数据科学基础

Foundations of Data Science

4.3 抽样相关定理

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

抽样分布定理

【定理】设某总体的均值为μ，方差为σ^2。X_1，…，X_n为总体一样本，$\bar{X}$为样本均值，S^2为样本方差，则

$$(1)\ \mathrm{E}(\bar{X}) = \mu$$

$$(2)\ \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$(3)\ \mathrm{E}(S^2) = \sigma^2$$

切比雪夫不等式

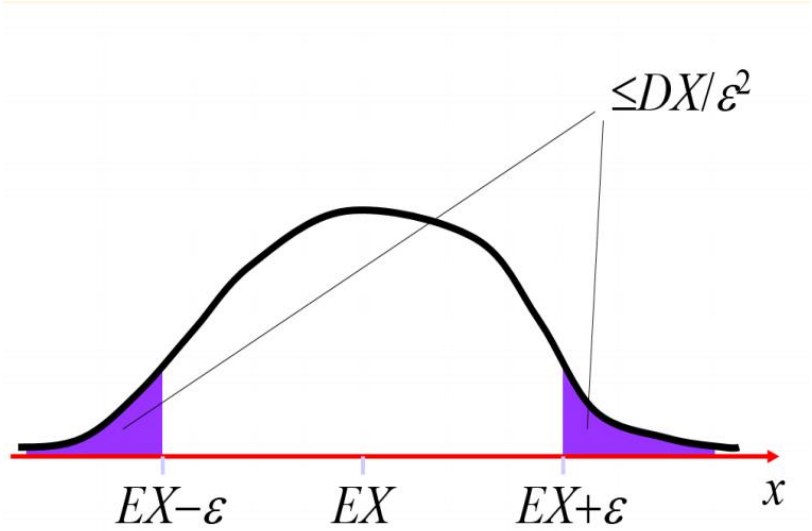
【切比雪夫不等式】 设随机变量具有数学期望

$E(X)=\mu$，方差$Var(X)=\sigma^2$，则对于任意$\varepsilon>0$，有

$$P\{|X-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$$

即等价于

$$P\{|X-\mu|<\varepsilon\}\geq 1-\frac{\sigma^2}{\varepsilon^2}$$



切比雪夫不等式

证明: 设X的概率密度为$f(x)$，则

$$
\begin{aligned}
&P\{|X-\mu| \geq \varepsilon\} \\
&= \int_{|x-\mu| \geq \varepsilon} f(x)dx \\
&\leq \int_{|x-\mu| \geq \varepsilon} \frac{(x-\mu)^2}{\varepsilon^2} f(x)dx \\
&\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \\
&= \frac{Var(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}
\end{aligned}
$$

注明：比雪夫不等式表明$|X-\mu| \geq \varepsilon$的概率由σ控制(ϵ给定)。

切比雪夫不等式示例

设X是掷一颗骰子所出现的点数，若给定ε =1，2，试计算$P(|X-EX|\geq\varepsilon)$，并对比切比雪夫不等式。

$$P(X=k)=\frac{1}{6},\ k=1,\ 2,\ 3,\ 4,\ 5,\ 6$$

$$E(X)=\frac{7}{2},\ E\left(X^2\right)=\frac{91}{6},\ D(X)=\frac{35}{12}$$

$$P(|X-EX|\geq 1)=\frac{2}{3}\leq\frac{D(X)}{\varepsilon^2}=\frac{35}{12}$$

$$P(|X-EX|\geq 2)=\frac{1}{3}\leq\frac{D(X)}{\varepsilon^2}=\frac{35}{48}$$

切比雪夫不等式示例

已知事件A的随机变量$X \sim \mathbb{B}(n，0.75)$，估计事件A发生频率在$0.74-0.76$之间的概率大于0.90的最小实验次数$n$.

$$E(X)=np=0.75n，\ Var(X)=np(1-p)=0.1875n$$

$$f_n(A)=\frac{X}{n}$$

$$P\left\{0.74<\frac{X}{n}<0.76\right\}=P\{|X-0.75n|<0.01n\}$$

$$\geq 1-\frac{0.1875n}{(0.01n)^2}=1-\frac{1875}{n}\geq 0.90$$

解得$n \geq 18750$。

更多详情，请咨询课程助教。

切比雪夫不等式练习

设某大楼有10000盏电灯，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.

大数定律

- 大数定律又称大数法则、大数律，是个数学与统计学的概念，意指数量越多，则其平均就越趋近期望值。
- 在重复试验中，随着试验次数的增加，事件发生的频率趋于一个稳定值；人们同时也发现，在对物理量的测量实践中，测定值的算术平均也具有稳定性。
- 历史上最早的大数定律是伯努利在1713年建立的。概率论的研究到现在约有300多年的历史，最终以事件的频率稳定值来定义其概率。

伯努利大数定律

【伯努利大数定理】设x_n是n重伯努利试验中事件A出现的次数，p是事件在每次试验中A出现的概率，则对任意的$\varepsilon > 0$，有

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

证明：$x_n \sim \mathbb{B}(n, p)$，则

$$E\left(\frac{x_n}{n}\right) = \frac{1}{n}E(n_A) = \frac{np}{n} = p, \quad Var\left(\frac{x_n}{n}\right) = \frac{1}{n^2}Var(x_n) = \frac{pq}{n}$$

$$\forall \varepsilon > 0: P\left\{\left|\frac{x_n}{n} - p\right| < \varepsilon\right\} \geq 1 - \frac{pq}{n\varepsilon^2}$$

伯努利大数定理建立了在大量重复独立试验中事件出现频率的稳定性，正因为这种稳定性，概率的概念才有客观意义。

独立同分布大数定律

【独立同分布的大数定律】 设X_1，X_2，$\cdots$，X_n，是相互独立有相同分布的随机变量序列，各有数学期望$E(X_i)=\mu$，$(i=1，2，\cdots)$，方差有$D(X_i)=\sigma^2(i=1，2，\cdots)$，则对任意的$\varepsilon>0$，有

$$\lim_{n\to+\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1$$

切比雪夫大数定律

【切比雪夫大数定律】设X_1，X_2，…，X_n，是相互独立的随机变量序列，各有数学期望$E(X_i)=\mu$，$(i=1，2，\cdots)$和有限的方差，并且方差有$D(X_i)\ (i=1，2，\cdots)$共同的上界，即$D(X_i)\leq c$则对任意的$\varepsilon>0$，

$$\lim_{n\to+\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}EX_i\right|<\varepsilon\right\}=1$$

辛钦大数定律

【辛钦大数定理】设X_1，X_2，…，X_n，是独立同分布的随机变量序列，只要数学期望$EX_i=\mu$，$(i=1，2，\cdots)$存在，则对任意的$\varepsilon>0$，有

$$\lim_{n\to+\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i-\mu\right|<\varepsilon\right\}=1$$

伯努利大数定理是辛钦大数定理的特殊情况。

中心极限定理

- 中心极限定理讨论随机变量和的分布以正态分布为极限的一组定理。
- 中心极限定理的第一版被法国数学家棣莫弗在1733年发表的卓越论文中，计算大量抛掷硬币出现次数的极限分布构建的一个概率分布。
- 这个超越时代的成果险些被历史遗忘，所幸著名法国数学家拉普拉斯在1812年发表的巨著中拯救了这个默默无名的理论。拉普拉斯扩展了棣莫弗的理论，指出二项分布可用正态分布逼近。
- 但同棣莫弗一样，拉普拉斯的发现在当时并未引起很大反响。直到十九世纪末经过高斯等人的推动，中心极限定理的重要性才被世人所知。

中心极限定理

【德莫佛-拉普拉斯中心极限定理】设$Y_n \sim \mathbb{B}(n，p)$，$0 < p < 1$，$n = 1，2，\cdots$，则对任意实数x，有

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

更多详情，请咨询课程助教。

中心极限定理计算

设某工厂有400台同类机器，各台机器发生故障的概率都是0.02，各台机器工作是相互独立的。试求机器出故障的台数不小于2的概率。

解:设故障台数为X，则X~$\mathbb{B}(400，0.02)$

(1)二项分布计算

$$P(X \geq 2) = 1 - P(\mathrm{X} = 0) - \mathrm{P}(\mathrm{X} = 1) = 0.9972$$

中心极限定理计算

(2) 泊松分布计算

$$\lambda = np = 400 * 0.02 = 8$$

查表得

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - 0.000335 - 0.002684 \approx 0.9969$$

(3)正态分布计算

$$\sqrt{npq} = \sqrt{400 * 0.02 * 0.98} = 2.8$$

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \Phi\left(\frac{1 - np}{\sqrt{npq}}\right) = \Phi\left(\frac{7}{2.8}\right) = 0.9938$$

中心极限定理练习

某保险公司的老年人寿保险有1万人参加，每人每年交200元。若老人在该年内死亡，公司付给受益人1万元。设老年人死亡率为0.017，试求保险公司在一年内这项保险亏本的概率。保险公司在一年内这项保险盈利10万元以上的概率。

中心极限定理

- 棣莫弗和拉普拉斯提出了二项式的中心极限定理雏形。
- 随后的一百多年，一大批数学家们前赴后继努力给出普适性的中心极限定理。贡献者包括泊松、狄利克莱、柯西、贝塞尔等知名数学家 。
- 普适性的中心极限定理雏形是切比雪夫从1887年开始的。
- 马尔科夫和李雅普诺夫都是切比雪夫的学生，马尔科夫沿着老师的基于矩法的思路推进证明。李雅普诺夫则沿着拉普拉斯当年提出的基于特征函数的思路，并于1901年给出了第一个普适性严格证明。

中心极限定理

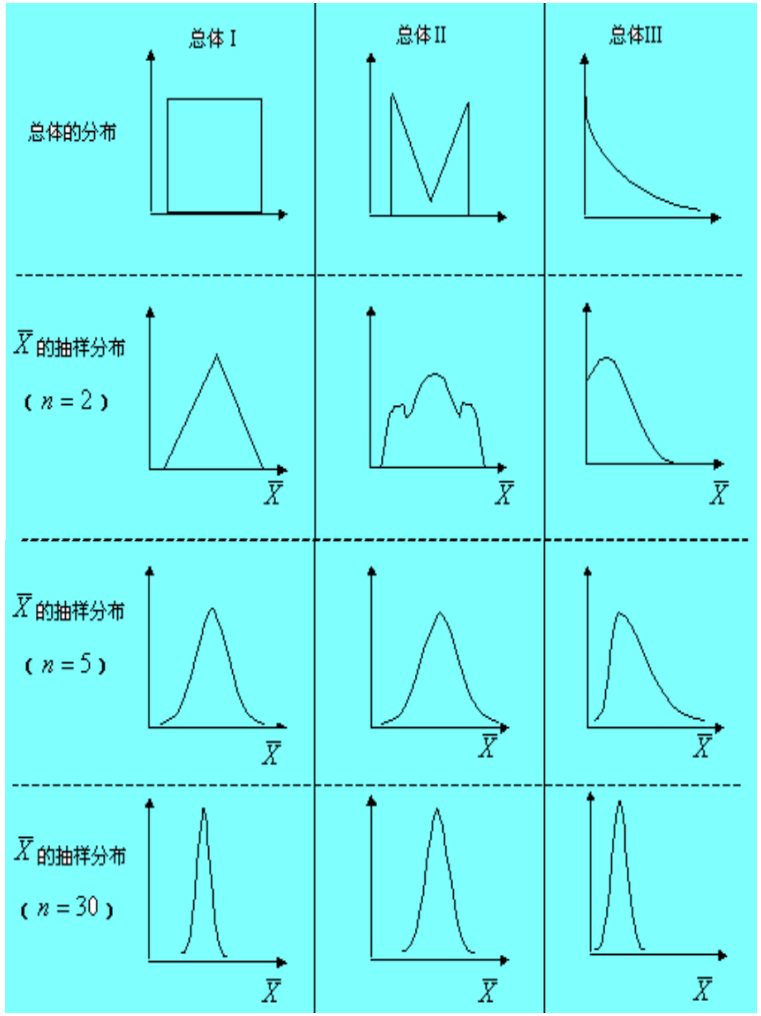
【独立同分布的中心极限定理】设X_1，X_2，…，X_n是相互独立有相同分布的随机变量序列，且有期望和方差$E(X_i)=\mu$，$Var(X_i)=\sigma^2$，则对任意实数x，

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

- 中心极限定理说明，对于任意分布，当样本容量足够大时，样本均值的抽样分布近似正态分布$N(\mu，\sigma^2/n)$。
- 在总体分布信息未知的情况下，经验上通常 $n\ge 30$ 认为样本容量足够大。

中心极限定理

原始总体分布与正态分布越接近， 其抽样分布收敛到正态分布的速度越快。当原始总体分布服从正态分布时，独立同分布的正态样本均值服从正态分布。



【独立同分布的正态样本】设X_1，X_2，…，X_n总体$N(\mu，\sigma^2)$的样本，$\bar{X}$和S^2分别是样本均值和样本方差，则有

$$\bar{X}\sim N(\mu，\sigma^2/n)$$

更多详情，请咨询课程助教。

中心极限定理计算

设某种电器元件的寿命服从均值为100小时的指数分布， 现随机取得16只， 设它们的寿命是相互独立的，求这16只元件的寿命的总和大于1920小时的概率。

中心极限定理计算

解: 记16只元件的寿命分别为X_1，…，X_{16}则16只电器元件的寿命总和为$X=\sum_{i=1}^{16}X_i$由题意知

$$E(X_i)=100，Var(X_i)=100^2$$

根据中心极限定理

$$Y=\frac{X-16*100}{4*100}=\frac{X-1600}{400}\sim\mathbb{N}(0，1)$$

$$P(X>1920)=1-P(X\leq 1920)\approx 1-\Phi(\frac{1920-1600}{400})$$

$$=1-\Phi(0.8)=0.2119$$

更多详情，请咨询课程助教。