データ全体を母集団と呼び、その母集団から抽出されたデータを標本、またはサンプルと呼びます。この２つは明確に区別される必要があります。そして、その標本の数を標本の大きさとか標本サイズといいます。

母集団と標本の違いの理解

硬貨を投げで母集団と標本の違いを理解

1. 硬貨を2回投げて表裏を記録
2. 表を1、裏を-1としてそれぞれの試行の平均を計算
3. 母集団の平均ゼロと比較
4. １～3を5回実行

例

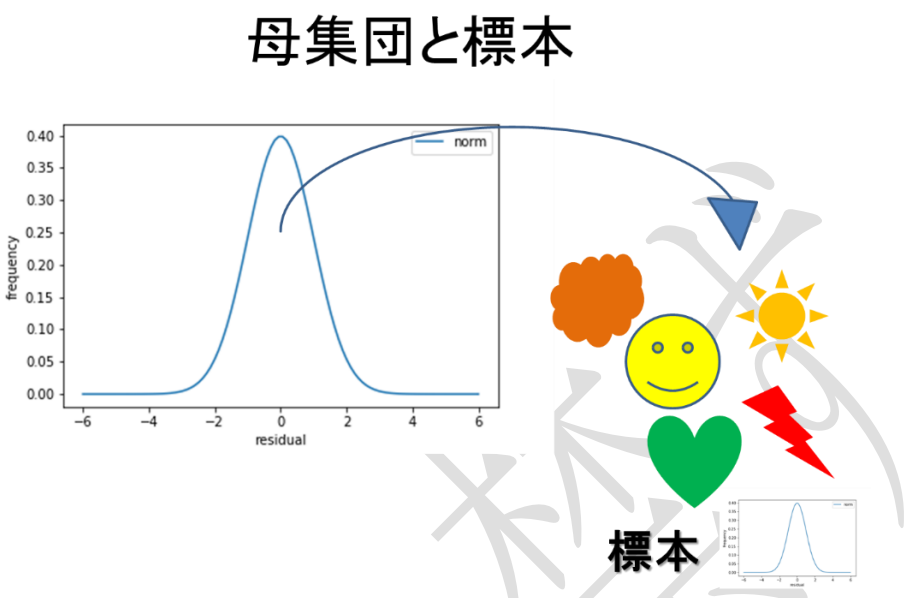
1回目:おもて、おもて　->(1+ 1)/2=1
2回目:うら、おもて　->(-1+ 1)/2=0
3回目:おもて、うら　->(1- 1)/2=0
4回目:うら、うら　->(-1-1)/2=-1
5回目:うら、おもて　->(-1+ 1)/2=0

投げる回数が2回だと平均がゼロになるとは限らないことが分かります。

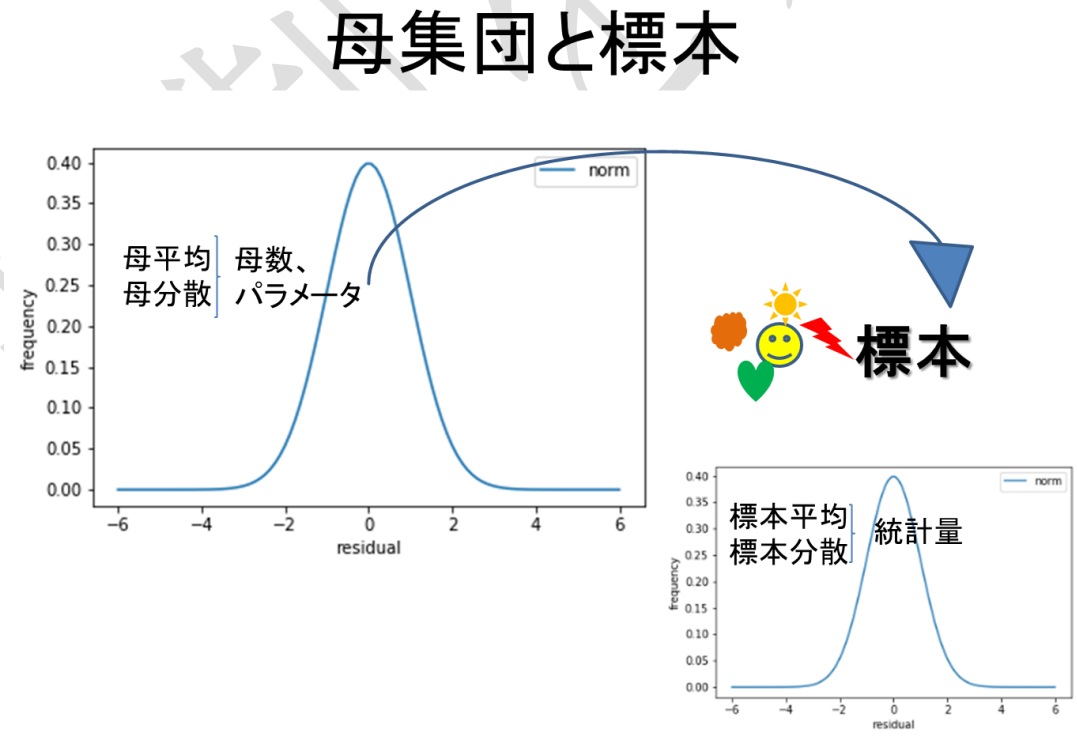
3.1 母集団と推定

母集団とは、調査対象となっているデータのもととなる集合のことです。母集団というときには２つのタイプに分類されます。定義により母集団が確定している場合と、ある特定の確率分布を前提としている場合があります。前者は選挙の当選予測などに相当します。後者は株価の予測などです。標本は、その母集団から抽出された個体の集合で、母集団の部分集合です。私たちは母集団について知りたいと思っているのですが、実際に知ることができるのは標本についてであって母集団についてではありません。したがって、推測統計では、部分集合である標本から集合全体の母集団を推測します。この過程では誤差が生じます。そこで、その大きさを確率の理論を用いて評価し、分析結果の信頼度を明らかにします。これが統計的手法を用いる１つのメリットです。

繰り返しになりますが、母集団は様々な理由から母集団すべてを把握できないために、その母集団から標本を得て、その標本を分析します。つまり、標本を分析しながら、母集団の特性を知ろうとしているのです。



母集団(確率分布)を特徴づける定数を母数(パラメータ)といいます。母平均、母分散は母数です。一方、標本から作られた関数を統計量といいます。標本平均、標本分散は統計量です。



3.2 大数の法則と中心極限定理

データ全体を母集団と呼び、その母集団から抽出されたデータを標本といいます。標本の大きさが大きくなるとそれにともない、標本から得られる統計量は真の統計量(母数)に近づいていきます。

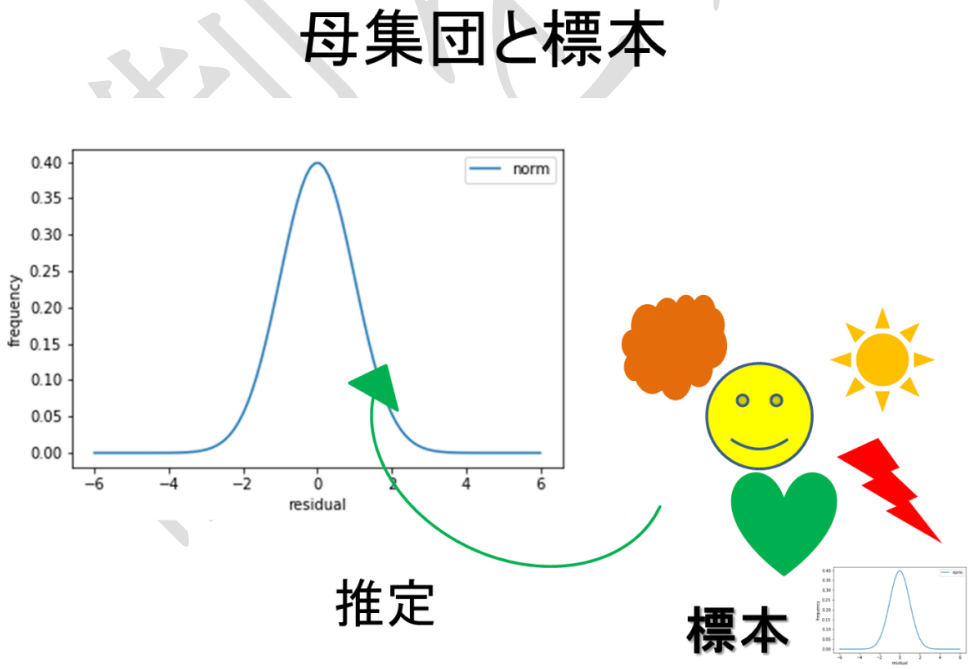
母集団が平均をもつときに、標本の大きさを大きくしていくと、母集団のもつ平均値(母平均)、または真の平均に標本の平均は近づいていきます。これを大数の法則といいます。真の平均と標本の平均の誤差は標本の大きさを大きくすれば正規分布に近づいていきます。これが中心極限定理です。

大数の法則により、Nが大きくなれば、観測データの平均$\bar{x}$は期待値 μ に近づきます。期待値はしたがって、理論的な確率分布の平均と同じです。

２０２１/03/12

3.3 推定の性質

推測統計では、部分集合である標本から統計量を用いて母集団の母数を推定量として推測します。そこで推定量の性質について明らかにします。



3.3.1 一致性

ある母数の推定量がデータの数の増加にしたがい母数に収束するとき、それを一致性とよび、そのような推定

量を一致推定量といいます。実際には標本の大きさは有限であり、推定量にはばらつきがあります。

3.3.2 不偏性

もう 1 つの推定量の基準に不偏性があります。推定量の期待値が母数に等しくなるとき不偏性があるといいます。その性質をもつ推定量を不偏推定量といいます。σ^2 の不偏推定量は、得られたデータが $x_1, x_2, \ldots x_n$ のとき

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

という形で得ることができます。$\bar{x}$は得られたデータの平均値です。これを不偏分散とよびます。

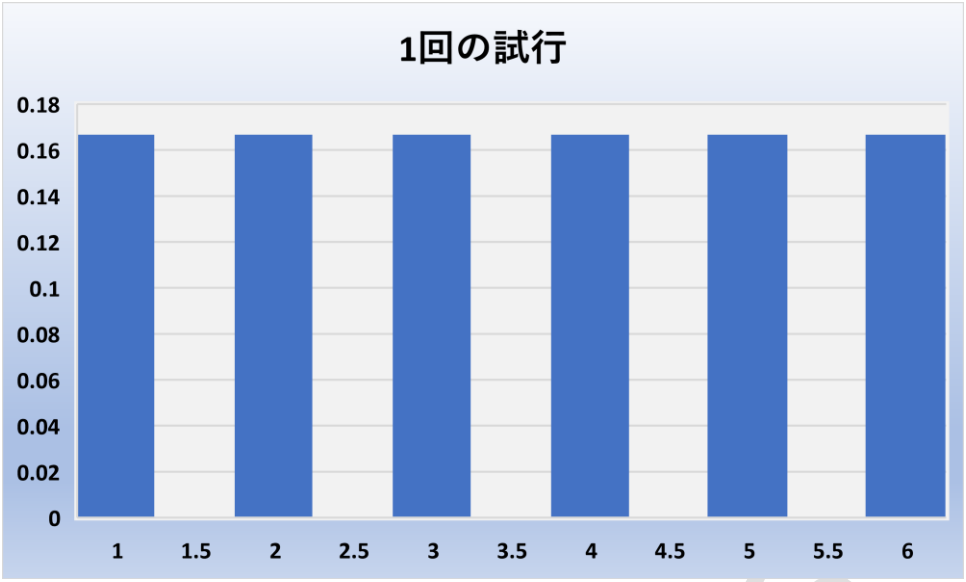
3.4 標本分布

母集団から n 個の標本を繰り返し抽出すると、それぞれのデータ集合は，同じ値になるとは限りません。したがって、これらのデータ集合を確率変数と見なすことができます。

標本平均や標本分散などは統計量です。それぞれの標本抽出によって得られるデータ(情報)の値は同じになるとは限らないため、それぞれの統計量は，標本抽出の際にそれぞれが異なる確率変数となります。したがって、それぞれの標本抽出で得られた統計量から分布が得られます。このような、統計量の確率分布を標本分布といいます。

例　サイコロを 1 回だけ振ることで得られた目の平均と 2 回だけ振ることで得られる目の平均を計算して、ヒストグラムにしてみましょう。

1 回だけの試行：さいころを一回だけ振ることを考えるとその出る目は 1, 2, 3, 4, 5, 6 のどれかです。したがってその目が出たときの平均はそれぞれ、1, 2, 3, 4, 5, 6 です。どの目も同じ確立で起こるとすると、平均が 1, 2, 3, 4, 5, 6 になる確率はそれぞれ 1/6 となります。



これは離散一様分布になります。

2 回だけの試行：２度サイコロを投げるときには最初の結果と、２番目の結果が同じになるとは限りません。最初が 1 の場合を考えると、２番目の結果は 1, 2, 3, 4, 5, 6 の可能性があります。そこでこれらの結果をつぎに様に表現します。

Xは1回目の試行の結果、Yは2回目の試行の結果です。それを$\{X,Y\}$で表しています。($\{X,Y\},Z$)のZは$Z=(X+Y)/2$です。

({X=1,Y=1},$\bar{x}$=1),({1,2},1.5),({1,3},2),({1,4},2.5),({1,5},3),({1,6},3.5)

({X=2,Y=1},$\bar{x}$=1.5),({2,2},2),({2,3},2.5),({2,4},3),({2,5},3.5),({2,6},4)

({X=3,Y=1},$\bar{x}$=2),({2,2},2.5),({3,3},3),({3,4},3.5),({3,5},4),({3,6},4.5)

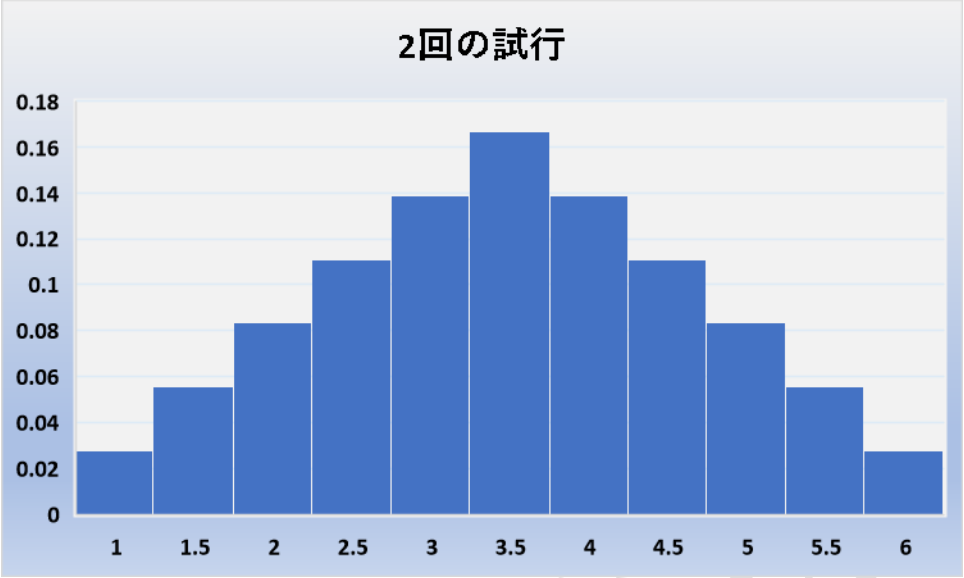
({X=4,Y=1},$\bar{x}$=2.5),({4,2},3),({4,3},3.5),({4,4},4),({4,5},4.5),({4,6},5)

({X=5,Y=1},$\bar{x}$=3),({5,2},3.5),({5,3},3),({5,4},4.5),({5,5},5),({5,6},5.5)

({X=6,Y=1},$\bar{x}$=3.5),({6,2},4),({6,3},3.5),({6,4},5),({6,5},5.5),({6,6},6)

そうすると平均の範囲は 1～６となります。また、平均の標本空間は

$\Omega=\{1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 5.0,\ 5.5,\ 6.0\}$となります。その頻度を数えて頻度図にしたものがつぎの図です。



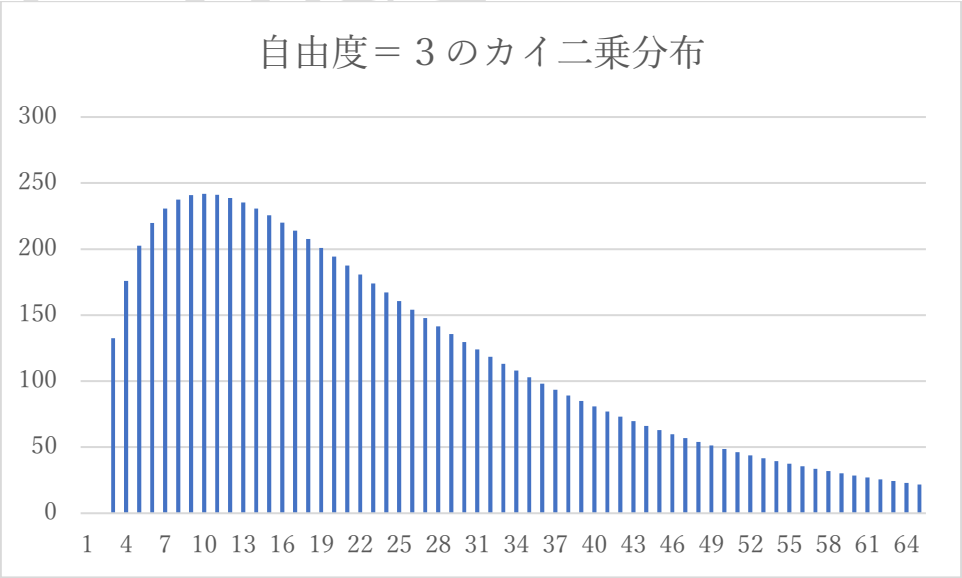
平均はどのように標本が得られるかでばらつくことが分かります。そのばらつき具合はベル型の分布をしています。サイコロを振る回数を増やしていくとこの分布は正規分布に近づいていきます。それは中心極限定理を説明しています。誤差が正規分布になるという説明が必要

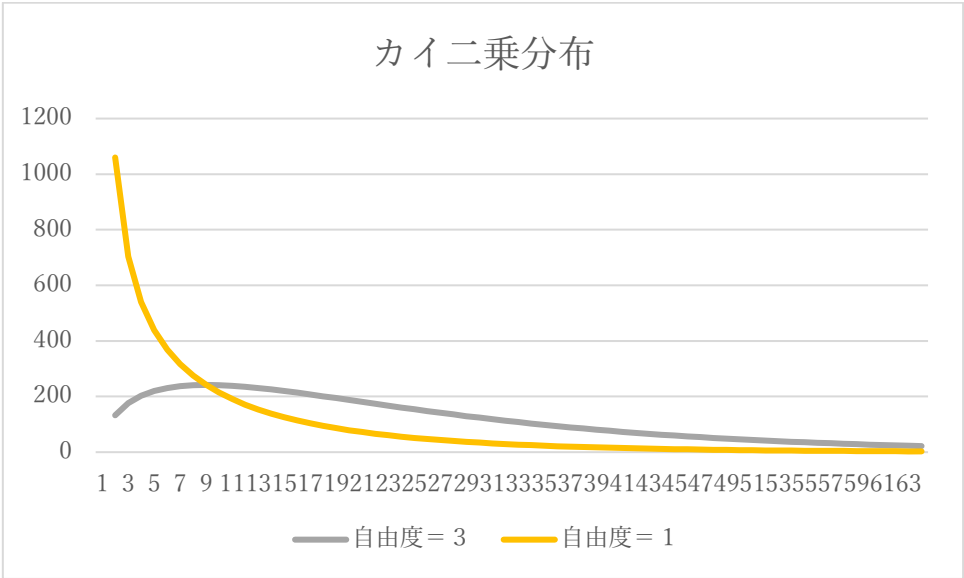
3.4.1 カイ二乗分布

確率変数 $X_1, X_2, \ldots, X_n$ が、互いに独立に、平均ゼロ、分散 1 の標準正規分布にしたがうとき、その統計量

$$Z = \sum_{i=1}^{n} X_i^2$$

がしたがう分布を自由度 n のカイ二乗分布といいます。カイ二乗分布は n が大きくなると正規分布にしたがいます。





3.4.2 t分布

確率変数が正規分布にしたがうとき、その母集団の平均と分散が既知であるというような場合は、まれです。ステューデントの t 分布は、標本の大きさが小さいときに、そのような母集団の平均を推定するのに用いられます。

確率変数 $X_1, X_2, \ldots, X_n$ は平均 μ 、分散 σ^2 の正規分布に独立にしたがいます。その標本平均が

$$\bar{X} = \frac{X_1 + X_2, \cdots + X_n}{n}$$

不偏分散が

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

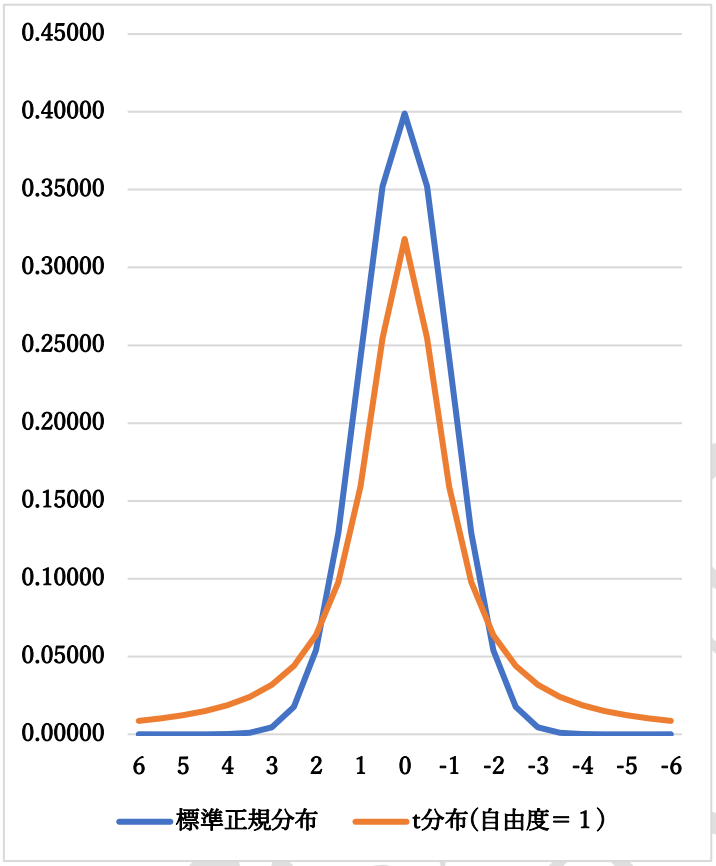
のとき、

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

は自由度 n の t 分布にしたがいます。

$\bar{X}$ の標準偏差 $S/\sqrt{n}$ を標本平均の標準誤差(standard error ， s.e.)といいます。

例：正規分布、t 分布などはエクセル関数を用いて描くことができます。自由度 1 の場合の t 分布と正規分布を比べてみます。



3.4.3 *F* 分布

カイ二乗分布にしたがう自由度が d_1 と d_2 の２つの確率変数 Z_1 と Z_2 の比は F 分布にしたがいます。

$$F = \frac{Z_1/d_1}{Z_2/d_2}$$

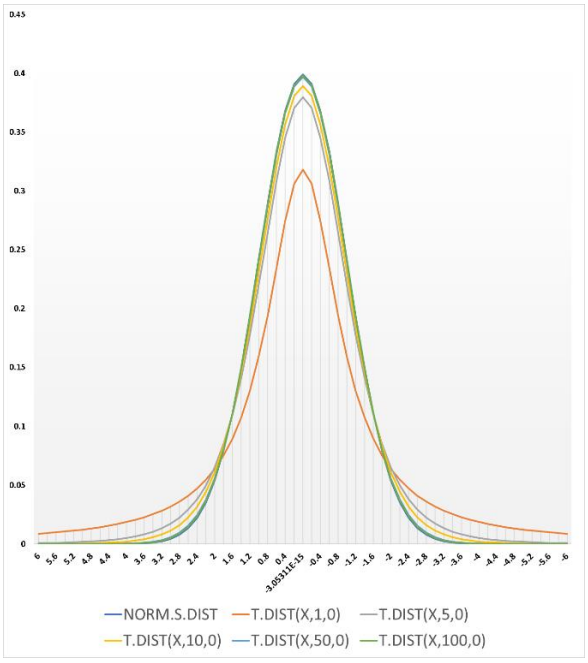
確率分布の分類

連続　vs　離散
(正規分布)　(2項分布)

母集団 vs 標本
(正規分布)　(t-分布)

練習問題 3.1:　t 分布について、ｎを 1, 5, 10, 50, 100 と変化させグラフに描いてみましょう。
練習問題 3.2: 練習問題 3.1 の結果からｔ分布の性質を記述してみましょう。
練習問題 3.3: 平均と期待値の違いを説明してみましょう。
練習問題 3.4: カイ二乗分布について自由度を変えて性質を調べてみよう
練習問題 3.5: カイ二乗分布と標本分散の関係についてエクセルで表示してみよう。
練習問題 3.6: 母分散と標本分散の関係についてエクセルで表示してみよう。

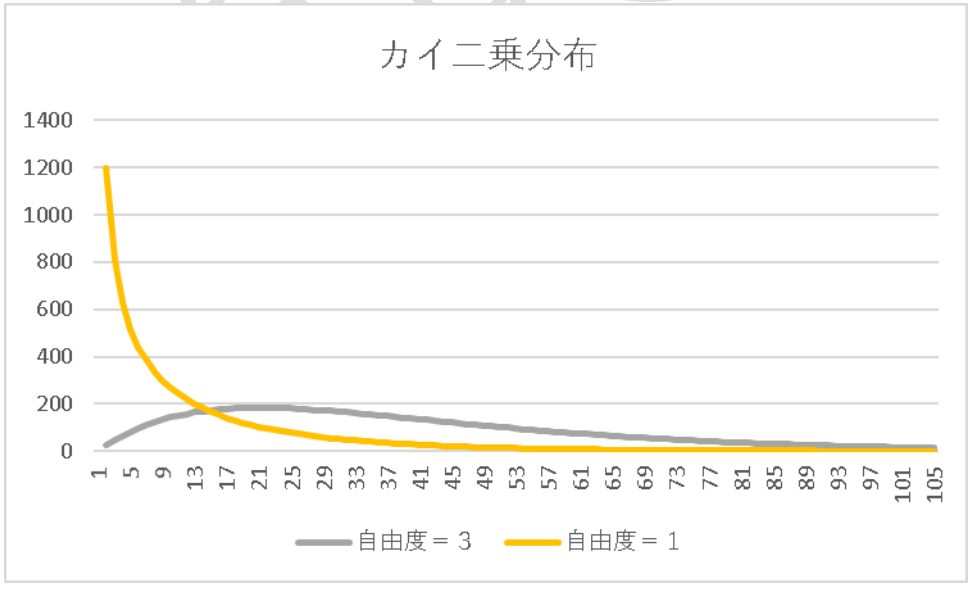
練習問題 3.1:　*t* 分布について、ｎを 1, 5, 10, 50, 100 と変化させグラフに描いてみましょう。



練習問題 3.2: 練習問題 3.1 の結果からｔ分布の性質を記述してみましょう。
自由度の数が大きくなると標準正規分布に近づいていきます。

練習問題 3.3: 平均と期待値の違いを説明してみましょう。
平均は手ものに得られたデータ、観測値、実験結果などから得られた平均的な結果でしか過ぎませんが、期待値は確率分布をもとに計算されています。

練習問題 3.4: カイ二乗分布について自由度を変えて性質を調べてみよう。



練習問題 3.5: カイ二乗分布と標本分散の関係についてエクセルで表示してみよう。

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

練習問題 3.6：母分散と標本分散の関係についてエクセルで表示してみよう。

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.$$