სოხუმის სახელმწიფო უნივერსიტეტის შრომები,
ჰუმანიტარულ და სოციალურ-პოლიტიკურ მეცნიერებათა სერია,
Proceedings of Sokhumi State University,
Humanities and Socio-Political Sciences Series, ტომი/ Vol. XX, 2022 წელი

მზევინარ ფაცაცია[1], ელენე ესიავა[2]
სოხუმის სახელმწიფო უნივერსიტეტი

https://doi.org/10.52340/sou.2022.20.29

რეგრესიის მოდელი - კორელაცია ასაკსა და კორონავირუსის ქეისს შორის ავსტრალიის მაგალითზე

აბსტრაქტი. ეს კვლევა ეხება თანამედროვე ერთ-ერთ უმნიშვნელოვანეს და აქტუალურ საკითხს - COVID-19-ს. კვლევის ფარგლებში, ჩვენ ვეცადეთ სტატისტიკურად აღგვეწერა და გაგვეკეთებინა ანალიზი ვირუსთან დაკავშირებული ერთ-ერთი ყველაზე დიდი მონაცემთა ბაზის, ავსტრალიის შემთხვევაზე. კვლევის მიზანია ცნობიერების ამაღლება ამ, ჯერ კიდევ ნაკლებად ცნობილი ფენომენის შესახებ და მისი გამოყენების პრაქტიკული რეკომენდაციების დოკუმენტად ქცევა მომავალში.

უფრო კონკრეტულად, კვლევის მთავარი ფოკუსი არის ვირუსით ინფიცირებული ადამიანების რაოდენობასა და მათ ასაკობრივ ჯგუფებს შორის კორელაციის ანალიზი, რათა განვსაზღვროთ, არსებობს თუ არა ამ ორ ცვლადს შორის კავშირი. ამისთვის ჩვენ გამოვთვალეთ T-ტესტი Excel-ის ინსტრუმენტების გამოყენებით. შედეგების უკეთესი ვიზუალიზაციისა და სანდოობის მაქსიმალური ხარისხის მისაღწევად, შევიყვანეთ კიდევ ერთი ცვლადი - სქესი, და შევადარეთ ინფიცირებული მამაკაცებისა და ქალების კროსტაბულაციები.

კვლევის შედეგად დადგინდა, რომ ასაკსა და COVID-19-ით ინფიცირებულ შემთხვევებს შორის კორელაცია არ არსებობს (რადგან 20%-იანი მნიშვნელობის დონეზე ნულოვანი ჰიპოთეზა უარყოფილი იქნა, ხოლო ქალების შემთხვევაში ნულოვანი ჰიპოთეზა დადასტურდა როგორც 10%, ასევე 20%-

[1] **მზევინარ ფაცაცია,** მათემატიკის დოქტორი, სოხუმის სახელმწიფო უნივერსიტეტის ასოცირებული პროფესორი.
[2] **ელენე ესიავა,** სოხუმის სახელმწიფო უნივერსიტეტის მათემატიკის საბაკალავრო პროგრამის სტუდენტი.

იანი მნიშვნელობის დონეზე). შესაბამისად, მოგვიხდა ე.წ. "scatterplot"-ების, ანუ გრაფიკების გამოყენება, მიღებული მონაცემების სტანდარტიზაციისა და სქესობრივ განსხვავებათა მინიმიზაციის შემდეგ. ორივე შემთხვევაში, ასაკსა და „კორონავირუსის“ შემთხვევებს შორის კორელაციის ხაზი უარყოფითი მიმართულების იყო, ხოლო საუკეთესო შესაბამისი ხაზიც უარყოფითი დახრილობის. უფრო მეტიც, ორივე შემთხვევაში მივიღეთ ხაზოვანი გაფანტულობა, რაც გვაძლევს საფუძვლიან მიზეზს ვივარაუდოთ, რომ ამ ორ ცვლადს შორის რეგულარული კორელაცია არსებობს (ასაკსა და COVID-19-ის შემთხვევებს შორის).

საკვანძო სიტყვები: რეგრესიის მოდელი, კორელაცია, ასაკი, კორონავირუსის შემთხვევა, კოეფიციენტი, გაფანტვის დიაგრამა, დადგენილების კოეფიციენტები, ავსტრალიის მაგალითი.

Mzevinar Patsatsia[3], Elene Esiava[4]
Sokhumi State University

Regression model - Correlation Between Age and Coronavirus Cases in Australia

Abstract. This study addresses one of the most important and topical issues of modernity - COVID-19. As part of the study, we tried to statistically describe and analyze one of the largest databases related to the virus - the case of Australia. The aim of the research is to raise an awareness about this still unknown phenomenon and to turn it into a kind of recommendation document for its practical implementation in the future.

More specifically, the main focus of the study is the correlation analysis of the number of people infected with the virus and their age groups, thus determining whether the relationship between these two variables exists, for which the T-test had been calculated using some of the Excel tools. In order to have a better visualization and maximum degree of reliability achieved of the results, we came up with another variable - gender and compared the cross-tabulations of infected men with the cross-tabulations of infected women.

The study found that there was no such correlation between age and COVID-19-infected cases (since we had to reject the null hypothesis at a 20% significance level, while for women the null hypothesis has been validated at both 10 and 20%

[3] **Mzevinar Fatsatsia**, Doctor of Mathematics, Associate Professor of Sokhumi State University.
[4] **Elene Esiava,** student of the undergraduate program of mathematics at Sokhumi State University.

significance). Therefore, we had to use the so-called "scatterplots", or graphs, after standardizing the data obtained and minimizing the statistical difference between the genders. Indeed, in both cases the line of reference between age and ‘coronavirus’ cases had a negative direction. And the best fit line had a negative slope too. Moreover, in both cases we obtained a linear scattering, which gives us legit reason to conclude and assume that there is a regular correlation between these two variables (age and COVID-19 cases).

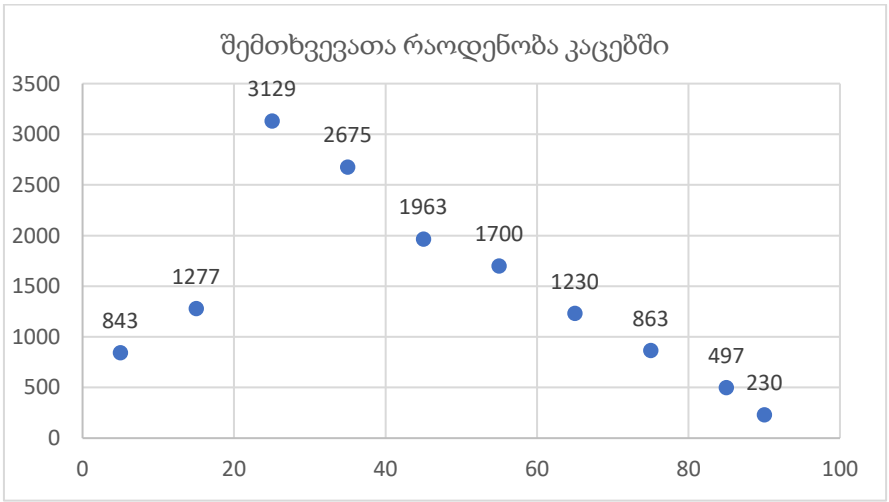
Key words: regression model, correlation, age, case of corona virus, coefficient, scatter diagram, coefficients of determination, Australian example.

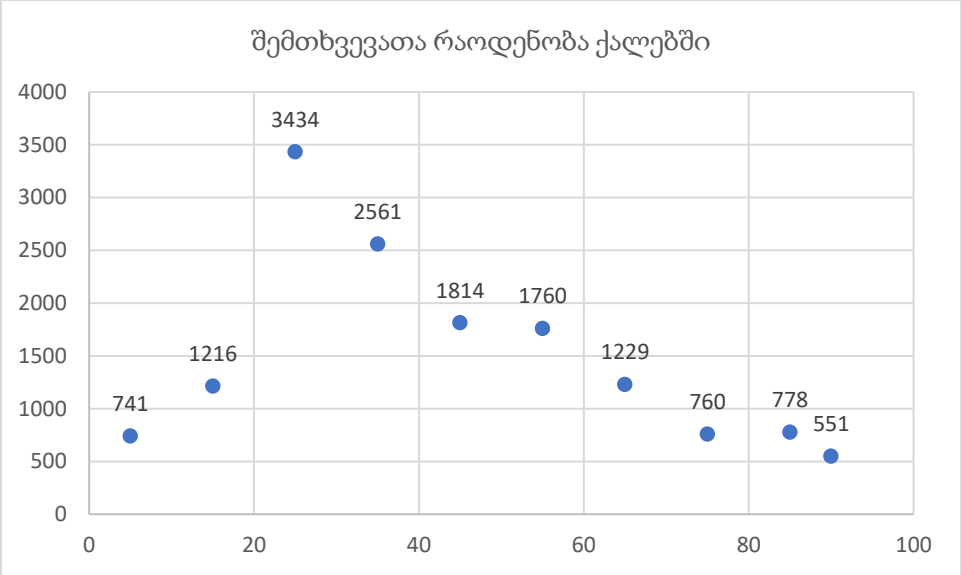
ძირითადი კვლევა. ქვემოთ მოცემულია ჩვენ მიერ გამოყენებული მონაცემები, რომელიც აღებულია ოფიციალური ვებ-საიტიდან (აკმაყოფილებენ შერჩევითობის ყველა პირობას):

Age	Number of Cases Male
5	843
15	1 277
25	3 129
35	2 675
45	1 963
55	1 700
65	1 230
75	863
85	497
90	230

Age	Number of cases female
5	741
15	1 216
25	3 434
35	2 561
45	1 814
55	1 760
65	1 229
75	760
85	778
90	551

წარმოვადგენთ თითოეული შემთხვევებისთვის (მდედრობითი და მამრობითი) გაფანტულობის დიაგრამები მათ სქესს და ასაკს შორის:





ამ დიაგრამებიდან აშკარად ჩანს, ასაკსა და კოვიდვირუსის გავრცელებათა რაოდენობას შორის უარყოფითი კავშირია ორივე შემთხვევაში. იმისათვის, რომ დავრწმუნდეთ, რომ ეს კავშირი ჭეშმარიტია, ქალებში უფრო გამოხატულია, თუ კაცებში, საჭიროა შევამოწმოთ ჰიპოთეზა კორელაციის კოეფიციენტის ნულის ტოლობის შესახებ, როგორც მამრობითი ისე მდედრობითი სქესისათვის.

ჯერ შევამოწმოთ მამრობითი სქესისთვის: ნულოვანი ჰიპოთეზა იქნება $H_0 : \rho = 0$, რაც ნიშნავს, რომ ცვლადებს შორის არ არსებობს კორელაცია, ეს კი გულისხმობს, რომ არ არსებობს კორელაცია ასაკსა და შემთხვევათა რაოდენობას შორის (მამაკაცი). ალტერნატიული ჰიპოთეზა იქნება $H_1 : \rho \neq 0$, რაც ნიშნავს, რომ ცვლადებს შორის არსებობს კორელაცია, რაც გულისხმობს, რომ არსებობს გარკვეული კორელაცია ასაკსა და შემთხვევათა რაოდენობას შორის (მამაკაცი).

პირველ რიგში, T-ტესტის ჩასატარებლად, Excel-ის გამოყენებით, გამოვთვალეთ კორელაციის კოეფიციენტი ასაკსა და შემთხვევათა რაოდენობას შორის (მამაკაცი). Excel-ის მიხედვით ეს რიცხვი უდრის r = -0,5190221 და n = 10. იხ. ქვემოთ მოცემული ცხრილი 1.

ცხრილი 1. ურთიერთკავშირი ასაკსა და კორონავირუსის შემთხვევათა რაოდენობას შორის (მამაკაცი)

	AGE	*NUMBER OF CASES (MALE)*
AGE	1	
NUMBER OF CASES (MALE)	-0,519022107	1

ცნობილია ტესტის სტატისტიკის ფორმულა :

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}},$$

რომელიც განაწილებულია სტიუდენტის კანონით თავისუფლების ხარისხით n-2=10-2=8.

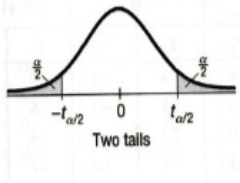
ამ გამოსახულებაში შესაბამისი მნიშვნელობის ჩასმით, ვღებულობთ:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = -0.5190221 \cdot \sqrt{\frac{10-2}{1-(-0.5190221)^2}} = -1.71745.... \approx -1.72$$

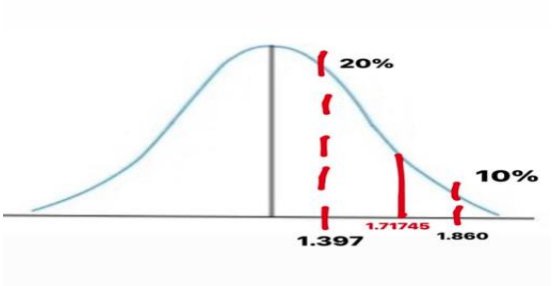
დავაკვირდეთ სტიუდენტის განაწილების ცხრილს თავისუფლების ხარისხით 8 (იხ. ცხ. 2)

ცხრილი 2.

T-TABLE

TABLE T	Two-tail probability	0.20	0.10	0.05	0.02	0.01
	One-tail probability	0.10	0.05	0.025	0.01	0.005
	df					
Values of t_α 	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355

ჩვენ ვხედავთ, რომ ჩვენ მიერ მიღებული $|t| = 1.71745$ სტატისტიკა არის 1.397- სა 1.860-ს შორის, ვინაიდან გვაქვს ორმხრივი ალტერნატივა ($p \neq 0$ შემთხვევა). (დიაგრამა 1. t მოდელი)



დიაგრამა 1.

ე. ი. P- მნიშვნელობის დიაპაზონი მერყეობს 20%-დან 10%-მდე, რაც იმას ნიშნავს, რომ შეგვიძლია უარვყოთ ნულოვანი ჰიპოთეზა 20%-იანი მნიშვნელობის დონეზე, მაგრამ ვერ უარვყოფთ მას 10% -იანი მნიშვნელობის დონეზე. გამოვთვალოთ დეტერმინაციის კოეფიციენტი:

$$d = r^2 = (-00.5190221)^2 = 0.26938 \approx 0.27 .$$

ეს იმას ნიშნავს, რომ მამაკაცებში ასაკსა და კორონავირუსის გავრცელებათა რაოდენობას შორის შიდა კავშირი 27%-ია, დანარჩენი 77% სხვა ფაქტორითაა განპირობებული.

ანალოგიურად ვმოქმედებთ მდედრობითი სქესის შემთხვევაში. ნულოვანი ჰიპოთეზა იქნება $H_0 : \rho = 0$, რაც ნიშნავს, რომ ცვლადებს შორის არ არსებობს კორელაცია, ეს კი გულისხმობს, რომ არ არსებობს კორელაცია ასაკსა და შემთხვევათა რაოდენობას შორის ქალებში. ალტერნატიული ჰიპოთეზა $H_1 : \rho \neq 0$ იქნება, რაც ნიშნავს, რომ ცვლადებს შორის არსებობს კორელაცია, რაც გულისმობს, რომ არსებობს გარკვეული კორელაცია ასაკსა და შემთხვევათა რაოდენობას შორის ქალებში.

Excel-ის გამოყენებით ვიპოვეთ კორელაციის კოეფიციენტი ასაკსა და შემთხვევათა რაოდენობას შორის (მდედრობითი სქესი):r = -0.4323523, n = 10.

ურთიერთკავშირი ასაკსა და კორონავირუსული შემთხვევების რაოდენობას შორის (მდედრობითი სქესი) (იხ. ცხრ. 3)

ცხრილი 3.

	AGE	*NUMBER OF CASES (FEMALE)*
AGE	1	
NUMBER OF CASES (FEMALE)	-0,4323523	1

ტესტის სტატისტიკის ფორმულა:

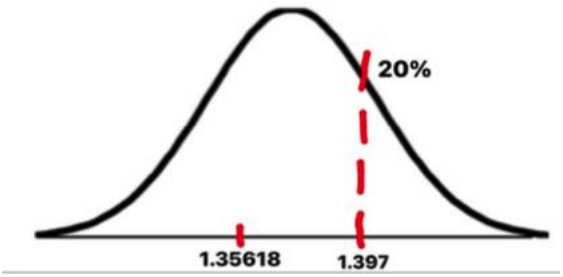
$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}},$$

რომელიც განაწილებულია სტიუდენტის კანონით თავისუფლების ხარისხით n-2=10-2=8.

ამ ტოლებაში შესაბამისი მნიშვნელობის ჩასმით, ვღებულობთ:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = -0.4323523 \cdot \sqrt{\frac{10-2}{1-(-0.4323523)^2}} = -1.35618.... \approx -1.35$$

ანუ t = -1.35618. ასევე ვიცით, რომ თავისუფლების ხარისხი არის n-2, ე. ი. 10-2 = 8. ვხედავთ რომ 8-ის თავისუფლების ხარისხით t სტატისტიკაა 1.397> 1.356. (იხილეთ დიაგრამა 2).



დიაგრამა 2. t მოდელი

რაც ნიშნავს, რომ ვერ უარვყოფთ ნულოვან ჰიპოთეზას 20%-იანი მნიშვნელობის დონეზეც კი, ე. ი. არ არსებობს კორელაცია 20%-იანი მნიშვნელობის დონეზე ქალების ასაკსა და მათში კორონავირუსის გავრცელებათა რაოდენობას შორის.

დეტერმინაციის კოეფიციენტი (ქალები)

$$d = r^2 = (-0.4323523)^2 = 0.186928 \approx 0.19.$$

ეს იმას ნიშნავს, რომ ქალებში ასაკსა და კორონავირუსის გავრცელებათა რაოდენობას შორის შიდა კავშირი 19%-ია, დანარჩენი 81% სხვა ფაქტორითაა განპირობებული.

ცხრილი 4. ასაკის აღწერითი სტატიკა და შემთხვევების რაოდენობა (კაცი)

Age		*Number of Cases Male*	
Mean	49,5	Mean	1440,7
Standard Err	9,32291085	Standard Err	295,003316
Median	50	Median	1253,5
Mode	#N/A	Mode	#N/A
Standard Dev	29,4816327	Standard Dev	932,882397
Sample Varia	869,166667	Sample Varia	870269,567
Kurtosis	-1,3048778	Kurtosis	-0,3890342
Skewness	-0,0834164	Skewness	0,65307774
Range	85	Range	2899
Minimum	5	Minimum	230
Maximum	90	Maximum	3129
Sum	495	Sum	14407
Count	10	Count	10
Confidence L	21,0898896	Confidence L	667,343865

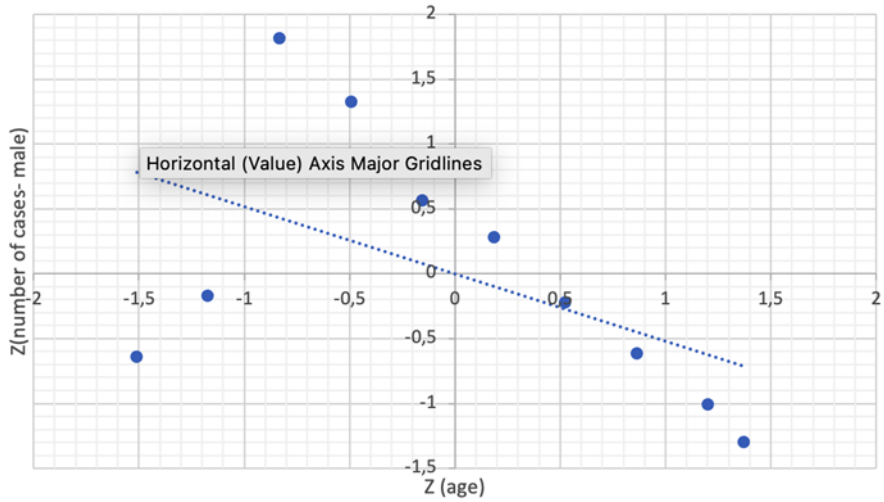
ცხრილი 5 -ში მოცემული ქულები ჩაწერილია სტანდარტული ქულებით

ცხრილი 5.

ასაკობრივი ჯგუფები	ასაკის მედიანა	დაინფიც. მამაკაცთა რაოდენობა	Z-ასაკთა სტანდ. ქულები	Z-დაინფიც. მამაკაცთა რაოდ. სტანდ. ქულები
(0;10]	5	843	-1.509416043	-0.640704056
(10; 20]	15	1277	-1.170221426	-0.175478089
(20;30]	25	3129	-0.83102681	1.809771889
(30;40]	35	2675	-0.491832194	1.323106938

(40;50]	45	1963	-0.152637577	0.559879084
(50;60]	55	1700	0.186557039	0.277956436
(60;70]	65	1230	0.525751655	-0.225859703
(70;80]	75	863	0.864946272	-0.619265072
(80;90]	85	497	1.204140888	-1.011598491
(90;100]	95	230	1.373738196	-1.297808936

ასაკის სტანდარტიზებული ქულებსა (Z- ქულები) და კორონავირუსის გავრცელებათა რაოდენობას შორის (მამაკაცი) აგებულია გაფანტულობის დიაგრამა და რეგრესიის წრფე (იხ. დიაგრ. 3).



დიაგრამა 3.

ანალოგიურად გაკეთებულია მდედრობითი სქესის მონაცემების შემთხვევების მიხედვით. (იხ. ცხრილი 6 და ცხრილი 7).

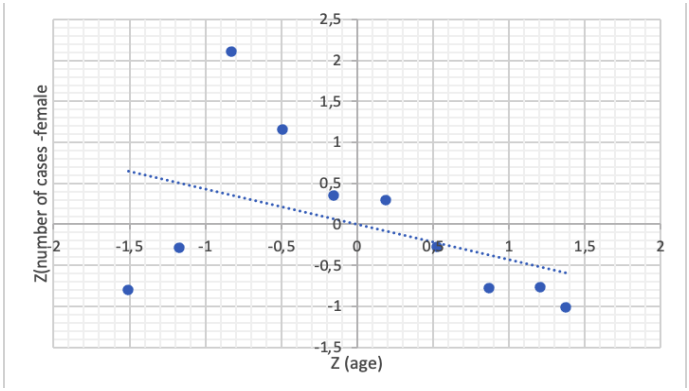
ცხრილი 6.

Age		*Number of cases female*	
Mean	49,5	Mean	1484,4
Standard Err	9,32291085	Standard Error	293,05863
Median	50	Median	1222,5
Mode	#N/A	Mode	#N/A
Standard Dev	29,4816327	Standard Deviation	926,73276
Sample Varia	869,166667	Sample Variance	858833,6
Kurtosis	-1,3048778	Kurtosis	0,8112683
Skewness	-0,0834164	Skewness	1,1730533
Range	85	Range	2883
Minimum	5	Minimum	551
Maximum	90	Maximum	3434
Sum	495	Sum	14844
Count	10	Count	10
Confidence L	21,0898896	Confidence Level(95,0%)	662,94468

ცხრილი 7. სტანდარტიზებული მნიშვნელობები ქალებისთვის:

ასაკობრივი ჯგუფები	ასაკის მედიანა	დაინფიც. ქალთა რაოდენობა	Z-ასაკთა სტანდ. ქულები	Z-დაინფიც. ქალთა რაოდ. სტანდ. ქულები
(0;10]	5	741	-1.509416043	-0.802173006
(10; 20]	15	1216	-1.170221426	-0.289619633
(20;30]	25	3434	-0.83102681	2.103734857
(30;40]	35	2561	-0.491832194	1.16171571
(40;50]	45	1814	-0.152637577	0.355658088
(50;60]	55	1760	0.186557039	0.297388863
(60;70]	65	1229	0.525751655	-0.275591856
(70;80]	75	760	0.864946272	-0.781670871
(80;90]	85	778	1.204140888	-0.762247796
(90;100]	95	551	1.373738196	-1.007194356

აგებულია გაფანტულობის დიაგრამა და რეგრესიის წრფე სტანდარტიზებული ასაკის ქულებსა (Z- ქულები) და კორონავირუსის გავრცელებათა რაოდენობას შორის (ქალი). (იხ. დიაგრამა 4).



დიაგრამა 4.

საბოლოოდ შეგვიძლია გავაკეთოთ შემდეგი დასკვნა: როგორც მდედრობით ასევე მამრობით სქესში კორონა ვირუსის გავრცელებათა რაოდენობასა და მათ ასაკს შორის კორელაციის კოეფიციენტი არის უარყოფითი, რაც იმას ნიშნავს, რომ რაც უფრო მეტია ასაკი მამაკაცების 27%-ში და ქალების 19%-ში დაინფიცირებათა რაოდენობა ნაკლებია. ჩვენი სიტყვებით, რომ ვთქვათ ასაკიანი კაცები უფრო ნაკლებად ინფიცირდებიან (27%), ვიდრე ასაკიანი ქალები (19%). მოხდა მიღებულ მონაცემთა სტანდარტიზება და სქესთაშორისი სტატისტიკური განსხვავების მინიმუმამდე დაყვანა. ორივე შემთხვევაში ასაკისა და კორონავირუსის შემთხვევების მიმართების წრფეს

ჰქონდა უარყოფითი მიმართულება, ე. ი. წარმოსახვით ღერძს ჰქონდა უარყოფითი დახრილობა. თანაც, ორივე შემთხვევაში მივიღეთ წრფივი გაფანტულობა, რაც გვაძლევს საკმაო საფუძველს, რომ დავასკვნათ, რომ ამ ორ ცვლადს (ასაკსა და ვირუსით დაინფიცირებულთა რაოდენობას) შორის არსებობს კანონზომიერი კავშირი.

გამოყენებული ლიტერატურა:

Australia COVID: 33,726 cases and 923 deaths**.** (0131). Worldometer - real time worldstatistics. https://www.worldometers.info/coronavirus/country/australia/

Australia: Number of COVID-19 cases by age and gender 2021. (2021, June 1). Statista. https://www.statista.com/statistics/1104012/australia-number-of-coronavirus-cases-by-age-group/;

(n.d.). Homepage. https://home.iitk.ac.in/~shalab/regression/Chapter3-Regression-MultipleLinearRegressionModel.pdf;

Sharpe, N. R., Veaux, R. D., & Velleman, P. F. (2014). **Business statistics**. Pearson College Division;

What do correlation coefficients positive, negative, and zero mean? **(**n.d.). Investopedia. https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp.