

12: 验证中心极限定理

许传奇 PB16021546

1 题目

自设若干个随机分布（相同或不同分布，它们有相同或不同的 μ 和 σ^2 ），通过Monte Carlo模拟，验证中心极限定理成立（N=2、5、10）。

2 原理与算法

2.1 原理

2.1.1 中心极限定理与误差

概率论中的大数法则和中心极限定理是Monte-Carlo方法应用于统计计算的基础。

1. 大数法则：

如随机量序列 f_i 有期待值 μ 存在，则：

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N f_i \rightarrow \mu \quad (1)$$

2. 中心极限定理：

当N有限时，平均值 $\langle f \rangle \approx \frac{1}{N} \sum_{i=1}^N f(x_i)$ 式满足：

$$P \left\{ \left| \frac{\langle f \rangle - \mu}{\sigma_f / \sqrt{N}} \right| < \beta \right\} \rightarrow \Phi(\beta) \quad (2)$$

其中的 $\Phi(\beta)$ 是Gauss正态分布，因此可得：

$$\sigma_s = |\langle f \rangle - \mu| \propto \frac{\sigma_f}{\sqrt{N}} = \frac{1}{\sqrt{N}} \sqrt{\langle f^2 \rangle - \langle f \rangle^2} \quad (3)$$

2.2 算法

由原理部分的叙述可以，运用中心极限定理，可以将任意分布的独立同分布的变量转化成标准正态分布，即：

$$P \left\{ \left| \frac{\langle X \rangle - \mu}{\sigma_X / \sqrt{n}} \right| < \beta \right\} \rightarrow \Phi(\beta) \quad (4)$$

可以在分布 $f(x)$ （期望为 μ ，方差为 σ_f ）下抽取 N 个独立同分布变量 X_1, X_2, \dots, X_N ，设变量 X 为：

$$X = \frac{1}{N} \sum_{i=1}^N X_i \quad (5)$$

由概率论与数理统计的知识可以求得 X 的期望、方差和标准差：

$$\begin{aligned} EX &= E\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N} \sum_{i=1}^N EX_i = \frac{1}{N} \sum_{i=1}^N \mu = \mu \\ \text{Var}(X) &= \text{Var}\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(X_i) = \frac{1}{N^2} \sum_{i=1}^N \sigma_f^2 = \frac{\sigma_f^2}{N} \\ \sigma_X &= \sqrt{\text{Var}X} = \frac{\sigma_f}{\sqrt{N}} \end{aligned} \quad (6)$$

因此， X 在中心极限定理的条件下，可以转化成标准正态分布：

$$\frac{\langle X \rangle - \mu}{\sigma_X / \sqrt{N}} \sim \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (7)$$

因此，我们分别抽取 Num 个 X_1, X_2, \dots, X_N ，验证对应的 Num 个 $X = \frac{1}{N} \sum_{i=1}^N X_i$ 和特定分布的 μ 和 σ 是否符合标准正态分布。

如果上述关系成立，则验证了中心极限定理；否则，验证不了中心极限定理。

3 源文件使用说明

编译并运行“12Central.Limit.Theorem.cpp”，程序运行，各种分布得到的数据输出到文件中。

编译并运行“plot.py”即可绘制出直方图。

4 计算结果及具体分析

总的随机点数为100000，16807随机数生成器的种子值由C语言自带的随机数生成。

4.1 离散随机变量

4.1.1 0-1分布

0-1分布为：

$$\begin{aligned} P(X = k) &= p^k (1 - p)^{1-k}, \quad k = 0, 1 \\ EX &= p, \quad \text{Var}(X) = p(1 - p) \end{aligned} \quad (8)$$

不失一般性，取 $p = 0.5$ ，则 $EX = 0.5$ ， $\text{Var}(X) = 0.25$ 。

得到的结果如下所示：

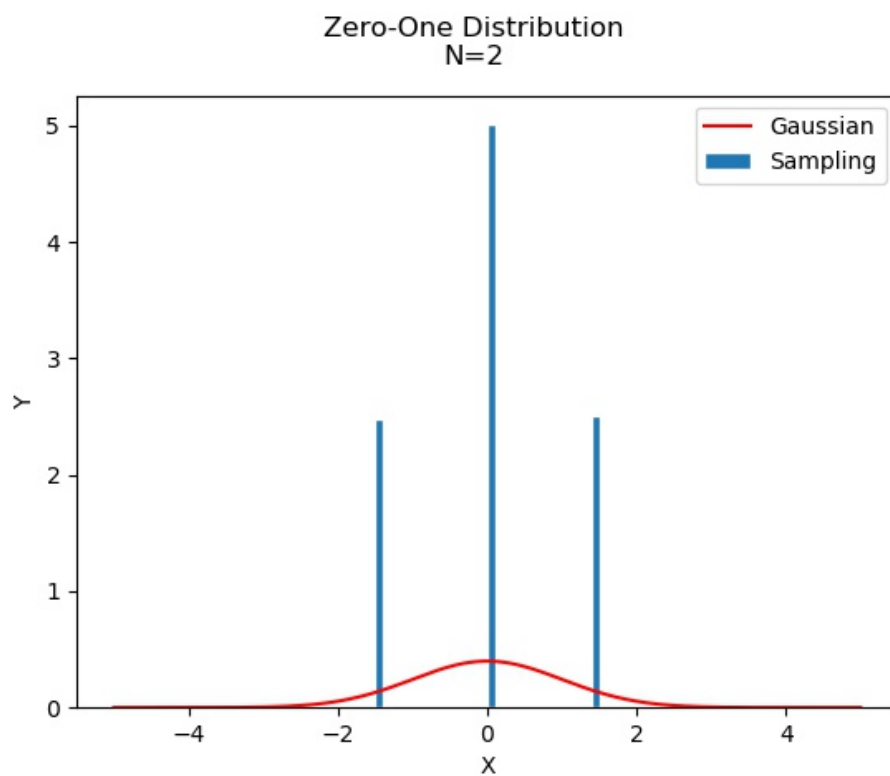


图 1: N=2时0-1分布的结果

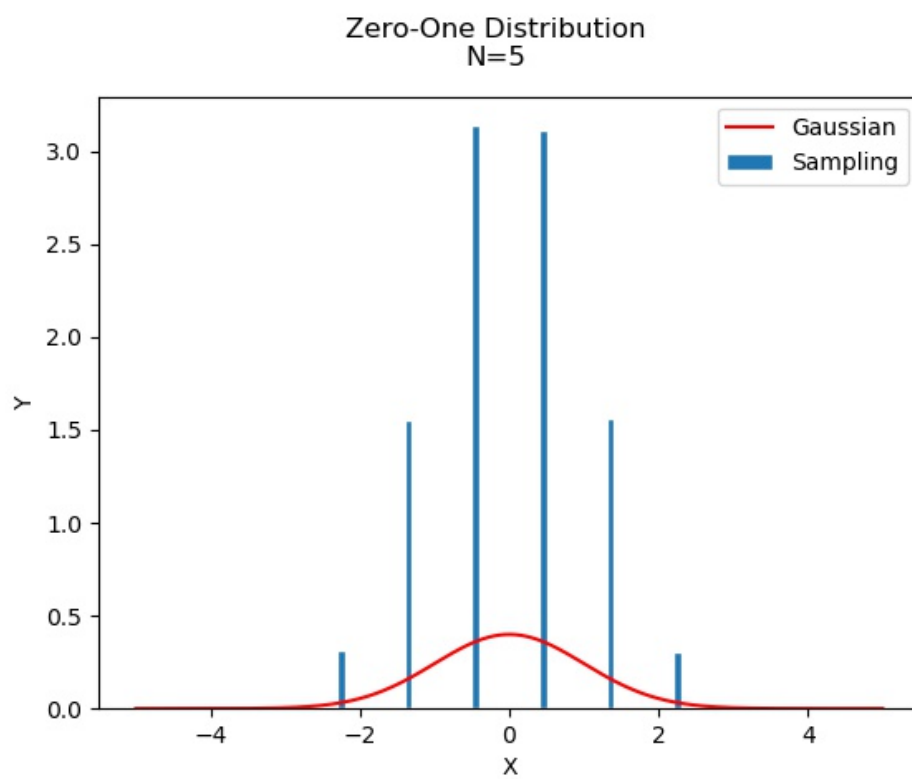


图 2: N=5时0-1分布的结果

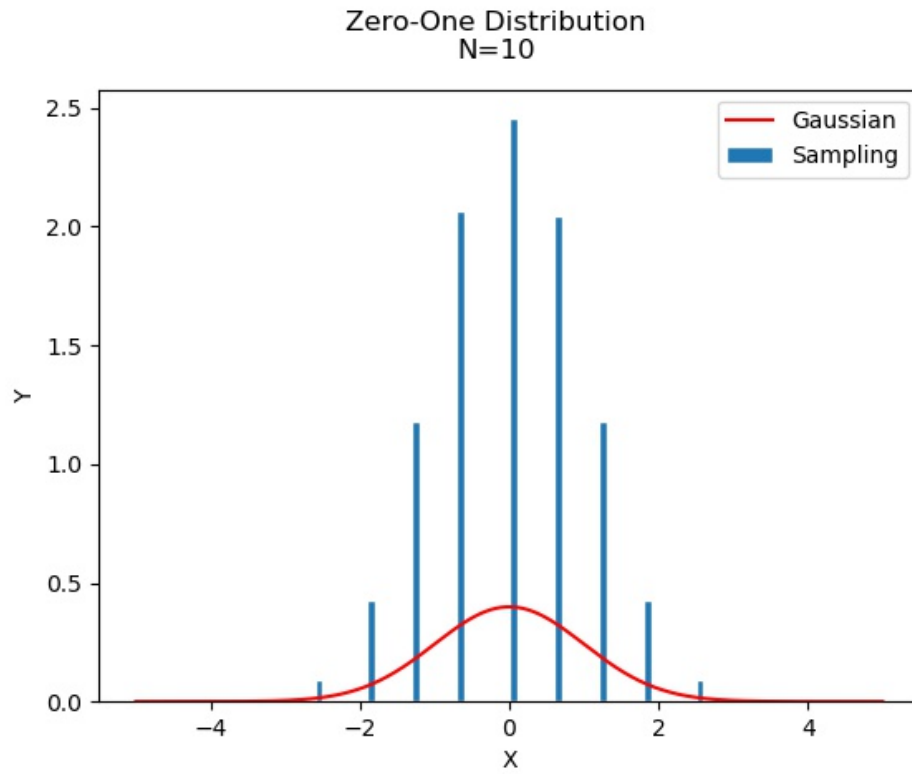


图 3: N=10时0-1分布的结果

4.1.2 二项分布

二项分布为:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n \quad (9)$$

$$EX = np, \quad \text{Var}(X) = np(q-p)$$

不是一般性, 取 $n = 5, p = 0.5$, 则 $EX = 2.5, \text{Var}(X) = 1.25$ 。

得到的结果如下所示:

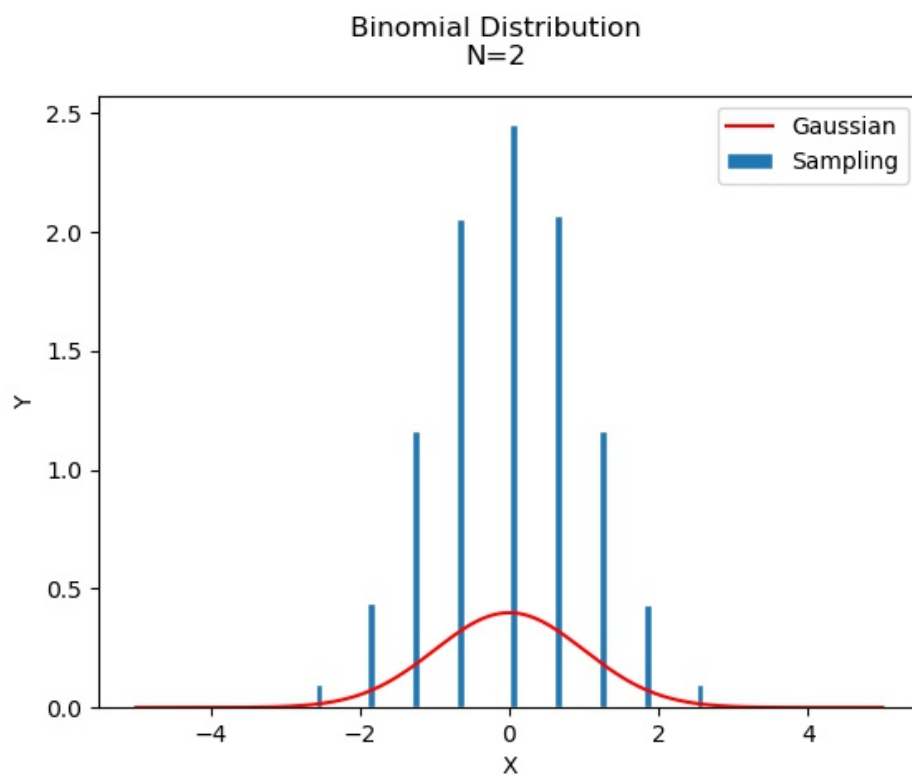


图 4: N=2时二项分布的结果

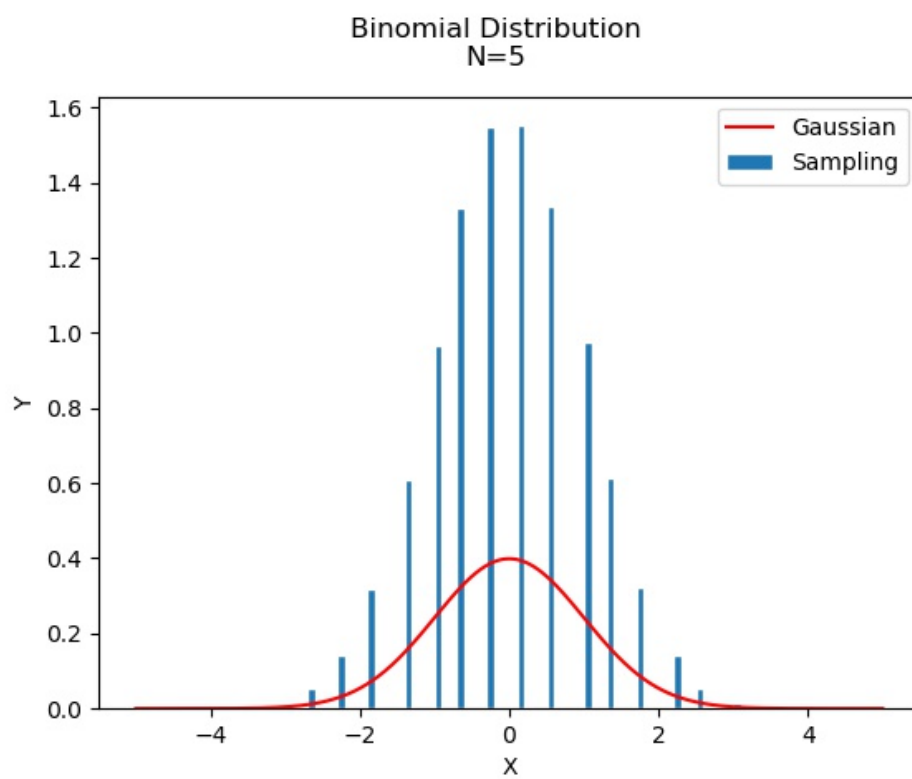


图 5: N=5时二项分布的结果

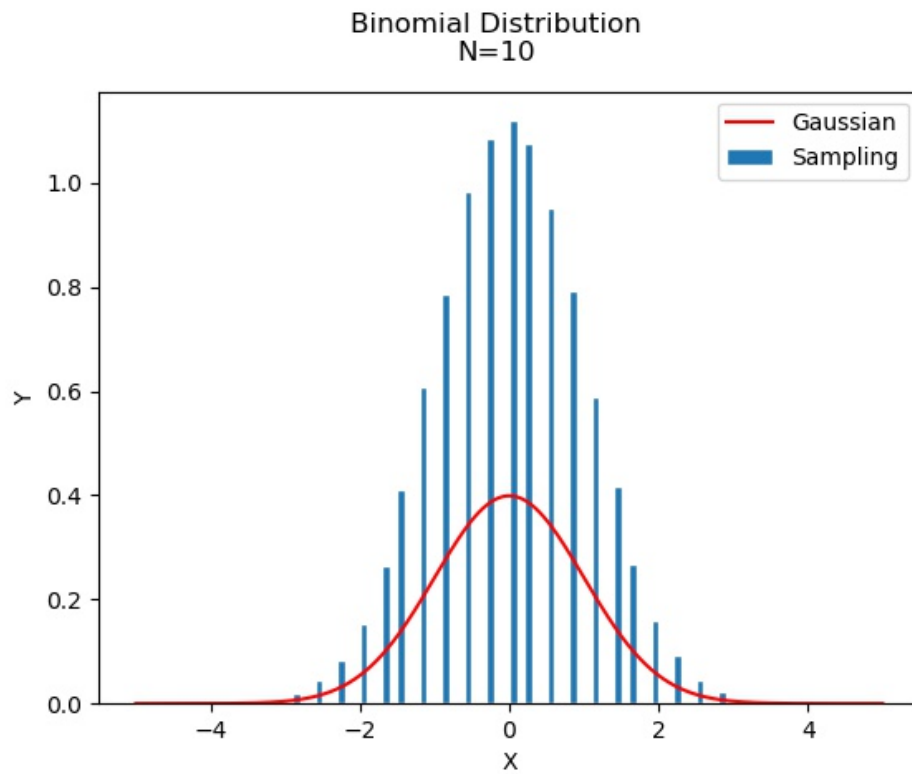


图 6: N=10时二项分布的结果

4.2 连续随机变量

4.2.1 均匀分布

均匀分布为:

$$f(x) = \frac{1}{b-a}, a < x < b$$

$$EX = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12} \quad (10)$$

不失一般性, 取 $a=0, b=2$, 则 $EX=1, Var(x)=\frac{1}{3}$ 。

得到的结果如下所示:

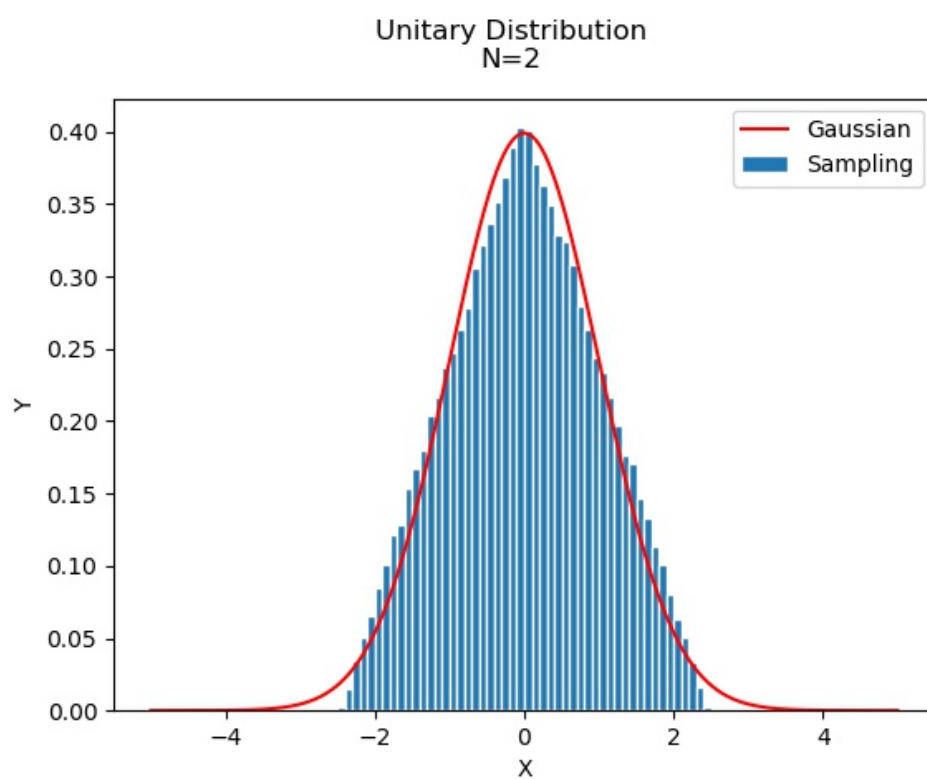


图 7: N=2时均匀分布的结果

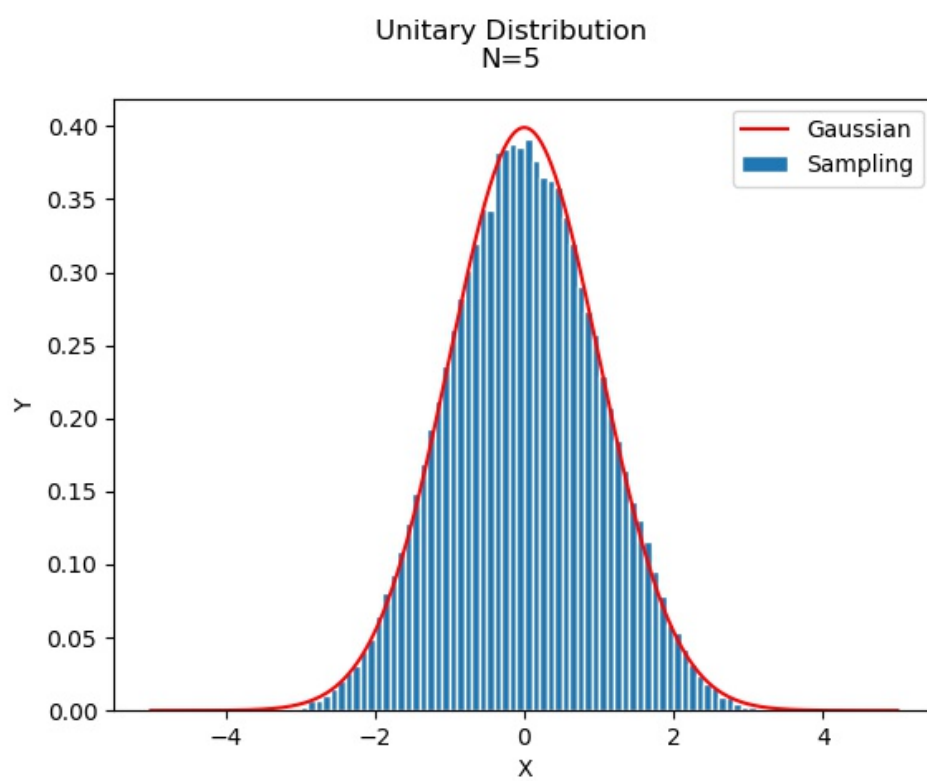


图 8: N=5时均匀分布的结果

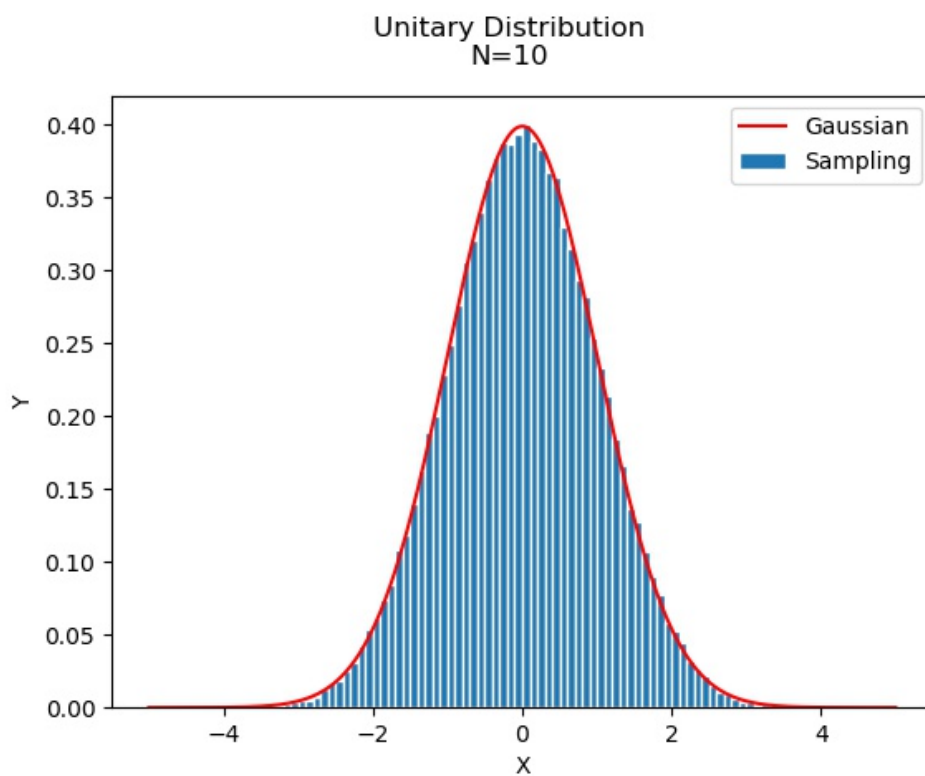


图 9: N=10时均匀分布的结果

4.2.2 指数分布

指数分布为:

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x}, x > 0 \\ EX &= \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2} \end{aligned} \quad (11)$$

不失一般性, 取 $\lambda = 1$, 则 $EX = 1, \text{Var}(X) = 1$ 。

得到的结果如下所示:

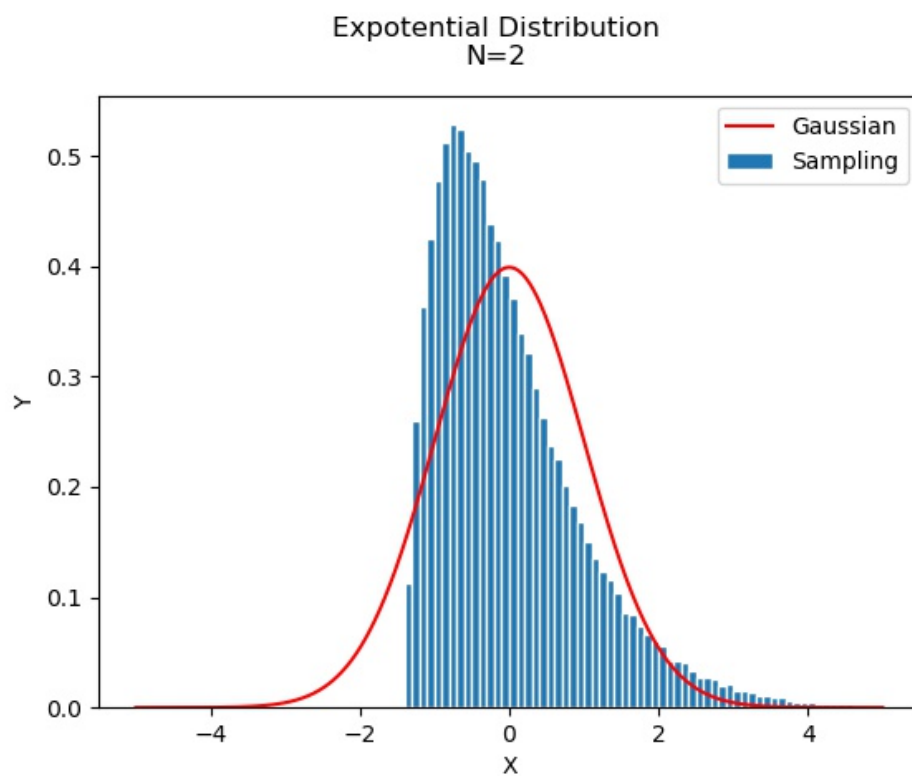


图 10: N=2时指数分布的结果

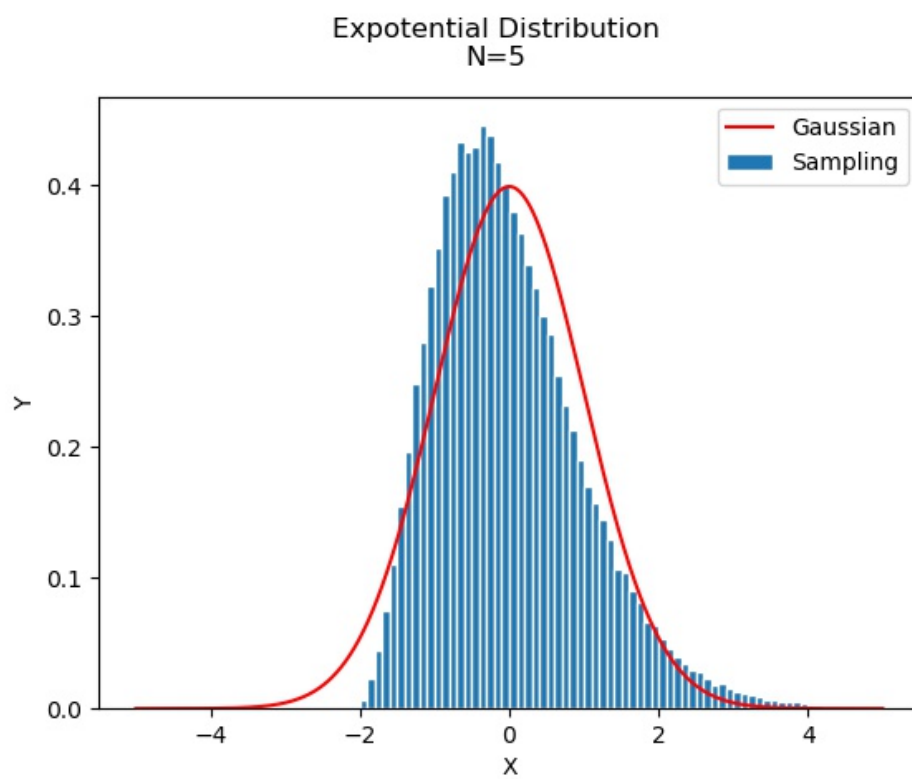


图 11: N=5时指数分布的结果

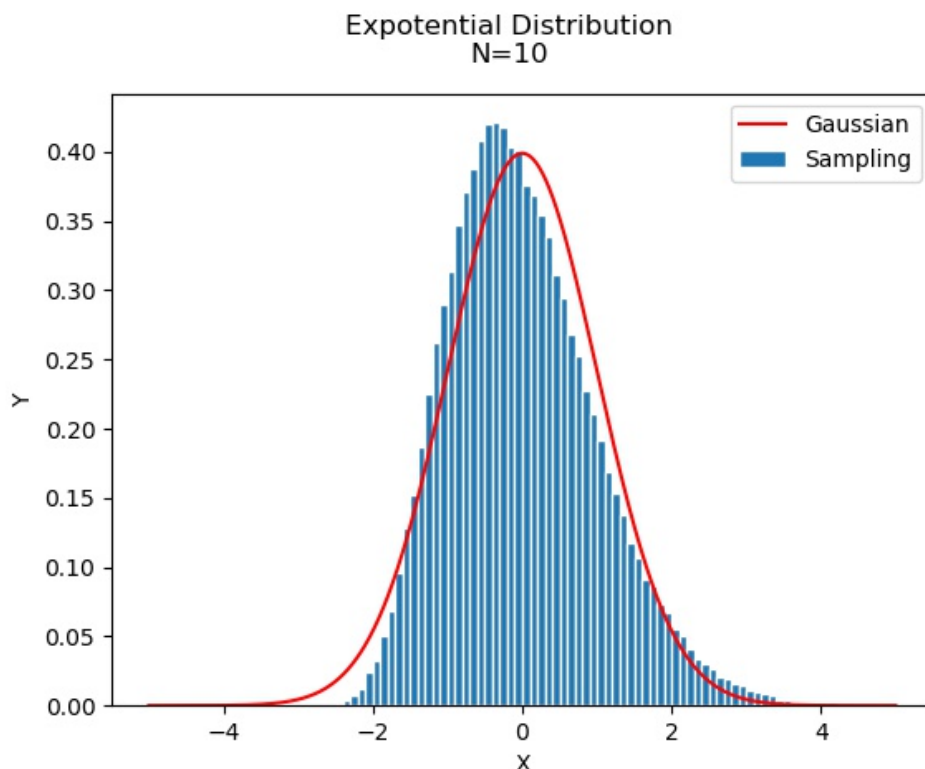


图 12: N=10时指数分布的结果

5 讨论

5.1 原理的讨论

对于中心极限定理：

$$P \left\{ \left| \frac{\langle f \rangle - \mu}{\sigma_f / \sqrt{N}} \right| < \beta \right\} \rightarrow \Phi(\beta) \quad (12)$$

要求N趋近于无穷时，才是标准正态分布。

与我做这题之前的判断不同，即便题中的N取值并不大，对于连续分布的函数也十分符合。说明一般情况下使用中心极限定理的时候，并不需要N取值非常大也可以有比较好的结果。

这也解答了我在概统学习中的一个困惑。概统中的题目要求使用中心极限定理来通过样本估计分布的期望，但很多题目中样本给的数据点并不大。曾经我以为是题目出的缺陷，但其实在N很小的时候也近似满足，甚至N是个位数的时候都有比较好的符合了。

5.2 结果的讨论

从结果的直方图来看，尽管每次抽样得到的分布不同，但最后的X都近似满足标准正态分布。

因为中心极限定理要求抽取的随机数点为无穷大，但实际上达不到，因此并不能完全符合标准正态分布。但当点数很多时，已经很接近标准正态分布了。

1. 从结果中也可以看出，对于离散分布的函数，拟合的不是很好。

比如0-1分布， X 仅能取几个值，而高斯分布是离散的。但当 N 增大的时候， X 能取的值也变多，从图中也可以看出， N 增大与标准正态分布也越来越靠近。

二项分布则同理，在 N 为有限值的时候也只能取有限个值，当 N 增大时，与标准正态分布越来越相似。

2. 对于连续分布的函数，则没有离散分布的函数那样的取值的限制，但仍然随着 N 的增大，与标准正态分布越来越相似。

从图中可以看出，均匀分布的 N 较小时，分布的图像类似于三角形，在标准正态分布的两个尾部基本上没有分布。但当 N 增大时，这里的分布逐渐被填满，而超过标准正态分布的部分则减少，这是为了满足归一性，与标准正态分布越来越相似。

指数分布也是一样。在 N 消失，指数分布与标准正态分布差别很大，但也可以看出随着 N 的增大，与标准正态分布越来越类似。