金融財務研究会　セミナー

時系列データ分析入門

（エクセル編）

2021 年 02 月 05 日

9:30～12:30

講師:森谷博之

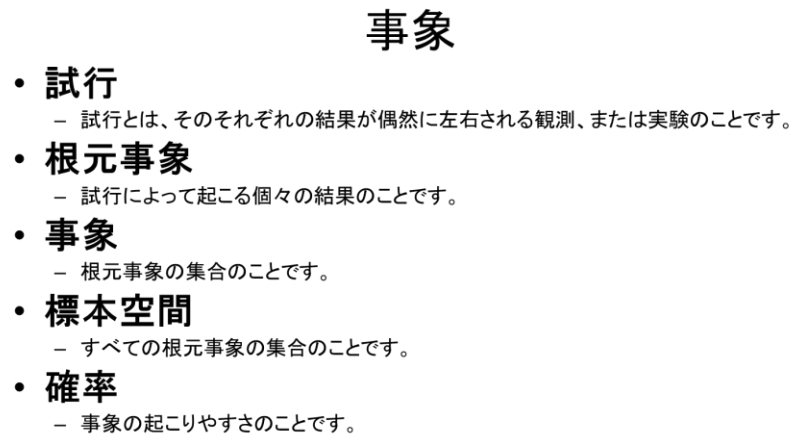
内容：

1. 統計学入門：確率と確率分布、母集団と標本、点推定と区間推定、線形回帰、最小二乗法
2. 経済時系列データ入門：移動平均と季節調整、トレンドと循環変動
3. 確率過程と時系列モデル：ランダムウォーク・モデル、自己回帰モデル、多項式モデル
4. モデルの最適化と予測：誤差二乗和、クロスバリデーション、尤度と情報量基準
5. 応用分野：リスク管理とバリューアットリスク

1. 統計学入門

確率と確率分布
５つのキーワードを中心に確率と確率分布について理解していきます。
事象についてのキーワード
たとえば、サイコロを投げるとき、硬貨を投げるときなどのように、その結果が偶然に左右されるような行為を試行といいます。そして、その結果の集合を事象、それ以上に分けられない事象を根元事象、すべての根元事象を標本空間、このような行為から得られる事象の起こりやすさを確率といいます。

試行
試行とは、そのそれぞれの結果が偶然に左右される観測、または実験のことです。
根元事象
試行によって起こる結果のことです。
事象
根元事象の集合のことです。
標本空間
すべての根元事象の集合のことです。
確率
事象の起こりやすさのことです。



確率には
- どれも同じような確からしさで起こるとする古典的な定義
- 頻度に基づく定義
- 主観に基づく定義
などがあります。

確率の定義

数学的には、確率は

- 任意の事象 $\boldsymbol{A}$ に対して　$0 \leq P(\boldsymbol{A}) \leq 1$
- 全事象Ωに対して $P(\Omega)=1$

と定義されます。

確率

- 確率はある事象の期待される割合を指します。
- その値はゼロから1までの値をとります。
- すべての事象の確率の和は1になります。

ある確率法則にしたがう変数を確率変数といいます。統計学的には確率分布から得られる変数です。

事象

- **試行**
  - サイコロを振る
- **根元事象**
  - {1},{2},{3},{4},{5},{6}
- **事象**
  - A={1，6}
- **標本空間(全事象)**
  - Ω={1，2，3，4，5，6}
- **確率**
  - {P(A)=2/6}

例：さいころの目の出る確率を均等とします。そうするとそれぞれの根元事象の出る確率は 1/6 になります。$\boldsymbol{A}$={1,2,3}、$\boldsymbol{B}$={3,4,5}とすると $\boldsymbol{A} \cap \boldsymbol{B}$ の確率はいくらでしょうか？また、$\boldsymbol{A}$ ={1,2,3,4}、$\boldsymbol{B}$={3,4,5,6}の場合はどうでしょうか？

$\boldsymbol{A}$={1,2,3}、B={3,4,5}のときは $\boldsymbol{A} \cap \boldsymbol{B}$ は{3}なので 1/6 になります。

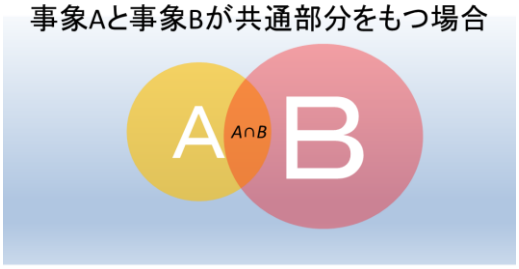
$\boldsymbol{A}$={1,2,3,4}、B={3,4,5,6}のときは $\boldsymbol{A} \cap \boldsymbol{B}$ は{3, 4 }なので 2/6 になります。

２つの事象 $\boldsymbol{A}$ と $\boldsymbol{B}$ の関係

- 和事象($\boldsymbol{A} \cup \boldsymbol{B}$)：$\boldsymbol{A}$ と $\boldsymbol{B}$ の少なくとも一方が起こる
- 積事象($\boldsymbol{A} \cap \boldsymbol{B}$)：$\boldsymbol{A}$ と $\boldsymbol{B}$ ば同時に起こる
- 余事象(^{c})：$\boldsymbol{A}^{c}$、A が起こらない事象 $\boldsymbol{B}^{c}$、$\boldsymbol{B}$ が起こらない事象
- 全事象(Ω)：標本空間全体の事象
- 空事象(∅)：何も起こらない事象

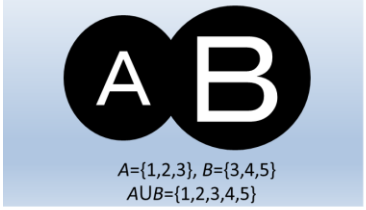
事象Aは標本空間の部分集合
事象A∈Ω

共通部分をもつ事象
事象Aと事象Bが共通部分をもつ場合

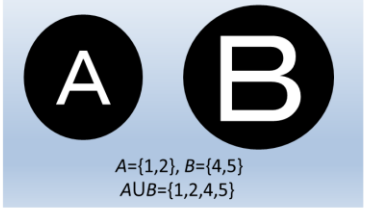


排反事象
事象Aと事象Bが共通部分をもたない場合

和事象
事象Aと事象Bの少なくとも一方が起きる。



$P(A \cup B)=P(A)+P(B)-P(A \cap B)$
$=3/6+3/6-1/6=5/6$



$P(A \cup B)=P(A)+P(B)$
$=2/6+2/6=4/6$

確率変数

変数Xがどのような値を取るかは事前にはわからないのですが、その値の確率が与えられるとき、変数Xを確率変数といいます。これらは

- 離散型確率変数
    - とびとびの値をとる確率変数
- 連続型確率変数
    - 連続的な値(実数値)をとる確率変数

に分類されます。

期待値と分散

$x_1, x_2, \cdots$の集合から得られる確率変数Xは離散型で、その期待値は、

$$E(X) \equiv \sum_i x_i f(x_i)$$

で表されます。fは離散型の分布です。
連続型の場合は、

$$E(X) \equiv \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

として定義されます。

期待値は分布の中心を表します。または「期待される」値という意味もあり、予測に近い意味の場合もあります。

離散型確率分布にしたがう確率変数について考えてみましょう。確率変数 X の実現値として得られるデータ、標本はヒストグラムとして要約することができます。$X=x_i$ の相対度数を N_i/N、その確率を p_i とすると、標本平均は $\bar{x} = \sum_i x_i N_i / N$ で表せます。期待値は $\mu = \sum_i x_i p_i$ です。

同じ議論が分散についても成り立ちます。また、確率変数 X の関数 $f(X)$ も確率変数なので、その期待値を考えることができ、同様の議論が成り立ちます。

独立な確率変数

一方の事象の起こる確率が、もう一方の事象の起こる確率に影響されないとき、それぞれの事象は独立といいます。これは事象 A,B について $P(A \cap B)=P(A)P(B)$ が成り立つことです。$\cap$ は A と B が同時起こることを表しています。$P()$ は確率を表します。たとえば、確率変数 X と Y が独立であると、その分散では $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)$ が成り立ちます。X と Y が独立でなければ $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X,Y)$ となります。

独立と無相関は混同しやすいのですが、独立は関係のことであり、無相関は平均的な性質のことです。したがって、独立ならば無相関になりますが、無相関であれば独立というわけではありません。

例：さいころをふる試行が独立だとします。サイコロを 2 回投げたときに事象 A、B を $A \in \{1,2,3\}$、$B \in \{3,4,5\}$ とすると $A \cap B$ の確率はいくらでしょうか？

1 回目の試行の結果は 1,2,3,4,5,6 のどれかです。したがってそれぞれの試行が独立であれば、その確率はそれぞれ 1/6 です。1,2 が出れば A です。3 は A と B に属するすべての $A \cap B$ です。4,5 は B です。 6 はどこにも属することがないので $\varnothing$ です。
つぎにそれぞれの試行は左から $\varnothing$、A、A、$A \cap B$、B、B、$\varnothing$ となります。したがって $A \cap B$ の確率は 1/6 です。つぎに 1 回目の試行が 3 として、2 回目の試行で出る目を考えてみます。これは 1,2,3,4,5,6 のどれかです。したがって、2 回目に $A \cap B$ が出る確率も 1/6 です。したがって、$A \cap B$ が 2 回続けて出る確率は 1/6・1/6=1/36 となります。これをさらに確かめてみましょう。すべての組み合わせを書いてみます。(1 回目の結果,2 回目の結果)とします。
$(A, A), (A, A), (A, A \cap B), (A, B), (A, B), (A, \varnothing)$
$(A, A), (A, A), (A, A \cap B), (A, B), (A, B), (A, \varnothing)$
$(A \cap B, A), (A \cap B, A), (A \cap B, A \cap B), (A \cap B, B), (A \cap B, B), (A \cap B, \varnothing)$
$(B, A), (B, A), (B, A \cap B), (B, B), (B, B), (B, \varnothing)$
$(B, A), (B, A), (B, A \cap B), (B, B), (B, B), (B, \varnothing)$

$(\varnothing, A), (\varnothing, A), (\varnothing, A \cap B), (\varnothing, B), (\varnothing, B), (\varnothing, \varnothing)$
すべてで 36 組あります。この中で$(A \cap B, A \cap B)$となっているのは 1 つなのでその確率は 1/36 です。

確率分布

確率変数がとりえる値とそれに対応する確率を確率分布といいます。

主な離散型確率分布

データが連続した値をとれないと、それは離散的なデータです。このようなデータの作る確率分布を離散型確率分布といいます。

離散一様分布

確率変数が離散値をとり、それぞれが一様に同じ確率をもつとき、それらは離散一様分布にしたがうといいます。確率は $1/n$ となります。

$f(x) = 1/N, x = 1, 2, \cdots, N$

ベルヌーイ分布

コインを投げたときには、一般的には表と裏しか出ません。このような事象が起きる行為をベルヌーイ試行といいます。この場合に、確率 p で表が出て、確率 1-p で裏が出るとき、その分布はベルヌーイ分布となります。結果が起こる確率は、一定かつ独立である必要があります。

[表,裏]、[1,0]、[上がる、下がる]など

ベルヌーイ分布の確率分布は
$$f(X = 1) = p, f(X = 0) = 1 - p$$
で与えられます。平均は p、分散は p(1-p)となります。ベルヌーイ分布にしたがう事象をくり返すと 2 項分布になります。

二項分布

二項分布とは、結果が成功か失敗、裏か表、上昇か下落というような 2 値で表される試行を n 回行ったときに得られる離散型確率分布です。それぞれの試行は独立でなければなりません。p と n について確率質量関数は
$$f(x) = {}_nC_x p^x (1-p)^{(n-x)} = \frac{n!}{k!(n-x)!} p^x (1-p)^{(n-x)}$$
となります。ここで、${}_n C_k$ は n 個から k 個を選ぶ組み合わせの数です。p は成功確率です。2 項係数を表しています。また、
$${}_nC_k = \frac{n!}{k!(n-k)!}$$

です。2 項分布では平均は $E(X)=np$、分散は $\mathrm{var}(X)=np(1-p)$となります。n=1 のとき、2 項分布はベルヌーイ分布になります。

主な連続型確率分布

確率変数 Xが連続な値をとるとき、その分布は連続型確率分布となります。これは全ての実数 x_i について、$X=x_i$ である確率がゼロである場合と同じです。

一様分布

確率変数の最小値と最大値を a, b としたときに、この区間の確率変数が生起する確率は等しくなります。$U(a,b)$と書くことがあります。

連続一様分布の確率密度関数は

$$f(x)=\begin{cases}\frac{1}{a-b} & for\ a\le x\le b\\ 0 & for\ others\end{cases}$$

となります。

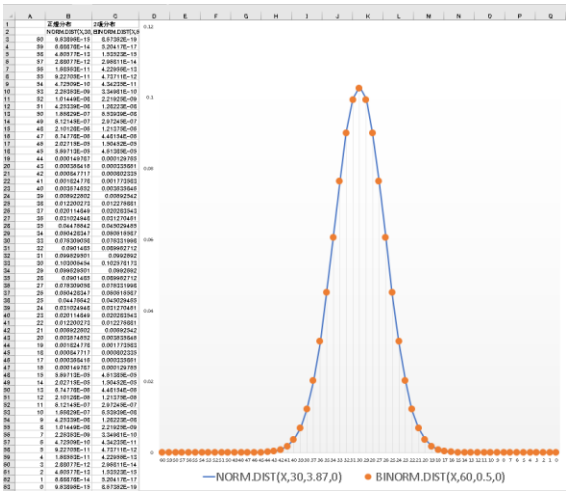
正規分布

確率変数 Xがとびとびの値ではなく連続していて、平均の周りに多く分布し、山が 1 つあり、山の裾が左右対称で、ベル型の形をしているような分布は正規分布です。正規分布では、分散は山の裾の広がり具合を表し、平均は分布の中心を示しています。正規分布の確率密度関数は平均と分散の関数として表されます。

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\bar{x})^2}{2\sigma^2}\right\}$$

ここで$\bar{x}$ は平均を、 σ^2 は分散を表します。

例：エクセルで、n=60, p=0.5 のときの 2 項分布と正規分布を描いてみましょう。



 # 練習問題 1　　エクセルを用いて乱数を発生させ、ヒストグラムを描きましょう。乱数は一様分布、ベルヌーイ分布、2 項分布、正規分布から発生させてみましょう。その際にデータ数を n=5,10,20,100 と変化させてみましょう。

Excel sheet: 分布.xlsx

母集団と標本

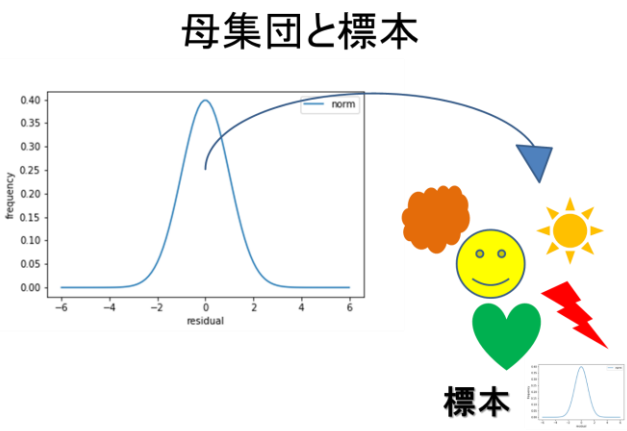
データ全体を母集団と呼び、その母集団から抽出されたデータを標本、またはサンプルと呼びます。この 2 つは明確に区別されます。そして、その標本の数を標本の大きさとか標本サイズといいます。

母集団と推定

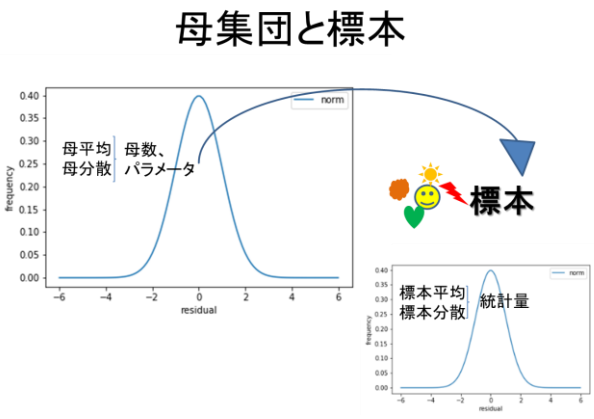
母集団とは、調査対象となっているデータのもととなる集合のことです。母集団というときには 2 つのタイプに分類されます。定義により母集団が確定している場合と、ある特定の仮想的集団を前提としている場合です。前者は選挙の時などの当選確率などが当てはまります。株価の予測などは後者に相当します。標本は、その母集団から抽出された個体の集合で、母集団の部分集合です。母集団について知りたいと思っているのですが、実は知ることができるのは標本についてであって母集団についてではありません。

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。**



母集団を特徴づける定数の値を知りたい場合があります。そのような定数を母数(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。



大数の法則と中心極限定理

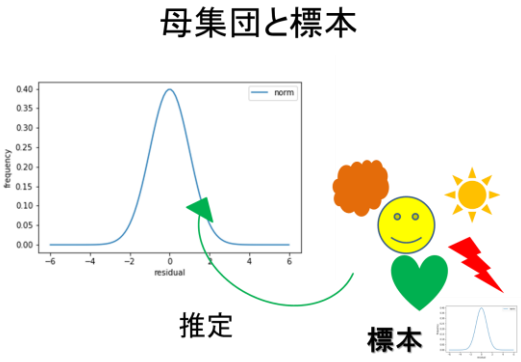
データ全体を母集団と呼び、その母集団から抽出られたデータを標本といいます。標本の大きさが大きくなるとそれにともない、標本から得られる統計量は真の統計量(母集団のもつ統計量)に近づいていきます。

母集団が平均値をもつときに、標本の大きさを大ききくしていくと、母集団のもつ平均値(母平均)、または真の平均値に標本の平均値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と理解できます。

推定の一般論

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。そこで確率の理論を用いて推測の信頼度を明らかにします。



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。

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。

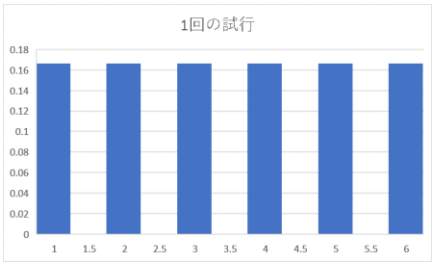
標本分布

母集団から n 個の標本を何度となく繰り返し抽出すると、それぞれの一組のデータは，同じ値の集合になるとは限らず、常に変化します。したがって、これらのデータの集合を確率変数と見なすことができます。

標本から計算される算術平均の結果（標本平均）や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)の値は同じとはならないので、それらから計算される統計量は，それぞれが同じ値になるとは限りません。したがって、それぞれの標本抽出で得られた統計量をヒストグラムで表すと分布が得られます。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それぞれの値の平均値と 2 回だけ振ることで得られた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：(X=1,$\bar{x}$=1),(2,2),(3,3),(4,4),(5,5),(6,6)



これは離散確率変数の一様分布になります。

2 回の試行：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)

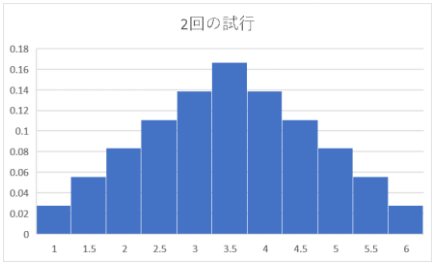
({X=2, Y =1},$\bar{x}$=1.5),({2,2},2),({2,3},2.5),({2,4},3),({2,5},3.5),({2,6},4)

({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)



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。これは試行の回数を増やしていくと正規分布に近づきます。これは中心極限定理を説明しています。

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。

t 分布

正規分布する確率変数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うとします。その標本平均が

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

不偏分散が

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

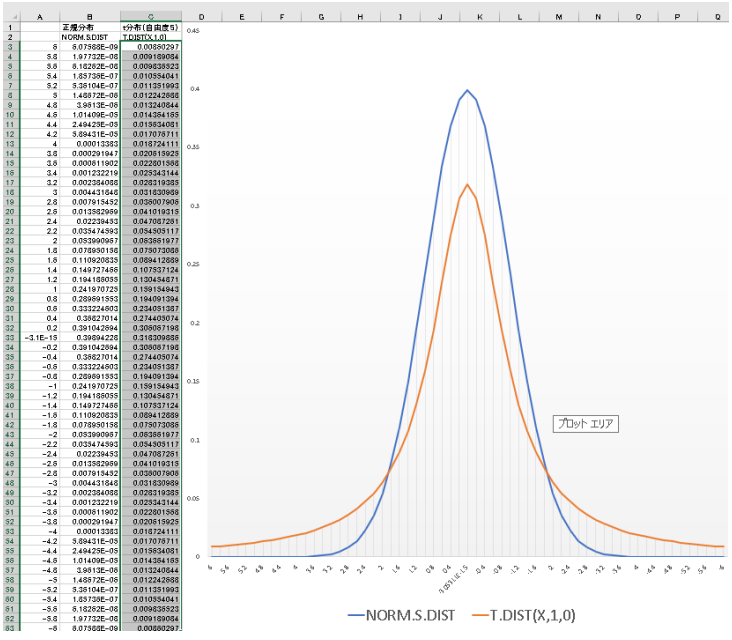
のとき、

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

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した。



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２　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。Excel sheet:分布.xlsx

統計的推定

確率変数と確率分布を基礎とする推測統計を学習します。ここでは、未知の母数を、観測値(得られたデータ)をもとにもとめていきます。これを統計的推定の問題といいます。得られたデータ Xから未知の母数を考えるとき、**その推定値とともに、その信頼度も考える必要があります。**つまり、推定値がどの程度の範囲にあるのだろうかと考える必要があるのです。母集団から手元にあるデータ $X_1, X_2, \ldots, X_n$ が得られるとすると母数の推定値は何度も計算することができ、かつその値はいつも変化する可能性があります。それらを確率変数ととらえるとき、推定量となります。

母数の推定値を表現する方法には 2 つあり、1 つの値としてとらえるのが点推定、上限、下限の間の区間としてとらえるのが区間推定です。

点推定

標本 $X_1, X_2, \ldots, X_n$ から得られる 1 つの値で推定したい母数を示す方法を点推定といいます。
- 平均、分散など

母数 θ に対してその推定量は θ に ^ をつけて表します。

標準誤差

母集団から得られた標本から統計量を推定するとき、そのばらつきの度合いを標準誤差といいます。これは標本のすべての組み合わせの標準偏差で表します。単に標準誤差といったときには平均のばらつきを表し、それは分散の推定量を標本の大きさで割り、その平方根をとったものです。推定量と標準誤差は組として示されます。

区間推定

標本から上限と下限の 2 つの値を求めて、その間に母数がふくまれるという表現の方法が区間推定です。

- 信頼区間

確率変数を X、区間の上限を $U(X)$、下限を $L(X)$、そして、母数を M とすると、

$$L(X) \leq M \leq U(X)$$

と表現します。

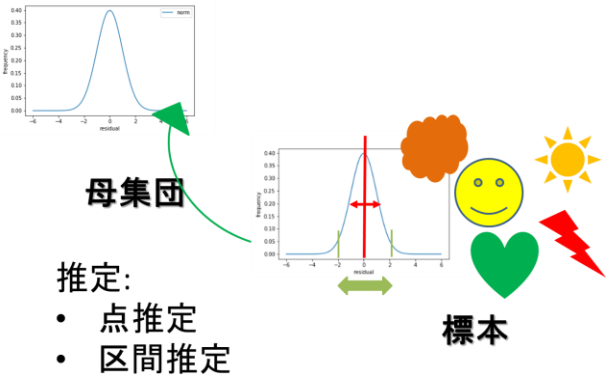
- 信頼係数

この信頼区間の中に母数が入る確率が信頼係数で 1- α で表します。したがって、

$$P(L(X) \leq M \leq U(X)) = 1 - \alpha$$

となります。

ここで問題となるのが $L(X)$, $U(X)$の決め方です。



平均の区間推定

平均値の区間推定を行ってみましょう。標本 $X_1, X_2, \cdots, X_n$ は独立に平均 μ，分散 σ^2 の正規分布にしたがうとします。

- 母分散が既知の場合
  - 正規分布を用います。
  - z_α: 確率 α における標準正規分布の臨界値

α は有意水準で、1-α が信頼係数となります。

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- 母分散が未知の場合
  - t 分布を用います。
  - $t_\alpha(n\text{-}1)$: 確率 α、自由度 n-1 の t 分布の臨界値

$$\bar{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

母分散の区間推定

分散の区間推定をする場合には、カイ二乗分布を用います。標本分散では

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$$

の関係があります。これを少し変形して、

$$z = \frac{\mathrm{var}(X)\cdot(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

とします。そうすると z は χ 二乗分布にしたがいます。信頼係数 1-α の信頼区間は

$$\frac{\sum(X_i - \bar{X})^2}{\chi_{\alpha/2}(n-1)} < \sigma^2 < \frac{\sum(X_i - \bar{X})^2}{\chi_{1-\alpha/2}(n-1)}$$

となります。標本の大きさが大きくなると信頼区間は狭くなります。

例：エクセルによるワインデータの主な要素の区間推定

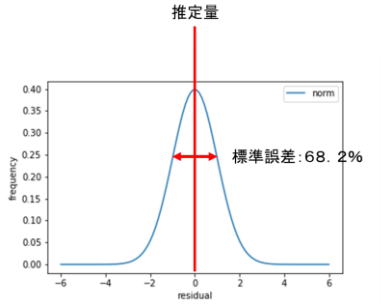
ワインデータ

要約統計量

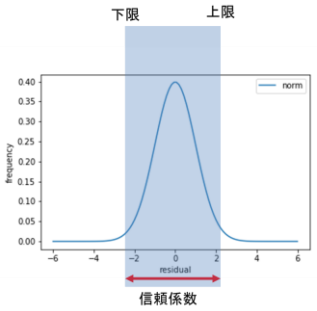
平均	8.31963727	0.52782051	0.27097561	2.538805503	0.087466542	15.8755472	46.46841776	0.99674668	3.3111132	0.658148843	10.4229831	5.636022514
分散	3.03141639	0.03206238	0.03794748	1.987897133	0.002215143	109.420672	1082.141524	3.562E-06	0.02383518	0.028732616	1.1356474	0.6521684
標準偏差	1.74109632	0.1790597	0.19480114	1.40992806	0.047065302	10.4604336	32.89591956	0.00188733	0.15438646	0.16950698	1.06566758	0.80756944
最大値	15.9	1.58	1	15.5	0.611	72	289	1.00369	4.01	2	14.9	8
最小値	4.6	0.12	0	0.9	0.012	1	6	0.99007	2.74	0.33	8.4	3
第1四分位範囲	7.1	0.39	0.09	1.9	0.07	7	22	0.9956	3.21	0.55	9.5	5
第3四分位範囲	9.2	0.64	0.42	2.6	0.09	21	62	0.99784	3.4	0.73	11.1	6
範囲	11.3	1.46	1	14.6	0.599	71	283	0.01362	1.27	1.67	6.5	5
歪度	0.98400508	0.6725605	0.31889339	4.542810386	5.678773525	1.25417641	1.514685264	0.06925907	0.19532987	2.428778802	0.86143898	0.217594626
尖度	1.13784198	1.23063553	-0.7865032	28.63049341	41.69291963	2.03746069	3.804575119	0.93557433	0.81532149	11.71493804	0.20012377	0.297613408
自由度	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598
信頼係数	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
上限	8.39129742	0.53519026	0.27899325	2.596835403	0.089403658	16.3060783	47.8223499	0.99682436	3.31746744	0.665125421	10.4668439	5.669260503
下限	8.24797713	0.52045076	0.26295797	2.480775604	0.085529425	15.4450161	45.11448562	0.996669	3.30475895	0.651172265	10.3791223	5.602784525

酒石酸濃度	酢酸濃度	クエン酸濃度	残糖濃度	塩化ナトリウム濃度	遊離二酸化硫黄濃度	総二酸化硫黄濃度	密度	pH	硫化カリウム濃度	アルコール度数	質
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5

- 点推定
  - 推定量
  - 標準誤差



- 区間推定
  - 上限、下限
  - 信頼係数



練習問題３　excel で正規乱数を発生させ基本統計量をとってみましょう。乱数の数を 10、100、1000、10000 といろいろと変えてやってみましょう。Excel sheet: 乱数 xlsx

統計的仮説検定

何度も何度も繰り返し、データを得て、まず仮説をたてます。しかし、得られる値が、いつでもその仮説からは想定出来ないのであれば、実際のデータは想定する仮説とは違うかも知れません。推測統計の手法の 2 つ目は統計的仮説検定です。これは、与えられた仮説に対して得られたデータからその正しさを統計的に判断する方法です。

仮説検定をなぜ統計的に行う必要があるのか？

実際に得られたデータを分析してみると、想定していた値とは異なる値が出るようなときがあります。データは実は想定していた特性をもっていないかもしれません。たまたま得られたデータがまれであるのかもしれないし、確率としてあり得るのかもしれません。判断が難しいときがあります。

ところで何度も何度も繰り返す観測の中で、確かに全部がおかしければ、それはおかしいという話になります。しかし、確率的に起きる可能性があるのであれば、それはおかしくないかもしれません。そして、それがどの程度の確率で起きればおかしいと考えたらよいのでしょう。そのような基準になる確率を**有意水準**といいます。そしてこのような判断のしかたが**統計的仮説検定**です。客観的に判断したいときに用います。

仮説検定の構造

帰無仮説と対立仮説を立て、検定に用いる統計量を定めて、標本をもとに帰無仮説の有無を判断します。

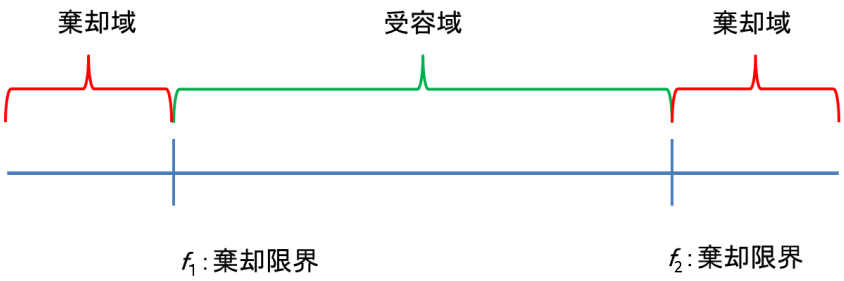
5.2.1 帰無仮説と対立仮説

「効果がある」とか「差がある」という検定をしたいときに、仮説検定では、「差がない」とか「効果がない」とかという仮説をたてます。それを帰無仮説といい、H_0 と書きます。それに対する仮説を対立仮説といいます。それを H_1 と書きます。実際には H_0 は棄却されてほしいのです。

- 対立仮説には片側検定と両側検定があります。

統計的仮説検定において確率変数 X_i が得られたとします。確率変数 X が取り得る全事象、または標本空間は、X_i の関数として表現されるある統計量 $F(X_i)$ によって 2 つの領域に分割されます。この統計量を検定統計量といい、確率変数は 2 つの領域に分けられます。1 つは受容域、もう一方は棄却域です。棄却域とは、帰無仮説が棄却される統計量がとる領域のことです。

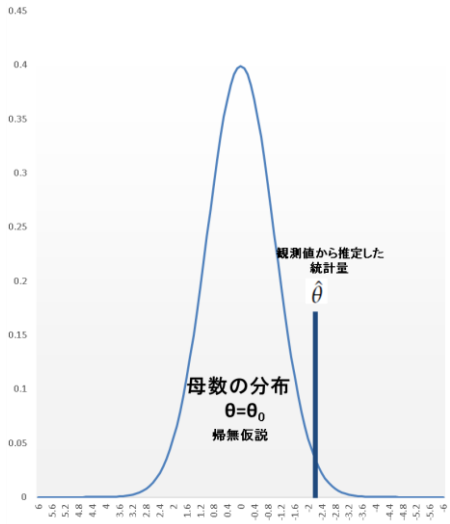
- 棄却域は片側検定の場合、$\{X_i \mid F(X_i) < f_1\}$、または $\{X_i \mid F(X_i) > f_1\}$、両側検定の場合、$\{X_i \mid F_i\ (X) < f_1\}$ または $\{X_i \mid F(X_i) > f_2\}$ などと表されます。f_1，f_2 を棄却限界といいます。



帰無仮説

確率変数 X_i が得られたとき、この X_i はもととなる集団、つまり母集団から得られたと仮定します。この母集団はある確率分布にしたがっている必要があります。この確率分布は統計量 θ により規定されます。それを θ_0 とします。したがって、X_i がこの母集団から得られたかどうかを特定するために統計量を用いることができます。その得られた値について帰無仮説(H_0)を立てます。

手元に得られたデータが x_i のとき, その標本から母数を推定できます。この推定量が得られる確率を母数の確率分布から推定します。この確率が有意水準 α より小さければ、H_0 に対してかなりまれな事象が起きたとして、帰無仮説を棄却します。そうでなければ棄却はしません。このような仮説を統計的仮説とよび、このような検定の方法を、統計的仮説の有意性検定といいます。



最も簡単な例は得られたデータの母集団の平均がゼロであることが既知の場合です。その場合の帰無仮説は $H_0: \mu = 0$ です。手元にあるデータから得られた平均の推定量が有意水準よりも小さく稀な事象であれば、帰無仮説は棄却されます。

帰無仮説と母数

得られた確率変数 Xi がある分布にしたがうとは、これらの母数により規定されている分布にしたがうということです。このとき、この母数が属する空間を母数空間と呼び Θ で表します。確率変数 X が母数 θ によって規定される分布にしたがうという仮説をたてるとき、この仮説を帰無仮説といいます。

帰無仮説が正しくないと結論付けるとき、帰無仮説を棄却するといいます。仮説が正しくないという判断には至らないとき、帰無仮説は「棄却するには十分ではない」とします。

対立仮説

帰無仮説が棄却されたときに採用される仮説が対立仮説です。

検定には２つのタイプがあります。１つは、差の大小を比べたいときです。購入したミルクの量が表示よりも少なそうだと思ったときには、少ないほうに大きくずれていることが判断の対象になります。そのような検定を片側検定といいます。

もう１つは明確に差があるか無いかを検定したいときです。その場合に、差の有無だけを判断したいのですから、比べたい両者を等しいとおいて、両者がたいへんに異なるときに、おかしいとすればよいのです。その際に、どちらの方向に大きくおかしいかは関係ありません。このような検定を両側検定といいます。

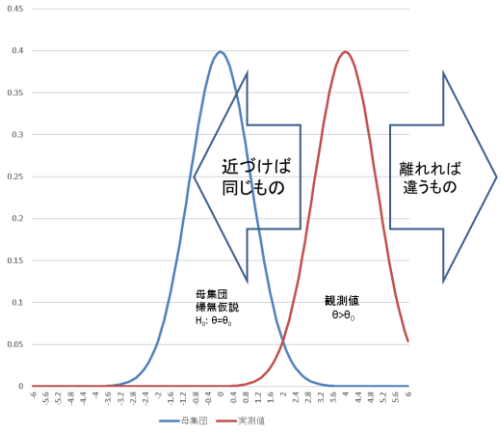
片側対立仮説と母数

$$H_1 : \theta < \theta_0 \text{ または } \theta > \theta_0$$

両側対立仮説と母数

$$H_1 : \theta \neq \theta_0$$

帰無仮説は比べたいものを等しいとして、対立仮説で片側検定、両側検定ので検定のタイプを指定します。



帰無仮説が棄却されるという意味

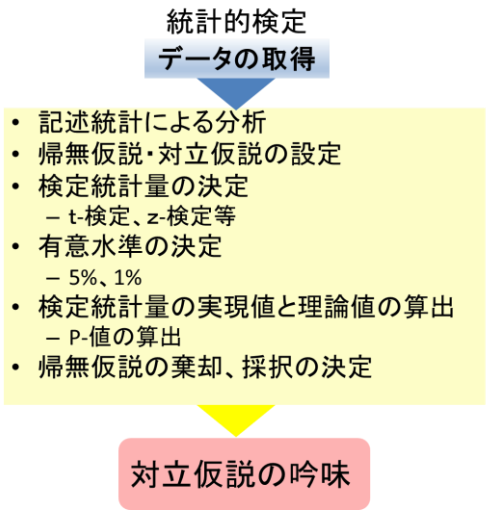
仮説検定では標本にもとづいて帰無仮説と対立仮説のどちらかを統計的な判断で選択しています。対立仮説が選択されたというときには、この選択が誤りであるという確率は α 以下であると保証されています。つまり、対立仮説が強く成り立っていると主張することができます。

もちろん、確率変数が、帰無仮説が仮定する条件を満たしていないために棄却された場合には、この限りではありません。

帰無仮説が棄却されないの意味

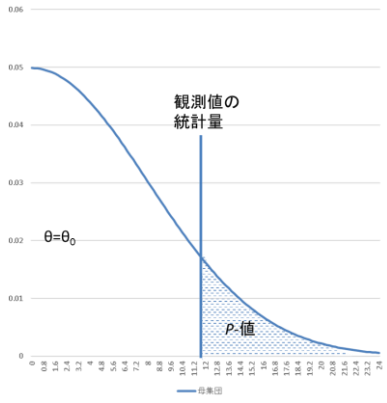
帰無仮説が棄却されなかったからといって、それを積極的に支持する理由にはなりません。それが誤りである確率が低いということはどこにも言及されていないのです。つまり、帰無仮説が棄却されないからといって、強く支持することはできないのです。帰無仮説が単に棄却される理由が不十分であったに過ぎないのです。ですので、帰無仮説を受容するとは言わずに、棄却するには十分ではないと表現するのです。

ここで注意が必要なのは確率変数が独立でないとか、一様でないとかという理由で棄却されてしまう場合があります。確率変数 X がある確率分布にしたがうとしたときは、確率変数は条件を満たしていなければなりません。確率変数 X がそれを満たさなければ、X はその確率分布にしたがいません。そうすると統計的検定には意味がなくなります。



p 値(有意確率)

p 値は、観察されたデータより極端な事象が現れる確率を表します。実現値の平均を $\bar{x}'$ とすると、p 値は $P(\bar{x} \geq \bar{x}')$ と書けます。観測されたデータをもとに棄却される有意水準を明確にできるので、固定した有意水準と観測データにもとづいた検定統計量を示すよりも、情報量が豊富であると考えられます。



一般に、

関係 (p は p 値を表す)	解釈
$0.01 \geq p$	帰無仮説を棄却する。
$0.1 \geq p \geq 0.01$	帰無仮説を棄却するに足る。
$p \geq 0.1$	帰無仮説を棄却するのは難しい。

「帰無仮説を棄却する」とは 0. 01 以下の確率でしか起こらないことが起こった、ということです。

「帰無仮説の棄却は難しい」は棄却するに十分な証拠がないということです。
統計学の目的は極力誤った判断を減らすことにあります。

両側検定の場合の p 値の使用には注意が必要です。片側検定の p 値の 2 倍とする方法と、観測データから算出した出現確率よりも、出現する確率が小さい事象の確率とする方式があります。後者の場合には、左右対称でない分布では前者と異なる結果となります。

第一種の誤り

第一種の誤りとは、帰無仮説が正しいときに棄却してしまう誤りことです。この確率を α で表します。

第二種の誤り

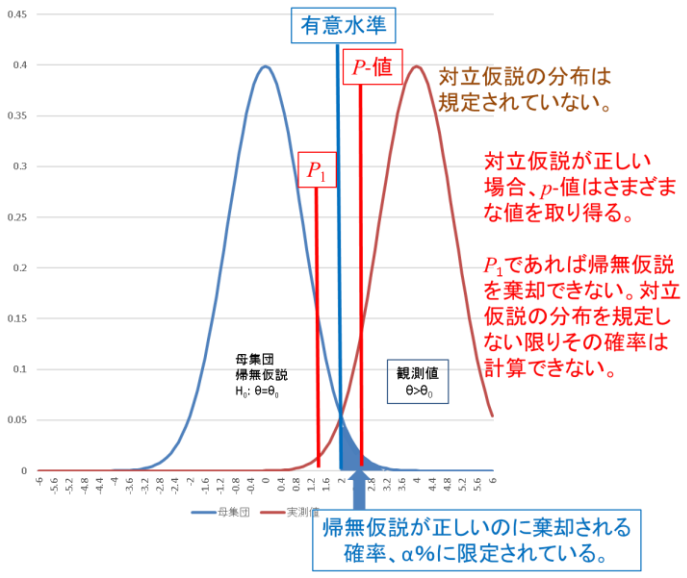
対立仮説が正しいときに帰無仮説を受容してしまう誤りのことです。この確率を β で表します。対立仮説が正しい時に正しい確率は $1-\beta$ になります。β は帰無仮説、対立仮説のかなでは規定されていません。++

有意水準

第一種の誤りと第二種の誤りの両方を同時に小さくする必要があります。ところがこの二つは二律背反の関係にあります。そこで一般的には、第一種の誤りを一定以下の α に抑えながら、第二種の誤りを小さくするよう試みられます。この α は有意水準と呼ばれ、第一種過誤の確率を表します。しかし、これは帰無仮説が正しくない確率ではないことに注意が必要です。β は帰無仮説からも対立仮説からも得られるものではありません。別途設定する必要があります。

また、α を小さくすると β が大きくなるのですから、$1-\beta$ は小さくなってしまいます。

	H_0 が正しい (H_1 が誤り)	H_1 が正しい (H_1 が誤り)
H_0 の棄却	第一種の過誤 (α)	正しい判断 ($1-\beta$)
H_0 の採択	正しい判断 ($1-\alpha$)	第二種の過誤 (β)



線形回帰モデル

２変量 Xと Yの関係を分析するために、Yを Xの一次式の形でとらえる方法があります。これを線形単回帰といいます。ここで、Xは説明変数、Yを被説明変数とよびます。

$$Y = \alpha + \beta X$$

を回帰直線、α, β を回帰係数と呼びます。これらの回帰係数は固定された母数ですが、決して既知ではありません。また、Xは非確率変数です。確率変数の場合には誤差項とは独立である必要があります。

$$E(Y|X_i) = \alpha + \beta X_i$$

となり、$E(|)$は条件付き平均です。$E(|)$は確定的な、またはシステマティックな部分です。
これを母集団線形回帰モデルといいます。

また、

$$Y_i - E(Y|X_i) = \epsilon_i$$

と書き、ϵ_iは攪乱項です。これは確率的または非システマティックな部分を表現しています。

確率的誤差項のもつ意味

ϵ_iはこのモデルでは説明できない部分です。ではなぜϵ_iが必要なのでしょうか。主に 7 つの理由があります。

- 理論のあいまいさ

理論がまだ確立されていない。または理論で説明できない部分が残っている状態です。

- データの不完全性

十分なデータ、情報があたえあられていない場合です。

- 主変数とそれ以外の変数

いくつかの変数による影響が確率的な効果として表れています。

- 内在的な確率的要素
- データに含まれるノイズ

データにはノイズが含まれていてそのノイズを誤差項で表現しています。

- 単純性の原理

回帰モデルを単純な形にしておきたい。

- 間違った関数形式

もともとのモデル自体に誤りがある。

線形単回帰モデル

母集団は一般に手に入りません。したがって、母数も未知です。そこで、母集団回帰式と同様に標本回帰式を導入します。それを

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

と書き、$\hat{Y}_i$は $E(Y|X_i)$の推定量、$\hat{\alpha}$, $\hat{\beta}$は α, β の推定量です。

これらは統計量です。統計量とは、与えられた標本データから得られる情報を用いて母数を推定する方法、関数です。推定量は確率変数ですが、推定値は実際のデータから計算された値です。

推測統計量を用いると

$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\epsilon}_i$

と書きます。$\hat{\epsilon}_i$を回帰残差項といいます。これは線形単回帰モデルです。$\hat{\epsilon}_i$ は互いに独立に正規分布 $N(0, \sigma^2)$にしたがいます。

最小二乗法

母集団線形回帰関数は

$Y_i = \alpha + \beta X_i + \epsilon_i$

でした。また、標本線形回帰関数は

$$\begin{aligned} Y_i &= \hat{\alpha} + \hat{\beta} X_i + \hat{\epsilon}_i \\ &= \hat{Y}_i + \hat{\epsilon}_i \end{aligned}$$

です。したがって、

$$\begin{aligned} \hat{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} X_i \end{aligned}$$

となり、$\hat{\epsilon}_i$の平方和は

$$\begin{aligned} \sum \hat{\epsilon}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \end{aligned}$$

となります。α と β は、これを最小になるように推定して得ます。

すると、

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

が得られます。

このようにして求めた統計量を最小二乗統計量といいます。

- これらの統計量は標本から計算されていて簡単に手に入ります。

- これらは点推定量です。

- 標本回帰直線が得られます。

    - 標本回帰直線は $\bar{Y}$ と $\bar{X}$ の平均を通ります。
    - $\bar{\hat{Y}} = \bar{Y}_i$ です。
    - $\hat{\epsilon}_i$=0 ですので、$Y_i - \bar{Y} = \beta (X_i - \bar{X})$が得られます。

$\bar{X}$は X_iの標本平均です。

古典的最小二乗法の仮定

統計モデルは幾つかの仮定のもとに成り立ちます。線形回帰モデルの仮定をここで列挙します。

1. 回帰関数は線形でなければならない。
2. X_iは確率変数であってはならない。確率変数の場合には誤差項と独立でなければならない。
3. ϵ_iの平均はゼロである。
4. ϵ_iの分散は一定である。
5. ϵ_iとϵ_jの共分散はゼロである。
6. 観測値の数nは説明変数の数よりも多い。
7. X_iは一定であってはならないが、外れ値などがあってもならない。

分析結果の評価

母数が未知であれば、推定したモデルがどれくらい正しいのかがわかりません。そこで、このモデルがどれくらい信頼できるのかを確かめる必要があります。

推定の信頼性

単純な標本回帰直線には、$\hat{\alpha}$, $\hat{\beta}$, $\hat{\epsilon}_i$の 3 つの推定量があります。これがどの程度信頼できるのかを見ていきます。

標準誤差

標準誤差(Standard Error,Std Err)は、母数の推定値と未知の母数との差です。これは統計量の正確さの測度です。回帰係数 α, β の推定値の標準誤差は

$$se(\hat{\alpha}) = \sqrt{\frac{\sum X_n^2}{N\sum(X_n - \bar{X})^2}}\sigma$$

$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum(X_n - \bar{X})^2}}$$

で与えられます。ここで σ は、次式で与えられるϵ_iの分散で推定の標準誤差と呼ばれます。

$$\sigma^2 = \sum\frac{\epsilon_i^2}{N} = \frac{\sum([Y_n - E(Y|X_n)]^2}{N}$$

この母数は未知ですので、その推定量は

$$\hat{\sigma}^2 = \sum\frac{\hat{\epsilon_i}^2}{N-2} = \frac{\sum([Y_n - E(Y|X_n)]^2}{N-2}$$

です。$E(\hat{\sigma}^2) = \sigma^2$です。標準誤差は、統計量のバラツキ具合、つまり精度の測度であり、標準誤差の推定値は標本回帰直線とデータとの適合度(goodness of fit)の目安となります。

決定係数

R^2(R-squared)は、標本回帰直線がデータをどの程度説明しているかを示す指標、適合度(goodness of fit)を表す測度として知られています。それは、決定係数(coefficient of determination)と呼ばれ、次式で与えられます。

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\mathrm{ESS}}{\mathrm{TSS}}$$

ここで$\bar{Y}$は Yの標本平均です。決定係数が 1 に近いほど、相対的なバラツキは少なくなります。

自動調整済み決定係数(Adj R-squared)

これは説明変数の数の効果を考慮した係数です。説明変数の数が多くなると決定係数は良くなる傾向があります。したがって、その分を調整します。

回帰係数の区間推定と仮説検定

- H_0: $\mu = 0$, H_1: $\mu \neq 0$ とします。

回帰係数の母平均をゼロとし、標本平均を $\hat{\theta}$ とすると有意水準 α の信頼区間は
$0 - t_{\alpha/2} se(\hat{\theta}) \leq \hat{\theta} \leq 0 + t_{\alpha/2} se(\hat{\theta})$
となります。これを採択域(the region of acceptance)といい、この外側の領域を棄却域(the reion(s) of rejection)といいます。これを危険域(the critical region(s))と呼ぶことがあります。そうすると大きな|t|値は帰無仮説の棄却域にいることになります。よって、

$$t_{\alpha/2} = \frac{\hat{\theta}}{se(\theta)}$$

ですから、$t_{\alpha/2}$ から p 値(p-value)を計算できます。

一般に、

関係 (p は p 値を表す)	解釈
$p <= 0.01$	帰無仮説を棄却する。
$0.01 < p <= 0.1$	帰無仮説を棄却するに足る。
$0.1 < p$	帰無仮説を棄却するのは難しい。

と解釈されます。

「帰無仮説を棄却する」とは 0.01 以下の確率でしか起こらないことが起こった、ということです。
「帰無仮説の棄却は難しい」は棄却するに十分な証拠がないということです。
このように表現する理由として、統計学の目的は極力誤った判断を減らすことにあるからです。

分散分析

Y_i の平均と $\hat{Y}$ の平均が同じ母集団から得られたのかどうかを検定します。Y_i の母平均(μ)と $\hat{Y}$ の母平均(μ_e)の差が有意であるかどうかを調べます。Y_i と $\hat{Y}$ をそれぞれ群ととらえます。そしてこの 2 つの群のそれぞれの平均に違いがあるかどうかを調べます。これらの標本の群の間で平均値に差があるからといって、母平均にも差があるとは限りません。Y とその予測値の 2 群の母平均に違いがなくても無作為に抽出された標本には違いがあるのかもしれません。このようなときに分散分析が用いられます。分散分析と名前がついていますが、実は平均値の差の検定です。

分散分析では非説明変数の全体平方和(TSS; 分散×自由度)を回帰モデルで説明できる部分とできない部分に分解します。説明できる部分が回帰平方和(ESS)、できない部分が残差平方和(RSS)です。したがって、TSS=ESS+RSS です。

- $H_0 : \mu_t = \mu_e$,
- $H_1 : \mu_t \neq \mu_e$
  - μ_t: 被説明変数の母平均
  - μ_e: 予測値の母平均

$$\begin{aligned} \text{TSS 全体平方和} &= \sum(Y_i - \bar{Y})^2 \\ \text{ESS 回帰平方和} &= \sum(\hat{Y}_i - \bar{Y})^2 \\ \text{RSS 残差平方和} &= \sum(\hat{Y}_i - Y_i)^2 \\ \text{ESS} &= \text{TSS-RSS} \end{aligned}$$

ここで回帰モデルで説明できる部分とできない部分を自由度て調整して

$$F = \frac{\mathrm{ESS}/df_p}{\mathrm{RSS}/df_r}$$

とします。df_p は群間の自由度です。この場合は 2-1=1 です。df_r は群内の自由度です。この場合は郡内のデータ数－2です。回帰分析では RSS が小さいと推定モデルは Y_iをよくとらえていることになります。つまり、回帰係数に意味があるということです。そうすると ESS は大きくなり TSS に近づきます。回帰係数に意味がなければ、RSS は TSS に近づきます。その際には ESS は小さくなります。それは Y_iの平均と$\hat{Y}_i$の平均が同じ母集団から得られたということです。したがって、帰無仮説はすべての回帰係数がゼロであると書き換えることができます。R^2=ESS/TSS を用いると

$$F = \frac{R^2}{(1-R^2)}\frac{df_r}{df_p}$$

と変形できます。

$F \geqq F_\alpha(df_p, df_r)$ならば帰無仮説を棄却します。
$F < F_\alpha(df_p, df_r)$ならば帰無仮説を棄却しません。
また、p-値を用いることもできます。

例　年齢と性別の同じ人の体重と身長の関係が

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	身長	113	111	118	112	114	118	124	118	122	119	118	116	118	114	119	117	119	119	120
2	体重	20	19	20	16	19	23	22	20	23	19	22	22	21	19	22	20	20	24	20

のように与えられたとして、エクセル分析ツールの回帰分析を用いて分析します。
結果はつぎのようになりました。

	D	E	F	G	H	I	J	K	L
1	概要								
2									
3	回帰統計								
4	重相関 R	0.632114							
5	重決定 R2	0.399568							
6	補正 R2	0.364249							
7	標準誤差	2.645181							
8	観測数	19							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	79.15657	79.15657	11.31296	0.003688			
13	残差	17	118.9487	6.996982					
14	合計	18	198.1053						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	94.54153	6.798192	13.90686	1.02E-10	80.1986	108.8845	80.1986	108.8845
18	X 値 1	1.106678	0.329028	3.363474	0.003688	0.412489	1.800866	0.412489	1.800866

回帰統計を見ると、決定係数 R2 と調整済み R2 はそれぞれ 0.40 と 0.36 程度です。特に大きいとはいえません。
分散分析表の有意 F は F検定の p-値のことです。小さいので、Yの平均とその推定値の平均が同じ母集団であるという帰無仮説を棄却します。つまり、回帰係数は有意であるという結論です。その下の表はそれぞれの回帰係数がゼロ

であるかないかの仮説検定をしています。P-値はゼロと判断できる程度に小さいので、帰無仮説は棄却され、回帰係数は有意であることが分かります。

練習問題 4：上述の例の有意 F について、エクセルシートを使って自分で求めてみましょう。
excel sheet: 分散分析.xlsx

2. 経済時系列データ分析

経済時系列データの多くは、内外の公官庁により作成されています。そして、そのほとんどが季節調整値です。季節調整前のデータを原系列といいます。季節調整値とはさまざまな基礎的なデータをもとに季節性と感がられる要素を修正した値のことです。これは季節変動よりも長期的変動や中期的な景気循環に興味があるからです。

移動平均と季節調整

季節調整の最も基本となる方法では、時系列データの前後の数値の算術平均をとります。これは時系列の平滑化と呼ばれる操作の 1 つです。代表的な平滑化の方法として移動平均があります。

原系列 $x_1, x_2, x_3, \ldots, x_T$ を$\{x_t\}$とします。ここで $1 \leqq t \leqq T$ で T は時系列の長さです。時刻 t のデータ x_t の前後 k 期ずつデータを取ると計 $2k+1$ 個のデータを用います。

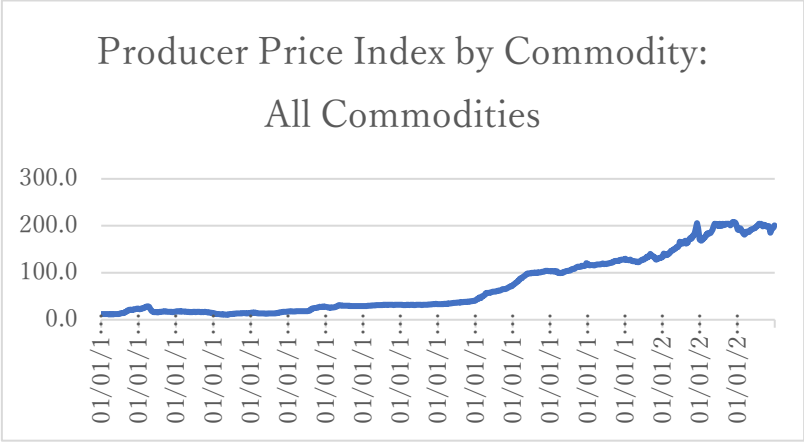
$$y_t=(x_{t-k}+\ldots+x_t+\ldots x_{t+k})/(2k+1)$$

これはそれぞれのデータを均等に扱った単純平均です。

これを一般化して、それぞれのデータに重みを用いた加重平均を用いることもあります。

エクセルで移動平均を計算する方法は 2 つあり、 1 つは分析ツールを用いる方法とエクセル関数を用いる方法です。

例　季節調整の無いデータ　　FRED よりダウンロード　生産者物価指数：全商品、月次データ
EXCEL シート：econometrics.xlsx

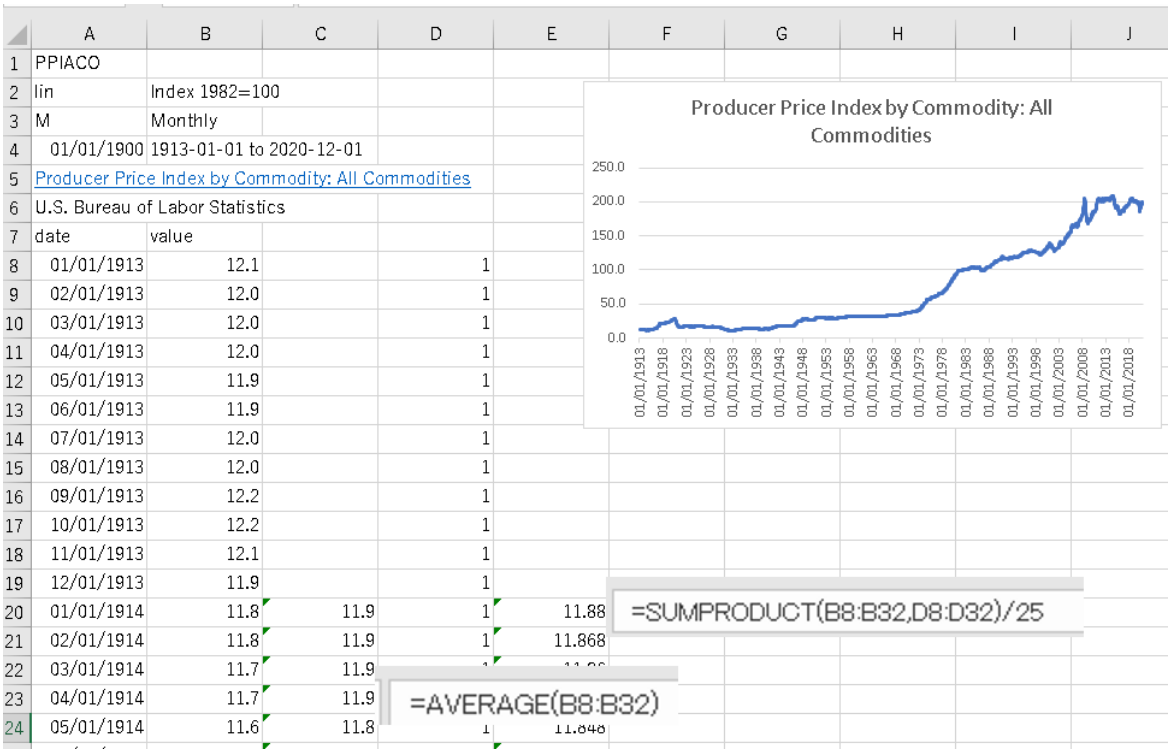


	A	B	C	D
1	PPIACO			
2	lin	Index 1982=100		
3	M	Monthly		
4	01/01/1900	1913-01-01 to 2020-12-01		
5	Producer Price Index by Commodity: All Commodities			
6	U.S. Bureau of Labor Statistics			
7	date	value		
8	01/01/1913	12.1		
9	02/01/1913	12.0		
10	03/01/1913	12.0		
11	04/01/1913	12.0		
12	05/01/1913	11.9		
13	06/01/1913	11.9		
14	07/01/1913	12.0		

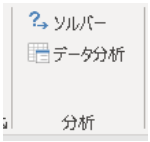
エクセル関数をつかった移動平均

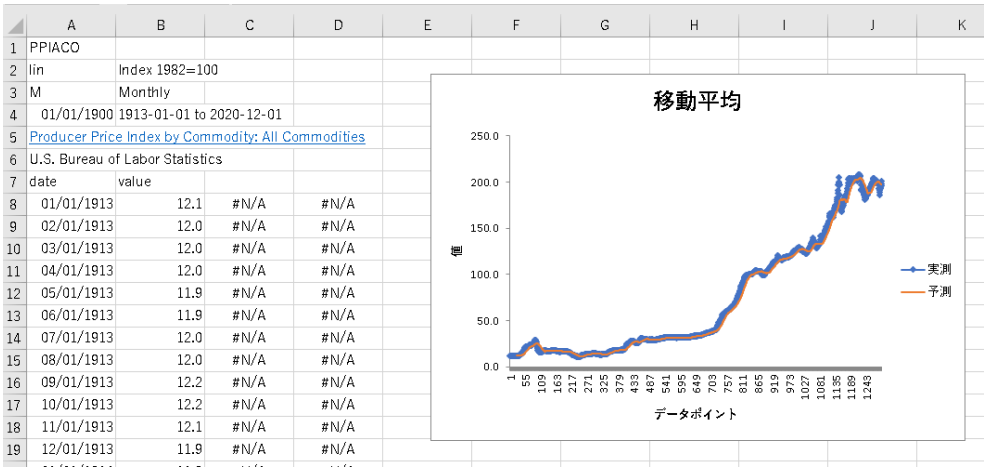
AVERAGE

SUMPRODUCT



データ分析ツールを使う例





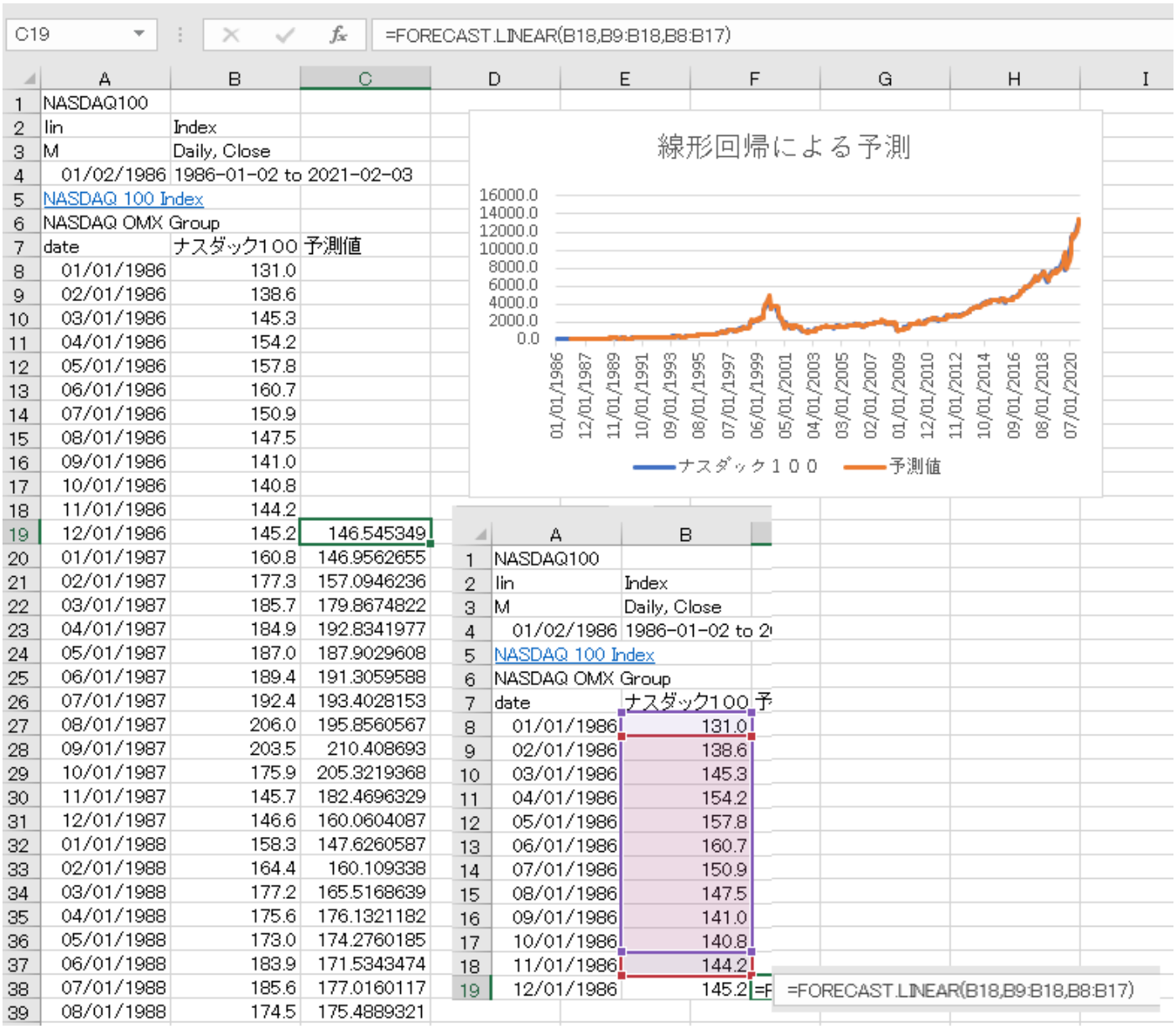
トレンドと循環変動

時間の経過とともに経済時系列が一方向に増加あるい減少する動きが見られるとき、その傾向をトレンドといいます。上向きの傾向を上昇トレンド、下向きの傾向を下落トレンドといいます。これは線形単回帰モデルを用いて表現されます。

$$x_t = a + bt$$

t は時間を表し、a,b は回帰係数です。

Excel シート: time_series_data_analysis.xlsx



	A	B	C
1	NASDAQ100		
2	lin	Index	
3	M	Daily, Close	
4	01/02/1986	1986-01-02 to 2021-02-03	
5	NASDAQ 100 Index		
6	NASDAQ OMX Group		
7	date	ナスダック100	予測値
8	01/01/1986	131.0	
9	02/01/1986	138.6	
10	03/01/1986	145.3	
11	04/01/1986	154.2	
12	05/01/1986	157.8	
13	06/01/1986	160.7	
14	07/01/1986	150.9	
15	08/01/1986	147.5	
16	09/01/1986	141.0	
17	10/01/1986	140.8	
18	11/01/1986	144.2	
19	12/01/1986	145.2	146.545349
20	01/01/1987	160.8	146.9562655
21	02/01/1987	177.3	157.0946236
22	03/01/1987	185.7	179.8674822
23	04/01/1987	184.9	192.8341977
24	05/01/1987	187.0	187.9029608
25	06/01/1987	189.4	191.3059588
26	07/01/1987	192.4	193.4028153
27	08/01/1987	206.0	195.8560567
28	09/01/1987	203.5	210.408693
29	10/01/1987	175.9	205.3219368
30	11/01/1987	145.7	182.4696329
31	12/01/1987	146.6	160.0604087
32	01/01/1988	158.3	147.6260587
33	02/01/1988	164.4	160.109338
34	03/01/1988	177.2	165.5168639
35	04/01/1988	175.6	176.1321182
36	05/01/1988	173.0	174.2760185
37	06/01/1988	183.9	171.5343474
38	07/01/1988	185.6	177.0160117
39	08/01/1988	174.5	175.4889321

	A	B
1	NASDAQ100	
2	lin	Index
3	M	Daily, Close
4	01/02/1986	1986-01-02 to 2
5	NASDAQ 100 Index	
6	NASDAQ OMX Group	
7	date	ナスダック100 予
8	01/01/1986	131.0
9	02/01/1986	138.6
10	03/01/1986	145.3
11	04/01/1986	154.2
12	05/01/1986	157.8
13	06/01/1986	160.7
14	07/01/1986	150.9
15	08/01/1986	147.5
16	09/01/1986	141.0
17	10/01/1986	140.8
18	11/01/1986	144.2
19	12/01/1986	145.2

=FORECAST.LINEAR(B18,B9:B18,B8:B17)

また、原系列の階差、差分を取る場合もあります。これを階差法といいます。

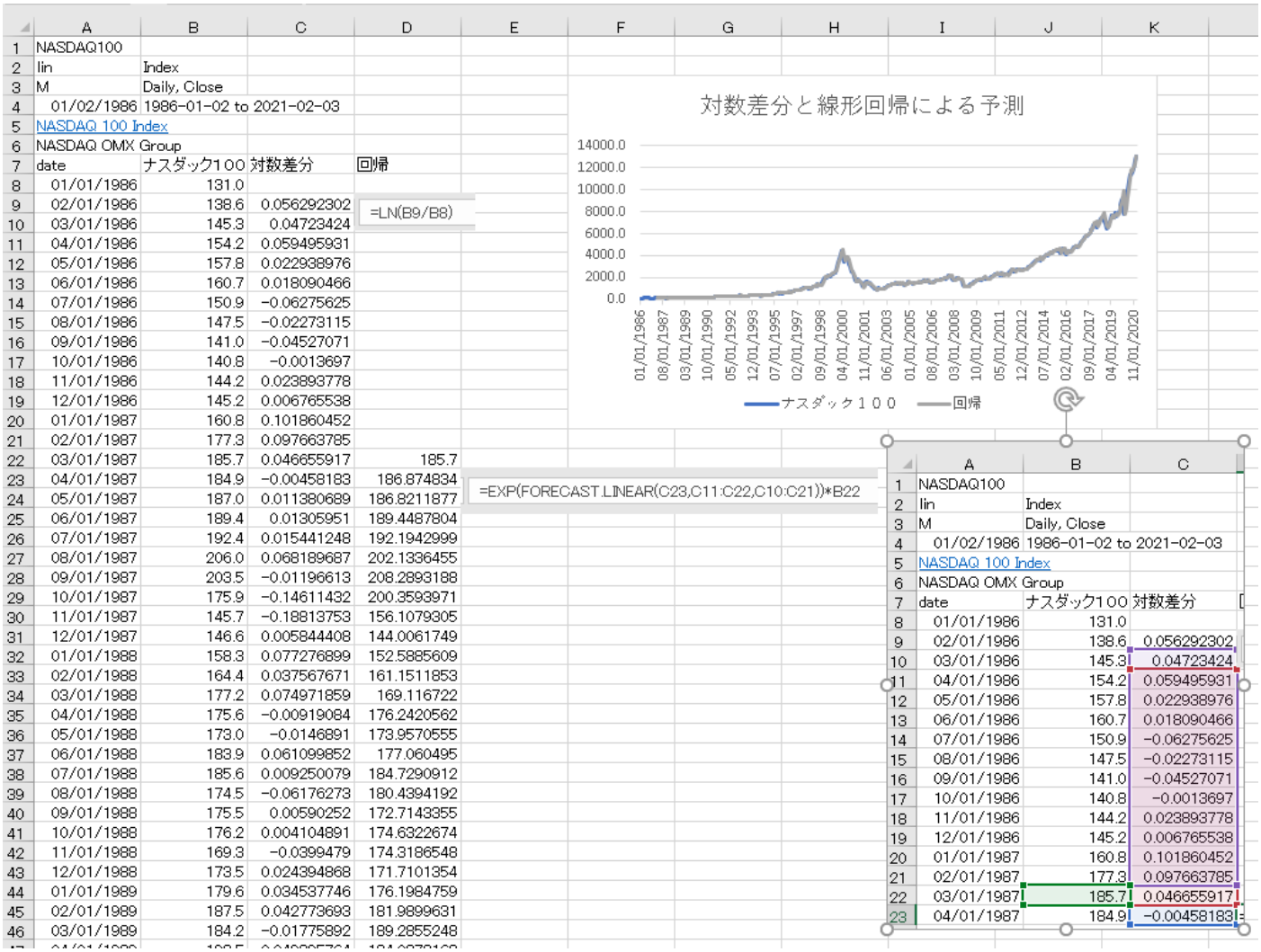
$\Delta y_t = y_t - y_{t-1}$

Δ は階差演算子です。

原系列$\{x_t\}$に関してトレンド部分$\{T_t\}$、季節変動部分$\{S_t\}$、循環部分$\{C_t\}$に分け、それ以外の不規則変動$\{I\}$をとしてつぎのように加法モデルを構成します。

$y_t = T_t + S_t + C_t + I_t$

これはトレンド部分の除去、季節変動部分の除去、循環部分の除去といった具合に扱います。このようなモデルを記述統計的手法といいます。



：3.　確率過程と時系列モデル

経済時系列は、経済学理論の検証、将来の時系列の動きの予測、さらにはこれらの分析に基づいたシステムの制御などを目的として分析されます。しかし、これらの目的を達成するには記述統計的手法では限界があります。そこで統計的時系列モデルが用いられます。これは確率変数の列を利用したモデルです。そして、実際の観測値を確率変数の列の実現系列と考えます。この２つの時系列の違いを分析しながら目的を達成していきます。これは無限次元のモデル(母集団)と有限次元のデータ(標本)を比較していることになります。

自己回帰モデル

Y_t の期待値が Y_t の b 倍となっている確率変数の列を考えます。

$E_{t-1}(Y_t)=a+b\ Y_{t-1}$

a,b は母回帰係数です。E_{t-1} は t-1 時点でも情報を用いた条件付期待値を表しています。

誤差項は

$v_t= Y_t - a - b\ Y_{t-1}$

です。

$Y_t = a + b\ Y_{t-1} + v_t$

を１次自己回帰過程（AR1）と呼びます。

時刻 t における確率変数の期待値と分散が時刻に依存せずに一定で、$a,\ b$ も時刻によらずに一定で$|b|<1$ ならば確率

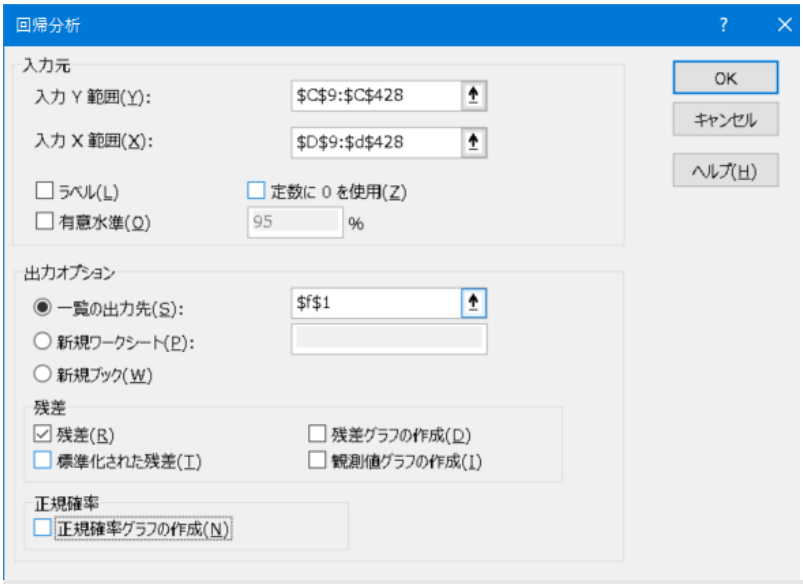
過程{ Y_t }は定常的です。

もし b=1 であると非定常過程となり線形回帰分析は使えません。そこで

$$Y_t \ - Y_{t-1} = \Delta Y_t = a + (b-1)\ Y_{t-1} + v_t$$

の回帰分析を行います。これで(b-1)がゼロであるかどうかを判定します。ゼロでなければ自己回帰モデルの可能性が高まります。

Excel シート: time_series_data_analysis.xlsx>自己回帰判定



	A	B	C	D
1	NASDAQ100			
2	lin	Index		
3	M	Daily, Close		
4	01/02/1986	1986-01-02 to 2021-02-03		
5	NASDAQ 100 Index			
6	NASDAQ OMX Group			
7	date	ナスダック100	対数価格	対数差分
8	01/01/1986	131.0	4.875487357	
9	02/01/1986	138.6	4.931779659	0.056292302
10	03/01/1986	145.3	4.9790139	0.04723424
11	04/01/1986	154.2	5.038509831	0.059495931
12	05/01/1986	157.8	5.061448807	0.022938976
13	06/01/1986	160.7	5.079539273	0.018090466
14	07/01/1986	150.9	5.016783025	-0.06275625
15	08/01/1986	147.5	4.99405188	-0.02273115
16	09/01/1986	141.0	4.948781167	-0.04527071
17	10/01/1986	140.8	4.947411464	-0.0013697
18	11/01/1986	144.2	4.971305242	0.023893778
19	12/01/1986	145.2	4.978070779	0.006765538

概要

回帰統計	
重相関 R	0.029274577
重決定 R2	0.000857001
補正 R2	-0.00153329
標準誤差	1.183998947
観測数	420

分散分析表

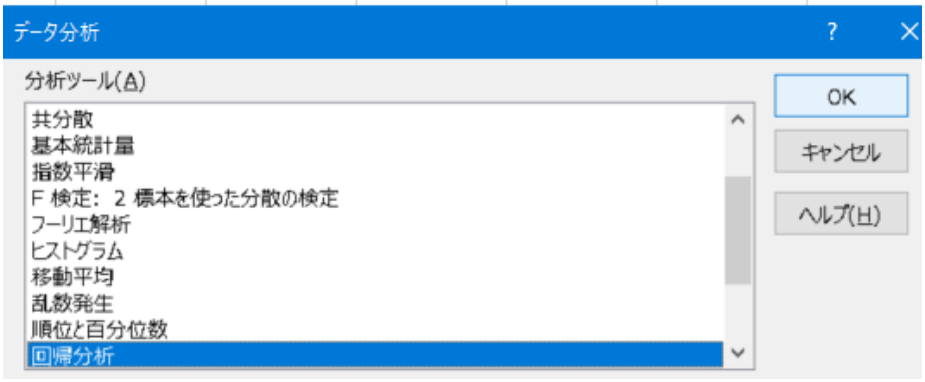
	自由度	変動	分散	観測された分散比	有意 F
回帰	1	0.502611632	0.502611632	0.358533634	0.549646093
残差	418	585.9747662	1.401853508		
合計	419	586.4773778			

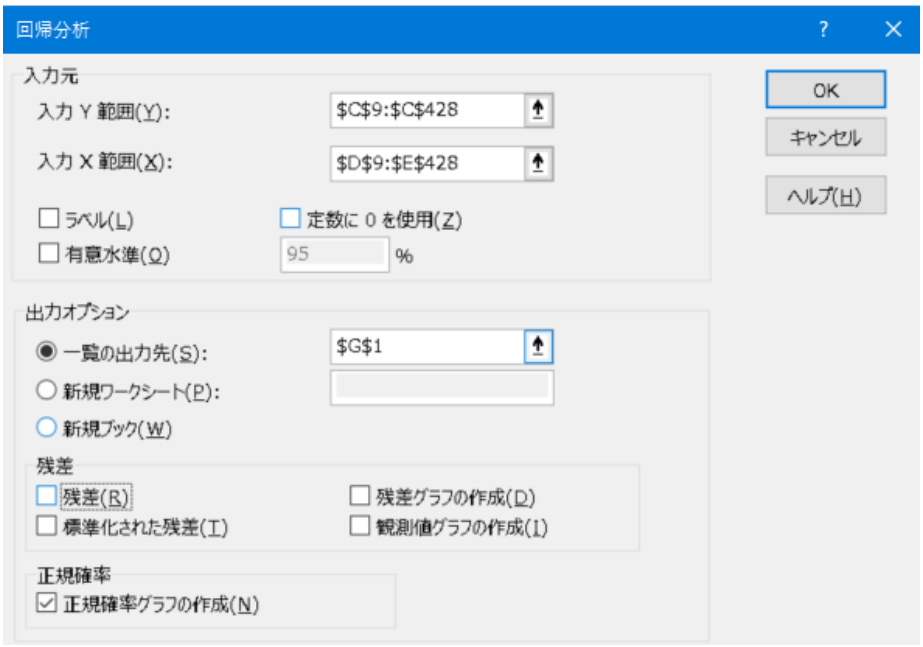
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	7.142483628	0.058840504	121.3871939	0	7.026823471	7.258143786	7.026823471	7.258143786
X 値 1	0.609793508	1.018398721	0.598776781	0.549646093	-1.3920275	2.611614519	-1.3920275	2.611614519

F 検定の p 値が高いため、モデルが正しくない可能性があります。そこでトレンド項を加えます。

Excel シート: time_series_data_analysis.xlsx>自己回帰判定（２）

	A	B	C	D	E
1	NASDAQ100				
2	lin	Index			
3	M	Daily, Close			
4	01/02/1986	1986-01-02 to 2021-02-03			
5	NASDAQ 100 Index				
6	NASDAQ OMX Group				
7	date	ナスダック100	対数価格	対数差分	トレンド項
8	01/01/1986	131.0	4.875487357		
9	02/01/1986	138.6	4.931779659	0.056292302	1
10	03/01/1986	145.3	4.9790139	0.04723424	2
11	04/01/1986	154.2	5.038509831	0.059495931	3
12	05/01/1986	157.8	5.061448807	0.022938976	4





	A	B	C	D	E	F	G	H	I	J	K	L	M
1	NASDAQ100						概要						
2	lin	Index											
3	M	Daily, Close					回帰統計						
4	01/02/1986	1986-01-02 to 2021-02-03					重相関 R	0.943545098					
5	NASDAQ 100 Index						重決定 R2	0.890277352					
6	NASDAQ OMX Group						補正 R2	0.889751105					
7	date	ナスダック100	対数価格	対数差分	トレンド項		標準誤差	0.392830992					
8	01/01/1986	131.0	4.875487357				観測数	420					
9	02/01/1986	138.6	4.931779659	0.056292302	1								
10	03/01/1986	145.3	4.9790139	0.04723424	2		分散分析表						
11	04/01/1986	154.2	5.038509831	0.059495931	3			自由度	変動	分散	観測された分散比	有意 F	
12	05/01/1986	157.8	5.061448807	0.022938976	4		回帰	2	522.1275272	261.0637636	1691.745798	7.9758E-201	
13	06/01/1986	160.7	5.079539273	0.018090466	5		残差	417	64.34985064	0.154316189			
14	07/01/1986	150.9	5.016783025	-0.06275625	6		合計	419	586.4773778				
15	08/01/1986	147.5	4.99405188	-0.02273115	7								
16	09/01/1986	141.0	4.948781167	-0.04527071	8			係数	標準誤差	t	P-値	下限 95%	上限 95%
17	10/01/1986	140.8	4.947411464	-0.0013697	9		切片	5.208613177	0.038568239	135.0492868	0	5.13280078	5.284425575
18	11/01/1986	144.2	4.971305242	0.023893778	10		X 値 1	0.517667083	0.33789133	1.532051985	0.126267877	-0.14651548	1.181849648
19	12/01/1986	145.2	4.978070779	0.006765538	11		X 値 2	0.009191827	0.000158099	58.13978475	3.946E-202	0.008881057	0.009502597

結果は回帰係数 0.52 となりました。p 値が高いですが、ここは良しとします。気になる人はトレンド項を増やしてください。

ランダムウォーク・モデル

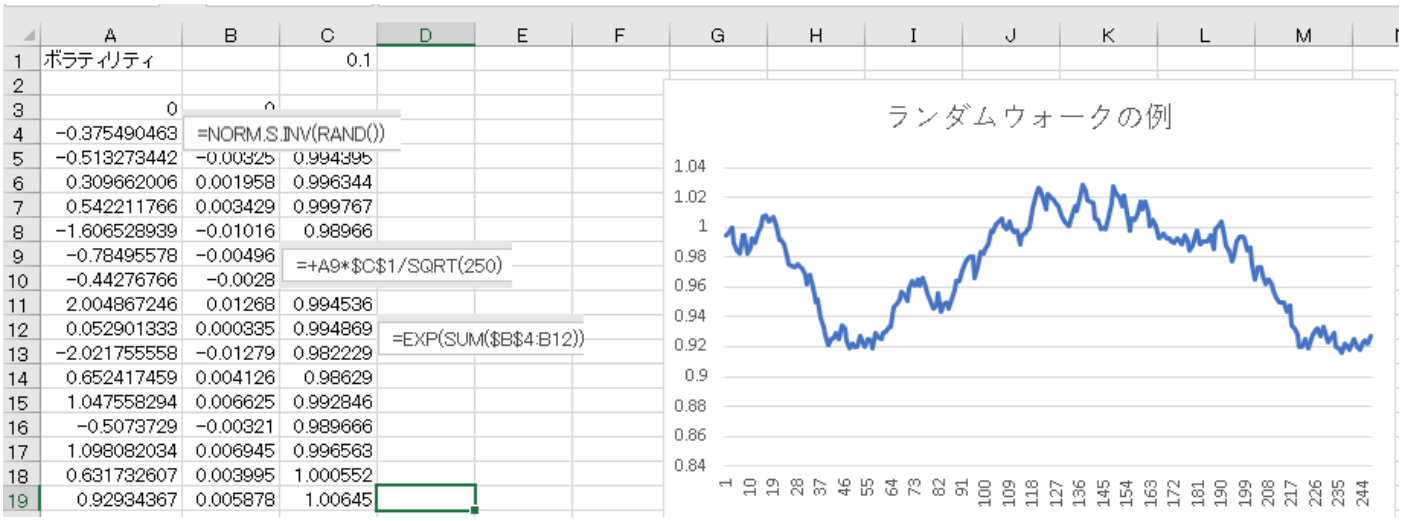
AR(1)モデルで b=1 とすると、これは独立な確率変数の和となり、ランダムウォーク・モデルと呼ばれる。これは定常確率過程ではなく、非定常確率過程です。なぜなら時間の経過に伴い、時刻 t の確率変数の分散が大きくなるからです。

$$Y_t = a + Y_{t-1} + v_t$$

確率的トレンド

ランダムウォークにしたがう時系列がトレンドを持つとき、そのトレンドを確率的トレンドといい、確定的トレンドと区別されます。このようなトレンドはランダムに発生するために、発生するタイミングと期間を予測することは不可能です。

エクセルシート：time_series_data_analysis.xlsx>ランダムウォーク



多項式モデル

従属変数を独立変数の多項式としてモデル化することができます。

$$Y_t = a + b_1 Y_{t-1} + b_2 Y_{t-1}^2 + +, \ldots, b_n Y_{t-1}^n + v_t$$

多項式回帰では従属変数と独立変数とが非線形な関係で表現されます。しかし、これは独立変数の線形和として表現されているために、重回帰分析とみなすことができます。

4. モデルの最適化と予測

モデルを最適さする際にはインサンプルとアウトオブサンプルという方法が使われます。

- インサンプル:得られたデータすべてを使ってモデルを最適化します。理論の検証などに用いられます。
- アウトオブサンプル：得られたデータの一部を使ってモデルを最適化し、残りのデータを推定パラメータの検証に使います。将来の被説明変数を予測する際などに用います。

誤差二乗和

一般にモデルのパラメータを決める際には誤差の二乗の和を最小化する方法が使われます。

クロスバリデーション

クロスバリデーションはモデルの最適化に際して、得られたデータをいくつかの区間に分けて、その中でアウトサンプルテストを繰り返します。予測を目的とする際に用いられます。

尤度と情報量基準

尤度

標本から母数を推定するために、点推定、区間推定を行うと同様に、ある分布を仮定して、その分布が再現される確率が最も高くなるようにパラメータを決めるという方法があります。パラメータを θ とすると

$$L(\theta|x)=f(x|\theta)$$

となる L を尤度関数といいます。これは母数の点推定です。尤度を最大にするのも、尤度の自然対数を最大にするのも同じなので、扱いやすい対数尤度を最大にします。

正規分布を仮定する場合には、母数は $\theta=(\mu, \sigma^2)$ なので最尤推定量は

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{(\sum x_i - \bar{x})^2}{n} \right)$$

となります。尤度そのものではなくその比率に大きな意味があります。たとえば

$$L(\theta_1|x) / L(\theta_2|x) > 1$$

であれば、θ_1 の方が θ_2 よりももっともらしいと判断できます。

この性質を利用したものとして情報量基準があります。これは統計的モデルの構築の際に用いられます。情報量基準は限られたデータにもとづいて良いモデルを選択するための方法を示唆してくれます。情報量基準は有限なデータを利用してモデルを構築する際に、自由度の大きなモデルを用いることによる不安定性の上昇の度合いを避けるために、パラメータ数の決定や、目的変数の選択に利用されます。情報量基準を最小にすることで情報量基準という意味で最良のモデルを選択できます。主な情報量基準に赤池情報量基準(AIC)とベイズ情報量基準(BIC)があります。

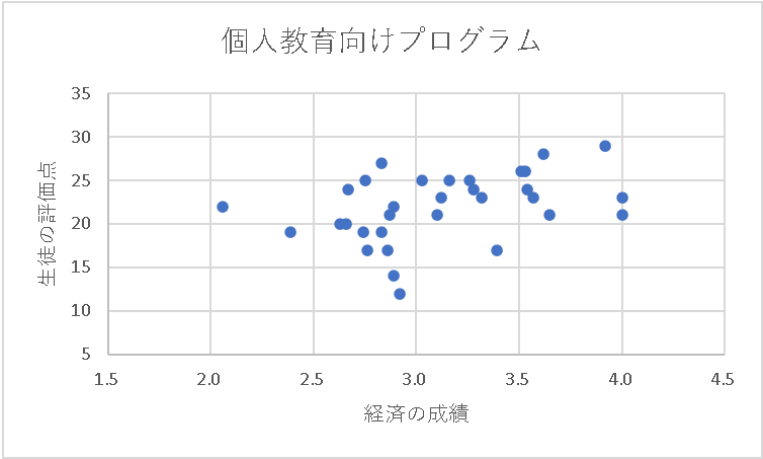
$$\mathrm{AIC} = -2 \ln L + 2n$$

$$\mathrm{BIC} = -2 \ln L + 2 \ln(n)$$

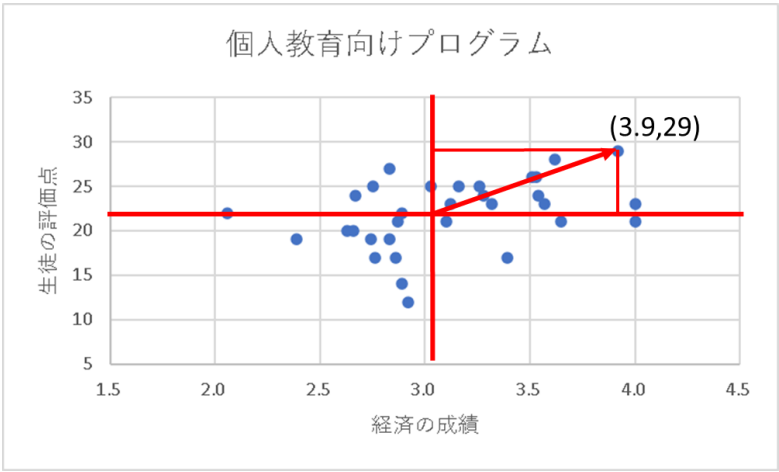
で表わされます。AIC は主に予測に用いられ、BIC は検証に用いられます。

情報と情報量

そこで個人教育向けプログラムに関するつぎのような情報をもとに参加者の成績を考えてみたいと思います。経済の成績と生徒の評価点の２つが与えられています。これらの２つの成績をもとに生徒の成績を考えてみたいと思います。散布図を取ってみるとつぎの様になりました。



２つの情報が得られましたが、これではより判断が難しくなってしまっています。そこで、１つの指数を作って判断してみたいと思います。情報という観点からこの指標をとらえます。情報の多い少ないを、不確実性の尺度として考えます。つまり考えたことがないような事態が起こるときは情報が多く、明らかな事態が起こるときは情報がないと考えます。その観点から行くと分散は情報の量の１つの目安になります。ここではこの性質を利用して、情報の量が減らないような指標を 1 つ作ります。分散は平均からのばらつきの程度を測定していますから、この考え方をデータの１点１点に当てはめます。経済の成績の平均は 3.1、生徒の評価点の平均は 22 ですから、それとデータ点との差の 2 乗を情報量と考えます。そうするとそれぞれのデータ点が経済の成績の平均と評価点の平均からどの程度離れているかということが情報の量となります。



たとえば、経済の成績が 3.9、評価点が 29 のデータ点(3.9,29)の情報量は

$(3.9-3.1)^2+(29-22)^2=50.5$

となります。この値の全データの平均が分散ということになります。

5. 応用分野

時系列データ分析は理論の検証、将来時系列の予測、システムの制御の他に、リスク管理、損失管理などに用いられます。

リスク管理

複数資産に投資をする際に、リスクを分散や標準偏差を用いて把握し、またトレンドは上昇率を用いてとらえます。このような統計量は、時間の経過とともにたびたび変化するために、さまざまな資産配分の方法があります。投資対象資産の性質により、平均・分散を用いるものや、分散のみを用いるもの、または均等に配分するものなど様々です。また、ドローダウンと呼ばれる投資のリターンが下落し始めてから回復するまでに被る最大損失を最小にするものなどがあります。

バリューアットリスク

バリューアットリスク(value at risk:VaR)は、保有ポジションを閉じるために被る最大の損失額のことです。一般的にはある一定期間後にある信頼度で被る損害額として算出されます。

もっとも単純な方法はポートフォリのリターンが正規分布すると仮定して、その損失額を算出します。エクセルではNORMINV 関数を用いて求めることができます。

例：NORMINV(0.01,0.0,0.025)=-0.058 となり 5.8 の損失が 1 日で見込まれます。0.025 の標準偏差ボラティリティに換算すると約 40%に相当します。

エクセルシート：time_series_data_analysis.xlsx>VaR