DESPRE METODĂ
Compararea statistică a coeficienților de corelație

Robert Balazs[i1]
Universitatea Babeş-Bolyai, Cluj-Napoca

Alături de cvasi experimente, studiile corelaționale reprezintă poate cele mai frecvente tipuri de studii utilizate în cercetarea aplicată. În condiții obişnuite un astfel de studiu presupune colectarea simultană a mai multor date (ex. indici de stres, vechime la locul de muncă, anxietate, venit, număr de absențe, etc) şi calcularea coeficienților de corelație pentru a surprinde existența unor eventuale asocieri a două sau mai multor variabile măsurate (Cohen, 2001). Aceste demers pare a fi unul simplu şi de obicei nici nu ridică probleme pentru cercetători. În cercetările publicate, de obicei, sunt comunicate: procedura de colectare a datelor (mai rar şi indicii de consistență internă deşi aceasta afectează valoarea coeficienților de corelație, implicit concluziile studiului), valoarea coeficienților de corelație precum şi probabilitățile calculate în baza ipotezei nule (Ho:ρ=0) a coeficienților r.

Greşeala metodologică apare atunci când concluziile studiului depăşesc acest nivel şi cercetătorul trece la a discuta diferențele existente între coeficienții de corelație calculați pentru diferitele perechi de variabile (ex. corelația între variabila X şi Y este mai mare decât cea existentă între X şi V), sau la a compara valoarea coeficienților de corelație obținuți în studiu, cu corelațiile obținute pentru aceleaşi două variabile în alte studii. În principiu efectuarea acestor comparații nu este o greşeală în sine, eroarea apare atunci când aceste comparații nu implică un demers inferențial, ci se rezumă doar la unul pur intuitiv (în cercetarea autohtonă cel din urmă demers este mai frecvent).

În continuare vom trece la prezentarea demersului inferențial care permite efectuarea comparațiilor menționate mai sus. Din moment ce compararea corelațiilor este un demers oarecum similar comparării mediilor, în discuția problemei vom respecta traseul didactic parcurs în predarea demersului de comparare a două medii care, adaptat la corelații, presupune abordarea următoarelor probleme: verificarea unor ipoteze nule diferite de Ho:ρ=0, compararea a doi coeficienți de corelație obținuți pe două eşantioane independente şi compararea a doi coeficienți de corelație obținuți pe eşantioane dependente.

a. Verificarea unor ipoteze nule diferite de Ho:ρ=0

În demersul obişnuit de verificare a Ho ceea ce interesează este dacă în populație coeficientul de corelație este diferit de 0 sau nu. Acest calcul se realizează pe baza formulei

$$t = r*\sqrt{n-2}/\sqrt{1-r^2},\ df=n-2$$

Rezultatul acestui calcul este afişat de obicei de softurile statistice sub forma probabilității calculate în baza ipotezei nule, *Ho:ρ=0.* Dacă valoarea acestei probabilități este mai mică decât cea setată (de obicei .05) atunci se conchide că probabilitatea, ca valoarea calculată a lui r să aparțină unei distribuții caracterizate de ρ=0, este neglijabilă, deci infirmăm Ho.

Să presupunem însă că scopul unui cercetător este să verifice în ce măsură s-a schimbat valoarea corelației existente între nivelul stresului ocupațional şi numărul greşelilor decizionale la manageri în ultimii zece ani. Având la îndemână un studiu care arată că la nivelul populației această corelație, cu zece ani în urmă, avea valoarea ρ=.9, cercetătorul încearcă să compare valoarea dată cu corelația obținută în studiul propriu r=.6 pe un eşantion de n=34. În acest caz obiectivul cercetătorului este de a verifica ipoteza nulă conform căreia Ho:ρ=.9.

Valorile posibile ale coeficientului de corelație fluctuează între valorile ±1, ca urmare a acestui fapt distribuția ipotezei nule Ho:ρ=0 este un simetrică şi va aproxima o distribuție normală cu atât mai bine cu cât volumul eşantionului este mai mare.

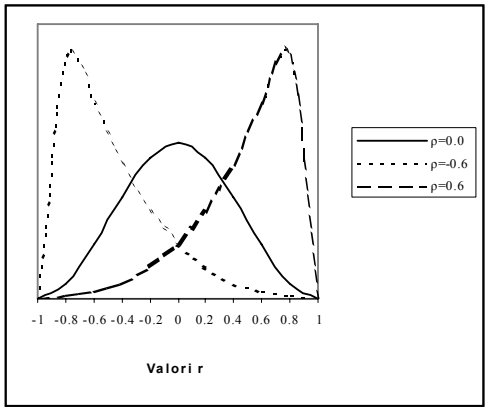
În cazul unei ipoteze nule Ho:ρ=.6, valorile mai mari decât .6 se vor distribui pe o plajă de valori mai restrânsă decât valorile mai mici, ca urmare distribuția va fi asimetrică înclinată spre stânga (Figura 1). Acesta este motivul pentru care demersul inferențial trebuie precedat de un proces de normalizare a

[1] Adresa de corespondență: robertbalazsi@psychology.ro

distribuției utilizând formula de transformare matematică a valorilor r în valori z′.

Formula a fost elaborată de Fisher, rezultatele acestei transformări nonlineare fiind trecute de obicei în anexele cărților de specialitate.

$$z' = .5[\ln(1+r) - \ln(1-r)]$$



Fi

Figura 1 Distribuția valorilor r în funcție de valoarea ρ

Utilizând această formulă distribuțiile se normalizează (Figura 2).

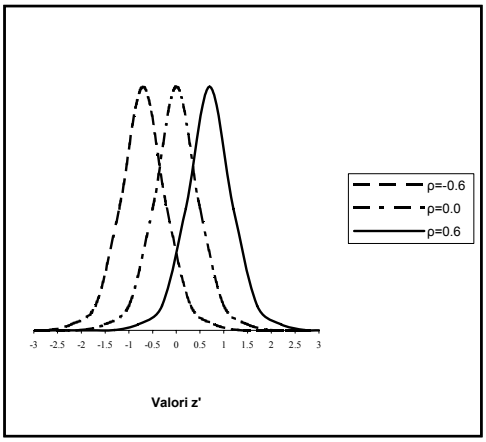


Figura 2 Distribuția valorilor z' în funcție de valoarea ρ

Compararea corelației obținută pe eşantion cu valoarea corelației obținută pe populație se efectuează utilizând valorile transformate, conform formulei z deja cunoscute de la compararea mediilor (Hays, 1963).

$$z=z'_r - z'_\rho/sd_{z'}^2,$$

unde z'_r şi z'_ρ sunt valorile transformate a coeficienților de corelație (r şi ρ) şi $sd_{z'}$ este eroarea standard a distribuției ipotetice a valorilor r, distribuție obținută prin selecția aleatoare repetată a unui eşantion de mărimea n şi calcularea repetată a valorii r. Calculul erorii standard depinde doar de numărul subiecților din eşantion $sd_{z'}=1/\sqrt{n-3}$. Revenind la datele problemei noastre $sd_{z'}=.179$, iar valoarea lui z pentru această comparație va fi z=.693-1.472/.179 = -4.35, valoare mai mare decât -1.96, valoarea critică corespunzătoare pragului de .05 setat. Pe baza rezultatului putem afirma că puterea asocierii celor două variabile a devenit semnificativ mai slabă pe parcursul celor zece ani.

b. Compararea a doi coeficienți de corelație obținuți pe două eşantioane independente.

În problema anterioare s-a pornit de la premisa că se cunoaşte valoarea corelației în populație, o supoziție care are un iz mai degrabă didactic. Putem asuma acest fapt doar în cazurile în care eşantionul pe care s-a calculat corelația este foarte mare (Minium, King, & Bear, 1993). În cele mai multe situații avem de a face cu comparații a doi coeficienți de corelație obținute pe două eşantioane diferite. Să presupunem că un cercetător doreşte să compare coeficienții de corelație, între stresul ocupațional şi greşelile manageriale, obținuți pe două eşantioane independente, unul de bărbați şi unul de femei. În acest scop selectează două eşantioane a câte 40 subiecți şi calculează pentru fiecare eşantion valoarea coeficientului de corelație a celor două variabile măsurate. Valoarea coeficientului de corelație pentru bărbați este $r_1=.64$ iar pentru femei este de $r_2=.82$. Şi de această dată se va recurge la o transformare a r în z′ şi la compararea valorilor transformate, utilizând formula

$$z=z'_1 - z'_2/sd_{z'_{12}}$$

unde z'_1 şi z'_2 sunt valorile transformate a coeficienților de corelație (r_1 şi r_2) şi $sd_{z'_{12}}$ este eroarea standard aproximată prin adunarea erorilor standard pentru cele două distribuții din care au fost selectate corelațiile comparate şi se calculează după formula, $sd_{z'_{12}} = \sqrt{(1/n_1-3+1/n_2-3)}$. Pentru datele problemei $sd_{z'_{12}} = 0.22$. Astfel vom obține z=.758-1.157/.22=-1.81. Valoarea calculată nu depăşeşte

[2] Este important să se păstreze distincția între valorile z care semnifică cote standard şi valorile z′ care semnifică valorile transformate, în baza formulei oferite de Fisher, al coeficienților de corelație.

valoarea critică de 1.96, ceea ce înseamnă că nu există diferențe semnificative în ceea ce priveşte nivelul corelației în cele două populații.

c. Compararea a doi coeficienți de corelația obținuți pe eşantioane dependente.

În anumite situații interesul cercetătorului este de a compara coeficienții de corelație calculați pe diferite perechi de variabile. Să presupunem că în studiul menționat anterior, alături de stresul ocupațional (notat cu x) şi greşelile manageriale (notat cu y) mai există o variabilă măsurată, numărul angajaților avuți în subordine (notat cu z). Corelațiile calculate pentru cele trei variabile sunt: r_{xy}=.45, r_{zy}=.23 şi r_{xz}=.34 pe un eşantion de 40. În această situație cercetătorul îşi poate propune să compare coeficientul de corelație calculat pentru numărul angajaților şi greşeli (r_{zy}) cu cel calculat pentru numărul de angajați şi numărul greşelilor decizionale (r_{zy}). În acest caz transformarea Fisher nu se mai poate aplica, din moment ce toate măsurătorile au fost efectuate pe acelaşi eşantion. Asemenea testului t pentru eşantioane dependente, calculul trebuie să ia în considerare şi faptul că toate măsurătorile au fost efectuate pe aceeaşi subiecți (Cohen & Cohen, 1983). Formula utilizată în acest caz este

$$t=(r_{xy}-r_{zy})\sqrt{(n-1)(1+r_{xz})}/\sqrt{2(n-1/n-3)*|\mathbf{R}| +ř^2(1-r_{xz})^3},$$

unde **R** reprezintă determinanta matricei de corelație a celor trei variabile şi ř reprezintă media aritmetică a coeficienților de corelație comparați (în cazul nostru rxy+rzy/2). Determinanta **R** se calculează după formula $\mathbf{R} = 1- r_{xy}^2 - r_{zy}^2 - r_{xz}^2 + 2r_{xy}\, r_{zy}\, r_{xz}$.

Aplicând aceste formule la datele problemei obținem **R=**0.7. Înlocuind valorile problemei obținem un t = 2.4. Gradul de libertate utilizat pentru identificarea valorii critice in tabelul Student este df=n-3. Pentru pragul bilateral de ά = .05 în acest tabel vom citi valoarea critică de 2.026. Valoarea calculată este mai mare decât valoarea critică, ca urmare putem afirma că există o diferență semnificativă statistic între cei doi coeficienți de corelație comparați.

CONCLUZII

Semnificația statistică a unui coeficient de corelație, aşa cum apare în output-ul softurile statistice uzuale, ne arată că valoarea corelației la nivelul populației diferă de 0. Aceste softuri nu ne permite definirea unor limite de încredere care să indice o plajă de valori între care poate varia în populație valoarea coeficientului de corelație, dar nu ne permit nici procesări statistice (ex. compararea statistică) asupra coeficienților de corelație. Poate aceasta este una din cauzele faptului că cercetătorii deseori recurg la un demers intuitiv atunci când doresc să efectueze aceste comparații. Aşa cum s-a arătat în această lucrare, compararea coeficienților de corelație este un proces similar comparării mediilor. Neincluderea acestor operații inferențiale în pachetele statistice nu argumentează nicidecum ignorarea lor în practica de cercetare.

BIBLIOGRAFIE

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis foe the behavioral sciences*. London: Lawrence Elrbaum Associates

Cohen, H. B. (2001). *Explaining psychological statistics*. New York: John Willey & Sons

Hays, W. (1963). *Statistics for psychologists*. New York: Holt, Rinehart and Winston

Minium, E.W, King, B.M & Bear, G. (1993). *Statistical reasoning in psychology*. New York: John Willey & Sons