PG14 標本平均の分布 (1) コイン投げ

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. コイン投げ

一様乱数を発生させて, 発生した乱数が0.5未満なら裏(Z=0), 発生した乱数が0.5以上なら表(Z=1)と解釈することで, 公平なコイン投げのシミュレーションになる。

In [2]:

```
if np.random.rand()<0.5:
    Z=0
else:
    Z=1
Z
```

Out[2]:

```
0
```

標本を n 個取り出して（コインを n 回投げて）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):
    if np.random.rand() < 0.5:
        Z = 0
    else:
        Z = 1
    Record.append(Z)
Coin = 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
Coin
```

Out[3]:

```
array([1, 0, 1, 1, 1])
```

n 回のコイン投げが終われば, その平均値（表の出る平均回数）が計算できる。 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(Coin)
sm
```

Out[4]:

```
0.8
```

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 = 50          # size 回のコイン投げを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):
        if np.random.rand() < 0.5:
            Z = 0
        else:
            Z = 1
        Record.append(Z)
        Coin = np.array(Record)
        sm = np.mean(Coin)
    SM_list.append(sm)
SM = np.array(SM_list)      # NumpyArray が便利
SM
```

Out[5]:

```
array([0.52, 0.64, 0.48, ..., 0.42, 0.42, 0.54])
```

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。

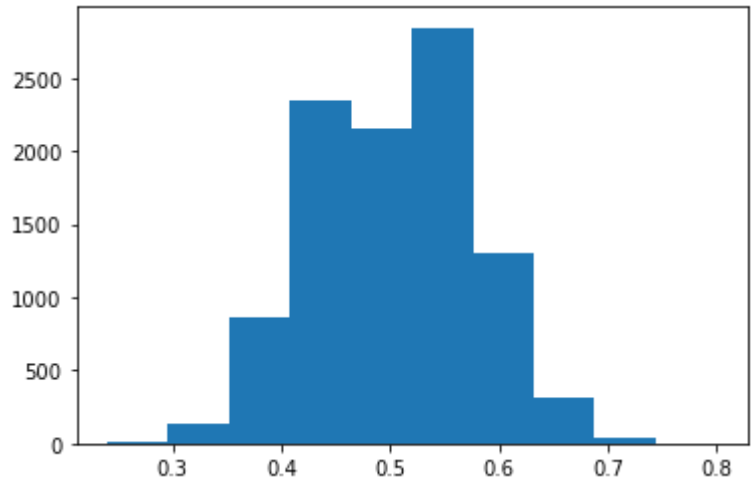
全自動でヒストグラムを描画すると、いかにもまずいものが出力される。

In [6]:

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

Out[6]:

```
(array([8.000e+00, 1.410e+02, 8.690e+02, 2.347e+03, 2.150e+03, 2.843e+03,
        1.301e+03, 3.070e+02, 3.200e+01, 2.000e+00]),
 array([0.24 , 0.296, 0.352, 0.408, 0.464, 0.52 , 0.576, 0.632, 0.688,
        0.744, 0.8  ]),
 <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。

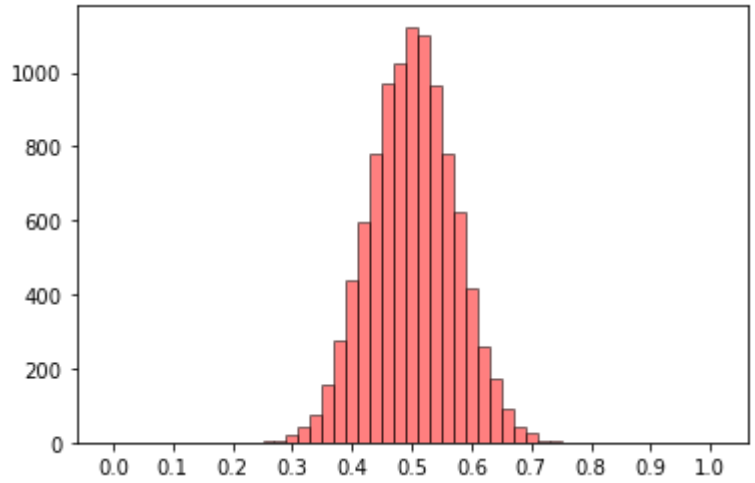
1) sm の値は 0.0 から 1.0 まで 0.2 刻みで現れる。

2）ここで、0.2 = 1/size であるから、 x 軸は $0 - 1/(2 * size)$ から $1 + 1/(2 * size)$ を6階級に等分割するのがよい。 6階級の 6 = size +1 である。

3）そうすることで、階級値が、0.0, 0.2, 0.4, 0.6, 0.8, 1.0 となる。

In [7]:

```
b = size + 1                # 階級の個数
F,x,_ =plt.hist(SM,
                range=(0-1/(2*size), 1+1/(2*size)),
                bins=b,
                color='red',
                alpha=0.5,
                ec='k')
plt.xticks(np.arange(0, 1.1, 0.1))    # 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) 一般 $size = n$ の場合は、本質的に二項分布である。つまり、

$$P(sm = k/n) = \binom{n}{k}(1/2)^k(1 - 1/2)^{n-k}$$

が理論的にわかる（確認せよ）。

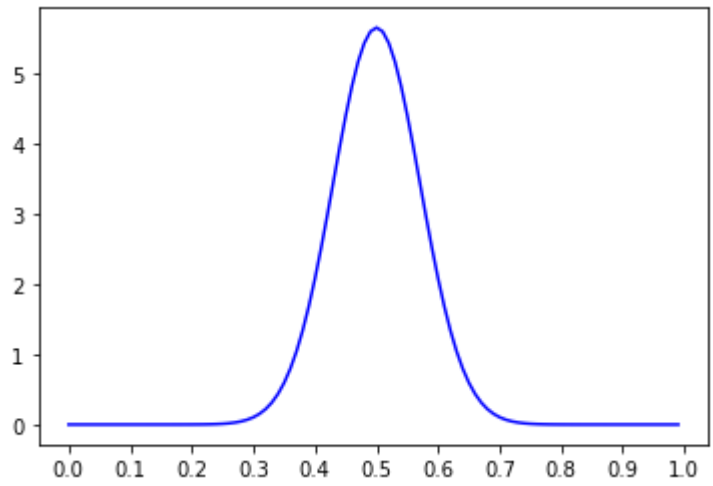
3) $size$ を大きくすると、正規分布 $N(\mu, \sigma^2/n)$ に近づく。 コイン投げの場合は、$\mu = 1/2, \sigma^2 = 1/4$ である（理論的にわかる）。

正規分布曲線を重ねて描画する

In [8]:

```
m = 1/2                  # 母平均の設定
s2 = 1/4                 # 母分散の設定
Z = stats.norm(m, np.sqrt(s2/size))

x = np.arange(0, 1, 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, 1.1, 0.1))     # 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 = 1/size,

縦の長さ = 1,

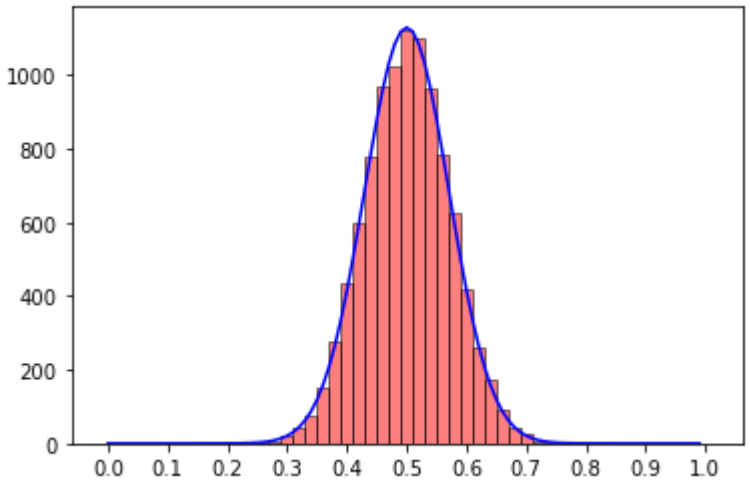
長方形の個数 = trial

であるから、ヒストグラム全体の面積 = trial/size 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 trial/size 倍すれば 両者の比較ができる。

In [9]:

```
plt.hist(SM, range=(0 - 1/(2*size), 1 + 1/(2*size)), bins=b,
         color='red', alpha=0.5, ec='k')
plt.plot(x, (trial/size)*Z.pdf(x), color='blue')
plt.xticks(np.arange(0, 1.1, 0.1))    # x軸の目盛
plt.show()
```



In [ ]:

```
```

In [ ]:

```
```