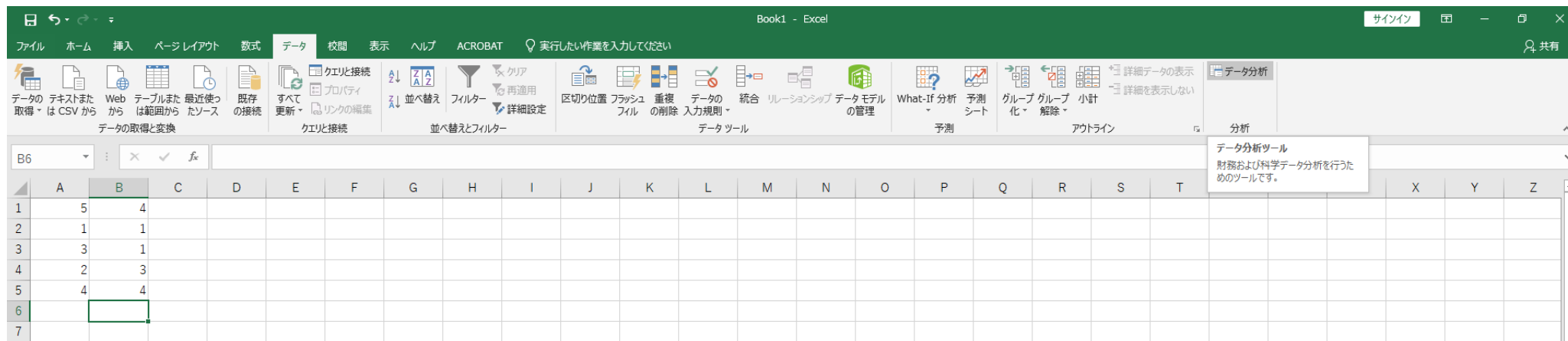


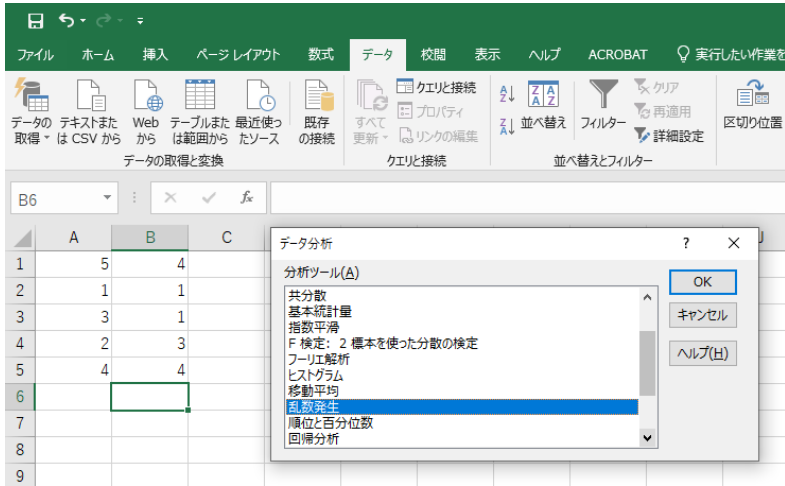
3.4.2 「分析ツール」による回帰分析

散布図による方法は、単回帰の場合には、比較的簡単に計算できるが、説明変数が2つ以上の重回帰には適用することは出来なくなる。この場合、「分析ツール」を使うと、簡単に、回帰分析を行うことができる。

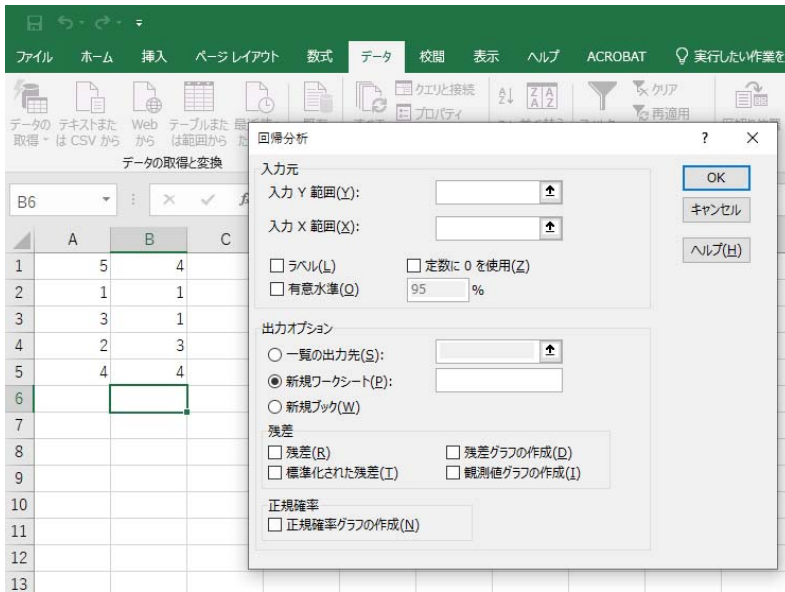
まず、「データ」タブを選ぶ。



「データ分析」のタブをマウスで選択すると、下記のような画面になり、様々なツールが利用できるようになる。主に利用するツールは、「ヒストグラム」と「回帰分析」である。

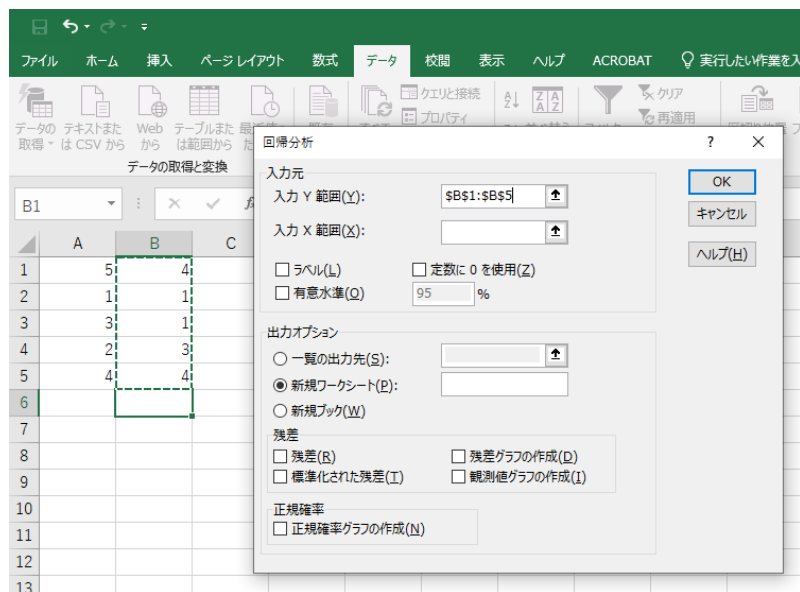


本節では、回帰分析の方法を解説する。まずは、「回帰分析」を選ぶと、下記の画面となる。

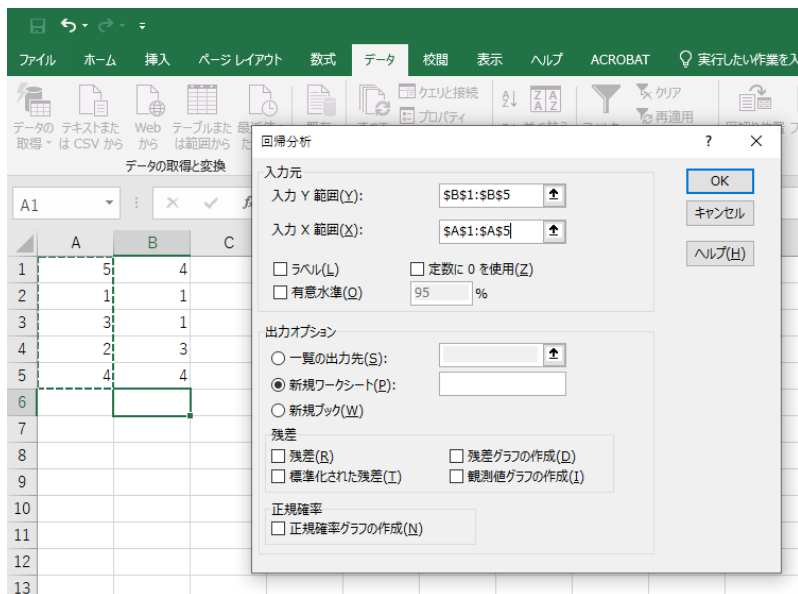


「入力 Y 範囲 (Y)」に B 列のデータ（被説明変数）を選択する。

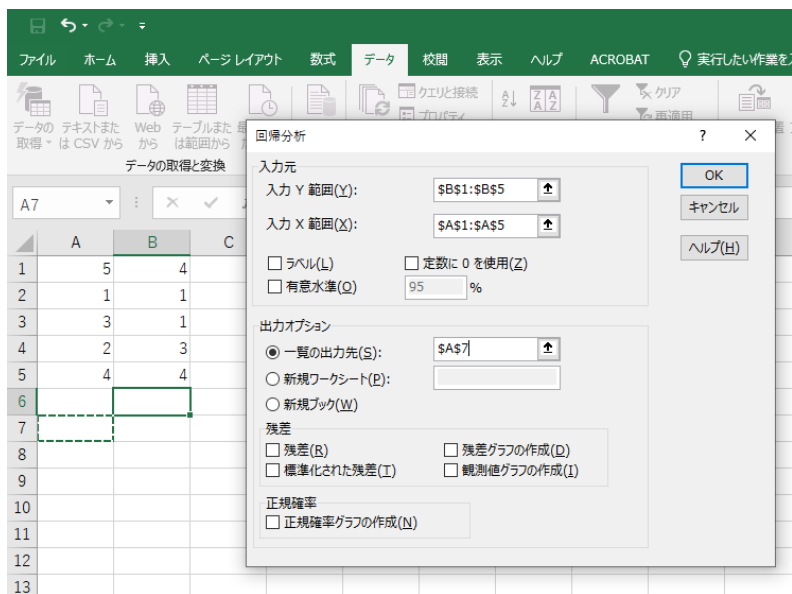
「入力 Y 範囲 (Y)」の右側の空欄をマウスの左ボタンをクリックして、さらに、B1 をマウスの左ボタンでクリック、さらにマウスの左ボタンを押し続けながら B5 でマウスボタンを離す（または、B1:B5 とタイプする）。下記の画面となる。



同様に、「入力 X 範囲 (X)」の右側の空欄をマウスの左ボタンでクリックして、さらに、A1 を左ボタンでクリック、マウスの左ボタンを押し続けながら A5 でマウスボタンを離す（または、A1:A5 と入力する）。下記の画面となる。



「一覧の出力先 (S)」にチェックを入れて、その右側の空欄をマウスの左ボタンでクリック、適当な場所をマウスでクリックして選択する（ここでは、A7 をクリックする。または、A7 とタイプする）。下のような表示になる。



このように入力した後、右側の「OK」ボタンをクリックする。下のような出力結果が得られる。

	A	B	C	D	E	F	G	H	I	J
1	5	4								
2	1	1								
3	3	1								
4	2	3								
5	4	4								
6										
7	概要									
8										
9	回帰統計									
10	重相関 R	0.7298								
11	重決定 R2	0.532609								
12	補正 R2	0.376812								
13	標準誤差	1.197219								
14	観測数	5								
15										
16	分散分析表									
17		自由度	変動	分散	割られた分	有意 F				
18	回帰	1	4.9	4.9	3.418605	0.161594				
19	残差	3	4.3	1.433333						
20	合計	4	9.2							
21										
22		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
23	切片	0.5	1.255654	0.398199	0.717129	-3.49605	4.496051	-3.49605	4.496051	
24	X 値 1	0.7	0.378594	1.848947	0.161594	-0.50485	1.904855	-0.50485	1.904855	
25										
26										
27										
28										

今までの授業では、下記の水色部分を扱った。

回帰統計	
重相関 R	0.7298
重決定 R2	0.532609
補正 R2	0.376812
標準誤差	1.197219
観測数	5

分散分析表					
	自由度	変動	分散	F 値	有意 F
回帰	1	4.9	4.9	3.418605	0.161594
残差	3	4.3	1.433333		
合計	4	9.2			

	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	0.5	1.255654	0.398199	0.717129	-3.49605	4.496051	-3.49605	4.496051
X 値 1	0.7	0.378594	1.848947	0.161594	-0.50485	1.904855	-0.50485	1.904855

Excel の「重決定 R2」は決定係数, 「補正 R2」は自由度修正済み決定係数, 「観測数」はデータ数 n のことである。

「残差 + 自由度」の 3, 「合計 + 自由度」の 4 はそれぞれ $n - k = 5 - 2 = 3$, $n - 1 = 5 - 1 = 4$ であり, 自由度を表す。

また, 「残差 + 変動」の 4.3, 「合計 + 変動」の 9.2 という数字は, それぞれ残差平方和, Y の平均からの差の

二乗和で、次のものである。

$$\sum_{i=1}^n \hat{u}_i^2 = 4.3 \quad \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = 43 - 5 \times 2.6^2 = 9.2$$

「切片+係数」の0.5, 「X値1+係数」の0.7は, 切片, 傾きを表す ($Y=0.7X+0.5$)。

得られた数値と今回得られた数値を比較すると, それぞれの数字がどのような意味かがわかるだろう。

3.4.3 決定係数 R^2 について

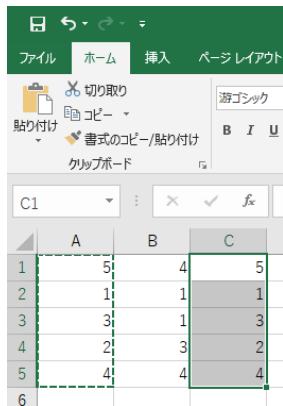
●説明変数を増やせば、必ず決定係数 R^2 は大きくなることを確認する。

都合により、A列のデータ（説明変数）をC列にコピーする。

コピーの方法としては、A1にマウスを持っていき、マウスの左ボタンを押し続けて、A5で左ボタンを離す。

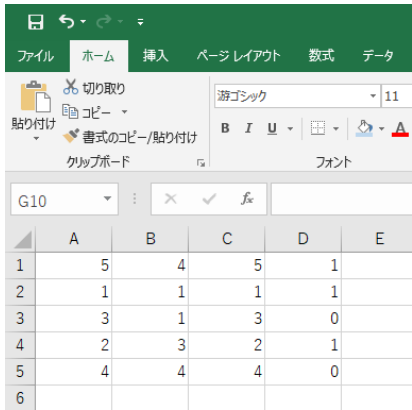
次に、A5にマウスがある状態で、マウスの右ボタンを押し、「コピー (C)」を選択する。C1で右ボタンを押し、「貼り付けのオプション」の一番左のアイコン「貼り付け (P)」を選ぶと、下記のように、A列がC列に

コピーできる。

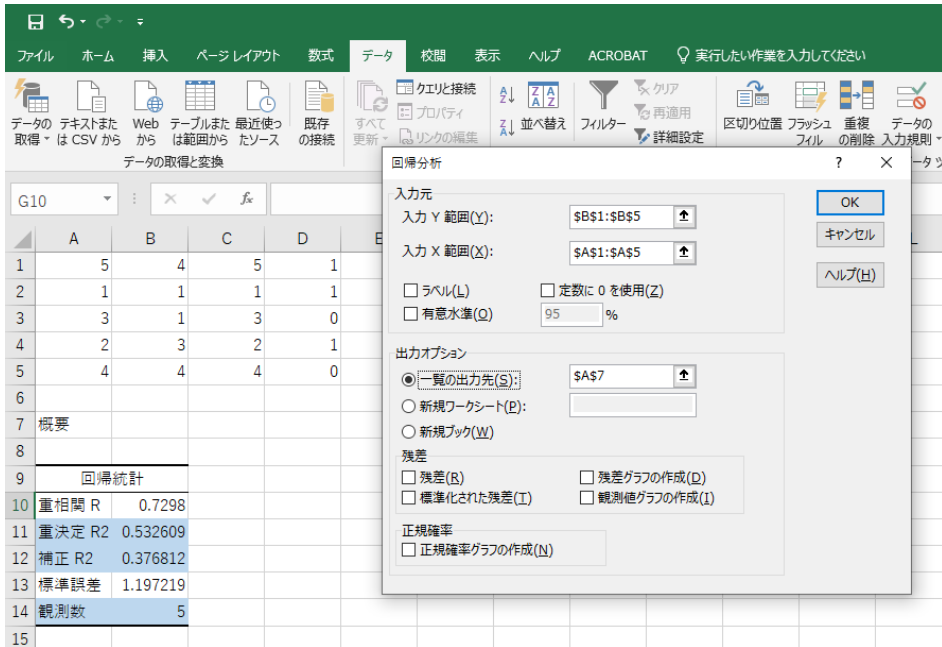


次に、D列に適当に、例えば、1, 1, 0, 1, 0というデータを入力する。

B列を被説明変数、C列・D列を説明変数として回帰分析する。



「データ」タブ、「データ分析」、「回帰分析」、「OK」と順番に選択していくと、下記のように前回のものが残ったままになっている。



「入力 X 範囲(X)」の欄を削除して、C1 にマウスを置いて、マウスの右ボタンを押し続けて、D5 に移動する

(選択範囲を C1 から D5 とする)。下記の画面になる。

回帰分析

入力元
入力 Y 範囲(Y): \$B\$1:\$B\$5
入力 X 範囲(X): \$C\$1:\$D\$5
 ラベル(L) 定数に 0 を使用(Z)
 有意水準(Q) 95 %

出力オプション
 一覧の出力先(S): \$A\$7
 新規ワークシート(P):
 新規ブック(V)

残差
 残差(B) 残差グラフの作成(D)
 標準化された残差(I) 観測値グラフの作成(I)

正規確率
 正規確率グラフの作成(N)

	A	B	C	D	E
1	5	4	5	1	
2	1	1	1	1	
3	3	1	3	0	
4	2	3	2	1	
5	4	4	4	0	
6					
7	概要				
8					
9	回帰統計				
10	重回相関 R	0.7298			
11	重決定 R2	0.532609			
12	補正 R2	0.376812			
13	標準誤差	1.197219			
14	観測数	5			
15					

次に、「一覧の出力先(S)」の欄を削除して、例えば、A26 でマウスの左ボタンを押す。

下記の画面となる。

データの取得と変換

	A	B	C	D	E				
1	5	4	5	1					
2	1	1	1	1					
3	3	1	3	0					
4	2	3	2	1					
5	4	4	4	0					
6									
7	概要								
8									
9	回帰統計								
10	重相関 R	0.7298							
11	重決定 R2	0.532609							
12	補正 R2	0.376812							
13	標準誤差	1.197219							
14	観測数	5							
15									
16	分散分析表								
17		自由度	変動	分散	F 値	有意 F			
18	回帰	1	4.9	4.9	3.418605	0.161594			
19	残差	3	4.3	1.433333					
20	合計	4	9.2						
21									
22		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
23	切片	0.5	1.255654	0.398199	0.717129	-3.49605	4.496051	-3.49605	4.496051
24	X 値 1	0.7	0.378594	1.848947	0.161594	-0.50485	1.904855	-0.50485	1.904855
25									
26									
27									

右の「OK」ボタンを押す。

A26 以下に下記の結果が出力される。

概要								
	A	B	C	D	E	F	G	H
25								
26	概要							
27								
28	回帰統計							
29	重相関 R	0.782718						
30	重決定 R2	0.612648						
31	補正 R2	0.225296						
32	標準誤差	1.334848						
33	観測数	5						
34								
35	分散分析表							
36		自由度	変動	分散	F 値	有意 F		
37	回帰	2	5.636364	2.818182	1.581633	0.387352		
38	残差	2	3.563636	1.781818				
39	合計	4	9.2					
40								
41		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%
42	切片	-0.23636	1.808885	-0.13067	0.907996	-8.01937	7.546642	-8.01937
43	X 値 1	0.781818	0.440886	1.77329	0.218182	-1.11516	2.678796	-1.11516
44	X 値 2	0.818182	1.272727	0.642857	0.58618	-4.65792	6.294285	-4.65792
45								
46								
47								
48								

D 列の変数を Z とすると,

$$Y_i = - 0. 236 + 0. 782 X_i + 0. 818 Z_i$$

という結果となった。

D 列の説明変数を加えたことにより, 決定係数は 0. 5326 から 0. 6126 に増えたが, 自由度修正済み決定係数は 0. 3768 から 0. 2253 へ低下した。

したがって、D列（説明変数）はB列（被説明変数）に影響を与える変数ではないと言える。

言い換えると、B列にとって、D列は重要ではない。

● 統計学の知識が必要な部分を薄黄色で表す。

26	概要								
27									
28	回帰統計								
29	重相関 R	0.782718							
30	重決定 R ²	0.612648							
31	補正 R ²	0.225296							
32	標準誤差	1.334848							
33	観測数	5							
34									
35	分散分析表								
36		自由度	変動	分散	割られた分散	有意 F			
37	回帰	2	5.636364	2.818182	1.581633	0.387352			
38	残差	2	3.563636	1.781818					
39	合計	4	9.2						
40									
41		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
42	切片	-0.23636	1.808885	-0.13067	0.907996	-8.01937	7.546642	-8.01937	7.546642
43	X 値 1	0.781818	0.440886	1.77329	0.218182	-1.11516	2.678796	-1.11516	2.678796
44	X 値 2	0.818182	1.272727	0.642857	0.58618	-4.65792	6.294285	-4.65792	6.294285
45									

水色は前述の通り、授業で既に解説済み。

● 決定係数を比較するためには、被説明変数が同じでなければならない。

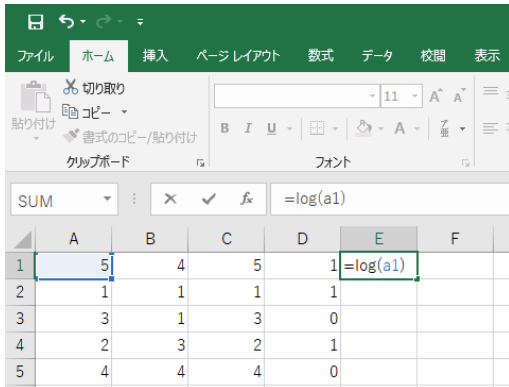
先ほどの例では、

$$Y = 0.5 + 0.7 X \quad R^2 = 0.5326$$

であった。

Y, X に対数を取って、 $\log Y = \alpha + \beta \log X$ を推定してみる。

E 列・F 列に A 列・B 列の対数を求める。E1 に「=log(a1)」とタイプする。



	A	B	C	D	E	F
1	5	4	5	1	=log(a1)	
2	1	1	1	1		
3	3	1	3	0		
4	2	3	2	1		
5	4	4	4	0		

Enter キーを押す。

	A	B	C	D	E	F
1	5	4	5	1	0.69897	
2	1	1	1	1		
3	3	1	3	0		
4	2	3	2	1		
5	4	4	4	0		

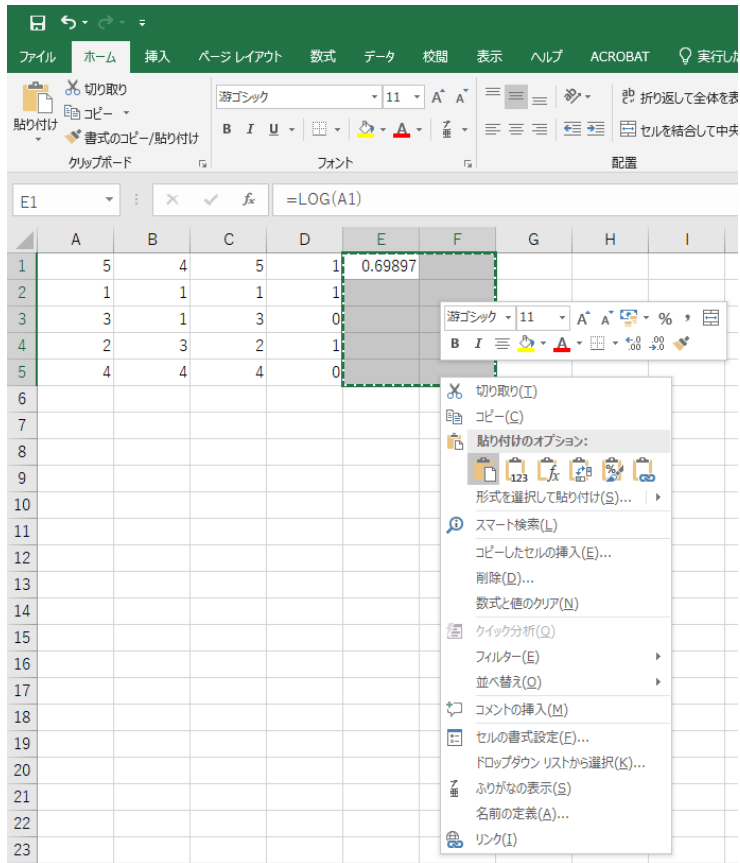
5 の常用対数の値（底が 10，すなわち， $\log_{10} 5$ ）が E1 に計算される。

E1 にマウスを置いて，マウスの右ボタンを押して，「コピー(C)」を選択する。

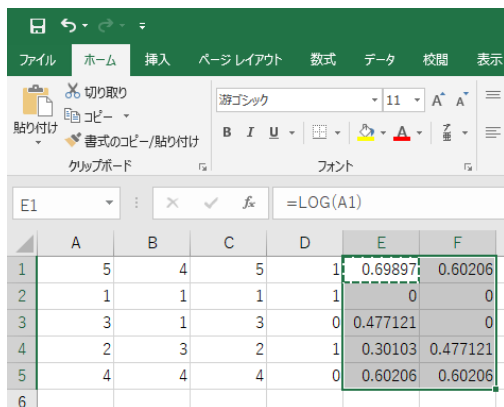
マウスを押し続けながら，F5 で，マウスの右ボタンを離すと，下記のようにになる。

	A	B	C	D	E	F
1	5	4	5	1	0.69897	
2	1	1	1	1		
3	3	1	3	0		
4	2	3	2	1		
5	4	4	4	0		
6						

すぐに，再度，右ボタンを押すと，下記のようにになる。



「貼り付けオプション：」の一番左を選択すると、下記のように対数が計算される。



「入力 Y 範囲 (Y)」を F1 から F5, 「入力 X 範囲 (X)」を E1 から E5, 「一覧の出力先 (S)」は適当なところ (ここでは, A46) を選択して, 「OK」ボタンを押すと, 下記の結果が得られる。

概要								
帰帰統計								
重相関 R	0.663151							
重決定 R2	0.43977							
補正 R2	0.253026							
標準誤差	0.268928							
観測数	5							
分散分析表								
	自由度	変動	分散	F 値	有意 F			
帰帰	1	0.170315	0.170315	2.35494	0.222445			
残差	3	0.216968	0.072323					
合計	4	0.387283						
	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	0.025354	0.235602	0.107614	0.921095	-0.72444	0.775143	-0.72444	0.775143
X 値 1	0.747636	0.487192	1.534581	0.222445	-0.80283	2.2981	-0.80283	2.2981

$$\log Y = 0.0254 + 0.7476 \log X \quad R^2 = 0.4398$$

となっている。対数を取る前は,

$$Y = 0.5 + 0.7 X \quad R^2 = 0.5326$$

で, R^2 の比較はできない。係数の意味も異なる (この点は後述)。

3.4.4 補足

3.4.3 節の冒頭で、「都合により、A列のデータ（説明変数）をC列にコピーする。」と述べた。

そして、C列・D列を説明変数として回帰分析を行った。

A列とD列を説明変数とするとどうなるかを見る。

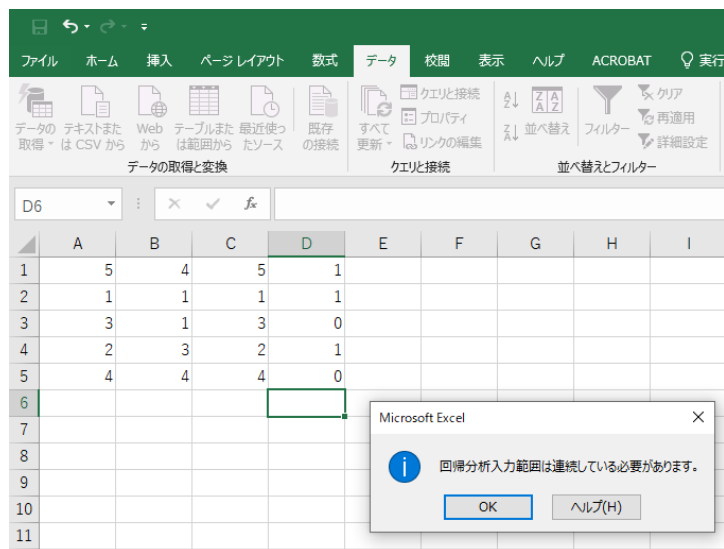
	A	B	C	D
1	5	4	5	1
2	1	1	1	1
3	3	1	3	0
4	2	3	2	1
5	4	4	4	0
6				
7				
8				
9				
10				
11				
12				
13				
14				

「入力 Y 範囲 (Y)」は B 列（これは今までと同様）、「一覧の出力先 (S)」を A7 にする。

「入力 X 範囲 (X)」に、A 列と D 列を選択する（グラフ作成の時と同様に、A1 から A5 までをマウスの左ボタンを押し続けて選択して、次に、Ctrl キーを押しながら D1 から D5 までをマウスの左ボタンを押し続けて選

択する)。

「OK」を押すと、下記の画面になる。



このように、計算結果が出力されない。

「入力 X 範囲(X)」の選択の際には、説明変数データを隣に並べておく必要がある（説明変数が3つであれば、3列連続に並べなければならない）。

これは、試行錯誤で説明変数の種類を変えて、数多くの式を推定する場合はかなり手間がかかる（推定の度に、毎回、説明変数を連続になるように並べ直すことになる）。

この状況を避けるためには、専門の計量経済ソフトを使うことを勧める。

時間の節約にもなり、簡単に推定結果を出すこともできるようになる。

専門の計量経済ソフト：

- ・ 有料 → STATA, EViews, TSP, SPSS など（しかし、高価）
- ・ 無料 → R, Python, Gretl など（ただし、R や Python は若干のプログラミングの知識が必要）

総合的には、Gretl がおすすめ。

<http://gretl.sourceforge.net/>

からダウンロード（windows 版, mac 版あり）

ただし、英語

第4章 統計学の基礎：復習

4.1 確率変数，確率分布について

確率変数は，通常，大文字のアルファベット（例えば， X ）で表すのに対して，実際に起こった値（すなわち，実現値）を小文字（例えば， x ）で表す。

確率変数には離散型確率変数と連続型確率変数がある。まず，離散型確率変数 X を考える。

X の取り得る値は分かっている。例えば、 $X = x_1, x_2, \dots, x_n$ の n 通りの値を取るものとする。それぞれの値には確率が割り当てられる。すなわち、 $\text{Prob}(X = x_i) = p_i$ と表記し、「確率変数 X が x_i を取る確率は p_i である」と読む。 p_i は確率であり、しかも、 X は x_1, x_2, \dots, x_n のいずれかの値を取るので、 $\sum_{i=1}^n p_i = 1$ となる。また、 p_i は x_i の関数であり、 $f(x_i)$ と表すことができる。 $f(x_i)$ を確率関数と呼ぶ。 $f(x_i)$ は、(i) $f(x_i) \geq 0$ 、(ii) $\sum_{i=1}^n f(x_i) = 1$ を満たす関数でなければならない。

X をサイコロを投げて出た目としよう。このとき、 X の取る値は $1, 2, 3, 4, 5, 6$ で、それぞれの目が出る確率は $\frac{1}{6}$ となる。したがって、 $x_i = i, p_i = \frac{1}{6}, i = 1, 2, 3, 4, 5, 6$ となる。

X が連続型確率変数の場合は、ある値 a から別の値 b までの区間に入る確率 $\text{Prob}(a < X < b)$ という意味になる（ただし、 $a < b$ ）。この場合、 $f(x), x = a, x = b, x$ 軸で囲まれた面積が

確率を表すことになる。すなわち，

$$\mathbf{Prob}(a < X < b) = \int_a^b f(x)dx,$$

となり， $f(x)$ を確率密度関数，または，密度関数と呼ぶ。 $f(x)$ は，(i) $f(x) \geq 0$ ，(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$ を満たす連続関数でなければならない。

離散型の $f(\cdot)$ と連続型の $f(\cdot)$ の違いは，前者は $f(\cdot)$ そのものが確率を表すのに対して，後者の $f(\cdot)$ は面積が確率を表す（すなわち，連続型の $f(\cdot)$ の高さは確率を表さない）。

分布関数（累積分布関数）： 分布関数（累積分布関数） $F(x)$ は，

$$F(x) = \mathbf{Prob}(X \leq x) = \begin{cases} \sum_{i=1}^r f(x_i) & X \text{ が離散型確率変数のとき} \\ \int_{-\infty}^x f(t)dt & X \text{ が連続型確率変数のとき} \end{cases}$$

ただし，離散型の場合， r は $x_r \leq x < x_{r+1}$ となる r である。すなわち，離散型の場合， $F(x)$ は **0** と **1** の間の階段状（階段関数）となる。

同時確率分布： 2つの確率変数 X, Y を考える。離散型の場合， X の取る値を x_1, x_2, \dots, x_n とし， Y の取る値を y_1, y_2, \dots, y_m としたとき， X が x_i を取り，かつ， Y が y_j を取る確率を同時確率分布と呼び，下記のように表す。

$$\mathbf{Prob}(X = x_i, Y = y_j) = p_{ij}$$

p_{ij} は x_i, y_j の関数となり , $p_{ij} = f(x_i, y_j)$ と表す。 $f(x_i, y_j)$ を同時確率関数と呼ぶ。

連続型の場合は , X が c と d の間の値 (ただし , $a < b$) を取り , かつ , Y が c と d の間の値 (ただし , $c < d$) を取る確率は , 下記のように表される。

$$\mathbf{Prob}(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x, y) dy dx$$

$f(x, y)$ を同時確率密度関数 (または , 同時密度関数) と呼ぶ。

4.2 期待値・分散・共分散の定義・定理

4.2.1 期待値の定義

定義 (期待値, 1 変数): 確率変数 X , ある関数 $g(\cdot)$ とするとき, $g(X)$ の期待値は次のように定義される。

$$\mathbf{E}(g(X)) = \begin{cases} \sum_{i=1}^n g(x_i)f(x_i), & X \text{ が離散型確率変数のとき} \\ \int_{-\infty}^{\infty} g(x)f(x)\mathbf{d}x, & X \text{ が連続型確率変数のとき} \end{cases} \quad (4.1)$$

ただし, $f(\cdot)$ は確率関数 (離散型のとき), または, 密度関数 (連続型のとき) を表す。

定義 (期待値, 2変数): 確率変数 X, Y , ある関数 $g(\cdot, \cdot)$ とするとき, $g(X, Y)$ の期待値は次のように定義される。

$$\mathbf{E}(g(X, Y)) = \begin{cases} \sum_{i=1}^n \sum_{j=1}^m g(x_i, y_j) f(x_i, y_j), & X, Y \text{ が離散型確率変数のとき} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \mathbf{d}y \mathbf{d}x, & X, Y \text{ が連続型確率変数のとき} \end{cases} \quad (4.2)$$

ただし, $f(\cdot, \cdot)$ は確率関数 (離散型のとき), または, 密度関数 (連続型のとき) を表す。

2変数 (X, Y) を n 変数 (X_1, X_2, \dots, X_n) に拡張することも出来る。

4.2.2 期待値の定理

定理 (1 変数) : X を確率変数とする。 $a + bX$ の期待値は ,

$$\mathbf{E}(a + bX) = a + b\mathbf{E}(X), \quad (4.3)$$

となる。ただし , a, b は定数とする。 $g(X) = a + bX$ に対応する。

定理 (2 変数) : X, Y を確率変数とする。 $X + Y$ の期待値は ,

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y), \quad (4.4)$$

となる。 $g(X, Y) = X + Y$ に対応する。

定理 (多変数) : n 個の確率変数 X_1, X_2, \dots, X_n を考える。このとき , $\sum_{i=1}^n c_i X_i$ の平均は ,

$$\mathbf{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbf{E}(X_i), \quad (4.5)$$

となる。

4.2.3 分散・共分散の定義・定理

定義 (1 変数) : X を確率変数とする。 X の分散 $\sigma^2 = \mathbf{V}(X)$ は ,

$$\sigma^2 = \mathbf{V}(X) = \mathbf{E}((X - \mu)^2), \quad (4.6)$$

である。ただし , $\mu = \mathbf{E}(X)$ とする。 $g(X) = (X - \mu)^2$ に対応する。

定義 (1 変数) : X を確率変数とする。 X の標準偏差 σ は ,

$$\sigma = \sqrt{V(X)} \quad (4.7)$$

である。

定理 (1 変数) : X を確率変数とする。 X の分散は ,

$$V(X) = E(X^2) - \mu^2, \quad (4.8)$$

と書き換えられる。ただし , $\mu = E(X)$ とする。

定理 (1 変数) : X を確率変数とする。 $a + bX$ の分散は ,

$$\mathbf{V}(a + bX) = \mathbf{V}(bX) = b^2 \mathbf{V}(X), \quad (4.9)$$

となる。ただし , a, b は定数とする。

定理 (1 変数) : X を平均 μ , 分散 σ^2 の確率変数とする。 $Z = \frac{X - \mu}{\sigma}$ について ,

$$\mathbf{E}(Z) = 0, \quad \mathbf{V}(Z) = 1, \quad (4.10)$$

となる。この変換を標準化 , または , 基準化と呼ぶ。

定義 (2 変数) : X, Y を確率変数とする。 X と Y の共分散 $\sigma_{XY} = \mathbf{Cov}(X, Y)$ は ,

$$\sigma_{XY} = \mathbf{Cov}(X, Y) = \mathbf{E}((X - \mu_X)(Y - \mu_Y)), \quad (4.11)$$

となる。 $\mathbf{Cov}(X, Y)$ について , $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$ に対応する。

定義 (2 変数) : X, Y を確率変数とする。 X と Y の相関係数 ρ_{XY} は ,

$$\rho_{XY} = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{V}(X)} \sqrt{\mathbf{V}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad (4.12)$$

となる。ただし , $\sigma_X^2 = \mathbf{V}(X)$, $\sigma_Y^2 = \mathbf{V}(Y)$ とする。

定理 (2 変数) : X, Y を確率変数とする。 X と Y の共分散は ,

$$\mathbf{Cov}(X, Y) = \mathbf{E}(XY) - \mu_X \mu_Y, \quad (4.13)$$

と書き換えられる。 $\mathbf{E}(XY)$ について , $g(X, Y) = XY$ に対応する。

定理 (2 変数) : X, Y を確率変数とする。 $X + Y$ の分散は ,

$$\mathbf{V}(X + Y) = \mathbf{V}(X) + 2\mathbf{Cov}(X, Y) + \mathbf{V}(Y), \quad (4.14)$$

となる。

定理 (2 変数) : X, Y を確率変数とする。 X と Y が独立のとき , X と Y の共分散は ,

$$\mathbf{Cov}(X, Y) = 0, \quad (4.15)$$

となる。

定理 (2 変数) : X, Y を確率変数とする。 X と Y が独立のとき , $X + Y$ の分散は ,

$$\mathbf{V}(X + Y) = \mathbf{V}(X) + \mathbf{V}(Y), \quad (4.16)$$

となる。

定理 (多変数) : n 個の独立な確率変数 X_1, X_2, \dots, X_n を考える。このとき, $\sum_{i=1}^n c_i X_i$ の分散は,

$$\mathbf{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbf{V}(X_i), \quad (4.17)$$

となる。

4.3 正規分布について

確率変数 X の密度関数 $f(x)$ が,

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right),$$

となるとき， $f(x)$ を正規分布と呼ぶ。ただし， $\exp(x) = e^x$ である。 e は自然対数の底と呼ばれ， $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 2.7182818284590452353602874713\dots$ と定義される。

上記の正規分布は，

$$\mathbf{E}(X) = \mu, \quad \mathbf{V}(X) = \sigma^2,$$

となる（期待値の定義通りに計算すればよい）。

確率変数 X が上記の密度関数 $f(x)$ となるとき， $X \sim N(\mu, \sigma^2)$ と表す。 $X \sim N(\mu, \sigma^2)$ とは，「 X は平均 μ ，分散 σ^2 の正規分布に従う」と言う意味である。すなわち， N は正規分布 (**Normal distribution**) のアルファベットの頭文字で， \sim は「に従う」と読む。

定理（標準化，基準化）： (4.10) のように X を基準化する。

$$X \sim N(\mu, \sigma^2) \quad \text{のとき,} \quad Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \quad (4.18)$$

基準化によって， X がどの分布に従う確率変数であっても，平均 0 ，分散 1 に変換することができるということを (4.10) の定理は示している。(4.18) では，さらに進んで， X が正規分布であれば， Z も正規分布となるということを言っている。この証明は，変数変換（置換積分）を利用して証明することになる（本書では証明略）。平均 0 ，分散 1 の正規分布 $N(0, 1)$ は，標準正規分布と呼ばれる。

標準正規分布の確率分布表があれば，一般の正規分布の確率を得ることができる。すなわち， μ と σ^2 が既知とするとき， Z が z より大きい確率 $\mathbf{Prob}(Z > z)$ について， $\mathbf{Prob}(Z > z) = \mathbf{Prob}(X > \mu + z\sigma)$ となる。同様に， X が x より大きい確率 $\mathbf{Prob}(X > x)$ について，

$\text{Prob}(X > x) = \text{Prob}\left(Z > \frac{x - \mu}{\sigma}\right)$ となる。453 ページの付表 1 を用いると、標準正規分布の確率、すなわち、 $\text{Prob}(Z > z)$ を求めることができる。

(4.5) 式と (4.16) 式によって、 n 個の独立な確率変数 X_1, X_2, \dots, X_n が同一の分布（平均、分散が同じ分布）に従うとき、 $\sum_{i=1}^n c_i X_i$ の平均、分散は、

$$\mathbf{E}\left(\sum_{i=1}^n c_i X_i\right) = \mu \sum_{i=1}^n c_i, \quad \mathbf{V}\left(\sum_{i=1}^n c_i X_i\right) = \sigma^2 \sum_{i=1}^n c_i^2$$

となる。ただし、すべての i について $\mu = \mathbf{E}(X_i)$, $\sigma^2 = \mathbf{V}(X_i)$ とする。

n 個の独立な確率変数 X_1, X_2, \dots, X_n が同一の正規分布に従うものとする。すなわち、すべ

での i について $X_i \sim N(\mu, \sigma^2)$ とする。このとき，

$$\sum_{i=1}^n c_i X_i \sim N\left(\mu \sum_{i=1}^n c_i, \sigma^2 \sum_{i=1}^n c_i^2\right)$$

となる。すなわち，正規分布に従う確率変数の加重和もまた正規分布となる。この証明はそれほど簡単ではなく，積率母関数を利用して証明することになる（本書では証明略）。

特に，標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ を考えると，

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

となる（すべての i について， $c_i = \frac{1}{n}$ の場合を考えればよい）。

4.4 統計値・統計量，推定値・推定量について

1. 理論標本，理論観測値 $\Rightarrow X_1, X_2, \dots, X_n \Rightarrow$ 確率変数
2. 実現された標本，実現された観測値，実現値，観測値 $\Rightarrow x_1, x_2, \dots, x_n \Rightarrow$ 観測データ

1. 理論観測値 X_1, X_2, \dots, X_n の関数 \Rightarrow 統計量

2. すべての i について， $\mu = \mathbf{E}(X_i)$ と仮定する。

3. 母平均 μ の推定に使われる統計量 $\Rightarrow \mu$ の推定量

(a) $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は μ の推定量

(b) $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ は σ^2 の推定量

4. 実現された標本を用いて実際に計算された推定量の値 \implies 推定値

(a) $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ は μ の推定値

(b) $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ は σ^2 の推定値

5. μ や σ^2 の推定量の候補は無数に考えられる。

4.5 大数の法則と中心極限定理

4.5.1 大数の法則

大数の法則：その n 個の確率変数 X_1, X_2, \dots, X_n は互いに独立ですべて同じ分布にしたがい、すべての $i = 1, 2, \dots, n$ について $E(X_i) = \mu$ とする。 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ (すなわち、標本平均) とする。

$n \rightarrow \infty$ のとき、

$$\bar{X} \rightarrow \mu$$

となる。

大数の法則：その2 n 個の確率変数 X_1, X_2, \dots, X_n を考える（互いに独立である必要はなく，同じ分布である必要もない）。

$$\mu = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{E} \left(\sum_{i=1}^n X_i \right) < \infty, \quad \sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{V} \left(\sum_{i=1}^n X_i \right) < \infty$$

とする。

$n \rightarrow \infty$ のとき，

$$\bar{X} \rightarrow \mu$$

となる。

4.5.2 中心極限定理

中心極限定理：その 1 n 個の確率変数 X_1, X_2, \dots, X_n は互いに独立ですべて同じ分布にしたがい、すべての $i = 1, 2, \dots, n$ について $\mathbf{E}(X_i) = \mu$, $\mathbf{V}(X_i) = \sigma^2$ とする。 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ とする。

$n \rightarrow \infty$ のとき,

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0, 1)$$

となる。 $\mathbf{E}(\bar{X}) = \mu$, $\mathbf{V}(\bar{X}) = \sigma^2/n$ に注意せよ。

中心極限定理：その 2 n 個の確率変数 X_1, X_2, \dots, X_n を考える（互いに独立である必要はなく，同じ分布である必要もない）。

$$\mu = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{E} \left(\sum_{i=1}^n X_i \right) < \infty, \quad \sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{V} \left(\sum_{i=1}^n X_i \right) < \infty$$

とする。

$n \rightarrow \infty$ のとき，

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0, 1)$$

となる。

4.6 推定量の望ましい性質

$\hat{\alpha}$, $\hat{\beta}$ の性質を求めるために

4.6.1 不偏性

ある母集団のある母数 θ に対して, θ の推定量として $\hat{\theta}$ を考える。このとき,

$$\mathbf{E}(\hat{\theta}) = \theta$$

となるとき, $\hat{\theta}$ は θ の不偏推定量であると言う。 $\hat{\theta}$ は不偏性を持つと言う。 $\mathbf{E}(\hat{\theta}) - \theta$ は偏りと定義される。

n 個の確率変数 X_1, X_2, \dots, X_n に関して, すべての $i = 1, 2, \dots, n$ について $\mathbf{E}(X_i) = \mu$ とするとき, 標本平均 \bar{X} は μ の不偏推定量である。

証明:

$$\mathbf{E}(\bar{X}) = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

このように, $\mathbf{E}(\bar{X}) = \mu$ なので, 標本平均 \bar{X} は μ の不偏推定量となる。

4.6.2 有効性 (最小分散性)

ある母数 θ に対して, $\hat{\theta}_1$ と $\hat{\theta}_2$ の 2 つの不偏推定量を考える。このとき, $\mathbf{V}(\hat{\theta}_1) \leq \mathbf{V}(\hat{\theta}_2)$ が成り立つとき, $\hat{\theta}_1$ は $\hat{\theta}_2$ より有効であると言う。

ある母数 θ に対して、可能なすべての不偏推定量を考え、 $\hat{\theta}$ が最も小さな分散を持つ不偏推定量であるとする。このとき、 $\hat{\theta}$ を最小分散不偏推定量、または、最良不偏推定量と言う。

一般に、有効推定量が存在するとは限らない。代わりに、推定量 $\sum_{i=1}^n c_i X_i$ (すなわち、線形推定量) の中で最も小さい分散を持つ推定量を求めることを考える。この推定量を最良線形不偏推定量と呼ぶ。

標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は不偏推定量の中で最も小さな分散を持つ推定量である。

証明：

期待値を取ると、

$$\mathbf{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbf{E}(X_i) = \mu \sum_{i=1}^n c_i$$

となる。 $\sum_{i=1}^n c_i X_i$ が不偏推定量になるためには $\sum_{i=1}^n c_i = 1$ が必要となる。分散は、

$$\mathbf{V}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n \mathbf{V}(c_i X_i) = \sum_{i=1}^n c_i^2 \mathbf{V}(X_i) = \sigma^2 \sum_{i=1}^n c_i^2$$

となる。

したがって、最良線形不偏推定量を得るためには、 $\sum_{i=1}^n c_i = 1$ の条件のもとで、 $\sum_{i=1}^n c_i^2$ を最小にする c_1, c_2, \dots, c_n を求めればよい。ラグランジェ未定乗数法を用いれば、 $c_i = \frac{1}{n}$ が得られる。

4.6.3 一貫性

ある母数 θ について推定量 $\hat{\theta}$ を考える。 n 個の標本から構成された推定量を $\hat{\theta}^{(n)}$ と定義する。数列 $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(n)}, \dots$ を考える。十分大きな n について、 $\hat{\theta}^{(n)}$ が θ に確率的に収束するとき、 $\hat{\theta}$ は θ の一貫推定量であると言う。

$$\hat{\theta} \rightarrow \theta, \quad \text{または,} \quad \text{plim } \hat{\theta} = \theta,$$

と表現する。 **plim** とは **probability limit** の略である。

$E(\hat{\theta}) = \theta$ とする。 $n \rightarrow \infty$ のとき $V(\hat{\theta}) \rightarrow 0$ が成り立てば、 $\hat{\theta}$ は θ の一貫推定量である。

μ の推定量 \bar{X} を調べる。

$$\mathbf{E}(\bar{X}) = \mu$$

である。

$$\mathbf{V}(\bar{X}) = \frac{\sigma^2}{n}$$

となる。 $n \rightarrow \infty$ のとき、

$$\mathbf{V}(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0$$

となるので、 \bar{X} は μ の一致推定量であると言える。

4.7 χ^2 分布

m 個の確率変数 Z_1, Z_2, \dots, Z_m は, 互いに独立な標準正規分布に従うものとする。このとき, $Y = \sum_{i=1}^m Z_i^2$ は, 自由度 m の χ^2 分布に従う。

$Y \sim \chi^2(m)$, または, $Y \sim \chi_m^2$ と表記する。

χ^2 (カイ二乗) 分布表から確率を求める。

$Y \sim \chi^2(m)$ のとき, $\mathbf{E}(Y) = m$, $\mathbf{V}(Y) = 2m$ となる。(証明略)

1. 2つの独立な χ^2 分布からの確率変数 X, Y を考える。 $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ とする。このとき, $Z = X + Y \sim \chi^2(n + m)$ となる。(証明略)
2. n 個の独立な確率変数 X_1, X_2, \dots, X_n が同一の正規分布 $N(\mu, \sigma^2)$ に従うものとする。

3. $\frac{X_i - \mu}{\sigma} \sim N(0, 1)$ なので, $\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$ となる。
 $\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \dots, \frac{X_n - \mu}{\sigma}$ はそれぞれ独立なので,

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

となる。

4. μ を \bar{X} に置き換えると,

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

となる。(証明は後述)

さらに,

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

を定義すると,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

となる。 S^2 は σ^2 の不偏推定量である (後述)。

5. すなわち,

$$\mathbf{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = n-1 \quad \mathbf{V}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1),$$

となる。

4.8 t 分布

正規分布の重要な定理： n 個の独立な確率変数 X_1, X_2, \dots, X_n が同一の正規分布 $N(\mu, \sigma^2)$ に従うものとする。このとき，

$$\sum_{i=1}^n c_i X_i \sim N\left(\mu \sum_{i=1}^n c_i, \sigma^2 \sum_{i=1}^n c_i^2\right)$$

となる。ただし， c_1, c_2, \dots, c_n は定数とする。

t 分布： Z を標準正規分布， Y を自由度 m の χ^2 分布に従い，両者は独立な確率変数とする。このとき， $U = \frac{Z}{\sqrt{Y/m}}$ は，自由度 m の t 分布に従う。

$U \sim t(m)$ ，または， $U \sim t_m$ と表記する。

$U \sim t(m)$ のとき , $m > 1$ について $E(U) = 0$, $m > 2$ について $V(U) = \frac{m}{m-2}$ となる。(証明略)

t 分布表から確率を求める。(表 9.1.3 を見よ)

1. ゼロを中心に左右対称。($E(U) = 0$)

2. t 分布は , 標準正規分布より裾野の広い分布 (なぜなら , $V(U) = \frac{m}{m-2} > 1$)

3. $m \rightarrow \infty$ のとき , $t(m) \rightarrow N(0, 1)$ となる。(期待値は $m > 1$ について $E(U) = 0$, 分散は $V(U) = \frac{m}{m-2} \rightarrow 1$)

4.9 標本平均 \bar{X} の分布

X_1, X_2, \dots, X_n の n 個の確率変数は、互いに独立で、平均 μ 、分散 σ^2 の正規分布に従うものとする。

1. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ なので、 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ となる。

2. $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$ である。(証明は略)

3. $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ と $\frac{(n-1)S^2}{\sigma^2}$ は独立。(証明は略)

すなわち、 \bar{X} と S^2 は独立。

4. したがって、

$$\frac{\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} / n - 1}} = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$$

を得る。

重要な結果は、

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$$

ただし、 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 、 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ である。

σ^2 を S^2 に置き換えると、正規分布から t 分布になる。

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \quad \Longrightarrow \quad \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n - 1)$$

4.10 区間推定 (信頼区間)

\bar{X} の分布を利用して, μ の信頼区間を求める。

1. \bar{X} の分布は以下の通り。

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n - 1)$$

となる。

2. $t_{\alpha/2}(n-1)$, $t_{1-\alpha/2}(n-1)$ を自由度 $n-1$ の t 分布の上から $100 \times \frac{\alpha}{2}$ % 点, $100 \times (1 - \frac{\alpha}{2})$ % 点の値とする。このとき,

$$\mathbf{Prob}\left(t_{1-\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha$$

となる。ただし, 自由度と α が決まれば, $t_{\alpha/2}(n-1)$, $t_{1-\alpha/2}(n-1)$ は t 分布表から得られる。

3. t 分布は左右対称なので,

$$t_{1-\alpha/2}(n-1) = -t_{\alpha/2}(n-1) \qquad t_{\alpha/2}(n-1) = |t_{1-\alpha/2}(n-1)|$$

$$t_{1-\alpha/2}(n-1) = -|t_{\alpha/2}(n-1)|$$

となる。

4. 書き直して,

$$\text{Prob}\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

となる。

5. μ が区間 $(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}})$ にある確率は $1 - \alpha$ である。

6. 推定量 \bar{X}, S^2 をその推定値 \bar{x}, s^2 で置き換える。ただし, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ とする。

7. 区間 $(\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}})$ を信頼係数 $1 - \alpha$ の信頼区間といい、 $\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ を信頼下限, $\bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ を信頼上限と呼ぶ。

4.11 仮説検定

\bar{X} の分布を利用して, μ の仮説検定を行う。

1. 帰無仮説 $H_0 : \mu = \mu_0$ 対立仮説 $H_1 : \mu \neq \mu_0$

2. 帰無仮説 $H_0 : \mu = \mu_0$ が正しいもとでの分布は,

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n-1)$$

となる。

3. $\text{Prob}\left(t_{1-\alpha/2}(n-1) < \frac{\bar{X} - \mu_0}{S / \sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha$

$t_{\alpha/2}(n-1)$, $t_{1-\alpha/2}(n-1)$ をそれぞれ自由度 $n-1$ の t 分布の上から $100 \times \frac{\alpha}{2}$ % 点, $100 \times \frac{1-\alpha}{2}$ % 点の値とする。

自由度と α が決まれば, $t_{\alpha/2}(n-1)$, $t_{1-\alpha/2}(n-1)$ は t 分布表から得られる。

4. α を有意水準と呼ぶ。慣習的に $\alpha = 0.01, 0.05$ が使われる。

5. $-t_{\alpha/2}(n-1) > \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$, または, $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{\alpha/2}(n-1)$ ならば, 帰無仮説 $H_0: \mu = \mu_0$ は, 分布の端にあり, 起こりにくいと考ええる。

⇒ 有意水準 α で帰無仮説 $H_0: \mu = \mu_0$ を棄却する。

6. 実際の検定手続:

(a) \bar{X}, S^2 を実績値で置き換えて,

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

を得る。ただし, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ とする。

(b) $-t_{\alpha/2}(n-1) > \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, または , $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha/2}(n-1)$ ならば , 有意水準 α で帰無仮説 $H_0 : \mu = \mu_0$ を棄却する。