Što treba znati kada izračunavamo koeficijent korelacije?
What we need to know when calculating the coefficient of correlation?

Martina Udovičić[1], Ksenija Baždarić[1], Lidija Bilić-Zulle[1,2], Mladen Petrovečki[1,3]

[1]Katedra za medicinsku informatiku, Medicinski fakultet Sveučilišta u Rijeci, Rijeka
[1]Department of Medical Informatics, School of Medicine, University of Rijeka, Rijeka, Croatia
[2]Zavod za laboratorijsku dijagnostiku, Klinički bolnički centar „Rijeka", Rijeka
[2]Institute of Laboratory Diagnosis, Rijeka Clinical Hospital Center, Rijeka, Croatia
[3]Klinički zavod za laboratorijsku dijagnostiku, Klinička bolnica „Dubrava", Zagreb
[3]Clinical Institute of Laboratory Diagnosis, Dubrava Clinical Hospital, Zagreb, Croatia

Sažetak

Korelacija je statistički postupak za izračunavanje povezanosti dviju varijabli. Vrijednost korelacije brojčano se iskazuje koeficijentom korelacije, najčešće Pearsonovim ili Spearmanovim, dok se značajnost koeficijenta iskazuje vrijednošću *P*. Koeficijent korelacije pokazuje u kojoj su mjeri promjene vrijednosti jedne varijable povezane s promjenama vrijednosti druge varijable. Predznak koeficijenta korelacije (+ ili –) govori nam o smjeru povezanosti. Prilikom izračunavanja korelacije najčešće se pogrješke odnose na uvjete za izračunavanje korelacije, tumačenje koeficijenta i značajnost korelacije, visoke koeficijente korelacije, pretpostavljanje uzročno-posljedične veze, jačinu povezanosti (koeficijent determinacije), te usporedbu dva koeficijenta korelacije.

Ključne riječi: korelacija, Pearsonov koeficijent korelacije, Spearmanov koeficijent korelacije, koeficijent determinacije, pogrješka, statistika

Abstract

Correlation is a statistical procedure applied to calculate association between two variables. The value of correlation is numerically shown by a coefficient of correlation, most often by Pearson's or Spearman's coefficient, while the significance of the coefficient is expressed by *P* value. The coefficient of correlation shows the extent to which changes in the value of one variable are correlated to changes in the value of the other. A sign preceding the coefficient of correlation (+ or -) indicates the direction of correlation. The most frequent errors in calculating correlation are related to conditions for calculation, interpretation of the coefficient and correlation significance, high correlation coefficients, assumption of causal relationship, the strength of correlation (coefficient of determination), and comparison of two correlation coefficients.

Key words: correlation, Pearson's correlation coefficient, Spearman's correlation coefficient, coefficient of determination, error, statistics

Pristiglo: 6. ožujka 2007. *Received: March 6, 2007*
Prihvaćeno: 5. travnja 2007. *Accepted: April 5, 2007*

Uvod

Statistički postupak izračunavanja korelacije jedan je od najčešće korištenih u biomedicini. Korelacija je sukladnost vrijednosti dviju skupina podataka, a iskazuje stupanj povezanosti ispitivanih pojava. Biomedicinska istraživanja često ispituju povezanosti dviju skupina podataka, kao npr. povezanost koncentracije glukoze u krvi s koncentracijom glikiranog hemoglobina ili povezanost biološke dobi i koncentracije kolesterola. Uporaba koeficijenta korelacije ovisi o vrsti podataka, odnosno o ljestvici koju slijede podatci. Najčešće se koriste Pearsonov i Spearmanov koeficijent korelacije (1).

Introduction

The statistical procedure of calculating correlation is one of the most frequently used procedures in biomedicine. Correlation is agreement of values from two data sets, and it expresses the degree of association between investigated phenomena. Biomedical studies often examine the correlation between two data sets as, e.g. the correlation between concentrations of blood glucose and glycated hemoglobin, or between biological age and cholesterol level. The use of the coefficient of correlation depends on the type of data, i.e. on the scale where the data are available. The Pearson's and Spearman's coefficients of correlation are used most frequently (1).

Pearsonov koeficijent korelacije koristi se za varijable na intervalnoj ili omjernoj ljestvici (brojčani podatci) koje su u linearnom odnosu. Linearni odnos varijabli može se očitati s točkastog dijagrama (engl. *scatter diagram*) i podrazumijeva kako točke slijede i rasipaju se oko ravne crte, tj. pravca. Ponekad podatci mogu biti međusobno povezani, ali nisu u linernom odnosu i tada ne možemo izračunavati Pearsonov koeficijent korelacije (1). Primjerice, promatramo li, sukladno Michaelis-Mentenovom modelu enzimske kinetike, povezanost brzine enzimske reakcije i koncentracije supstrata u otopini, uviđamo kako je ta povezanost vrlo visoka no nije linearna, već se odnos dviju varijabli opisuje krivuljom.

Pearsonov koeficijent korelacije označava se malim slovom *r* ili r_p te može poprimati vrijednosti od -1 do +1. Vrijednost koeficijenta korelacije od 0 do 1 je pozitivna korelacija i označava sukladan rast vrijednosti obje skupine podataka. Primjer pozitivne korelacije jest duljina trajanja šećerne bolesti i stupanj oštećenja kapilara u oku. Što je trajanje bolesti duže, to je veći stupanj oštećenja kapilara. Vrijednost koeficijenta korelacije od 0 do -1 označava negativnu korelaciju, odnosno sukladan porast vrijednosti jedne varijable, a pad vrijednosti druge varijable, npr. s porastom nadmorske visine opada koncentracija kisika u zraku. Potpune povezanosti tj. vrijednosti koeficijenta korelacije r = ±1 nisu svojstvene biološkim sustavima i najčešće se odnose na teoretske modele. Kada koeficijent korelacije ima vrijednost 0, tada on označava nepostojanje linearne povezanosti, što upućuje na činjenicu kako poznavajući vrijednosti jedne varijable ne možemo ništa zaključiti o vrijednostima druge. Primjerice, ukoliko bi promatrali povezanost veličine zjenice oka i koncentracije kalcijevih iona u krvi, mogli bismo zaključiti kako nema povezanosti, tj. svakoj veličini zjenice oka može se pridružiti bilo koja koncentracija kalcijevih iona (jasno, u fiziološkim granicama) (2).

Spearmanov koeficijent korelacije (rho, r_s) ili korelacija ranga izračunava se kada jedan od skupa podataka slijedi ordinalnu ljestvicu ili kada raspodjela podataka značajno odstupa od normalne raspodjele te postoje podatci koji značajno odstupaju od većine izmjerenih (engl. *outliers*) (3). Za razliku od Pearsonovog koeficijenta korelacije koji podrazumijeva linearnu povezanost, za Spearmanov koeficijent korelacije to nije uvjet, a može se računati i na manjim uzorcima (N < 35). U slučaju dobivenog $r_s = 0$ može se zaključiti da povezanosti među varijablama zaista nema (1).

Postupak izračunavanja korelacije često se koristi neispravno te je stoga potrebno prije izračunavanja razumjeti pojam i vrste korelacije, uvjete za izračunavanje korelacije te tumačenje povezanosti kako bi se izbjeglo pogrješno zaključivanje.

U nastavku su navedene neke od najčešćih pogrješaka prilikom izračunavanja korelacija i njhova tumačenja.

The Pearson's coefficient of correlation is employed for variables on an interval or ratio scale (numerical data) that are in linear relation. The linear relation of variables may be read from a scatterplot and it implies that the points follow and scatter around the straight line. The data may sometimes be interconnected without being in linear relation and then the Pearson's correlation coefficient cannot be calculated (1). For instance, if we observe - in accordance with the Michaelis Menten model of enzyme kinetics, - the correlation between enzyme reaction velocity and substrate concentration in a solution, we can conclude that this correlation is very high but not linear, and the relation between the two variables is described by a curve.

Pearson's coefficient of correlation is denoted by a small letter *r* or r_p, and its values may range from -1 to +1. The value of the correlation coefficient from 0 to 1 is positive correlation and it designates proportional growth of values in both data sets. An example of positive correlation is the duration of diabetes mellitus and the degree of damage of eye capillaries. The longer the duration of the disease, the higher the damage to eye capillaries. The correlation coefficient value from 0 to -1 indicates negative correlation, i.e. a rise in the value of one variable that is proportional to a decline in the value of the other; e.g. oxygen concentration in the air drops with the rise in altitude above sea level. Perfect correlations, i.e. the values of the coefficient of correlation r = ± 1 are not characteristical for biological systems and most frequently refer to theoretical models. The zero value of the coefficient of correlation indicates absence of linear correlation, i.e. by knowing the values of one variable, we can conclude nothing on the values of the other. Thus, for instance, if we observe the correlation between the size of the pupil of the eye and calcium ion concentration in the blood, we can conclude that there is no correlation, i.e. each size of the pupil could be associated to any calcium ion concentration (understandably, within physiological limits) (2).

Spearman's coefficient of correlation (rho, r_s) or rank correlation is calculated when one of the data sets is on ordinal scale, or when data distribution significantly deviates from normal distribution and data are available that considerably diverge from most of those measured (outliers) (3). Linear correlation, implied by the Pearson's coefficient of correlation, is not required for the Spearman's correlation coefficient which can also be calculated for small samples (N<35). In case of $r_s = 0$, it may be concluded that there is no actual correlation between variables (1).

The procedure of calculating correlation is frequently applied incorrectly. Prior to calculation, it is therefore necessary to understand the concept and types of correlation, conditions for calculating correlation and interpreting association in order to avoid wrong conclusions.

What follows are some of the most frequent mistakes made while calculating correlations, and their explanations.

Uvjeti za izračunavanje korelacije

Pitanje: Je li ispravno računati Pearsonov koeficijent korelacije za stupanj opeklina na tijelu i trajanje bolničkog liječenja izraženo u danima?
Odgovor: Nije ispravno.
Tumačenje: Prvi korak u izračunavanju korelacije jest provjeriti zadovoljavaju li izmjereni podatci uvjete za izračunavanje Pearsonove korelacije. Stupanj opeklina na tijelu označava se na ljestvici od 1 do 4 i takvi su podatci kategorički (svrstavaju ispitanike u unaprijed utvrđene "razrede") te slijede ordinalnu mjernu ljestvicu. Duljina bolničkog liječenja izražena u danima slijedi omjernu ljestvicu i bila bi pogodna za računanje Pearsonovog koeficijenta korelacije, ali samo onda kada bi i druga varijabla sljedila intervalnu ili omjernu ljestvicu. Pearsonov koeficijent korelacije računa se samo ako su zadovoljeni sljedeći uvjeti: podatci obje ispitivane varijable slijede intervalnu ili omjernu ljestvicu, podatci barem jedne varijable su normalno, tj. simetrično raspodijeljeni, ispitivani uzorak je velik (N > 35) i zadovoljen je uvjet linearne povezanosti, što treba očitati iz točkastog grafikona (1).
Ukoliko uvjeti za izračunavanje Pearsonovog koeficijenta korelacije nisu zadovoljeni, može se koristiti Spearmanov koeficijent korelacije. U opisanom primjeru stupanj opeklina slijedi ordinalnu ljestvicu pa stoga nije zadovoljen uvjet za Pearsonovu korelaciju, već je potrebno izračunati Spearmanov koeficijent korelacije.

Tumačenje i značajnost koeficijenta korelacije

Pitanje: U istraživanju povezanosti raspoloženja i količine tekućine unesene pijenjem tijekom dana dobivena je povezanost r = 0,12; P = 0,003. Je li ispravno zaključiti kako postoji značajna povezanost raspoloženja i količine popijene tekućine?
Odgovor: Nije ispravno.
Tumačenje: Nakon izračuna koeficijenta korelacije važno je znati kako rezultat protumačiti, odnosno objasniti što vrijednosti koeficijenta korelacije zaista znače. U prikazu rezultata korelacija obvezno se navode koeficijent povezanosti (korelacije) "r" i to brojem s dva decimalna mjesta, te značajnost koeficijenta korelacije "P" brojem s tri decimalna mjesta (4). Ukoliko je koeficijent korelacije značajan s obzirom na postavljenu granicu značajnosti (uobičajeno P < 0,05), zaključujemo da je koeficijent korelacije značajan i da se smije tumačiti. Ukoliko je vrijednost P > 0,05 zaključujemo da koeficijent korelacije nije značajan i tada se bez obzira na njegovu vrijednost ne smije tumačiti. Prilikom tumačenja vrijednosti koeficijenta korelacije vrijede ista pravila i za Pearsonov i Spearmanov koeficijent te se uobičajeno smatra kako vrijednosti r od 0 do 0,25 ili od 0 do –0,25 upućuju kako nema povezanosti, dok vri-

Conditions for calculating correlation

Question: Is it correct to calculate the Pearson's correlation coefficient for the degree of burns on the body and the duration of hospitalization expressed by the number of days?
Answer: It is not correct.
Explanation: Initial step in calculating correlation is to check if the measured data meet the conditions for calculating the Pearson's correlation. The degree of burns on the body can be ranked on a scale from 1 to 4; such data are categorical (classifying subjects in predefined "classes") and they follow an ordinal scale. The duration of hospital therapy expressed in the number of days is on a ratio scale and is suitable for calculating the Pearson's correlation coefficient if the other variable is on an interval or ratio scale. The Pearson's coefficient of correlation can be calculated only if the following conditions are met: the data for both examined variables are on an interval or ratio scale, the data for at least one variable have normal, i.e symmetrical distribution, the examined sample is large (N > 35), and the condition of linear correlation is met, which may be read from a scatterplot (1).
Unless the conditions for calculation of the Pearson's coefficient of correlation are met, the Spearman's correlation coefficient can be applied. In the example described above, the degree of burns is measured on an ordinal scale, and therefore the condition for Pearson's correlation is not fulfilled but rather the Spearman's rank coefficient of correlation should be calculated.

Interpretation and significance of the coefficient of correlation

Question: In a study of correlation between the mood and the amount of liquid consumed by daily drinking, the correlation r = 0.12; *P* = 0.003 was obtained. Is it correct to conclude that there is a significant correlation between the mood and the amount of the consumed liquid?
Answer: It is not correct.
Explanation: After calculating the coefficient of correlation, it is important to know how to interpret the result, that is, the real meaning of the correlation coefficient. In presenting the results of correlation, the coefficient of correlation "r" should be expressed by a number with two decimal places, and the significance of the coefficient of correlation "P" in a number with three decimal places (4). If the coefficient of correlation is significant in regard to the set limit of significance (commonly *P* < 0.05), we may conclude that the coefficient of correlation is significant and may be interpreted. If the value is *P* > 0.05, we can conclude that the coefficient of correlation is not significant and in this case it may not be interpreted regardless of its value. When interpreting the value of the corrrelation

jednosti r od 0,25 do 0,50 ili od –0,25 do –0,50 upućuju na slabu povezanost među varijablama. Vrijednosti r od 0,50 do 0,75 ili od –0,50 do –0,75 upućuju na umjerenu do dobru povezanost, te vrijednosti r od 0,75 do 1 ili od –0,75 do –1 upućuju na vrlo dobru do izvrsnu povezanost među varijablama (1).
Sukladno navedenome, pogrješno je zaključiti kako postoji značajna povezanost raspoloženja i količine popijene tekućine tijekom dana. Ispravno zaključivanje glasi: nema povezanosti između ispitivanih varijabli (r = 0,12) i to smijemo tvrditi jer je koeficijent korelacije značajan (P=0,003) (5,6).

Visoka vrijednost koeficijenta korelacije

Pitanje: U istraživanju povezanosti visine tijela i biološke dobi dobivena je korelacija r = 0,97. Možemo li zaključiti kako su visina i dob nesumnjivo izvrsno povezani?
Odgovor: Ne, barem ne nesumnjivo.
Tumačenje: Ukoliko je izračunat koeficijent korelacije za biološke varijable r > 0,95, treba posumnjati na pogrješku u mjerenju, uzorkovanju ispitanika ili mogućem prepravljanju izmjerenih rezultata. Zbog prirodne raznolikosti u biološkim sustavima upravo je nemoguće dobiti tako visoki koeficijent korelacije ukoliko su mjerenja učinjena ispravno (reprezentativan uzorak, dovoljno osjetljiv instrument i sl.) (1). Uvijek je potrebno voditi računa o vrsti podataka koji se mjerenjem prikupljaju i statistički obrađuju. Primjerice, ukoliko uspoređujemo vrijednosti glukoze izmjerene u seriji uzoraka krvi s pomoću dva različita instrumenta, tj. biokemijska analizatora, za očekivati je kako će koeficijent korelacije biti vrlo visok (pa i do r = 0,99), što je tada znak dobre usklađenosti dvaju instrumenata.

Povezanost i uzročno posljedična veza

Pitanje: U istraživanju povezanosti koncentracije alkohola u krvi i prometnih nesreća utvrđeni su r = 0,78 i P=0,002. Možemo li zaključiti kako uzimanje alkohola uzrokuje prometne nesreće, tj. promatrane prometne nesreće su posljedica uzimanja alkohola?
Odgovor: Ne, ne možemo.
Tumačenje: Korelacija govori o povezanosti, a ne o uzročno posljedičnoj vezi među varijablama. Dakle, ukoliko postoji visoka povezanost između uzimanja alkohola i prometnih nesreća ne možemo zaključiti da jedna varijabla utječe na drugu, odnosno da uzimanje alkohola uzrokuje nesreće u prometu. Moguće je da veća količina alkohola uzorkuje više prometnih nesreća, no postoji mogućnost značajnog utjecaja ostalih neispitivanih čimbenika ili rijetkih događaja (7,8). U opisanom primjeru to bi moglo biti stanje na cesti, ispravnost vozila, moguća bolest vozača nevezana za alkohol, djelovanje drugih farmakološki aktivnih tvari i sl.

coefficient, the same rules are valid for both Pearson's and Spearman's coefficient, and r values from 0 to 0.25 or from 0 to -0.25 are commonly regarded to indicate the absence of correlation, whereas r values from 0.25 to 0.50 or from -0.25 to -0.50 point to poor correlation between variables. r values ranging from 0.50 to 0.75 or -0.50 to -0.75 indicate moderate to good correlation, and r values from 0.75 to 1 or from -0.75 to -1 point to very good to excellent correlation between the variables (1).
Accordingly, it is wrong to conclude that there is a significant correlation between the mood and the amount of liquid taken during a day. Correct conclusion is as follows: there is no correlation between the examined variables (r = 0.12), which may be claimed because the correlation coefficient is significant (P = 0.003) (5,6).

High value of the correlation coefficient

Question: The correlation value obtained in a study of correlation between body height and biological age was r = 0.97. May we conclude that height and age are definitely excellently correlated?
Answer: No, at least not beyond doubt.
Interpretation: If the correlation coefficient calculated for biological variables is r > 0.95, an error in measurement and sampling or possible alteration of measured results should be suspected. Due to natural variety of biological systems, it is virtually impossible to obtain such a high correlation coefficient if measurements have been done correctly (representative sample, sufficiently sensitive instrument, etc.) (1). The type of data collected by measurements and processed statistically should always be taken into account. For example, if comparison is made of the values of glucose measured in a series of blood samples by two different instruments, i.e. biochemical analyzers, the coefficient of correlation may be expected to be very high (even up to r = 0.99), which in this case indicates good agreement between the two instruments.

Correlation and causal relationship

Question: r = 0.78 and *P* = 0.002 were determined in a study of correlation between blood alcohol level and traffic accidents. Are we allowed to conclude that alcohol consumption is the cause of traffic accidents, i.e. that the observed traffic accidents are the consequence of alcohol consumption?
Answer: No, we are not.
Explanation: Correlation provides information on association rather than a cause- and-effect relationship between variables. Thus, if there is a high correlation between alcohol consumption and traffic accidents, we may not conclude that one variable affects the other, i.e. that alcohol consumption causes traffic accidents. It is possible that in-

U istraživanjima se korelacija treba ponajprije koristiti za postavljanje hipoteza, a ne za njihovo testiranje kao što se to često sasvim pogrješno koristi (9). Primjerice, utvrdi li se povezanost između varijabli, uzročno posljedična veza dokazuje se znanstvenim pokusom. Jedini pokus kojim se dokazuje uzročno posljedična veza u biomedicini jest randomizirani kontrolirani klinički pokus (10).

Jačina (udio) povezanosti

Pitanje: Usporedbom katalitičke koncentracije dvaju enzima u krvi ispitanika dobivena je povezanost r = 0,52; P = 0,002. Možemo li zaključiti kako vrijednosti enzima imaju 52% zajedničkih vrijednosti katalitičke koncentracije?
Odgovor: Ne, ne možemo.
Tumačenje: Koeficijent korelacije nije mjera jačine povezanosti. Vrijednost koeficijenta korelacije r = 0,52 ne može se tumačiti kao povezanost od 52%, tj. 52% zajedničkih vrijednosti dviju katalitičkih koncentracija enzima. Udio zajedničkih vrijednosti, tj. jačina linearne povezanosti izražava se koeficijentom determinacije. Koeficijent determinacije računa se jednostavno, tj. kvadriranjem koeficijenta korelacije i označava kao r^2. Može se računati samo za Pearsonovu korelaciju (3). Stoga je jačina povezanosti (koeficijent determinacije) u ovom primjeru $r^2 = 0,52 \times 0,52 = 0,27$, tj. katalitičke koncentracije dva enzima imaju 27% zajedničkih vrijednosti. Dvostruko veća povezanost ne znači i dvostruko veću jačinu povezanosti, npr. ako povezanost iznosi $r_1 = 0,26$, jačina povezanosti biti će $r_1^2 = 0,07$ (7%) dok za dvostruko veću povezanost $r_2 = 0,52$ jačina povezanosti iznosi $r_2^2 = 0,27$ (27%).

Usporedba dvaju koeficijenata korelacije istih obilježja u dva uzorka ispitanika

Pitanje: Ispitana je povezanost vremena provedenog u radu s računalom i brzine pisanja teksta na računalu u žena ($N_1 = 60$) i muškaraca ($N_2 = 40$). Koeficijent korelacije za žene iznosi $r_1 = 0,70$, a za muškarce $r_2 = 0,50$; oba su statistički značajna. Možemo li zaključiti da je $r_1 > r_2$, odnosno da je u žena veća povezanost vremena provedenog u radu s računalom i brzine pisanja teksta na računalu?
Odgovor: Ne, ne možemo.
Tumačenje: Dva se koeficijenta korelacije nikako ne smiju izravno uspoređivati već je potrebno posebno ispitati značajnost razlike između korelacija dviju skupina podataka. Postupak utvrđivanja značajnosti razlike dvaju koeficijenata korelacije uzima u obzir vrijednost koeficijenata korelacije i veličine oba uzorka (8).
Usporedbom dvaju koeficijenta korelacije u opisanom primjeru utvrđeno je da povezanost vremena provedenog u radu s računalom i brzine pisanja teksta na računalu u žena nije značajno veća od povezanosti istih varijabli u muškaraca (P = 0,132) (11).

creased amount of alcohol causes the increased number of accidents, yet there is a possibility of a considerable effect of other uninvestigated factors or rare events (7,8). In the example described above, these factors or events could be road condition, proper operation of a vehicle, potential illness of a driver unrelated to alcohol, action of other pharmacologically active substances, and the like.
In research, correlation should be primarily employed to build hypotheses rather than to test them, the latter being a frequent and entirely wrong application (9). If, e.g., correlation is established between variables, causal relationship should be demonstrated by scientific experiment (10). The only experiment to demonstrate such relationship in biomedicine is a randomized controlled clinical trial (10).

The strength of correlation

Question: By comparing catalytic concentration of two enzymes in the blood, the correlation r = 0.52; *P* = 0.002 was obtained. Can we conclude that enzyme values share 52% of catalytic concentration values?
Answer: No, we cannot.
Explanation: The coefficient of correlation is not a measure of the strength of correlation. The correlation coefficient value r = 0.52 cannot be interpreted as 52% correlation, i.e. 52% of the joint values for the two catalytic enzyme concentrations. The proportion of shared values, i.e. the strength of linear correlation is expressed by the coefficient of determination. The coefficient of determination is calculated simply by squaring the correlation coefficient, and is denoted by r^2. It can be calculated only for the Pearson's correlation (3). Therefore the strength of correlation (coefficient of determination) in this example is $r^2 = 0.52 \times 0.52 = 0.27$, ie. the catalytic concentrations of two enzymes share 27% of common values. Twice as high correlation does not imply the twofold strength of correlation; e.g., if the correlation was $r_1 = 0.26$, the strength of correlation would be $r_1^2 = 0.07$ (7%); also, it would be $r_2^2 = 0.27$ (27%) for the twofold higher correlation, $r_2 = 0.52$.

Comparison of two correlation coefficients with the same properties on two subject samples

Question: Correlation between the time spent at computer work and the speed of typing a text into computer has been examined for women ($N_1 = 60$) and men ($N_2 = 40$). The coefficient of correlation, for women is $r_1 = 0.70$ and for men $r_2 = 0.50$: both are statistically significant. Can we conclude that $r_1 > r_2$, i.e. that the correlation between the time spent at computer work and computer typing speed is higher in women?
Answer: No, we cannot.

Zaključak

Utvrđivanje povezanosti, tj. korelacije među pojavama (varijablama) važno je oruđe u znanstvenom radu. Primijećene povezanosti dviju pojava omogućuju samo postavljanje hipoteze u znanstvenom pokusu kojim će se tek pokušati utvrditi i uzročno posljedična sveza (koju korelacija nikad ne dokazuje). Osim u biološkim sustavima, osobito u laboratorijskoj medicini, koeficijent korelacije značajan je u proučavanju i usporedbi dvaju analitičkih sustava (metoda, instrumenata i sl.) kada upravo na temelju njegove visoke vrijednosti možemo složeniju metodu zamijeniti, primjerice, jednostavnijom ili jeftinijom. Često korištena u obradbi podataka u znanstvenim radovima, korelacija se nerijetko i zlorabi i to uglavnom zbog neznanja ili zanemarivanja pravila uporabe testa korelacije. Posljedica su tada pogrješni zaključci o znanstvenim hipotezama koji vode u zabludu, a ne k novom znanju.

Adresa za dopisivanje:

Martina Udovičić
Katedra za medicinsku informatiku
Medicinski fakultet u Rijeci
Braće Branchetta 20
51000 Rijeka
e-pošta: umartina@medri.hr
tel: +385 51 651 255
faks: +385 51 651 255

Explanation: The two coefficients of correlation should by no means be directly compared but the significance of difference between the correlations for two data sets should be examined. The procedure of establishing the significance of the difference between two coefficients of correlation takes into account the value of correlation coefficients and the size of both samples (8).
By comparison of the two correlation coefficients in the example above, the correlation between the time spent in computer work and the typing speed of women has not been found to differ significantly from the correlation of the same variables in men (P = 0.132) (11).

Conclusion

Determination of association, i.e. correlation between phenomena (variables) is an important tool in scientific study. The associations observed between two phenomena only allow us to pose a hypothesis in a scientific experiment that is, actually, itself an attempt to establish causal relationship (which is never demonstrated by correlation). Aside from biological systems (particularly laboratory medicine), the coefficient of correlation is important in the study and comparison of two analytical systems (methods, instruments, etc.) when we can, on the basis of its high value, replace a more complex method by, e.g. an easier or cheaper one. Being frequently used in data processing in scientific studies, correlation is also often misused mostly due to ignorance or negligence of the rules for using a correlation test. The consequence is faulty conclusions on scientific hypotheses that lead to fallacies rather than new insights.

Corresponding author:

Martina Udovicic
Department of Medical Informatics
Rijeka University School of Medicine
B. Branchetta St. 20
HR-51000 Rijeka, Croatia
e-mail: umartina@medri.hr
phone: +385 51 651255
fax: +385-51 651255

Literatura/References

1. *Dawson B, Trapp RG. Basic and Clinical Biostatistics. 4th Ed. New York: Lange Medical Books/McGraw-Hill; 2004.*
2. *Ažman J, Frković V, Bilić-Zulle L, Petrovečki M. Korelacija i regresija. Acta Med Croat 2006;60(Suppl 1):81-91.*
3. *Petrie A, Sabin C. Medical Statistics at Glance. 2nd Ed. Oxford: Blackwell Publishing; 2005.*
4. *Lang T. Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles. CMJ 2004;45(4):361-70.*
5. *Petrovečki M, Gabela O, Marcelić T. Statistical management of automimmune disease data. "New trends in classification, monitoring and management of autoimmune diseases", 5th FESCC Postgraduate Course in Clinical Chemistry, Dubrovnik, October 2005:77-80.*
6. *Petrovečki M, Gornik O, Marcelić T. Processing and presentation of biochemical research data. Congress of the Croatian Society of Biochemistry and Molecular Biology, Vodice, October 2006:42.*
7. *Rumsey D. Statistics for Dummies. Indianapolis: Wiley Publishing Inc.; 2003.*
8. *Petz B. Osnove statističke metode za nematematičare. Jastrebarsko: Naklada Slap, 2002.*
9. *Zou KH, Tuncali KT, Silverman SG. Correlation and Simple Linear Regression. Radiology 2003; 227:617-28.*
10. *Marušić M, urednik. Uvod u znanstveni rad u medicini. Zagreb: Medicinska naklada; 2004.*
11. *Usporedba dva koeficijenta korelacije. MedCalc Manual. Dostupno na URL: http://www.medcalc.be/manual/mpage08-06.php Pristupljeno: 28. veljače 2007.*