大数定律与中心极限定理

第 8 讲 · 概率论与数理统计

复旦大学经济学院　ECON130001

上一讲回顾：数字特征与变量的和

平均水平、离散程度与线性关联

期望 EX、方差 $D(X) = E(X^2) - (EX)^2$、协方差与相关系数分别描述随机变量的不同特征。

和的期望与方差

$$E(X+Y) = EX + EY,$$
$$D(X+Y) = D(X) + D(Y) + 2\mathrm{Cov}(X,Y).$$

第一式要求期望有限；第二式要求方差有限。独立时协方差为0。

本讲的问题

对独立同分布的 $X_1, \ldots, X_n$，均值 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 会怎样变化？

大数定律研究均值是否接近期望；中心极限定理研究标准化后的波动呈现何种分布。

连续出现黑色后，红色更可能出现吗？

背景 一则广为引用的记述称，1913年蒙特卡洛赌场曾连续出现26次黑色。

转折 部分赌客因此押注红色，认为此前黑色出现过多，接下来红色应当更容易出现。这种判断称为**赌徒谬误**。

问题 **公平的单零轮盘有37格，其中18格为红色。各次旋转相互独立，即使前26次全是黑色，下一次出现红色的概率仍为** $18/37$**。**

如果继续旋转，红色次数占总次数的比例会怎样变化？

大数定律说明，旋转次数越多，红色频率偏离 $18/37$ **超过给定幅度的概率趋于0。**

案例来源

1913年蒙特卡洛赌场连续出现26次黑色的记述，见 Peacock（2013），Significance。下述轮盘概率按公平单零轮盘模型计算。核对日期：2026-11-19。

本讲学习目标

1. 利用切比雪夫不等式估计偏离期望的概率界限。
2. 理解依概率收敛，并辨认不同大数定律的适用条件。
3. 用中心极限定理对和或均值进行标准化，计算近似概率。
4. 给定允许误差和目标概率，用切比雪夫不等式求一个足够的样本量，用正态近似估算所需的样本量。

两个基本问题

随着独立重复试验次数增加，经验频率为何接近理论概率？

为什么许多独立随机因素相加后，在适当标准化下可以用正态分布近似？

抛硬币次数增加时，两种概率怎样变化？

公平硬币的独立重复试验

比较「正面次数恰好占一半」与「正面频率在0.4至0.6之间（含端点）」的概率。

抛10次： $X \sim B(10, 1/2)$

$$P(X = 5) = \binom{10}{5} 2^{-10} \approx 0.2461$$

$$P(4 \le X \le 6) = \sum_{k=4}^{6} \binom{10}{k} 2^{-10} \approx 0.6563$$

抛100次： $X \sim B(100, 1/2)$

$$P(X = 50) = \binom{100}{50} 2^{-100} \approx 0.0796$$

$$P(40 \le X \le 60) = \sum_{k=40}^{60} \binom{100}{k} 2^{-100} \approx 0.9648$$

频率接近概率的含义

在这两组计算中，「恰好一半」的概率减小，而频率落在给定邻域内的概率增大。大数定律描述的是频率接近理论概率的趋势。

从样本均值提出问题

设定

$X_1, X_2, \ldots$ 独立同分布，期望存在且有限，记为 μ。记

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

需要研究的概率

给定允许的误差 $\varepsilon > 0$，考察

$$P(|\overline{X}_n - \mu| \geq \varepsilon).$$

当样本量增加时，这个概率是否趋于0?

先研究方差有限的情形：利用期望和方差，可以在不求出完整分布的情况下给出概率界限。

切比雪夫不等式

定理

设 $EX = \mu$，$D(X) = \sigma^2 < \infty$。对任意 $\varepsilon > 0$，

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

等价地，

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

依据与用途

在事件 $|X - \mu| \geq \varepsilon$ 上，平方离差至少为 ε^2，因此

$$\sigma^2 = E[(X - \mu)^2] \geq \varepsilon^2 P(|X - \mu| \geq \varepsilon).$$

这个界限适用于任何具有有限方差的分布；在具体问题中，它可能较保守。

切比雪夫不等式的两种应用

由方差给出概率下界

10000盏灯独立亮起，每盏亮起概率为0.7。令 X 为亮灯数，估计 $P(6800 < X < 7200)$ 的下界。

$EX = 7000$，$DX = 2100$。

$$P(|X - 7000| < 200) \geq 1 - \frac{2100}{200^2} = 0.9475.$$

由尾概率给出方差下界

设 X 有有限方差，$EX = 10$，$P(X \leq 7) = 0.2$，$P(X \geq 13) = 0.3$。

$$P(|X - 10| \geq 3) = 0.2 + 0.3 = 0.5.$$

由切比雪夫不等式，

$$0.5 \leq \frac{DX}{9} \quad \Longrightarrow \quad DX \geq \frac{9}{2}.$$

样本均值的期望与方差

独立同分布、有限方差

设 $EX_i = \mu$、$D(X_i) = \sigma^2 < \infty$，则

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu,$$

$$D(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}.$$

均值随样本量增加而更集中

样本均值的标准差为 $\sigma/\sqrt{n}$。要使它减半，样本量需变为原来的4倍。

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0.$$

用切比雪夫不等式确定足够的样本量

例

从同一总体独立抽样，已知总体标准差不超过2。希望 $P(|\overline{X}_n - \mu| < 1) \geq 0.99$，用切比雪夫不等式给出足够的样本量。

计算

$$P(|\overline{X}_n - \mu| < 1) \geq 1 - \frac{\sigma^2}{n} \geq 1 - \frac{4}{n}.$$

$$1 - \frac{4}{n} \geq 0.99 \quad \Longleftrightarrow \quad n \geq 400.$$

因此，取400个或更多独立样本即可得到所要求的概率保证。这里给出的是充分条件。

依概率收敛的含义

定义

若对任意 $\varepsilon > 0$，

$$P(|X_n - a| \geq \varepsilon) \longrightarrow 0,$$

则称 X_n 依概率收敛于常数 a，记为 $X_n \xrightarrow{P} a$。

对频率的解释

固定任意误差范围，频率落在这个范围之外的概率随试验次数增加而趋于0。

回到轮盘的例子

设前26次均为黑色，之后 m 次出现黑色 B_m 次，$p = 18/37$。总频率满足

$$\frac{26 + B_m}{26 + m} - p = \frac{26(1-p)}{26+m} + \frac{m}{26+m}\left(\frac{B_m}{m} - p\right).$$

第一项随试验次数增加而趋于0；第二项由后续独立试验的大数定律趋于0，无需改变下一次的概率。

三种大数定律的条件与结论

切比雪夫大数定律

X_i 相互独立，方差一致有界：存在有限常数 C，使所有 $D(X_i) \le C$。则

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \xrightarrow{P} 0.$$

不要求同分布；各项期望可以不同。

伯努利大数定律

独立重复试验中事件 A 的概率均为 p。若发生次数为 N_n，则

$$\frac{N_n}{n} \xrightarrow{P} p.$$

这是独立同分布0—1变量的均值收敛。

辛钦大数定律

X_i 独立同分布，且 $E|X_i| < \infty$。记 $EX_i = \mu$，则

$$\overline{X}_n \xrightarrow{P} \mu.$$

不要求方差有限。

用概率界限刻画总长度与累积误差的波动

部件总长度

10个部件的长度独立同分布，期望为2 mm，标准差为0.01 mm。总长度 Y 满足 $|Y-20|<0.1$ 时合格。

$EY=20$，$DY=10\times 0.01^2=0.001$。

$$P(|Y-20|<0.1)\geq 1-\frac{0.001}{0.1^2}=0.9.$$

独立均匀舍入误差

设每次舍入误差独立服从 $U(-0.5,0.5)$。1500次误差之和为 S，则 $ES=0, DS=125$。

$$P(|S|\geq 15)\leq \frac{125}{225}=\frac{5}{9}.$$

若希望 n 次误差之和 S_n 满足 $P(|S_n|<10)\geq 0.9$，以下条件足够：

$$1-\frac{n/12}{10^2}\geq 0.9 \quad\Longleftrightarrow\quad n\leq 120.$$

两个结果均为切比雪夫不等式提供的界限。第18页将用正态近似重新估计舍入误差。

应用：保单平均收益为何更稳定

简化模型

每份保单收取500元，出险时赔付100000元。忽略其他费用，假设各保单独立同分布、出险概率为 p，单份收益为 X_i。

$$EX_i = 500(1-p) - 99500p = 500 - 100000p.$$

大数定律的结论

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} 500 - 100000p.$$

在该模型下，保单数增加时，每份保单的平均收益依概率收敛于其期望。期望收益为正的条件是 $p < 0.005$。

共同灾害可能使多份保单同时出险；这时需要重新考虑保单之间的相关性，不能直接沿用上述独立模型。

应用：以样本平均估计期望

固定预测模型的平均损失

对一个已确定的预测模型，在独立同分布的测试样本上记录损失 $L_1, \ldots, L_n$。若 $E|L_i| < \infty$，则

$$\frac{1}{n}\sum_{i=1}^{n} L_i \xrightarrow{P} E(L_1).$$

平均测试损失可以用来估计该模型在总体上的平均损失。

蒙特卡洛积分

要计算 $I = E[g(X)]$，从 X 的分布独立抽取 $X_1, \ldots, X_n$。若 $E|g(X)| < \infty$，则

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{P} I.$$

这说明随机抽样得到的平均值可以作为积分的近似。

从均值收敛到波动的近似分布

工具	回答的问题
切比雪夫不等式	已知有限方差时，偏离期望至少 ε 的概率有多大的上界？
大数定律	样本量增加时，样本均值是否依概率接近期望？
中心极限定理	在适当标准化后，和或均值的分布是否趋近标准正态？

下一步

利用标准化后的极限分布，估计有限样本下的概率。大数定律本身并不等同于切比雪夫上界；极限分布也不等同于有限样本的精确分布。

中心极限定理：独立同分布情形

定理

设 $X_1, X_2, \ldots$ 独立同分布，$EX_i = \mu$，$0 < D(X_i) = \sigma^2 < \infty$。令

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$

则对任意实数 z，

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z),$$

其中 Φ 为标准正态分布函数。

结论的含义

各项本身可以不是正态变量。随着项数增加，和经过中心化、标准化后的分布趋近标准正态分布。

本讲的主要例题均满足独立同分布条件。

舍入误差：严格界限与正态近似

沿用第13页的模型：1500次独立 $U(-0.5, 0.5)$ 误差之和为 S，$ES = 0$、$DS = 125$。

切比雪夫上界

$$P(|S| \geq 15) \leq \frac{5}{9} \approx 0.5556.$$

中心极限定理近似

$$P(|S| \geq 15) \approx 2\left[1 - \Phi\left(\frac{15}{\sqrt{125}}\right)\right] \approx 0.1797.$$

相加次数的近似范围

要求 $P(|S_n| < 10) \geq 0.9$，按正态近似有

$$2\Phi\left(\frac{10}{\sqrt{n/12}}\right) - 1 \geq 0.9 \quad \Longrightarrow \quad n \leq \frac{1200}{z_{0.95}^2} \approx 443.53.$$

其中 $\Phi(z_{0.95}) = 0.95$，$z_{0.95} \approx 1.64485$，故近似得到 $n \leq 443$。

上界与近似值回答不同问题；正态近似给出 $n \leq 443$；切比雪夫不等式给出的充分条件是 $n \leq 120$。

样本均值的近似分布与标准误

独立同分布、正的有限方差

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$

因此，样本量充分大时可采用

$$\overline{X}_n \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

记号 $\dot{\sim}$ 表示近似服从。

标准误

样本均值的标准差 $\sigma/\sqrt{n}$ 称为它的**标准误**，记为 $\mathrm{SE}(\overline{X}_n)$。它描述不同重复样本得到的均值之间的波动。

总体标准差 σ 描述单个观测的波动；两者对应不同的随机变量。

按正态近似估算样本量

记 $\Phi(z_q)=q$。对独立同分布且具有正的有限方差的样本，近似计算 $P(|\overline{X}_n-\mu|<d)\approx 2\Phi(d\sqrt{n}/\sigma)-1$。

器件寿命：误差小于1的概率约为95%

寿命以小时计，总体标准差为20，取 $d=1$。

$$\frac{\sqrt{n}}{20}\geq z_{0.975}\approx 1.95996,$$
$$n\geq (20z_{0.975})^2\approx 1536.58.$$

向上取整，取 $n=1537$。

与第10页比较：标准差不超过2

误差小于1的目标概率为99%，按标准差上限2计算：

$$\frac{\sqrt{n}}{2}\geq z_{0.995}\approx 2.57583,$$
$$n\geq (2z_{0.995})^2\approx 26.54.$$

正态近似给出 $n\geq 27$；切比雪夫不等式给出的充分条件是 $n\geq 400$。

正态近似得到的样本量依赖近似效果，不具有切比雪夫界限的分布无关保证。

二项分布的正态近似与连续性修正

棣莫弗—拉普拉斯定理

固定 $0 < p < 1$，若 $X_n \sim B(n, p)$，则对任意实数 z，

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le z\right) \longrightarrow \Phi(z).$$

整数计数的连续性修正

将整数 k 对应为宽度1的区间 $[k-0.5, k+0.5]$。对整数 $a \le b$，

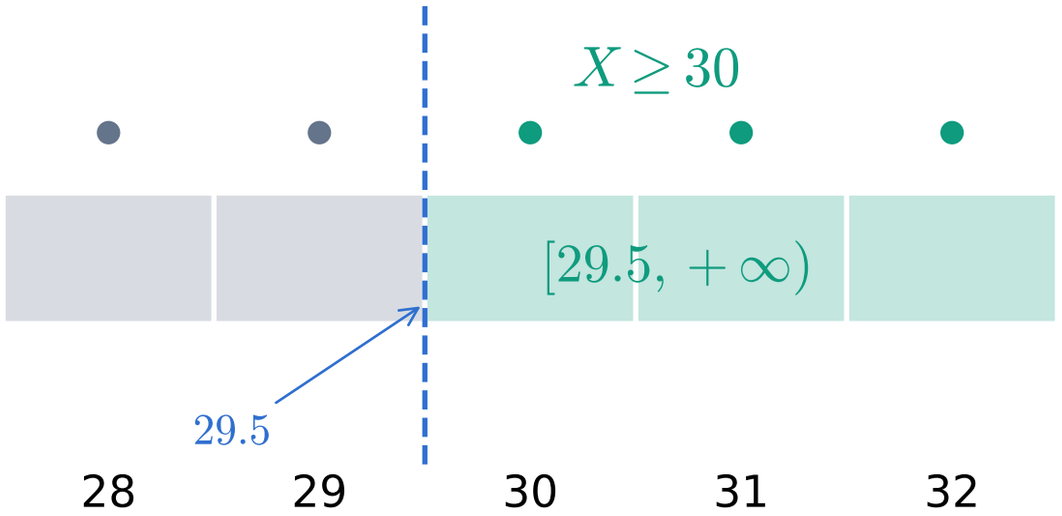
$$P(a \le X_n \le b) \approx \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

例如，$X_n \ge 30$ 对应连续变量从29.5开始的区间。

例：木柱数量的正态近似

模型

每根木柱短于3 m的概率为0.2。独立抽取100根，令 Y 为短于3 m的根数，则 $Y \sim B(100, 0.2)$，$EY = 20$、$DY = 16$。求 $P(Y \geq 30)$。



整数30对应的区间从29.5开始。

采用连续性修正

$$P(Y \geq 30) \approx 1 - \Phi\left(\frac{29.5 - 20}{4}\right) \approx 0.00877.$$

方法	结果
不作连续性修正	0.00621
作连续性修正	0.00877
二项分布直接计算	0.01125

本例的连续性修正改善了近似结果，但与二项分布直接计算仍有差异。

二项分布的两种近似

近似分布	适用情形	需要的参数
泊松分布 $\mathrm{Poisson}(np)$	试验次数多、单次成功概率小；常用于 np 适中的稀有事件计数	$\lambda = np$
正态分布 $N(np, np(1-p))$	np 与 $n(1-p)$ 都较大，分布不集中在计数范围的边缘	均值 np，方差 $np(1-p)$

选择近似方法

常见的初步经验检查是 $np \geq 5$ 且 $n(1-p) \geq 5$。计算整数区间概率时配合连续性修正。

这些条件用于初步判断；需要精确结果时，可以直接计算二项概率。

综合例题：两种家长会人数

题目

每个学生有0、1、2名家长参会，概率分别为0.05、0.8、0.15。400个学生的参会情况相互独立。求两类人数的近似概率。

（1）参会家长总数超过450

令 X_i 为第 i 个学生的参会家长数。$EX_i = 1.1$，$E(X_i^2) = 1.4$，$DX_i = 0.19$。总数 $S = \sum_i X_i$，$ES = 440$、$DS = 76$。

$$\begin{aligned} P(S > 450) &= P(S \geq 451) \\ &\approx 1 - \Phi\left(\frac{450.5 - 440}{\sqrt{76}}\right) \\ &\approx 0.1142. \end{aligned}$$

（2）恰有1名家长参会的学生数不多于340

令 Y 为这类学生数，则 $Y \sim B(400, 0.8)$，$EY = 320$、$DY = 64$。

$$P(Y \leq 340) \approx \Phi\left(\frac{340.5 - 320}{8}\right) \approx 0.9948.$$

第一问汇总每名学生的参会家长数；第二问统计满足指定条件的学生数。

应用：寿命之和与体重之和

依次更换元件的总使用时间

100个元件依次使用，失效后立即更换下一个。各寿命独立同分布，期望1000小时、标准差100小时。总使用时间为它们的和 S。

$ES = 100000$，$DS = 1000000$。

$$P(S > 98000) \approx 1 - \Phi(-2) \approx 0.9772.$$

这里通过中心极限定理近似求和后的分布。

正态体重模型下的载重概率

教学模型：10位乘客的体重相互独立，均服从 $N(70, 10^2)$（单位kg）。总重 T 超过800 kg的概率为多少?

由独立正态变量的可加性，$T \sim N(700, 1000)$。

$$P(T > 800) = 1 - \Phi\left(\frac{100}{\sqrt{1000}}\right) \approx 0.000783.$$

这一正态分布结论来自题设，无需使用大样本近似。

应用：两组样本比例的差异

A/B测试的一个教学模型

网页A、B各分配1000名不同用户。假设所有用户独立，完成注册的概率（转化率）在两个版本中均为0.10。样本转化比例分别为 $\hat{p}_A, \hat{p}_B$。

真实转化率相同，样本比例仍可能有差异

记 $W = \hat{p}_B - \hat{p}_A$。两组独立，因此

$$EW = 0, \qquad D(W) = 2 \times \frac{0.1 \times 0.9}{1000} = 0.00018.$$

比例差的相邻取值间隔为0.001，作连续性修正后，

$$P(W \geq 0.03) \approx 1 - \Phi\left(\frac{0.0295}{\sqrt{0.00018}}\right) \approx 0.0139.$$

即使真实转化概率相同，B的样本比例也可能高出A至少3个百分点。后续将学习如何结合抽样波动进行参数估计与假设检验。

本讲总结：三个工具的分工

工具	条件与核心结论
切比雪夫不等式	方差有限时，$P(\lvert X-EX\rvert \geq \varepsilon) \leq D(X)/\varepsilon^2$。给出严格概率界限。
大数定律	独立同分布、有限期望时，$\overline{X}_n \xrightarrow{P} \mu$。描述均值的集中趋势。
中心极限定理	独立同分布、$0<\sigma^2<\infty$ 时，$(\overline{X}_n-\mu)/(\sigma/\sqrt{n})$ 的分布趋于标准正态。

样本均值的期望、方差与标准误

$$E(\overline{X}_n)=\mu, \qquad D(\overline{X}_n)=\frac{\sigma^2}{n}, \qquad \mathrm{SE}(\overline{X}_n)=\frac{\sigma}{\sqrt{n}}.$$

方差公式在独立同分布、有限方差下精确成立；用正态分布计算概率通常是近似。

下一讲：样本及抽样分布

总体参数在实际问题中通常未知。下一讲研究样本构成的统计量及其分布，为参数估计和假设检验作准备。