon^2 \leq D(X)$ 无实际意义。

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

注1：此不等式只用期望和方差就描述了r.v的变化情况。



注2：从不等式看出，方差是描述r.v与其期望离散程度的量。

例1 已知随机变量X的期望$E(X)=14$, 方差$D(X)=35/3$, 试估计$P\{10<X<18\}$的大小.

解

$$P\{10<X<18\}=P\{10-E(X)<X-E(X)<18-E(X)\}$$

$$=P\{-4<X-14<4\}=P\{|X-14|<4\}$$

由契比雪夫不等式, 得

$$P\{|X-14|<4\}\geq 1-35/3/4^2\approx 0.271,$$

所以 $P\{10<X<18\}\geq 0.271.$

二、大数定律

定义1 设 $X_1, X_2, \cdots, X_n, \cdots$ 是一系列 r.v. 若 $\forall \varepsilon > 0$ 存在一常数a，满足 $\lim\limits_{n\to\infty} P\left(\left|X_n - a\right| \geq \varepsilon\right) = 0$

(或 $\lim\limits_{n\to\infty} P\left(\left|X_n - a\right| < \varepsilon\right) = 1$)

则称 r.v. 序列 $X_1, X_2, \cdots, X_n, \cdots$ 依概率收敛于常数 a，记作

$$X_n \xrightarrow[n\to\infty]{P} a$$

定理1.(Chebyshev大数定律)

设 $r.v.$ 序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，

(指任意给定 $n > 1, X_1, X_2, \cdots, X_n$ 相互独立)

$$E(X_k) = \mu_k, D(X_k) = \sigma_k^2 \le \sigma^2, \quad k = 1,2,\cdots$$

则 $\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left| \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \right| \ge \varepsilon \right) = 0$$

或 $$\lim_{n\to\infty} P\left(\left| \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \right| < \varepsilon \right) = 1$$

推论1(辛钦大数定律)设随机变量$X_1, X_2,\ldots, X_n, \ldots$

相互独立且服从同一分布, $E(X_i)=\mu$, $D(X_i)=\sigma^2<\infty$

则

$$\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} \mu$$

即n个相互独立同分布的随机变量的算术平均值依概率收敛于随机变量的期望值。

需要指出的是辛钦大数定律可以去掉$D(X_i)=\sigma^2<\infty$

这个条件。此定理在应用中很重要，使算术平均值法则有了理论依据。

推论2(贝努里Bernoulli 大数定律)

设 μ_n 是 n 次独立重复贝努里试验中事件 A 发生的次数, p 是每次试验中 A 发生的概率, 则

$\forall \varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{\mu_n}{n} - p\right| \geq \varepsilon\right) = 0$$

或

$$\lim_{n\to\infty} P\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

即：

$$\frac{\mu_n}{n} \xrightarrow{P} p$$

贝努里(Bernoulli)大数定律的意义

在概率的统计定义中, 事件 A 发生的频率 μ_n/n "稳定于"事件 A 在一次试验中发生的概率是指：频率 μ_n/n 与 p 有较大偏差 $\left(\left|\frac{\mu_n}{n}-p\right|\geq\varepsilon\right)$ 是小概率事件,因而在 n 足够大时, 可以用频率近似代替 p .

在 Bernoulli 定理的证明过程中，Y_n 是相互独立的服从 0 -1分布$B(1,p)$的 r.v. 序列 $\{X_k\}$ 的算术平均值, Y_n 依概率收敛于其数学期望 p .

结果同样适用于服从其它分布的独立 r.v. 序列。

易得小概率原理：即当A的概率很小时，A发生的频率也很小，即A很少发生。

2.3中心极限定理

在第2章, 我们提到在自然现象和社会现象中, 大量的随机变量都是服从或近似服从正态分布的.中心极限定理就是以此为背景的关于“在一定的条件下大量的相互独立的随机变量和的极限分布是正态分布”的一系列定理.

定理一

林德伯格-列维中心极限定理 (Lindberg-levi)

[独立同分布的中心极限定理]

定理二

棣莫弗-拉普拉斯中心极限定理
(De Moivre-Laplace)

[二项分布以正态分布为极限分布]

定理三

林德伯格-列维中心极限定理
(Lindberg-levi)

[独立不同分布的中心极限定理]

定义 当随机变量X的均值、方差都存在时, 则

$$Y=\frac{X-EX}{\sqrt{DX}}$$

称为随机变量X的标准化随机变量.

根据第二章知识若 $X\sim N(\mu,\sigma^2)$ 则X的标准化随机变量

$$Y=(X-EX)/\sqrt{DX}=(X-\mu)/\sigma\sim N(0,1)$$

若$X_1, X_2, \ldots, X_n$为独立同分布的随机变量，

$X_i \sim N(\mu,\sigma^2)$ ，则 $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$

其标准化随机变量

$$\frac{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i}{\sqrt{D(\sum_{i=1}^{n} X_i)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

更一般地，有下面的关于标准化随机变量的近似分布的中心极限定理.

定理 1 独立同分布的中心极限定理

设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$

独立同分布, 且有期望和方差：

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

则对于任意实数 x,

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

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

则 Y_n 为 $\sum_{k=1}^{n}X_k$ 的标准化随机变量.

$$\lim_{n\to\infty}P(Y_n\le x)=\Phi(x)$$

即 n 足够大时，Y_n 的分布函数近似于标准正态随机变量的分布函数

$$Y_n\overset{\text{近似}}{\sim}N(0,1)$$

$\sum_{k=1}^{n}X_k=\sqrt{n}\sigma Y_n+n\mu$ 近似服从 $N(n\mu,n\sigma^2)$

中心极限定理的意义

在第二章曾讲过有许多随机现象服从正态分布 是由于许多彼次没有什么相依关系、对随机现象谁也不能起突出影响，而均匀地起到微小作用的随机因素共同作用(即这些因素的叠加)的结果.

若联系于此随机现象的随机变量为X，则它可被看成为许多相互独立的起微小作用的因素X_k的总和 $\sum_k X_k$，而这个总和服从或近似服从正态分布.

定理 2(德莫佛—拉普拉斯(DeMoivre-Laplace）中心极限定理)

设 $Y_n \sim B(n, p)$，$0 < p < 1$，$n = 1,2,\ldots$

则对任一实数 x，有

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

即对任意的 $a < b$,

$$\lim_{n\to\infty} P\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \le b\right) = \frac{1}{\sqrt{2\pi}}\int_{a}^{b} e^{-\frac{t^2}{2}}\,dt$$

$Y_n \sim N(np, np(1\text{-}p))$ (近似)

定理表明, 当n充分大时, 二项分布$B(n,p)$ 可近似地用正态分布$N(np, np(1\text{-}p))$来代替. 因此, 当$X\sim B(n, p)$, 且n充分大时, 有

$$P\{a < X \le b\} \approx \Phi(\frac{b-np}{\sqrt{npq}}) - \Phi(\frac{a-np}{\sqrt{npq}})$$

其中q=1-p.

利用微分中值定理可以进一步证明:

对于随机变量$X\sim B(n, p)$, n充分大时,

$$P\{X=k\}\approx\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}=\frac{1}{\sqrt{npq}}\varphi(\frac{k-np}{\sqrt{npq}})$$

中心极限定理的应用

例1每颗炮弹命中飞机的概率都为0.01,求
(1)500发炮弹中命中5发的概率.
(2)500发炮弹至少命中2发的概率.

解 (1)500发炮弹命中飞机的炮弹数$X \sim B(n,p)$,
其中：$n=500$, $p=0.01$, $np=5$,

下面用两种方法计算并加以比较

1°用二项分布公式计算：

$$P\{X=5\}=C_{500}^{5}\times 0.01^{5}\times 0.99^{495}=0.17635$$

2°用正态分布近似计算

$$P\{X=5\}\approx\frac{1}{\sqrt{npq}}\varphi\left(\frac{5-np}{\sqrt{npq}}\right)\approx 0.1793$$

(2)要求的是P$\{X\geq 2\}$

1°用二项分布公式计算：

$$P\{X\geq 2\}=1-P\{X=0\}-P\{X=1\}$$

$$=1-C_{500}^{0}\times 0.01^{0}\times 0.99^{500}-C_{500}^{1}\times 0.01\times 0.99^{499}\approx 0.96024$$

2°用正态分布近似计算

$$P\{X \geq 2\} = 1 - P\{X < 2\} = 1 - P\{X \leq 1\}$$

$$\approx 1 - \Phi\left(\frac{1 - np}{\sqrt{npq}}\right)$$

$$= 1 - \Phi(-1.7978)$$

$$= 0.96327$$