Homework 8

抹茶奶绿

April 11, 2025

Exercise 1.
判断下述说法正确与否，并尝试给出说明：由概率的频率解释知，当试验次数 $n \to \infty$ 时，事件 A 的频率 $\frac{m}{n}$ 的极限就是概率，即有 $\lim_{n\to\infty} \frac{m}{n} = P(A)$.

Solution.
频率解释是一种主要的概率解释之一，它由大数定律保证. 但并非所有概率问题都适合用这种方法来解释，比如一些一次性的事件或主观概率评估. 这是概率哲学中“客观概率”与“主观概率”的不同观点所反映的内容.

Exercise 2.
证明 $Chebyshev$ 弱大数定律：设随机变量 X_i 两两不相关，且 $\mathrm{Var}(X_i) < c$（c 为常数），那么对于任意 $\epsilon > 0$，

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| \geq \epsilon\right) = 0$$

Solution.
记

$$S_n = \sum_{i=1}^{n} X_i, \quad \bar{X}_n = \frac{S_n}{n}$$

由题设，两两不相关，则

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \leq nc$$

根据 Chebyshev 不等式，有

$$P\left(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon\right) \leq \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} \leq \frac{nc}{n^2\epsilon^2} = \frac{c}{n\epsilon^2}$$

因此

$$0 \leq \lim_{n\to\infty} P\left(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{c}{n\epsilon^2} = 0$$

□

Exercise 3.
利用中心极限定理证明 $Khinchin$ 弱大数定律.

Solution.
设 $\{X_i\}$ 为独立同分布随机变量，均有有限的期望 μ 和有限方差 σ^2. 由 CLT

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

于是

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| \geq \epsilon\right) = 2\left[1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right]$$

当 $n \to \infty$ 时，$\frac{\epsilon\sqrt{n}}{\sigma} \to \infty$，所以

$$1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right) \to 0$$

□

Exercise 4.

设随机变量 $\{X_i,\ i=1,2,\dots,n\}$ 独立同分布，其公共期望为 μ，公共方差为 σ^2，令

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

证明：S^2 依概率收敛至 σ^2.

Solution.

首先

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu) - (\bar{X}_n-\mu)\right]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n-1}(\bar{X}_n-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{n}{n-1}(\bar{X}_n-\mu)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{n-1}(\bar{X}_n-\mu)^2 \end{aligned}$$

再对随机变量列 $\{X_n-\mu\}$ 和 $\{(X_n-\mu)^2\}$ 使用 $Khinchin$ 弱大数定律知

$$\bar{X}-\mu \xrightarrow{P} 0, \quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 \xrightarrow{P} \sigma^2$$

□

Exercise 5.

设随机变量序列 $\{X_n\}$ 与 $\{Y_n\}$ 分别以概率 1 收敛于 a 和 b（其中 $b \neq 0$）. 证明：序列 $\left\{\frac{X_n}{Y_n}\right\}$ 以概率 1 收敛至 $\frac{a}{b}$.

Solution.

由于 $X_n \xrightarrow{a.s.} a$ 和 $Y_n \xrightarrow{a.s.} b$，存在事件 A，$P(A)=1$，且对于 A 中的每个样本点，当 n 足够大时，有 $|X_n-a|<\delta$ 和 $|Y_n-b|<\delta$.

选择 δ 足够小使得 $b-\delta>0$，则在事件 A 上，当 n 足够大时，有 $Y_n > b-\delta > 0$，因此

$$\begin{aligned} \left|\frac{X_n}{Y_n}-\frac{a}{b}\right| &= \left|\frac{b(X_n-a)-a(Y_n-b)}{Y_n b}\right| \\ &\le \frac{|b||X_n-a|+|a||Y_n-b|}{|Y_n||b|} \\ &\le \frac{|b|\delta+|a|\delta}{(b-\delta)|b|} \\ &= \frac{(|b|+|a|)\delta}{(b-\delta)|b|} \end{aligned}$$

当 δ 足够小时，上式可以小于任意给定的 $\varepsilon > 0$. □

Exercise 6.

设独立同分布的随机变量 $\{X_i,\ i=1,2,\dots\}$ 的期望为 2，随机变量 $\{Y_i,\ i=1,2,\dots\}$ 的期望为 5，并且对于任意 n 有 $Y_1+\cdots+Y_n\neq 0$. 证明：

$$\frac{Y_1+\cdots+Y_n}{X_1+\cdots+X_n}\xrightarrow{a.s.}\frac{5}{2}$$

Solution.

由 $Kolmogrov$ 强大数定律，几乎处处有

$$\frac{1}{n}\sum_{i=1}^{n}X_i\to 2,\quad \frac{1}{n}\sum_{i=1}^{n}Y_i\to 5$$

那么由 **Exercise 5.** 可知

$$\frac{Y_1+\cdots+Y_n}{X_1+\cdots+X_n}=\frac{\frac{1}{n}(Y_1+\cdots+Y_n)}{\frac{1}{n}(X_1+\cdots+X_n)}\xrightarrow{a.s.}\frac{5}{2}$$

□

Exercise 7.

设 X 表示抛 40 次均匀硬币出现正面（记为 1）次数，求 $P(X=20)$ 及其正态近似值.

Solution.

由于抛硬币的试验服从二项分布，设 $X\sim \mathrm{B}(n=40,p=0.5)$，则

$$P(X=20)=\binom{40}{20}(0.5)^{40}\approx 0.12537$$

运用连续性校正，有近似

$$P(X=20)\approx P(19.5<X<20.5)\approx \Phi\left(\frac{20.5-20}{\sqrt{10}}\right)-\Phi\left(\frac{19.5-20}{\sqrt{10}}\right)\approx 0.12563$$

Exercise 8.

某保险公司向 10000 个投保人提供内容相同的汽车保险，假定这 10000 个投保人在一年内因交通事故造成的损失是独立同分布的，且一辆车在一年内发生事故的概率为 0.001，事故损失为 1000 元.

I 如果忽略运营成本，保险公司每份保险卖 2 元合理吗？

II 保险公司至少有 20% 的毛利润（毛利润 = 保费 - 保险赔付 - 运营成本（忽略））的概率多大？

III 以 95% 的概率可以保证保险公司至少还有多少毛利润？

Solution.

设每份保险保费为 2 元，共有 10000 份保险. 记每个投保人的赔付随机变量为 X_i，其中

$$P(X_i=1000)=0.001,\quad P(X_i=0)=0.999$$

则

$$E[X_i]=1000\times 0.001=1,\quad \mathrm{Var}(X_i)=1000^2\times 0.001\times 0.999=999.$$

I 保险公司每份保险的平均赔付为 1 元，保费为 2 元，因此其平均毛利润为 1 元，从均值来看收费合理.

II 记总赔付为

$$S = \sum_{i=1}^{10000} X_i$$

保险公司总毛利润为

$$\Pi = 10000 \times 2 - S = 20000 - S$$

若要求毛利润至少为 20%，则要求

$$\Pi \geq 0.2 \times 20000 = 4000$$

即

$$20000 - S \geq 4000 \quad \Rightarrow \quad S \leq 16000.$$

由于 n 较大，可以用正态近似：

$$E[S] = 10000, \quad \mathrm{Var}(S) = 10000 \times 1000 = 10^7, \quad \sigma_S = \sqrt{10^7} \approx 3162.3.$$

因此

$$P(S \leq 16000) \approx \Phi\left(\frac{16000 - 10000}{3162.3}\right) = \Phi\left(\frac{6000}{3162.3}\right) \approx \Phi(1.897) \approx 0.9713$$

III 设 p 为毛利润下限，要求有 95% 的概率满足

$$20000 - S \geq p$$

即

$$P\big(S \leq 20000 - p\big) = 0.95$$

由正态近似，有分位点 $z_{0.95} \approx 1.645$，故

$$20000 - p = E[S] + 1.645\,\sigma_S = 10000 + 1.645 \times 3162.3 \approx 10000 + 5200 = 15200$$

即保险公司有 95% 的概率至少获得 4800 元的毛利润.

Exercise 9.

假设某物理量的真值为 m，多次测量时，每次测量产生一个随机误差，一个合理的假设是：在选择适当的单位下，随机误差服从 $\mathrm{U}(-1,1)$.

I 设进行 n 次独立测量，令算术平均值记为 $\bar{X}_n$，问 $P\big(|\bar{X}_n - m| < \epsilon\big)$ 等于多少？请利用中心极限定理给出当 $n = 25$ 且 $\epsilon = 0.2$ 的概率近似值.

II 为使得 $P\big(|\bar{X}_n - m| < \epsilon\big) > 1 - \alpha$，至少需要进行多少次测量？请利用中心极限定理讨论，并给出当 $\epsilon = 0.2,\ \alpha = 0.05$ 时所需的测量次数.

III 利用 Chebyshev 不等式求解 (2) 问题，并比较两种方法得出的结果.

Solution.

对于 $X_i \sim \mathrm{U}(-1,1)$，有

$$E[X_i] = 0, \quad \mathrm{Var}(X_i) = \frac{(1-(-1))^2}{12} = \frac{4}{12} = \frac{1}{3}$$

记 $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$，则

$$E[\bar{X}_n] = 0, \quad \mathrm{Var}(\bar{X}_n) = \frac{1}{3n}$$

I 由中心极限定理，$\bar{X}_n \approx N(0, 1/(3n))$，故

$$P(|\bar{X}_n - m| < \epsilon) = P(|\bar{X}_n| < \epsilon) \approx 2\Phi\left(\epsilon\sqrt{3n}\right) - 1$$

对于 $n = 25$ 且 $\epsilon = 0.2$，

$$\epsilon\sqrt{3n} = 0.2\sqrt{75} \approx 0.2 \times 8.66 \approx 1.732$$

因此

$$P(|\bar{X}_{25}| < 0.2) \approx 2 \times 0.9582 - 1 \approx 0.9164$$

II 要满足

$$P(|\bar{X}_n| < \epsilon) = 2\Phi(\epsilon\sqrt{3n}) - 1 \geq 1 - \alpha$$

即要求

$$\Phi(\epsilon\sqrt{3n}) \geq 1 - \frac{\alpha}{2}$$

令 $z_{1-\alpha/2}$ 为标准正态分布上分位点，则需有

$$\epsilon\sqrt{3n} \geq z_{1-\alpha/2} \quad \Longrightarrow \quad n \geq \frac{z_{1-\alpha/2}^2}{3\epsilon^2}$$

对于 $\epsilon = 0.2,\ \alpha = 0.05$，有 $z_{0.975} \approx 1.96$，故

$$n \geq \frac{(1.96)^2}{3 \times 0.04} = \frac{3.8416}{0.12} \approx 32.01$$

即至少需要 33 次测量.

III 利用 Chebyshev 不等式，

$$P(|\bar{X}_n| < \epsilon) \geq 1 - \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = 1 - \frac{1}{3n\epsilon^2}$$

要求

$$1 - \frac{1}{3n\epsilon^2} \geq 1 - \alpha \quad \Longrightarrow \quad n \geq \frac{1}{3\alpha\epsilon^2}$$

对于 $\epsilon = 0.2,\ \alpha = 0.05$ 得

$$n \geq \frac{1}{3 \times 0.05 \times 0.04} = \frac{1}{0.006} = 166.67$$

即需要至少 167 次测量.

比较：中心极限定理给出的结果 (33 次) 要远小于 *Chebyshev* 不等式给出的保守结果 (167 次)，这反映了 *Chebyshev* 不等式对分布信息没有利用，从而结果较为宽松，而中心极限定理利用了误差分布近似正态的性质，因此给出了更精确的估计.

Exercise 10.

假设某批次电子元件中合格品占 80%，从中任意购买 5000 个. 试问把误差限 ϵ 定为多少才能保证买到的元件的合格率与该批次的合格率之差不超过 ϵ 的概率至少为 0.99？此时的合格品数的范围是多少？

Solution.

令 $X \sim \mathrm{B}(n = 5000, p = 0.8)$，则

$$E[X] = 5000 \times 0.8 = 4000, \quad \mathrm{Var}(X) = 5000 \times 0.8 \times 0.2 = 800$$

注意合格率为 $X/5000$，其均值为 0.8，方差为 $\mathrm{Var}(X/5000) = \dfrac{800}{5000^2} \approx 0.000032$

要求

$$P\left(\left|\frac{X}{5000} - 0.8\right| < \epsilon\right) \geq 0.99$$

利用正态近似，

$$\frac{X}{5000} \sim N\Big(0.8, \frac{800}{5000^2}\Big)$$

标准差约为

$$\sigma_{率} = \frac{\sqrt{800}}{5000} \approx \frac{28.28}{5000} \approx 0.00566$$

令 $z_{0.995} \approx 2.5758$（因为两侧共 1%），则要求

$$\epsilon \geq 2.5758 \times 0.00566 \approx 0.0146$$

即当误差限取约 0.0146 时，

$$P\Big(\Big|\frac{X}{5000} - 0.8\Big| < 0.0146\Big) \geq 0.99$$

对应的合格品数范围为

$$5000 \times (0.8 - 0.0146) \leq X \leq 5000 \times (0.8 + 0.0146)$$

即大约

$$3927 \leq X \leq 4073$$

Exercise 11.
尝试找出一个公开报道的抽样调查实例，指出其是否与课上选举问题的讨论相符，并简要说明理由.

Solution.

Exercise 12.
假设某只股票的价格每天“上涨 70%”或“下降 50%”的概率均为 0.5，并且不同日子之间相互独立. 设初始股价为 $Y_0 = 1$，记 Y_n 为经过 n 天后的股价.

I 验证 $\ln Y_n$ 近似正态分布，并给出正态分布的参数.

II 当 $n \to \infty$ 时，$E(Y_n)$ 会怎样变化？

III 证明：$P(\lim_{n\to\infty} Y_n = 0) = 1$.

IV 尝试给前两问的结果一个直观解释.

Solution.

I 令每天的价格变动因子 R_i 定义为

$$R_i = \begin{cases} 1.7, & p = 0.5, \\ 0.5, & 1 - p = 0.5 \end{cases}$$

那么

$$Y_n = \prod_{i=1}^{n} R_i, \quad \ln Y_n = \sum_{i=1}^{n} \ln R_i$$

又

$$E[\ln R_i] = -0.08125, \quad \mathrm{Var}(\ln R_i) \approx 0.3744$$

由中心极限定理，当 n 较大时，

$$\ln Y_n \approx N\Big(-0.08125\, n,\ 0.3744\, n\Big).$$

II
$$E(Y_n) = Y_0\Big[E(R_1)\Big]^n, \quad E(R_1) = 0.5(1.7) + 0.5(0.5) = 1.1$$
故
$$E(Y_n) = (1.1)^n$$
因此 $E(Y_n) \to \infty$.

III 由 $Kolmogrov$ 强大数定律，几乎处处有
$$\frac{1}{n}\ln Y_n \to -0.08125$$
从而
$$Y_n^{1/n} \to e^{-0.08125}$$
因此
$$Y_n \to 0 \quad a.s.$$
这反映了股票价格的分布极右尾稀有极值对期望的拉高作用，而大多数路径的实际表现远低于均值.

IV 直观解释：$E(Y_n)$ 趋于无穷大，是由于极少数情形下股票可能出现巨额上涨，从而抬高了总体均值；而第三问的结果表明，对于" 绝大多数" 情形，股票价格实际上远低于 $E(Y_n)$，即样本路径普遍较低. 这揭示了" 均值-极端值悖论"：期望值受到少数极端事件的影响，但在概率意义下，几乎所有结果却远低于均值.

Exercise 13.

设 X 的累积分布函数 $F(x)$ 未知，样本 $\{X_i,\ i = 1, \dots, n\}$ 来自该总体，相互独立，其经验分布函数定义为
$$F_n(x) = \frac{1}{n}\#\{i:\ X_i \le x\}.$$

I 求 $F_n(x)$ 的期望和方差.

II 证明：$F_n(x)$ 依概率 1 收敛于 $F(x)$.

Solution.

I 对任意固定 x，记示性函数
$$I_i(x) = \begin{cases} 1, & X_i \le x, \\ 0, & X_i > x. \end{cases}$$
则 $I_i(x)$ 为 $Bernoulli$ 随机变量，其参数为 $p = F(x)$. 故
$$E[I_i(x)] = F(x), \quad \text{Var}(I_i(x)) = F(x)(1 - F(x))$$
由独立性，
$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I_i(x)$$
因此
$$E[F_n(x)] = F(x)$$
$$\text{Var}(F_n(x)) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(I_i(x)) = \frac{F(x)(1 - F(x))}{n}$$

II 由 $Kolmogrov$ 强大数定律
$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I_i(x) \xrightarrow{a.s.} E[I_i(x)] = F(x)$$

Exercise 14.

设随机变量 $\{X_i,\ i=1,\dots,n\}$ 独立同分布，记

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

I 当 $X_i\sim N(1,0)$ 时，请分别对 $n=1,\ 25,\ 100,\ 1000$ 模拟 $\bar{X}$ 的分布；

II 当 $X_i\sim U(0,1)$ 时，请分别对 $n=1,\ 25,\ 100,\ 1000$ 模拟 $\bar{X}$ 的分布；

III 当 X_i 的概率密度函数为

$$f(x)=\frac{1}{\pi}\frac{1}{1+x^2},\quad x\in\mathbb{R}$$

请分别对 $n=1,\ 25,\ 100,\ 1000$ 模拟 $\bar{X}$ 的分布，并观察以上实验结果是否与中心极限定理相符，说明原因.

Solution.

代码如下：

```python
import random
from matplotlib import pyplot as plt
import numpy as np

def normal(n):
    return [random.normalvariate(0, 1) for ii in range(0, n)]

def uniform(n):
    return [random.uniform(0, 1) for ii in range(0, n)]

def cauchy(n):
    return np.random.standard_cauchy(n)

def graph(src, n, sample):
    expectancy = [sum(src(n)) / n for ii in range(0, sample)]
    plt.figure(dpi=300)
    plt.hist(expectancy, bins=25)
    plt.title(f"x~{src.__name__} distribution, n={n}")
    plt.savefig(f"{src.__name__}_{n}.jpg")

spl = 5000
for i in [1, 25, 100, 1000]:
    graph(normal, i, spl)
    graph(uniform, i, spl)
    graph(cauchy, i, spl)
```

对于前两个，可以看出 n 越大，$\bar{X}$ 的分布越接近正态分布，符合中心极限定理的要求；而对于柯西分布，由于其期望，方差均不存在，因此并不符合中心极限定理.

另外绘图如下：

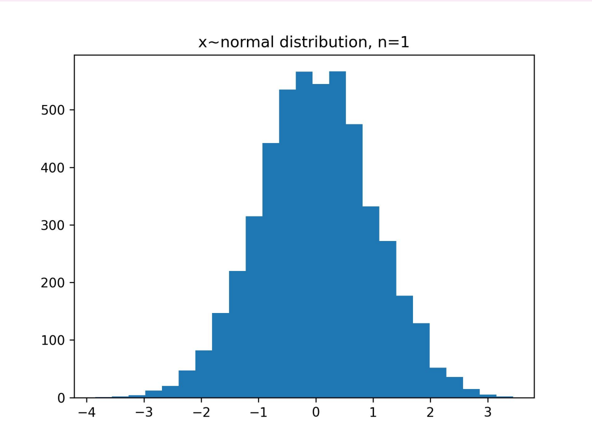

Figure 1: 正态分布，$n = 1$

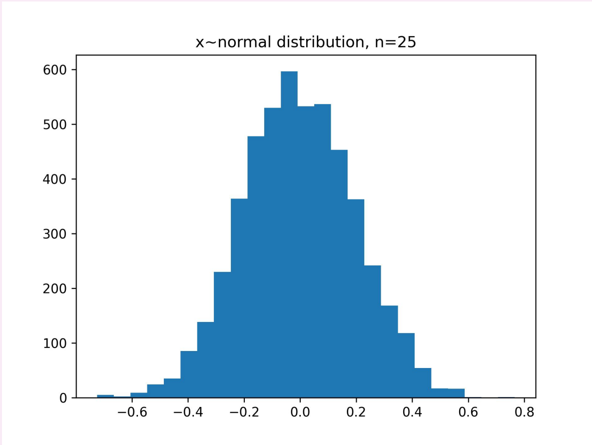

Figure 2: 正态分布，$n = 25$

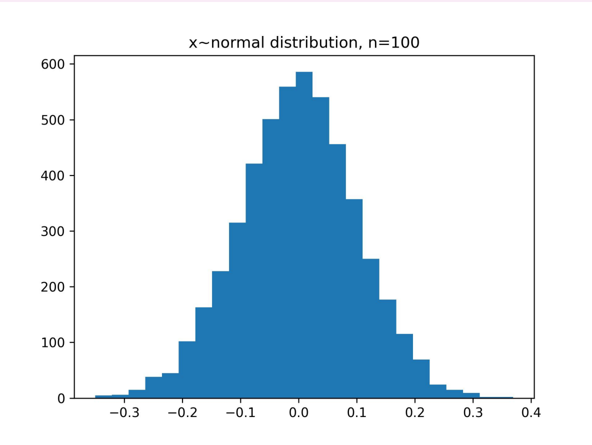

Figure 3: 正态分布，$n = 100$

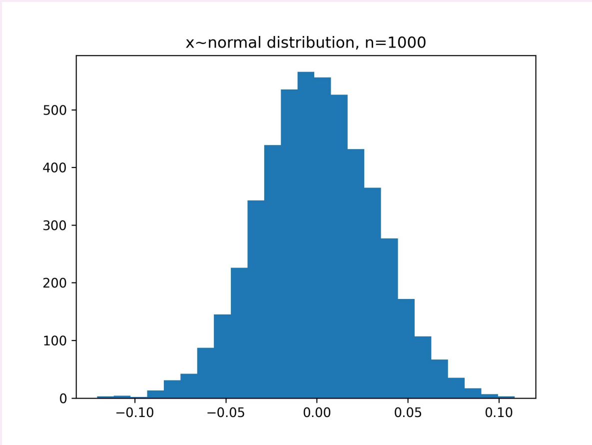

Figure 4: 正态分布，$n = 1000$

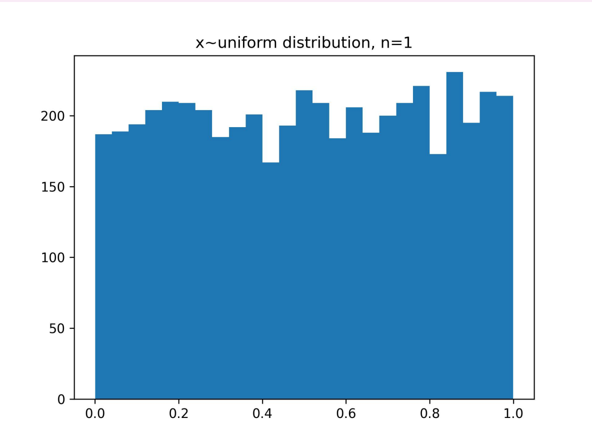

Figure 5: 均匀分布，$n = 1$

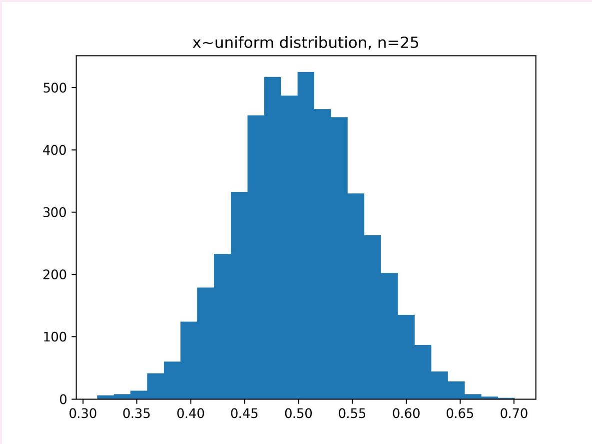

Figure 6: 均匀分布，$n = 25$

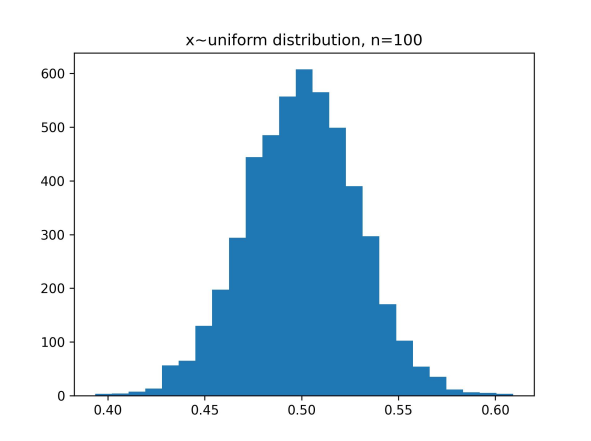

Figure 7: 均匀分布，$n = 100$

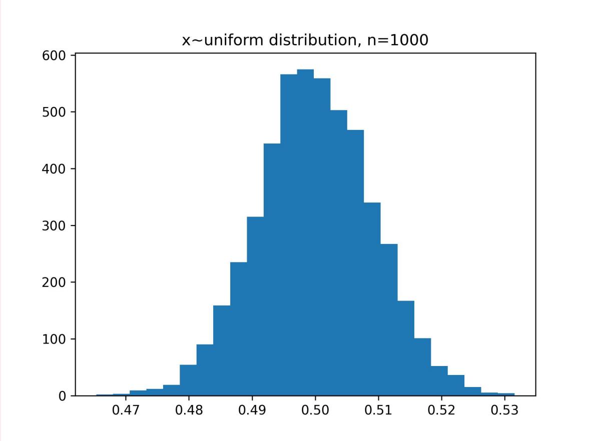

Figure 8: 均匀分布，$n = 1000$

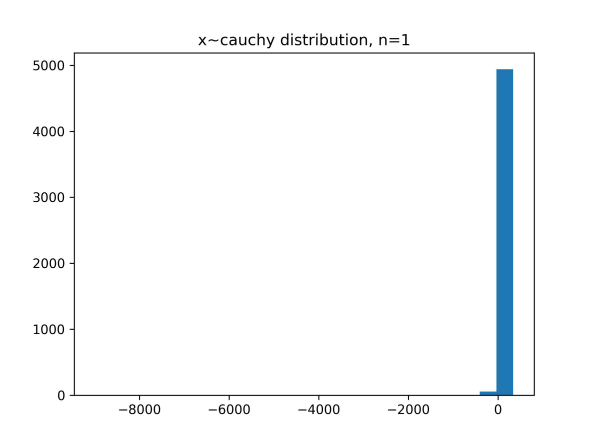

Figure 9: 柯西分布，$n = 1$

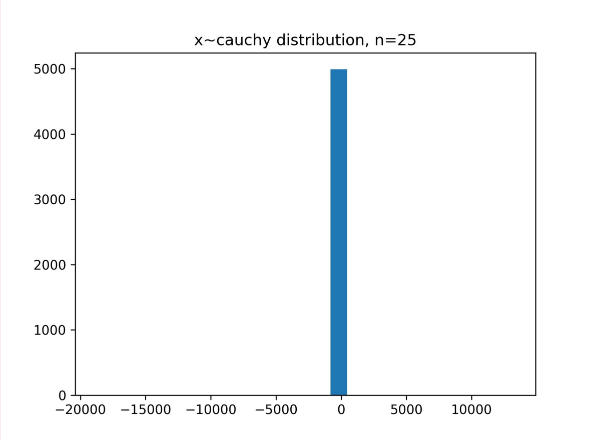

Figure 10: 柯西分布，$n = 25$

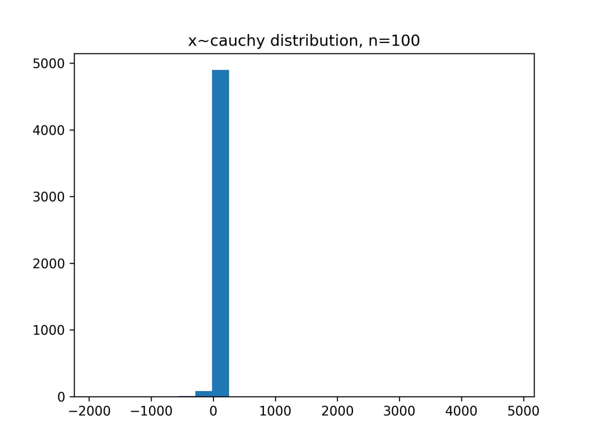

Figure 11: 柯西分布，$n = 100$

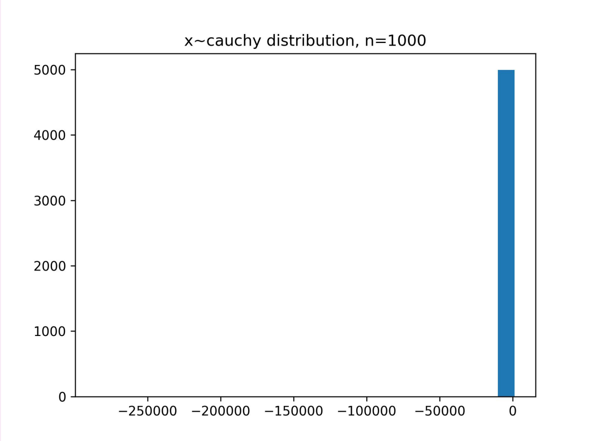

Figure 12: 柯西分布，$n = 1000$