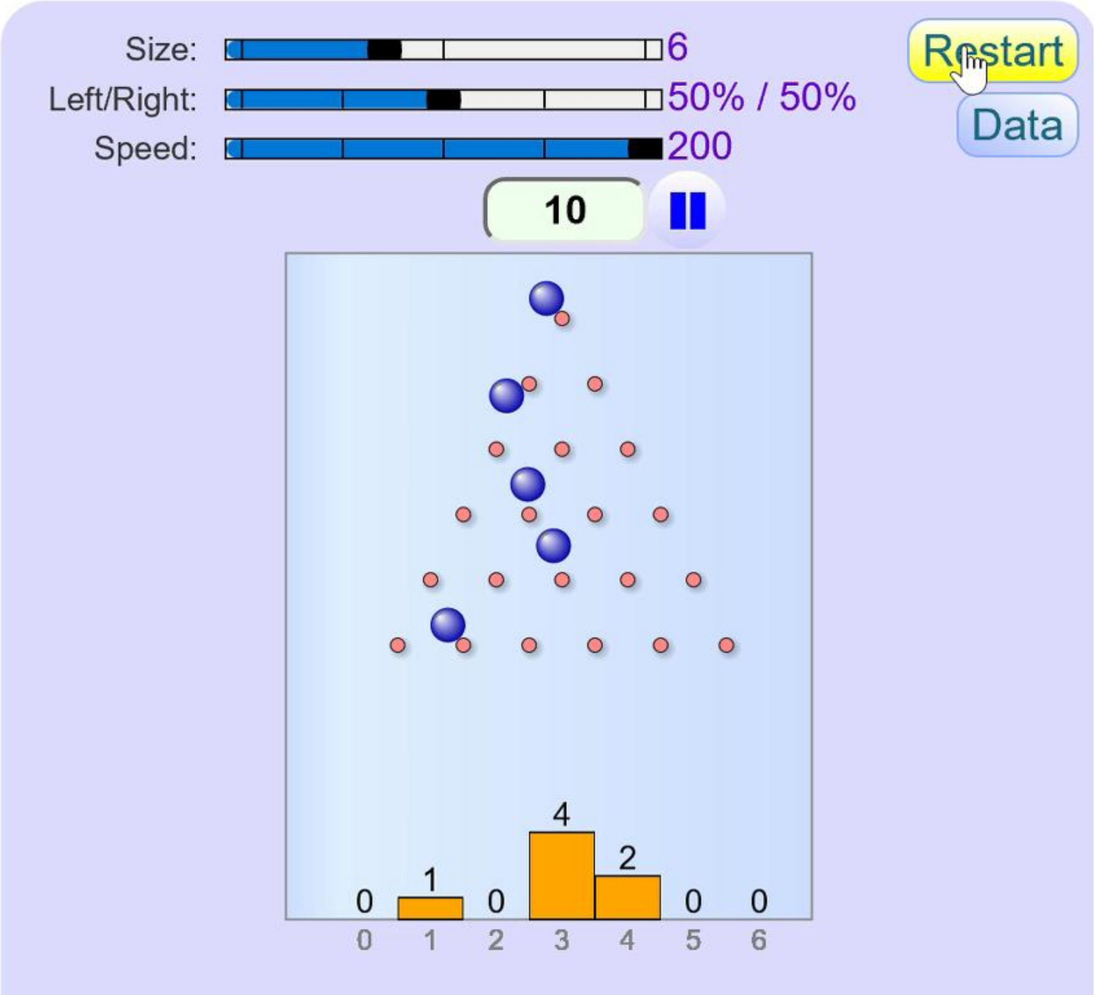
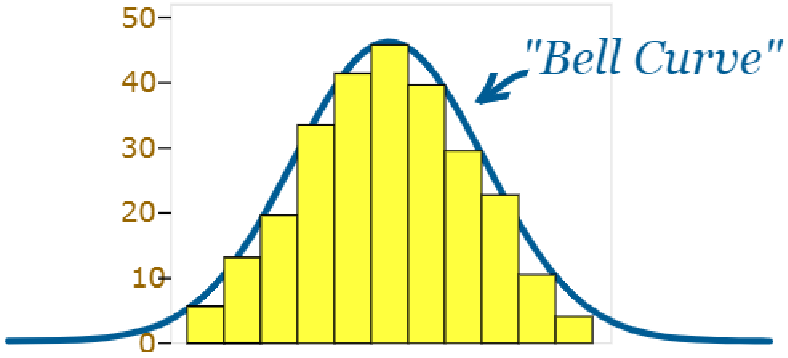
6.2. 中心极限定理(Central Limit Theorem)—引例：高尔顿板





https://www.mathsisfun.com/data/quincunx.html

定理4.7 棣莫弗-拉普拉斯中心极限定理

设 $X_1, X_2, \dots, X_n, \dots$ 是一列 i.i.d. 随机变量序列，$X_i \sim B(1,p)$，$0 < p < 1$，

$S_n = \sum_{i=1}^{n} X_i$，则$\forall x \in \mathbb{R}$, 有

$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x).$$

$\Phi(x)$为标准正态分布函数.

即：$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} N(0,1)$ 或 $\frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} N(0,1)$. $(\bar{X} = S_n/n)$

定理4.6 林德伯格-莱维中心极限定理

设 $X_1, X_2, \ldots, X_n, \ldots$ 是一列 i.i.d. 随机变量序列，记它们相同的期望和方差分别为μ, σ^2. $S_n = \sum_{i=1}^n X_i$，则$\forall x \in \mathbb{R}$, 有

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(S_n/n - \mu)}{\sigma} \le x\right) = \Phi(x).$$

即：$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathcal{L}} N(0,1)$ 或 $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{\mathcal{L}} N(0,1)$. $(\bar{X} = S_n/n)$

$$\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{\mathcal{L}} N(0,1) \quad \text{或} \quad \frac{\bar{X} - E(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} \xrightarrow{\mathcal{L}} N(0,1)\right)$$

- 一般而言，当$n \geq 30$时，中心极限定理的近似程度能满足一般的要求。
- 中心极限定理和大数定律都是普适性结论, 即它们不依赖总体是什么分布, 其结论都是一样的
  - ✓简单随机样本的样本均值依概率收敛于总体期望, 而它的分布收敛于正态分布.

- 利用正态分布估计二项分布 $X \sim B(n,p)$，当n较大时，由定理4.7，近似有

$$\mathbb{P}(X=k) = \mathbb{P}(k-1/2 < X \leq k+1/2) \approx \Phi\left(\frac{k+1/2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-1/2-np}{\sqrt{np(1-p)}}\right)$$

$$\overset{\text{中值定理}}{\approx} \frac{1}{\sqrt{np(1-p)}}\varphi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) = \frac{1}{\sqrt{2\pi np(1-p)}}\exp\left\{-\frac{(k-np)^2}{2np(1-p)}\right\} \tag{4.29}$$

$$\mathbb{P}(k_1 \leq X \leq k_2) \approx \Phi\left(\frac{k_2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1-np}{\sqrt{np(1-p)}}\right), \tag{4.30}$$

$$\mathbb{P}(k_1 \leq X \leq k_2) = \mathbb{P}(k_1-1/2 < X < k_2+1/2) \approx \Phi\left(\frac{k_2+1/2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1-1/2-np}{\sqrt{np(1-p)}}\right), \tag{4.31}$$

这是对公式(4.30)的连续性修正，因为正态分布是连续的，而二项分布分布函数是阶梯函数，因此将下限 k_1 和上限 k_2 分别修正为 $k_1-1/2$ 和 $k_2+1/2$. 这一修正在近似计算二项分布两头的概率时比较有效. 区间不在两头时可以不作修正，直接用 k_1 和 k_2，即公式(4.30).

- 例：设 $X \sim B(20, 0.3)$, 求 $\mathbb{P}(X=3)$ 和 $\mathbb{P}(1 \le X \le 3)$.
  - 直接计算：
$$\mathbb{P}(X=3) = \binom{20}{3} \cdot 0.3^3 \cdot 0.7^{17} = 0.07160,$$
$$\mathbb{P}(1 \le X \le 3) = \sum_{k=1}^{3} \binom{20}{k} 0.3^k 0.7^{20-k} = 0.10629,$$
  - 正态逼近：
$$\mathbb{P}(X=3) \approx \frac{1}{\sqrt{2\pi \times 20 \times 0.3 \times 0.7}} \exp\left\{-\frac{(3-6)^2}{2 \times 20 \times 0.3 \times 0.7}\right\} = 0.06668.$$

使用修正的中心极限定理近似式(4.31)有
$$\mathbb{P}(1 \le X \le 3) \approx \Phi\left(\frac{3.5-6}{\sqrt{4.2}}\right) - \Phi\left(\frac{0.5-6}{\sqrt{4.2}}\right) = 0.99636 - 0.88875 = 0.10761,$$
而使用未修正的正态分布近似式(4.30)计算有
$$\mathbb{P}(1 \le X \le 3) \approx \Phi\left(\frac{3-6}{\sqrt{4.2}}\right) - \Phi\left(\frac{1-6}{\sqrt{4.2}}\right) = 0.99265 - 0.92838 = 0.06427.$$

由上可见, 即使在 $n = 20$ 场合, 用正态逼近 $\mathbb{P}(X=3)$ 的误差不超过 0.005, 用修正的正态分布来近似 $\mathbb{P}(1 \le X \le 3)$, 其误差不超过 0.0014, 但是用不修正的中心极限定理来估计误差要大一点.

- 设计算机按四舍五入进行运算, 求计算机运行 n 步后累积误差绝对值超过 $0.6\sqrt{n}$的概率.

例子(Page 32续)

- 为了解某地区初中生人群沉迷于网络游戏的比例 p, 需要对该地区人群进行调查. 问要调查多少青少年, 才能使调查得到的网络游戏沉迷比例与真实网络游戏沉迷率之差的绝对值小于 0.5% 的概率不小于 95%.

  ✓ 利用切比雪夫不等式，$n \geq 2 \times 10^5$

  ✓ 利用中心极限定理，$n \geq 38416$ $[\mathrm{p}(1-\mathrm{p}) \leq 1/4]$

  若p能有一个大致范围，调查人群规模还能减少。如：第一步先调查少量青少年，比如 200 人, 大致确定 p的一个范围, 例如 $\mathrm{p} \in (0.2, 0.3)$, 此时 $\mathrm{p}(1-\mathrm{p}) \leq 0.3 \times 0.7 = 0.21$, 由此得到调查人数 n 只要满足 $n \geq 32269.44$, 即只要调查 32270个青少年即可满足精度要求.

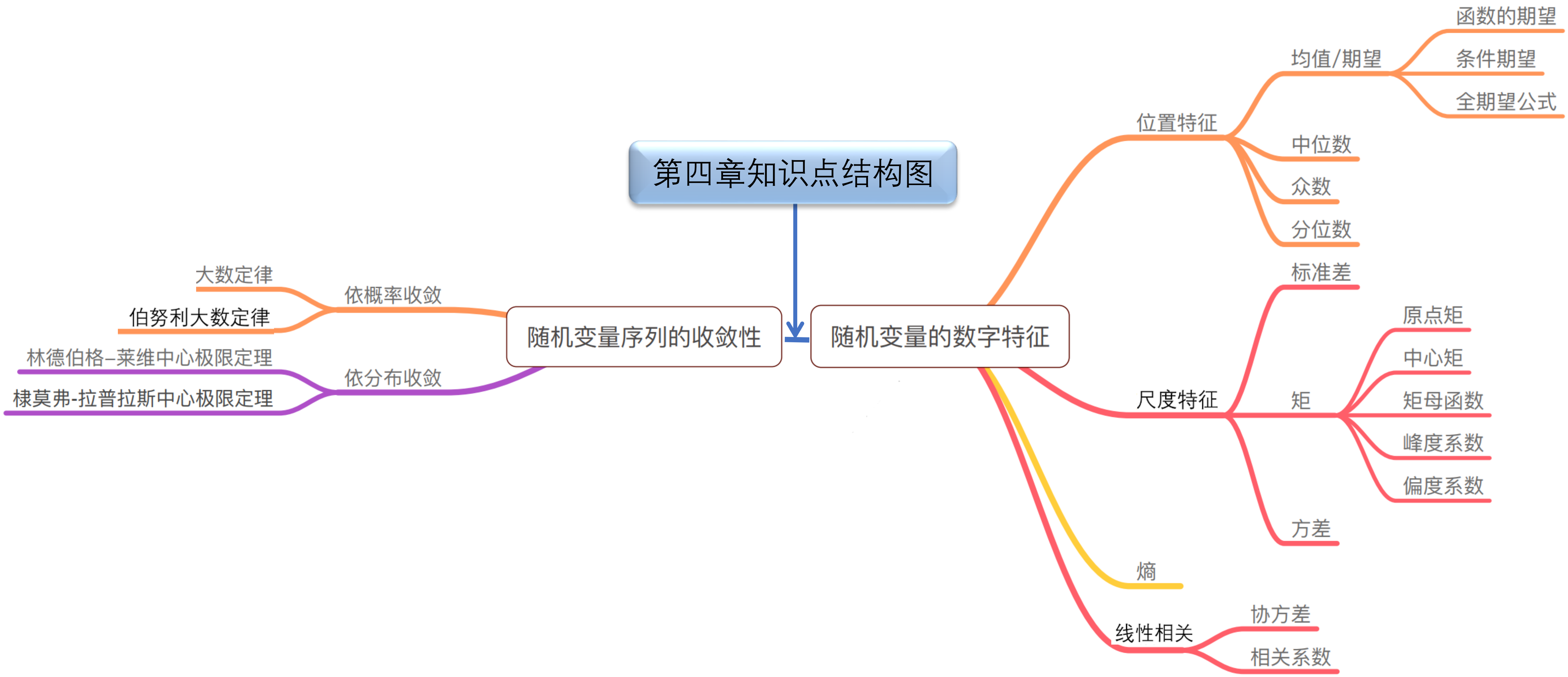
例子

- 某统计学老师教两个班的《概率论与数理统计》，设期末总评的平均分都是80 分, 甲班有 40 名学生, 乙班有 60 名学生, 每个学生成绩的方差都是 25.

  (1) 分别估计两个班平均成绩超过 81 分的概率,

  (2) 估计甲班平均成绩超过乙班 1.7 分的概率,

  (3) 估计乙班平均成绩超过甲班 1.7 分的概率.


第四章知识点结构图
随机变量序列的收敛性
依概率收敛
大数定律
伯努利大数定律
依分布收敛
林德伯格–莱维中心极限定理
棣莫弗-拉普拉斯中心极限定理
随机变量的数字特征
位置特征
均值/期望
函数的期望
条件期望
全期望公式
中位数
众数
分位数
尺度特征
标准差
矩
原点矩
中心矩
矩母函数
峰度系数
偏度系数
方差
熵
线性相关
协方差
相关系数

Thank you!