Student's t-table modification for the linear correlation coefficients estimation in the small samples cases

Ilya Igushkin
Engineering Institute
Kazan Federal University
ilyha133@mail.ru

Natalya Verzun
Department of information systems and technologies St. Petersburg State University of Economics
verzun.n@unecon.ru

Anatoly Shikhalev
Engineering Institute
Kazan Federal University
shihalev_48@mail.ru
Mikhail Kolbanyov
Department of information systems and technologies
St. Petersburg State University of Economics
mokolbanev@mail.ru

Dmitry Vorontsov
Engineering Institute Kazan Federal University
DPVoroncov@kpfu.ru
Irina Akhmetova
Institute of Management, Economics an d Finance Kazan Federal University
iraahmetova@mail.ru

VIII International Conference on Information Technology and Nanotechnology (ITNT-2022)

Abstract

In the linear correlation coefficient calculations for the statistical significance estimation is often used the famous Chaddock's scale of the relationship between the studied phenomena with the characteristics like "weak", "medium", "visible", "high", "very high", and for the significance evaluation used the Student's t-test table with the fixed alpha-level (α = 0.10; 0.05; 0.01) and with the available degrees of freedom. If the calculated values of the linear correlation coefficient are less than critical, then as usual the researchers will increase the number of initial observations N. However, in an unstable economics period this is not always possible. Therefore, we have the task of estimating the confidence interval for the calculated value of the linear correlation coefficient, especially to its lower bound (of confidence level): what if the calculated module of linear correlation coefficient will be met the reliability requirements according to the famous t-Student criterion? Moreover, this means that the significance in (both) cases is assessed on a step by step manner which is not fully expedient. For the decision of this problem we propose to use the modified Student's scale and then it is also possible to use the Chaddock's scale. As the raw data we use statistical aggregates with the limited size; after some modifications we create on this variation series and apply the consent criteria. For the noted problem solving we also must note the received equations for LCC error and non-strict inequality. Some part of the results was obtained with the program made in FoxPro 2.5 which was created by the member of the author's team (Anatoly Shikhalev).

INTRODUCTION

In the last decades due to financial and other economic crises there has been an objective reduction in the number of experimental data N due to unstable economic indicators (see, for example, in [8] - [10], etc.), widely noted in the literature. If the size of the study sample N cannot be expanded due to objectively reasonable causes, is bound to occur the question about the degree of linkage significance between the studied indicators, for example, the linear correlation coefficient, which involves a preliminary analysis of the hypothesis that the both studied statistical universes are distributed by the Gaussian normal law of distribution (NLR) of random variables (RV), and only then it could be the calculation of the linear correlation coefficient module and its sign. And if the value of the linear correlation coefficient module will be small and, moreover, insignificant on Student's t-test, and using the possible methods for increasing the sample (N number) is almost impossible, then we have the question of finding its minimum module significance in the Student's t- criterion.

The basic (reference) data description

As the initial (i.e. source) data which could describe some main characteristics of turbulent and unstable economy [8], [9], [10] we will take the statistical aggregates of limited volume: semester attendance (now big number of students must work for paying his dues, etc.) X = {xi}, i = 1, n and academic performance (attainment records) Y = {yi}, i = 1, n (see in Table 1).

Table I. Number of absence (or disruptions in university attendance) and results of semester examinations (in the traditional scale and in grade points) .

i	*Number of absences in the university X, pcs.*	*Results of examinations Y (in traditional five point grading scale / in grade points)*
1	1	4 / 83
2	3	4 / 80
3	0	4 / 80
4	2	3 / 68
5	2	3 / 68
6	1	5 / 90
7	2	4 / 78
8	2	4 / 80
9	1	4 / 75
10	2	5 / 90
N = 10	16	40 /787

On the authors of this article opinion the example in the Table 1 also reflects the socio-economic phenomena as unstable or weakly stable.

Materials and proposed methods modification

For testing the studied sets elements with the names like X and Y it is necessary and enough to make the so-called empirical data adjustment (or data fitting) and then to use the Pearson's χ2 consent criterion which operates with only two parameters: empirical frequencies fi and the theoretical frequencies fitheor. We will get the empirical frequencies from the variation series (VS) which we have constructed sequentially for the X and Y sets and the theoretical frequencies we will receive basing on the famous rules as a result of further calculations. If the elements of both variables are distributed on the normal distribution law (NDL) then we could estimate the further application of the linear correlation coefficient for studying the direction and value of the linkage (i.e. the correlation) between the X and Y sets is correct. The creation of variable set for the X variable elements (columns 1 - 4 in the Table II) for which it is necessary to estimate the value of step h = R/n, where R = x_{max} - x_{min} is the range of the studied sample within the X named set and the number of the created variable set intervals for the set X is determined by the famous approximate Sturges's formula: n≈1 + 3,322 • lg (N) = 4.322 ≈ 4 (in some sources it is recommended to round precisely "with a disadvantage"). Then h = (3 - 0)/4 = 0.75 (i.e. omissions). The step value from the expression (3) will be the ratio from dividing the R range of sample X to the number of intervals found: h = R/n = (3 - 0)/4 = 0.75 (skip), as a result of which random variable for the original set X which is displayed in random variables look as n = 4 row intervals: for i = 1 (0 - 0.75); for i = 2 (0.75 - 1.50); for i = 3 (1.50-2.25); for i = 4 (2.25 - 3.00). Empirical frequencies will be: f_1 = 1; f_2 = 3; f_3 = 5; f_4 = 1, and the midpoints of the intervals $x_i^{av.}$ are 0.375; 1.125: 1.875; 2.625. In other words, for filling the column 5 of Table II we will need the values of two parameters: weighted average and standard deviation as well as the constant value for this variable con = Nh/σ. So x_{weigh}^{av}= ∑ x_i^{av}· f_i/ ∑ f_i= 1.575 (skip); dispersion D = D = σ^2= ∑(x_i^{av}- x_{weight}^{av})^2·f_i / ∑ f_i= 0.36 (skip.^{2}); σ = (D) 1/2 = (0.36) 1/2 = 0.6 (skip); con = Nh/a = 10 • 0.75/0.6 = 12.5 . Next, we must use here Pearson's statistical table [11] for which purpose it is necessary to estimate the value of degrees of freedom df = n-r-1, where "n" is the number of variants-rows of variable set X, "r" is the number of variables which we have used when were looking for the theoretical frequencies (there are two of them i.e. xweighav and σ, therefore, r = 2).

Table II. Variation series for variable X (1-4 rows) and f_i^{theor} obtaining

i	x_i^{beg} - x_i^{end} (+)	f_i	x_i^{av}	x_i^{av}– x_{weigh}^{av}	$t = (x_i^{av} - x_{weigt}^{av})/\sigma$	$\varphi(t)$ on App.1 [4]	$f_i^{theor} = \varphi(t)\cdot con$
1	2	3	4	5	6	7	8
1	0.00–0.75	/ 1	0.375	-1.200	-2.000	0.0540	
2	0.75–1.50	/// 3	1.125	-0.450	-0.750	0.3011	
3	1.50–2.25	///// 5	1.875	0.300	0.500	0.3521	
4	2.25–3.00	/ 1	2.625	1.050	1.750	0.0863	
	Sums:	10	-	-	-	-	9.919≈10

With the creation and filling out a standard table for calculating the Pirson's χ2 criterion as a function of fi and fitheor.., we get χ2calc. = 0.50, which must be compared with its corresponding table value χ2tab. (df = 1; α = 0,05) = 3,84. As we could see, then df = 4 - 2 - 1 = 1, and χ2table (df = 1; α = 0.05) = 3.84. If χ2calc = 0.50 < χ2calc = 3.84, then we could estimate the discrepancies between fi and fitheor as the random and our preliminary "null hypothesis" about the normal distribution of the X set elements as the variable set is not refuted. The same is with the parameters of elements of another random variable under the Y name of the set (academic performance): χ2calc = 1.33 < χ2calc = 3.84, which are also distributed according to the Gauss distribution law. Therefore, the association of X variables (missing semester classes, in pcs) and Y (semester final performance, scores) could be investigated using a statistical apparatus i.e. the linear correlation coefficient (LCC). Further basing on this (2) results we are ascertained that the received linear correlation coefficient is non-significant: tρ = 0.571 < t (k = 8; α = 0.05) = 2.306; validity < 95%. But is it really insignificant? Maybe there are other critical values for comparing the resulting linear correlation coefficient module with some other value, which propose to be found in this article. So, in order to estimate the minimum critical value of the linear correlation coefficient (LCC) according to the Student's t-test with reliability P = 95%, after calculating the LCC module according to formula (1) it is necessary and sufficient to write the minimum critical value from the 3rd column of the table, and then compare it with the calculated critical value tρ on formula (2).

If the non-strict inequality will be realized tρ=tcalc.. ≥ tmincrit.(k; α = 0.05) (10), then the calculated value of the linear correlation coefficient module which is equal to | 0.190 | ≈ | 0.20 | could be considered as the statistically significant.

Table IV. Minimum significant values of linear correlation modules coefficients with the n volumes and α = 0,05

Degrees of freedom (number) df = k	*Significance level α = 0,05*	*Statistically least significant value of LCC $\lvert\rho_{XY}\rvert$*	*Standard deviation value of LCC σ_ρ*	*Upper fiducial (right) confidence intervals limits for the minimum LCC*	*Remarks*
1	2	3	4	5	6
		0.9460	0.074440	$\lvert\rho_{XY}\rvert$>1	
		0.8188	0.190294	$\lvert\rho_{XY}\rvert$>1	
		0.7340	0.230632	$\lvert\rho_{XY}\rvert$>1	
1	12.706	0.6753	0.243244	$\lvert\rho_{XY}\rvert$>1	meaningless
2	4.3027	0.6313	0.245568	$\lvert\rho_{XY}\rvert$>1	meaningless
3	3.1825	0.5962	0.243636	$\lvert\rho_{XY}\rvert$>1	meaningless
4	2.7764	0.5671	0.239840	$\lvert\rho_{XY}\rvert$>1	meaningless
5	2.5706	0.5425	0.235242	$\lvert\rho_{XY}\rvert$>1	meaningless
6	2.4469	0.5211	0.230355	$\lvert\rho_{XY}\rvert$>1	meaningless
7	2.3646	0.5023	0.225439	$\lvert\rho_{XY}\rvert$>1	meaningless
8	2.3060	0.4856	0.220612	0.9711	meaningless
9	2.2622	0.4705	0.215950	0.9410	meaningless
10	2.2281	0.4569	0.211475	0.9137	meaningless
11	2.2010	0.4444	0.207204	0.8889	alm.wh.domain
12	2.1788	0.4350	0.203133	0.8660	alm.wh.domain
13	2.1604	0.4224	0.199260	0.8448	alm.wh.domain
14	2.1448	0.4126	0.195573	0.8252	alm.wh.domain
15	2.1315	0.4035	0.192063	0.8070	relat. acceptab.
16	2.1199	0.3950	0.188720	0.7900	relat. acceptab.
17	2.1098	0.3870	0.185532	0.7740	relat. acceptab.
18	2.1009	0.3795	0.182493	0.7590	relat. acceptab.
19	2.0930	0.3725	0.179589	0.7449	relat. acceptab.
20	2.0860	0.3658	0.176814	0.7316	relat. acceptab.
21	2.0796	0.3594	0.174159	0.7189	relat. acceptab.
22	2.0739	0.3534	0.171617	0.7069	relat. acceptab.
23	2.0687	0.3477	0.169178	0.6955	relat. acceptab.
24	2.0639	0.3423	0.166837	0.6846	relat. acceptab.
25	2.0595	0.3371	0.164588	0.6743	relat. acceptab.
26	2.0555	0.3322	0.162427	0.6644	relat. acceptab.
27	2.0518	0.3275	0.160345	0.6549	relat. acceptab.
28	2.0484	0.2892	0.143109	0.5785	relat. acceptab.
29	2.0452	0.2412	0.120587	0.4824	relat. acceptab.
30	2.0423	0.1745	0.088141	0.3490	relat. acceptab.
40	2.0211	0.0869	0.044339	0.1739	relat. acceptab.
60	2.0003	0.0617	0.031487	0.1234	relat. acceptab.
120	1.9799				relat. acceptab.
∞(≈500)	1.9600				small correlat.
∞(≈ 1000)	1.9600				small correlat.