大数定律与中心极限定理

Didnelpsun

目录

1 依概率收敛 **1**

2 大数定律 **1**

2.1 切比雪夫大数定律 1

2.2 伯努利大数定律 2

2.3 辛钦大数定律 2

3 中心极限定理 **3**

3.1 列维-林德伯格定理 3

3.2 棣莫弗-拉普拉斯定理 4

这些定理与定律是针对极大量数据的概率分析，是概率论向数理统计的过渡。

1 依概率收敛

定义：设随机变量 X 与随机变量序列 $\{X_n\}$（$n=1,2,3\cdots$），如果对任意的 $\epsilon>0$，有 $\lim\limits_{n\to\infty}P\{|X_n-X|\geqslant\epsilon\}=0$ 或 $\lim\limits_{n\to\infty}P\{|X_n-X|<\epsilon\}=1$，则称随机变量序列 $\{X_n\}$ **依概率收敛于随机变量** X，记为 $\lim\limits_{n\to\infty}X_n=X(P)$ 或 $X_n\xrightarrow{P}X(n\to\infty)$。

即在某项后面的全部项全部落在区域内的概率为 1。（不是严格的极限，可能存在超过范围的点，但是不影响后面的点在区域内）

通常使用伯努利试验类似的频率来估计概率，从而来极限逼近真实概率，但是进行试验时很可能不凑巧出现了概率很小的情况从而破坏了试验的概率随着频率变大而逼近真实概率，所以这种就是依概率收敛。比如抛硬币试验概率：$\frac{6}{10}$，$\frac{16}{20}$，$\frac{21}{40}$，其中 $\frac{15}{20}$ 就是一个特殊的情况，但是不影响后面的收敛。

2 大数定律

在满足一定的条件下，大数定律均为 $\frac{1}{n}\sum\limits_{i=1}^{n}X_i\xrightarrow{P}E\left(\frac{1}{n}\sum\limits_{i=1}^{n}X_i\right)$。

所以大数定律一般是考定律成立条件与结论正确性。

2.1 切比雪夫大数定律

定义：假设随机变量序列 $\{X_n\}$（$n=1,2,3\cdots$）是**相互独立**的（不一定同分布），若期望 EX 存在，且方差 $DX_i(i\geqslant1)$**存在且一致有上界**，即存在常数 C，使得 $DX_i\leqslant C$ 对一切 $i\geqslant1$ 均成立，则 $\forall\epsilon>0$，$\lim\limits_{n\to\infty}P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}EX_i\right|<\epsilon\right\}=1$，$\{X_n\}$ 服从大数定律：$\frac{1}{n}\sum\limits_{i=1}^{n}X_i\xrightarrow{P}\frac{1}{n}\sum\limits_{i=1}^{n}EX_i$。即 $\overline{X}\xrightarrow{P}E\overline{X}$。

即均值收敛于期望的均值。

证明：$E\left(\frac{1}{n}\sum\limits_{i=1}^{n}X_i\right)=\frac{1}{n}\sum\limits_{i=1}^{n}EX_i$，又 X_i 不相关，则 $Cov(X_i,X_j)=0$。

则 $D\left(\frac{1}{n}\sum\limits_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum\limits_{i=1}^{n}DX_i\leqslant\frac{nC}{n^2}=\frac{C}{n}$。

又根据切比雪夫不等式 $1\geqslant\lim\limits_{n\to\infty}P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}EX_i\right|<\epsilon\right\}\geqslant1-$

$\dfrac{D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)}{\epsilon^2}\geqslant 1-\dfrac{C}{n\epsilon^2}\xrightarrow{n\to+\infty}1$。

根据夹逼定理 $\lim\limits_{n\to\infty}P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}EX_i\right|<\epsilon\right\}=1$。

例题：设 $X_1,X_2,\cdots,X_n,\cdots$ 为相互独立的随机变量序列，X_n 服从参数为 n 的指数分布（$n\leqslant 1$），则下列随机变量序列中不服从切比雪夫大数定律的是()。

$A.X_1,\frac{1}{2}X_2,\cdots,\frac{1}{n}X_n,\cdots$　　$B.X_1,X_2,\cdots,X_n,\cdots$

$C.X_1,2X_2,\cdots,nX_n,\cdots$　　$D.X_1,2^2X_2,\cdots,n^2X_n,\cdots$

解：切比雪夫大数定律要求有两点，一个是随机变量序列有解，一个是方差存在上界，即 $DX_i\leqslant C$。因为题目说明相互独立，所以只用考虑方差上界。

$\because X_n\sim E(n)$，$\therefore EX_n=\dfrac{1}{n}$，$DX_n=\dfrac{1}{n^2}$。

对于 A，$D\left(\frac{1}{n}X_n\right)=\dfrac{1}{n^2}DX_n=\dfrac{1}{n^4}\leqslant 1$。对于 B，$DX_n=\dfrac{1}{n^2}\leqslant 1$。

对于 C，$D(nX_n)=n^2\dfrac{1}{n^2}=1$，对于 D，$D(n^2X_n)=n^4\dfrac{1}{n^2}=n^2\xrightarrow{n\to\infty}\infty$。

所以选择 D。

2.2 伯努利大数定律

定义：假设 μ_n 是 n 重伯努利试验中事件 A 发生的次数，在每次试验中事件 A 发生的概率为 p（$0<p<1$），则 $\dfrac{\mu_n}{n}\xrightarrow{P}p$，即对任意的 $\epsilon>0$，有 $\lim\limits_{n\to\infty}P\left\{\left|\frac{\mu_n}{n}-p\right|<\epsilon\right\}=\lim\limits_{n\to\infty}P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}EX_i\right|<\epsilon\right\}=\lim\limits_{n\to\infty}P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}EX_i\right|<\epsilon\right\}=1$。

可以看作通过 n 重伯努利试验，一个事件的试验概率 $\dfrac{\mu_n}{n}$ 会逼近一个固定的事件概率 p。

证明：n 重伯努利试验，则 $\mu_n\sim B(n,p)$，$E_{\mu_n}=np$，$D_{\mu_n}=np(1-p)$。

则 $E\left(\dfrac{\mu_n}{n}\right)=p$，$D\left(\dfrac{\mu_n}{n}\right)=\dfrac{p(1-p)}{n}$。

切比雪夫不等式：$\forall\epsilon>0$，$1\geqslant P\{|\dfrac{\mu_n}{n}-p|<\epsilon\}\geqslant 1-\dfrac{p(1-p)}{n\epsilon^2}\xrightarrow{n\to+\infty}1$。

根据夹逼定理 $\lim\limits_{n\to\infty}P\left\{\left|\frac{\mu_n}{n}-p\right|<\epsilon\right\}=1$。

2.3 辛钦大数定律

辛钦大数定律类似切比雪夫大数定律的特殊化，将序列约束为同分布。（但是对方差没有要求，所以不能按切比雪夫定律的证明来做）

定义：假设随机变量序列 $\{X_n\}$（$n=1,2,3\cdots$）是**相互独立**的**同分布**的，如果 $EX_i=\mu$（$i=1,2,\cdots$）**存在**，则 $\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} \mu$，即对任意的 $\epsilon>0$ 有 $\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\epsilon\right\}=1$。也可以转换为即 $\overline{X}\xrightarrow{P}E\overline{X}$。

即能用平均数可以来逼近期望。

例题：假设随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，根据辛钦大数定律，当 $n\to\infty$ 时，$\frac{1}{n}\sum_{i=1}^{n}X_i$ 依概率收敛于数学期望，只要 $\{X_n\}$()。

A. 有相同的数学期望　　*B*. 服从同一离散型分布

C. 服从同一泊松分布　　*D*. 服从同一连续型分布

解：辛钦大数定律要求三点：随机变量序列独立、拥有同样分布、期望存在。

已知题目表示变量相互独立，所以只用证明有同样分布、有期望就可以。

对于 BD 而言满足是有分布的，但是此时不一定有期望，所以 BD 不行。

对于 A 有相同期望，只要求有期望就可以了，相同期望不一定同一分布。

对于 D 服从同一分布，且泊松分布期望存在。

例题：将一枚骰子重复投掷 n 次，当 $n\to\infty$ 时，n 次掷出的点数的算术平均值 $\overline{X}$ 依概率收敛于何值？

解：根据题目，投掷是独立事件，发生概率是离散的同一分布，且期望存在 $=\frac{1}{6}\sum_{i=1}^{6}i=3.5$，所以使用辛钦大数定律。

所以根据辛钦大数定律 $\overline{X_n}\xrightarrow{P}E\overline{X_n}=EX_i=3.5$。

3　中心极限定理

中心极限定理总结来看均为：若 X_i 独立同分布于某一分布 F，则 $\sum_{i=1}^{n}X_i \overset{n\to\infty}{\sim} N(n\mu,n\sigma^2)$。

3.1　列维-林德伯格定理

定义：假设 $\{X_n\}$ 是独立分布的随机变量序列，若 $EX_i=\mu$，$DX_i=\sigma^2>0$（$i=1,2,\cdots$）存在，则对任意的实数 x，有 $\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}\,\mathrm{d}t=\Phi(x)$。（正态分布标准化）

定理要求：独立、同分布、期望方差存在。

$\dfrac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$，$\sum\limits_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$。

例题：已知手套是使用寿命服从指数分布，单位为小时，且平均寿命为 20 小时。若一个人需要带手套进行工作，发现手套坏了就立刻换新继续工作，为保证该工人有 95% 的把握能工作 2000 小时，求应该为其准备手套的副数。

解：题目中需要为其准备手套，且所有使用寿命加在一起大于 2000 的概率为 95%。这个问题是概率分布的和的问题，所以使用列维-林德伯格定理。

假设第 i 副手套的使用寿命为 X_i，则 $X_i \sim E(\lambda)$，又平均寿命为 20 小时，则 $E(X_i) = \dfrac{1}{\lambda} = 20$，即 $\lambda = \dfrac{1}{20}$，$D(X_i) = \dfrac{1}{\lambda^2} = 400$。

又根据中心极限定理，$\sum\limits_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) = N(20n, 400n)$，

保证该工人有 95% 的把握能工作 2000 小时，则 $P\{\sum\limits_{i=1}^{n} X_i \geqslant 2000\} \approx 0.95$，则标准化 $P\left\{\dfrac{\sum X_i - 20n}{20\sqrt{n}} \geqslant \dfrac{2000 - 20n}{\sqrt{20\sqrt{n}}}\right\} \approx 0.95$，$1 - P\left\{Z < \dfrac{100 - n}{\sqrt{n}}\right\} \approx 0.95$，$Z = \dfrac{\sum X_i - 20n}{20\sqrt{n}} \sim N(0,1)$，即 $\Phi(\dfrac{100 - n}{\sqrt{n}}) \approx 0.05$，查表，$\dfrac{n - 100}{\sqrt{n}} \approx 1.64$，$n \approx 118$。

3.2 棣莫弗-拉普拉斯定理

定义：假设随机变量 $Y_n \sim B(n,p)$（$0 < p < 1$，$n \geqslant 1$），则对任意实数 x，有 $\lim\limits_{n\to\infty} P\left\{\dfrac{Y_n - np}{\sqrt{np(1-p)}} \leqslant x\right\} = \dfrac{1}{\sqrt{pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t = \Phi(x)$。

$\dfrac{Y_n - np}{\sqrt{np(1-p)}} \sim N(0,1)$，$X \sim N[np, np(1-p)]$。即二项分布的极限分布是正态分布。其中正态分布的期望和方差就是二项分布的期望和方差。

例题：生产线生产的产品成箱包装，每箱质量是随机的。假设每箱平均钟 50 千克，标准差为 5，若用载重为 5 吨的汽车承运，试用中心极限定理说明每辆汽车最多可以装多少箱才能保证不超载的概率大于 0.977。（$\Phi(2) = 0.977$）

解：设 X_i 为第 i 箱质量，所以 $EX_i = 50$，$DX_i = 25$。

记 $T_n = \sum\limits_{i=1}^{n} X_i$，$ET_n = 50n$，$DT_n = 25n$。

根据中心极限定理得到：$P\{T_n \leqslant 5000\} = P\left\{\dfrac{T_n - 50n}{5\sqrt{n}} \leqslant \dfrac{5000 - 50n}{5\sqrt{n}}\right\} \approx \Phi\left(\dfrac{5000 - 50n}{5\sqrt{n}}\right) > 0.977 = \Phi(2)$。

即 $\dfrac{5000 - 50n}{5\sqrt{n}} \geqslant 2$，即 $n \leqslant 98$，即选 98。