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：

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}f_i\to\mu \tag{1}$$

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：

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

   其中的$\Phi(\beta)$是Gauss正态分布，因此可得：

$$\sigma_S=|\langle f\rangle-\mu|\propto\frac{\sigma_f}{\sqrt{N}}=\frac{1}{\sqrt{N}}\sqrt{\langle f^2\rangle-\langle f\rangle^2} \tag{3}$$

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：

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

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, $X_2, \ldots$, X_N，设变量X为：

$$X = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{5}$$

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：

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

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：

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

因此，我们分别抽取Num个X_1, $X_2, \ldots$, 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$\mu$和$\sigma$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：

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

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

得到的结果如下所示：

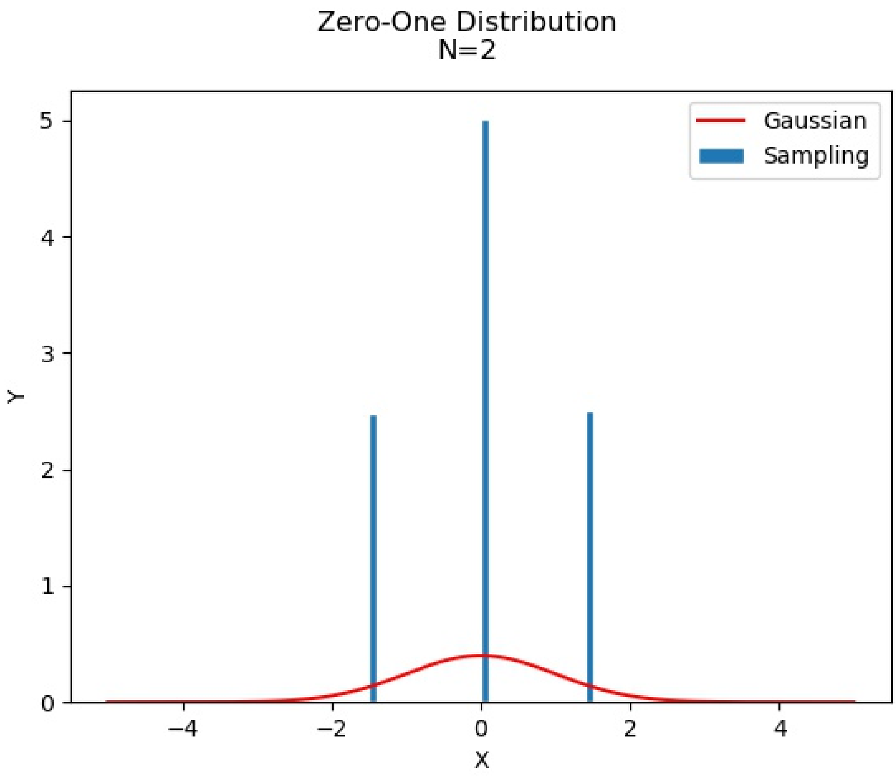

图 1: N=2时0-1分布的结果

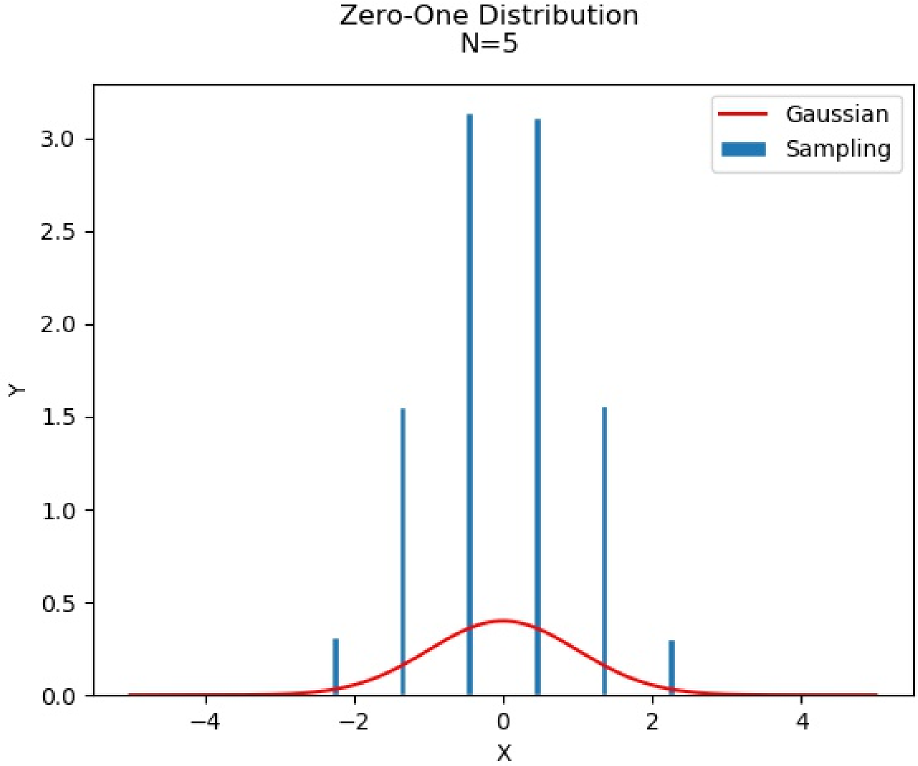

图 2: N=5时0-1分布的结果

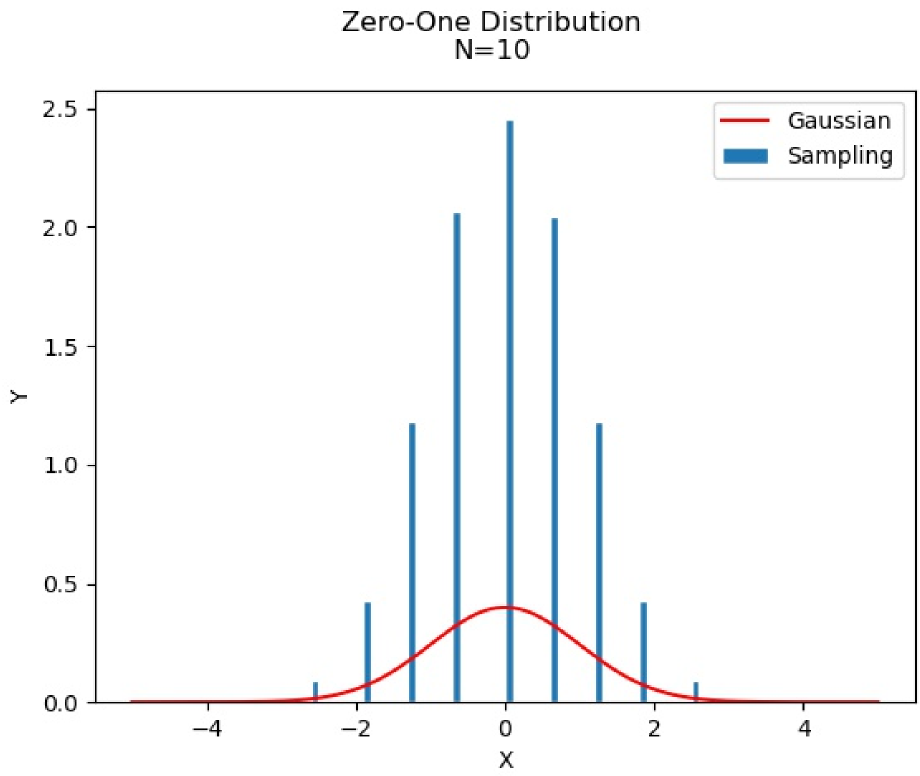

图 3: N=10时0-1分布的结果

4.1.2 二项分布

二项分布为：

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,1,\dots,n \\ EX=np,\ Var(X)=np(q-p) \tag{9}$$

不是一般性，取$n=5,p=0.5$，则$EX=2.5,Var(X)=1.25$。
得到的结果如下所示：

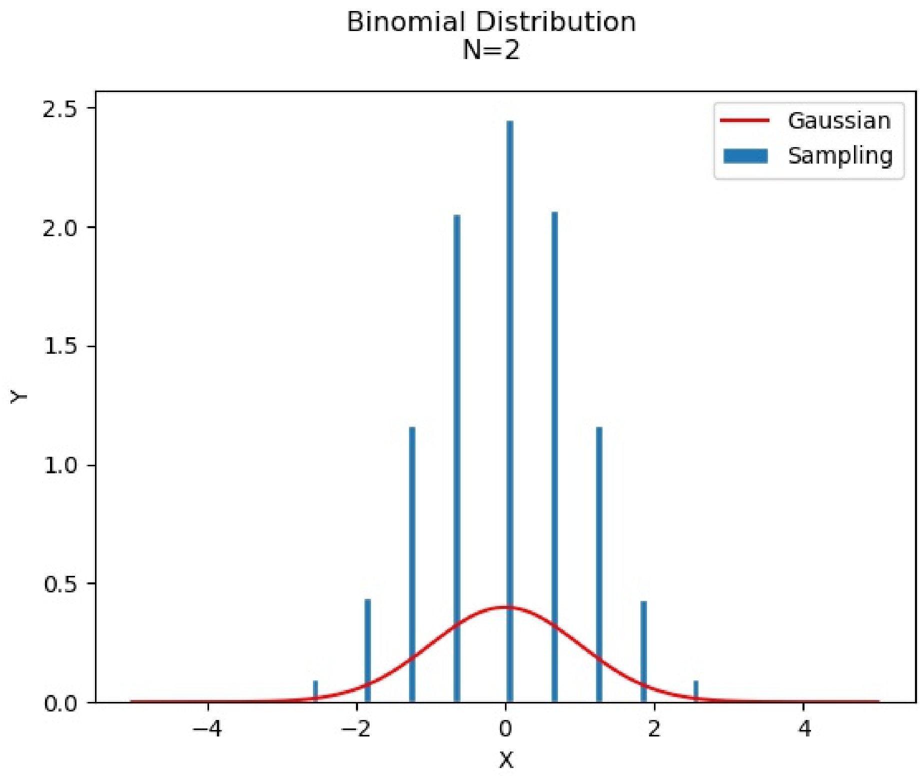

图 4: N=2时二项分布的结果

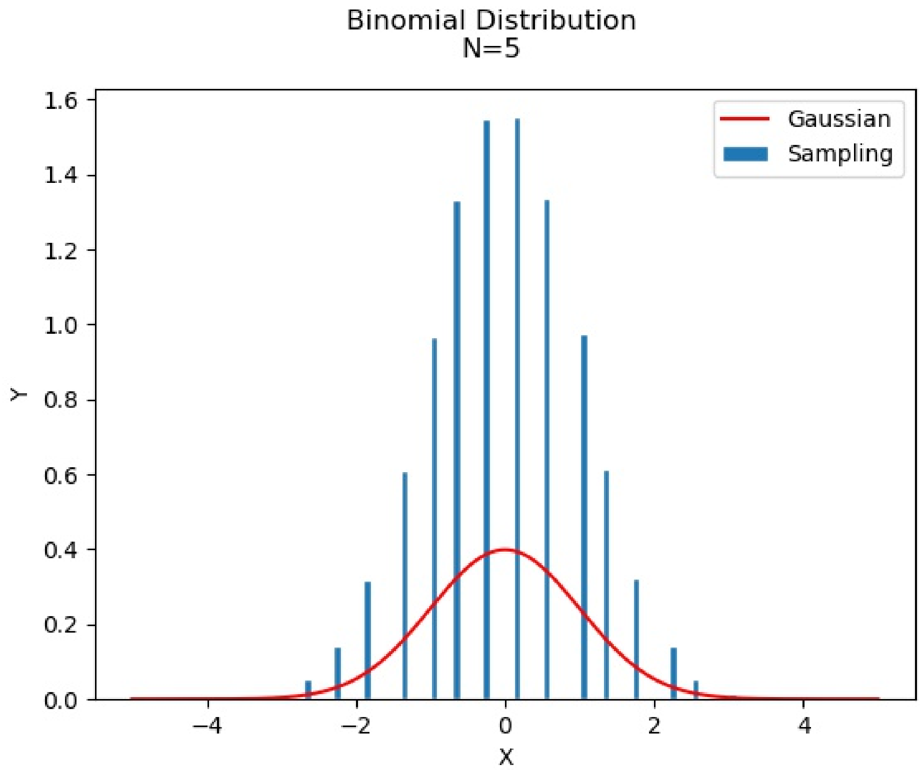

图 5: N=5时二项分布的结果

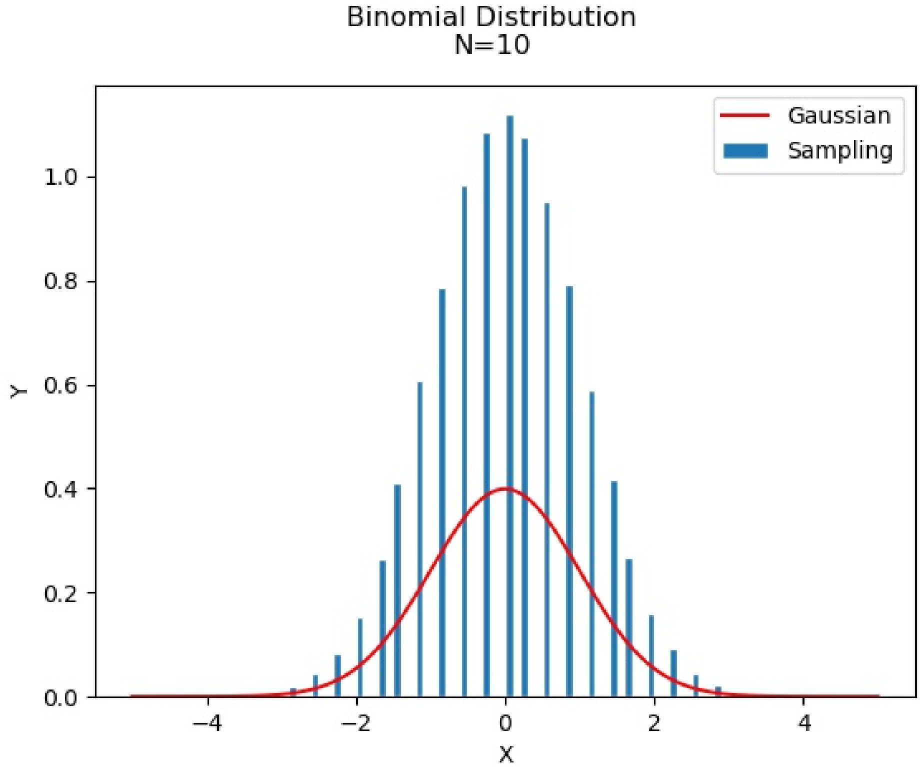

图 6: N=10时二项分布的结果

4.2 连续随机变量

4.2.1 均匀分布

均匀分布为：

$$f(x) = \frac{1}{b-a},\ a < x < b$$
$$EX = \frac{a+b}{2},\ Var(X) = \frac{(b-a)^2}{12} \tag{10}$$

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：

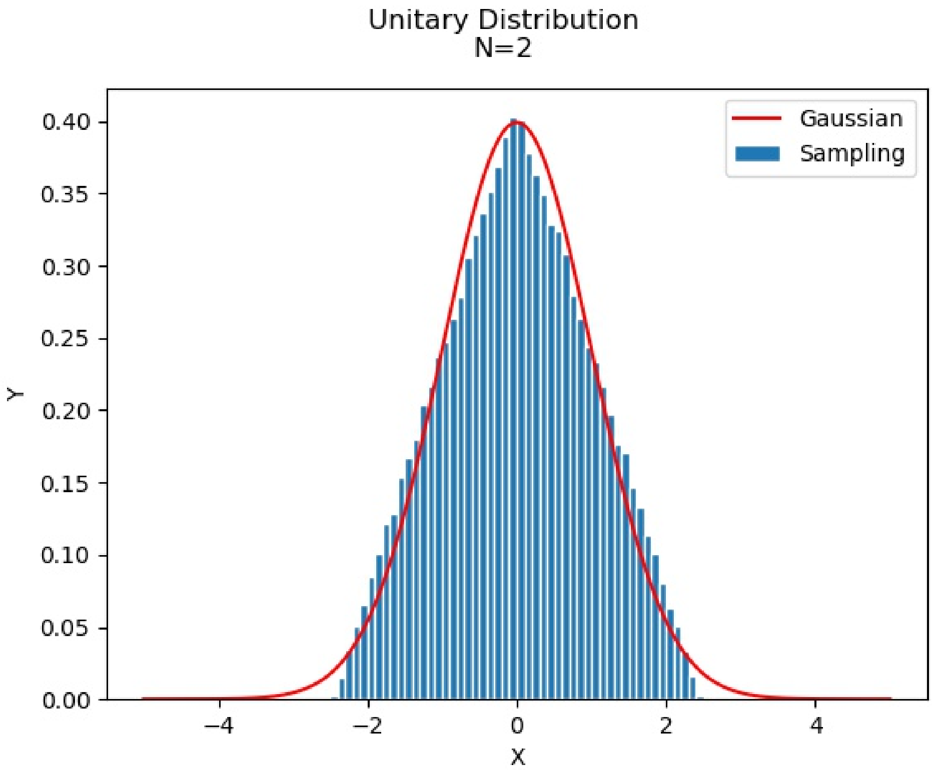

图 7: N=2时均匀分布的结果

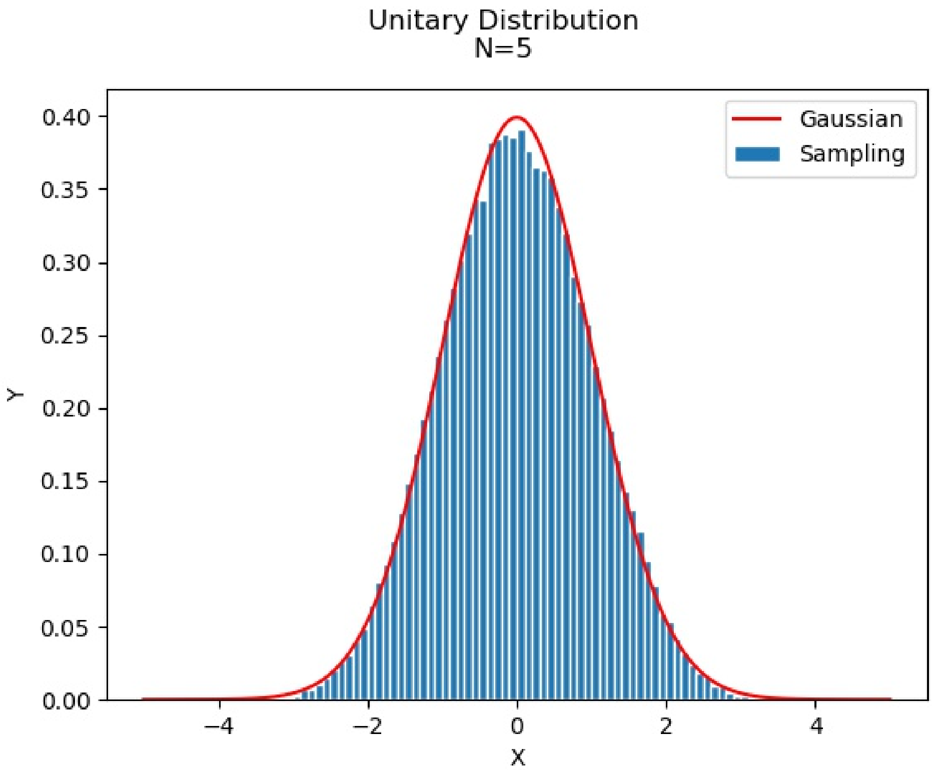

图 8: N=5时均匀分布的结果

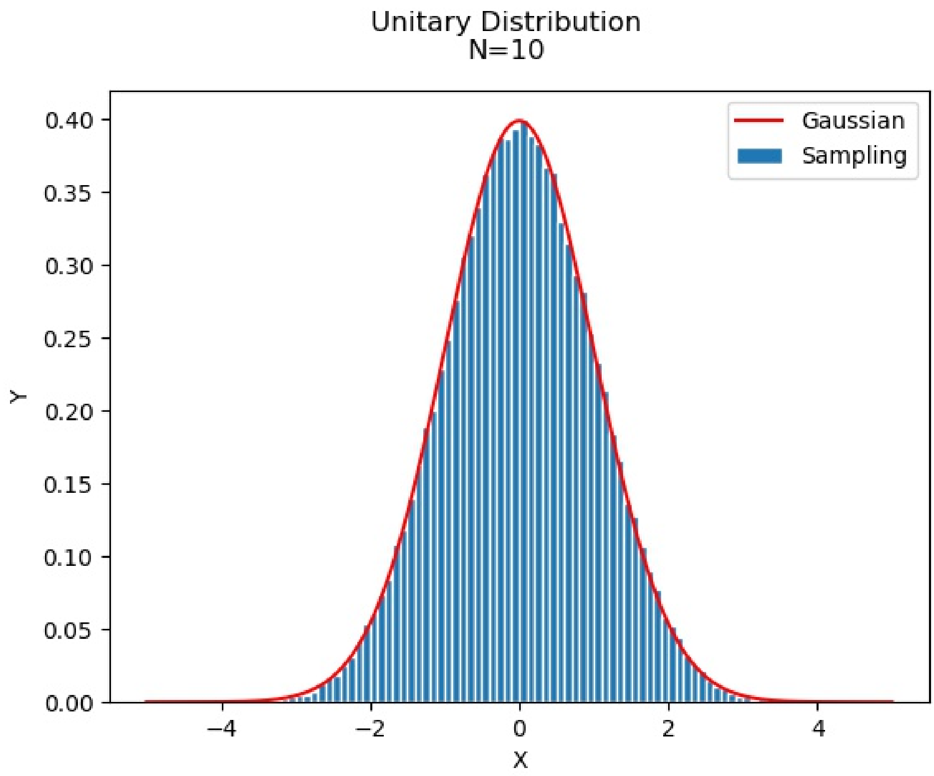

图 9: N=10时均匀分布的结果

4.2.2 指数分布

指数分布为：

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

不失一般性，取$\lambda = 1$，则$EX = 1, Var(X) = 1$。
得到的结果如下所示：

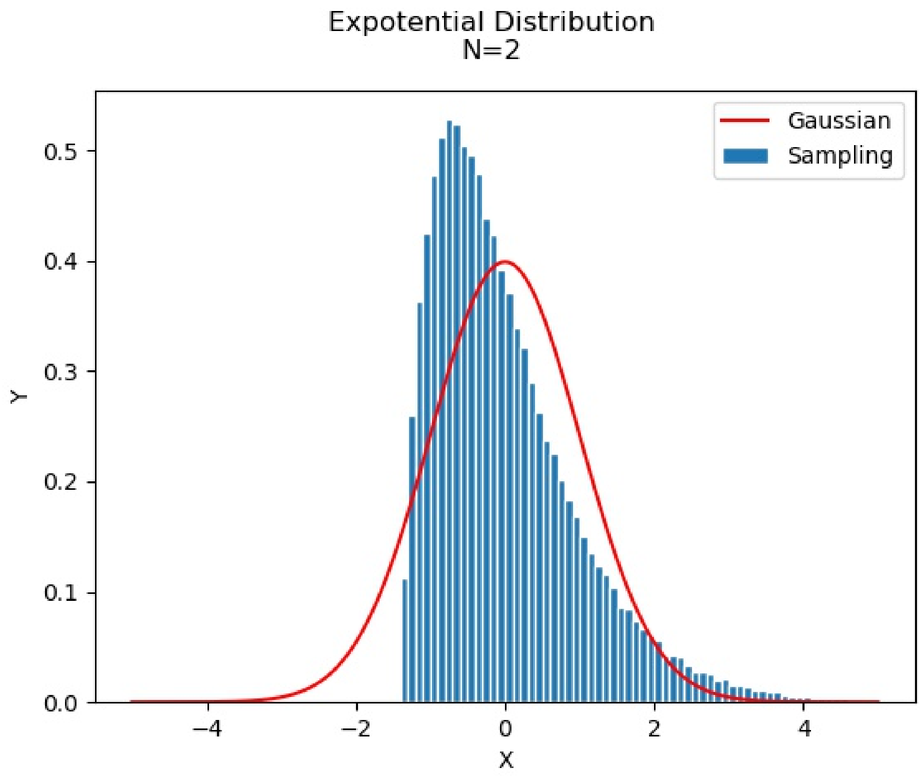

图 10: N=2时指数分布的结果

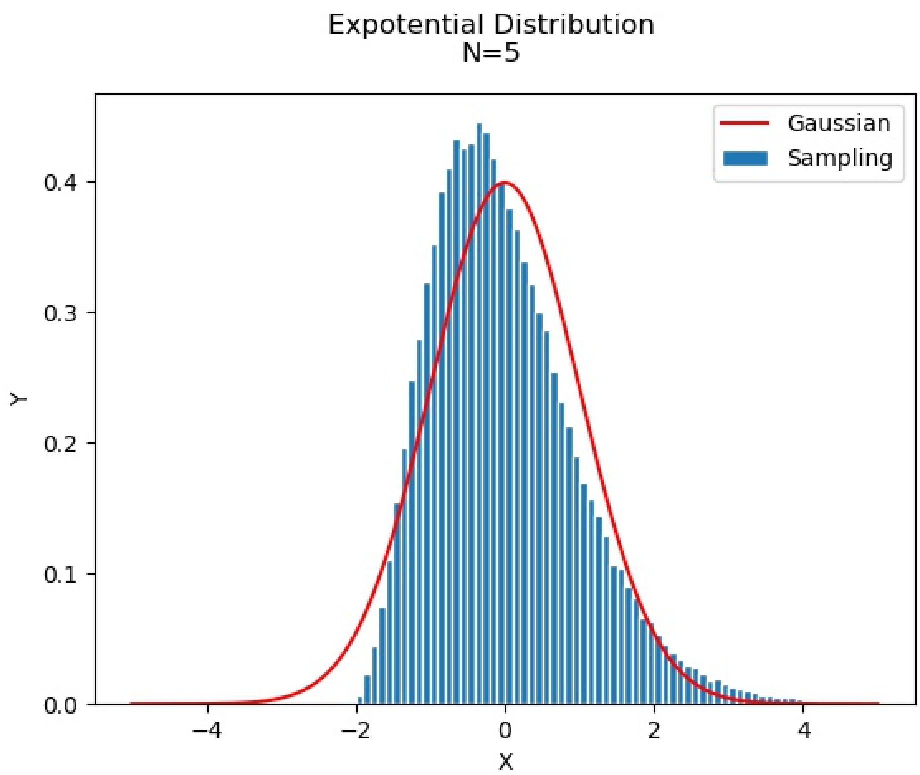

图 11: N=5时指数分布的结果

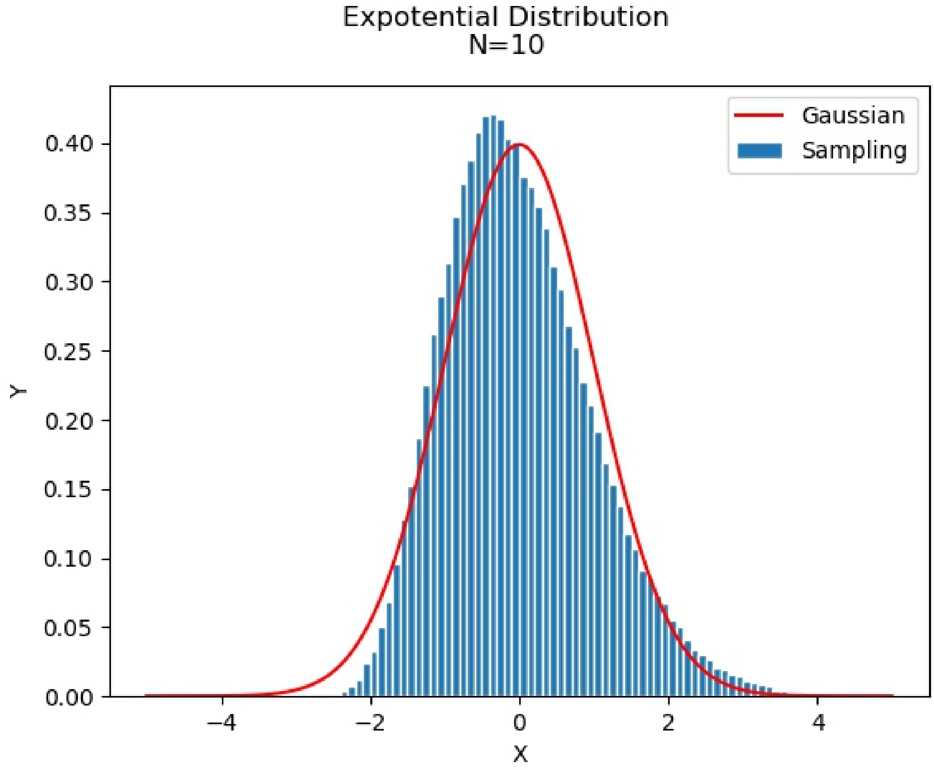

图 12: N=10时指数分布的结果

4.3 不同分布函数的平均值的分布

这里仅举一例，求均匀分布的随机变量和指数分布的随机变量的平均值的分布。
设前int(N/2)个X_i为[0,4]内的均匀分布，后面的X_i为$\lambda = 1$的指数分布。设$a = int(N/2), b = N - a$
则：

$$\begin{aligned} EX &= E\left[\frac{1}{N}(\sum_{i=1}^{a} X_i + \sum_{i=a+1}^{N} X_i)\right] = \frac{1}{N}(a\sum_{i=1}^{a} EX_i + b\sum_{i=a+1}^{N} EX_i) = \frac{1}{N}(aE_1X + bE_2X) \\ &= \frac{1}{N}(2a+b) = 1 + \frac{1}{N}int(N/2) \end{aligned} \tag{12}$$

$$\begin{aligned} Var(X) &= Var\left[\frac{1}{N}(\sum_{i=1}^{a} X_i + \sum_{i=a+1}^{N} X_i)\right] = \frac{1}{N^2}\left[a\sum_{i=1}^{a} Var(X_i) + b\sum_{i=a+1}^{N} Var_1(X_i)\right] = \frac{1}{N^2}\left[aVar_1(X) + bVar_2(X)\right] \\ &= \frac{1}{N^2}(\frac{4}{3}a + b) = \frac{1}{3N^2}\left[3N + int(N/2)\right] \end{aligned} \tag{13}$$

得到的结果如下所示：

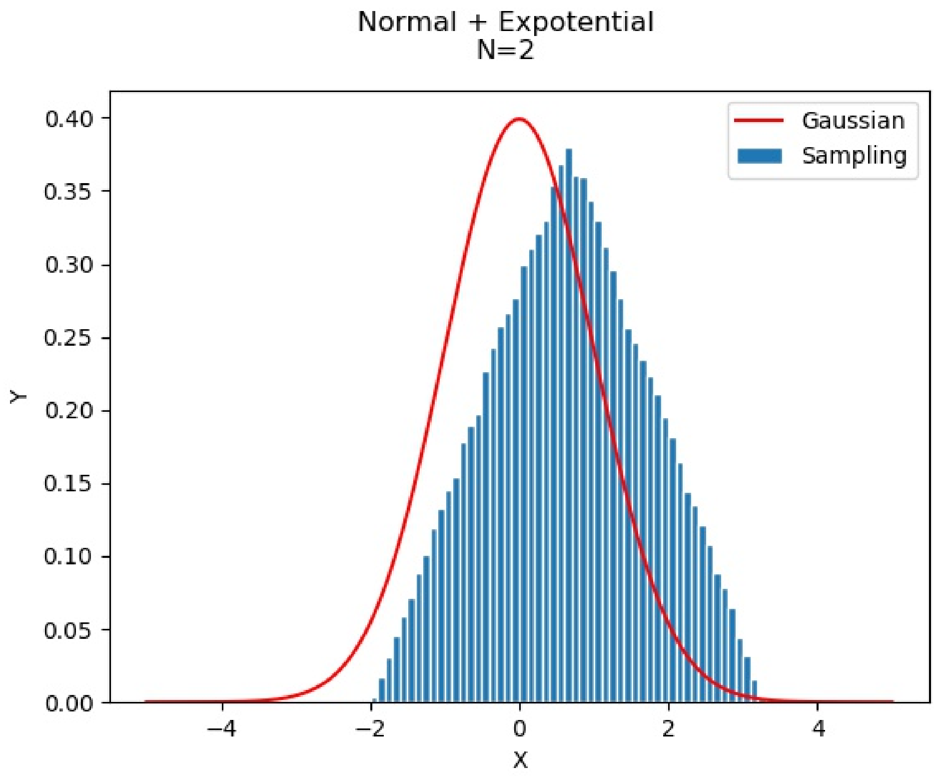

图 13: N=2时和函数分布的结果

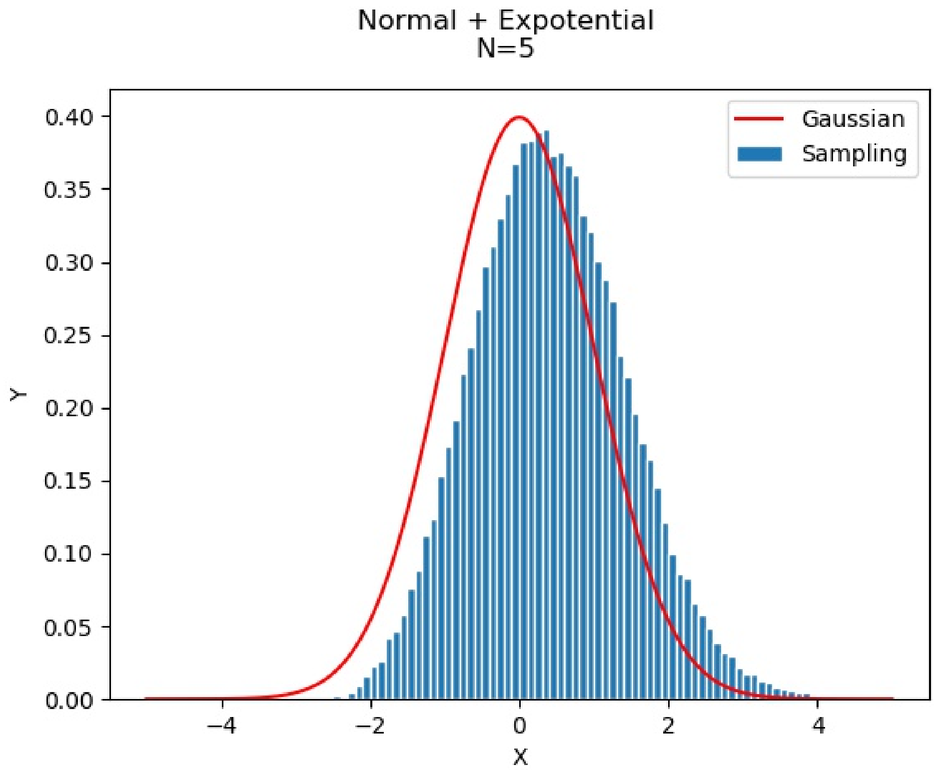

图 14: N=5时和函数分布的结果

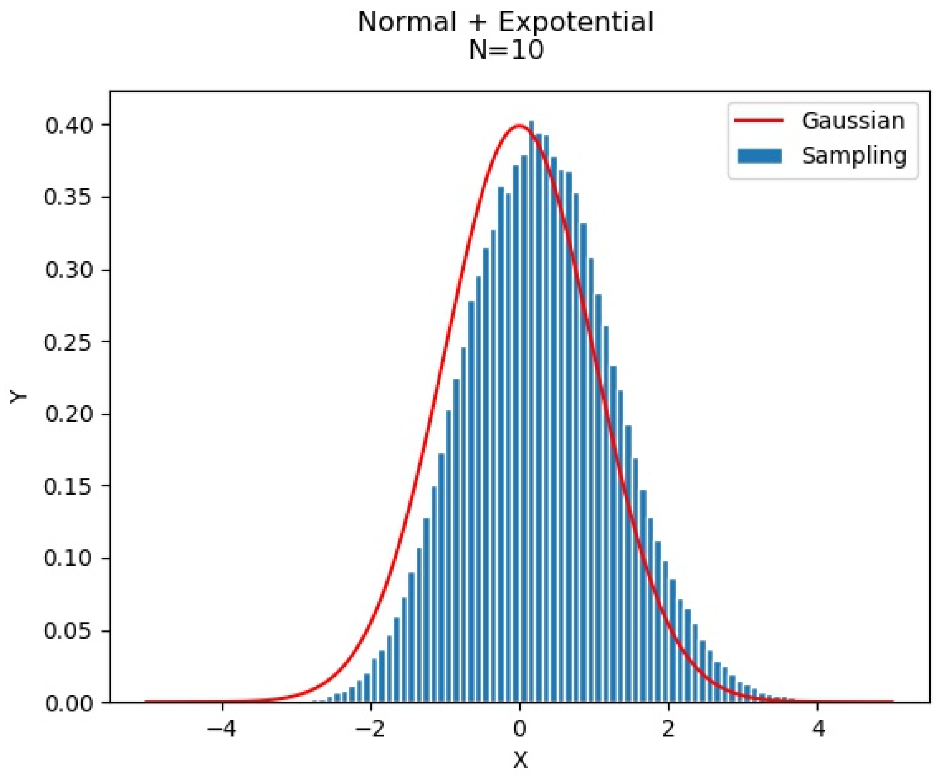

图 15: N=10时和函数分布的结果

5 讨论

5.1 原理的讨论

对于中心极限定理：

$$P\left\{\left|\frac{\langle f\rangle-\mu}{\sigma_f/\sqrt{N}}\right|<\beta\right\}\to\Phi(\beta) \tag{14}$$

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。

此外，4.3中X_i的分布不一样，期望和方差也不一样，但仍然满足中心极限定律。可以看到N的增加会使图像与标准正态分布更相似。

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。

3. 对于不同分布函数的平均值的函数，其变量也满足中心极限定理，与正态分布的相似程度与和函数内部的函数有关。

   当然，当N趋于无穷时，内部的分布函数形式对总体是正态分布无影响。这也是可以直接思考得出来的，因为不同分布的和函数本质上也有一种“线性”性质。每一个函数在N趋向无穷时都遵从正态分布，它们的平均值自然也遵从正态分布。

4. 从上面结果可以大胆推断，不管各种变量怎么组合，只要把它们的某种组合当作变量，这个变量存在且有意义，这个变量的期望和方差存在且可以求出来，当N趋向于无穷时，仍然满足中心极限定理，即：

   $$P\left\{\left|\frac{\langle f\rangle-\mu}{\sigma_f/\sqrt{N}}\right|<\beta\right\}\to\Phi(\beta) \tag{15}$$

   对于生活中的各种分布，很多都不是理想意义上的数学分布。但中心极限定理给我们理论上的计算的形式，并且误差不大。因此中心极限定理在日常生活中也非常有用。