PG15 標本平均の分布 (2) 一様母集団

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats          # 統計ライブラリを使用
```

母集団から取り出した n 個の標本 $X_1, X_2, \ldots, X_n$ の 標本平均 $\bar{X}$ がどのような確率分布に従うかを, 母集団分布が

1) コイン投げの分布 $P(X=0) = P(X=1) = 1/2$

2) 区間 $[0, 2]$ 上の一様分布

の場合について確認する。標本平均の分布は n によって変化するが, CLTによれば n が大きいほど正規分布 $N(\mu, \sigma^2/n)$ に近づくはずである。 ここで, μ は母平均, σ^2 は母分散である。

【演習】方法をまねして、次の場合も確認せよ。

1）サイコロ振りの分布 $P(X=1) = P(X=2) = \cdots = P(X=6) = 1/6$

2) 指数分布

1. 一様母集団

区間 [0,2] 上の一様分布に従う母集団からの標本抽出を考える。

一様乱数(デフォルトで0以上1以下)を発生させて, 発生した乱数を2倍することで、 母集団から取り出した標本とする。

In [2]:

```
Z = 2 * np.random.rand()
Z
```

Out[2]:

```
0.4866838456366436
```

標本を n 個取り出して標本平均 SM を調べる

n は標本数（サンプルサイズ）なので size と言う名の変数を使うことにする。

In [3]:

```
size = 5          # 自分で設定
Record = []       # コイン投げを記録するためのリストを準備。初期値は空
for _ in range(size):
    Z = 2 * np.random.rand()
    Record.append(Z)
Sample = np.array(Record)      # コイン投げの結果(0または1)をリストからアレイに変換
Sample
```

Out[3]:

```
array([1.44415335, 0.70035064, 0.40540713, 0.21443857, 1.89966294])
```

n 個の標本を抽出したら, その平均値を計算する。 この値が標本平均の実現値であり、ただ1個の数値である。

In [4]:

```
sm = np.mean(Sample)
sm
```

Out[4]:

```
0.9328025256933141
```

以上の事前チェックののち、いよいよ標本平均の実現値を大量に収集する。

標本平均の実現値 sm は、n 回のコイン投げのたびに異なる値が出る。 したがって、sm の値の出方の傾向が大事になるが、これが「標本平均の分布」にあたる。

ここでは n 回のコイン投げを1セットとして、これを多数回（trial 回）繰り返すことで、 標本平均の実現値 sm を収集して、その分布を可視化する。

In [5]:

```
size = 15         # size 個の標本を抽出する。これを1セットとして
trial = 10000     # trial 回の標本平均を収集する
SM_list = []      # 標本平均を記録するためのリストを準備。初期値は空
for i in range(trial):
    Record = []
    for _ in range(size):
        Z = 2 * np.random.rand()
        Record.append(Z)
    Sample = np.array(Record)
    sm = np.mean(Sample)
    SM_list.append(sm)
SM = np.array(SM_list)      # NumpyArray が便利
SM
```

Out[5]:

```
array([1.16511907, 1.09635032, 1.1171347 , ..., 1.11277884, 0.99408638,
       1.05220882])
```

ヒストグラムまたは度数折れ線による標本分布の可視化

SM に現れる数値の分布が、すなわち標本分布である。これを可視化する。

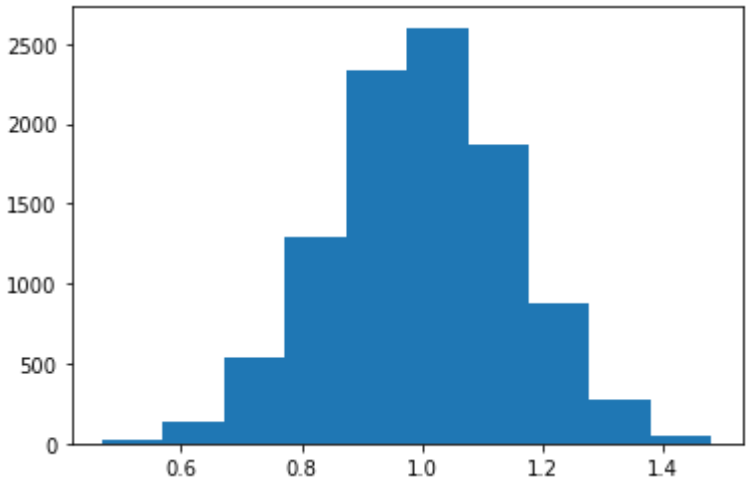
全自動でヒストグラムを描画すると、いかにもまずいものが出力される。

In [6]:

```
plt.hist(SM)
```

Out[6]:

```
(array([  21.,  140.,  533., 1296., 2338., 2602., 1868.,  881.,  271.,
         50.]),
 array([0.46872314, 0.56993618, 0.67114921, 0.77236225, 0.87357528,
        0.97478832, 1.07600136, 1.17721439, 1.27842743, 1.37964046,
        1.4808535 ]),
 <BarContainer object of 10 artists>)
```



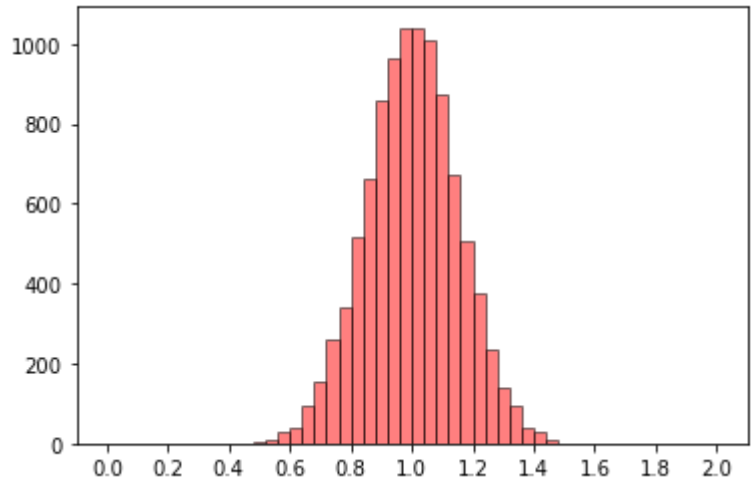
これを見ながら座標軸や階級の設定をすること。

1) sm の値は 0.0 から 2.0 までの値である。

2）sm は連続量であるから、階級幅は任意にとって大丈夫だが、 見やすさの観点から、階級の個数は b = trial/200 程度にしておこう。

In [7]:

```
b = int(trial/200)
F,x,_ =plt.hist(SM, range=(0, 2), bins=b,
                color='red', alpha=0.5, ec='k')
plt.xticks(np.arange(0, 2.1, 0.2))   # x軸の目盛
plt.show()
```



観察

1) $size = 1$ とすれば母集団分布が再現されるはず。

2) $size = 2$ とすれば山型グラフ、 $size = 3$ とすると放物線の組合せになる。 trial を大きくとり、階級幅を狭く(bins を大きく）とればより明瞭になる。

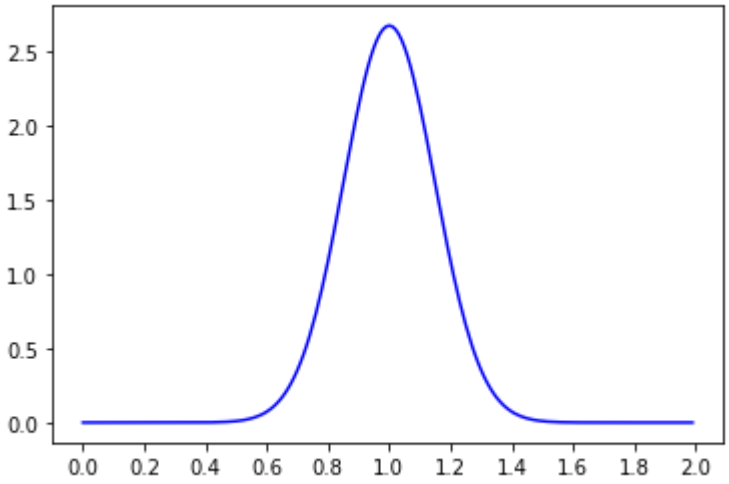
3) $size$ を大きくすると、正規分布 $N(\mu, \sigma^2/n)$ に近づく。 [0,2] 上の一様分布の場合は、$\mu = 1, \sigma^2 = 1/3$ である（理論的にわかる）。

正規分布曲線を重ねて描画する

In [8]:

```
m = 1                    # 母平均の設定
s2 = 1/3                 # 母分散の設定
Z = stats.norm(m, np.sqrt(s2/size))

x = np.arange(0, 2, 0.01)                  # x の範囲の指定、始点、終点、刻み幅
plt.plot(x, Z.pdf(x), color='blue')
plt.xticks(np.arange(0, 2.1, 0.2))     # x軸の目盛
plt.show()
```



ヒストグラムでは、

積み上げる長方形の横の長さ = 2/b = 400/trial,

縦の長さ = 1,

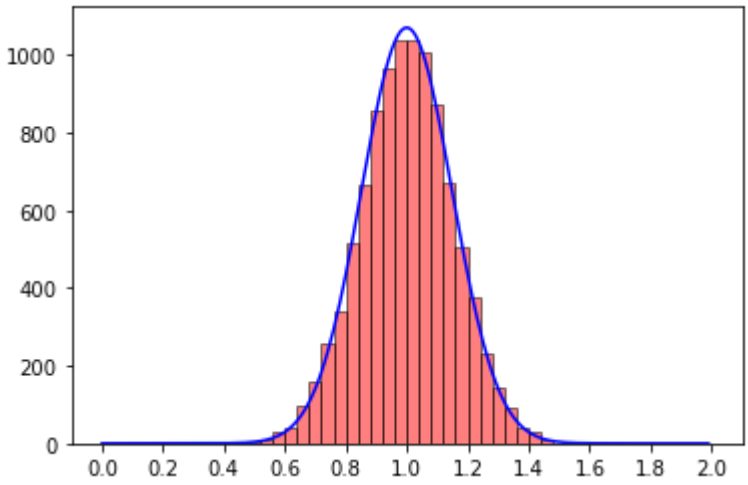
長方形の個数 = trial

であるから、ヒストグラム全体の面積 = 400 となる。

一方、密度関数の面積は 1 であるから、密度関数の値を 400 倍すれば 両者の比較ができる。

In [9]:

```
plt.hist(SM, range=(0, 2), bins=b,
         color='red', alpha=0.5, ec='k')
plt.plot(x, 400*Z.pdf(x), color='blue')
plt.xticks(np.arange(0, 2.1, 0.2))
plt.show()
```



In [ ]:

```
```

In [ ]:

```
```