プログラミング技法II

担当: 新田 直子
大学院工学研究科 電気電子情報工学専攻
naoko@comm.eng.osaka-u.ac.jp
http://www2c.comm.eng.osaka-u.ac.jp/~prog2/

例：正規分布の信頼区間と信頼度

```
from scipy.stats import norm

param = [(0.0, 1.0), (-1.0, 6.0), (9.0, 9.0), (1.0, 0.5)]

for mean, std in param:
    print("mean: ", mean, ", std: ", std)
    for i in range(1,4):
        prob = norm.cdf(x=mean+i*std, loc=mean, scale=std) ¥
            - norm.cdf(x=mean-i*std, loc=mean, scale=std)
        print("-", i, "*std to ", i, "*std: ", prob)
```

長い行を改行

例：適合度検定

- 課題5-1: 2回目課題の19枚目スライドを参考に、
  平均0、分散1の正規分布に従うN個の乱数を
  生成し、χ^2を計算せよ。
  これをM回繰り返し得られるχ^2の分布が、
  自由度$k-1$のχ^2分布に近似するか確認せよ。
  期待度数の算出には、モジュールscipy.statsを
  用いてよい。

例：適合度検定

- 課題5-2: weight-height.csvの男性／女性の体重／身長（いずれかでよい）、marathon_results.csvの時間(time)のデータがそれぞれ正規分布に従うかカイ二乗検定により判定せよ。ただし、データの標準化には課題4-2で作成した関数を用いよ。また、求めたχ^2に対するp値は、モジュールscipy.statsを用いて求めてよい。

  ※χ^2分布の自由度に注意せよ。

※適合度検定

ある母集団の互いに重なり合わないk個の事象$A_1, \cdots, A_k$について、それぞれが起こる確率が$p_1, \cdots, p_k (\sum_{i=1}^{k} p_i = 1)$とする。

この母集団からn個の標本をとるとき、それらがA_iに入る観測度数をf_i、A_iに入る期待度数を$e_i (= np_i)$とする。

このとき、統計量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{{f_i}^2}{e_i} - n$$

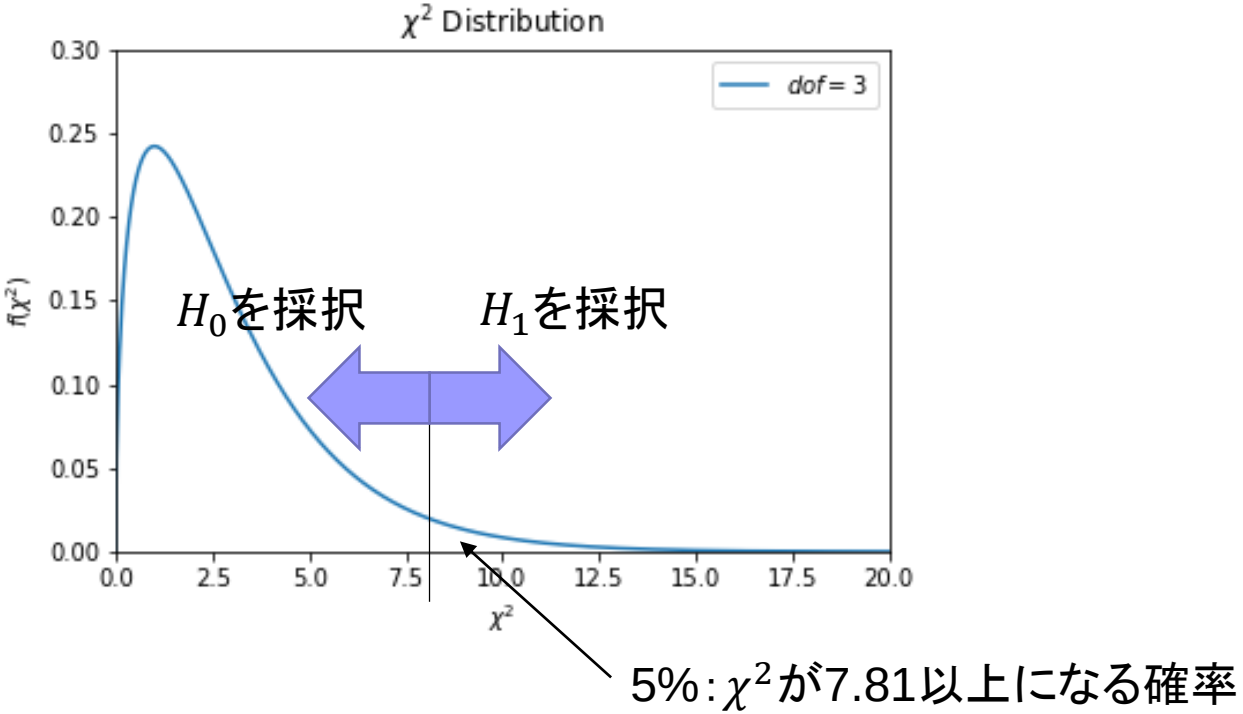
の分布は、nが大きく、各e_iが5以上であれば近似的に自由度$k-1$のχ^2分布と一致する。

※適合度検定

観測データが理論値に当てはまっているか？

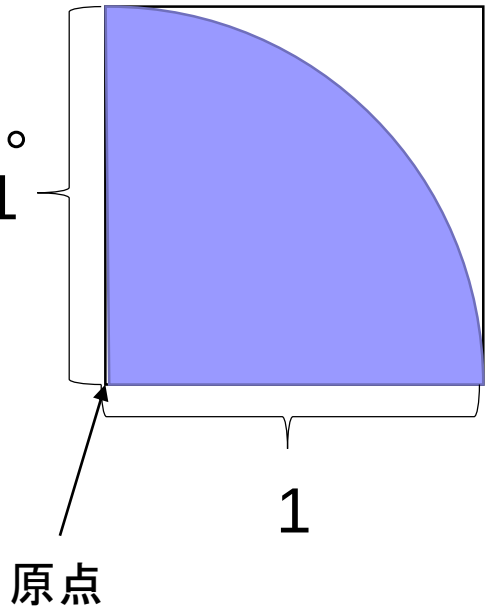
帰無仮説H_0: 観測データは理論値に当てはまっている

対立仮説H_1: 観測データは理論値に当てはまっていない



例：πを求める（シミュレーション）

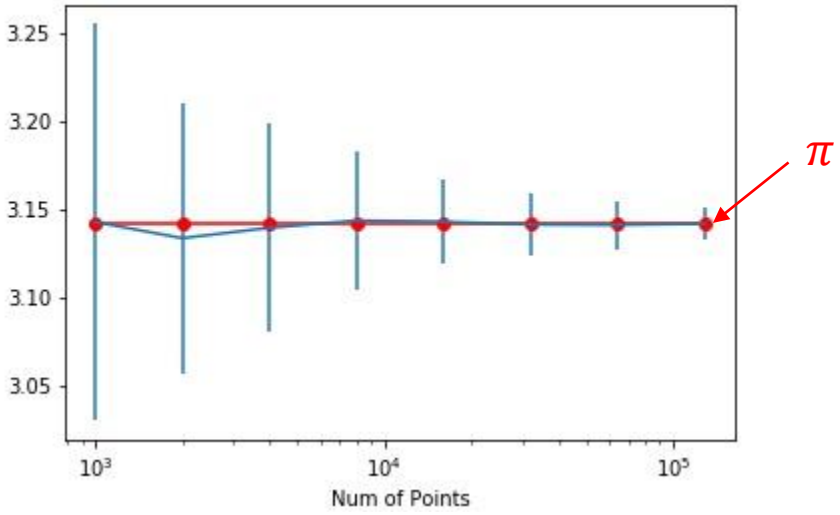
課題6-1：　右図のような1×1の正方形内にランダムに点を打ち、原点から距離が1以下の点（半円内に入る）の割合を算出する。半円の面積が$\pi/4$、正方形の面積が1なので、算出した割合$\times 4 = \pi$となる。点数Nを与えるとπの推定値を出力する関数を作成せよ。



例：πを求める（シミュレーション）

課題6-2:N個の点を打ち、πを推定することをM回繰り返し、推定値πが正規分布に従うか課題5-2と同様に判定せよ。正規分布に従うとき、95%信頼区間の幅が求める精度を満たすまでNを増やし、Nによる平均と95%信頼区間の変化をエラーバー付きグラフで示せ。

グラフの例（精度$mean \pm 0.01$のとき)



例：大数の法則と中心極限定理

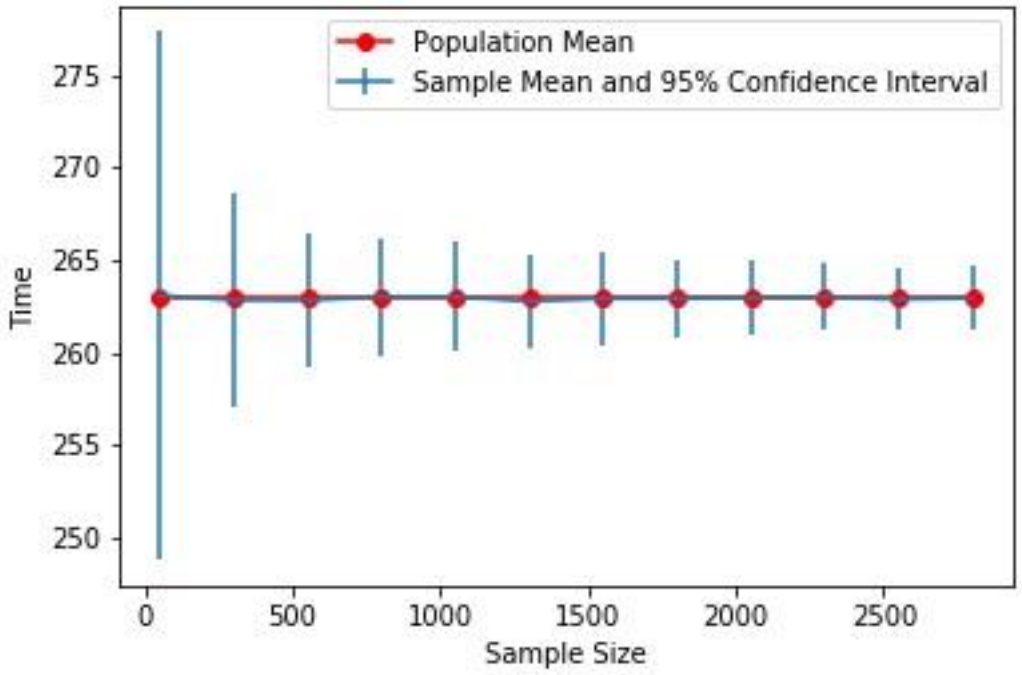
- 課題7-1: marathon_results.csvの時間(time)はあるマラソンにおける各選手のタイムである。一部の選手からこのマラソンにおける平均タイムを求めることを考える。N個のランダムに抽出した選手の平均タイムを得る（標本平均）。これをM回繰り返し、標本平均の平均と標準偏差を算出せよ。

  Nを変化させ、各Nに対して得られたM個の平均が正規分布に従うか、課題5-2と同様にカイ二乗検定により判定せよ。また、標本平均が正規分布に従うとき、その平均及びその95%信頼区間をエラーバー付きグラフで示せ。また、全選手の平均タイム（母平均）を同じグラフに表示し、それらの関係について考察せよ。

例：大数の法則と中心極限定理

- 課題7-1（続き）：

グラフの例



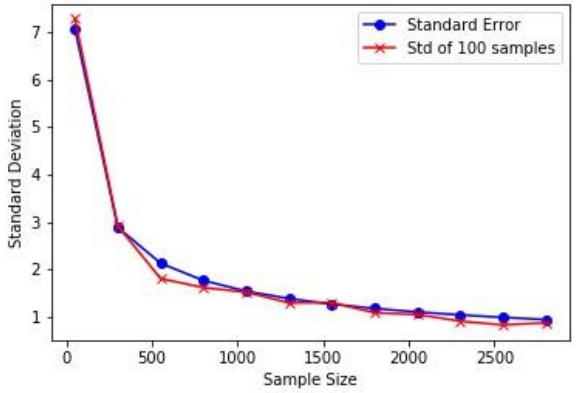
例：大数の法則と中心極限定理

課題7-2: 標本平均の標準偏差である標準誤差は、
母標準偏差σにより、

$$SE = \frac{\sigma}{\sqrt{N}}$$

で求められる。全選手のタイムの標準偏差σを用いて、
各Nに対する標準誤差SEを求め、課題7-1で得た、
各Nに対して得られたM個の標本平均の標準偏差と
と共にグラフに表示し、比較せよ。

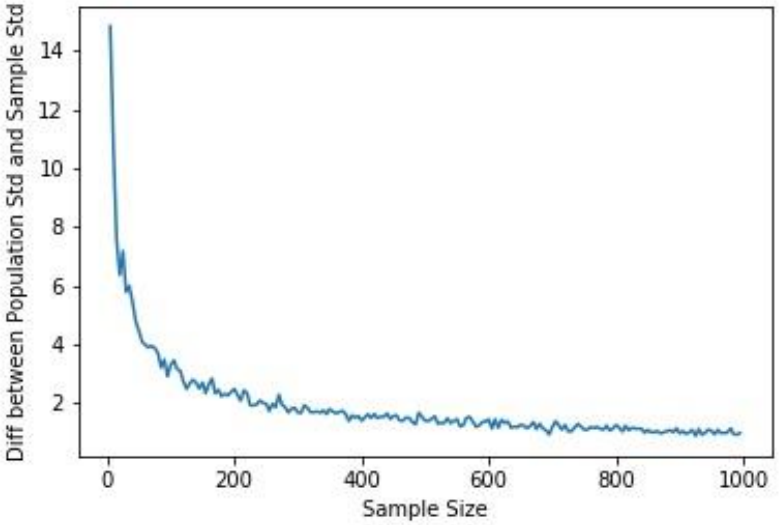
グラフの例



例：大数の法則と中心極限定理

課題7-3: N人の選手のタイムの標準偏差と母標準偏差σを比較する。これらの差をM回算出し、その平均をグラフにプロットせよ。

グラフの例



例：大数の法則と中心極限定理

課題7-4: 課題7-1～7-3の結果を踏まえ、適当な人数の選手のタイムを抽出し、平均タイムとその95%信頼区間を算出せよ。これを何回も（10000回など）繰り返し、全選手の平均タイムが求めた信頼区間に入らない割合を出力せよ（約5%となるはず）。

例：就職活動問題

課題8: 1～ Nまでの番号が付いたカードを裏返して並べる。
この中から最大の番号が付いたものを選びたい。
左端から順番に表向け、M枚までの最大値を閾値とし、
M枚目以降に閾値を超える最初のカードを選んだとき、
それが最大値であればあなたの勝ちである。
Mをどう設定すれば勝率が最大となるか、
シミュレーションにより決定せよ。
また、横軸をM（もしくはM/N）、縦軸を勝つ確率としたグラフをプロットせよ。

レポートの提出

- 課題5-1、5-2、6-1、6-2、7-1～7-4、8に取り組む。
- 各課題に対し、プログラム作成時の考え方、
  ソースプログラム、実行結果、考察を
  レポートに記載する。
- レポートとソースコードを入れたフォルダを圧縮し
  下記アドレスに提出する。
  prog2@nanase.comm.eng.osaka-u.ac.jp
- 読みやすいレポートとするよう心がけること。
- Subjectは「【Report3】学籍番号 氏名」とする。
- 提出期限：6月6日（木）