Pearsonov korelacijski koeficijent

Tabak, Ana

Master's thesis / Diplomski rad

2018

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:544508

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of School of Applied Mathematics and Computer Science

Sveučilište J.J.Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Ana Tabak

Pearsonov korelacijski koeficijent

Diplomski rad

Osijek, 2018.

Sveučilište J.J.Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Ana Tabak

Pearsonov korelacijski koeficijent

Diplomski rad

Voditelj: prof.dr.sc. Mirta Benšić

Lektorirala: Ivana Rašić, prof.

Osijek, 2018.

Sadržaj

1 Uvod **1**

2 Povijesne činjenice **2**

3 Koeficijent korelacije **3**
3.1 Osnovni pojmovi 3
3.2 Korelacijski koeficijent i njegova svojstva 5
3.3 Matrica kovarijanci i korelacijska matrica 7

4 Procjena očekivanja i matrice kovarijanci normalnog slučajnog vektora **9**
4.1 Pearsonov korelacijski koeficijent bivarijantnog uzorka normalne distribucije 14
4.1.1 Populacijski korelacijski koeficijent jednak je 0 14
4.1.2 Populacijski korelacijski koeficijent različit je od nule 16
4.1.3 Asimptotska distribucija uzoračkog koeficijenta korelacije 18

5 Literatura **20**

6 Sažetak **21**

7 Summary **22**

8 Životopis **23**

1 Uvod

Koeficijent korelacije se u statistici koristi kao mjera povezanosti dviju varijabli. Povezanost znači da je vrijednost jedne varijable moguće, s određenom vjerojatnošću, predvidjeti na osnovi saznanja o vrijednosti druge varijable. Postoji više mjera povezanosti koje se koriste u različitim slučajevima. U praksi se prilikom rada s linearnim modelima najčešće koristi Pearsonov koeficijent korelacije, kao mjera jakosti i smjera linearne statističke povezanosti dviju varijabli. Pearsonov koeficijent korelacije koristi se prvenstveno u slučajevima kada između varijabli promatranog modela postoji linearna povezanost i varijable imaju normalnu distribuciju. Vrijednost Pearsonovog koeficijenta korelacije kreće se od $+1$ do -1.

U prvom dijelu rada definirat ćemo neke osnovne pojmove kao što su: ishodišni i centralni moment, kvarijanca, korelacijski keoficijent, matrica kovarijanci i korelacijska matrica. Također, uvodimo osnovna svojstva korelacijskog koeficijenta.
Glavni dio ovoga diplomskig rada je Pearsonov korelacijski koeficijent kao procjenitelj za populacijski korelacijski koeficijent. Na temelju prepostavke o normalnoj distribuciji populacije, pokazujemo da je Pearsonov korelacijski koeficijent procjenitelj populacijskog korelacijskog koeficijenta dobiven primjenom metode maksimalne vjerodostojnosti. Taođer kreiramo statističke testove za testiranje pretpostavke o iznosu populacijskog koeficijenta korelacije i diskutiramo o njihovoj robusnosti na odstupanje od normalnosti populacijske distribucije.

2 Povijesne činjenice

Karl Pearson bio je važna osoba u ranom razvoju statistike. Utemeljio je Odjel za primijenjenu statistiku na Sveučilištu u Londonu 1911. godine. Bio je to prvi odjel sveučilišne statistike na svijetu. Sadašnji odjeli Državne znanosti i računarstva, kao i grupe za genetiku te biometrija u biologiji i dio antropologije, dio su njegove ostavštine UCL-u (University College London).

Rođen je u Londonu 27. ožujka 1857. Pohađao je privatnu školu, nakon čega upisuje studij matematike na King's College Cambridge. Dvije godine provodi na sveučilištima u Berlinu i Heidelbergu, studirajući srednjovjekovnu i njemačku književnosti iz 16. stoljeća. Na tom je polju bio dovoljno obrazovan, stoga mu je ponuđen njemački odjel na Sveučilištu Cambridge.

Njegov sljedeći korak u karijeri bio je u Lincoln's Inn[1], gdje je proučavao pravo do 1881. godine, iako se njime nikada nije bavio. Nakon toga, vratio se na matematiku, kao zamjenik profesora matematike na King's Collegeu u Londonu 1881. godine, zatim kao profesor na Sveučilištu u Londonu 1883. godine. Godine 1884. postavljen je na katedru Goldfieldsove primijenjene matematike i mehanike na Sveučilištu u Londonu. 1891. započinje rad na profesorskom fakultetu na Gresham Collegeu gdje upoznaje W.F.R. Weldona[2], zoologa koji je imao neke zanimljive probleme koji zahtijevaju kvantitativna rješenja. Suradnja u biometrijskoj i evolucijskoj teoriji bila je plodonosna i trajala je sve do Weldonove smrti 1906. godine. Weldon upoznaje Pearsona sa Galtonom[3]. Galton je bio zainteresiran za evoluciju i grane evolucije kao što je nasljeđe, no više ga je zanimala evolucija statistike. Pokrenuo je ideju o koeficijentu korelacije te je vjerovao kako nasljednosti mogu biti znanstveno utemeljene jedino uvođenjem novih statističkih koncepata kao što su regresija i korelacija.

Nakon Galtonove smrti, Pearson je formirao Odjel za primijenjenu statistiku u kojem je uključio biometrijska i Galtonova istraživanja. Ostao je u odjelu do umirovljenja 1933. godine, a nastavio je raditi sve do svoje smrti 1936. godine.

Pearson se oženio Mariom Sharpe 1890. godine. Imali 2 kćeri i jednog sina. Sin Egon Sharpe Pearson naslijedio ga je kao voditelj Odjela za primijenjenu statistiku na Sveučilišnom fakultetu.

Osim svog profesionalnog života, Pearson je bio aktivan kao istaknuti slobodni mislilac i socijalist.

[1] Odvjetnička komora; Priznaje se kao jedno od najprestižnijih svjetskih stručnih tijela sudaca i odvjetnika

[2] Walter Frank Raphael Weldon; engleski evolucijski biolog i utemeljitelj biometrije

[3] Francis Galton-engleski viktorijanski statističar, polimat, sociolog, psiholog,antropolog, eugenik, tropski istraživač, zemljopisac, izumitelj, meteorolog, proto-genetičar i psihometrij

3 Koeficijent korelacije

3.1 Osnovni pojmovi

U ovo diplomskom radu, usredotočit ćemo se najviše na problem procjene koeficijenta korelacije. No, da bismo mogli govoriti o njegovoj procjeni, prvo ćemo ju definirati te navesti neka važna svojstva koeficijenta korelacije.

Koeficijent korelacije jedna je od numeričkih karakteristika slučajnog vektora, stoga prvo navodimo definicije momenata slučajnog vektora koje će nam poslužiti u definiciji koeficijenta korelacije, ali i u daljnjim dijelovima ovoga rada.

Definicija 3.1.1. Neka je (X, Y) diskretan ili neprekidan dvodimenzionalni slučajni vektor. Očekivanje

$$E(X^k Y^l), k, l \in N_0 \tag{3.1}$$

slučajne varijable $X^k Y^l$(ako postoji) nazivamo **ishodišni moment** reda (k, l) slučajnog vektora (X, Y) i pišemo

$$\mu_{kl} = E(X^k Y^l).$$

Definicija 3.1.2. Očekivanje $E((X - EX)^k (Y - EY)^l)$ (ako postoji) nazivamo **centralni moment** reda (k, l) slučajnog vektora (X, Y) i pišemo

$$m_{kl} = E((X - EX)^k (Y - EY)^l). \tag{3.2}$$

Možemo primjetiti da momenti reda $(0, 1), (0, 2), (1, 0), (2, 0)$ daju očekivanje i varijancu komponenata slučajnog vektora. Vrijedi da je:

$$\mu_{10} = EX,$$
$$\mu_{01} = EY,$$
$$m_{20} = E(X - EX)^2 = VarX$$
$$m_{02} = E(Y - EY)^2 = VarY.$$

Centralni moment reda $(1, 1)$ nazivamo korelacijski moment ili kovarijanca dvodimenzionalnoga slučajnog vektora.

Definicija 3.1.3. Kovarijanca dvodimenzionalnoga slučajnog vektora (X, Y) definirana je izrazom

$$\mathrm{Cov}\,(X, Y) = E((X - EX)(Y - EY)). \tag{3.3}$$

Kovarijanca se, kao numerička karakteristika slučajnog vektora, može se povezati s pojmom nezavisnosti slučajnih varijabli, o čemu govori Teorem 3.1.1.

Teorem 3.1.1. Neka je (X, Y) neprekidan ili diskretan dvodimenzionalni slučajni vektor za koji postoje EX i EY. Ako su slučajne varijable X i Y nezavisne, onda je $\mathrm{Cov}\,(X, Y) = 0$.

Dokaz:
Pretpostavimo da su slučajne varijable X i Y nezavisne.
Zbog nezavisnosti je

$$E(XY) = EXEY.$$

Dakle, vrijedi

$$\mathrm{Cov}\,(X, Y) = E(XY) - EXEY.$$

Q.E.D.
(vidi npr. [2].)
Kao posljedica tog teorema proizlazi zaključak: ako dvodimenzionalni slučajni vektor ima kovarijancu različitu od nule, onda su njegove komponente nužno zavisne.

Definicija 3.1.4. Neka je (X, Y) slučajni vektor za koji je $\mathrm{Cov}\,(X, Y) = 0$. Tada kažemo da su njegove komponente X i Y nekorelirane.

Valja napomenuti da nekoreliranost slučajnih varijabli nije garancija za nezavisnost. O tome svjedoči Primjer 3.1.1.

Primjer 3.1.1. Neka je (X, Y) slučajni vektor za koji je $Y = X^2$ a X neprekidna slučajna varijabla s parnom funkcijom gustoće $f(x)$. Očigledno slučajne varijable X i Y nisu nezavisne. Međutim, $\mathrm{Cov}\,(X, Y) = 0$.
Zaista: Kako imamo slučajni vektor (X, X^2) i znamo da je $E(X) = 0$, vrijedi

$$\mathrm{Cov}\,(X, X^2) = E(XX^2) - E(X)E(X^2) = E(X^3). \tag{3.4}$$

Znamo da je $f(x)$ parna funkcija gustoće, a iz toga slijedi da je $x^3 f(x)$ neparna funkcija, stoga vrijedi

$$E(X^3) = \int x^3 f(x)dx = 0. \tag{3.5}$$

3.2 Korelacijski koeficijent i njegova svojstva

Definicija kovarijance pokazuje nam da važnu ulogu u njezinom iznosu ima odstupanje pojedine komponente slučajnog vektora od njezinog očekivanja (devijacija).
Kako bismo smanjili utjecaj devijacije na numeričku karakteristiku kojom želimo dobiti nove informacije o slučajnom vektoru, od velikog je interesa proučavanje kovarijance standardiziranog oblika slučajnog vektora (X, Y).
Ukoliko X i Y imaju varijance $\sigma_X^2 \neq 0$ i $\sigma_Y^2 \neq 0$, možemo ih standardizirati. Ako označimo $\mu_X = EX, \mu_Y = EY$, postupak standardizacije daje vektor

$$(X_s, Y_s) = (\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y})$$

čija je kovarijanca

$$\mathrm{Cov}\,(X_s, Y_s) = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_X \sigma_Y}.$$

Tako nastaje koeficijent korelacije slučajnog vektora (X, Y).

Definicija 3.2.1. Neka su X i Y slučajne varijable sa standardnim devijacijama σ_X, σ_Y. Koeficijent korelacije od X i Y, u oznaci $\mathrm{Corr}\,(X, Y)$ ili ρ_{XY} definiran je na sljedeći način:

$$\rho_{X,Y} = \mathrm{Corr}\,(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_X \sigma_Y}. \tag{3.6}$$

Sljedeći teorem daje dovoljne uvjete za postojanje kovarijance slučajnog vektora i jednu korisnu nejednakost.

Teorem 3.2.1. Neka je (X, Y) slučajni vektor za koji je $0 < E(X^2) < \infty$ i $0 < E(Y^2) < \infty$. Tada postoji kovarijanca i vrijede nejednakosti:

$$|\rho_{X,Y}| \leq 1, \tag{3.7}$$

$$(E(XY))^2 \leq E(X^2)E(Y^2). \tag{3.8}$$

(Za dokaz teorema vidi npr. [2]. i [6].)

Da je koeficijent korelacije korisna numerička karakteristika za opisivanje veze među komponentama slučajnog vektora, vidljivo je i iz sljedećeg teorema koji pokazuje da se linearna veza među komponentama može uočiti na temelju vrijednosti koeficijenta korelacije.

Teorem 3.2.2. Neka je (X, Y) slučajni vektor za koji je $0 < \sigma_X < \infty$ i $0 < \sigma_Y < \infty$. Veza je među komponentama linearna, tj. postoje realni brojevi a $(a \neq 0)$ i b takvi da je

$$Y = aX + b \tag{3.9}$$

onda i samo onda ako je $|\rho_{X,Y}| = 1$. Pritom je koeficijent korelacije 1 ako je $a > 0$, odnosno -1 ako je $a < 0$.
(Za dokaz teorema vidi npr. [2]. i [6].)

Navest ćemo još neka bitna svojstva koeficijenta korelacije koja slijede izravno iz njegove definicije:

- $\rho_{X,Y} \in [-1, 1]$
- ako su X i Y nezavisne slučajne varijable tada je $\rho_{X,Y} = 0$
- ako je $\rho_{X,Y} = 0$, kažemo da su slučajne varijable X, Y nekorelirane.

3.3 Matrica kovarijanci i korelacijska matrica

Nerijetko je korisno i slučajni vektor zapisivati u matričnom obliku, kao kod vektora realnih brojeva, tj. slučajni vektor (X, Y) zapisujemo kao $[X, Y]'$. Nadalje ćemo tu oznaku koristit za definiranje očekivanja slučajnog vektora.
Ako za slučajni vektor $\mathbf{Z} = [X, Y]'$ postoje EX i EY , zapisivat ćemo ih u vektorskom obliku kao $[EX, EY]'$ i zvati očekivanje slučajnog vektora $\mathbf{Z} = [X, Y]'$. Varijance i kovarijancu također zapisujemo matrično. Označimo li $E\mathbf{Z} = [EX, EY]'$ tada je

$$E(\mathbf{Z}-E\mathbf{Z})(\mathbf{Z}-E\mathbf{Z})' = \begin{bmatrix} \text{E(X-EX)}^2 & \text{E[(X-EX)(Y-EZ)]} \\ \text{E[(X-EX)(Y-EZ)]} & \text{E(Y-EY)}^2 \end{bmatrix} = \begin{bmatrix} \text{VarX} & \text{Cov(X,Y)} \\ \text{Cov(X,Y)} & \text{VarY} \end{bmatrix}. \tag{3.10}$$

Matricu $E(\mathbf{Z} - E\mathbf{Z})(\mathbf{Z} - E\mathbf{Z})'$ zovemo matrica kovarijanci slučajnog vektora (X, Y).
Za matricu kovarijanci vrijedi da je simetrična i pozitivno semidefinitna.

Zaista, simetričnost proizlazi iz činjenice da je $\text{Cov}\,(X, Y) = \text{Cov}\,(Y, X)$.
Pozitivnu semidefinitnost ćemo dokazivati korištenjem Sylvesterovog kriterija (vidi npr. [4].).

$$det \begin{bmatrix} \text{VarX} & \text{Cov(X,Y)} \\ \text{Cov(X,Y)} & \text{VarY} \end{bmatrix} = \text{Var } X \text{ Var } Y - [\text{Cov}\,(X, Y)]^2 =$$
$$= \text{Var } X \text{ Var } Y - [\text{Corr}\,(X, Y)\sqrt{\text{Var } X \text{ Var } Y}]^2 =$$
$$= \text{Var } X \text{ Var } Y - [\text{Corr}\,(X, Y)]^2 \text{ Var } X \text{ Var } Y =$$
$$= \text{Var } X \text{ Var } Y(1 - [\text{Corr}\,(X, Y)]^2).$$

Obzirom da je $\text{Corr}\,(X, Y) \in [-1, 1]$ te da je $\text{Var } X > 0$,$\text{Var } Y > 0$ slijedi pozitivna semidefinitnost.

Neka je $\mathbf{Z}_s$ standardizirana varijanta slučajnog vektora $\mathbf{Z}$, tj.

$$\mathbf{Z}_s = [\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}]'.$$

Njegovu matricu kovarijanci zovemo korelacijska matrica slučajnog vektora $\mathbf{Z} = [X, Y]'$ i označavamo $\text{Corr}\,(X, Y)$.
Vrijedi

$$\text{Corr}\,(X, Y) = \begin{bmatrix} 1 & \rho_{X,Y} \\ \rho_{X,Y} & 1 \end{bmatrix}. \tag{3.11}$$

Očekivanje slučajnog vektora, matrica kovarijanci i korelacijska matrica mogu se na analogan način definirati za n-dimenzionalan slučajni vektor $\mathbf{Z} = [X_1, ..., X_n]'$.

Ako za svaku komponentu slučajnog vektora $\mathbf{Z} = [X_1, ..., X_n]$ postoji matematičko očekivanje, onda postoji i matematičko očekivanje slučajnog vektora $\mathbf{Z}$ koje se definira kao vektor očekivanja njegovih komponenti

$$E\mathbf{Z} = \begin{bmatrix} \text{EX}_1 \\ . \\ . \\ . \\ \text{EX}_n \end{bmatrix}. \tag{3.12}$$

Ukoliko postoje EX_i^2, za sve $i = 1, ..., n$ matrica kovarijanci slučajnog vektora $\mathbf{Z} = [X_1, ..., X_n]$ se definira izrazom

$$\mathrm{Cov}\,(\mathbf{Z}) = E[(\mathbf{Z} - E\mathbf{Z})(\mathbf{Z} - E\mathbf{Z})'] = \begin{bmatrix} \mathrm{Cov}(\mathrm{X}_1, X_1) & \mathrm{Cov}(\mathrm{X}_1, X_2) & ... & \mathrm{Cov}(\mathrm{X}_1, X_n) \\ \mathrm{Cov}(\mathrm{X}_2, X_1) & \mathrm{Cov}(\mathrm{X}_2, X_2) & ... & \mathrm{Cov}(\mathrm{X}_2, X_n) \\ . & & & \\ . & & & \\ . & & & \\ \mathrm{Cov}(\mathrm{X}_n, X_1) & \mathrm{Cov}(\mathrm{X}_n, X_2) & ... & \mathrm{Cov}(\mathrm{X}_n, X_n) \end{bmatrix}. \tag{3.13}$$

Korelacijska matrica slučajnog vektora $\mathbf{Z} = [X_1, ..., X_n]$ definira se kao

$$\mathrm{Corr}\,(\mathbf{Z}) = \mathrm{Corr}\,(X_1, ..., X_n) = \begin{bmatrix} 1 & \mathrm{Corr}\,(X_1, X_2)... & \mathrm{Corr}\,(X_1, X_n) \\ \mathrm{Corr}\,(X_2, X_1) & 1\ ... & \mathrm{Corr}\,(X_2, X_n) \\ . & & \\ . & & \\ . & & \\ \mathrm{Corr}\,(X_n, X_1) & \mathrm{Corr}\,(X_n, X_2)... & 1 \end{bmatrix} \tag{3.14}$$

gdje je $\mathrm{Corr}\,(X_i, X_j) = \frac{\mathrm{Cov}\,(X_i, X_j)}{\sqrt{\mathrm{Var}\,(X_i)\,\mathrm{Var}\,(X_j)}}, i, j = 1, ..., n.$

Kao u dvodimenzionalnom slučaju, vidljivo je da su matrice $\mathrm{Cov}\,(\mathbf{Z})$ i $\mathrm{Corr}\,(\mathbf{Z})$ pozitivno semidefinitne i simetrične.

4 Procjena očekivanja i matrice kovarijanci normalnog slučajnog vektora

Neka je $\mathbf{X} = (X_1, ..., X_p)$ p-dimenzionalni slučajni vektor s matricom očekivanja $\boldsymbol{\mu}$ i matricom kovarijanci $\boldsymbol{\Sigma}$. Dakle imamo

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_p \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & ...\rho_{1p}\sigma_1\sigma_p \\ . & & \\ . & & \\ . & & \\ \rho_{p1}\sigma_p\sigma_1 & ... & \sigma_p^2 \end{bmatrix}$$

gdje je $\mu_j = EX_j$, $\sigma_j^2 = \text{Var } X_j$, $\rho_{jk} = \text{Corr}\,(X_j, X_k)$, $j, k = 1, ..., p$.
S obzirom da se radi o normalnom slučajnom vektoru, pretpostavimo da je $\boldsymbol{\Sigma}$ pozitivno definitna matrica. Zaključivanje o parametrima $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ provodimo na temelju n-dimenzionalnog jednostavnog slučajnog uzorka iz distribucije slučajnog vektora $\mathbf{X}$. Za danih n realizacija $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$ slučajnog vektora $\mathbf{X}$ funkcija vjerodostojnosti za naš uzorak dana je formulom

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pn}(det(\boldsymbol{\Sigma}))^{\frac{1}{2}n}} exp[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})]. \tag{4.1}$$

Metodom maksimalne vjerodostojnosti (vidi npr [5].) tražimo $\max L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ po p-parametara za matricu očekivanja $\boldsymbol{\mu}$ i $(p + \frac{p(p-1)}{2}) = \frac{p(p+1)}{2}$ parametara za matricu kovarijanci $\boldsymbol{\Sigma}$. Da bismo odredili procjenitelja za $\boldsymbol{\mu}$ i za $\boldsymbol{\Sigma}$, prvo ćemo naći ML procjenitelja očekivanja $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}$. L je funkcija od $\boldsymbol{\mu}$ i $\boldsymbol{\Psi}$.
Logaritam funkcije vjerodostojnosti iznosi

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Psi}) = -\frac{1}{2}pn\log(2\pi) + \frac{1}{2}n\log det(\boldsymbol{\Psi}) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{\mu})'\boldsymbol{\Psi}(\boldsymbol{x_i} - \boldsymbol{\mu}). \tag{4.2}$$

ML procjenitelji od $\boldsymbol{\mu}$ i $\boldsymbol{\Psi}$ su $\hat{\boldsymbol{\mu}}$ i poz. def. matrica $\hat{\boldsymbol{\Psi}}$ koja maksimizira $\log L(\boldsymbol{\mu}, \boldsymbol{\Psi})$.

Neka je aritmetička sredina uzorka

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{1i} \\ . \\ . \\ . \\ \frac{1}{n}\sum_{i=1}^{n} x_{pi} \end{pmatrix} = \begin{pmatrix} \overline{x}_1 \\ . \\ . \\ . \\ \overline{x}_p \end{pmatrix}, \tag{4.3}$$

i neka je matrica A definirana kao

$$A = [A_{jk}]_{j,k=1,...,p} = \sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}})' = [\sum_{i=1}^{n}(x_{ji} - \overline{x}_j)(x_{ki} - \overline{x}_k)], j, k = 1, ..., p. \tag{4.4}$$

Za određivanje procjenitelja $\hat{\boldsymbol{\mu}}$ i $\hat{\boldsymbol{\Psi}}$ koristimo sljedeću lemu.

Lema 4.1. Neka je $\overline{\boldsymbol{x}}$ aritmetička sredina uzorka. Tada, za bilo koji vektor $\boldsymbol{b}$, vrijedi

$$\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{b})(\boldsymbol{x_i} - \boldsymbol{b})' = \sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}})' + n(\overline{\boldsymbol{x}} - \boldsymbol{b})(\overline{\boldsymbol{x}} - \boldsymbol{b})'. \tag{4.5}$$

Dokaz:
$\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{b})(\boldsymbol{x_i}-\boldsymbol{b})' = \sum_{i=1}^{n}[(\boldsymbol{x_i}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{b})][(\boldsymbol{x_i}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{b})]' =$

$$= \sum_{i=1}^{n}[(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})' + (\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}-\boldsymbol{b})' + (\overline{\boldsymbol{x}}-\boldsymbol{b})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})' + (\overline{\boldsymbol{x}}-\boldsymbol{b})(\overline{\boldsymbol{x}}-\boldsymbol{b})'] =$$

$$= \sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})' + [\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})](\overline{\boldsymbol{x}}-\boldsymbol{b})' + (\overline{\boldsymbol{x}}-\boldsymbol{b})\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})' + n(\overline{\boldsymbol{x}}-\boldsymbol{b})(\overline{\boldsymbol{x}}-\boldsymbol{b})'.$$

Izrazi $[\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})](\overline{\boldsymbol{x}}-\boldsymbol{b})'$ i $(\overline{\boldsymbol{x}}-\boldsymbol{b})\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})'$ su 0 jer je $\sum(x_i-\overline{x}) = \sum x_i - n\overline{x} = 0$. Q.E.D.

Ako stavimo da je $\boldsymbol{b} = \boldsymbol{\mu}$, imamo

$$\begin{aligned}\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})(\boldsymbol{x_i}-\boldsymbol{\mu})' &= \sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})' + n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})' = \\ &= A + n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'. \end{aligned} \tag{4.6}$$

Korištenjem dobivenih rezultata i svojstva traga matrice (invarijantnost na cikličke permutacije) imamo

$$\begin{aligned}\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Psi}(\boldsymbol{x_i}-\boldsymbol{\mu}) &= tr\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Psi}(\boldsymbol{x_i}-\boldsymbol{\mu}) = \\ &= tr\sum_{i=1}^{n}\boldsymbol{\Psi}(\boldsymbol{x_i}-\boldsymbol{\mu})(\boldsymbol{x_i}-\boldsymbol{\mu})' = \\ &= tr\boldsymbol{\Psi}A + tr\boldsymbol{\Psi}n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})' = \\ &= tr\boldsymbol{\Psi}A + n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Psi}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'. \end{aligned} \tag{4.7}$$

Sada izraz (4.2) možemo napisati kao

$$\log L(\boldsymbol{\mu},\boldsymbol{\Psi}) = -\frac{1}{2}pn\log(2\pi) + \frac{1}{2}n\log(det(\boldsymbol{\Psi})) - \frac{1}{2}tr\boldsymbol{\Psi}A - \frac{1}{2}n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Psi}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}). \tag{4.8}$$

Kako je $\boldsymbol{\Psi}$ pozitivno semidefinitna, $n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Psi}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \geq 0$ i 0 je kada je $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$.

Napomena 4.1. Metoda maksimalne vjerodostojnosti ima vrlo važno svojstvo invarijantnosti. Neka je $\phi = g(\boldsymbol{\theta})$, gdje je g bijektivna funkcija, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$,$k \in N$ te neka je $\widehat{\boldsymbol{\theta}}$ ML-procjenitelj za $\boldsymbol{\theta}$. Tada je $g(\widehat{\boldsymbol{\theta}})$ ML-procjenitelj za $g(\boldsymbol{\theta}) = \phi$. (Vidi npr. [3].)

Teorem 4.1. Ako je $\boldsymbol{X_1}, ..., \boldsymbol{X_n}$ j.sl. uzorak iz $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gdje je $p < n$, ML procjenitelji od $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ su $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} = (1/n)\sum_{i=1}^{n}\boldsymbol{x_i}$ i $\widehat{\boldsymbol{\Sigma}} = (1/n)\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})'$.

Dokaz:
Neka je dana funkcija vjerodostojnosti i njen logaritam

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pn}(det(\boldsymbol{\Sigma}))^{\frac{1}{2}n}}exp[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu})], \tag{4.9}$$

$$\log L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}pn\log(2\pi) + \frac{1}{2}n\log det(\boldsymbol{\Sigma}^{-1}) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu}). \tag{4.10}$$

Računajući parcijalnu derivaciju po $\boldsymbol{\mu}$ dobivamo

$$\frac{\partial \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu}). \tag{4.11}$$

Izjednačujemo sa nulom

$$\sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu}) = 0. \tag{4.12}$$

Inverzna matrica matrice $\boldsymbol{\Sigma}$ je $\boldsymbol{\Sigma}^{-1}$ pa vrijedi da je

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = I$$

gdje je I jedinična matrica. Sređivanjem izraza (4.12) dobivamo

$$\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \boldsymbol{x_i} - n\boldsymbol{\mu} = 0. \tag{4.13}$$

Rješavanjem jednadžbe i korištenjem Napomene 4.1. slijedi

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}. \tag{4.14}$$

U nastavku dokaza pretpostavimo da je $p = 2$, tj. neka je (X, Y) dvodimenzionalni normalni sl.vektor a podaci $(x_1, y_1), ..., (x_n, y_n)$ njegove realizacije. Tada imamo

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

i

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

te

$$\boldsymbol{\Sigma^{-1}} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & \text{-}\rho\sigma_1\sigma_2 \\ \text{-}\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2(1-\rho^2)} & \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} \\ \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} & \frac{1}{\sigma_2^2(1-\rho^2)} \end{bmatrix} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{bmatrix} = \boldsymbol{\Psi}$$

gdje je $\det(\Sigma^{-1}) = \det(\boldsymbol{\Psi}) = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}$.

Sada logaritamska funkcija vjerodostojnosti izgleda ovako:

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = konst. + \frac{n}{2}\log\det(\boldsymbol{\Psi}) - \frac{1}{2}\sum_{i=1}^{n}\left[(x_i - \mu_1, y_i - \mu_2)' \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{bmatrix} \begin{bmatrix} \mathrm{x}_i - \mu_1 \\ \mathrm{y}_i - \mu_2 \end{bmatrix}\right] =$$

$$= konst. + \frac{n}{2}\log(\psi_{11}\psi_{22} - \psi_{12}^2) - \frac{1}{2}\sum_{i=1}^{n}\left[(x_i - \mu_1, y_i - \mu_2) \begin{bmatrix} \psi_{11}(x_i - \mu_1) + \psi_{12}(y_i - \mu_2) \\ \psi_{12}(x_i - \mu_1) + \psi_{22}(y_i - \mu_2) \end{bmatrix}\right] =$$

$$= konst. + \frac{n}{2}\log(\psi_{11}\psi_{22} - \psi_{12}^2) - \frac{1}{2}\sum_{i=1}^{n}[\psi_{11}(x_i - \mu_1)^2 + \psi_{12}(y_i - \mu_2)(x_i - \mu_1) +$$

$$+\psi_{12}(y_i - \mu_2)(x_i - \mu_1) + \psi_{22}(y_i - \mu_2)^2].$$

Slijedi

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = konst. + \frac{n}{2}\log(\psi_{11}\psi_{22} - \psi_{12}^2) - \frac{1}{2}\sum_{i=1}^{n}[\psi_{11}(x_i - \mu_1)^2 + 2\psi_{12}(x_i - \mu_1)(y_i - \mu_2) + \psi_{22}(y_i - \mu_2)^2].$$

ML procjenitelje za μ_1 i μ_2 smo već izračunali:

$$\widehat{\mu_1} = \overline{x} \qquad i \qquad \widehat{\mu_2} = \overline{y}.$$

Procjenitelje za parametre matrice $\boldsymbol{\Sigma}$ dobivamo deriviranjem

$$l(\boldsymbol{\Psi}) = konst. + \frac{n}{2}\log(\psi_{11}\psi_{22} - \psi_{12}^2) - \frac{1}{2}\sum_{i=1}^{n}[\psi_{11}(x_i - \overline{x})^2 + 2\psi_{12}(x_i - \overline{x})(y_i - \overline{y}) + \psi_{22}(y_i - \overline{y})^2]$$

po elementima matrice $\boldsymbol{\Psi}$ te primjenom inverzne transformacije $\boldsymbol{\Sigma} = \boldsymbol{\Psi}^{-1}$.

$$\frac{\partial l}{\partial \psi_{11}} = \frac{n}{2}\frac{1}{det(\Psi)}\psi_{22} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$$

$$\frac{\partial l}{\partial \psi_{22}} = \frac{n}{2}\frac{1}{det(\Psi)}\psi_{11} - \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2 = 0$$

$$\frac{\partial l}{\partial \psi_{12}} = -\frac{n}{2}\frac{1}{det(\Psi)}2\psi_{12} - \frac{1}{2}\sum_{i=1}^{n}2(x_i - \overline{x})(y_i - \overline{y}) = 0$$

Izračunavanjem prethodnih izraza dobivamo sljedeće:

$$\frac{\psi_{22}}{det(\Psi)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\frac{\psi_{11}}{det(\Psi)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$\frac{\psi_{12}}{det(\Psi)} = -\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Slijedi

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\widehat{\sigma_2^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$\widehat{\rho\sigma_1\sigma_2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Kako $\boldsymbol{\Sigma}^{-1}$ možemo zapisati i na sljedeći način

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{\psi_{22}}{\det(\Psi)} & -\frac{\psi_{12}}{\det(\Psi)} \\ -\frac{\psi_{12}}{\det(\Psi)} & \frac{\psi_{11}}{\det(\Psi)} \end{bmatrix}$$

tada slijedi da je

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}(x_i - \overline{x})^2 & \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) & \sum_{i=1}^{n}(y_i - \overline{y})^2 \end{bmatrix}$$

što odgovara obliku navedenom u iskazu teorema za slučaj $p = 2$. Odavde također možemo dobiti procjenu za koeficijent korelacije ρ:

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{4.15}$$

Analognim razmatranjem, uz korištenje izraza

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Psi}) = -\frac{1}{2}pn\log(2\pi) + \frac{1}{2}n\log det(\boldsymbol{\Psi}) - \frac{1}{2}\sum_{i=1}^{n} tr[(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})'\boldsymbol{\Psi}]$$

dokazuje se oblik od $\widehat{\boldsymbol{\Sigma}}$ naveden u dokazu teorema za prirodan broj $p > 2$.
Q.E.D.

Primjenom Leme 4.1. možemo dobiti drugi izraz za računanje procjene od $\boldsymbol{\Sigma}$

$$\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}})' = \sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i}' - n\overline{\boldsymbol{x}}(\overline{\boldsymbol{x}})'.$$

Korolar 4.1. Ako je $\boldsymbol{X_1}, ..., \boldsymbol{X_n}$ j.sl. uzorak iz $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gdje je $\sigma_{jk} = \sigma_j\sigma_k\rho_{jk}$, ML procjenitelj od $\boldsymbol{\mu}$ je $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} = (1/n)\sum_{i=1}^{n}\boldsymbol{x_i}$, ML procjenitelj od σ_j^2 je $\widehat{\sigma_j^2} = (1/n)\sum_{i=1}^{n}(x_{ji} - \overline{x_j})^2 = (1/n)(\sum_i^n x_{ji}^2 - n\overline{x_j}^2)$. x_{ji} je j-ta komponenta od $\boldsymbol{x_i}$, $\overline{x_j}$ j-ta komponenta od $\overline{\boldsymbol{x}}$ i ML procjenitelj od ρ_{jk} je

$$\begin{aligned}\widehat{\rho_{jk}} &= \frac{\sum_i(x_{ji} - \overline{x_j})(x_{ki} - \overline{x_k})}{\sqrt{\sum_i(x_{ji} - \overline{x_j})^2}\sqrt{\sum_i(x_{ki} - \overline{x_k})^2}} \\ &= \frac{\sum_i x_{ji}x_{ki} - n\overline{x_j}\overline{x_j}}{\sqrt{\sum_i x_{ji}^2 - n\overline{x_j}^2}\sqrt{\sum_i x_{ki}^2 - n\overline{x_k}^2}}.\end{aligned} \tag{4.16}$$

Dokaz:

Skup parametara $\mu_j = \mu_j, \sigma_j^2 = \sigma_{jj}$ i $\rho_{jk} = \sigma_{jk}/\sqrt{\sigma_{jj}\sigma_{kk}}$ je bijekcija skupa parametara μ_j i σ_{jk}. Koristeći Napomenu 4.1. i Teorem 4.1., procjena od μ_j je $\widehat{\mu_j}$, od σ_j^2 je $\widehat{\sigma_{jj}}$ i od ρ_{jk} je

$$\widehat{\rho_{jk}} = \frac{\widehat{\sigma_{jk}}}{\sqrt{\widehat{\sigma_{jj}}\widehat{\sigma_{kk}}}}.$$

Q.E.D.
Formula

$$\widehat{\rho_{jk}} = \frac{\sum_i(x_{ji} - \overline{x_j})(x_{ki} - \overline{x_k})}{\sqrt{\sum_i(x_{ji} - \overline{x_j})^2}\sqrt{\sum_i(x_{ki} - \overline{x_k})^2}} \tag{4.17}$$

definira tzv. Pearsonov korelacijski koeficijent. Obično se u literaturi označava sa r_{jk}, stoga ćemo tako nastaviti označavati u sljedećem poglavlju.

4.1 Pearsonov korelacijski koeficijent bivarijantnog uzorka normalne distribucije

U daljnjem razmatranju izvodimo distribuciju Pearsonovog korelacijskog koeficijenta jednostavnog slučajnog uzorka iz dvodimenzionalne normalne distribucije, i to prvo kada je populacijski korelacijski koeficijent jednak 0, a zatim za bilo koju vrijednost populacijskog korelacijskog koeficijenta. Također ćemo izvesti izraze za procjenu populacijskog korelacijskog koeficijenta pouzdanim intervalom.

4.1.1 Populacijski korelacijski koeficijent jednak je 0

Neka je (X, Y) dvodimenzionalni slučajni vektor, $(X_1, Y_1), ..., (X_n, Y_n)$ nezavisne jednakodistribuirane kopije (X, Y), a podaci $(x_1, y_1), ..., (x_n, y_n)$ nezavisne realizacije od (X, Y).

Pretpostavimo da (X_i, Y_i), $i = 1, ..., n$ ima distribuciju

$$\mathcal{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 \end{pmatrix} \right]. \tag{4.18}$$

Neka je

$$A = \begin{bmatrix} \sum_{i=1}^n (x_i - \overline{x})^2 & \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \\ \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) & \sum_{i=1}^n (y_i - \overline{y})^2 \end{bmatrix} = \begin{bmatrix} \mathrm{a}_{11} & \mathrm{a}_{12} \\ \mathrm{a}_{12} & \mathrm{a}_{22} \end{bmatrix} \tag{4.19}$$

Kako je općenito r_{jk} dan izrazom (4.17), vrijedi da je r

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} \tag{4.20}$$

Pearsonov korelacijski koeficijent ovog uzorka.
Sada r možemo zapisati kao

$$r = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}}. \tag{4.21}$$

$\overline{x}$ i $\overline{y}$ su definirani kao

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i \tag{4.22}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i. \tag{4.23}$$

Promatramo li a_{jk} kao statistike, znamo da su one distribuirane isto kao (vidi npr. [1].)

$$\sum_{i=1}^n Z_{ji} Z_{ki}, j, k = 1, 2 \tag{4.24}$$

gdje (Z_{1i}, Z_{2i}) imaju distribuciju

$$\mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 \end{pmatrix} \right], \tag{4.25}$$

a $(Z_{11}, Z_{21}), ..., (Z_{1n}, Z_{2n})$ su međusobno nezavisni.

Uvjetna distribucija od Z_{2i}, uz uvjet $Z_{1i} = z_{1i}$, je $\mathcal{N}(\beta z_{1i}, \sigma^2)$, gdje je $\beta = \rho\sigma_2/\sigma_1$ i $\sigma^2 = \sigma_2^2(1 - \rho^2)$. Zajednička distribucija od $\boldsymbol{Z_2} = (Z_{21}, ..., Z_{2n})$ uz uvjet $\boldsymbol{Z_1} = \boldsymbol{z_1}$, gdje su

$\boldsymbol{Z_1} = (Z_{11}, ..., Z_{1n})$ i $\boldsymbol{z_1} = (z_{11}, ..., z_{1n})$, je $\mathcal{N}(\beta z_{1i}, \sigma^2 I)$, zato što su Z_{2i} nezavisni. Preciznije, funkciju gustoće uvjetne distribucije komponente Z_{2i}, uz uvjet $Z_{1i} = z_{1i}$ računamo pomoću funkcija gustoće od $f(z_{1i}, z_{2i})$ od (Z_{1i}, Z_{2i}) i funkcije gustoće $f_{Z_{1i}}(z_{1i})$, tj.

$$f_{Z_{2i}|Z_{1i}=z_{1i}}(z_{2i}) = \frac{f(z_{1i}, z_{2i})}{f_{Z_{1i}}(z_{1i})} = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(z_{2i}-(\rho\frac{\sigma_2}{\sigma_1}z_{1i}))}{2\sigma_2^2(1-\rho^2)}}, \forall i = 1, ..., n \tag{4.26}$$

Analogan rezultat vrijedi za uvjetnu distribuciju komponenata Z_{1i}, uz uvjet $Z_{2i} = z_{2i}$.

Lema 4.1.1 Ako su (Z_{1i}, Z_{2i}), $i = 1, ..., n$, nezavisne, s distribucijom (4.25), tada je uvjetna distribucija od $b = \frac{\sum_i Z_{2i}Z_{1i}}{\sum_i Z_{1i}^2}$ uz uvijet $Z_{1i} = z_{1i}$ $\mathcal{N}(\beta, \sigma^2/c^2)$ $(c^2 = \sum_{i=1}^n z_{1i}^2)$ a od $V/\sigma^2 = \frac{\sum_i (Z_{2i}-bZ_{1i})^2}{\sigma^2}$ je, uz uvjet $Z_{1i} = z_{1i}$, χ^2 sa $n-1$ stupnjeva slobode. Također b i V su nezavisni.

Ako je $\rho = 0$, tada je $\beta = 0$ a b ima $\mathcal{N}(0, \sigma^2/c^2)$ distribuciju i

$$T = \frac{cb/\sigma}{\sqrt{\frac{V/\sigma^2}{n-1}}} = \frac{cb}{\sqrt{\frac{V}{n-1}}} \tag{4.27}$$

ima uvjetnu $t-$distribuciju sa $n-1$ stupnjeva slobode. Ta slučajna varijabla je

$$T = \sqrt{n-1}\frac{a_{12}\sqrt{a_{11}}/a_{11}}{\sqrt{a_{22} - a_{12}^2/a_{11}}} = \sqrt{n-1}\frac{a_{12}/\sqrt{a_{11}a_{22}}}{\sqrt{1-[a_{12}^2/(a_{11}a_{12})]}} = \sqrt{n-1}\frac{R}{\sqrt{1-R^2}}, \tag{4.28}$$

gdje je R statistika Pearsonovog korelacijskog koeficijenta.
Tako $R\sqrt{n-1}/\sqrt{1-R^2}$ ima uvjetnu $t-$distribuciju sa $n-1$ stupnjeva slobode. Funkcija gustoće od T je

$$f_T(t) = \frac{\Gamma(\frac{1}{2}n)}{\sqrt{n-1}\Gamma[\frac{1}{2}(n-1)]\sqrt{\pi}}(1 + \frac{t^2}{n-1})^{-\frac{1}{2}n}, \tag{4.29}$$

i funkcija gustoće od $W = R/\sqrt{1-R^2}$ je

$$f_W(w) = \frac{\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-1)]\sqrt{\pi}}(1+w^2)^{-\frac{1}{2}n}. \tag{4.30}$$

Kako je $W = R(1-R^2)^{-\frac{1}{2}}$, $dw/dr = (1-r^2)^{-\frac{3}{2}}$. Funkcija gustoće od R je

$$f_R(r) = \frac{\Gamma[\frac{1}{2}(n)]}{\Gamma[\frac{1}{2}(n-1)]\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(n-3)}. \tag{4.31}$$

To je također i marginalna funkcija gustoće od R. Tako smo dokazali sljedeći teorem.

Teorem 4.1.1. Neka su $\boldsymbol{X_1}$,...,$\boldsymbol{X_n}$ međusobno nezavisni, sa distribucijom $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ako je $\rho = 0$, funkcija gustoće od R je (4.31).

Najvažnija upotreba teorema 4.1.1. je definiranje statističkog testa za testiranje hipoteze da par varijabli nije koreliran. Kako bismo korištenjem procjene koeficijenta korelacije potvrdili zavisnost slučajnih varijabli, potrebno je odbaciti statističku hipotezu

$$\mathcal{H}_0 : \rho = 0.$$

Ako je nul-hipoteza istinita, statistika T ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode. Označimo li s T_{n-1} slučajnu varijablu koja ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode, pripadnu p-vrijednost statističkog testa određujemo na slijedeći način:

$p = P\{T_{n-1} \geq t\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \rho > 0$

$p = P\{T_{n-1} \leq t\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \rho < 0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α i donosimo odluku:

ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su slučajne varijable X i Y zavisne

ako je $p > \alpha$, nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze, tj. kažemo da nemamo dovoljno argumenata kojima bismo potkrijepili tvrdnju da su X i Y zavisne varijable.

4.1.2 Populacijski korelacijski koeficijent različit je od nule

Da bi pronašli distribuciju uzoračkog koefijenta korelacije kada je popluacijski korelacijski koeficijent različit od nule, moramo prvo izvesti zajedničku funkciju gustoće od statistika a_{11}, a_{12} i a_{22}. Na temelju Leme 4.1.1. znamo da su, pod uvjetom da je z_1 fiksan, slučajne varijable $b = a_{12}/a_{11}$ i $V/\sigma^2 = \frac{a_{22} - \frac{a_{12}^2}{a_{11}}}{\sigma^2}$ nezavisne. Pri tome $b \sim \mathcal{N}(\beta, \sigma^2/c^2)$ $(c^2 = \sum_{i=1}^n z_{1i}^2)$ a V/σ^2 ima χ^2 distribucijom sa $n-1$ stupnjeva slobode.
Označimo li funkciju gustoće χ^2 distribucije sa stupnjem slobode $n-1$ kao $g_{n-1}(v)$, a gustoću slučajne varijable b kao $f_b(x)$ zajedničku funkciju gustoće od b, V i Z_1 možemo prikazati kao

$$f(x, v, z_1) = f_{b,V}(x, v|z_1) f_{Z_1}(z_1) = f_b(x|z_1) g_{n-1}(v) f_{Z_1}(z_1). \tag{4.32}$$

S obzirom na to da slučajna varijabla $a_{11}/\sigma_1^2 = Z_1' Z_1/\sigma_1^2$ ima χ^2 distribuciju sa stupnjem slobode n, funkcija gustoće slučajne varijable a_{11} je

$$f_{a_{11}}(p) = \frac{1}{\sigma_1^2} g_n(\frac{p}{\sigma_1^2}). \tag{4.33}$$

Odavde slijedi da je zajednička funkcija gustoće od b, V i a_{11} oblika:

$$\begin{aligned} f(x, v, p) = \frac{(p)^{\frac{1}{2}n-1}}{(2\sigma_1^2)^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)} exp(-\frac{1}{2\sigma_1^2}p) \frac{\sqrt{p}}{\sqrt{2\pi\sigma^2}} exp[-\frac{p}{2\sigma^2}(x-\beta)^2] * \\ * \frac{1}{(2\sigma^2)^{\frac{1}{2}(n-1)}\Gamma[\frac{1}{2}(n-1)]} v^{\frac{1}{2}(n-3)} exp(-\frac{1}{2\sigma^2}v). \end{aligned} \tag{4.34}$$

Znamo da su $b = a_{12}/a_{11}$ i $V = a_{22} - a_{12}^2/a_{11}$. Determinanta Jakobijeve matrice je

$$|\frac{\partial(x, v)}{\partial(q, s)}| = \begin{vmatrix} \frac{\partial x}{\partial q} & \frac{\partial x}{\partial s} \\ \frac{\partial v}{\partial q} & \frac{\partial v}{\partial s} \end{vmatrix} = \begin{vmatrix} \frac{1}{p} & 0 \\ -2\frac{q}{p} & 1 \end{vmatrix} = \frac{1}{p}. \tag{4.35}$$

Funkcija gustoće od a_{11}, a_{12} i a_{22} je

$$f(p, q, s) = \frac{p^{\frac{1}{2}(n-3)} (\frac{ps-q^2}{p})^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}Q}}{2^n \sigma_1^n (\frac{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2}{\sigma_1^2})^{\frac{1}{2}n} \sqrt{\pi}\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}, \tag{4.36}$$

gdje je

$$Q = \frac{1}{1-\rho^2}(\frac{p}{\sigma_1^2} - 2\rho\frac{q}{\sigma_1\sigma_2} + \frac{s}{\sigma_2^2}). \tag{4.37}$$

Funkciju gustoće od a_{11}, a_{12} i a_{22} također možemo zapisati u sljedećem obliku

$$f(p,q,s) = \frac{|A|^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}Q}}{2^n |\mathbf{\Sigma}|^{\frac{1}{2}n} \sqrt{\pi}\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}. \tag{4.38}$$

Prethodni izraz je poseban slučaj Wishart distribucije (vidi npr. [1].).
Funkcija gustoće od a_{11}, a_{22} i $R = a_{12}\sqrt{a_{11}a_{22}}$ je

$$f(p,s,r) = \frac{p^{\frac{1}{2}n-1} s^{\frac{1}{2}n-1} (1-r^2)^{\frac{1}{2}(n-3)} e^{-\frac{1}{2}Q}}{2^n [\sigma_1^2\sigma_2^2(1-\rho^2)]^{\frac{1}{2}n} \sqrt{\pi}\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}, \tag{4.39}$$

gdje je

$$Q = \frac{1}{(1-\rho^2)}(\frac{p}{\sigma_1^2} - 2\rho r \frac{\sqrt{p}\sqrt{s}}{\sigma_1\sigma_2} + \frac{s}{\sigma_2^2}). \tag{4.40}$$

Da bismo pronašli funkciju gustoće od R, moramo integrirati (4.38), s obzirom na to da su a_{11} i a_{22} unutar raspona od 0 do ∞. Dio eksponenta ćemo razviti u red:

$$exp[\frac{\rho r\sqrt{p}\sqrt{s}}{(1-\rho^2)\sigma_1\sigma_2}] = \sum_{i=0}^{\infty} \frac{(\rho r\sqrt{p}\sqrt{s})^i}{i![\sigma_2\sigma_1(1-\rho^2)]^i}. \tag{4.41}$$

Tada je funkcija gustoće (4.38)

$$\begin{aligned} f(p,s,r) = {} & \frac{(1-r^2)^{\frac{1}{2}(n-3)}}{\sigma_1^n\sigma_2^n(1-\rho^2)^{\frac{1}{2}n} 2^n \sqrt{\pi}\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)} \sum_{i=0}^{\infty} \frac{(\rho r)^i}{i![(1-\rho^2)]^i \sigma_1^i \sigma_2^i} \\ & \{exp[-\frac{p}{2(1-\rho^2)\sigma_1^2}]p^{\frac{1}{2}(n+i)-1}\}\{exp[-\frac{s}{2(1-\rho^2)\sigma_2^2}]s^{\frac{1}{2}(n+i)-1}\}. \end{aligned} \tag{4.42}$$

Kako je

$$\int_0^{\infty} p^{\frac{1}{2}(n+i)-1} exp[-\frac{p}{2(1-\rho^2)\sigma_1^2}]dp = \Gamma[\frac{1}{2}(n+i)][2\sigma_1^2(1-\rho^2)]^{\frac{1}{2}(n+i)}, \tag{4.43}$$

integral izraza (4.41) je

$$\frac{(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{\sqrt{\pi}\Gamma(\frac{1}{2}n)\Gamma[(\frac{1}{2}(n-1))]} \sum_{i=0}^{\infty} \frac{(2\rho r)^i}{i!} \Gamma^2[\frac{1}{2}(n+i)]. \tag{4.44}$$

Ako koristimo formulu gama funkcije

$$\Gamma(z)\Gamma(z+1) = \frac{\sqrt{\pi}\Gamma(2z)}{2^{2z-1}}, \tag{4.45}$$

možemo prilagoditi prethodni izraz. Tako smo dokazali sljedeći teorem:

Teorem 4.1.2. Korelacijski koeficijent n uzorka bivarijantne normalne distribucije s korelacijom ρ je distribuiran funkcijom gustoće

$$f_R(r) = \frac{2^{n-2}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{(n-2)!\pi} \sum_{i=0}^{\infty} \frac{(2\rho r)^i}{i!} \Gamma^2[\frac{1}{2}(n+i)]. \tag{4.46}$$

4.1.3 Asimptotska distribucija uzoračkog koeficijenta korelacije

U bivarijantnom normalnom modelu, standardni test za hipotezu $\mathcal{H} : \rho = 0$, ili ekvivalentno, $\mathcal{H} : X, Y$ su nezavisni, odbacuje $\mathcal{H}$ u korist alternative $\rho \neq 0$ kada je uzorački korelacijski koeficijent

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})/n}{\widehat{\sigma_1}\widehat{\sigma_2}} \tag{4.47}$$

dovoljno velik po apsolutnoj vrijednosti, gdje su

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \widehat{\sigma_2^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

To je posljedica činjenice da

$$\sqrt{n}R = \frac{\sqrt{n}\sum_{i=1}^{n} x_i y_i/n}{\widehat{\sigma_1}\widehat{\sigma_2}} - \frac{\sqrt{n}(\overline{x})(\overline{y})}{\widehat{\sigma_1}\widehat{\sigma_2}} \tag{4.48}$$

konvergira po distribuciji prema normalnoj slučajnoj varijabli sa očekivanjem 0 i varijancom 1 za $n \to \infty$. Zaista, već smo do sada koristili činjenicu da je distribucija od R neovisna o iznosu očekivanja slučajne varijable X i Y pa, bez smanjenja općenitosti, možemo pretpostaviti da je $EX = EY = 0$. Pod pretpostavkom $\mathcal{H}$, varijable X i Y su nezavisne pa vrijedi

$$\text{Var } XY = E(X^2Y^2) = E(X^2)E(Y^2) = \text{Var } X \text{ Var } Y. \tag{4.49}$$

Primjenom centralnog graničnog teorema (vidi npr. [6].) i teorema "Slutsky" (vidi npr. [7].) vidimo da prvi član u izrazu (4.48) konvergira po distribuciji prema normalnoj slučajnoj varijabli sa očekivanjem 0 i vrijancom 1. S obzirom na to da drugi član u tome izrazu konvergira prema 0 po vjerojatnosti slijedi da

$$\sqrt{n}R \xrightarrow{D} Z \sim \mathcal{N}(0, 1). \tag{4.50}$$

Područje odbacivanja hipoteze $\mathcal{H}$,

$$|\sqrt{n}R| \geq u_{\alpha/2} \tag{4.51}$$

ima asimptotski nivo značajnosti α za testiranje $\mathcal{H}$.
Razmotrimo što će se dogoditi nivou od (4.51) kada pretpostavka normalnosti nije opravdana. Pretpostavimo da je $(X_1, Y_1), ..., (X_n, Y_n)$ uzorak iz neke bivarijantne distribucije F s konačnim drugim momentima i neka je ρ korelacijski koeficijent. U normalnom slučaju, hipoteza

$$\mathcal{H}_1 : \rho = 0$$

je ekvivalentna hipotezi

$$\mathcal{H}_2 : \text{X,Y su nezavisni.}$$

U općenitom slučaju to ne vrijedi, stoga postaje potrebno razlikovati $\mathcal{H}_1$ i $\mathcal{H}_2$. Ako su X i Y nezavisne tada i (4.48) i (4.49) ostaju valjane i asimptotski nivo od (4.51) nastavlja biti α. Kada je $\rho = 0$, ali X i Y nisu nezavisne broj

$$\gamma^2 = \frac{\text{Var}\,(XY)}{\text{Var } X \text{ Var } Y} \tag{4.52}$$

može preuzeti bilo koju vrijednost između 0 i ∞. To možemo vidjeti ako stavimo da su $i) Y = X$ i $ii) Y = 1/X$. Ako je X simetričan oko 0, onda je, u slučaju $i)$

$$\gamma^2 = \frac{E(X^4) - [E(X^2)]^2}{[E(X^2)]^2} \tag{4.53}$$

što može biti proizvoljno veliko (uključujući ∞), stavljajući dovoljno težine u rep distribucije X. U slučaju *ii*)

$$\gamma^2 = \frac{1}{\mathrm{Var}\,(1/X)}. \tag{4.54}$$

Ovdje nazivnik može biti proizvoljno velik, stavljajući pri tome dovoljno težine u ishodište, čineći time $\mathrm{Var}\,(1/X)$ velikom bez prevelikog mjenjanja $\mathrm{Var}\ X$.
Slijedi,

$$\sqrt{n}R \to \mathcal{N}(0, \gamma^2), \tag{4.55}$$

i vidimo da asimptotski nivo od (4.51) sada može poprimiti bilo koju vrijednost $\alpha(\gamma)$ između 0 i 1. Nivo normalnog teorijskog testa je asimptotski robusan prema nenormalnosti pod $\mathcal{H}_2$, ali ne pod $\mathcal{H}_1$.

5 Literatura

[1] Anderson T. W.: *An Introduction to Multivariate Statistical Analysis, Third Edition* , Stanford University Department of Statistics, Stanford, CA (1984).

[2] Benšić M., Šuvak N.: *Uvod u vjerojatnost i statistiku* , Sveučilište J.J. Strossmayera, Odjel za matematiku, Osijek (2014).

[3] Lehmann E.L.: *Elements of Large-Sample Theory* , Springer, New York (2001).

[4] Pandžić P., Tambača J., Diferencijalni račun funkcija više varijabli, materijali sa predavanja, Zagreb: `https://web.math.pmf.unizg.hr/nastava/difraf/dif/p_o16.pdf`, 25.1.2018

[5] Papić I., Statistika, materijali sa vježbi, Osijek: `http://www.mathos.unios.hr/statistika/Vjezbe/STATvjezbe3.pdf`, 25.1.2018.

[6] Šarapa N.: *Teorija vjerojatnosti, Treće, prerađeno izdanje* , Školska knjiga, Zagreb (2002).

[7] Šuvak N., Vjerojatnost, materijali sa predavanja, Osijek: `http://www.mathos.unios.hr/images/homepages/nsuvak/vjerojatnost/v9.pdf`, 25.1.2018.

6 Sažetak

Koeficijent korelacije numerička je karakteristika dvodimenzionalnog vektora koja se koristi kao pokazatelj povezanosti među njegovim komponentama. Na temelju jednostavnog slučajnog uzorka iz tog slučajnog vektora, koeficijent korelacije procjenjuje se izrazom koji je poznat pod nazivom Pearsonov korelacijski koeficijent. U ovom diplomskom radu izvedena je funkcija gustoće Pearsonovog korelacijskog koeficijenta u modelu jednostavnog slučajnog uzorka iz dvodimenzionalnog normalnog slučajnog vektora. Taj rezultat iskorišten je za definiranje statističkog testa o iznosu populacijskog korelacijskog koeficijenta.

Ključne riječi: Korelacija, Pearsonov korelacijski koeficijent, metoda maksimalne vjerodostojnosti

7 Summary

Correlation coefficient is the numerical characteristic of the two-dimensional random vector used as an indicator of connections among its components. Based on simple random sample from that random vector, the correlation coefficient is estimated using the expression that is known as Pearson's correlation coefficient. In this graduate thesis the density function of the Pearson Correlation Coefficient in a simple random sample model was derived from a two-dimensional normal random vector. This result was used to define the statistical test of the amount of the population correlation coefficient.

Key words: Correlation, Pearson's correlation coefficient, maximum likelihood method

8 Životopis

Rođena sam 5. svibnja 1990. u Osijeku. Osnovnu sam školu pohađala u Petrijevcima, nakon koje upisujem Isusovačku klasičnu gimnaziju s pravom javnosti u Osijeku. Nakon završene gimnazije upisujem Preddiplomski studij matematike na Odjelu za matematiku, na Sveučilištu Josipa Jurja Strossmayera u Osijeku. Preddiplomski studij završavam izradom završnog rada pod nazivom Primjene kongruencija, te stječem akademski stupanj prvostupnice matematike. Akademsko obrazovanje nastavljam na Odjelu za matematiku u Osijeku, smjer Financijska matematika i statistika. Tijekom završne godine diplomskog studija obavila sam stručnu studentsku praksu u tvrtki Farmeron d.o.o. Tijekom apsolventske godine obavljam stručnu praksu kroz Erasmus program u trajanju od tri mjeseca.