VII. STATISTIČKI I GRAFIČKI PRIKAZ BIVARIJACIONIH PODATAKA[1]

Neophodni matematički pojmovi za razumevanje teksta u ovoj glavi:[2]

Osnovni pojmovi teorije verovatnoće

Operator sabiranja (sumacioni operator) **Σ**

Operator dvostrukog sabiranja (dvojni sumacioni operator) **ΣΣ**

U glavama IV i V razmotrili smo statistički opis i grafički prikaz tzv. univarijacionih podataka, tj. podataka koje smo prikupili posmatrajući promene, tj. variranja na jednoj kvantitativnoj varijabli ili prateći raznolikost na jednoj kategoričkoj varijabli.[3] Sređivanjem podataka na jednoj varijabli dobijamo tzv. univarijacionu distribuciju/raspodelu učestalosti za kvantitativnu varijablu ili raspored učestalosti po kategorijama kategoričke varijable.
U ovoj glavi prikazaćemo kako statistički možemo opisati uzorak i grafički prikazati podatke koje dobijemo posmatrajući *zajedničke* promene na dvema kvantitativnim ili dvema kategoričkim varijablama. Premda postoji i mogućnost da od dveju varijabli na kojima istovremeno posmatramo promene jedna bude kvantitativna a druga kategorička, takav slučaj nećemo razmatrati u ovoj glavi.[4] Sređivanjem podataka dobijenih na uzorku pri posmatranju zajedničkih promena na dvema varijablama dobijamo tzv. bivarijacionu ili *zajedničku* distribuciju učestalosti. U tom slučaju distribucija svake od tih dveju varijabli predstavlja tzv. univarijacionu ili *marginalnu* distribuciju.

Kada statistički posmatramo zajedničke promene na dvema varijablama prevashodno se bavimo pitanjima statističke nezavisnosti i povezanosti tih dveju varijabli. Statistička nezavisnost dveju varijabli prosto znači da na osnovu pojedinačnih distribucija dveju varijabli možemo potpuno da definišemo njihovu zajedničku distribuciju. Ako su dve varijable, X i Y, statistički nezavisne, njihova zajednička distribucija, u oznaci $f(x,y)$, predstavljaće proizvod njihovih pojedinačnih, marginalnih distribucija:

$$f(x, y) = f(x)f(y).$$

---

[1] Manji deo teksta u ovoj glavi preuzet je iz Tenjović, 2020, glava XII.

[2] Osnovni pojmovi teorije verovatnoće prikazani su u Glavi 3, a ostale neophodne pojmove čitalac kojem je to potrebno može pronaći pod odrednicama **Operator sabiranja (sumacioni operator)** i **Operator dvostrukog sabiranja (dvojni sumacioni operator)** u Matematičkom pojmovniku u Dodatku **

[3] Prvi deo složenice univarijacioni, tj. uni- potiče od latinske reči *unus* (jedan) i ovde znači jednostruk ili jednočlan. Drugi deo složenice (-varijacioni) potiče od latinskog *variatio*, što znači menjanje ili promena. Dakle, univarijacioni slučaj (Glave IV i V) odnosio se na situaciju kada posmatramo promene, tj. variranja samo na jednoj promenljivoj, tj. variabli. Umesto termina univarijacioni često se sreće i termin univarijatni (nikako univarija**n**tni!) od engleskog *univariate*, što je isto što i univarijacioni. Pridev bivarijacioni (lat. *bis* = dvaput, udvojeno) u ovoj glavi označava da posmatramo zajedničke promene, tj. variranja na dvema varijablama istovremeno. Pri razmatranju statističkih postupaka često se, međutim, termini univarijacioni, bivarijacioni i multivarijacioni (lat. *multus* = mnogi, višestruk, mnogočlan) koriste kako bi se definisao broj **zavisnih** varijabli u analizi. To početnike može da zbuni, mada se iz konteksta uvek može razlučiti o kojoj od dve različite upotrebe ovih termina je reč.

[4] O istovremenom posmatranju jedne kategoričke i jedne kvantitativne varijable može se pročitati u Tenjović, 2020 (glave IX i X).

Setimo se da smo slično tome u Glavi 3 ustanovili da je verovatnoća zajedničkog dešavanja dva statistički nezavisna događaja jednaka proizvodu verovatnoća svakog od ovih događaja:

$$P(A \cap B) = P(A)P(B).$$

Odnos između dveju varijabli možemo da prikažemo tabelarno i grafički. Tabelarni prikaz odnosa između dveju varijabli prikazaćemo posebno za dve kvantitativne i za dve kategoričke varijable. Odnos između dveju kvantitativnih varijabli možemo prikazati i grafički korišćenjem Dekartovog koordinatnog sistema. Za istovremeni grafički prikaz dveju kategoričkih varijabli najčešće se koristi klasterski i „naslagani" (engl. s*tacked*) štapićasti dijagram.

Pretpostavimo da smo na 20 ispitanika prikupili podatke o njihovim merama na dvema ***kvantitativnim varijablama***, X i Y, koje mogu uzeti vrednosti od 1 do 10. U Tabeli 7.1 dat je prikaz zajedničke distribucije ovih dveju varijabli.

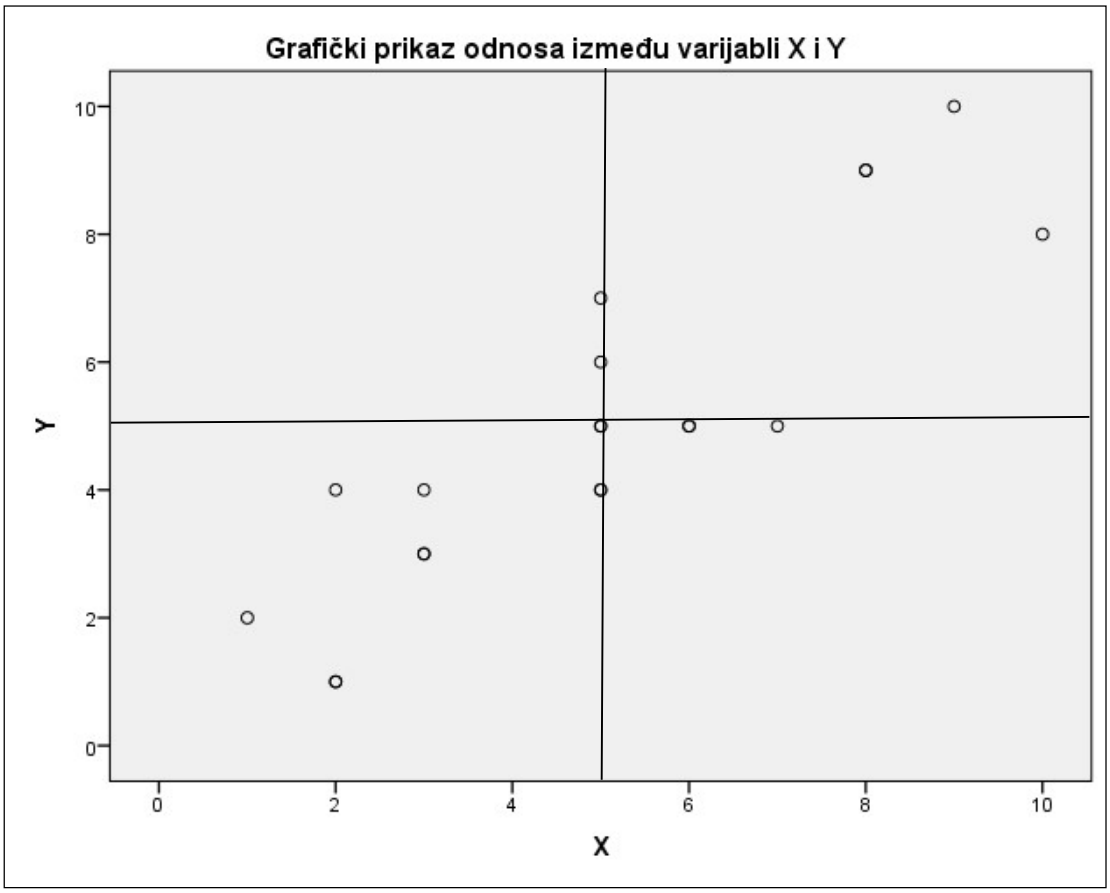
Tabela 7.1
Primer tabelarnog prikaza zajedničke distribucije dveju kvantitativnih varijabli

X→ Y ↓	1	2	3	4	5	6	7	8	9	10	f_Y
1		2									2
2	1										1
3			2								2
4		1	1		2						4
5					2	2	1				5
6					1						1
7					1						1
8										1	1
9								2			2
10									1		1
f_X	1	3	3	0	6	2	1	2	1	1	n = 20

U prvom redu Tabele 7.1 su vrednosti koje može uzeti varijabla X, a u prvoj koloni su vrednosti koje može uzeti varijabla Y. Brojevi u ćelijama tabele gde se sustiču kolone sa vrednostima na varijabli X i redovi sa vrednostima na varijabli Y predstavljaju *zajedničke frekvencije*. To su zajedničke frekvencije jer pokazuju koliko ispitanika na obema varijablama istovremeno ima određene rezultate. Tako, cifra 2 u ćeliji u kojoj se sustiču kolona 5 i red 4 označava da dva ispitanika na varijabli X imaju meru 5 dok istovremeno na varijabli Y imaju meru 4. Red f_X prikazuje marginalnu distribuciju varijable X, a kolona f_Y marginalnu distribuciju varijable Y. Treba uočiti da marginalne frekvencije za varijablu X predstavljaju zbirove zajedničkih frekvencija po kolonama a marginalne frekvencije varijable Y predstavljaju zbirove zajedničkih frekvencija po redovima. Naravno, zbir marginalnih frekvencija na svakoj od varijabli jednak je ukupnom broju parova rezultata, a u ovom slučaju jednak je i veličini uzorka pošto za sve ispitanike imamo podatke na obema varijablama. Kada bi ove dve varijable bile statistički nezavisne, onda bismo zajedničke frekvencije u svim ćelijama tabele mogli dobiti kao količnik proizvoda marginalnih frekvencija u redu i koloni kojoj ćelija pripada i veličine uzorka. Na primer, zajednička

frekvencija u ćeliji gde se ukršta red 5 i kolona 5 bila bi (5 * 6) / 20, tj. 1.5 a ne 2 kao što je ovde slučaj.
Zajednička distribucija dveju ***kvantitativnih*** varijabli uobičajeno se grafički prikazuje dijagramom raspršenja (engl. *scatterplot*). Dijagram raspršenja za varijable X i Y čiju smo zajedničku distribuciju prikazali u Tabeli 7.1 prikazan je na Grafiku 7.1.

Grafik 7.1
Primer grafičkog prikaza zajedničke distribucije dveju kvantitativnih varijabli



Pretpostavimo sada da smo na uzorku mladih osamnaestogodišnjaka iz Srbije (n = 473) ispitivali postoji li na tom uzrastu veza između toga kog su pola i da li imaju romantičnog partnera. Zajedničku distribuciju ovih dveju ***kategoričkih*** varijabli možemo prikazati pomoću tzv. tabele kontingencije, kao što je Tabela 7.2.

Tabela 7.2.
Primer tabelarnog prikaza zajedničke distribucije dveju kategoričkih varijabli

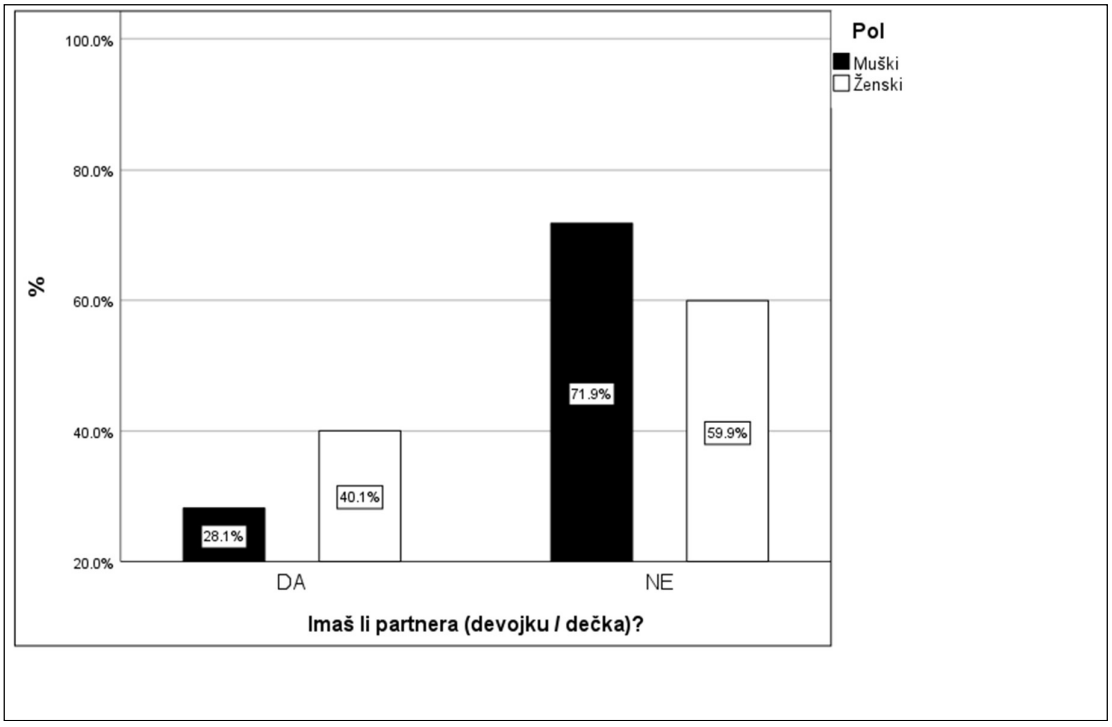
*Pol * Ima li partnera (devojku / dečka)?*

	Imaš li partnera (devojku / dečka)?	Ukupno (f_{j+})

			DA	NE	
Pol	Muški	f_{jk}	34	87	121
		%	28.1%	71.9%	
	Ženski	f_{jk}	141	211	352
		%	40.1%	59.9%	
Ukupno		(f_{+k})	175	298	n = 473
		%	37.0%	63.0%	

Oznakom f_{jk}, $j = 1, 2, \ldots, r$, $k = 1, \ldots, c$, označene su zajedničke frekvencije u tabeli kontingencije. Prvi indeks, undeks *j*, jeste oznaka reda u kojem je ćelija, a drugi indeks, indeks *k*, predstavlja oznaku kolone u kojoj se nalazi ćelija. Oznaka f_{j+} označava marginalnu frekvenciju koja predstavlja zbirove frekvencija u redu *j* po svim kolonama (umesto oznake kolone stoji znak + pošto se sabiraju frekvencije po svim kolonama). Oznaka f_{+k} označava marginalnu frekvenciju koja predstavlja zbirove frekvencija za sve redove u koloni *k*.

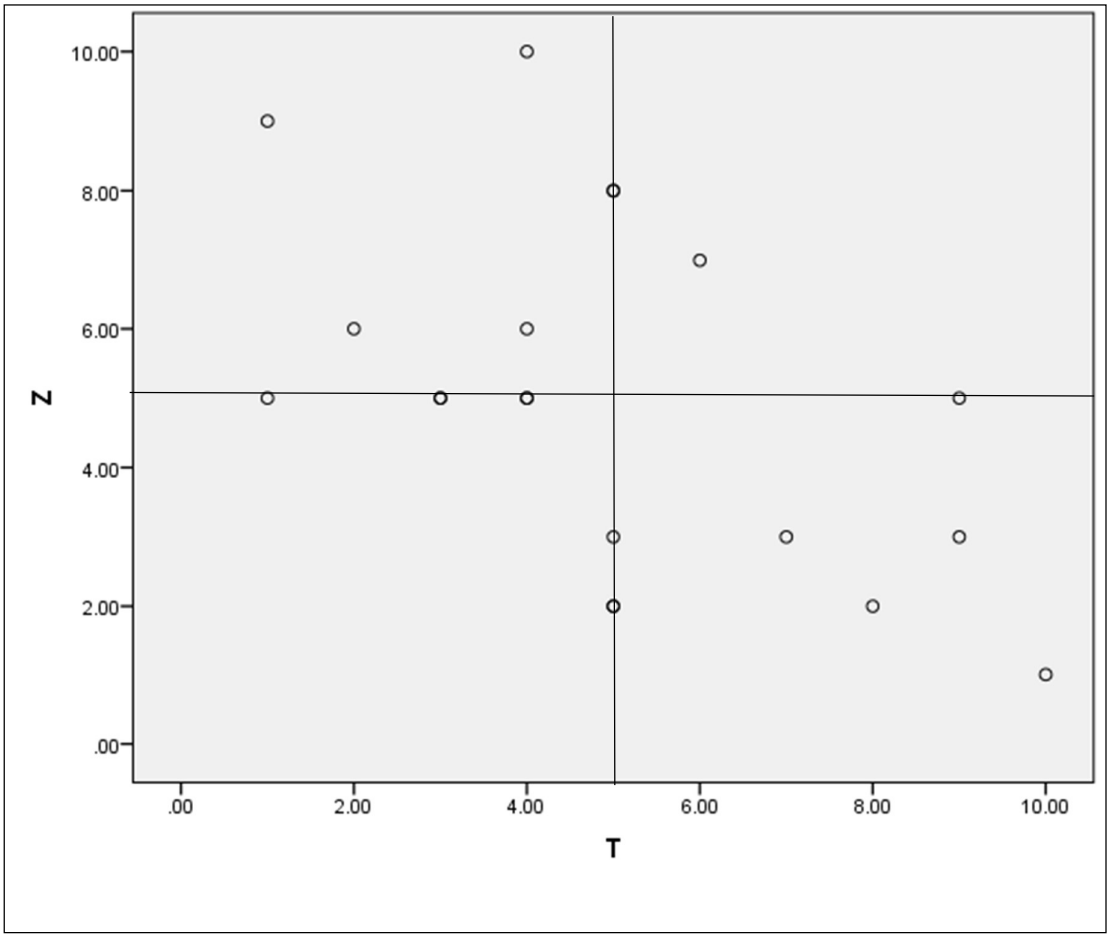
Grafik 7.2
Primer grafičkog prikaza zajedničke distribucije dveju kategoričkih varijabli klasterskim štapićastim dijagramom



Od grafičkih prikaza ka formulama koje iskazuju povezanost dveju kvantitativnih, odnosno dveju kategoričkih varijabli

Grafički prikaz odnosa dveju kvantitativnih varijabli na Grafiku 7.1 možemo podeliti na četiri kvadranta vertikalmom i horizonatlnom linijom, pri čemu vertikalnu linija povučemo iz aritmetičke sredine varijable X a horizontalnu liniju povučemo iz aritmetičke sredine varijable Y. U tom slučaju lako možemo po broju tačaka, tj. ispitanika po kvadrantima vizuelno orijentaciono odrediti kakva je veza između varijabli. Ako je većina tačaka istovremeno raspoređena u obliku šire ili uže elipse u donjem levom i gornjem desnom kvadrantu (Grafik 7.1) onda je korelacija pozitivna, a ako je većina tačaka raspoređena u obliku šire ili uže elipse u gornjem levom i donjem desnom kvadrantu onda je korelacija negativna (Grafik 7.3). Što je oblik elipse po kojem su raspoređene tačke uži to je korelacija veća. Što se elipsa više približava krugu tačke su relativno podjednako raspoređene po svim kvadrantima i korelacija se bliži nuli. U Ekstremnom slučaju, ako su sve tačke raspoređene na jednoj pravoj liniji, korelacija je maksimalna, tj. jednaka +1 ili -1.

Grafik 7.3.
Primer negativne korelacije dveju varijabli



Matematički, sledeći ovu logiku možemo lako doći do obrasca za koeficijent linearne korelacije: potrebno je vrednost svake tačke (x_i i y_i)-iskazati kao odstupanje od njene aritmetičke sredine (x_i-M_X; y_i-M_Y) a potom sva odstupanja na obema varijablama međusobno izmnožiti. Ove proizvode zovemo unakrsnim proizvodima (engl. *Cross-*

product). Na kraju, sabraćemo sve unakrsne proizvode. Tačke u gornjem levom kvadrantu manje su od aritmetičke sredine na varijabli X i njihova odstupanja od aritmetičke sredine varijable X biće negativna. Tačke u donjem desnom kvadrantu manje su od aritmetičke sredine na varijabli Y te će njihova odstupanja od aritmetičke sredine varijable Y biti negativna. Tačke u donjem levom kvadrantu imaće negativna odstupanja na obema varijablama, a tačke u gornjem desnom kvadrantu pozitivna odstupanja na obema varijablama. Stoga će tačke u gornjem levom i donjem desnom kvadrantu davati negativne unakrsne proizvode, dok će tačke u donjem levom i gornjem desnom kvadrantu davati pozitivne unakrsne proizvode.

Zbir unakrsnih proizvoda podeljen sa n ili češe sa n-1 daje meru linearne povezanosti dveju kvantitativnih varijabli koja se zove kovarijansa.[5]
Dakle, kovarijansa uzorka[6], u oznaci S_{XY}, definiše se na sledeći način:

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{n-1}$$

Međutim, kovarijansa govori jasno samo o smeru povezanosti varijabli. Ukoliko je kovarijansa negativna, varijable su u linearnoj negativnoj vezi a ako je kovarijansa pozitivna varijable su pozitivno povezane. Na osnovu kovarijanse ne možemo zaključiti ništa precizno o jačini povezanosti jer njena veličina zavisi od jedinica kojima su iskazane vrednosti na varijablama. Stoga je kovarijansu potrebno sameriti, tj. staviti u odnos sa maksimalno mogućom kovarijansom za date jedinice (ili skale) kojima su iskazane varijable. Matematički se može dokazati da je maksimalna kovarijansa za varijable iskazane datim jedinicama jednaka proizvodu standardnih devijacija tih varijabli. Podelom dobijene kovarijanse maksimalno mogućom kovarijansom za varijable iskazane datim skalama zapravo dobijamo koeficijent linearne korelacije kao standardizovanu kovarijansu. Na taj način dobijamo obrazac za računanje koeficijenta korelacije preko kovarijanse:[7]

$$r_{XY} = \frac{S_{XY}}{S_X * S_Y}$$

Dakle, koeficijent linearne korelacije predstavlja količnik empirijski dobijene kovarijanse, S_{XY} i maksimalno moguće kovarijanse za varijable iskazane datim skalama, tj. proizvoda standardnih devijacija dveju varijabli. Iz ovog obrasca mogu se jasno uočiti različiti odnosi koji postoje između kovarijanse i koeficijenta linearne korelacije (na primer, u pogledu predznaka, u pogledu uslova kada su ove dve mere jednake i slično).
Opšte je prihvaćeno da je indeks, tj. formulu za koeficijent korelacije matematički definisao Karl Pirson 1895. godine, deceniju nakon što je, baveći se problemima herediteta, Sir Frensis Golton definisao teoriju „regresije“ i korelacije. Pre toga, Ogist Brave (Auguste Bravais), francuski mornarički oficir i astronom, definiše tzv. bivarijacionu normalnu distribuciju (zajedničku distribuciju dveju varijabli koja je normalna i ima pet parametara) nazvavši

---

[5] Deljenjem kovarijanse sa n-1 dobijamo manje pristrasan ocenitelj kovarijanse populacije. Pristrasnost ocenitelja parametara objašnjena je u glavi IX.

[6] Kovarijansa populacije, u oznaci σ_{XY}, matematički se definiše na sledeći način:

$$\sigma_{XY} = E\{[X - E(X)][Y - E(Y)]\} = E(XY) - \mu_X\mu_Y$$

pri čemu je *E* oznaka matematičkog očekivanja.

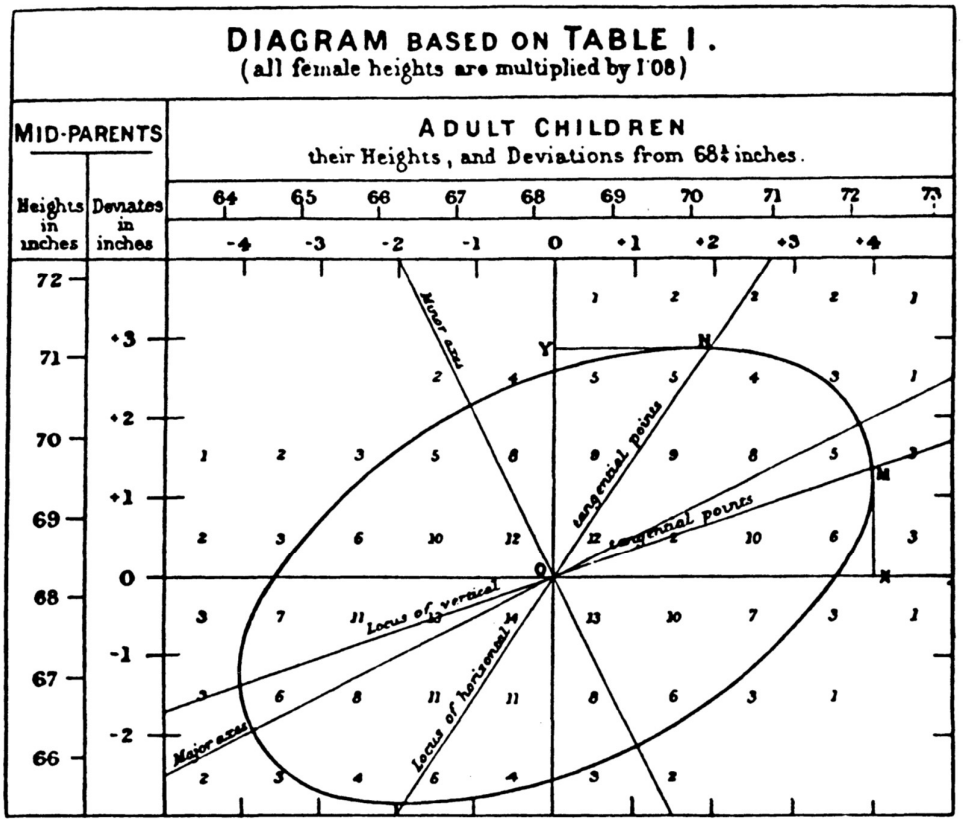
[7] Koeficijent linearne korelacije populacije, u oznaci ρ_{XY}, matematički se definiše na sledeći način:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Pri tome, σ_{XY} je kovarijansa populacije između varijabli X i Y, a σ_X i σ_Y standardne devijacije populacije za varijable X i Y.

jedan od parametara ove distribucije korelacijom. Braveu pojedini autori pripisuju i prvu definiciju formule za koeficijent korelacije. Stoga ćemo u ovoj knjizi koeficijent linearne korelacije najčešće zvati Brave-Pirsonovim koeficijentom.

Grafik 7.4.
Originalni Goltonov dijagram odnosa između visine roditelja i dece



Preuzeto iz: American Statistician, February 1988, *42*(1), str.60.

Na osnovu Goltonovog dijagrama, koji po nekim elementima nalikuje onome sa Grafika 7.1, i koji je prikazan na Grafiku 7.4, matematičar Karl Pirson je definisao koeficijent linearne korelacije, u oznaci r_{XY}:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - M_X)^2}\sqrt{\sum_{i=1}^{n}(y_i - M_Y)^2}}$$

U brojiocu ove formule je suma unakrsnih proizvoda dveju varijabli, a u imeniocu su sume kvadriranih odstupanja rezultata na varijablama od odgovarajućih aritmetičkih sredina. Dakle, Pirson je definisao koeficijent linearne korelacije kao funkciju sirovih skorova na varijablama i odgovarajuih aritmetičkih sredina. Uočimo identičnost Pirsonove formule sa formulom koja je izvedena kao funkcija kovarijanse i standardnih devijacija varijabli: ukoliko brojilac Pirsonove formule podelimo sa n-1 dobićemo kovarijansu, a ukoliko svaku potkorenu sumu kvadriranih odstupanja podelimo sa n-1 dobićemo varijanse, čiji pozitivni kvadratni korenovi predstavljaju standardne devijacije varijabli.
Prema tome, brojilac formule za koeficijent linearne korelacije ukazuje na zajedničko/deljeno variranje dveju varijabli a imenilac na njihovo pojedinačno variranje. Na prvi pogled, što je zajedničko variranje/kovariranje dveju varijabli veće a njihovo

pojedinačno variranje manje, koeficijent linearne korelacije bi trebalo da bude veći. Ipak, ovo ne treba sasvim jednostavno posmatrati budući da zajedničko variranje dveju varijabli i njihovo pojedinačno variranje nisu nezavisni i ne mogu se jednostavno odvojeno posmatrati. Naime, veće ili manje variranje jedne varijable može da vodi većem ili manjem zajedničkom variranju varijabli. Na primer, smanjenje varijabilnosti jedne od varijabli dovodi do manje korelacije sa drugim varijablama. To je fenomen poznat kao „restrikcija raspona". To se, na primer, dešava kada koreliramo uspeh na testu znanja psihologije na prijemnom ispitu sa uspehom na studijama. U tom slučaju, za varijablu uspeh na testu znanja psihologije uzimamo u račun samo podatke za one koji su primljeni na studije, te na varijabli uspeh na testu znanja psihologije dolazi do restrikcije raspona. Postoje načini da izvršimo korekciju koeficijenta korelacije za restrikciju raspona, ali ovi postupci, koji nisu sasvim bez problema, nadilaze namenu ove knjige, te ih nećemo prikazivati.[8]
Ako se obe varijable standardizuju, koeficijent linearne korelacije jednak je prosečnom unakrsnom proizvodu, tj. kovarijansi standardizovanih varijabli:[9]

$$r_{XY} = \frac{\sum_{i=1}^{n} (z_{x_i} z_{y_i})}{n-1}.$$

Dakle, i kovarijansa i korelacija pokazuju koliko mere jedne varijable "idu ruku pod ruku" sa merama druge varijable.[10] Kovarijansa i korelacija pokazuju stepen kovariranja dveju varijabli, tj. stepen u kojem promene jedne varijable bivaju praćene promenama u drugoj varijabli. Drugačije rečeno, njima se matematički iskazuju stepen u kojem jedinice posmatranja (ili entiteti) imaju isti relativni položaj u odnosu na ostale entitete u pogledu dveju varijabli.

A sada bi bilo dobro da na osnovu svega do sada prikazanog o koeficijentu linearne korelacije, čitalac proceni smer i visinu koeficijenta linearne korelacije za podatke prikazane na Grafiku 7.5. Podaci prikazani na tom dijagramu raspršenja predstavljaju čuvene podatke o inteligenciji 34 para jednojajnih blizanaca odgajanih odvojeno iz istraživanja Džejmsa Šildsa (James Shields) iz 1962. godine, koji su za ovu knjigu preuzeti iz Farnsworth, 2014. Na Grafiiku 7.5. na X osi date su vrednosti skora inteligencije za blizanca koji je prvi rođen a na Y osi za blizanca koji je rođen drugi. Koliki je, dakle Brave-Pirsonov koeficijent u ovom slučaju?

Grafik 7.5.

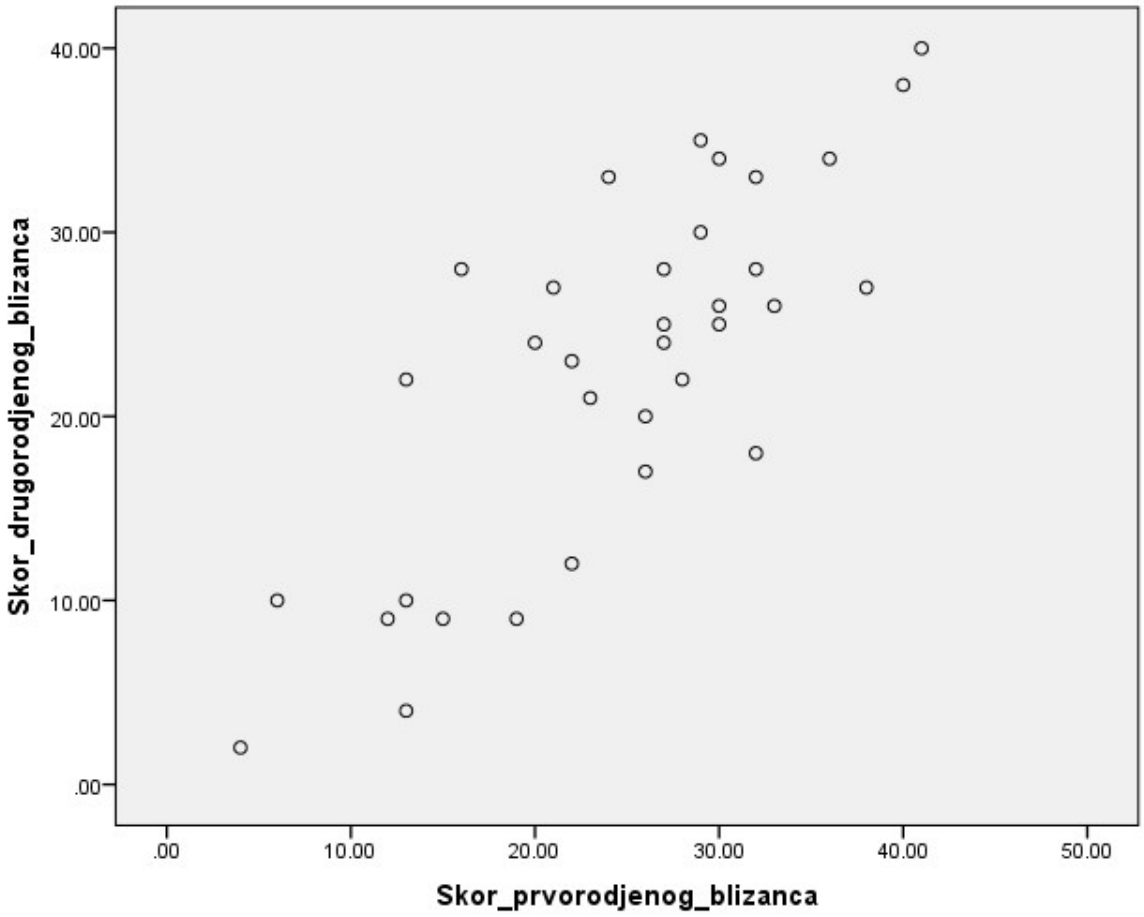
---

[8] O tome se može pročitati, na primer, u Nunnally, & Bernstein, 1994.

[9] Sva dosadašnja određenja koeficijenta linearne korelacije podrazumevala su da smo povezanost dveju kvantitativnih varijabli grafički predstavili dijagramom raspršenja. Na dijagramu raspršenja varijable predstavljamo kao ose koje definišu prostor u koji smeštamo jedinice posmatranja prema njihovim rezultatima na dvema varijablama.Ako pak varijable centriramo i predstavimo ih kao vektore u dvodimenzionalnom potprostoru tzv. „prostora ispitanika", tada je koeficijent linearne korelacije jednak kosinusu ugla α koji ove dve vektorski predstavljene centrirane varijable zaklapaju jedna sa drugom:
$r_{XY} = \cos(\alpha)$

[10] Reč kovarijansa jasno ukazuje na meru koja govori o zajedničkom variranju dveju varijabli. Reč korelacija potiče od novolatinske reči correlatio (co+relatio) što znači saodnos, zajednički odnos, uzajamni odnos, uzajamnu vezu. Dakle, korelacija takođe podrazumeva da posmatramo udruženo ili zajedničko variranje dveju ili više varijabli.

Dijagram raspršenja skorova inteligencije 34 para identičnih blizanaca gajenih odvojeno



Podaci za grafik preuzeti su iz Farnsworth, D. L. (2014). Identical Tweens Raised Apart. *Teaching Statistics*, *37*(1), 1–6, strana 2.

Pretpostavljam da nije bilo nimalo teškoća u pogađanju. Očigledno, Brave-Pirsonov koeficijent je u ovom slučaju pozitivan i visok i iznosi čak 0.79. Sudeći po ovim podacima veliki deo varijabilnosti u inteligenciji posledica je genetskih faktora budući da je koeficijent determinaciije, tj. kvadrat koeficijenta korelacije .6241 a blizanci su odgajani u različitim sredinama. U ovom slučaju koeficijent determinacije se zove koeficijent heritabilnosti.

Prosečna korelacija

Brave–Pirsonov koeficijent korelacije je indeksni broj koji pokazuje tip (s obzirom na predznak) i jačinu linearne povezanosti između varijabli. Koeficijent linearne korelacije može se kretati u granicama od -1 do +1. Ipak, kada se distribucije varijabli X i Y razlikuju po obliku, maksimalni koeficijent linearne korelacije manji je od jedinice. Koeficijenti korelacije nisu brojevi na linearnoj skali, te se ne može reći za r = 0.50 da je dva puta veće od r = 0.25. Stoga, ukoliko se želi računanje prosečne vrednosti koeficijenata linearne korelacije, najsmisaonije je naći kvadratni koren proseka kvadriranih koeficijenata linearne korelacije. Drugi način za računanje prosečne vrednosti koeficijenata korelacije jeste da se r

koeficijenti pretvore u Fišerove z-statistike (u oznaci z_F) transformacijom

$$z_F = 0.5\ln [(1 + r)/(1- r)],$$

a potom na osnovu dobijenih koeficijenata z_F izračuna prosečno z_F. Zatim se dobijeno prosečno z_F pretvori nazad u r transformacijom

$$r = (e^{2z_F} - 1) / (e^{2\,z_F} + 1),$$

gde je e = 2.71828.

Koeficijent korelacije nema osobinu tranzitivnosti. Ako je korelacija između dveju varijabli (X i Y) jednaka 0.80, a korelacija varijable X sa trećom varijablom Z iznosi isto 0.80, to nikako ne znači da je korelacija varijable Y sa varijablom Z jednaka 0.80. Korelacija varijable Y sa varijablom Z može u ovom slučaju biti čak jednaka 0.28.[11]

Kovarijansa, linearna korelacija i statistička nezavisnost varijabli

Koeficijent linearne korelacije između varijabli jednak nuli ili **kovarijansa među varijablama jednaka nuli ne znače nužno** i **statističku nezavisnost** dveju kvantitativnih varijabli. Statistička nezavisnost dveju kvantitativnih varijabli postoji onda kada je njihova zajednička, bivarijaciona distribucija jednaka proizvodu njihovih marginalnih, tj. univarijacionih distribucija. S druge strane, kada su varijable statistički nezavisne, kovarijansa i linearna korelacija nužno su jednake nuli.[12]

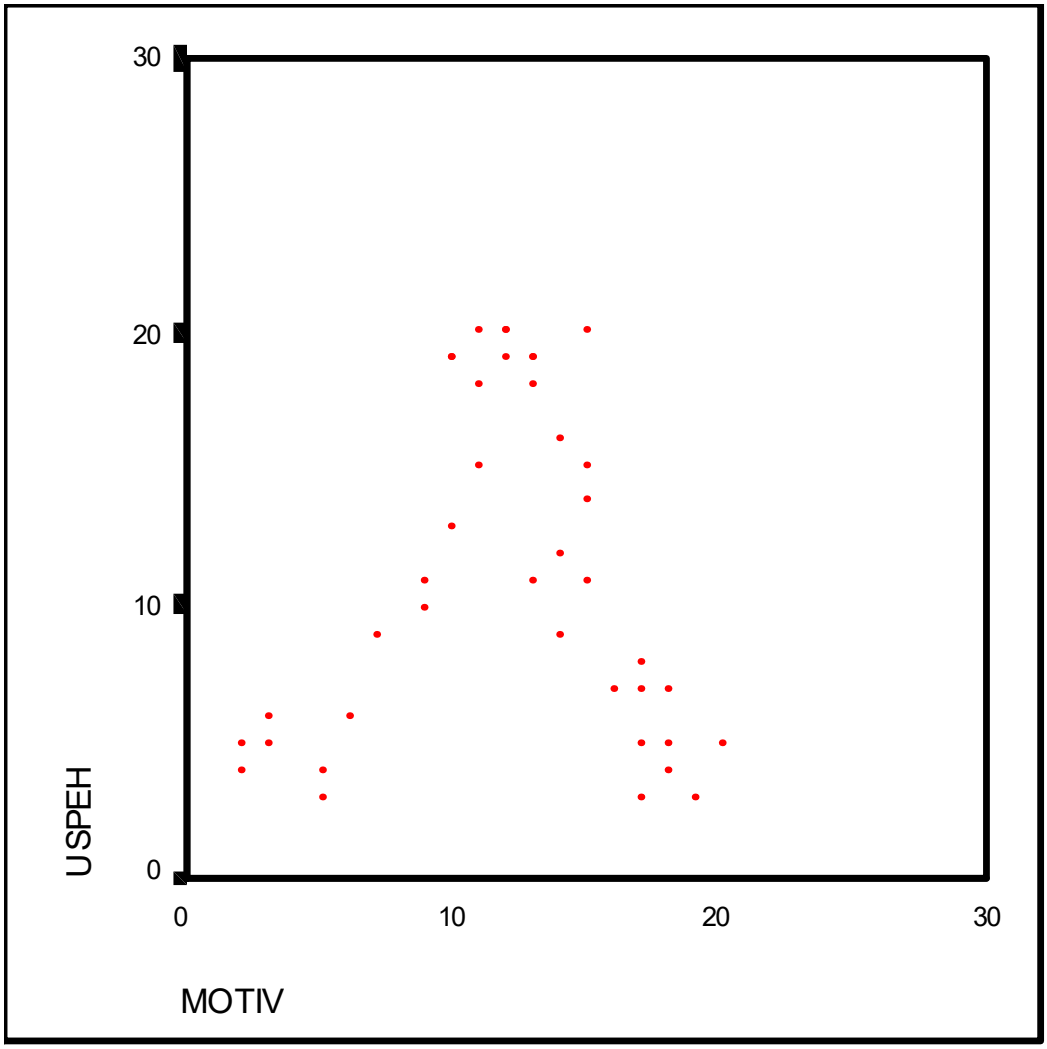
Linearna i nelinearna povezanost između kvantitativnih varijabli

Pored linearne povezanosti, između dveju varijabli može postojati i nelinearna povezanost. Ako je određena promena vrednosti na jednoj varijabli generalno praćena isto tolikom promenom vrednosti na drugoj varijabli, tada su dve varijable u linearnoj vezi. Međutim, veza između dveju kvantitativnih varijabli može biti i nelinearna, tj. znatno složenija od linearne. Na primer, porast vrednosti na jednoj varijabli može biti konzistentno praćen ili porastom ili opadanjem vrednosti na drugoj varijabli, ali promene na dvema varijablama ne moraju biti podjednako velike. U tom slučaju reč je o monotonoj povezanosti među varijablama. Na primer, određene promene na jednoj varijabli mogu biti praćene znatno većim, tj. bržim promenama, na drugoj varijabli. Ukoliko među dvema varijablama postoji monotona povezanost i pri tome su promene na jednoj varijabli praćene podjednakim promenama na drugoj varijabli, tada možemo reći da među tim varijablama postoji linearna povezanost. Prema tome, linearna povezanost među varijablama je specijalni slučaj monotone povezanosti. S druge strane, monotona povezanost koja nije linearna predstavlja samo jedan tip nelinearne povezanosti među varijablama. Tipova nelinearne povezanosti između dveju varijabli može biti beskonačno mnogo. Na primer, porast vrednosti na varijabli X može biti u određenom rasponu vrednosti varijable X praćen porastom vrednosti na varijabli Y, dok u nekom drugom rasponu vrednosti varijable X, porast vrednosti na istoj varijabli biva praćen stagniranjem ili padom vrednosti na varijabli Y.

[11] Najniža korelacija varijable Y sa trećom varijablom Z, ukoliko X i Y, kao i X i Z imaju jednaku korelaciju r sa Z jednaka je $2r^2$-1.

[12] O statističkoj nezavisnosti varijabli može se podrobnije pročitati u Vuković (1997), str. 94–97 ili u Stilson (1966), str. 165–182.

Grafik 7.6.
Dijagram raspršenja stepena motivisanosti i uspeha u obavljanju aktivnosti



Sa Grafika 7.6 može se uočiti jasna nelinearnost veze ovih dveju varijabli. Do određenog nivoa, sa porastom motivisanosti generalno raste i uspešnost, a posle tog optimalnog nivoa, sa daljim porastom motivacije uspešnost, prosečno gledano, opada. Adekvatna mera povezanosti ovih dveju varijabli bio bi korelacioni razmer a ne koeficijent linearne korelacije.

Dobro poznati primeri nelinearnih veza u psihologiji postoje između jačine motivacije i uspešnosti u obavljanju neke aktivnosti (Grafik 7.6), te između inteligencije i dužine zadržavanja na istom radnom mestu. Tip povezanosti među dvema kvantitativnim varijablama najbolje se može uočiti ako se ta veza prikaže grafički. Stoga je pre računanja koeficijenta linearne korelacije između dveju varijabli neophodno grafički prikazati vezu između njih kako bi se izbeglo računanje ovog koeficijenta kada među varijablama postoji nelinearna povezanost.

Ukoliko se pregledom dijagrama raspršenja ustanovi da je generalni trend tačaka izrazito nelinearan (ne formira manje ili više izduženu elipsu ili krug), onda nema smisla ni

računati koeficijent linearne korelacije. Tada se za utvrđivanje nelinearne povezanosti među varijablama može koristiti korelacioni racio o kojem se podrobnije može pročitati u udžbenicima statistike koji obrađuju nelinearne veze među varijablama.[13]

Ocenjivanje koeficijenta linearne korelacije u populaciji

Ukoliko na osnovu koeficijenta korelacije dobijenog na uzorku (r) **ocenjujemo** kolika je linearna korelacija između varijabli u populaciji (ρ), r je utoliko pristrasniji ocenitelj parametra što je uzorak manji.[14] Nepristrasna ocena linearne korelacije populacije (u oznaci $r_{Adj.}$, supskript Adj., od engleskog Adjusted = korigovan) može se dobiti iz r na sledeći način:

$$r_{Adj.} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}\ .$$

Tumačenje koeficijenta linearne korelacije

Koeficijent linearne korelacije treba pre svega posmatrati kao pokazatelj linearne veze između dveju varijabli. Pri njegovom tumačenju treba uzeti u obzir smer i jačinu veze. Ukoliko koeficijent linearne korelacije ima pozitivan predznak, veza između varijabli je pozitivna, a ukoliko je predznak ovog koeficijenta negativan veza između varijabli je negativna ili inverzna. Pozitivna veza među varijablama znači da je porast vrednosti na jednoj varijabli praćen porastom vrednosti na drugoj varijabli, a negativna veza ukazuje na to da je povećanje vrednosti na jednoj varijabli praćeno smanjivanjem vrednosti na drugoj varijabli.

Tumačenje visine koeficijenta linearne korelacije nije uvek jednostavno pošto u psihologiji i srodnim oblastima ne postoje precizno definisani pragovi za nisku/slabu, umerenu/osrednju i visoku/jaku korelaciju/povezanost. Često se koeficijent oko .10 (između 0 i .20) tumači kao da ukazuje na nisku povezanost, onaj oko .30 (između .20 i .50) smatra umerenim, dok se koeficijent preko .50 posmatra kao da ukazuje na jaku ili snažnu povezanost između varijabli. Prema drugom kriterijumu nizak koeficijent korelacije je onaj koji je manji od .30, umerenim/osrednjim koeficijentom se smatra onaj između .30 i .49, a visokim koeficijent koji je .50 i više. Budući da nema jednoznačnih kriterijuma za kategorije visine koeficijenta korelacije možda je najbolje rešenje da istraživač koji se bavi određenom

[13]Korelacioni racio, u oznaci $\eta_{Y.X}$, definiše se na sledeći način:

$$\eta_{Y.X} = \sqrt{\frac{S_Y^2 - S_{Y.X}^2}{S_Y^2}} = \sqrt{1 - \frac{S_{Y.X}^2}{S_Y^2}}, \text{ gde je } S_{Y.X}^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + ... + n_k S_k^2}{n}$$

U ovom obrascu S^2_Y je varijansa varijable Y, $S^2_{Y.X}$ je prosečna (uslovna) varijansa niza vrednosti na varijabli Y za pojedine vrednosti varijable X, n_1, n_2,..., n_k su brojevi rezultata na Y varijabli za svaku od k pojedinih vrednosti X varijable, dok je $n = n_1 + n_2 + ... + n_k$. Razumevanje korelacionog racija podrazumeva poznavanje regresione analize . Korelacioni racio je asimetrična mera asocijacije: korelacioni racio u slučaju kada je Y zavisna a X nezavisna varijabla, nije nužno jednak korelacionom raciju u slučaju kada je X zavisna a Y nezavisna varijabla. Korelacioni racio može se koristiti i u slučaju kada je nezavisna varijabla kategorička. U tom slučaju ovaj koeficijent naziva se Fišerov intergrupni koeficijent ili eta-koeficijent. Korelacioni racio jednak je koeficijentu linearne korelacije u slučaju kada je regresija Y varijable na X varijablu tačno linearna. Korelacioni racio jednak je 1 akko su X i Y u striktnoj funkcionalnoj (linearnoj ili nelinearnoj) vezi. Stepen u kojem je korelacioni racio veći od koeficijenta linearne korelacije predstavlja pokazatelj nelinearnosti veze između dveju varijabli (Tenjović, 2020).

[14] Ocenjivanje parametara objašnjeno je u glavi **IX**.

oblašću zaključak o jačini veze između varijabli donosi uzimajući u obzir visinu korelacija koje se sreću u datoj oblasti. Na primer, u oblasti u kojima se retko sreću koeficijenti korelacije veći od 0.30 takvi koeficijenti se mogu smatrati relativno visokim.

Koeficijent linearne korelacije može se posmatrati i probabilistički: što više u skupu jedinica posmatranja postoji jedinica posmatranja čiji se parovi podataka slažu po svom redosledu u skupu podataka koeficijent je utoliko veći. Šta zapravo znači ovo slaganje parova podataka po redosledu? Ako su, na primer, oba rezultata za ispitanika A veći, odnosno manji nego parovi rezultata za ispitanika B, onda postoji takvo slaganje. Ukoliko samo kod 50% ispitanika važi isti relativni redosled parova podataka ovaj će koeficijent biti jednak nuli. Ako pak kod oko 70% ispitanika postoji isti redosled parova podataka onda će koeficijent linearne korelacije biti približno 0.75.

Treba voditi računa i o tome da visoka korelacija između dva testa ne znači nužno da ova dva testa mere istu osobinu. Ona to može značiti, ali ne nužno samo na osnovu visoke korelacije između testova. Naime, velika većina ispitanika u tom su slučaju na ovim testovima na gotovo istim pozicijama u odnosu na ostale ispitanike u uzorku. Ali to ne znači da njihovi skorovi na ova dva testa mere istu osobinu i da su međusobno zamenljivi. Oni čak mogu biti po veličini sasvim različiti jer visoka korelacija između dva testa ne znači da su testovi jednaki po dobijenim skorovima, aritmetičkim sredinama i varijansama (cf. Branch, 1990).

Pokatkad se kvadrat koeficijenta linearne korelacije pogrešno tumači. Ovaj kvadrat prosto označava proporciju varijanse u jednoj varijabli (recimo Y) koja se može objasniti (ili predvideti) na osnovu varijabilnosti (varijanse) druge varijable (recimo X). Ali, na osnovu ovoga ne može se ništa reći o odnosima relativnih nivoa razvoja osobina koje su u korelaciji. Na primer, na osnovu korelacije od 0.71 između inteligencije dece na uzrastima od 4 i 17 godina besmisleno bi bilo reći da deca od 4 godine dostižu 50% (0.71^2) nivoa njihove inteligencije u odraslom dobu. (Falk, & Well, 1997)

Zapamtite:

- Linearnost veze među varijablama, koja je pretpostavka za računanje koeficijenta linearne korelacije, treba uvek proveriti, barem grafički (dijagramom raspršenja).
- Koeficijent linearne korelacije između dveju varijabli na uzorku koji je homogen u pogledu jedne od ovih varijabli manji je (zbog restrikcije raspona) nego na neselekcionisanom uzorku. Na primer, ukoliko su ispitanici probrani u pogledu inteligencije tako da su na studije primljeni samo kandidati sa IQ-om iznad 120, onda će (zbog restrikcije raspona na varijabli inteligencija) koeficijent korelacije između inteligencije i uspeha na studijama biti niži no što bi bio kada bi se računao na uzorku svih prijavljenih kandidata (što praktično nije moguće).
- Veoma je važno ne tumačiti korelaciju automatski kao uzročno-posledičnu vezu. Varijable, između kojih postoji povezanost, mogu zaista biti u uzročno-posledičnoj vezi, ali nam sama korelacija direktno ne kazuje da li je određena veza uzročno posledična. Međutim, varijable koje su u kauzalnoj vezi moraju biti u korelacionoj. Postojanje korelacije možemo tumačiti na dva načina: 1. jedna varijabla je važna komponenta druge varijable (ona čini r^2*100% komponenti druge varijable) ili 2. dve varijable imaju r^2*100% zajedničkih elemenata, tj. neki treći faktor je sadržan u obema. Na primer, dobijena je korelacija između broja roda i broja novorođene dece. Prema tome, mogli bismo (pogrešno!) zaključiti da rode, ipak, donose decu! Takav zaključak bi bio fatalan po nas i po ceo ljudski rod. Naravno, uzrok ove korelacije je veličina mesta jer što je mesto veće više je i dimnjaka (roda) a i dece! Isto tako, iz korelacije pušenja i učestalosti raka pluća, a ona postoji, ne smemo odmah zaključiti da je pušenje uzrok raka pluća (ono to može biti ali ne mora nužno, obe stvari mogu biti posledica dublje genetske ili psihološke osobine). Ipak, za neke oblike raka pluća uzročno-posledična veza sa pušenjem je prilično verovatna.

  Da bi X varijabla bila uzrok Y varijable moraju biti ispunjeni sledeći uslovi:

  1. X vremenski prethodi Y;
  2. X i Y kovariraju, tj. postoji korelacija između ovih varijabli;
  3. Varijabla Y nije uzrokovana nekom drugom varijablom koja je uključena u odnos između varijabli X i Y. Drugim rečima, Y varijabla se ne može posmatrati kao posledica delovanja drugih varijabli, osim varijable X, niti postoji varijabla koja se može posmatrati kao uzrok obeju varijabli.
- Koeficijent linearne korelacije je invarijantan na određene linearne transformacije. Dakle, ako su b, c, d i e konstante tako da su b i d veći od 0, korelacija između varijabli Z i W, pri čemu je Z = bX + c a W = dY + e, ista je kao i korelacija između varijabli X i Y. Prema tome, svejedno je da li ćemo korelaciju između dveju varijabli računati na osnovu sirovih (netransformisanih) rezultata, na osnovu devijacionih (centriranih) skorova ili korišćenjem standardizovanih mera.
- Koeficijent linearne korelacije ima veoma ograničenu maksimalnu vrednost (daleko manju od jedan) kada su distribucije dveju varijabli asimetrične u suprotnim smerovima ili ako je jedna od varijabli dihotomna (svedena na dve kategorije) pri čemu je proporcija slučajeva u kategorijama nejednaka.

Testiranje statističke značajnosti koeficijenta linearne korelacije:[15]

Nulta hipoteza se pri tzv. dvosmernom testiranju statističke značajnosti koeficijenta linearne korelacije formuliše na sledeći način:[16]

$$H_0: \rho_{XY} = 0$$

(Rečima: koeficijent linearne korelacije u populaciji jednak je nuli).
Alternativna hipoteza se, pri dvosmernom testiranju, formuliše ovako:

$$H_1: \rho_{XY} \neq 0$$

Test statistik za testiranje H_0, u oznaci t, računa se na sledeći način:

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

Statistik t ima, ako je nulta hipoteza tačna, Studentovu T-distribuciju uzorkovanja sa n-2 stepeni slobode, pri čemu je n broj parova rezultata.
Na osnovu toga računamo „malo“ $p = P(|T| \geq t_{dobijeno}$ ako je H_0 tačna), pri čemu je $t_{dobijeno}$ konkretna vrednost t-statistika dobijena na uzorku . Ako je verovatnoća da t uzme na slučaj neku vrednost jednaku dobijenom t ili još ekstremniju, tj. ako je p manje od vrednosti za odbacivanje H_0 (najčešće 0.05), odbacujemo nultu hipotezu i zaključujemo da je t-statistik statistički značajan, tj. da je koeficijent linearne korelacije statistički značajan, te da u populaciji postoji linearna povezanost između varijabli. Ako je p veće od od vrednosti za odbacivanje H_0 (najčešće 0.05), ne odbacujemo H_0 i možemo da kažemo da nemamo razloga da tvrdimo postojanje linearne povezanosti varijabli u populaciji.[17]

Uticaj iznimaka (autlajera) na koeficijent linearne korelacije.

Iznimci ili autlajeri (engl. *Outlier*) su podaci koji su toliko udaljeni od glavnine podataka da se može posumnjati da su generisani nekim drugim mehanizmom a ne onim kojim je generisana glavnina podataka. Autlajeri najčešće nastaju pri uzorkovanju, prikupljanju i unosu podataka, premda pokatkad predstavljaju stvarne podatke pojedinih ispitanika koji su u datom pogledu pravi iznimci. Autlajere ćemo u slučaju korelacije najlakše uočiti na dijagramu raspršenja. Postoje dva tipa autlajera u ovom slučaju: iznimci pomerenosti i strukturni iznimci. Iznimci pomerenosti prate osnovni trend podataka na dijagramu raspršenja ali su veoma udaljeni od glavnine podataka. Strukturni iznimci narušavaju trend glavnine podataka.
Monte-Karlo simulacija koju je izveo Eledum (2017) pokazuje sledeće:

---

[15] Ovaj odeljak o testiranju statističke značajnosti koeficijenta korelacije korisnik treba da preskoči dok ne savlada sadržaj Glave X i da se tek potom vrati na njega.
[16] Da li je testiranje statističkih hipoteza dvosmerno (engl. *two-tailed*) ili jednosmerno (engl. *one-tailed*) zavisi od toga kako su formulisane nulta i alternativna hipoteza. Ako se u ovim hipotezama precizira smer korelacije (na primer $H_0: \rho_{XY} = 0$, $H_0: \rho_{XY} > 0$) onda se vrši tzv. jednosmerno testiranje. Jednosmerno testiranje se relativno retko koristi. Odluku o tome da li će testiranje biti dvosmerno ili jednosmerno donosimo pre prikupljanja podataka, tj. u fazi planiranja istraživanja.
[17] Za testiranje ove hipoteze može se koristiti i statistik F koji je jednak t^2, a pod H_0 ima Snidikorovu distribuciju **Uslov**uzorkovanja sa $df_1=1$ i $df_2 = n - 2$.

a) autlajeri mogu povećati ili smanjiti koeficijent korelacije zavisno od pozicije na kojoj se nalaze u odnosu na glavninu podataka. Premda Eledum ne koristi prethodno navedenu podelu iznimaka, možemo pretpostaviti da iznimci pomerenosti u principu povećavaju a strukturni iznimci smanjuju koeficijent korelacije;
b) efekat autlajera na vrednost koeficijenta korelacije je utoliko manji ukoliko je uzorak veći; ključni element od kojeg zavisi veličina efekta autlajera na vrednost koeficijenta korelacije je njegova pozicija u odnosu na glavninu podataka;
c) na autlajere su osetljiviji koeficijenti korelacije bliži jedinici nego oni niži.

Statistički uslovi za primenu koeficijenta linearne korelacije

Za primenu Brave-Pirsonovog koeficijenta korelacije poželjno je da budu ispunjeni sledeći uslovi (prilagođeno prema Havlicek, & Peterson (1977)):

1. Obe varijable su (barem teorijski) kontinuirane;
2. Varijable su u linearnom odnosu;
3. Parovi podataka na varijablama su statistički nezavisni (ne sme više od jednog para podataka poticati od iste jedinice posmatranja);
4. Zajednička distribucija dveju varijabli treba da bude bivarijaciona normalna kako bi se statistička značajnost koeficijenta linearne korelacije testirala standardnim t-testom. Premda je ovaj postupak dosta robustan na neispunjenost uslova o bivarijacionoj normalnosti, onda kada postoje ekstremna odstupanja najbolje je pre testiranja značajnosti koeficijenta linearne korelacije normalizovati raspodele varijabli inverznom normalnom transformacijom pomoću rangova (i to pretvaranjem u tzv. *rankit* skorove procedurom koja postoji u svim poznatim statističkim paketima (Bishara, & Hittner, 2012).

Testiranje povezanosti ili statističke nezavisnosti dveju kategoričkih varijabli

Statističku nezavisnost ili povezanost dveju kategoričkih varijabli analiziramo tako što za tabele kontingencije, poput tabele 7.2, računamo hi-kvadrat statistik i sa njim povezane mere asocijacije a potom testiramo njihovu statističku značajnost. Za tabele 2x2 (pa i za veće tabele) možemo koristiti količnik šansi i njegovu statističku značajnost.
U Tabeli 7.2 prikazane su u svakoj ćeliji empirijske zajedničke frekvencije dveju kategoričkih varijabli: pola ispitanika i da li trenutno ima romantičnog partnera (momka/devojku). Svaka od ovih varijabli ima po dve kategorije tako da se radi o tabeli 2x2. Da bismo izračunali hi-kvadrat statistik i testirali povezanost ili statističku nezavisnost ovih kategoričkih varijabli potrebno je za svaku ćeliju tabele kontingencije izračunati očekivanu frekvenciju. Očekivane frekvencije pokazuju kakav bi bio raspored frekvencija u tabeli kontingencije kada bi dve kategoričke varijable bile statistički nezavisne, tj. kada ne bi bile u vezi. Pretpostavka da su dve kategoričke varijable statistički nezavisne predstavlja u ovom slučaju nultu hipotezu. Očekivane frekvencije za ćeliju *jk*, u oznaci ϕ_{jk}, računaju se po sledećoj formuli:

$\phi_{jk} = (f_{j+} * f_{+k})/n$ gde su, kao što smo već uz Tabelu 7.2 napomenuli, f_{j+} i f_{+k} odgovarajuće marginalne frekvencije, tj. zbirovi kolona za red *j* u kojem je ćelija *jk*, odnosno zbirovi redova za kolonu *k* kojoj ćelija *jk* pripada.[18]

[18]Ako su kategoričke varijable nezavisne (a pod tom pretpostavkom računamo zajedničke očekivane frekvencije) onda je verovatnoća zajedničkog događanja neke kategorije jedne varijable i (istovremeno) neke

Razliku očekivane i empirijske frekvencije u datoj ćeliji zovemo i rezidualom.

Pirsonov hi-kvadrat statistik za testiranje nulte hipoteze računamo po sledećem obrascu:

$$\chi^2 = \sum_{k=1}^{c} \sum_{j=1}^{r} \frac{(f_{jk} - \phi_{jk})^2}{\phi_{jk}}$$

Ovaj statistik, ako su varijable statistički nezavisne, tj. ako je nulta hipoteza tačna, ima hi-kvadrat distribuciju uzorkovanja sa (r-1) (c-1) stepeni slobode.[19] Potom računamo „malo p" ili „značajnost", „malo" $p = P(\chi^2 \geq \chi^2_{\text{dobijeni}} \mid H_0 \text{ tačno})$, pri čemu je χ^2_{dobijeni} konkretna vrednost hi-kvadrata koju smo dobili na uzorku. Ako je verovatnoća p manja od α, tj. vrednosti za odbacivanje H_0 (uobičajeno 0.05), odbacujemo H_0, a ako je veća od te vrednosti ne odbacujemo H_0. Ako H_0 ne odbacimo ne možemo tvrditi da su varijable povezane, tj. pretpostavljamo da su statistički nezavisne. Ako H_0 odbacimo tada možemo doneti sledeće sinonimne statističke zaključke: H_0 odbacujemo ili hi-kvadrat statistik je statistički značajan. Na osnovu ovog statističkog zaključka zaključujemo da su varijable u vezi, tj. da nisu statistički nezavisne.

Za primer iz Tabele 7.2 Pirsonov hi-kvadrat statistik bismo izračunali na sledeći način:

$\chi^2 = (34 - 44.8)^2 / 44.8 + (87 - 76.2)^2 / 76.2 + (141 - 131.2)^2 / 131.2 + (211 - 221.8)^2 / 221.8 = 5.52$. „Malo" p u ovom slučaju za 1 stepen slobode je .019. Dakle, odbacili bismo nultu hipotezu ili bismo rekli hi-kvadrat je statistički značajan. Ove dve kategoričke varijable su u vezi, tj. nisu statistički nezavisne.

---

kategorije druge varijable jednaka proizvodima verovatnoće događanja date kategorije jedne varijable i verovatnoće događanja date kategorije druge varijable . Dakle:
$p_{jk} = p_{1j} * p_{2k} = (f_{1j}/n) * (f_{2k}/n)$,,
a $\phi_{jk} = p_{jk} * n = (f_{1j}/n) * (f_{2k}/n)*n = (f_{1j} * f_{2k})/n$.

[19] Broj stepeni slobode je ovoliki jer moramo oceniti (r - 1) verovatnoća za Q_1 i (c - 1) verovatnoća za Q_2.
Verovatnoće poslednje kategorije svake varijable determinisane su činjenicom da je $\sum_{j=1}^{r} p_j = 1$ i $\sum_{k=1}^{c} p_k = 1$.

Zapamtite:

- Opisani hi-kvadrat test podrazumeva nezavisnost opservacija, tj. nezavisnost frekvencija u ćelijama kontingencijske tabele. To znači da se jedan ispitanik sme pojaviti samo jednom u jednoj ćeliji tabele, a nikako ponovo u istoj ili drugoj kategoriji! U slučajevima tabela sa tzv. višestrukim odgovorima, više od jedne frekvencije u tabeli potiče od istog ispitanika. Na primer, isti ispitanik može na neko pitanje odabrati više od jednog odgovora i prema tome više frekvencija potiče od istog ispitanika. **Velika je greška** primeniti opisani hi-kvadrat test na takvim podacima, tj. na podacima sa **zavisnim** frekvencijama. U tim slučajevima potrebno je primeniti tzv. Rao-Skotov hi-kvadrat test (postupak se može naći u Decady & Thomas, 2000).

- Opisani hi-kvadrat test smemo koristiti ako su, u slučaju kada nominalne varijable imaju po **dve** kategorije sve **očekivane** frekvencije bar **5** ili, ako je reč o varijablama sa **više od dve** kategorije, ukoliko očekivane frekvencije nisu manje od 5 u **20% kategorija** i **nijedna** očekivana frekvencija **nije manja od 1**. U situacijama kada ovaj test ne treba koristititi, a koje smo upravo naveli, treba, ako je to smisleno, sažeti kategorije na manji broj ili:
  ako je n>40 a tabela 2x2 primeniti Yates-ovu korekciju[1]
  ako je n < 40 a tabela 2x2 primeniti Fisher-ov egzaktni test.

- Prikazani postupak testiranja hipoteze o nezavisnosti podrazumeva da je uzorak *E* izvučen slučajno iz populacije *P*, tj. da je pri tom fiksirano samo n, tj. veličina uzorka.

XII.3.Koeficijenti asocijacije dveju kategoričkih varijabli

U slučaju odbacivanja nulte hipoteze hi-kvadrat testom smisaono je upotrebiti neku od mera asocijacije kategoričkih varijabli zasnovanih na hi-kvadrat statistiku od kojih ćemo u ovoj knjizi prikazati samo tri:[20]

1. **Kramerov V koeficijent**, u oznaci V, računa se na sledeći način:

$$V = \sqrt{\chi^2 / \left[n(q-1)\right]} \text{ gde je } q = \min(r, c)$$

Gornja granica ovog koeficijenta je 1. Kramerov V-koeficijent, međutim, ima tendenciju da potcenjuje stvarnu jačinu povezanosti među varijablama u populaciji. Zgodna osobina ovog statistika je što se Kramerovi V-koeficijenti iz tabela različitih dimenzija mogu međusobno upoređivati.

2. **Fi-koeficijent**

U slučaju kada je tabela 2 x 2, pošto je tada q = 2, V-koeficijent je identičan fi-koeficijentu :

$$\phi = \sqrt{\chi^2 / n}$$

[20] Podrobniji prikaz ostalih koeficijenata asocijacije može se naći u Momirović (1988).

3. **Pirsonov koeficijent kontingencije, C-koeficijent**:

$$C = \sqrt{\chi^2 / (n + \chi^2)}$$

C-koeficijent, čak ni kada je veza među varijablama savršena, ne dostiže uvek vrednost 1, već maksimalna vrednost koju može dostići zavisi od dimenzija tabele:

$$C_{max} = \sqrt{(q-1)/q} \quad \text{gde je } q = \min(r,c)$$

Jasno je stoga da se C-koeficijenti iz tabela kontingencije koje su nejednakih dimenzija ne mogu direktno porediti.

Uočimo da su svi prikazani koeficijenti asocijacije dveju kategoričkih varijabli zapravo matematičke funkcije hi-kvadrat statistika. Stoga zaključak o statističkoj značajnosti hi-kvadrat statistika automatski važi i za Kramerov V-koeficijent, fi-koeficijent i C koeficijent.

Količnik šansi

Kao mera povezanosti dveju kategoričkih varijabli, posebno onih dihotomnog tipa, često se koristi i količnik šansi (engl. *Odds Ratio)* ili količnik unakrsnih proizvoda. Šanse smo detaljno objasnili u Glavi **III**. Ponovićemo ovde samo ukratko: šanse predstavljaju količnik verovatnoće da se neki događaj desi i verovatnoće da se taj događaj desi. Na primer, ako je verovatnoća da se neki događaj desi jednaka 0.8, šanse tog događaja su 0.8 / 0.2, tj. 4. Ako su verovatnoće da se događaj desi i da se ne desi jednake 0.5, šanse tog događaja jednake su 1. Količnik šansi, u oznaci OR dobija se za tabele 2x2 na sledeći način:

$$OR = \frac{\frac{f_{11}}{f_{12}}}{\frac{f_{21}}{f_{22}}} = \frac{f_{11} * f_{22}}{f_{12} * f_{21}}$$

Pri tome su f_{11}, f_{12} , f_{22} i f_{21} empirijski dobijene frekvencije u ćelijama tabele. Iz desnog oblika obrasca postaje jasno zašto se količnik šansi u ovom slučaju zove i količnik unakrsnih proizvoda – dobija se deljenjem proizvoda frekvencija, pri čemu su u istom proizvodu ćelije koje su dijagonalno jedna naspram druge u tabeli. Ukoliko su kategoričke varijable (svaka sa po dve kategorije) statistički nezavisne količnik šansi jednak je 1. Inače, ovaj količnik varira od 0 do ∞. Međutim, on nije simetričan oko 1 pa može biti različit za tabele u kojima je prisutan isti stepen asocijacije ali drugačijeg smera. Da bi bio simetričan oko 1 dovoljno je uzeti logaritam od OR. Log OR varira od -∞ do ∞, a kada su varijable nezavisne jednak je 0. Količnik šansi se može generalizovati i na tabele kontingencije veće od 2x2.
Statistička značajnost količnika šansi određuje se pravljenjem intervala poverenja. Ukoliko interval poverenja za količnik šansi ne obuhvata 1 varijable su u vezi u populaciji, a ako obuhvata 1 ne možemo to da tvrdimo. Sam postupak formiranja ovog intervala poverenja nećemo prikazivati u ovoj knjizi.
Umesto količnika šansi, u medicini se ponekad koristi *količnik rizika* ili *relativni rizik* (u oznaci RR) koji predstavlja količnik verovatnoće nekog događaja u jednoj grupi i

verovatnoće tog istog događaja u drugoj grupi. Npr. ako je verovatnoća oboljevanja od neke bolesti u jednoj grupi 0.60, a u drugoj grupi 0.20, relativni rizik jednak je 0.60 / 0.20, tj. 3. To praktično znači da je verovatnoća oboljevanja od te bolesti u prvoj grupi 3 puta veća.
Kramerov V-koeficijent (fi-koeficijent) za primer iz Tabele 2.2. iznosi .108, C-koeficijent je .107, šanse za devojke da imaju partnera141 / 211 = 0.66, šanse za momke da imaju partnera 34 / 87 = 0.39, a količnik šansi za devojke u odnosu na momke 0.66 / 0.39 = 1.69. Dakle, u ovom uzrastu šanse da imaju partnera veće su za devojke 1, 69 puta nego za momke.

Reference

Bishara, A. J. & Hittner, J.B. (2012). Testing the Significance of a Correlation With Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. *Psychological Methods*, *17*(3), 399–417.

Branch, W. (1990). On interpreting correlation coefficients. *American Psychologist*, *45*(2), 296.

Decady, Y. J. & Thomas, D. R. (2000). A simple test of association for contingency tables with multiple column responses. *Biometrics*, *56*, 893−896.

Eledum, H., Y. (2017). A Monte Carlo Study of the Effects of Variability and Outliers on the Linear Correlation Coefficient. *Journal of Modern Applied Statistical Methods*, *16*(2), 231-255. doi: 10.22237/jmasm/1509495180

Falk, R., & Well, A. D. (1997). Many Faces of the Correlation Coefficient. Journal of Statistics Education, *5*(3), 1–15.

Farnsworth, D. L.(2014). Identical Tweens Raised Apart. *Teaching Statistics*, *37*(1), 1–6.
Havlicek, L. L., & Peterson, N. L. (1977). Effect of the Violation of Assumptions Upon Significance Levels of the Pearson *r*, *Psychological Bulletin*, *84*(2), 373–377

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.