Mjere asimetrije podataka

Šimić, Mihaela

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

Mihaela Šimić

Mjere asimetrije podataka

Diplomski rad

Voditelj rada:
prof.dr.sc. Miljenko Marušić

Zagreb, rujan 2017.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____ , predsjednik
2. _____ , član
3. _____ , član

Povjerenstvo je rad ocijenilo ocjenom _____ .

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentoru prof.dr.sc. Miljenku Marušiću na susretljivosti i pomoći oko izrade ovog rada.

Sadržaj

Uvod

U ovom radu bavit ćemo se asimetrijom u rasporedu podataka, odnosno proučavat ćemo na koje sve načine i kojim sve mjerama možemo detektirati postoji li asimetrija u rasporedu podataka.

Na početku prvog poglavlja dan je pregled osnovnih pojmova i oznaka koji će se koristiti u nastavku rada.
Zatim opisujemo koncept asimetrije te definiramo tri osnovne, odnosno najčešće korištene mjere asimetrije: koeficijent asimetrije, Pearsonova mjera asimetrije i Bowlyjeva mjera asimetrije.

U drugom poglavlju istražujemo postojanje drugih mjera, odnosno statističkih veličina kojima bismo mogli detektirati asimetriju. Bavimo se testiranjem asimetrije na simuliranim podacima te uspoređujemo jakost pojedinih statističkih veličina u detekciji asimetrije u rasporedu. Glavni zaključak poglavlja je da su mnoge statistike vjerojatne, no nisu sve dovoljno učinkovite.

Na početku trećeg poglavlja „Testovi omjera vjerodostojnosti za testiranje simetrije" definiramo pojam procjenitelja metodom maksimalne vjerodostojnosti te opisujemo metodu omjera vjerodostojnosti. U nastavku opisujemo postavljene hipoteze te definiramo dvije testne statistike i njihove asimptotske distribucije. Na kraju izvodimo simulacijsku studiju pomoću koje uspoređujemo jakost jednog od predloženih testova. Glavni zaključak poglavlja je da novopredloženi test mnogo bolji u odnosu na poznati neograničeni test omjera vjerodostojnosti za testiranje simetrije.

1 Asimetrija podataka

U deskriptivnoj statistici razvijene su određene metode i postupci za egzaktno proučavanje statističkih podataka. Kada govorimo o brojnijem nizu statističkih podataka, onda tablični i grafički prikaz tih podataka omogućuje vrlo jasan i pregledan uvid u bitna svojstva promatranog statističkog obilježja X. No, bitna se svojstva promatranog statističkog obilježja mogu izraziti još sažetije, odnosno karakterizacijom pomoću *mjere* (broja) ili više njih, koje će se na određeni način definirati pomoću danog niza $x_1, \ldots, x_n$ statističkih podataka o obilježju X.

Jedna od najvažnijih mjera jest *aritmetička sredina* niza izmjerenih vrijednosti $x_1, \ldots, x_n$ obilježja X koja pripada *mjerama centralne tendencije*, to jest mjeri „srednju vrijednost" podataka. Obično se označava s $\overline{x}$ i definira izrazom:

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

Nadalje, uvodimo pojam *disperzije* koji označava raspršenost članova numeričkog niza od neke srednje vrijednosti.
Mjere disperzije ili *raspršenosti* su veličine pomoću kojih se utvrđuje veličina raspršenosti članova numeričkog niza od neke srednje vrijednosti, odnosno utvrđuje se reprezentativnost srednjih vrijednosti. Najvažnije mjere disperzije su *varijanca*:

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

i *standardna devijacija*:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Pretpostavimo sada da su izmjerene vrijednosti $x_1, \ldots, x_n$ obilježja X poredane po veličini (u rastućem poretku), to jest vrijedi $x'_1 \leq x'_2 \leq \ldots \leq x'_n$.
Medijan podataka je mjera centralne tendencije numeričkih podataka, a ima značenje izmjerene vrijednosti koja se nalazi na sredini niza podataka kada

je on uređen po veličini. Izračunava se na sljedeći način:

$$Me = \begin{cases} \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}), & \text{za parno } n \\ x'_{\frac{n+1}{2}}, & \text{za neparno } n \end{cases}$$

Postotna vrijednost x'_p, za neki izabrani broj $p \in \langle 0, 100\rangle$, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti manje ili jednako x'_p, dok je barem $(100-p)\%$ vrijednosti veće ili jednako x'_p. Dvadeset pet postotnu vrijednost nazivamo *donji kvartil* i označavamo s $Q_1 = x'_{(\frac{n+1}{4})}$, a sedamdeset pet postotnu vrijednost *gornji kvartil* te označavamo $Q_3 = x'_{(\frac{3(n+1)}{4})}$. Donji i gornji kvartil su mjere koje pripadaju grupi mjera disperzije podataka.

Raspon podataka je također mjera disperzije te je definiran kao razlika najveće i najmanje vrijednosti u danom nizu statističkih podataka.

$$R = max\{x_1, \ldots, x_n\} - min\{x_1, \ldots, x_n\} = x'_n - x'_1.$$

Mod (Mo) podataka je vrijednost iz niza izmjerenih vrijednosti varijable kojoj pripada najveća frekvencija, odnosno izmjerena je najviše puta. Mod ne mora biti jedinstven.

Skup izmjerenih vrijednosti može se grafički prikazati pomoću *dijagrama pravokutnika* (eng. *box plot*) koji prikazuje odnos pet numeričkih karakteristika skupa izmjerenih vrijednosti: minimalna vrijednost, donji kvartil, medijan, gornji kvartil i maksimalna vrijednost, koje čine *karakterističnu petorku* danog niza statističkih podataka.

1.1 Koncept asimetrije

Asimetrija je pojam suprotan simetriji i pokazuje da se lijevi krak krivulje ne preklapa s desnim krakom krivulje preko osi simetrije (okomice s vrha krivulje).

Mjere asimetrije su veličine kojima se utvrđuje postoji li simetrija ili asimetrija te u slučaju asimetrije, smjer i njezina jačina (veličina). Mjeri se način rasporeda podataka prema nekoj srednjoj vrijednosti, odnosno osi simetrije.

Raspored podataka (distribucija) može biti:

1. negativno asimetričan (lijevostran) – mjere asimetrije su manje od nule
2. simetričan – mjere asimetrije su jednake nuli
3. pozitivno asimetričan (desnostran) – mjere asimetrije su veće od nule

Prema jačini asimetrija može biti jaka (velika) ili slaba (manja).
Neke od najvažnijih mjera asimetrije su:

1. Koeficijent asimetrije, α_3
2. Pearsonova mjera asimetrije, S_k
3. Bowleyjeva mjera asimetrije, S_{kQ}

Za utvrđivanje asimetrije definiraju se tzv. *statistički momenti.*
Momenti oko sredine ili *centralni momenti k-tog reda* (μ_k) danog statističkog niza predstavljaju aritmetičku sredinu odstupanja vrijednosti numeričkog obilježja od aritmetičke sredine podignutih na k-tu potenciju, to jest

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k. \tag{1}$$

Ishodišni ili *pomoćni momenti k-tog reda* (m_k) koriste se radi lakšeg izračunavanja momenata oko sredine, a definirani su formulom:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k. \tag{2}$$

Očigledno je $m_0 = \mu_0 = 1,\ m_1 = \overline{x},\ \mu_1 = 0\ i\ \mu_2 = s^2$.
Iz (1) i (2) slijedi da je

$$\mu_2 = m_2 - m_1^2 \tag{3}$$

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3 \tag{4}$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4, \tag{5}$$

pa se lako vidi da formule (3), (4) i (5) omogućuju da se centralni momenti μ_2, μ_3 i μ_4 izraze pomoću ishodišnih momenata, koji su definirani jednostavnijim formulama.

Uloga momenta trećeg reda μ_3 može se vidjeti iz sljedećeg rezoniranja.
Ako su podaci $x_1, \ldots, x_n$ simetrično raspoređeni oko točke $\overline{x}$, tada svakoj vrijednosti x_i odgovara simetrična vrijednost x_i', tako da je

$$x_i - \overline{x} = -(x_i' - \overline{x}) \quad i$$
$$(x_i - \overline{x})^3 = -(x_i' - \overline{x})^3,$$

iz čega slijedi da je

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 = 0.$$

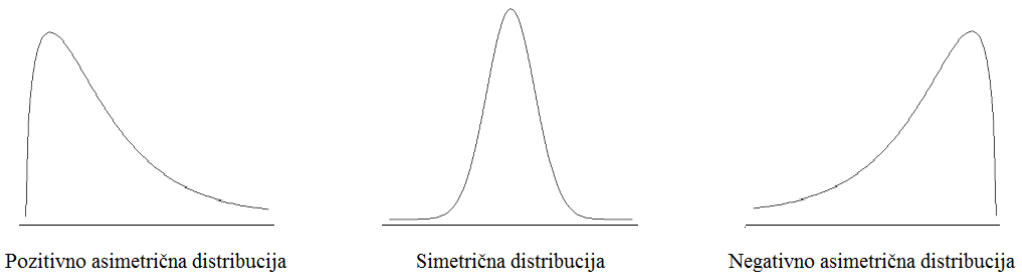
Ako je $\mu_3 > 0$, slijedi da je $\sum_{i=1}^{n}(x_i - \overline{x})^3 > 0$, to jest podaci su „razvučeniji" desno od $\overline{x}$, odnosno „zbijeniji" su lijevo od $\overline{x}$, a ako je $\mu_3 < 0$, onda su podaci „razvučeniji" lijevo od $\overline{x}$, a „zbijeniji" desno od $\overline{x}$.

1.1.1 Koeficijent asimetrije

Koeficijent asimetrije α_3 je standardizirana *mjera smjera i veličine asimetrije* i definira se kao omjer trećeg momenta oko sredine i standardne devijacije podignute na treću potenciju,

$$\alpha_3 = \frac{\mu_3}{\sigma^3}.$$

S obzirom da koristi sva odstupanja vrijednosti numeričke varijable od aritmetičke sredine, koeficijent asimetrije α_3 je *potpuna mjera asimetrije* te u pravilu zauzima vrijednosti u intervalu [-2,2], osim u slučaju vrlo jake asimetrije, kada može prijeći tu granicu.



Slika 1: Raspored podataka u uzorku

Dakle, u simetričnom rasporedu koeficijent α_3 jednak je nuli, u pozitivno asimetričnom je pozitivan, a u negativno asimetričnom negativan.

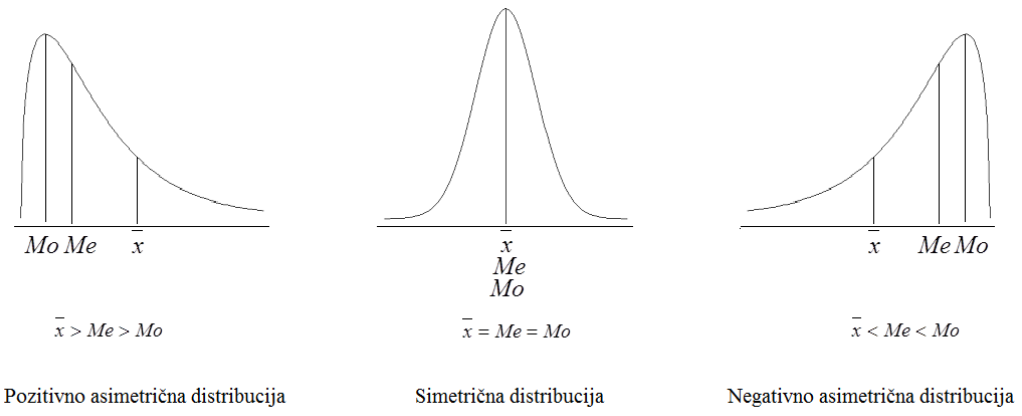
1.1.2 Pearsonova mjera asimetrije

Pearsonova mjera asimetrije S_k jest standardizirano odstupanje medijana ili moda od aritmetičke sredine.

$$S_k = \frac{\overline{x} - Mo}{\sigma},$$

$$S_k = \frac{3(\overline{x} - Me)}{\sigma}.$$

U pravilu se izračunava za neprekidne distribucije. Ako se izračunava za diskretne distribucije, mjeru je potrebno interpretirati s oprezom ili zaključak o asimetriji temeljiti na drugim mjerama asimetrije.
Zauzima vrijednosti u intervalu [-3,3] ovisno o obliku krivulje i jačini asimetrije.



Slika 2: Odnosi srednjih vrijednosti u rasporedu podataka

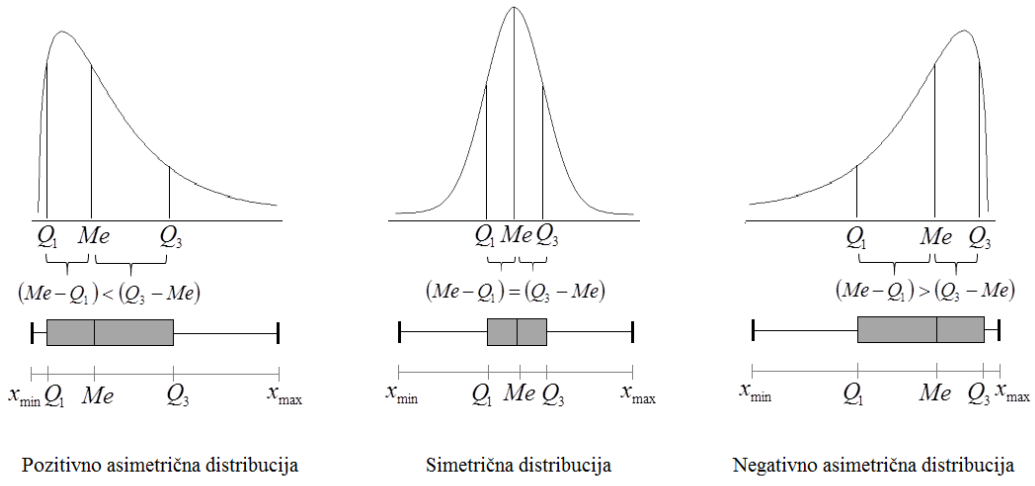
U simetričnom rasporedu podataka sve tri vrijednosti su jednake pa je razlika moda i aritmetičke sredine ili medijana i aritmetičke sredine jednaka nuli. U pozitivno asimetričnom rasporedu podataka ta razlika je pozitivna, a u negativno asimetričnom rasporedu podataka razlika je negativna. Pearsonova mjera je nepotpuna mjera asimetrije i manje je informativna od koeficijenta asimetrije, no izračunava se brže i jednostavnije.

1.1.3 Bowlyjeva mjera asimetrije

Bowlyjeva mjera asimetrije S_{kQ} je mjera asimetrije koja se temelji na odnosima kvartila i medijana.

$$S_{kQ} = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Me}{Q_3 - Q_1}$$

U pravilu zauzima vrijednosti u intervalu [-1,1] ovisno o obliku krivulje i jačini asimetrije.



Slika 3: Odnosi kvartila i medijana u rasporedu podataka

U simetričnom rasporedu vrijednosti razlika gornjeg kvartila i medijana jednaka je razlici medijana i donjeg kvartila, to jest $Q_1 + Q_3 - 2Me = 0$. U pozitivno asimetričnom rasporedu razlika gornjeg kvartila i medijana veća je od razlike medijana i donjeg kvartila, a u negativno asimetričnom rasporedu razlika gornjeg kvartila i medijana manja je od razlike medijana i donjeg kvartila.
Također, kao i Pearsonova mjera, Bowlyjeva mjera je nepotpuna mjera asimetrije.

2 Testiranje asimetrije

Postavlja se pitanje na koje se sve načine može testirati postoji li simetrija ili asimetrija u rasporedu podataka, odnosno uz već navedene mjere asimetrije, postoje li još neke koje bi bile relevantne te ukoliko postoje, koliko bi bile pouzdane.
Navodimo primjer zadatka na temelju kojeg će se u daljnjem nastavku rada ispitivati asimetrija.

Primjer 2.1. *Potrošačka organizacija provodila je istraživanje o potrošnji goriva pretpostavljajući da proizvođač automobila krivo navodi kupce pretjerujući u prosječnoj efikasnosti goriva (mjereno u miljama po galonu, eng. milles per gallon; mpg) određenog modela automobila. Model je reklamiran s 27 mpg. Istraživači su odabrali slučajan uzorak od 10 automobila navedenog modela. Svaki automobil je, na slučajan način, dodijeljen drugom vozaču. Svaki automobil vožen je 5000 milja te je izračunata ukupna potrošnja goriva.*
Nadalje, koristeći jednostrani t-test, na razini značajnosti 0.05 testiraju se sljedeće hipoteze:

$\mathcal{H}_0$*: populacija je normalno distribuirana s* μ*=5 i* σ *=1*
$\mathcal{H}_1$*: populacija je pozitivno asimetrična (desnostrana) s* μ*=5 i* σ *=1*

Za provođenje jednostranog t-testa u ovoj situaciji, izmjereni podaci (odnosno potrošnja goriva) populacije automobila bi trebali biti normalno distribuirani. Međutim box-plot i histogram, dani na slici ispod, navode na to da je distribucija uzorka od 10 vrijednosti desnostrana, odnosno pozitivno asimetrična.

Slika 4: Distribucija uzorka duljine 10

Jedna moguća testna statistika koja bi mjerila postojanje asimetrije je omjer aritmetičke sredine i medijana, $\frac{\overline{x}}{Me}$.
Koje bi sve vrijednosti (male, velike, blizu jedan) navedene statistike mogle ukazivati da je populacijska distribucija potrošnje goriva desnostrana?

Iako je raspored izmjerenih vrijednosti u uzorku desnostran, moguće je da podaci dolaze iz normalne distribucije te da je asimetrija posljedica uzoračke varijabilnosti.

Kako bi se pojava asimetrije istražila, generirano je 10000 uzoraka duljine 10 iz normalne distribucije s istim očekivanjem i standardnom devijacijom kao u originalnom uzorku. Izračunata je testna statistika, odnosno omjer aritmetičke sredine i medijana, za svaki od generiranih uzoraka.

U originalnom uzorku vrijednost testne statistike je 1.03.

Postavlja se pitanje, je li moguće da originalan uzorak 10 automobila dolazi iz normalne distribucije ili simulirani uzorci ukazuju na to da je raspored originalnog uzorka zaista desnostran.

Dana je tablica karakteristične petorke originalnog uzorka 10 automobila.

Minimum	Q_1	Medijan	Q_3	Maksimum
23	24	25.5	28	32

Tablica 1: Karakteristična petorka originalnog uzorka mpg-a.

Koristeći samo minimalnu vrijednost, donji kvartil, medijan, gornji kvartil i maksimalnu vrijednost, definirat ćemo različite testne statistike koje bi mjerile postojanje asimetrije.
Testne statistike prikazane su Tablicom 2 te su izračunate vrijednosti u originalnom uzorku.

	Testna statistika	Vrijednost
A	$\frac{\overline{x}}{Me}$	1.03
B	$\frac{x_{max}-Me}{Me-x_{min}}$	2.76
C	$\frac{Q_3-Me}{Me-Q_1}$	3.79
D	$\frac{x_{max}-Q_3}{Q_1-x_{min}}$	4.68
E	$\frac{\frac{1}{2}(x_{min}+x_{max})}{Me}$	1.15
F	$\frac{\frac{1}{2}(Q_1+Q_3)}{Me}$	1.07
G	$\frac{\frac{1}{2}(x_{min}+x_{max})}{\frac{1}{2}(Q_1+Q_3)}$	1.13

Tablica 2: Testne statistike i njihove vrijednosti u originalnom uzorku

2.1 Omjer aritmetičke sredine i medijana

U ovom potpoglavlju obrađujemo statističku veličinu

$$\frac{\overline{x}}{Me}$$

i njeno utvrđivanje postojanja asimetrije, točnije pozitivne asimetrije u rasporedu podataka.

Intuitivno, ukoliko je raspored podataka simetričan, omjer aritmetičke sredine i medijana jednak je 1. S obzirom da je u pozitivno asimetričnom rasporedu općenito aritmetička sredina veća od medijana, omjer tih veličina bit će veći od 1. Uzimajući u obzir podatke iz uzorka, taj omjer bi vrlo malo vjerojatno bio točno jednak 1, (čak i ako bi uzorak bio iz savršeno normalne distribucije) zbog postojanja određene varijabilnosti uzorkovanja.

Omjer aritmetičke sredine i medijana nema jednostavno definiranu distribuciju, no možemo ju odrediti opetovanim uzorkovanjem iz normalne distribucije (parametarski bootstrap). Označimo taj omjer s θ.
Neka je $x_1, \ldots, x_{10}$ realizacija slučajnog uzorka $X_1, \ldots, X_{10}$ iz Primjera 2.1 te pretpostavljamo da taj slučajni uzorak dolazi iz normalne distribucije s parametrima $\mu = 5$ i $\sigma = 1$. Promatramo parametar θ iz $\mathcal{N}(5,1)$ i njegov procjenitelj $\hat{\theta} = t(X_1, \ldots, X_{10})$. Zanima nas distribucija tog procjenitelja. S obzirom da nam je poznata distribucija parametra, možemo odrediti uzoračku distribuciju od $\hat{\theta}$ opetovanim uzorkovanjem iz normalne distribucije s $\mu = 5$ i $\sigma = 1$. Dakle, osnovna ideja metode je generirati uzorak iz već postojećeg i na temelju generiranih uzoraka procijeniti distribuciju određene statističke veličine.

Pomoću R-a, programskog jezika i statističkog softvera za analizu podataka, modeliranje i grafiku, generiramo 10000 uzoraka duljine 10 iz normalne distribucije s parametrima $\mu = 5$ i $\sigma = 1$ te za svaki od 10000 uzoraka računamo omjer aritmetičke sredine i medijana.[1]

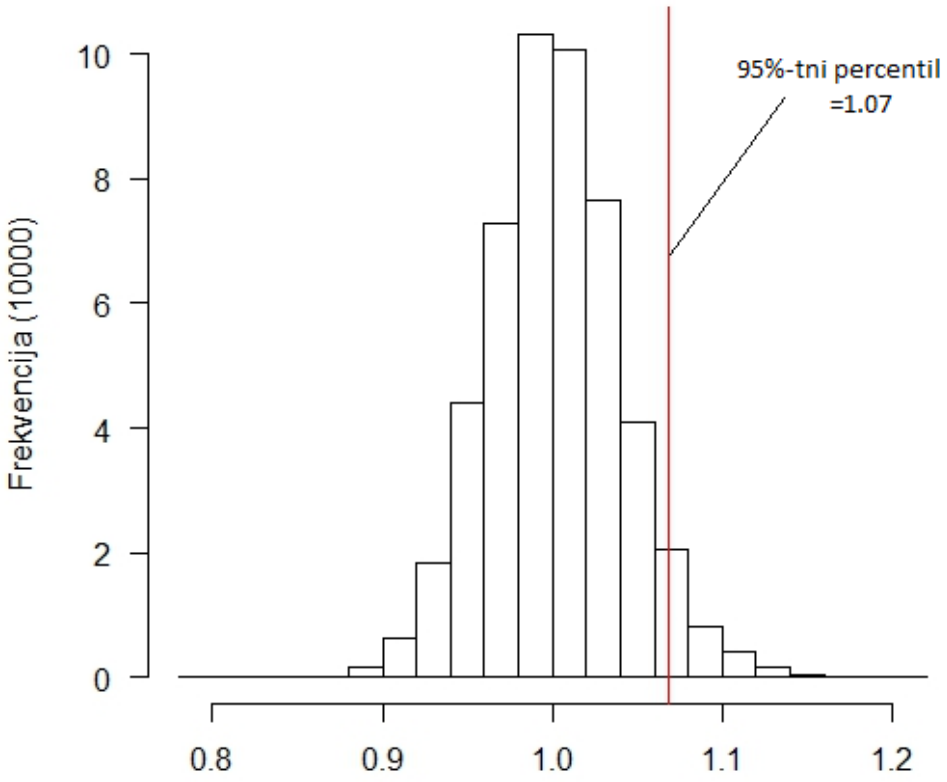
Slika 5 prikazuje histogram rezultata 10000 uzoraka. Distribucija ovog omjera centrirana je u 1, najniža vrijednost jednaka je 0.845, a najviša jednaka 1.190. 95%-tni percentil ove distribucije približno je jednak 1.07.

Dakle, testiranjem hipoteza iz Primjera 2.1:

$\mathcal{H}_0$: populacija je normalno distribuirana s μ=5 i σ =1
$\mathcal{H}_1$: populacija je pozitivno asimetrična (desnostrana) s μ=5 i σ =1,

na razini značajnosti od 0.05 nultu hipotezu ćemo odbaciti za svaki uzorak s omjerom aritmetičke sredine i medijana većim od 1.07.

[1] Vrijednosti $\mu = 5$ i $\sigma = 1$ su izabrane tako da bi se izbjegle negativne vrijednosti te vrijednosti blizu 0.



Slika 5: Simulirana distribucija omjera aritmetičke sredine i medijana normalne populacije (n=10).

Vrijednost testne statistike u originalnom uzorku iz Primjera 2.1 iznosi 1.03, što je manje od kritične vrijednosti testa. Dakle, na razini značajnosti od 0.05 ne bismo odbacili nultu hipotezu $\mathcal{H}_0$.

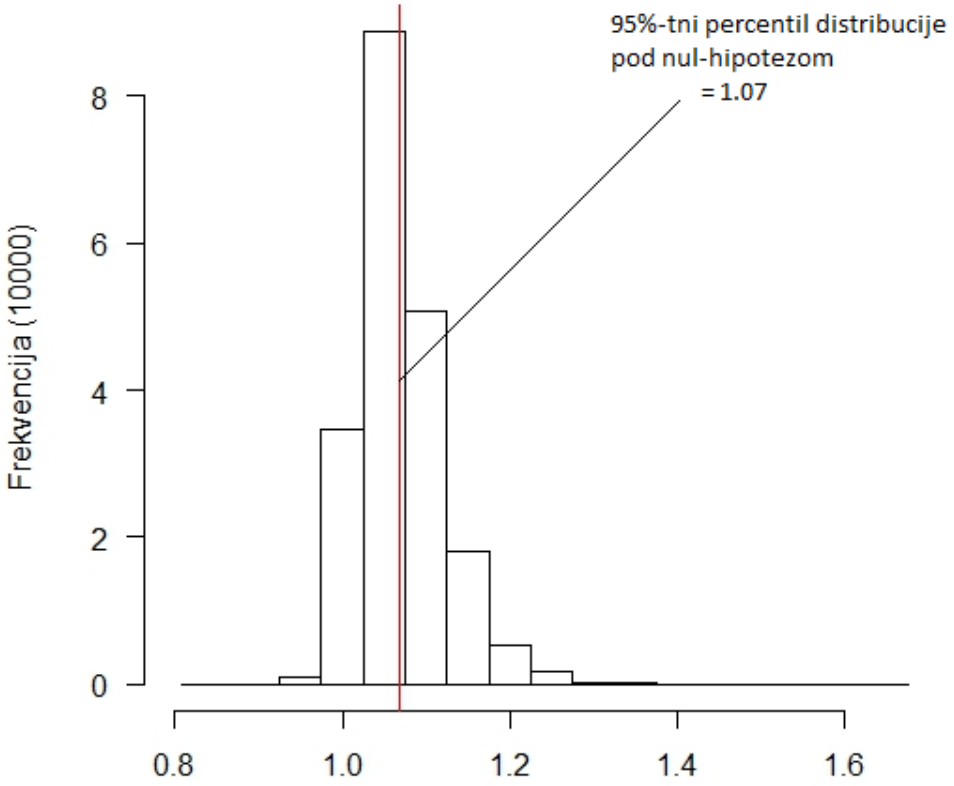
Nadalje, ispitujemo koliko učinkovito ova testna statistika detektira asimetriju, odnosno ispitujemo jakost testa. Prethodno opisanim postupkom, generiramo 10000 uzoraka duljine 10 iz jako pozitivno asimetrične distribucije (χ^2 distribucije s jednim stupnjem slobode, reskalirane tako da očekivanje bude jednako 5, a standardna devijacija jednaka 1).[2]

Za svaki od uzoraka računamo omjer aritmetičke sredine i medijana. Ako je taj omjer veći od 1.07, onda odbacujemo nul-hipotezu i točno zaključujemo da je raspored podataka pozitivno asimetričan.

U ovom kontekstu, jakost testa je udio slučajeva kada je omjer bio veći od 1.07. Jakost testa je vjerojatnost odbacivanja nul-hipoteze pretpostavljajući da je alternativna hipoteza istinita, što je u ovom slučaju vjerojatnost zaključivanja da je distribucija pozitivno asimetrična.

Kako je bilo i očekivano, većinom je omjer bio veći od 1, no *značajno* veći od 1 je samo onda kada je veći od 1.07, a to se dogodilo u 43% slučajeva. Dakle, iako je populacija iz jako pozitivno asimetrične distribucije, jakost ove testne statistike za utvrđivanje asimetrije je samo 0.43 za uzorke duljine 10.

[2]Za generiranjne slučajnog uzorka iz navedene distribucije u R-u, korištena je sljedeća formula: $\frac{rchisq(1)}{\sqrt{2}} + (5 - \frac{1}{\sqrt{2}})$.



Slika 6: Simulirana distribucija omjera aritmetičke sredine i medijana jako pozitivno asimetrične populacije (n=10).

2.2 Istraživanje drugih statističkih veličina

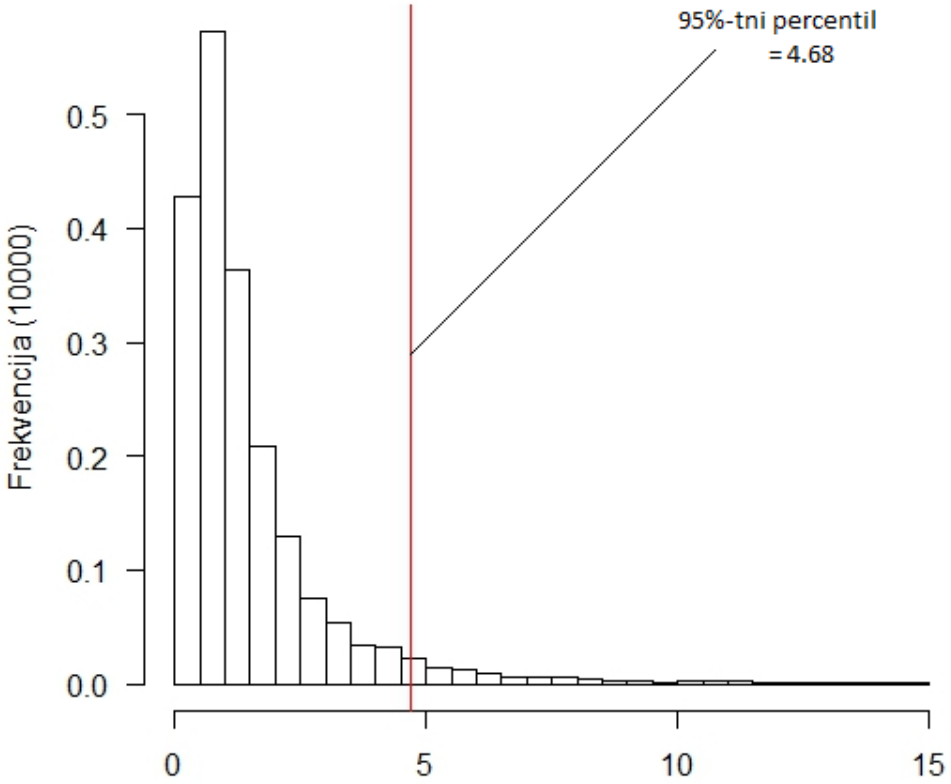
U ovom potpoglavlju obrađujemo statističku veličinu

$$\frac{x_{max} - Q_3}{Q_1 - x_{min}}$$

izvedenu iz karakteristične petorke uzorka (najmanja vrijednost, donji kvartil, medijan, gornji kvartil, najveća vrijednost). Dakle, uspoređuje se udaljenost gornjeg kvartila do najveće vrijednosti s udaljenošću donjeg kvartila do najmanje vrijednosti te koliko ta veličina dobro detektira asimetriju u rasporedu podataka. Promatrajući box-plot, uspoređuje se desni brk s lijevim. Intuitivno, ukoliko je raspored podataka pozitivno asimetričan, omjer $\frac{x_{max}-Q_3}{Q_1-x_{min}}$ bit će veći od 1.

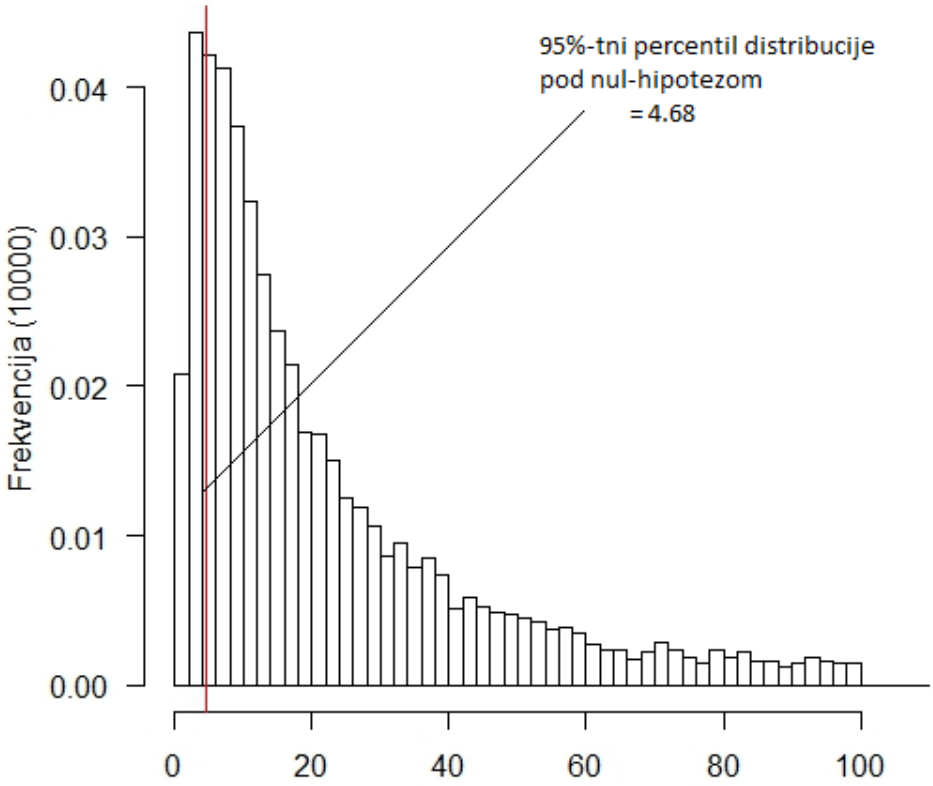
Koristeći postupak naveden u prethodnom potpoglavlju, generiramo 10000 uzoraka duljine 10 iz normalne distribucije s parametrima $\mu = 5$ i $\sigma = 1$ te za svaki od uzoraka računamo omjer $\frac{x_{max}-Q_3}{Q_1-x_{min}}$. Dobivena uzoračka distribucija je jako pozitivno asimetrična, s više od 99% vrijednosti koje se nalaze između 0 i 15.
Slika 7 prikazuje procijenjenu uzoračku distribuciju omjera $\frac{x_{max}-Q_3}{Q_1-x_{min}}$ za vrijednosti između 0 i 15. 95%-tni percentil za tu distribuciju približno iznosi 4.68, stoga svaki omjer veći od 4.68 daje dovoljno uvjerljiv dokaz da je distribucija podataka pozitivno asimetrična, odnosno tada odbacujemo hipotezu $\mathcal{H}_0$ na razini značajnosti 0.05.

Vrijednost testne statistike u originalnom uzorku iz Primjera 2.1 iznosi 4.68. Dakle, na razini značajnosti od 0.05 ipak ne odbacujemo nultu hipotezu.



Slika 7: Distribucija omjera $(x_{max} - Q_3)/(Q_1 - x_{min})$ normalne populacije (n=10). Vrijednosti veće od 15 su isključene iz grafa, ali su uključene u numeričko izračunavanje.

Nadalje, ispitujemo koliko učinkovito ova testna statistika detektira asimetriju, odnosno ispitujemo jakost testa. Generiramo 10000 uzoraka duljine 10 iz jako pozitivno asimetrične distribucije, to jest χ^2 distribucije s jednim stupnjem slobode reskalirane tako da očekivanje bude jednako 5, a standardna devijacija jednaka 1. Za svaki od uzoraka računamo omjer $\frac{x_{max}-Q_3}{Q_1-x_{min}}$. Dobivena uzoračka distribucija navedenog omjera ekstremno je pozitivno asimetrična s nekim vrijednostima omjera većih od 200000, no većina vrijednosti se nalazi između 0 i 100 kako je prikazano na Slici 8.



Slika 8: Distribucija omjera $(x_{max} - Q_3)/(Q_1 - x_{min})$ jako pozitivno asimetrične populacije (n=10). Vrijednosti veće od 100 su isključene iz grafa, ali su uključene u numeričko izračunavanje.

Dakle, Slika 8 pokazuje da se većina vrijednosti u uzoračkoj distribuciji omjera, kada se uzorci generiraju iz pozitivno asimetrične distribucije, nalazi iznad kritične vrijednosti koja iznosi 4.68. Od 10000 uzoraka, njih 86% imaju omjere koje daju uvjerljiv dokaz da je distribucija pozitivno asimetrična. Stoga je jakost ovog testa 0.86 te s obzirom da je dvostruko veća od jakosti dobivene sa statističkom veličinom $\frac{\overline{x}}{Me}$ (0.43), možemo zaključiti da je omjer $\frac{x_{max}-Q_3}{Q_1-x_{min}}$ bolja mjera za utvrđivanje asimetrije podataka. Naravno,

to možda ne bi bila istina ukoliko bi se promijenila veličina uzorka ili jačina pozitivne asimetrije.

Tablica 3 prikazuje 10 statističkih veličina testiranih koristeći metode opisane u prethodnom potpoglavlju i njihove procijenjene jakosti za utvrđivanje asimetrije u rasporedu podataka. Svakoj statističkoj veličini pridruženo je slovo A-J te su rangirane po njihovim jačinama. Statističke veličine od A do G definirane su koristeći karakterističnu petorku uzorka te vrijednosti veće od 1 ukazuju na to da je raspored podataka pozitivno asimetričan.

		Jako	Srednje	Slabo	
Ime	Statistika	Jakost (Rang)	Jakost (Rang)	Jakost (Rang)	Prosječan rang
A	$\frac{\overline{x}}{Me}$	0.429 (7)	0.211 (6)	0.097 (5)	6
B	$\frac{x_{max}-Me}{Me-x_{min}}$	0.838 (2)	0.305 (2)	0.100 (4)	2.67
C	$\frac{Q_3-Me}{Me-Q_1}$	0.253 (9)	0.092 (10)	0.061 (10)	9.67
D	$\frac{x_{max}-Q_3}{Q_1-x_{min}}$	0.857 (1)	0.261 (5)	0.089 (6)	4
E	$\frac{\frac{1}{2}(x_{min}+x_{max})}{Me}$	0.572 (5)	0.309 (1)	0.118 (1)	2.33
F	$\frac{\frac{1}{2}(Q_1+Q_3)}{Me}$	0.125 (10)	0.100 (8)	0.069 (8)	8.67
G	$\frac{\frac{1}{2}(x_{min}+x_{max})}{\frac{1}{2}(Q_1+Q_3)}$	0.504 (6)	0.280 (3)	0.112 (2)	3.67
H	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^3}{\sigma^3}$	0.674 (3)	0.272 (4)	0.104 (3)	3.33
I	$\frac{3(\overline{x}-Me)}{\sigma}$	0.634 (4)	0.197 (7)	0.082 (7)	6
J	$\frac{(Q_3-Me)-(Me-Q_1)}{Q_3-Q_1}$	0.261 (8)	0.094 (9)	0.061 (9)	8.67

Tablica 3: 10 testnih statistika, procjena jakosti kada se uzorkuje iz distribucija s različitim jačinama asimetrije, rang i prosječan rang (rang 1 = najveća jakost).

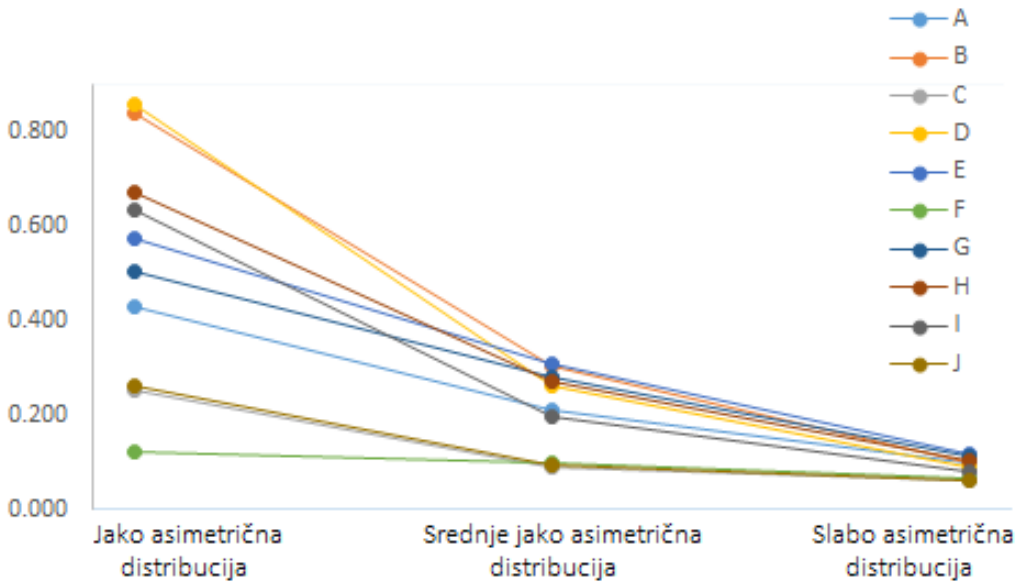
Nadalje, statistike H, I i J su standardne mjere asimetrije. Statistika H je koeficijent asimetrije i uključuje računanje trećeg momenta oko sredine. Statistika I je Personova mjera asimetrije, a statistika J Bowleyeva mjera

asimetrije. Svaka od ovih statistika jednaka je 0 ukoliko se radi o simetričnoj distribuciji, a veća od 0 ako je distribucija pozitivno asimetrična.

U prvom stupcu Tablice 3 uzorkovano je iz jako pozitivno asimetrične distribucije (χ^2 distribucije s jednim stupnjem slobode, reskalirane tako da očekivanje bude jednako 5, a standardna devijacija jednaka 1). U drugom stupcu uzorkovano je iz srednje jake pozitivne distribucije, točnije iz χ^2 distribucije s 5 stupnjeva slobode reskalirane tako da očekivanje bude jednako 5, a standardna devijacija 1. U trećem stupcu uzorkovano je iz slabo jake pozitivne distribucije, odnosno iz χ^2 distribucije s 20 stupnjeva slobode, također reskalirane tako da očekivanje bude jednako 5, a standardna devijacija 1.

Dakle, za svaku od 10 navedenih statistika, generirano je 10000 uzoraka duljine 10 iz svake od tri pozitivno asimetrične distribucije (jake, srednje i slabe jakosti) te je izračunat udio slučajeva kada je uzorak bio veći od 95%-tnog percentila.
Procjene jakosti (i rangova) statistika prikazani su Tablicom 3 i Slikom 9.

Primijetimo, u Tablici 3 (prvi stupac) statistike imaju poprilično velik raspon ranga jakosti u slučaju kada se uzorkuje iz jako pozitivno asimetrične distribucije. Statistika F ima samo 13% šanse detektirati asimetriju, dok statistika D ima 85% šanse. Također, primijetimo da standardne mjere asimetrije dobro detektiraju asimetriju, iako nemaju toliko veliku jakost kao statistike B ili D.



Slika 9: Jakost testnih statistika za svaki tip distribucije.

Kako je bilo i očekivano, jakost statistika se smanjuje kako distribucija postaje slabije pozitivno asimetrična, a više simetrična. Također, može se primijetiti da vrijednosti jakosti konvergiraju k 0.05, razlog tomu je 95%-tni percentil korišten kao kritična vrijednost. Ukoliko je nulta hipoteza istinita, svaka statistika bi trebala ići iznad te vrijednosti, odnosno nalaziti se u području odbacivanja u otprilike 5% slučajeva, čak i kad je uzorkovano iz normalne distribucije.

Nadalje, statistika D, koja je imala najveću jakost u jako pozitivno asimetričnoj distribuciji, pada na šesto mjesto u slabo pozitivno asimetričnoj distribuciji. Statistika B, koja je u početku bila na drugom mjestu, pada na četvrto mjesto.
Koristeći prosječan rang kao ukupnu mjeru jakosti prikazanu u Tablici 3, čini se da od ukupno 10 testiranih statistika, statistike B i E imaju najbolji uspjeh u detektiranju asimetrije u rasporedu podataka te obje statistike sadrže iste komponente: minimalnu vrijednost, medijan i maksimalnu vrijednost. Na drugom kraju spektra, tri najgore statistike (C, F i J) sadrže kombinaciju donjeg kvartila, medijana i gornjeg kvartila, što za posljedicu ima slabu detekciju asimetrije u rasporedu podataka te potom i loš rang u Tablici 3. Obrazloženje lošeg ranga je činjenica da kvartili ignoriraju 25% vanjskog ruba distribucije gdje je asimetrija najočitija.

2.3 Veličina uzorka

U ovom potpoglavlju ispitujemo mijenja li se jakost testa povećanjem veličine uzorka.

Koristit ćemo testnu statistiku E,

$$E = \frac{\frac{1}{2}(x_{min} + x_{max})}{Me},$$

te veličine uzoraka 10 i 100.

Generiramo 10000 uzoraka duljina 10 i 100 iz normalne distribucije s parametrima $\mu = 5$ i $\sigma = 1$, za svaki od uzoraka računamo vrijednost statistike te za svaku veličinu uzorka odredimo kritičnu vrijednost. Nadalje, generiramo 10000 uzoraka duljina 10 i 100 iz pozitivno asimetrične distribucije s parametrima $\mu = 5$ i $\sigma = 1$, računamo vrijednosti statistike te koliko puta je ta vrijednost bila iznad kritične vrijednosti.

Tablica 4 prikazuje rezultate korištenja testne statistike E za veličine uzoraka 10 i 100.

	Jakost (jako asimetrična distribucija)	Jakost (srednje jako asimetrična distribucija)	Jakost (slabo asimetrična distribucija)
n=10	0.56	0.30	0.12
n=100	1.00	0.98	0.41

Tablica 4: Usporedba jakosti testne statistike E za različite veličine uzorka

Dakle, povećanjem uzorka definitivno se povećava jakost statistike. Ako imamo veliki uzorak (npr. uzorak duljine 100), općenito se možemo osloniti na centralno granični teorem pri testiranju aritmetičke sredine bez obaziranja na oblik distribucije.

Kao još jedan primjer uzeli smo testnu statistiku B,

$$B = \frac{x_{max} - Me}{Me - x_{min}},$$

te prethodno opisanim postupkom dobili rezultate prikazane u Tablici 5.

	Jakost (jako asimetrična distribucija)	Jakost (srednje jako asimetrična distribucija)	Jakost (slabo asimetrična distribucija)
n=10	0.83	0.28	0.09
n=100	1.00	0.99	0.45

Tablica 5: Usporedba jakosti testne statistike B za različite veličine uzorka

Na temelju dobivenih rezultata, možemo potvrditi prethodni zaključak da se povećanjem uzorka povečava i jakost testne statistike.

3 Testovi omjera vjerodostojnosti za testiranje simetrije

U ovom poglavlju proučavamo dva testa omjera vjerodostojnosti za testiranje simetrije diskretne distribucije oko nekog parametra θ u odnosu na jednostrane alternative. Ukratko ćemo opisati metodu omjera vjerodostojnosti te potom opisati navedene testove.

Metoda omjera vjerodostojnosti

Procjenitelj metodom maksimalne vjerodostojnosti (MLE) u nekom statističkom modelu s klasom dopuštenih distribucija $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, pri čemu θ može biti i vektorski parametar, odnosno $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k,\ k \in \mathbb{N}$, određuje se na sljedeći način:
Ako je riječ o klasi diskretnih distribucija, tada se funkcija vjerodostojnosti definira formulom

$$\mathbf{L}(\theta) = P(X_1 = x_1 \mid \theta) \cdot \ldots \cdot P(X_n = x_n \mid \theta), \quad \theta \in \Theta, \tag{6}$$

a ako je riječ o klasi neprekidnih distribucija, definira se formulom

$$\mathbf{L}(\theta) = f(x_1 \mid \theta) \cdot \ldots \cdot f(x_n \mid \theta), \quad \theta \in \Theta. \tag{7}$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \in \Theta$, za koju je

$$\mathbf{L}(\hat{\theta}) = \max_{\theta \in \Theta} \mathbf{L}(\theta)$$

zovemo procjena metodom maksimalne vjerodostojnosti.
Statistika $\hat{\theta}(X_1, \ldots, X_n)$ je *procjenitelj metodom maksimalne vjerodostojnosti.*

Opće načelo za definiranje kritičnog područja $\mathcal{C}$ nekog testa sastoji se u tome da se u $\mathcal{C}$ uključe one točke $(x_1, \ldots, x_n) \in \mathbb{R}^n$ kojima pripada mala vjerojatnost pod uvjetom da je $\mathcal{H}_0$ istinita u usporedbi s vjerojatnošću te točke uz uvjet da je istinita alternativna hipoteza $\mathcal{H}_1$. To je načelo bilo jednostavno operacionalizirati u slučaju jednostavnih hipoteza, što je i učinjeno Neyman-Pearsovom lemom, gdje je ključnu ulogu u definiranju najboljeg kritičnog područja imao omjer $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta_1)}$. Mala vrijednost tog omjer u nekoj

točki $(x_1, \ldots, x_n) \in \mathbb{R}^n$ upućivala je na veliku mogućnost da hipoteza $\mathcal{H}_0$ nije istinita, što bi značilo da tu točku treba uključiti u kritično područje testa.

Ako je riječ o testiranju jednostavne hipoteze, $\mathcal{H}_0 : \theta = \theta_0$, prema složenoj alternativnoj hipotezi, $\mathcal{H}_1 : \theta \in \Theta_1$, tada za svaki $\theta \in \Theta_1$ funkcija vjerodostojnosti ima određenu vrijednost $\mathbf{L}(\theta)$, tako da se ovdje ne može govoriti o omjeru $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta)}$ kao konstantnoj veličini pridruženoj točki $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Taj omjer je sada funkcija nepoznatog parametra θ. Međutim, ako postoji

$$\max_{\theta \in \Theta} \mathbf{L}(\theta) = \mathbf{L}(\theta_1),$$

tada je vrijednost omjera $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta_1)}$ u točki $(x_1, \ldots, x_n)$ određeni fiksirani broj te ako je taj broj dovoljno malen, onda to upućuje na to da tu točku treba uključiti u kritično područje testa.
Primijetimo najprije da je $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta_1)} \leq 1$. Ako je $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta_1)} \approx 0$, tada to označava da je vjerojatnost $\mathbf{L}(\theta_0)$ da se dobije baš izmjereni niz podataka $x_1, \ldots, x_n$, uz uvjet da nepoznati parametar ima vrijednost θ_0, zanemarivo mala u odnosu na najveću moguću vjerojatnost da se dobije taj niz podataka pri variranju parametra θ po cijelom skupu Θ dopuštenih vrijednosti. Nadalje, za $\theta_1 \in \Theta$ je ona vrijednost parametra nepoznate distribucije za koju dobiveni niz podataka ima najveću vjerojatnost pa je $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta_1)}$ najmanja vrijednost omjera $\frac{\mathbf{L}(\theta_0)}{\mathbf{L}(\theta)}$ pri variranju parametra θ po skupu Θ.

Definicija 3.1. *Veličina*

$$\lambda(x_1, \ldots, x_n) = \frac{\boldsymbol{L}(\theta_0)}{\max_{\theta \in \Theta} \boldsymbol{L}(\theta)} = \frac{\boldsymbol{L}(\theta_0)}{\boldsymbol{L}(\theta_1)},$$

zove se omjer vjerodostojnosti *u točki* $(x_1, \ldots, x_n) \in \mathbb{R}^n$.
Ako je i nul-hipoteza složena hipoteza, to jest $\mathcal{H}_0 : \theta \in \Theta_0$, *i ako postoji*

$$\max_{\theta \in \Theta_0} \boldsymbol{L}(\theta) = \boldsymbol{L}(\theta_0),$$

onda se omjer vjerodostojnosti definira formulom

$$\lambda(x_1, \ldots, x_n) = \frac{\max_{\theta \in \Theta_0} \boldsymbol{L}(\theta)}{\max_{\theta \in \Theta} \boldsymbol{L}(\theta)} = \frac{\boldsymbol{L}(\theta_0)}{\boldsymbol{L}(\theta_1)},$$

Očigledno je $\lambda(x_1, \ldots, x_n) \leq 1$.
Ako se dobije $\lambda(x_1, \ldots, x_n) \approx 1$, tada točka $(x_1, \ldots, x_n)$ ne bi trebala pripadati kritičnom području. Ako je $\lambda(x_1, \ldots, x_n) \approx 0$, tada niz izmjerenih

podataka $x_1, \ldots, x_n$ upućuje na činjenicu da je njegova maksimalna vrijednost $\mathbf{L}(\theta_0)$, uz uvjet da je hipoteza $\mathcal{H}_0$ istinita, zanemarivo mala u odnosu na njegovu maksimalno moguću vrijednost $\mathbf{L}(\theta_1)$ i da bi stoga točka $(x_1, \ldots, x_n)$ trebala pripadati kritičnom području hipoteze $\mathcal{H}_0$.

Sada se čini razumnim smatrati da će se dobiti dobar test ako se kritično područje $\mathcal{C}$ odabere tako da se u njega uključe one točke iz prostora $\mathbb{R}^n$ (prostor vrijednosti slučajnog uzorka) za koje je pripadni omjer vjerodostojnosti manji od zadanog broja c $(0 < c \leq 1)$.

Definicija 3.2. *Ako kritično područje $\mathcal{C}$, za testiranje parametarske hipoteze $\mathcal{H}_0 : \theta \in \Theta_0$, u odnosu na alternativnu hipotezu $\mathcal{H}_1 : \theta \in \Theta_1$, ima oblik*

$$\mathcal{C} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \ : \ \lambda(x_1, \ldots, x_n) \leq c\},$$

onda se kaže da je test dobiven metodom omjera vjerodostojnosti *(LR-test).*

Zajednička pretpostavka koja se temelji na mnogim statističkim analizama jest da je osnovna distribucija simetrična. Valjanost nekih često korištenih procedura ovisi upravo o toj pretpostavci.

Neka je X slučajna varijabla na $\mathbb{R}^1$ s funkcijom distribucije $F(x)$.

Definicija 3.3. *Distribucija $F(x)$ je simetrična oko ishodišta ako vrijedi*

$$F(x) + F(-x) = 1, \ \forall x. \tag{8}$$

Kao alternativu simetriji oko 0, definirat ćemo pozitivnu asimetriju (eng. *positive biasedness*) u različitim značenjima te ćemo tu činjenicu označiti s $X \succ 0$ $(\mathcal{B}_i)$ ili $F(\cdot) \succ 0$ $(\mathcal{B}_i)$ (Yanagimoto i Sibuya).

Definicija 3.4.

- $X \succ 0$ $(\mathcal{B}_0)$ *ako i samo ako* $\mathbb{P}(X > 0) \geqq \mathbb{P}(X < 0)$*, ili ekvivalentno* $1 - F(0) \geqq F(0 - 0)$*.*
- $X \succ 0$ $(\mathcal{B}_1)$ *ako i samo ako* $\mathbb{P}(X > \alpha_1) \geqq \mathbb{P}(X < -\alpha_1), \ \forall \alpha_1 > 0$*, ili ekvivalentno* $F(x) + F(-x) \leqq 1, \ \forall x \geqq 0$*.*
- $X \succ 0$ $(\mathcal{B}_2)$ *ako i samo ako* $\mathbb{P}(\alpha_2 \geqq X > \alpha_1) \geqq \mathbb{P}(-\alpha_1 > X \geqq -\alpha_2), \ \forall \alpha_2 > \alpha_1 > 0$*, ili ekvivalentno* $F(x + y) - F(x) \geqq F(-x) - F(-x - y), \ \forall x, y \geqq 0$*.*

- $X \succ 0$ $(\mathcal{B}_3)$ *ako i samo ako* $\frac{\mathbb{P}(\alpha_3 \geqq X > \alpha_2)}{\mathbb{P}(\alpha_2 \geqq X > \alpha_1)} \geqq \frac{\mathbb{P}(-\alpha_2 > X \geqq -\alpha_3)}{\mathbb{P}(-\alpha_1 > X \geqq -\alpha_2)}$, $\forall \alpha_3 > \alpha_2 > \alpha_1 > 0$ *tako da su nazivnici pozitivni, ili ekvivalentno ako je* $\frac{F(x+y)-F(y)}{F(-y)-F(-x-y)}$ *neopadajuća u* $x > 0$ *i* $y > 0$.
- $X \succ 0$ $(\mathcal{B}_4)$ *ako i samo ako* $\frac{\mathbb{P}(\alpha_3 \geqq X > \alpha_2)}{\mathbb{P}(\alpha_2 \geqq X > \alpha_1)} \geqq \frac{\mathbb{P}(-\alpha_2 > X \geqq -\alpha_3)}{\mathbb{P}(-\alpha_1 > X \geqq -\alpha_2)}$, $\forall \alpha_3 > \alpha_2 > \alpha_1$ *tako da su nazivnici pozitivni, ili ekvivalentno ako je* $\frac{F(x+y)-F(y)}{F(-y)-F(-x-y)}$ *neopadajuća u* $x > 0$ *i* y.

Negativna asimetrija se definira na sličan način i označava s $X \prec 0$ $(\mathcal{B}_i)$.

Kako je simetrija (oko 0) ekvivalentna izrazu (8), mnogi su testovi zasnovani u terminima empirijske kumulativne distribucije.

Ako nam je cilj odbaciti hipotezu o simetriji, tada je važno poznavati strukturu distribucije koja nas vodi do odbacivanja.
Ekvivalentno Definiciji 3.3, kažemo da slučajna varijabla X ima simetričnu distribuciju (oko 0) ako i samo ako X i $-X$ imaju jednaku distribuciju.
Možemo razmatrati različite jednostrane alternative simetriji uzimajući u obzir različite tipove stohastičkog poretka između slučajnih varijabli X i $-X$. Najčešće, pozitivnu asimetriju povezujemo sa situacijom kada je X stohastički veći od $-X$ što povlači $\mathbb{E}[g(X)] \geq \mathbb{E}[g(-X)]$ za sve neopadajuće funkcije g.

Razmatramo problem kada su podaci diskretni ili grupirani te bez smanjenja općenitosti pretpostavljamo da je $\theta = 0$.

Neka slučajna varijabla X poprima $(2k+1)$ vrijednosti $-k, -(k-1), \ldots, -1, 0, 1, \ldots, (k-1), k$ s pripadajućim vjerojatnostima redom $p_{-k}, p_{-(k-1)}, \ldots, p_{-1}, p_0, p_1, \ldots, p_{k-1}, p_k$. Označimo s $\boldsymbol{p}$ vektor dimenzije $(2k+1)$ koji sadrži sve $p_i,\ i = 0, \pm 1, \ldots, \pm k$.

Pretpostavimo da imamo slučajan uzorak duljine n iz te populacije. Neka je n_i broj slučajeva kada je slučajna varijabla X poprimila vrijednost i, za $i = 0, \pm 1, \ldots, \pm k$ tako da je $\sum_{i=-k}^{k} n_i = n$.

Testiramo nul-hipotezu o simetriji oko 0:

$$\mathcal{H}_0 :\ p_j = p_{-j}, \quad j = 1, 2, \ldots, k, \tag{9}$$

u odnosu na alternative $\mathcal{H}_1 - \mathcal{H}_0$ i $\mathcal{H}_2 - \mathcal{H}_0$, pri čemu su:

$$\mathcal{H}_1 : \sum_{i=j}^{k} p_i \geq \sum_{i=j}^{k} p_{-i}, \quad j = 1, 2, \ldots, k \tag{10}$$

i

$$\mathcal{H}_2 : p_j \geq p_{-j}, \quad j = 1, 2, \ldots, k. \tag{11}$$

Ako je hipoteza $\mathcal{H}_1$ (10) istinita, kažemo da je X pozitivno asimetrična slučajna varijabla po kriteriju $\mathcal{B}_1$ (tip I), i ako je hipoteza $\mathcal{H}_2$ (11) istinita, kažemo da je X pozitivno asimetrična slučajna varijabla po kriteriju $\mathcal{B}_2$ (tip II).

3.1 Testiranje $\mathcal{H}_0$ u odnosu na $\mathcal{H}_1 - \mathcal{H}_0$

Određujemo procjenitelje maksimalne vjerodostojnosti (MLE) za vektor $\boldsymbol{p}$ pod ograničenjima obje hipoteze $\mathcal{H}_0$ i $\mathcal{H}_1$. Funkcija vjerodostojnosti proporcionalna je

$$L(\boldsymbol{p} \mid \boldsymbol{n}) = p_0^{n_0} \prod_{i=1}^{k} [p_{-i}^{n_{-i}} p_i^{n_i}]. \tag{12}$$

Neograničen MLE za p_i je $\hat{p}_i = n_i/n$, $i = 0, \pm 1, \ldots, \pm k$. MLE za $\boldsymbol{p}$ u uvjetima istinitosti hipoteze $\mathcal{H}_0$ dan je s

$$\hat{p}_{-i}^{(0)} = \hat{p}_i^{(0)} = \frac{n_{-i} + n_i}{2n}, \quad i = 1, 2, \ldots, k \tag{13}$$

i

$$\hat{p}_0^{(0)} = \hat{p}_0 = \frac{n_0}{n}.$$

U uvjetima istinitosti alternativne hipoteze $\mathcal{H}_1$, ograničenja na vektor $\boldsymbol{p}$ su

$$\sum_{i=j}^{k} p_i \geq \sum_{i=j}^{k} p_{-i}, \quad j = 1, 2, \ldots, k.$$

Ta ograničenja ekvivalentna su

$$\sum_{i=j}^{k} p_i \geq \sum_{i=j}^{k} p_{-i}, \quad j = -k, -(k-1), \ldots, k, \tag{14}$$

što zapisujemo kao $\boldsymbol{p} \gg \boldsymbol{p}'$, gdje $\boldsymbol{p}'$ označava obrnuti $(2k+1)$ dimenzionalni vektor $(p_k, p_{k-1}, \ldots, p_1, p_0, p_{-1}, \ldots, p_{-(k-1)}, p_{-k})$. Esencijalno, ovo je dvostrani problem procjenjivanja $\boldsymbol{p}$ i $\boldsymbol{p}'$ unutar stohastičkog poretka $\boldsymbol{p} \gg \boldsymbol{p}'$. Dvostrani problem proučavali su Barlow i Brunk (1972) te se ovdje mogu primijeniti njihovi rezultati. Pretpostavljamo da su operacije množenja i dijeljenja vektora definirane na sljedeći način:
$x, y \in \mathbb{R}^n,\ x \cdot y = (x_1 \cdot y_1, \ldots, x_n \cdot y_n)$ *i* $\frac{x}{y} = (\frac{x_1}{y_1}, \ldots, \frac{x_n}{y_n})$.

Nadalje, maksimiziranje produkta vjerodostojnosti, $L(\boldsymbol{p} \mid \boldsymbol{n})$, unutar $\mathcal{H}_1$ ekvivalentno je maksimiziranju

$$L^2(\boldsymbol{p} \mid \boldsymbol{n}) = \prod_{i=-k}^{k} p_i^{n_i} \prod_{i=-k}^{k} p_{-i}^{n_{-i}}. \tag{15}$$

Teorem 3.5. *Ako je* $\hat{p}_i > 0$ *za* $i = -k, -(k-1), \ldots, -1, -0, 1, \ldots, k-1, k$, *tada je MLE za* $\boldsymbol{p}$, *u uvjetima istinitosti hipoteze* $\mathcal{H}_1$, *dan s*

$$\hat{\boldsymbol{p}}^{(1)} = \hat{\boldsymbol{p}} E_{\hat{p}}(\frac{\hat{\boldsymbol{p}} + \hat{\boldsymbol{p}}'}{2\hat{\boldsymbol{p}}} \mid I), \tag{16}$$

gdje $E_{\boldsymbol{w}}(\boldsymbol{x} \mid I)$ *označava najmanju kvadratnu projekciju s težinama* $\boldsymbol{w}$ *vektora* $\boldsymbol{x}$ *na konus* $I = \{\boldsymbol{x} : x_{-k} \leq \ldots \leq x_{-1} \leq x_0 \leq x_1 \leq \ldots \leq x_k\}$ *neopadajućih vektora.*

Dokaz. S obzirom da vrijedi sljedeće

$$\sup_{\boldsymbol{p} \gg \boldsymbol{p}'} L^2(\boldsymbol{p} \mid \boldsymbol{n}) \leq \sup_{\boldsymbol{p} \gg \boldsymbol{q}} L(\boldsymbol{p} \mid \boldsymbol{n}) L(\boldsymbol{q} \mid \boldsymbol{n}')$$

te je rješenje desne strane dano s

$$\hat{\boldsymbol{p}}^{(1)} = \hat{\boldsymbol{p}} E_{\hat{\boldsymbol{p}}}(\frac{\hat{\boldsymbol{p}} + \hat{\boldsymbol{p}}'}{2\hat{\boldsymbol{p}}} \mid I)$$

i

$$\hat{\boldsymbol{q}}^{(1)} = \hat{\boldsymbol{p}}' E_{\hat{\boldsymbol{p}}'}(\frac{\hat{\boldsymbol{p}}' + \hat{\boldsymbol{p}}}{2\hat{\boldsymbol{p}}'} \mid A),$$

pri čemu je A konus neopadajućih vektora, rezultat direktno slijedi provjeravanjem da je $\hat{q}^{(1)} = \hat{p}^{(1)'}$. □

Koristeći (13) i (16), možemo dobiti procjenitelje maksimalne vjerodostojnosti za funkciju distribucije od X pod ograničenjima hipoteza $\mathcal{H}_0$ i $\mathcal{H}_1$. Slijedi da test omjera vjerodostojnosti za testiranje $\mathcal{H}_0$ u odnosu na $\mathcal{H}_1 - \mathcal{H}_0$ odbacuje $\mathcal{H}_0$ za dovoljno velike vrijednosti od

$$T_1 = 2 \sum_{i=-k}^{k} n_i \{\ln \hat{p}_i^{(1)} - \ln \hat{p}_i^{(0)}\}. \tag{17}$$

Asimptotska nul-distribucija statistike T_1

T_1 je statistika testa omjera log-vjerodostojnosti za navedeni problem. Pokazujemo da je distribucija od T_1 tipa $\bar{\chi}^2$ (mješavina nezavisnih χ^2 distribucija).

Proširujući $\ln\hat{p}_i^{(1)}$ i $\ln\hat{p}_i^{(0)}$ oko $\hat{p}_i$ pomoću Taylorovog teorema s ostatkom drugog stupnja i koristeći činjenicu da je $\sum_{i=-k}^{k} \hat{p}_i^{(1)} = \sum_{i=-k}^{k} \hat{p}_i^{(0)} = 1$, slijedi da je pod $\mathcal{H}_0$ (uz pretpostavku $\hat{p}_i > 0,\ \forall i$),

$$T_1 = \sum_{i=-k}^{k} n_i [\frac{1}{\alpha_i^2}(\hat{p}_i^{(0)} - \hat{p}_i)^2 - \frac{1}{\beta_i^2}(\hat{p}_i^{(1)} - \hat{p}_i)^2],$$

pri čemu su α_i i β_i koeficijenti iz Taylorova proširenja i za $i = -k, \ldots, k$, α_i je uvijek između $\hat{p}_i$ i $\hat{p}_i^{(0)}$, a β_i između $\hat{p}_i$ i $\hat{p}_i^{(1)}$ te konvergiraju gotovo sigurno k p_i (pod uvjetima $\mathcal{H}_0$).

Posebno, vrijedi da je $\hat{\boldsymbol{p}}^{(0)} - \hat{\boldsymbol{p}} = (\hat{\boldsymbol{p}}' - \hat{\boldsymbol{p}})/2$ i $\hat{\boldsymbol{p}}^{(1)} - \hat{\boldsymbol{p}} = \hat{\boldsymbol{p}} E_{\hat{\boldsymbol{p}}}(\frac{\hat{\boldsymbol{p}}' - \hat{\boldsymbol{p}}}{2\hat{\boldsymbol{p}}} \mid I)$, stoga možemo zapisati

$$T_1 = n \sum_{i=-k}^{k} \hat{p}_i \ [\ \frac{1}{\alpha_i^2}\ \hat{p}_i^2\ (\frac{\hat{p}_i{}' - \hat{p}_i}{2})^2 - \frac{1}{\beta_i^2}\ \hat{p}_i^2\ E_{\hat{p}}(\frac{\hat{\boldsymbol{p}}' - \hat{\boldsymbol{p}}}{2\hat{\boldsymbol{p}}} \mid I)_i^2\ ]$$

$$= (\frac{1}{4}) \sum_{i=-k}^{k} \hat{p}_i \ [\ \frac{1}{\alpha_i^2} \ \hat{p}_i^2 \ \psi_i^2 - \frac{1}{\beta_i^2} \ \hat{p}_i^2 \ E_{\hat{p}}(\boldsymbol{\psi} \mid I)_i^2 \ ],$$

pri čemu je $\boldsymbol{\psi} = \sqrt{n}(\hat{\boldsymbol{p}}' - \hat{\boldsymbol{p}})/\hat{\boldsymbol{p}}$. Po centralnom graničnom teoremu, slučajni vektor $\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ po distribuciji konvergira k $\boldsymbol{p}(\boldsymbol{U} - \bar{U}\boldsymbol{E})$, gdje su $U_{-k}, U_{-(k-1)}, \ldots, U_k$ nezavisne normalne slučajne varijable s očekivanjem 0 i odgovarajućim varijancama $p_{-k}^{-1}, p_{-(k-1)}^{-1}, \ldots, p_k^{-1}$, $\bar{U} = \sum_{i=-k}^{k} p_i U_i$ i $\boldsymbol{E} = (1, 1, \ldots, 1)^T$.

S obzirom da pretpostavljamo da je hipoteza $\mathcal{H}_0$ istinita, možemo zapisati

$$\boldsymbol{\psi} = \frac{\sqrt{n}[(\hat{\boldsymbol{p}} - \boldsymbol{p})' - (\hat{\boldsymbol{p}} - \boldsymbol{p})]}{\hat{\boldsymbol{p}}}$$

$$\xrightarrow{\mathcal{L}} \frac{\boldsymbol{p}'(\boldsymbol{U} - \bar{U}\boldsymbol{E})' - \boldsymbol{p}(\boldsymbol{U} - \bar{U}\boldsymbol{E})}{\boldsymbol{p}}$$

$$= (\boldsymbol{U}' - \mathrm{U}) = \boldsymbol{V}.$$

Nadalje, neprekidnost od $E_{\hat{\boldsymbol{p}}}(\boldsymbol{\psi} \mid I)$ u $\hat{\boldsymbol{p}}$ i $\boldsymbol{\psi}$ povlači

$$T_1 \xrightarrow{\mathcal{L}} (\frac{1}{4}) \sum_{i=-k}^{k} p_i \ [\ V_i^2 - E_{\boldsymbol{p}}(\boldsymbol{V} \mid I)_i^2 \ ]. \tag{18}$$

Primijetimo da je $V_0 = 0$ i $V_{-i} = -V_i$, što slijedi iz pretpostavke da pod uvjetima hipoteze $\mathcal{H}_0$ vrijedi $p_i = p_{-i}$ i maxmin formule za $E_{\boldsymbol{p}}(\boldsymbol{V} \mid I)$ pri čemu su $E_{\boldsymbol{p}}(\boldsymbol{V} \mid I)_0 = 0$ i $E_{\boldsymbol{p}}(\boldsymbol{V} \mid I)_{-i} = -E_{\boldsymbol{p}}(\boldsymbol{V} \mid I)_i$.

Neka su $\boldsymbol{V}_r$ i $\boldsymbol{p}_r$ restrikcije od $\boldsymbol{V}$ i $\boldsymbol{p}$ na $\{1, \ldots, k\}$ i neka je $J = \{\boldsymbol{x} = (x_1, \ldots, x_k) \ : \ 0 \leq x_1 \leq x_2 \ldots \leq x_k\}$.
Tada je

$$E_p(\boldsymbol{V} \mid I)_i = \begin{cases} (E_{p_r}(\boldsymbol{V}_r \mid J)_i, & i = 1, 2, \ldots, k \\ 0, & i = 0 \\ -E_{p_r}(\boldsymbol{V}_r \mid J)_{-i}, & i = -k, \ldots, -1 \end{cases}$$

Dakle, sada imamo:

$$\begin{aligned} T_1 &\xrightarrow{\mathcal{L}} \textstyle\sum_{i=1}^{k} [V_i^2 - E_p(\boldsymbol{V}_r \mid J)_i^2](\frac{p_i}{2}) \\ &= \textstyle\sum_{i=1}^{k} [V_i - E_{p_r}(\boldsymbol{V}_r \mid J)_i]^2(\frac{p_i}{2}) \end{aligned} \tag{19}$$

Primijetimo da su $V_1, V_2, \ldots, V_k$ nezavisne, normalne slučajne varijable s očekivanjem 0 i varijancom $(2/p_i)$. Asimptotska distribucija od T_1 je tipa $\bar{\chi}^2$ koja ovisi o konusu J i o nepoznatim vrijednostima parametara $p_0, p_1, \ldots, p_k$. Sljedećim teoremom dana je najmanje povoljna distribucija.

Teorem 3.6. *Ako $\boldsymbol{p}$ zadovoljava $\mathcal{H}_0$ i ako je $p_i > 0,\ i = -k, \ldots, k$, tada za svaki realan broj t vrijedi:*

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{p}}[T_1 \geq t] = \sum_{l=0}^{k} p(l, k, \boldsymbol{p}_r)\ \mathbb{P}[\chi^2_{k-l} \geq t] \tag{20}$$

pri čemu je $p(0, k, \boldsymbol{p}_r)$ vjerojatnost da je $E_{\boldsymbol{p}_r}(\boldsymbol{V}_r \mid J)$ jednaka nuli i $p(l, k, \boldsymbol{p}_r)$ za $l = 1, 2, \ldots, k$ vjerojatnost da $E_{\boldsymbol{p}_r}(\boldsymbol{V}_r \mid J)$ ima l različitih ne-nul vrijednosti.
χ^2_λ je χ^2-slučajna varijabla s λ stupnjeva slobode ($\chi^2_0 \equiv 0$).
Nadalje,

$$\sup_{\boldsymbol{p}} \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{p}}[T_1 \geq t] = \frac{1}{2}\ \mathbb{P}[\chi^2_{k-1} \geq t] + \frac{1}{2}\ \mathbb{P}[\chi^2_k \geq t]. \tag{21}$$

Vrlo je vjerojatno da će test temeljen na najmanje povoljnoj distribuciji, danoj u Teoremu 3.6, biti konzervativan (ovisno o pravim vrijednostima parametara $p_1, p_2, \ldots, p_k$). Međutim, ako vrijednosti parametara $p_1, p_2, \ldots, p_k$ ne variraju previše (na primjer ako je omjer najveće i najmanje vrijednosti p_i manji od 3), tada će kritična vrijednost testa temeljenog na jednakim

težinama ($p_1 = p_2 = \ldots = p_k$) imati značajnu razinu razumno blizu opaženoj vrijednosti.

Ako imamo dodatnu informaciju da vrijedi $p_1 \geq p_2 \geq \ldots \geq p_k$, tada je najmanje povoljna distribucija dana s:

$$\sum_{l=0}^{k} \binom{k}{l} (\frac{1}{2})^k \, \mathbb{P}[\chi_l^2 \geq t] \tag{22}$$

Kritična vrijednost odabrana iz ove distribucije rezultirat će manje konzervativnim testom u odnosu na kritičnu vrijednost dobivenu iz najmanje povoljne distribucije dane izrazom (21).

Druga alternativa jest aproksimacija $\mathbb{P}_{\mathbf{p}}[T_1 \geq t]$ s

$$\sum_{l=0}^{k} p(l, k, \hat{\boldsymbol{p}}^{(0)}) \cdot \mathbb{P}[\chi_{k-l}^2 \geq t],$$

pri čemu je $\hat{\boldsymbol{p}}^{(0)}$ procjenitelj za $\boldsymbol{p}$ dan s (13). Ovaj izraz ima istu asimptotsku distribuciju kao i T_1 u uvjetima hipoteze $\mathcal{H}_0$ i osigurava dobru aproksimaciju.

3.2 Testiranje $\mathcal{H}_0$ u odnosu na $\mathcal{H}_2 - \mathcal{H}_0$

Određujemo procjenitelje maksimalne vjerodostojnosti (MLE) za vektor $\boldsymbol{p}$ pod ograničenjima $\mathcal{H}_0$ i $\mathcal{H}_2$. Kao u prethodnom potpoglavlju, funkcija vjerodostojnosti proporcionalna je

$$L(\boldsymbol{p} \mid \boldsymbol{n}) = p_0^{n_0} \prod_{i=1}^{k} [p_{-i}^{n_{-i}} p_i^{n_i}].$$

Kako bi pronašli procjenitelje maksimalne vjerodostojnosti za $\boldsymbol{p}$, parametriziramo na sljedeći način. Neka su

$$\theta_i = \frac{p_i}{p_{-i} + p_i} \quad i \quad \varphi_i = (p_{-i} + p_i), \quad i = 1, 2, \ldots, k, \tag{23}$$

tako da vrijedi

$$p_i = \theta_i \varphi_i \quad i \quad p_{-i} = \varphi_i (1 - \theta_i), \quad i = 1, 2, \ldots, k. \tag{24}$$

Funkcija vjerodostojnosti u terminima novih parametara proporcionalna je

$$L_0(\boldsymbol{\theta}, \boldsymbol{\varphi} \mid \boldsymbol{n}) = [\prod_{i=1}^{k} \theta_i^{n_i} (1 - \theta_i)^{n_{-i}}][\prod_{i=1}^{k} \varphi_i^{n_i + n_{-i}} (1 - \sum_{i=1}^{k} \varphi_i)^{n_0}]. \tag{25}$$

Procjenitelji maksimalne vjerodostojnosti pod ograničenjima $\mathcal{H}_0$ su

$$\hat{\theta}_i^{(0)} = \frac{1}{2} \quad i \quad \hat{\varphi}_i^{(0)} = \frac{n_i + n_{-i}}{n}, \quad i = 1, 2, \ldots, k. \tag{26}$$

Pod alternativnom hipotezom $\mathcal{H}_2$ vrijedi da je $\theta_i \geq \frac{1}{2}$, $i = 1, 2, \ldots, k$ i nema ograničenja na vrijednosti od φ_i. Tako da je MLE za $\boldsymbol{\varphi}$ isti kao i procjenitelj pod nul-hipotezom $\mathcal{H}_0$:

$$\hat{\varphi}_i^{(2)} = \hat{\varphi}_i^{(0)} = \frac{n_i + n_{-i}}{n}, \quad i = 1, 2, \ldots, k. \tag{27}$$

MLE za $\boldsymbol{\theta}$ dan je s

$$\hat{\theta}_i^{(2)} = (\frac{n_i}{n_i + n_{-i}}) \vee \frac{1}{2}, \tag{28}$$

pri čemu $a \vee b$ ($a \wedge b$) označava maksimnum (minimum) od a i b.

Koristeći (27) i (28), MLE za $\boldsymbol{p}$ u uvjetima istinitosti hipoteze $\mathcal{H}_2$ dan je u sljedećem teoremu.

Teorem 3.7. *Procjenitelj maksimalne vjerodostojnosti za $\boldsymbol{p}$ podvrgnut $\mathcal{H}_2$ dan je s*

$$\hat{p}_i^{(2)} = \begin{cases} (\frac{n_i+n_{-i}}{n})(\frac{n_i}{n_i+n_{-i}} \vee \frac{1}{2}), & i = 1, 2, \ldots, k \\ \frac{n_0}{n}, & i = 0 \\ (\frac{n_i+n_{-i}}{n})(\frac{n_i}{n_i+n_{-i}} \wedge \frac{1}{2}), & i = -k, -(k-1), \ldots, -1 \end{cases} \tag{29}$$

Koristeći (13) i (29), možemo dobiti procjenitelje maksimalne vjerodostojnosti za funkciju distribucije od X pod ograničenjima hipoteza $\mathcal{H}_0$ i $\mathcal{H}_2$. Slijedi da test omjera vjerodostojnosti za testiranje $\mathcal{H}_0$ u odnosu na $\mathcal{H}_2 - \mathcal{H}_0$ odbacuje $\mathcal{H}_0$ za dovoljno velike vrijednosti od

$$T_2 = 2 \sum_{i=-k}^{k} n_i \{\ln \hat{p}_i^{(2)} - \ln \hat{p}_i^{(0)}\}. \tag{30}$$

Asimptotska nul-distribucija statistike T_2

Testna statistika omjera log-vjerodostojnosti za testiranje hipoteze $\mathcal{H}_0$ u odnosu na $\mathcal{H}_2 - \mathcal{H}_0$ dana je s

$$T_2 = 2 \sum_{i=1}^{k} [n_i \{\ln \hat{\theta}_i^{(2)} - \ln(\frac{1}{2})\} + n_{-i} \{\ln(1 - \hat{\theta}_i^{(2)}) - \ln(\frac{1}{2})\}] \tag{31}$$

Istim postupkom kao u Potpoglavlju 3.1, vrijedi sljedeće

$$T_2 \xrightarrow{\mathcal{L}} \sum_{i=1}^{k} [Z_i \vee 0]^2, \tag{32}$$

pri čemu su $Z_1, Z_2, \ldots, Z_k$ nezavisne standardno normalne slučajne varijable.

Teorem 3.8. *U uvjetima istinitosti hipoteze $\mathcal{H}_0$ vrijedi*

$$\lim_{n\to\infty} \mathbb{P}[T_2 \geq t] = \sum_{l=0}^{k} \binom{k}{l} \, (\frac{1}{2})^k \, \mathbb{P}[\chi_l^2 \geq t], \tag{33}$$

za svaki realan broj t ($\chi_0^2 \equiv 0$).

Primijetimo da asimptotska nul-distribucija ne ovisi o $\boldsymbol{p}$.

3.3 Simulacijska studija

U ovom potpoglavlju napravit ćemo simulaciju pomoću koje ćemo usporediti jakosti predloženog testa T_2.

Fokusirat ćemo se na pomaknutu binomnu distribuciju:

$$p_j = \binom{2k}{j+k} p^{j+k}(1-p)^{k-j}, \qquad j = 0, \pm 1, \ldots, \pm k.$$

Neka je $k = 3$, tada imamo ukupno 7 ćelija. Dakle, distribucija je simetrična kada je $p = 0.5$, $\mathcal{H}_0$, dok zadovoljava alternativnu hipotezu $\mathcal{H}_2$ kada je $p > 0.5$.
U našoj simulaciji, duljina uzorka je fiksna, $n = 100$, a broj replikacija 10000. Razina značajnosti je $\alpha = 0.05$.

U prethodnom potpoglavlju opisana je testna statistika T_2 i dana je formulom (31):

$$T_2 = 2\sum_{i=1}^{k}[n_i\{\ln\hat{\theta}_i^{(2)} - \ln(\frac{1}{2})\} + n_{-i}\{\ln(1-\hat{\theta}_i^{(2)}) - \ln(\frac{1}{2})\}].$$

Nadalje, koristit ćemo još dva testa kako bi što bolje usporedili jakost predloženog testa.

Označimo s $X_1, \ldots, X_n$ nezavisne, jednako distribuirane slučajne varijable s navedenom distribucijom. Za pomaknutu binomnu distribuciju, uniformno najjači test jest odbacivanje hipoteze $\mathcal{H}_0$ u korist alternativne hipoteze $\mathcal{H}_2$ ako je suma $\sum_{i=1}^{n} X_i$ prevelika. Taj ćemo test označiti s *UMP* te ćemo ga koristiti kao mjerilo pomoću kojega ćemo bolje usporediti jakost testa T_2. Dakle, testna statistika je sljedećeg oblika:

$$UMP = \sum_{i=1}^{n} X_i.$$

Kritična vrijednost za ovaj test dobivena je na sljedeći način. Ako sa Z označimo binomnu slučajnu varijablu s parametrima $2k = 2 \cdot 3$ i $p = 0.5$, tada možemo zapisati $X = Z - k$, odnosno $X = Z - 3$.

$$\overline{X} \sim N(\mathbb{E}[X],\ \frac{1}{n}\sigma^2(X))$$

$$\frac{\overline{X} - \mathbb{E}[X]}{\sqrt{\sigma^2}} \cdot \sqrt{n} \sim N(0,1)$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i \ - \mathbb{E}[X]}{\sqrt{\sigma^2}} \cdot \sqrt{n} \sim N(0,1)$$

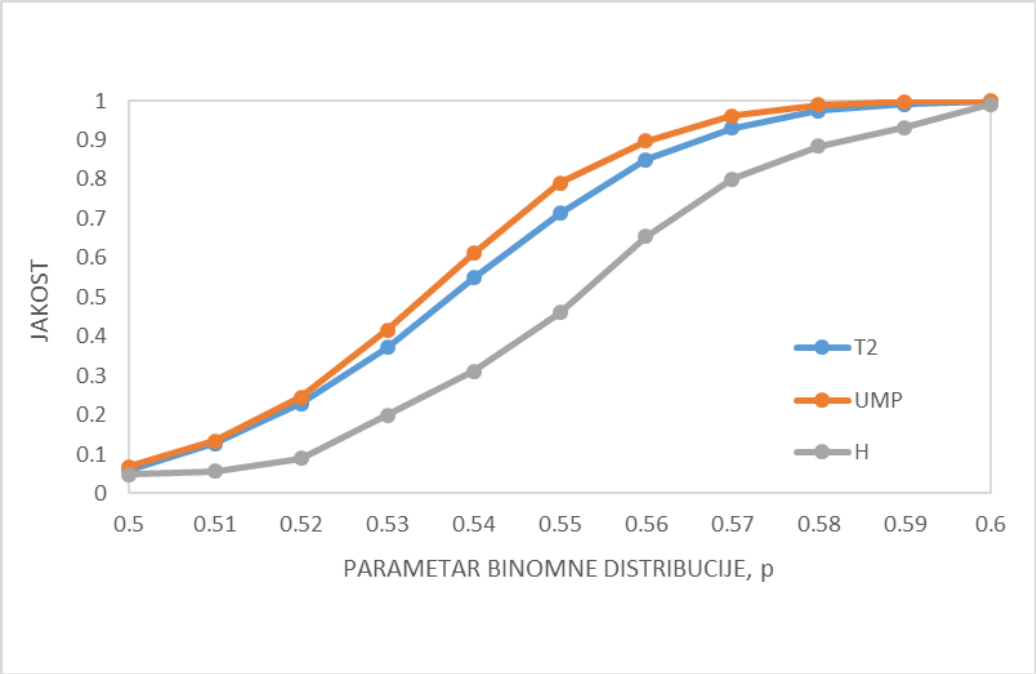
Vrijedi: $\mathbb{E}[X] = \mathbb{E}[Z] - k = 2k \cdot p - k = 0$ i $\sigma^2(X) = \sigma^2(Z) = 2k \cdot p \cdot (1-p) = 1.5$.

Drugi test koji razmatramo jest neograničen test omjera vjerodostojnosti za testiranje simetrije u odnosu na asimetriju (koji je esencijalno ekvivalentan uobičajenom χ^2 testu pripadnosti za simetriju). Označit ćemo ga s H.
Ako je $p = 0.5$, slučajna varijabla $X = Z - k$ je simetrična oko 0. Testna statistika je sljedećeg oblika:

$$H = \sum_{j=-3}^{3} \frac{(f_j - f'_j)^2}{f'_j} \sim \chi^2(k - r - 1),$$

pri čemu su $f'_j = n \cdot p_j$ i $f_j = n_j$ (broj slučajeva kada je slučajna varijabla poprimila vrijednost j), $j = -3, -2, \ldots, 2, 3$.
Dakle, za razliku od testa T_2, navedeni test nema nikakva ograničenja na alternativnu hipotezu. Kritična vrijednost ovog testa je 95%-tni percentil χ^2 distribucije s 3 stupnja slobode.

Generiramo 10000 uzoraka duljine 100 iz pomaknute binomne distribucije s parametrima 7 i $p \in \{0.5, 0.51, 0.52, \ldots, 0,59, 0.60\}$. Za svaki od uzoraka računamo vrijednosti 3 navedene testne statistike (T_2, UMP, H). Jakost pojedinog testa je udio slučajeva kada je vrijednost testne statistike bila iznad kritične vrijednosti.



Slika 10: Krivulje jakosti testova za $k = 3$, $n = 100$, $\alpha = 0.05$.

Na razini značajnosti od 5% dobivamo da su pod nultom hipotezom p-vrijednosti sljedeće: 0.0537 za test T_2, 0.0581 za *UMP* test i 0.0762 za test H. Funkcije jakosti tri navedena testa prikazane su na Slici 10.
Sa Slike 10 jasno se vidi da je krivulja jakosti testa T_2 bliže krivulji jakosti *UMP* testa u odnosu na krivulju jakosti testa H. S obzirom da smo uzeli *UMP* test kao mjerilo, možemo zaključiti da je novopredloženi test T_2 mnogo bolji u odnosu na neograničeni test omjera vjerodostojnosti za testiranje simetrije što je posljedica činjenice da test T_2 sadrži ograničenja na parametre pod alternativnom hipotezom, odnosno više je restriktivan pa je stoga i jakost testa veća u odnosu na test H.
Također, slično ponašanje možemo očekivati i kod testa T_1 opisanog u Potpoglavlju 3.1. Iako u uvjetima alternativne hipoteze tog testa postoje neka ograničenja na parametre, ona su restriktivnija kod testa T_2 tako da pretpostavljamo da bi test T_1 bio lošiji od testa T_2, no ipak bolji u odnosu na neograničen test H.

Dakle, ukoliko imamo diskretnu distribuciju i želimo provjeriti postoji li simetrija oko određene točke, ili se ipak radi o asimetriji (pozitivnoj ili negativnoj) možemo s dovoljnom pouzdanošću koristiti test T_2 opisan u Potpoglavlju 3.2.

Literatura

[1] R. E. Barlow, H. D. Brunk, *The Isotonic Regression Problem and Its Dual*, Journal of the American Statistical Association 67 (1972), 140-147

[2] R. Dykstra, S. Kochar, T. Robertson, *Likelihood Ratio Test for Symmetry Against One-sided Alternatives*, Annals of the Institute of Statistical Mathematics 47 (1995), 719-730

[3] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[4] T. Robertson, F. T. Wright, *Likelihood Ratio Tests for and Against a Stochastic Ordering Between Multinomial Populations*, The Annals of Statistics 9 (1981), 1248-1257

[5] J. Tabor, *Investigating the Investigative Task: Testing for Skewness - An Investigation of Different Test Statistics and their Power to Detect Skewness*, Journal of Statistics Education 2 (2010), 1-13

[6] T. Yanagimoto, M. Sibuya, *Test for Symmetry of a One-dimensional Distribution against Positive Biasedness*, Annals of the Institute of Statistical Mathematics 24 (1972), 423-23

Sažetak

U ovom diplomskom radu obradili smo mjere asimetrije podataka, odnosno načine mjerenja rasporeda članova statističkog skupa prema nekoj vrijednosti, to jest prema osi simetrije.

Za razumijevanje tematike u prvom poglavlju dan je pregled osnovnih pojmova statistike. Opisan je koncept asimetrije te tri najčešće korištene mjere asimetrije. Nadalje, istražene su druge statističke veličine te njihova učinkovitost, odnosno jakost tih statističkih veličina u detekciji asimetrije pri čemu je bilo važno razumijevanje koncepta uzorkovanja distribucije te kako koristiti uzorkovanje za testiranje hipoteze za nepoznatu statistiku.

Obrađeni su testovi omjera vjerodostojnosti za testiranje simetrije u odnosu na jednostrane alternative te su opisane pripadne dvije testne statistike i njihove asimptotske distribucije. Na kraju, u simulacijskoj studiji ispitujemo jakost drugog testa te zaključujemo da uspješno konkurira poznatom neograničenom testu omjera vjerodostojnosti za testiranje simetrije.

Naposljetku, naglasimo opširnost ove tematike te da je ovim radom obuhvaćen samo mali dio. Uz već trenutno bogatu literaturu, konstantno se nadopunjava te se istražuju novi testovi za testiranje asimetrije i njihove učinkovitosti u primjeni.

Summary

In this paper we discussed measures of skewness and ways to measure the distribution of members of a statistical set according to the point of symmetry.

At the beginning of the first chapter, a basic statistical results are given. We described the concept of skewness and the most common measures of skewness. Furthermore, we have investigated different ways to measure skewness. Using simulations, we have estimated the power of each statistic to detect skewness where it was important to understand the concept of a sampling distribution and how to use the sampling distribution to test a hypothesis for an unfamiliar statistic.

Then we studied likelihood ratio tests for symmetry of a descrete distribution against one-sided alternatives. In the process, two test statistics and their asymptotic null distributions had been obtained. In addition, we performed a simulation study to compare the power of the second test and had concluded that this test successfully compete with the known unrestricted likelihood ratio test for testing symmetry vs. non-symmetry.

At the end, let us emphasize extensiveness of this topic and that this paper contain only a small part of it. Along with the already rich literature, new tests for testing skewness and their efficiency are being investigated, so it is continuously being updated.

Životopis

Mihaela Šimić rođena 29. rujna 1992. godine u Zagrebu. Završava osnovnu školu dr. Vinka Žganca te potom i XV. gimnaziju u Zagrebu. Godine 2011. upisuje Preddiplomski studij Matematika na Prirodoslovnom matematičkom fakultetu Sveučilišta u Zagrebu. Navedeni studij završava 2014. godine te tako stječe naziv bacc. univ. math. Iste godine upisuje diplomski sveučilišni studij Financijska i poslovna matematika.