計量経済学における相関有無の判定基準への疑問

大嶋　一人＊

(2019年1月7日受理)

1．まえがき

計量経済学[1]の分野においては特定の２変量に相関があるか否かを検定することがよく行われている。相関の有無を検定する方法としては、２変量の相関係数を算出し、相関係数から、統計学におけるｔ分布に従う量を構成し、t検定を行うものがよく使われている。この解析方法は、２変量が、２つの独立な正規分布に従うと仮定した場合に証明される数学の定理に基づいていると考えられる。しかしながら、経済学に現れる２変量がともに正規分布に従うという保証があると考えることは難しいように思える。そこで、本論文では２変量が別の分布、具体的には独立な２つの有限区間の一様分布、に従うと仮定したときに、２変量の相関の有無に関する相関係数に対する判定基準がどのように変更されるかを見る。そのことにより、計量経済学の標準的な教科書で行われている相関係数を用いたｔ検定が妥当なものか否かを検討する。

2．統計学における相関係数に対する定理

計量経済学において相関の有無を検定する際に使う数学の定理は次の２つのものであると考えられる。

定理１[2]　２つの独立な正規母集団より大きさｎの標本をとりその相関係数をｒ($=\frac{s_{xy}}{s_x s_y}$)とすればｒの確率密度関数f(r)は次の形となる。

$f(r)=c(n)(1-r^2)^{(n-4)/2}/\pi^{1/2}$,

ここで

$c(n)=\Gamma((n-1)/2)/\Gamma((n-2)/2)$

ただし、Γ(n)はガンマ関数である。

定理２[3]　ｒが定理１の確率密度関数に従うとき、変量$T=(n-2)^{1/2}r/(1-r^2)^{1/2}$は自由度n-2のｔ分布に従う。

3．正規分布と一様分布

このセクションでは大きさｎの標本をとりだすときの元の母集団が２つの独立な正規母集団であるときと、２つの独立な有限区間で一様分布をする場合とで相関係数ｒの確率密度関数に差異が生ずるか否かを検討する。有限区間で一様分布をする場合については、理論的解析は困難と考え、数式処理ソフトMathematica5が発生する区間[0,1]の一様な疑似乱数を用いて、数値的シミュレーションを行う。有効数字については考慮しないことにする。

n=4の場合

母集団が正規分布に対しては$f(r)=1/2$ $(-1<r<1)$と一様分布となる。一方、母集団が区間[0,1]の一様分布の場合はMathematicaの疑似乱数により、１０万個の相関係数ｒを発生させた。区間[-1,1]を４０等分して（区間幅 0.05)、値の小さい方から区間に１から４０の名前を付ける。例えば、区間１は区間[-1,-0.95]を表す。ｒの分布の様子を表１に示す。結果はほとんど一様であり、正規母集団の場合とほとんど差異はないと考えられる。

n=8の場合

母集団が正規分布に対しては$f(r)=15(1-r^2)^2/16$ $(-1<r<1)$となる。表２にf(r)を区間[-1,1]を４０等分して区間積分した値と、母集団が区間[0,1]の一様分布の場合にmathematicaの疑似乱数により、１０万個の相関係数ｒを発生させ、区間[-1,1]を２０等分した時の、ｒの分布のシミューレーション値を載せる。この場合も両者にはほとんど差異はないと考えられる。

n=11の場合

n=4,8の場合と同様に表３に結果を示す。r=∓1付近以外では正規分布の場合と一様分布の場合とで大きな差は生じないと考えられる。r=∓1付近は度数が少ないため、相対的ズレが大きくなっているとも考えられるが、r=∓1付近では若干の考慮が必要になる可能性のあることがわかる。

n=20の場合

同様に表４に結果を示す。この場合は、正規分布と、一様乱数とでｒの分布の仕方が大きく異なることがわかる。一様分布の場合は、正規分布の場合に比べてｒは０付近により多く集まっている。従って、ｒの絶対値の大きい方から5%の値は正規分布の方がより大きくなると考えられる。

＊電子メディア工学科

表-１　n=4 の場合の度数分布の割合

区間名	正規分布	一様分布	相対ズレ%
1	0.025	0.02488	−0.48
2	0.025	0.02581	3.24
3	0.025	0.02472	−1.12
4	0.025	0.02452	−1.92
5	0.025	0.02496	−0.16
6	0.025	0.0239	−4.4
7	0.025	0.02525	1
8	0.025	0.02498	−0.08
9	0.025	0.02486	−0.56
10	0.025	0.0254	1.6
11	0.025	0.02408	−3.68
12	0.025	0.02483	−0.68
13	0.025	0.02536	1.44
14	0.025	0.02541	1.64
15	0.025	0.02516	0.64
16	0.025	0.02527	1.08
17	0.025	0.02519	0.76
18	0.025	0.02454	−1.84
19	0.025	0.0248	−0.8
20	0.025	0.02474	−1.04
21	0.025	0.02485	−0.6
22	0.025	0.02514	0.56
23	0.025	0.0251	0.4
24	0.025	0.02479	−0.84
25	0.025	0.02492	−0.32
26	0.025	0.02423	−3.08
27	0.025	0.02495	−0.2
28	0.025	0.02526	1.04
29	0.025	0.02518	0.72
30	0.025	0.02501	0.04
31	0.025	0.02543	1.72
32	0.025	0.02626	5.04
33	0.025	0.02508	0.32
34	0.025	0.0254	1.6
35	0.025	0.02401	−3.96
36	0.025	0.02506	0.24
37	0.025	0.02501	0.04
38	0.025	0.02535	1.4
39	0.025	0.02476	−0.96
40	0.025	0.02555	2.2
計	1	1	0(平均値)

表-2　n=8 の場合の度数分布の割合

区間名	正規分布	一様分布	相対ズレ%
1	0.0001504	0.00019	26.29
2	0.0010077	0.00115	14.12
3	0.0026003	0.00299	14.99
4	0.0048016	0.00512	6.63
5	0.0074923	0.00796	6.24
6	0.0105596	0.01078	2.09
7	0.0138983	0.01356	−2.43
8	0.0174098	0.01673	−3.9
9	0.0210028	0.0204	−2.87
10	0.0245928	0.0238	−3.22
11	0.0281026	0.02806	−0.15
12	0.0314618	0.0319	1.39
13	0.0346071	0.0345	−0.31
14	0.0374823	0.03767	0.5
15	0.0400381	0.03943	−1.51
16	0.0422325	0.04358	3.19
17	0.0440301	0.04315	−2
18	0.045403	0.04542	0.04
19	0.04633	0.04614	−0.41
20	0.046797	0.04736	1.2
21	0.046797	0.04751	1.52
22	0.04633	0.04694	1.32
23	0.045403	0.04513	−0.6
24	0.0440301	0.04497	2.13
25	0.0422325	0.04124	−2.35
26	0.0400381	0.03931	−1.82
27	0.0374823	0.03761	0.34
28	0.0346071	0.03424	−1.06
29	0.0314618	0.03138	−0.26
30	0.0281026	0.02787	−0.83
31	0.0245928	0.0243	−1.19
32	0.0210028	0.02074	−1.25
33	0.0174098	0.01713	−1.61
34	0.0138983	0.01447	4.11
35	0.0105596	0.01052	−0.38
36	0.0074923	0.0076	1.44
37	0.0048016	0.00474	−1.28
38	0.0026003	0.00309	18.83
39	0.0010077	0.00118	17.1
40	0.0001504	0.00014	−6.94
計	1	1	2.1775

表-3　n=11 の場合の度数分布の割合

区間名	正規分布	一様分布	相対ズレ%
1	0.0000038	0.00001	−163.15789
2	0.00007618	0.00009	−18.141244
3	0.00038008	0.0005	−31.552983
4	0.00109515	0.00139	−26.923253
5	0.00237098	0.00277	−16.829328
6	0.00430929	0.00483	−12.083429
7	0.00695749	0.00667	4.13209361
8	0.0103086	0.00971	5.80680209
9	0.0143054	0.01482	−3.597243
10	0.0188464	0.01851	1.78495628
11	0.0237946	0.02275	4.3900717
12	0.0289862	0.02848	1.74634826
13	0.0342401	0.03488	−1.8688614
14	0.03963	0.03817	3.68407772
15	0.0441803	0.04206	4.79919783
16	0.0485009	0.04878	−0.5754532
17	0.0521685	0.05229	−0.2328992
18	0.0550463	0.0566	−2.822533
19	0.0570267	0.05669	0.59042519
20	0.0580358	0.05855	−0.8860048
21	0.058036	0.05868	−1.1096561
22	0.0570267	0.05852	−2.618598
23	0.0550463	0.05594	−1.6235424
24	0.0521685	0.05316	−1.9005722
25	0.0485009	0.04899	−1.0084349
26	0.0441803	0.04326	2.08305512
27	0.03963	0.03964	−0.0252334
28	0.0342401	0.0339	0.99327981
29	0.0289862	0.02782	4.02329384
30	0.0237946	0.02276	4.34804535
31	0.0188464	0.01901	−0.8680703
32	0.0143054	0.01388	2.97370224
33	0.0103086	0.0104	−0.8866383
34	0.00695749	0.00673	3.26971365
35	0.00430929	0.00421	2.30409186
36	0.00237098	0.00272	−14.720495
37	0.00109515	0.00126	−15.052733
38	0.00038008	0.00044	−15.766625
39	0.00007618	0.00013	−70.648464
40	0.0000038	0	100
計	1	1	−6.449

表-４　n=20 の場合の度数分布の割合

区間名	正規分布	一様分布	相対ズレ%
1	0	0	−
2	0.00000003	0	100
3	0.00000101	0	100
4	0.00001042	0	100
5	0.00005858	0	100
6	0.00022499	0	100
7	0.00066485	0.00001	98.4958968
8	0.00162159	0.00008	95.0665705
9	0.00341343	0.00024	92.9689491
10	0.00638989	0.00095	85.1327644
11	0.0108633	0.00234	78.4595841
12	0.0170288	0.0056	67.1145354
13	0.0248905	0.01134	54.4404492
14	0.0342115	0.02096	38.7340514
15	0.0445027	0.03501	21.3306159
16	0.055055	0.05146	6.5298338
17	0.0650138	0.07098	−9.1768209
18	0.0734853	0.08827	−20.119262
19	0.0796561	0.10342	−29.83312
20	0.0829083	0.1103	−33.03855
21	0.0829083	0.11156	−34.558301
22	0.0796561	0.10155	−27.485528
23	0.0734853	0.08761	−19.221123
24	0.0650138	0.06979	−7.3464403
25	0.055055	0.05271	4.25937699
26	0.0445027	0.03413	23.308024
27	0.0342115	0.02165	36.7171857
28	0.0248905	0.01117	55.1234407
29	0.0170288	0.00568	66.644743
30	0.0108633	0.00189	82.6019718
31	0.00638989	0.00086	86.5412394
32	0.00341343	0.00037	89.1604632
33	0.00162159	0.00006	96.2999278
34	0.00066485	0.00001	98.4958968
35	0.00022499	0	100
36	0.00005858	0	100
37	0.00001042	0	100
38	0.00000101	0	100
39	0.00000003	0	100
40	0	0	−
計	1	1	−

n=4,8,11,20 の場合の解析から正規分布と一様分布とで r の分布の仕方が必ずしも同じではないことがわかった。そこで正規分布と一様分布とで上位 2.5%となる r の値を n=4 から n=20 まで求めてみる。表５において、正規分布とあるのは$\int_r^1 f(r)dr = 0.025$となる r の値であり、これは t 分布における上位 2.5%を与える T の値$t_{n-2}(0.025)$と同じものである。一様分布とあるのは Mathematica による１万組の一様分布の疑似乱数$x_1, x_2, \cdots, x_n$と$y_1, y_2, \cdots, y_n$から作った１万個の r のうち、大きい方から２５０番目の r の値である。同様に１０万個とあるのは同様に作成した１０万組の r のうち、大きい方から２５００番目の r の値である。n=11 あたりまでは正規分布と一様分布とでほとんど差はないが、n=12 からは両者で無視できない差があることがわかる。なお、表５の Mathetica によるシミュレーションは SeedRandom[n]なる疑似乱数発生のための種を各 n に対して与えてあるので再現可能である。n=20 では正規分布の上位 2.5%に対応する r の値は一様分布１０万個の場合の上位約３５０番目に相当しこれは全体の 0.35%に相当する。

表−５　上位 2.5%となる r の値

n	正規分布	一様分布	10万個
4	0.95	0.94917	0.951455
5	0.8783393	0.891548	0.885194
6	0.811401	0.807051	0.812652
7	0.7544924	0.753817	0.759711
8	0.706734	0.710232	0.705412
9	0.6663838	0.667059	0.669485
10	0.631897	0.630183	0.635067
11	0.6020689	0.603257	0.601
12	0.575983	0.557351	0.554823
13	0.5529425	0.501279	0.509538
14	0.532413	0.474213	0.472945
15	0.5139775	0.436332	0.444536
16	0.497309	0.417143	0.416123
17	0.482146	0.386996	0.39404
18	0.4682774	0.363005	0.374223
19	0.4555305	0.354283	0.351806
20	0.4437635	0.330753	0.336694

４．まとめと考察

計量経済学の教科書において、２つの変量の相関の有無を判定する基準として、相関係数の値から、t 分布による帰無仮説により、相関の有無を判定するものがあった。これは、母集団として独立な２つの正規母集団を仮定したときの相関係数の値の分布に対する数学の定理に基づくものと考えられる。現実の変数は正規分布に従うものばかりとは限らないと思われるため、２つの独立な一様分布を母集団としたときの相関係数の分布をコンピュータシミュレーションにより調べた。データ数 n が小さいときは、２つの分布においてほとんど差異は見られなかったが、n=12 以上では両者に無視できない差異があることがわかった。２つの変量の相関を判定するときに、相関なしと仮定する帰無仮説を棄却する 5%棄却域はデータ数 n が１２以上では正規母集団の方が一様分布の場合よりも狭いことがわかった。データ数 n=20 の一様分布に対して、正規分布の 5%両側棄却域を設定したならば、相関なしという帰無仮説棄却した場合に間違っている確率は 0.7%以下である。従って、計量経済学の教科書に書いてある相関有無の判定基準は、現実の変数の分布が仮に一様分布に近いものであるとすると、データ数 n がある程度大きい場合には、かなり厳しめのものである可能性がある。つまり、t 検定で相関なしという帰無仮説が棄却されたならば、棄却された場合に間違っている確率はかなり小さいと考えられる。計量経済学の教科書に書かれてある相関の有無に関する t 検定による判定基準に関しては疑問が残る。

謝辞

中村充氏（現九州大学経済学部経済工学科学生）には、初期の段階において、プログラムのミスに気付かされてくれたことを感謝します。

参考文献

１）白砂堤津耶、初歩からの計量経済学、日本評論社、２００７年.
２）丸山儀四郎、確率および統計、共立出版、１９５６年.
３）国沢清典、確率統計演習２、培風館、１９６６年.

Study of Criteria for the Presence or Absence of Correlation in Econometrics

Kazuto Oshima

In econometrics it is important to distinguish whether two variables are correlated or not. Distribution of correlation coefficient of n pairs of samples from two independent normal populations is theoretically known. Based on the distribution and the value of the sample correlation coefficient we decide whether the two variables are correlated or not. We compute distribution of correlation coefficient of n pairs of samples from two independent uniform populations by computer simulation. We see that we cannot distinguish the two distributions of the correlation coefficients when $n<11$. When $n>11$, however, the two distributions are different definitely.