MANUALIA UNIVERSITATIS STUDIORUM ZAGRABIENSIS

UDŽBENICI SVEUČILIŠTA U ZAGREBU

Sveučilište u Zagrebu

Gordana Medunić

OSNOVNE METODE (GEO)STATISTIČKE ANALIZE PODATAKA IZ OKOLIŠA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Zagreb, 2022.

Odlukom Senata Sveučilišta u Zagrebu od 13. prosinca 2022. (dopis klasa: 032-01/22-02/32, ur. broj: 380-062/250-22-4) rukopis je prihvaćen kao **sveučilišni priručnik** (*Manualia Universitatis studiorum Zagrabiensis*).

Autorica

dr. sc. Gordana Medunić, dipl. ing. geol., redovita profesorica Prirodoslovno-matematičkoga fakulteta Sveučilišta u Zagrebu, znanstvena savjetnica

Recenzenti

dr. sc. Tomislav Malvić, dipl. ing. geol., redoviti profesor Rudarsko-geološko-naftnoga fakulteta Sveučilišta u Zagrebu, znanstveni savjetnik u trajnom zvanju
dr. sc. Željka Fiket, dipl. ing. geol., viša znanstvena suradnica Instituta Ruđer Bošković

Lektorica

dr. sc. Tomislava Bošnjak Botica, viša znanstvena suradnica Instituta za hrvatski jezik i jezikoslovlje

SADRŽAJ

Predgovor

Priručnik je namijenjen studentima 1. godine diplomskoga studija geologije na Geološkom odsjeku PMF-a Sveučilišta u Zagrebu, s ciljem uspješna usvajanja nastavnoga gradiva iz predmeta Geostatistika. Nastavni materijali obuhvaćaju teoriju i riješene zadatke. Ovdje nisu predstavljeni matematički aspekti statističkih principa, o čemu postoji mnoštvo objavljenih udžbenika i knjiga na hrvatskom i engleskom jeziku. Priručnik nastoji predstaviti bitne značajke nekoliko statističkih metoda koje se najčešće koriste u geologiji i srodnim znanostima, poput testova usporedbe skupina podataka i korelacijske analize.

Uporaba i razumijevanje (geo)statističkih postupaka postaju sve potrebnije vještine u geološkoj struci i akademskoj znanstvenoj grani geologije (Barudžija i sur. 2020; Fiket i sur. 2017, 2019, 2020; Ivšinović i Malvić 2020; Malvić i sur. 2020a, 2020b; Pavlović i sur. 2004). Svrha je ovoga djela pružanje uvida u načine primjene osnovnih (geo)statističkih metoda na stvarnim podatcima iz okoliša. Doprinos se ogleda i u činjenici da je primjena kvantitativnih metoda u geoznanostima budućnost geologije i srodnih disciplina (Medunić i sur. 2022). Podatci u danim primjerima preuzeti su iz članka Medunić i sur. (2009). Obrada dotičnih podataka obavljena je s pomoću besplatnoga računalnog programa PAST (Hammer i sur. 2001).

Autorica je nadasve zahvalna recenzentima na savjetima koji su doveli do znatnoga poboljšanja kvalitete djela. Zahvale upućuje i Europskoj komisiji na novčanoj potpori za dvomjesečni boravak u Indiji (svibanj – srpanj 2022.) u okviru Erasmus+ programa, tijekom kojega je ovo djelo napisano, te kolegama domaćinima (dr. sc. Binoyu K. Saikiji, CSIR-NEIST, Jorhat, Assam, i dr. sc. Sanchiti Chakravarty, CSIR-NML, Jamshedpur, Jharkhand) na gostoljubivosti.

1. Uvod u (geo)statističku analizu geoloških podataka

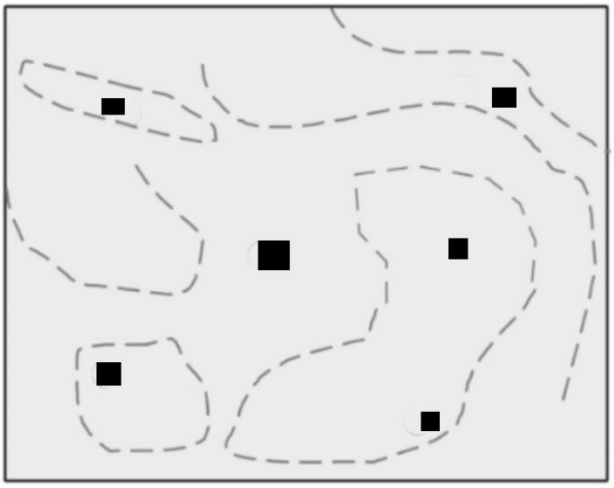
Statistička analiza brojčanih podataka te njihovo razumijevanje i tumačenje od velike su važnosti u svim znanstvenim disciplinama. Razvoj i napredak računala omogućuje sve veću upotrebu brojnih statističkih alata. Prije samo pedesetak godina uporaba statističkih metoda bila je vrlo rijetka, a danas se one rutinski koriste u svim istraživanjima. Potrebno je prije svega dobro razumjeti probleme koje želimo riješiti nekim geološkim istraživanjem da bi se mogli prikupiti smisleni podatci na kojima će se primijeniti (geo)statističke metode analize. Loš je pristup prvo skupiti uzorke pa tek onda pokušati odgonetnuti što učiniti s njima, naprotiv, treba dobro znati kako su geološki uzorci uzeti i zašto su oni važni. Stoga treba imati na umu sljedeće: a) stvarni problemi gotovo uvijek zahtijevaju postupnu upotrebu nekoliko (geo)statističkih metoda, b) gotovo je uvijek riječ o više razumnih/smislenih rješenja koja one nude te c) problemi s kojima se susrećemo u praksi općenito su daleko složeniji i nejasniji nego oni u udžbenicima (Zhang 2011). Nepoznat netko nekoć je negdje zabilježio sljedeće: „Sve na ovom svijetu povezano je sa svim ostalim u tanahan i zamršen splet uzajamnih odnosa. Najbolje računalo što ga je čovjek ikada konstruirao ipak ne može izračunati čak ni sićušan dio svih odnosa koji postoje u ekosustavu običnoga ribnjaka. Znanstvenici su to pokušali, no ostalo im je jedino u očaju dići ruke od svega nakon što su shvatili o kakvoj se zamršenosti i detaljima tu radi“.

Naziv statistika potječe od latinske riječi *status* što znači 'stanje'. Glavni zadatci statistike su: (a) izvesti zaključke o **populaciji** na temelju analize informacija sadržanih u ograničenom broju **uzoraka**, (b) procijeniti veličinu **nesigurnosti** povezane s tim zaključcima i (c) osmisliti postupak i veličinu **uzorkovanja** da bi opažanja poslužila donošenju pravilnih i točnih zaključaka.

Na početku treba odrediti: (1) koju **varijablu** (mjereno svojstvo uzorka, npr. koncentracija Cu u tlu) izabrati te (2) na koji način uzeti uzorke iz prirode, a da za ishod imaju mogućnost procjene, predviđanja i kartiranja dotičnih vrijednosti. S obzirom na to da je priroda (ali i unutrašnjost Zemlje) jako promjenjiva (varijabilna) u gotovo svakom smislu, osnovna je zadaća (geo)statistike opisati tu promjenjivost, koja može biti kilometarskih do mikrometarskih razmjera. Mjerenjem i analizom uzoraka dobivamo uvid u procese koji oblikuju okoliš te čimbenike koji utječu na njegovu promjenjivost. Time dobivamo mogućnost predviđanja prostornih odnosa i upravljanja resursima od interesa (kakvoća vode, stanje nutrijenata u

poljoprivrednom tlu, rudonosni potencijal istraživanoga terena itd.). Upravo je glavni zadatak geostatistike istraživanje i procjena prostorne promjenjivosti.

Zamah razvoja geostatistike 1960-ih godina dogodio se zbog rudarske industrije, a daljnji razvoj prenio se na naftno inženjerstvo, hidrogeologiju, meteorologiju, istraživanje tla, poljoprivredu, ribarstvo, probleme zagađenja okoliša i njegove zaštite (npr. Davis 1986). To se znanstveno polje bavi proučavanjem prostornih pojava s pomoću odgovarajućih metoda kartiranja (npr. Malvić 2008). Danas geostatistika obuhvaća lepezu znanstvenih polja i grana koje se bave analizom prostornih, ponekad i vremenskih podataka, poput oceanografije, hidrogeologije, daljinskih istraživanja, agronomije i znanosti o okolišu. Uspjeh geostatistike leži u njezinoj sposobnosti uporabe prvoga zakona geografije, koji prema Tobleru (1970) glasi: „Sve je povezano sa svim ostalim, ali bliske stvari više su povezane od onih udaljenijih“. Jedno od glavnih obilježja takvih podataka jest njihova strukturirana raspodjela u prostoru (i vremenu) kao odraz utjecaja različitih čimbenika (npr. geologije, ljudske aktivnosti itd.) koji djeluju u različitim razmjerima. Bitno je naglasiti da se geostatistika uglavnom bavi predviđanjem vrijednosti (sl. 1.1) gdje nemamo uzorke na temelju onih izmjerenih (npr. Malvić 2008; Zhang 2011).



Slika 1.1. Pravokutnici crne boje predstavljaju područja gdje su uzeti uzorci, a crtkane izolinije pokazuju područja vrijednosti procijenjenih na temelju geostatističkoga računa mjerenih podataka.

Interpolacija i ekstrapolacija moguće su zbog postojanja autokorelacije u podatcima, koja se može kvantificirati i modelirati s pomoću variograma (najčešće). Različite tehnike krigiranja omogućuju procjenu vrijednosti varijable i odgovarajuću varijancu pogreške predviđanja na

neuzorkovanim lokacijama (npr. Malvić 2008). Nadalje, za razliku od statistike koja se između ostaloga bavi analizom i tumačenjem nesigurnosti uzrokovanih ograničenim uzorkovanjem pojave koja se proučava, geostatistika nije uvijek ograničena pretpostavljenim modelom raspodjele populacije, npr. da su svi uzorci populacije simetrično raspodijeljeni i međusobno neovisni. Većina geoloških podataka (npr. svojstva stijena, koncentracije onečišćujućih tvari u sedimentu itd.) uglavnom ne udovoljavaju pretpostavci prostorne simetričnosti (izotropnosti) jer mogu imati jako asimetrične raspodjele, anizotropiju i/ili posjedovati tek malu prostornu zavisnost (tj. vrijednosti podataka s lokacija koje su bliže jedna drugoj teže biti međusobno sličnije u odnosu na vrijednosti podataka s udaljenijih lokacija). Većina geoznanstvenika dobro je upoznata s činjenicom da su blisko smješteni uzorci obično međusobno sličniji jer su takvi uzorci bili pod utjecajem sličnih (geo)fizičkih i (geo)kemijskih procesa u prirodi. U usporedbi sa statistikom koja ispituje statističku raspodjelu skupa izmjerenih podataka geostatistika obuhvaća ne samo statističku raspodjelu podataka uzorka nego i prostornu korelaciju između podataka uzorka. Zbog te razlike kartiranje u geoznanostima učinkovitije se rješava korištenjem geostatističkih metoda. One pružaju alate za razumijevanje pojava kroz njihovo rigorozno ispitivanje (pravila stacionarnosti) te opisivanje na temelju rijetkih, često pristranih, ali i skupih podataka.

Nadalje, geostatistika nudi sredstvo za kvantificiranje nesigurnosti, uz korištenje postojećih podataka kao podršku optimizaciji uzorkovanja (Motulsky 1999). Može se definirati kao skup alata za kvantificiranje geoloških informacija u svrhu izrade jednodimenzionalnih ili višedimenzionalnih numeričkih geoloških modela koji se koriste za procjenu i predviđanje učinkovitosti nekoga ležišta (Malvić i sur. 2008). Kao njezine glavne prednosti mogu se izdvojiti mogućnosti: (1) modeliranja heterogenosti ležišta, (2) integriranja različitih tipova podataka možebitno različitih uporišta i različitih stupnjeva pouzdanosti te (3) procjena i kvantificiranje nesigurnosti modela ležišta. Stoga je to vjerojatnosni (probabilistički) pristup proučavanju prirodnih pojava koje variraju u prostoru, na „inteligentan“ i matematički robustan način. Također je to poželjna metoda za rad s velikim skupovima podataka i integraciju različitih tipova podataka, a sve potpomognuto matematičkom strogošću i ponovljivošću te potrebom donošenja odluka suočenih s neizvjesnošću (npr. Davis 1986; Reimann i sur. 2008; Zhang 2011).

Takav skup matematičkih alata sastoji se od komponenata analize podataka i algoritama interpolacije/ekstrapolacije (npr. Mesić i Medunić 2014). Mnogi geoznanstvenici izbjegavaju interpretirati vjerojatnosnu komponentu geostatističkih rješenja jer ne žele uvoditi nesigurnosti u svoj model. No, sama geostatistika se posljednjih godina udaljila od ideje o jedinstvenom

deterministočkom odgovoru na probleme geoznanosti, nego se usredotočuje na **neizvjesnost** povezanu s tim odgovorom. Na primjer, takva rješenja neće uputiti na „bušenje (u potrazi za nekim resursom) dva metra s lijeve strane“, već radije na „bušenje između nultoga i desetoga metra s lijeve strane s najvećom vjerojatnošću pronaći resurs nakon dva metra s lijeve strane“. Također, takav model neće dati precizne količine resursa u nekom ležištu, već će samo procijeniti dotični volumen i **nesigurnost** povezanu s tom odlukom.

Na znanstvenicima je da dadnu stručno mišljenje o potencijalnim lokacijama bušotina, ali uz alate koji im omogućuju kvantificiranje neizvjesnosti i **rizika** povezanih s odlukama koje moraju donijeti (npr. Mesić Kiš i Malvić 2014). Promjena u „filozofiji“ jest prihvaćanje činjenice da postoji nesigurnost u volumenima ležišta i da nikada nećemo dobiti točan odgovor, tj. bilo koja deterministička naznaka nosit će stupanj pogreške. Stoga se odluke trebaju donositi kvantificiranjem neizvjesnosti i „**vjerojatnosti** ishoda“ (Reimann i sur. 2008; Zhang 2011).

Tradicionalno su se odluke često donosile „ekspertnom procjenom“, npr. „promatranjem“ konturnih karata i odlučivanjem o tome koji bi dio ležišta bio najbolji na temelju uočene korelacije između dviju bušotina. Do poteškoća dolazi pri prelasku s 1D ili 2D na 3D skupove podataka pri čemu kvalitativno donošenje odluka postaje nemoguće i nepouzdano. Geostatistički alati su ti koji pružaju **kvantifikacijski** okvir rješavanja različitih (koreliranih ili autokoreliranih) podataka, uzorkovanih u različitim volumenima, s različitim razinama preciznosti i pouzdanosti te za različita geološka obilježja. Kao i svi alati, geostatistika ima odgovarajuće namjene i ograničenja. Geostatistika je vrlo korisna za cjeloživotni vijek ležišta, ali najveći utjecaj ima u ranoj fazi kad postoji golema nesigurnost oko geoloških količina nekoga resursa. Tijekom vremena dodatni podatci postaju sve dostupniji (npr. zapisi bušotina, 3D seizmički podatci, podatci o proizvodnji, itd.) pa se i njihova nesigurnost smanjuje, što je poznato kao informacijski učinak. Kako vrijeme prolazi, sve više informacija postaje dostupno, što dodatno ograničava model i smanjuje nesigurnost. Na kraju vijeka ležišta ono će u potpunosti biti poznato te će neizvjesnost nestati (Zhang 2011.).

Zbog svoje prostorne komponente geološki podatci imaju posebna svojstva koja je potrebno prepoznati i razumjeti prije njihove statističke analize da bi se odabrale ispravne metode analize podataka. Ta svojstva uključuju sljedeće (Reimann i sur., 2008; Zhang, 2011):

a) Podatci su prostorno ovisni, tj. što su dva uzorkovana mjesta bliža jedno drugom, to je veća vjerojatnost da uzorci pokazuju slične vrijednosti.

b) Na bilo kojem uzorkovanom mjestu mnoštvo različitih procesa moglo je utjecati na izmjerenu analitičku vrijednost (npr. u slučaju uzoraka tla to mogu biti izvorni materijal stijenske podloge, topografija, vegetacija, klima, Fe/Mn-oksihidroksidi, sadržaj organske tvari,

raspodjele veličina zrna, pH, mineralogija, prisutnost mineralizacije ili onečišćenja, analitičke pogreške, itd.). Međutim, većina statističkih testova pretpostavlja da uzorci potječu iz iste raspodjele (populacije) što u geoznanostima tek povremeno vrijedi (npr. ako su mjerenja provedena u istom taložnom okolišu) jer različiti procesi utječu na različite uzorke u različitim omjerima. Mješavina rezultata uzrokovanih cijelim nizom različitih temeljnih procesa može oponašati npr. lognormalnu raspodjelu (kada su logaritmi izvornih podataka normalno raspodijeljeni), ali i niz drugih (npr. Malvić i Medunić 2015). Međutim, ako podatci potječu iz više populacije, nije ispravno smatrati ih vrijednostima i varijablama iste populacije s jedinstvenom razdiobom (npr. normalnom).

c) Poput brojnih znanstvenih podataka (npr. psihološka istraživanja) geološki podatci su po svojoj prirodi neprecizni te u sebi sadrže nesigurnost (npr. Mesić i Medunić 2014). One se neizbježno unose tijekom uzorkovanja, pripreme i analize uzorka, što degradira primjenu (geo)statističkih metoda.

d) Podatci iz okoliša najčešće su podatci o sastavu uzorka (tlo, sediment, stijena, minerali itd.). Pojedinačne varijable nisu međusobno neovisne, već su povezane tako što se, na primjer, izražavaju u postotcima (ili dijelovima na milijun, tj. ppm, odnosno mg/kg). Oni se zbrajaju dajući npr. 100 % ili 1. Postotci su omjeri koji sadrže sve varijable koje se istražuju u nazivniku. Dakle, pojedinačne varijable postotnih podataka ne mogu slobodno i neovisno varirati pa ovo ima ozbiljne posljedice za analizu podataka.

Navedena svojstva ne uklapaju se dobro u pretpostavke „klasične“ statistike. Međutim, to su najšire korištene statističke metode koje se poučavaju u svim osnovnim statističkim predmetima na sveučilištima. Umjesto njih geološki podatci često preferiraju korištenje robustnih neparametarskih statističkih metoda kao prikladnijih.

Zaključno, tehnike kvantitativne geologije obuhvaćaju (geo)statističke odnosno geomatematičke metode obrade geoloških podataka (Barudžija i sur. 2020; Ivšinović i Malvić 2020; Malvić i Cvetković 2013; Malvić i sur. 2020a, 2020b). One se temelje na postavci da se informacija o nekoj pojavi može „izvući“ iz opažanja manjega broja uzoraka (npr. 10-30) prikupljenih iz neusporedivo većega skupa potencijalnih opažanja (npr. nekoliko desetaka do nekoliko tisuća) neke pojave koju nazivamo populacijom. Međutim, geoznanstvenici izvode opažanja većinom tamo gdje za to imaju mogućnosti. Podatci iz dubokih bušotina preskupi su da bi se odbacili samo zato što se njihove lokacije ne uklapaju u klasični statistički dizajn uzorkovanja. Paleontolozi se moraju zadovoljiti fosilima koje prikupe s izdanaka, dok su im oni zatrpani duboko pod površinom zauvijek nedostupni. Uzorci intruzivnih stijena mogu se uzeti samo s njihovih vršnih dijelova duž „zidova“ kanjona, dok se njihovi dublji dijelovi

nalaze na nedostupno velikim dubinama. Problem je geoznanosti uglavnom u tome što raspolaže s **premalo** podataka. Stoga se iz njih pokušava izvući što je moguće više spoznaja, ali pri tome prepoznati i nedostatke te nesavršenost toga znanja (sustavne pogreške ili pristranost). Dostupnost osobnih računala doprinosi sve većoj važnosti poznavanja principa koji stoje iza relevantnih statističkih izračuna. Te naprave mogu brzo izvesti bilo koji statistički test ili izračun koji se izabere bez obzira na možebitnu (ne)primjerenost postupka u slučaju analiziranih podataka. Čak ni najbolji računalni programi ne nude savjet o ispravnom izboru metoda za dani skup podataka. Stoga analitičari moraju imati statističko znanje, ali isto tako i zdrav razum da bi izveli točne izračune (tzv. GIGO princip – engl. *garbage in, garbage out*). Računala su savršen alat za analiziranje podataka, ali programi za analizu podataka mogu se zloupotrijebiti. Ako se unese netočan podatak ili upotrijebi neprikladna metoda analize, rezultati će biti beskorisni. Otuda princip GIGO – loši podatci daju loše rezultate (Davis 1986). Analizom podataka želi se doći do najsnažnijega mogućeg zaključka na temelju ograničene količine podataka. U tu svrhu potrebno je nadvladati dva problema:

1) Bitne razlike među skupinama podataka mogu ostati nezamijećene zbog geološke promjenjivosti i eksperimentalne nepreciznosti, a to otežava raspoznavanje ili razlučivanje stvarnih razlika među uzorcima od nasumične ili slučajne promjenjivosti.

2) Obilježje je ljudskoga mozga uočiti obrasce ili modele čak i u nasumičnim podatcima. Ljudi imaju prirodnu sklonost (pogotovo s vlastitim podatcima) donošenja zaključaka kako su razlike među uzorcima stvarne (posljedica neke zakonitosti), a ne rezultat nasumične promjenjivosti (Motulsky 1999).

Statistička je analiza nužna kad opažene razlike među uzorcima nisu zamjetne u odnosu na eksperimentalnu nepreciznost i geološku promjenjivost. U brojnim situacijama geoznanstvenici ne mogu izbjeći velike količine promjenjivosti kad je posrijedi terenski i laboratorijski rad, a brinu se o relativno malim razlikama među skupinama podataka, stoga je, primjerice, besmisleno trošiti dragocjene resurse (vodu i struju) čisteći danima sito do zadnjega zrnca nakon sijanja sedimenta/tla.

Što se tiče mjernih sustava, geoznanstvenici moraju biti svjesni prirode brojnih sustava u kojima se obavljaju mjerenja. Trebaju razumjeti ne samo geološko značenje opažanih/mjerenih varijabli nego i matematičko značenje korištenih metoda obrade podataka (npr. Davis 1986; Motulsky 1999; Malvić i Cvetković 2013; Malvić i Medunić 2015).

2. Priprema podataka za (geo)statističku analizu

Temeljna je ideja (geo)statistike jednostavna: procijeniti vrijednosti izvan područja poznatih vrijednosti (prikupljeni podatci, tj. uzorci) u svrhu postavljanja općih zaključaka o većoj populaciji iz koje ti uzorci potječu. Podatci predstavljaju populaciju ili katkad ciljanu populaciju. Budući da je fizički i financijski nemoguće prikupiti sve podatke koji nas zanimaju (npr. svu riječnu vodu unutar razdoblja promatranja), uzima se i mjeri samo podskup podataka nazvan uzorak, i to na način da se zaključci o njemu mogu proširiti na cijelu populaciju. Statistički parametri izračunati na temelju uzorka samo su zaključci ili procjene obilježja populacije. U istraživanjima okoliša populacija je gotovo uvijek označena granicama neke fizičke regije, a jedinice su sva mjesta unutar nje gdje netko može mjeriti njezina svojstva (ako se radi o manjim poljima, ona moraju biti istih dimenzija). Populacija se uzorkuje uzimanjem podskupa njezinih jedinica definiranih dimenzija. Taj podskup treba biti odabran tako da udovoljava elementima slučajnosti, tj. da je otklonjena bilo kakva mogućnost pristranosti (Pentecost 1999).

(Geo)statistički testovi temelje se na pretpostavci da je svaki subjekt (ili svaka eksperimentalna jedinica) uzorkovan neovisno o ostalima. Primjerice, prikupljanje deset uzoraka tla iz središta grada i deset iz prigradskih naselja ne predstavlja dvadeset subjekata iz jedne populacije. Gradski podatci mogu biti međusobno sličniji i pod većim antropogenim utjecajem u usporedbi s onim prigradskima. Dotični uzorci uzeti su iz dviju populacija (gradska i prigradska) pa to treba uzeti u obzir.

Podatci su **neovisni** kada bilo koji slučajni čimbenik, koji uzrokuje previsoku ili prenisku vrijednost, utječe samo na jednu vrijednost. Ako faktor koji nije uzet u obzir prilikom analize podataka može utjecati na više od jedne vrijednosti, tada podatci nisu međusobno neovisni. U svrhu ekstrapolacije uzoraka na populaciju statističari su izumili tri osnovna pristupa (npr. Motulsky 1999; Zhang 2011).

1) Populacija slijedi određenu raspodjelu, npr. Gaussovu. Tada statistički testovi omogućuju donošenje zaključaka o obilježjima populacije.

2) Sve vrijednosti podataka rangiraju se od najnižih do najviših nakon čega se uspoređuju istovjetni rangovi skupina podataka. Na tom se principu temelji većina često korištenih neparametarskih testova za analizu podataka koji nisu normalno raspodijeljeni.

3) Ponovljeno uzorkovanje.

2.1. Uzimanje uzoraka

Skupovi geoloških podataka obično su malobrojni ($n < 30$), a u statistici je uvijek bolje raditi s većim skupovima podataka jer je u tom slučaju moguće donijeti sigurnije (pouzdanije) zaključke. Katkad se koriste za opisivanje prilično velikih prirodnih fenomena kao što su granitno tijelo, golemo klizište ili sedimentna jedinica širokoga prostiranja. Geostatističkim metodama predviđaju se prostorna obilježja populacije na temelju prikupljenih uzoraka. Uzorci moraju biti reprezentativni, što znači da osnovni statistički parametri skupa vrijede za cijelu populaciju te zato skup mora biti dovoljno velik. Drugim riječima, uzorci trebaju što bolje predstavljati istraživanu populaciju. Stoga postupak odabira uzoraka mora sadržavati određeni stupanj nasumičnosti da bi svaka sastavnica populacije imala jednaku šansu ili vjerojatnost biti uzorkovana (načelo slučajne varijable). To je u geoznanostima vrlo teško izvedivo (npr. Helsel i sur. 2020). U geološkom kontekstu obično nema problema oko nasumičnoga uzorkovanja ako je populacija jednolike građe ravnomjerno dostupna. Daleko je češći slučaj nejednolike populacije, čiji dijelovi nisu jednako dostupni. Tada je gotovo nemoguće postići nasumični uzorak populacije te nije uputno o njoj donositi zaključke (Medunić 2022).

Pravilna strategija uzorkovanja ovisi o vrsti analizirane varijable, ciljevima istraživanja i traženoj pouzdanosti rezultata. Nakon odabrane strategije uzorkovanja na kvalitetu skupa podataka mogu utjecati brojni poremećaji. Na primjer, uzorci nisu reprezentativni za veću populaciju. Kemijske ili fizičke promjene, onečišćenje drugim materijalom ili pak premještanje prirodnim ili antropogenim procesima može rezultirati pogrešnim rezultatima i tumačenjima. Stoga je preporučljivo ispitati kvalitetu uzorka, primijenjenu metodu analize podataka i valjanost zaključaka temeljenih na analizi u svim stadijima provedenoga istraživanja (Swan i Sandilands 1995). Uspješna strategija terenskoga uzorkovanja (npr. Medunić i sur. 2016) obuhvaća odluke o veličini uzorka i prostornoj shemi uzorkovanja.

1) Veličina uzorka podrazumijeva volumen, težinu i broj uzoraka uzetih na terenu. Prva dva su od ključne važnosti ako se uzorci kasnije trebaju analizirati u laboratoriju. Većina (geo)statističkih metoda također zahtijeva neku minimalnu brojnost uzorka. U slučaju nejednolike populacije uzorak mora biti dovoljno velik da opisuje njezinu promjenjivost. Nadalje, uzorak treba biti što manji da se skrati vrijeme i napor za provedbu uzorkovanja i analize. Preporučljivo je prikupiti manji probni uzorak prije definiranja konačne veličine uzorka.

2) Po prostornoj shemi uzorkovanja u većini područja uzorci se uzimaju onako kako dopuštaju raspoloživi izdanci na terenu. Uzorkovanje u kamenolomima obično je u obliku klastera, dok je duž usjeka cesta, obalnih klifova ili strmih klanaca shema uzorkovanja poprečna. Na primjer, pravilna shema uzorkovanja znači da se radi o pravokutnoj ili sličnoj mreži točaka uzorkovanja, a jednolika strategija uzorkovanja podrazumijeva pravilan razmještaj točaka uzorkovanja unutar svakoga pojedinačnog četverokuta mreže.

2.2. Tipovi podataka

Većina geoloških podataka sastoji se od brojčanih vrijednosti mjerenja, ali je moguć i kategorički prikaz. Raspoložive metode analize podataka mogu zahtijevati određene tipove podataka (Helsel i sur. 2020).

a) Nominalni – npr. učestalost minerala, tipova stijena te drugih obilježja kao što su struktura ili tekstura. Stoga se informacija katkad prikazuje kao popis naziva, npr. razne fosilne vrste iz sloja vapnenca ili pak minerali prepoznati u nekom izbrusku. U nekim je istraživanjima te podatke potrebno pretvoriti u kategorički oblik (1 = ima, 0 = nema).

b) Ordinalni – npr. Mohsova ljestvica (skala) tvrdoće, Mercallijeva ljestvica (skala) jačine potresa, rangovi, zastupljenost teških minerala, stratigrafsko mjerilo, školske ocjene, brojevi kuća, itd. To su brojčani podatci koji predstavljaju opažanja koja se mogu rangirati, ali intervali duž mjerne ljestvice nisu konstantni. Dakle, ta ljestvica služi samo za označavanje njihova redoslijeda, tj. određuje se samo je li nešto veće ili manje od nečega drugog, ali razlike između pojedinih jedinica ljestvice nisu jednake.

c) Intervalni – npr. temperaturna ljestvica, sferičnost, zaobljenost, ljestvica kisikovih izotopa, 'apsolutna' vremenska ljestvica itd. Kod ovakvih podataka poznat je ne samo redoslijed već i razlika među brojevima na ljestvici jer je u tim ljestvicama neka definirana razlika jednaka na svakom njihovu dijelu. Na primjer, razlika od 1 ^{o}C uvijek je jednaka bez obzira na to radi li se o razlici između 0 ^{o}C i 1 ^{o}C ili između 153 ^{o}C i 154 ^{o}C.

d) Omjerni – npr. mm, g, ml, K, itd. Ti podatci imaju sva svojstva intervalnih, a k tomu i svojstvo da jednaki brojčani odnosi (omjeri) znače i jednake odnose u mjerenoj pojavi. Primjeri su mjerenje duljine, težine, otpora itd.

Osim tih standardnih tipova podataka geolozi se često susreću s posebnim vrstama podataka, s time da ni jedan nije samo geološki. Takvi su na primjer idući podatci.

a) Zatvoreni – izražavaju se kao proporcije/omjeri koji se zbrajaju čineći neki fiksni ukupni zbroj, kao što je 100 %. Podatci o sastavu predstavljaju glavninu zatvorenih podataka, poput primjerice elementnoga sastava uzoraka stijena.

b) Prostorni – prikupljaju se u 2D ili 3D području/prostoru. Raspodjela neke fosilne vrste, prostorna varijabilnost debljine pješčenjačkoga sloja i 3D koncentracija neke kemijski inertne tvari (engl. *tracer*) u podzemnoj vodi primjeri su ovoga tipa podataka. Može se reći da je ovaj tip podataka najvažniji u geološkim znanostima.

c) Usmjereni – izražavaju se kutovima. Primjeri su nagib i pružanje sloja, orijentacija izduženih fosila ili smjer toka lave. To je vrlo čest tip podataka u geologiji.

2.3. Neke od metoda (geo)statističke analize podataka

Metode analize podataka koriste se za opisivanje obilježja uzorka što je preciznije moguće. Nakon što se definiraju obilježja uzorka, postavljaju se istraživačke hipoteze. Određena metoda koja se koristi za opisivanje podataka ovisi o njihovu tipu i zahtjevima istraživanja (npr. Davis 1986).

• Univarijatne metode – svaka varijabla istražuje se zasebno pod pretpostavkom da su varijable međusobno neovisne. Te metode obuhvaćaju izračunavanje mjera središnje tendencije i mjera varijabilnosti. Primjeri su sadržaj Na u krhotinama vulkanskog stakla koje je bilo pod utjecajem kemijskoga trošenja ili veličina ljuštura puža u sloju sedimenta.

• Bivarijatne metode – dvije varijable istražuju se istovremeno da bi se uočili njihovi međusobni odnosi. Primjerice, može se računati koeficijent korelacije u svrhu ispitivanja linearnoga odnosa među dvjema varijablama. Primjer je dijagram odnosa oksida odabranih elemenata u odnosu na sadržaj SiO_2 u magmatskim stijenama.

• Multivarijatna analiza – obuhvaća istovremeno opažanje i analiziranje više varijabli. Budući da je grafički prikaz višedimenzionalnih skupova podataka složen, većina metoda služi se smanjenjem dimenzija. One se često koriste za geokemijske podatke, npr. u tefrokronologiji

kad se slojevi vulkanskoga pepela koreliraju s pomoću geokemijskoga otiska prsta (engl. *fingerprint*) krhotina vulkanskoga stakla. Drugi važan primjer jest usporedba fosilnih zajednica u oceanskim taložinama s ciljem rekonstrukcije paleookoliša.

• Analiza vremenskih nizova – ovim metodama ispituju se nizovi podataka u funkciji vremena. Vremenski nizovi rastavljaju se u dugoročni trend, sustavnu (periodičnu, cikličnu, ritmičnu) i nepravilnu (nasumičnu, slučajnu) komponentu. Primjeri su istraživanje cikličnih klimatskih varijacija u taložnim stijenama.

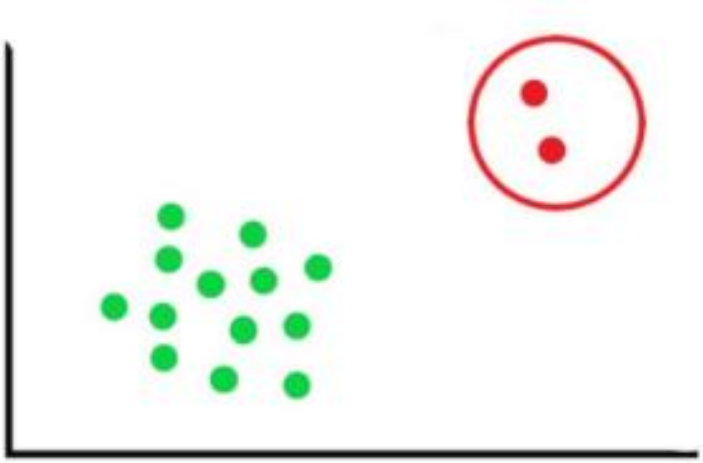
• Prostorna analiza – ovo je analiza parametara u 2D ili 3D prostoru. Stoga su dva ili tri od potrebnih parametara koordinate. Ove metode obuhvaćaju opisne alate kojima se ispituje prostorni model. Druge tehnike podrazumijevaju prostornu regresijsku analizu za otkrivanje prostornih trendova. Konačno, 2D i 3D interpolacijske tehnike pomažu u procjeni površina/ploha koje predstavljaju kontinuiranu raspodjelu neke varijable unutar područja.

• Obrada slika – obrada i analiza slika sve je važnija u geološkim znanostima. U te metode ubraja se obrada slika u svrhu povećanja signala u odnosu na šum te izdvajanja određene komponente slike. Uglavnom je to analiza dijela spektra koloritne ili sive slike da bi se po tome prepoznala neka fizikalna svojstva, npr. stijena od detritusa ili vegetacije.

• Analiza usmjerenih podataka – metode analize kružnih i sferičnih podataka naširoko se koriste u geološkim znanostima. Strukturni geolozi mjere i analiziraju orijentaciju strija na rasjednoj plohi. Također su česte kružne statističke metode u paleomagnetskim istraživanjima. Mikrostrukturna istraživanja obuhvaćaju analizu oblika zrna i orijentaciju osi c u kvarcu u mikroskopskim izbruscima.

2.4. Ekstremne vrijednosti

Ekstremne vrijednosti (engl. *outlier*) neobično su visoka ili niska opažanja čije vrijednosti jako odstupaju od ostalih u nizu (sl. 2.1).



Slika 2.1. Dvije crvene točke predstavljaju ekstremne odnosno neobično visoke vrijednosti u odnosu na glavninu zelenom bojom označenih podataka.

Obično se odbacuju prije (geo)statističke analize jer doprinose neželjenoj asimetričnosti raspodjele podataka. To nije uvijek uputno jer u skupu podataka mogu imati najveći značaj, npr. u istraživanjima onečišćenja okoliša (Medunić i sur. 2009), tako da ih treba detaljnije ispitati (npr. niske vrijednosti stratosferskoga ozona nad Antarktikom mogle su biti uočene čak deset godina ranije da nisu bile sustavno uklanjane iz računalne obrade podataka). Dakle, ako se ekstremne vrijednosti izbrišu, preuzima se rizik da se vidi samo ono što se očekuje od podataka. Ekstremne vrijednosti mogu imati tri uzroka: 1) pogreška mjerenja ili bilježenja, 2) opažanje pripada populaciji koja nije slična glavnini podataka (npr. poplava izazvana slomom brane, a ne padalinama) te 3) iznimno rijedak događaj u populaciji (npr. Reimann i sur. 2008; Helsel i sur. 2020).

Postoje grafičke metode za prepoznavanje ekstremnih vrijednosti. Osim toga, moguće je obaviti logaritamsku transformaciju podataka, ali ne treba ih odbaciti ni ako tada nije postignuta simetričnost raspodjele podataka. U tom slučaju treba upotrijebiti postupke koji nisu

pod utjecajem ekstremnih vrijednosti, a to su neparametarske metode. Za razliku od parametarskih metoda koje zahtijevaju da podatci budu normalno raspodijeljeni, neparametarske se ne temelje na pretpostavkama o vrsti raspodjele podataka.

Svrha matematičkih transformacija podataka jest učiniti raspodjelu podataka što simetričnijom i učiniti da podatci imaju što nižu varijancu ili raspršenost (tj. veću jednolikost) oko svojih aritmetičkih sredina. Općenito je statistička obrada međusobno sličnijih (jednolikijih) podataka lakša (u smislu donošenja zaključaka o populaciji na temelju uzoraka) u odnosu na obradu podataka velikih varijanca. Premda je matematičko transformiranje podataka svojevrsno „vidi ono što želiš vidjeti“, njihova primjena ipak ima daleko veću važnost te nije samo proizvoljan izbor. Transformacijom u nove jedinice mijenja se samo udaljenost između točaka na liniji. Učinak je smanjenje ili povećanje udaljenosti prema krajnjim vrijednostima na jednoj strani medijana kako bi one postale što sličnije onima na drugoj strani. Postoji cijeli niz transformacija za pozitivnu i negativnu asimetriju, pri čemu na raznim uzorcima iz iste populacije treba izvoditi istu transformaciju (Helsel i sur. 2020).

Nikakav matematički račun ne može sigurno upućivati na to je li ekstremna vrijednost došla iz iste ili neke druge populacije. Međutim, statistički račun može odgovoriti na sljedeće pitanje: ako svi podatci doista pripadaju Gaussovoj ili normalnoj raspodjeli podataka, kolika je vjerojatnost da se jedna (ekstremna) vrijednost nađe toliko daleko od ostalih? Ako je ta vjerojatnost (p) mala, tada se zaključuje da je to doista ekstremna vrijednost (možda zbog pogrešna mjerenja) i može se isključiti iz analize.

Statističari su osmislili nekoliko metoda za prepoznavanje ekstremnih vrijednosti (npr. Grubbsov i Dixonov test), gdje se prvo računa koliko je ekstrem daleko od ostalih vrijednosti (npr. koliko je daleko u odnosu na aritmetičku sredinu svih ostalih vrijednosti ili u odnosu na prvu najbližu vrijednost). Zatim se ta vrijednost standardizira dijeljenjem s nekom mjerom raspršenja, kao što je standardna devijacija svih ili preostalih (tj. bez te ekstremne vrijednosti) podataka ili pak raspon podataka. Konačno se izračuna vrijednost p (vjerojatnost) da bi se dobio odgovor na pitanje: ako svi podatci doista potječu iz Gaussove populacije, kolika je vjerojatnost da se pukim slučajem nađe ekstrem toliko daleko od ostalih vrijednosti? Ako je vrijednost p mala, odstupanje od ostalih vrijednosti statistički je značajno (dakle, nije rezultat slučajnosti koja je dala većinu ostalih podataka) te se može isključiti iz analize. Drugim riječima, to je vjerojatnost da se inače normalno opažanje pogrešno proglasi ekstremom kada uzorak potječe iz populacije normalno raspodijeljenih vrijednosti podataka. Međutim, ono što se čini

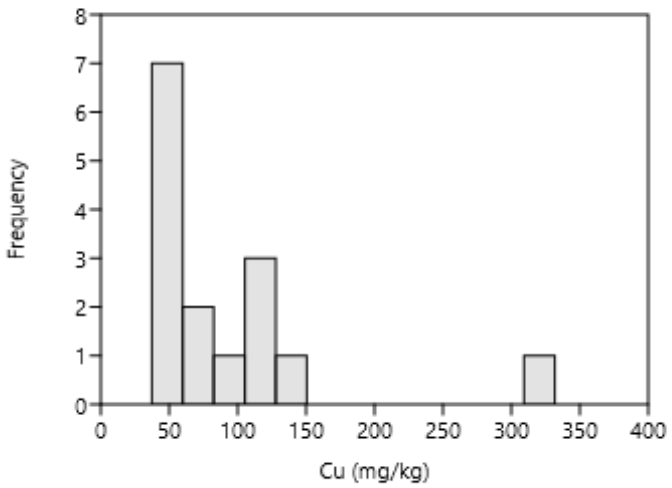
ekstremnom vrijednošću, pod pretpostavkom da je riječ o normalnoj raspodjeli, uopće to ne mora biti ako uzorak zapravo potječe iz, primjerice, lognormalne raspodjele. U tom slučaju ti testovi nisu upotrebljivi. Ovaj problem u kombinaciji s pojavom višestrukih ekstrema razlog je sve veće upotrebe robusnih neparametarskih statističkih metoda koje nisu osjetljive na ekstremne vrijednosti ili im pridjeljuju manju težinu u izračunima pa je time izbjegnut problem odbacivanja (ili prihvaćanja) ekstrema (Miller i Miller 2010).

2.5. Grafička analiza podataka

Dijagrami ili grafikoni pružaju daleko bolji i brži sažet pregled podataka nego tablice s brojevima. Jedna je od njihovih zadaća analiza podataka s ciljem njihova početnoga ispitivanja (engl. *exploratory data analysis*, EDA). Dotični postupci (EDA) svojevrsni su „prvi pogled“ u podatke. To su induktivni postupci jer se podatci ne testiraju nego se sumiraju. Njihovi su rezultati vodič za izbor odgovarajućih postupaka testiranja nulte hipoteze (npr. Helsel i sur. 2020).

2.5.1. Histogrami

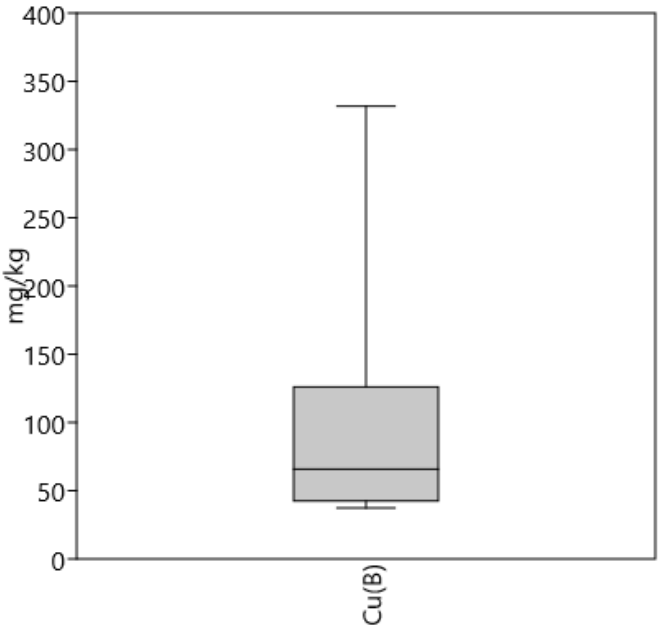
Ovo su dobro poznati dijagrami u kojima visina stupaca odgovara broju podataka (frekvenciji) koji pripadaju određenim intervalima ili kategorijama (sl. 2.2). Nedostatak im je u tome što njihov vizualni dojam ovisi o broju kategorija, tj. o izboru širine stupaca i njihovim središnjim vrijednostima pa su moguće praznine. Ti su dijagrami korisni za prikaz podataka koji prirodno tvore skupine ili kategorije. Primjer je broj vodocrpilišta koja premašuju neku vrijednost (npr. koncentracija arsena u podzemnoj vodi), grupiranih prema geološkim jedinicama.



Slika 2.2. Primjer histograma na temelju podataka za Cu u skupini B (Medunić i sur. 2009).

2.5.2. Kutijasti dijagrami

Kutijasti (engl. *box-plot* ili *box-whisker*) dijagrami najčešći su način slikovnoga prikaza podataka (sl. 2.3). Središnja je vrijednost podataka medijan (središnja linija u kutiji, engl. *box*), a tzv. zalistci (engl. *whisker*) jesu linije koje se obično protežu od kutije do najviše (engl. *max*) odnosno najniže (engl. *min*) vrijednosti. Raspršenje je prikazano interkvartilnim rasponom koji odgovara visini (ili duljini, ovisno o orijentaciji dijagrama) kutije. Na asimetriju podataka upućuje pomak medijana prema granicama kutije (kvartili Q_1 i Q_3) te nejednake duljine tzv. zalistaka. Ti su dijagrami još korisniji pri međusobnom uspoređivanju mjerenih svojstava nekoliko skupova podataka.



Slika 2.3. Primjer kutijastoga dijagrama na temelju podataka za Cu u tlu skupine B (Medunić i sur. 2009).

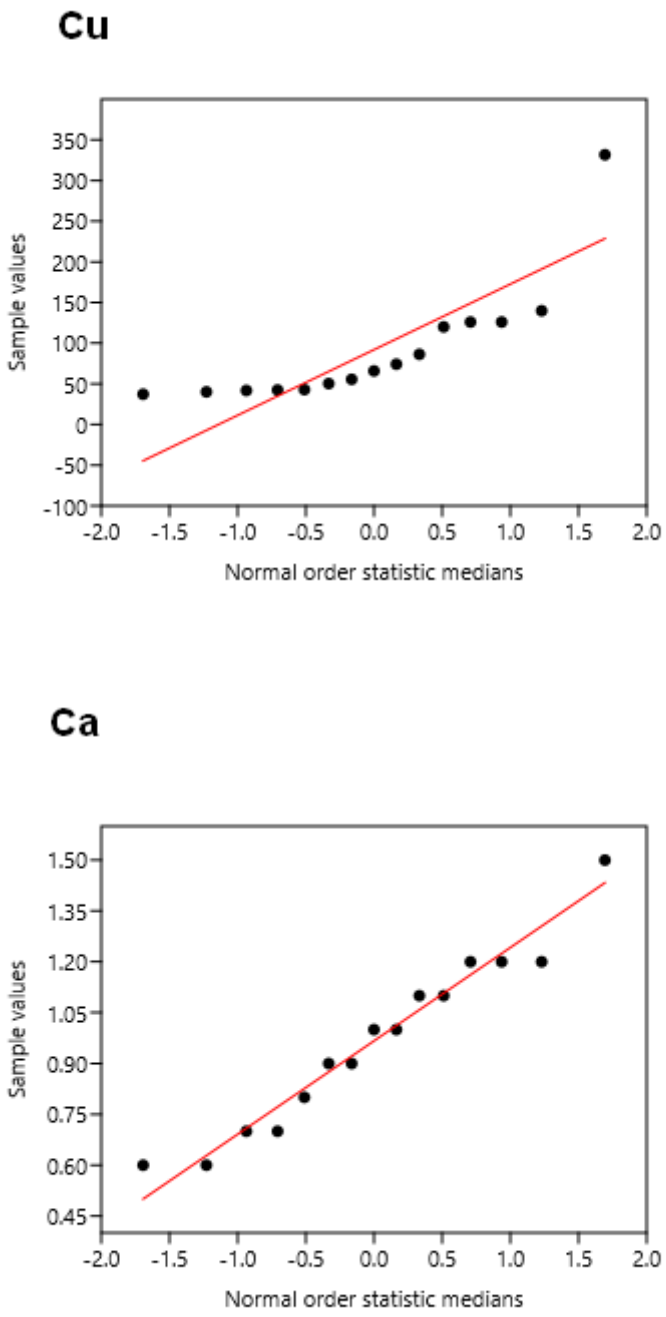
2.5.3. Dijagrami normalne vjerojatnosti

Dijagrami normalne vjerojatnosti (engl. *normal probability plots*) koriste se za određivanje vrste raspodjele podataka. Budući da je teoretska raspodjela prikazana kao ravan pravac, lako je uočiti odstupanja podataka od njega (sl. 2.4). Tri su tipične situacije u kojima dolazi do odstupanja od linearnosti:

1) asimetrija ili iskošenost raspodjele podataka

2) prisutnost ekstremnih vrijednosti i

3) jako naglašeni „repovi“ raspodjele podataka (velika zastupljenost podataka oko ekstremnih vrijednosti).



Slika 2.4. Prikaz pravaca normalne vjerojatnosti za Cu i Ca u skupini B. Vidi se da Cu, za razliku od Ca, znatno odstupa od teoretske normalne raspodjele na što upućuju veće udaljenosti nekih točaka od pravca (Medunić i sur. 2009).

3. Osnovni statistički parametri (opisna statistika)

Opisna (deskriptivna) statistika podrazumijeva sažimanje (sumiranje) niza podataka, što obično znači prikazivanje podataka u obliku tablica, dijagrama te računanje brojčanih pokazatelja središnjih vrijednosti i varijabilnosti. Osim toga na početku svake (geo)statističke analize podataka treba ispitati vrstu raspodjele podataka (simetrija ili odstupanje od nje) te uočiti ekstremne vrijednosti koje odskaču od ostalih.

3.1. Mjere središnje tendencije (lokacije) podataka

Nakon analitičkih mjerenja provedenih na uzorcima potrebno je dati kratak sažetak o informaciji koju ona pružaju. Riječ je o jednom jedinom broju koji na temelju tih mjerenja predstavlja vrijednost tipičnu za dotični uzorak, a to je **mjera središnje tendencije ili lokacije**. To je obično aritmetička sredina (engl. *mean*), medijan (engl. *median*), mod (engl. *mode*), a u nekim situacijama korisni su i kvartili (engl. *quartile*).

1) Aritmetička sredina dobar je osnovni statistički parametar u slučaju simetrično raspodijeljenih podataka. Jako je bitno imati na umu činjenicu da na aritmetičku sredinu jako utječu ekstremne vrijednosti, koje dovode do asimetrije podataka. Moguće rješenje toga problema jest upotreba tzv. odrezane sredine (engl. *truncated mean*) na način da se odstrani 25 % (ili neki drugi postotak) gornjih (visokih) i donjih (niskih) vrijednosti podataka.

2) Medijan je 50-i percentil (Q_2) koji dijeli niz rangiranih vrijednosti na dvije jednake polovice. Dakle, nalazi se točno u sredini svih vrijednosti poredanih (rangiranih) po veličini od najmanje do najveće. Time je preostalih 50 % vrijednosti (tj. pojedinih mjerenja) veće i 50 % manje od medijana. Riječ je o robusnoj mjeri lokacije jer na nju ne utječu ekstremne vrijednosti pa se ponajprije koristi za asimetrično raspodijeljene podatke.

3) Mod je najtipičnija ili najčešća vrijednost, tj. ona koja ima najveću učestalost (frekvenciju) pojavljivanja u nizu mjerenja. Katkad je kao rezultat moguće imati više takvih vrijednosti, npr. specifičnih geokemijskih procesa u uzvodnom drenažnom bazenu (Pavlović i sur. 2004).

4) Kvartili (donji Q_1 i gornji Q_3) zapravo su medijani donjih odnosno gornjih 50 % rangiranih podataka.

3.2. Mjere raspršenja ili varijabilnosti podataka

Mjera lokacije djelomice upućuju na položaj naših podataka u brojčanom smislu, npr. koliko iznosi aritmetička sredina vrijednosti Cu u tlu nekoga lokaliteta, što nam omogućuje da ih okvirno usporedimo s nekim drugim podatcima. Sama za sebe mjera lokacije ne ocrtava raspon vrijednosti u uzorku. Međutim, raspon ili „raspršenje“ mjerenja može biti od velike važnosti. Primjerice, tvornica ne smije ispuštati otpadnu vodu u kojoj je sadržaj suspendiranih čestica > 50 mg/kg. Ako se analizom utvrdi da je taj sadržaj 20 mg/kg (prosječna vrijednost), iz dotičnoga broja ne možemo zaključiti udovoljava li baš uvijek otpadna voda tom zakonu (npr. vrijednosti pojedinih mjerenja mogu iznositi 15, 45, 11, 22, 28, 29, 31, 41, 33 mg/kg). Stoga je potrebno poznavati **raspon** dotičnih vrijednosti. Dakle, slično mjerama lokacije, mjere raspona, raspršenja ili varijabilnosti (promjenjivosti) podataka su sljedeće.

1) Raspon (engl. *range*) = *maks* (najveća vrijednost u nizu podataka) – *min* (najmanja vrijednost u nizu podataka)

2) Standardna devijacija ili odstupanje (engl. *standard deviation, SD*) najbolja je mjera raspršenja podataka i uz aritmetičku sredinu predstavlja najčešće korišteni osnovni statistički parametar. Ona predstavlja prosječno odstupanje (devijaciju) podataka oko njihove aritmetičke sredine. Ekstremi na nju imaju još veći utjecaj nego na aritmetičku sredinu, tako da ona može dati nerealno veliku sliku raspršenja premda zapravo glavnina podataka može imati vrlo malo raspršenje.

Vrijednost standardne devijacije uvijek je veća u skupu heterogenijih (međusobno različitijih) podataka. Suprotno tomu, vrijednost standardne devijacije uvijek je manja u skupu homogenijih (međusobno sličnijih) podataka. Studente često zanima zašto je standardna devijacija toliko važna. Važna je zato što neizravno upućuje na vrstu raspodjele (simetrija nasuprot asimetriji) u određenom skupu podataka. Općenito, veća vrijednost SD znači da su podatci asimetrično raspodijeljeni, a manja SD znači da su podatci simetrično raspodijeljeni. U analizi skupa podataka važni su sljedeći parametri: a) mjere središnje tendencije kao što su aritmetička sredina i medijan; znajući središnju vrijednost i raspršenje podataka, možemo

dobro razumjeti i njihovu raspodjelu (simetrija vs asimetrija) i b) mjere raspršenja podataka, a to je prije svega standardna devijacija.

3) Interkvartilni raspon (*IQR*) kao mjera raspršenja kojom se mjeri raspon središnjih 50 % podataka grupiranih oko medijana, tj. od 25-og do 75-og percentila (odnosno od donjega Q1 do gornjega Q3 kvartila). Prednost mu je otpornost na ekstreme.

4) Koeficijent varijacije (engl. *coefficient of variation*, *CV*) ili relativna standardna devijacija (*RSD*) koristan je pokazatelj preciznosti (ili raspršenja) rezultata iskazanih u različitim mjernim jedinicama.

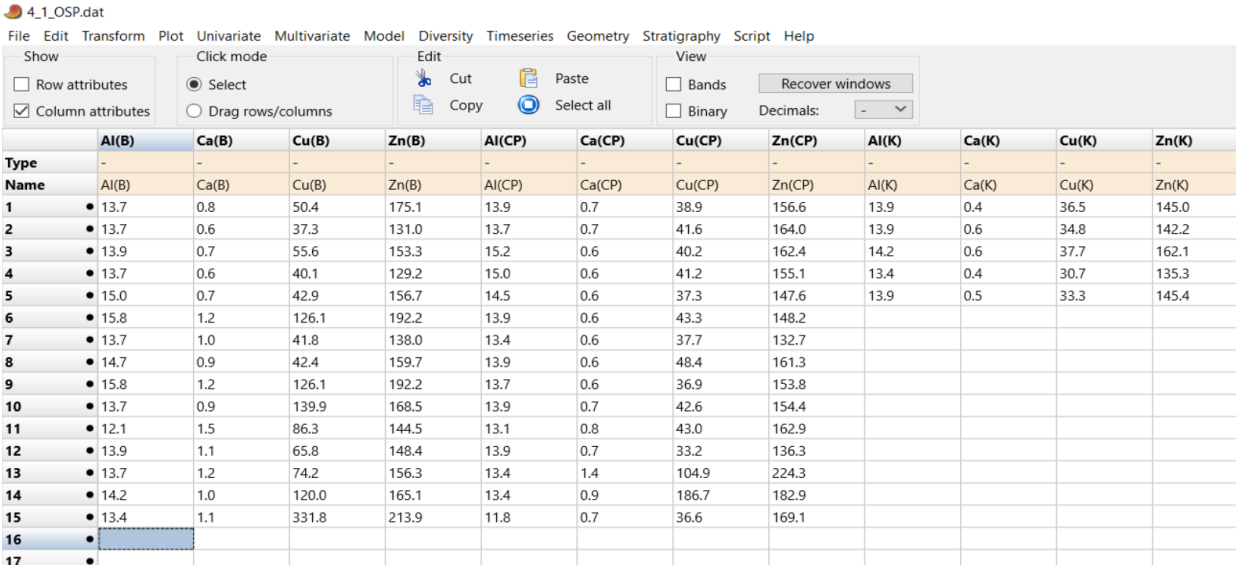
PRIMJER 3.1. Ovaj je primjer prikaz kako osnovne statističke parametre primijeniti na konkretnim podatcima iz okoliša preuzetima iz rada Medunić et al. 2009. Ukratko, riječ je o trima skupinama podataka:

1) podatci B – površinsko tlo (n = 15) iz dvorišta kemijske tvornice Bakrotisak

2) podatci CP – površinsko tlo (n = 15) oko susjednoga manjeg postrojenja za obradu (protektiranje) guma CroatiaProtect

3) podatci K – površinsko tlo (n = 5) iz šume (kontrolna, tj. lokacija koja nije pod izravnim utjecajem dviju tvornica) udaljene od B i CP 500-700 m.

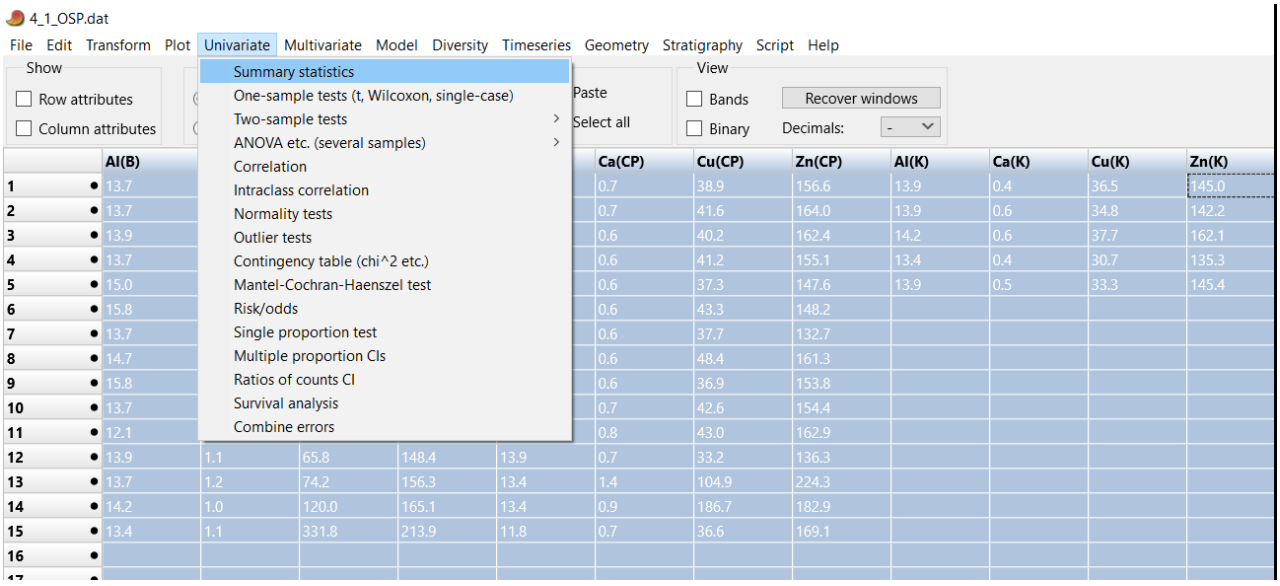
Cilj je dotičnoga istraživanja bio odrediti elementni sastav tla te na temelju statističke analize rezultata zaključiti je li tlo pod antropogenim utjecajem (B i CP) onečišćeno u odnosu na ono kontrolno (K), koje je uglavnom pod prirodnim utjecajem trošenja stijenske podloge. Radi jednostavnosti ovaj primjer bavit će se samo s četirima varijablama: Al i Ca (većinom prirodnoga porijekla) te Cu i Zn (potencijalno toksični elementi u tragovima koji mogu potjecati od antropogenih aktivnosti tvornica B i CP). Pokretanjem programa PAST (*Hammer* et al. 2001) otvorit će se radni list, slično kao u Excelu. Potrebno je prvo označiti *Column attributes* (sl. 3.1, kvačica u gornjem lijevom kutu) te unijeti podatke u stupce jedan pokraj drugoga.



	Al(B)	Ca(B)	Cu(B)	Zn(B)	Al(CP)	Ca(CP)	Cu(CP)	Zn(CP)	Al(K)	Ca(K)	Cu(K)	Zn(K)
Type	-	-	-	-	-	-	-	-	-	-	-	-
Name	Al(B)	Ca(B)	Cu(B)	Zn(B)	Al(CP)	Ca(CP)	Cu(CP)	Zn(CP)	Al(K)	Ca(K)	Cu(K)	Zn(K)
1	13.7	0.8	50.4	175.1	13.9	0.7	38.9	156.6	13.9	0.4	36.5	145.0
2	13.7	0.6	37.3	131.0	13.7	0.7	41.6	164.0	13.9	0.6	34.8	142.2
3	13.9	0.7	55.6	153.3	15.2	0.6	40.2	162.4	14.2	0.6	37.7	162.1
4	13.7	0.6	40.1	129.2	15.0	0.6	41.2	155.1	13.4	0.4	30.7	135.3
5	15.0	0.7	42.9	156.7	14.5	0.6	37.3	147.6	13.9	0.5	33.3	145.4
6	15.8	1.2	126.1	192.2	13.9	0.6	43.3	148.2				
7	13.7	1.0	41.8	138.0	13.4	0.6	37.7	132.7				
8	14.7	0.9	42.4	159.7	13.9	0.6	48.4	161.3				
9	15.8	1.2	126.1	192.2	13.7	0.6	36.9	153.8				
10	13.7	0.9	139.9	168.5	13.9	0.7	42.6	154.4				
11	12.1	1.5	86.3	144.5	13.1	0.8	43.0	162.9				
12	13.9	1.1	65.8	148.4	13.9	0.7	33.2	136.3				
13	13.7	1.2	74.2	156.3	13.4	1.4	104.9	224.3				
14	14.2	1.0	120.0	165.1	13.4	0.9	186.7	182.9				
15	13.4	1.1	331.8	213.9	11.8	0.7	36.6	169.1				
16												
17												

Slika 3.1. Izvorni podatci nekoliko odabranih varijabli (Al i Ca u %, Cu i Zn u mg/kg) mjerenih u uzorcima tla s lokacija B, CP i K (više detalja je u tekstu ovoga djela te u radu Medunić i sur. 2009).

Unesene podatke (stupci) treba označiti i potom → *Univariate*, *Summary statistics* (sl. 3.2).



Slika 3.2. Prikaz načina dobivanja osnovnih statističkih parametara odabranih varijabli u PAST-u.

Time se dobiju osnovni statistički parametri četiriju odabranih varijabli (sl. 3.3).

Univariate statistics

	Al(B)	Ca(B)	Cu(B)	Zn(B)	Al(CP)	Ca(CP)	Cu(CP)	Zn(CP)	Al(K)	Ca(K)	Cu(K)	Zn(K)
N	15	15	15	15	15	15	15	15	5	5	5	5
Min	12.1	0.6	37.3	129.2	11.8	0.6	33.2	132.7	13.4	0.4	30.7	135.3
Max	15.8	1.5	331.8	213.9	15.2	1.4	186.7	224.3	14.2	0.6	37.7	162.1
Sum	211	14.5	1380.7	2424.1	206.7	10.8	812.5	2411.6	69.3	2.5	173	730
Mean	14.06667	0.9666667	92.04667	161.6067	13.78	0.72	54.16667	160.7733	13.86	0.5	34.6	146
Std. error	0.2437146	0.06666667	19.52727	6.166132	0.2061437	0.05363013	10.44371	5.550611	0.128841	0.04472136	1.228007	4.413049
Variance	0.8909524	0.06666667	5719.717	570.3178	0.6374286	0.04314286	1636.067	462.1392	0.083	0.01	7.54	97.375
Stand. dev	0.9439027	0.2581989	75.62881	23.88133	0.7983912	0.2077086	40.44832	21.49742	0.2880972	0.1	2.745906	9.867877
Median	13.7	1	65.8	156.7	13.9	0.7	41.2	156.6	13.9	0.5	34.8	145
25 prcntil	13.7	0.7	42.4	144.5	13.4	0.6	37.3	148.2	13.65	0.4	32	138.75
75 prcntil	14.7	1.2	126.1	175.1	13.9	0.7	43.3	164	14.05	0.6	37.1	153.75
Skewness	0.321144	0.2532335	2.510206	0.7205729	-0.5152243	2.83863	3.007245	1.874184	-1.007859	0	-0.5002625	1.254118
Kurtosis	0.9709134	-0.3692308	7.492085	0.176447	2.234057	9.004312	9.19906	5.271789	2.550443	-3	-0.5334995	2.613314
Geom. mean	14.0373	0.9338983	74.41953	160.028	13.75791	0.6999505	47.1628	159.5752	13.85758	0.4919019	34.51113	145.7413
Coeff. var	6.710209	26.71023	82.16355	14.77744	5.793841	28.84841	74.67382	13.37126	2.078623	20	7.936145	6.75882

Slika 3.3. Prikaz osnovnih statističkih parametara odabranih varijabli po skupinama (B, CP, K).

Sada je potrebno proučiti dobivene vrijednosti međusobnom usporedbom nekoliko odabranih osnovnih statističkih parametara te njihovim tumačenjima. Promotrimo prvo Al i Ca: vrijednosti njihovih aritmetičkih sredina i medijana (gledano zasebno za svaki element) međusobno su slične (14 i 13,7 odnosno 0,97 i 1) što upućuje na simetričnu raspodjelu podataka i odsutnost ekstrema, a njihove SD (0,94 i 0,26) i RSD (6,7 i 26,7) upućuju na nisku do umjerenu varijabilnost koja je obično rezultat prirodnih geokemijskih procesa u okolišu (trošenje stijenske podloge). Osnovni statistički parametri za Cu i Zn (u skupinama B i CP) pokazuju da su njihove vrijednosti djelomice rezultat antropogenih procesa, koji obično dovode do povišene (ili jako visoke) varijabilnosti podataka (npr. Reimann i sur. 2008.). To se vidi iz sljedećega.

1) Vrijednosti aritmetičke sredine i medijana međusobno se razlikuju, više u slučaju Cu u B (92 nasuprot 66 odnosno 54 nasuprot 41 za Cu te 162 nasuprot 157 odnosno 161 nasuprot 157 za Zn), dok su im vrijednosti u K gotovo identične pa se skupina K s pravom može smatrati nezagađenim tlom.

2) Međusobno uspoređivanje vrijednosti SD ovih četiriju elemenata bilo bi svojevrsno uspoređivanje „krušaka, jabuka, banana i jagoda“ jer te vrijednosti dotičnih elemenata mogu biti različitih redova veličina (npr. Cu može imati raspon vrijednosti 30-50 mg/kg, a Zn 130-150 mg/kg;) te različite mjerne jedinice (mg/kg vs µg/kg).

Međutim, ima smisla uspoređivati npr. SD za Cu u trima skupinama podataka jer je riječ o istoj vrsti tla nad jednakom podinom i općenito sličnim geokemijskim procesima u tlu, s tim da je ovdje uključen i dodatni antropogeni čimbenik u tlima B i CP. Na antropogeni utjecaj upućuju velike vrijednosti SD za Cu i Zn u B i CP skupinama u odnosu na skupinu K. U svrhu usporedbe varijabilnosti četiriju promatranih elementa u trima skupinama podataka (B, CP i K) koristi se RSD, a sl. 3.3 pokazuje da najveću i najmanju varijabilnost imaju Cu (tipični kemijski element koji može upućivati na prisutnost antropogenih aktivnosti na nekom lokalitetu) u skupini B odnosno Al (tipični referentni element koji služi za tumačenje prirodnih geokemijskih procesa u tlima) u skupini K. Ti rezultati potvrđuju radnu hipotezu o vjerojatnoj onečišćenosti tla B potencijalno toksičnim elementima u tragovima (npr. Cu, Zn, Pb, Ni, itd.) i prirodnom geokemijskom sastavu tla K. Ako promatramo samo Cu i Zn, vidimo da je varijabilnost Cu daleko veća u odnosu na Zn, prije svega u skupinu B, a najmanje u skupini K. Razlog je visoke varijabilnosti Cu nekoliko ekstrema (neobično visoke vrijednosti Cu u tlu B, tj. 120 – 331 mg/kg) koji su povećali ne samo aritmetičku sredinu nego još više i SD. Njihovim uklanjanjem vrijednosti aritmetičke sredine i medijana postaju sličnije i iznose 54 odnosno 47 mg/kg, a RSD se osjetno smanji na 30 %. Time možemo zaključiti da nekoliko osnovnih statističkih parametara (aritmetička sredina, medijan, SD i RSD) služi kao svojevrsna brza dijagnostika općega stanja podataka. Njihovom brzom vizualnom usporedbom po skupinama možemo steći preliminarni uvid u glavne geokemijske procese u uzorcima te suziti izbor varijabli i/ili skupina, koje ćemo dalje obraditi primjerenim statističkim metodama da bismo iz njih izvukli bitne činjenice.

3.3. Normalna, simetrična ili Gaussova raspodjela podataka

Normalna raspodjela podataka središnji je pojam u statistici. Otkrili su je neovisno jedan od drugoga De Moivre, Laplace i Gauss, ali naziva se prema potonjem (Gaussova raspodjela). Njezina su osnovna obilježja sljedeća:

– približno 68 % vrijednosti populacije nalazi se unutar ± 1 SD oko aritmetičke sredine

– približno 95 % vrijednosti populacije nalazi se unutar ± 2 SD oko aritmetičke sredine

– približno 99,7 % vrijednosti populacije nalazi se unutar ± 3 SD oko aritmetičke sredine

– krivulja raspodjele je zvonolika

– aritmetička sredina, medijan i mod imaju identičnu vrijednost, a smješteni su u središtu krivulje

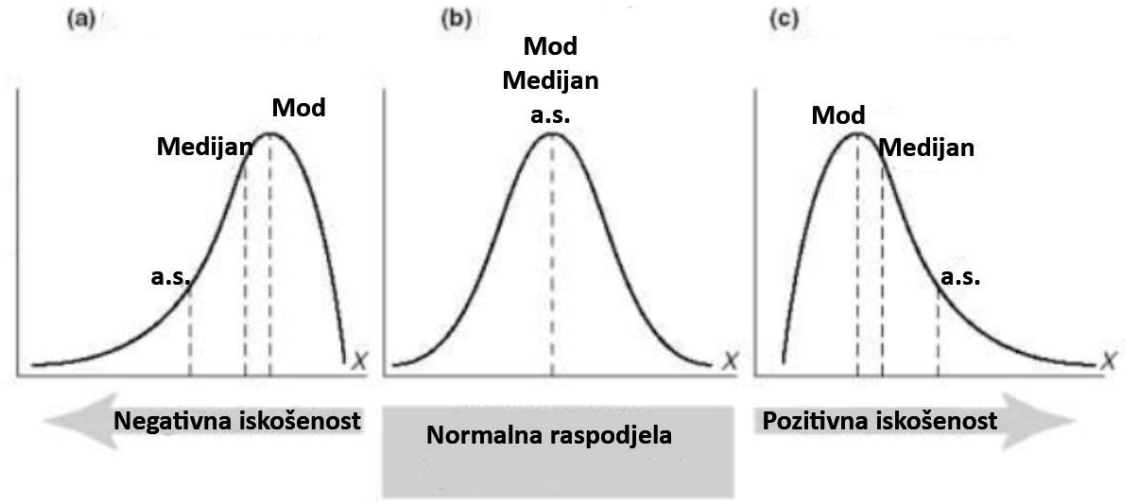
– krivulja je unimodalna

– krivulja je simetrična oko aritmetičke sredine

– krivulja je kontinuirana i nikada ne dodiruje os x

– ukupna površina pod krivuljom jednaka je 1,00 ili 100 %.

U geoznanostima (npr. volumen naftnih polja, koncentracija Se u ugljenu, koncentracija Ca u podzemnoj vodi, granulometrijski sastav sedimenta itd.) česta je pojava da raspodjele podataka jako odstupaju od simetrične pa se nazivaju asimetrične (sl. 3.4) ili iskošene (engl. *skewed*).



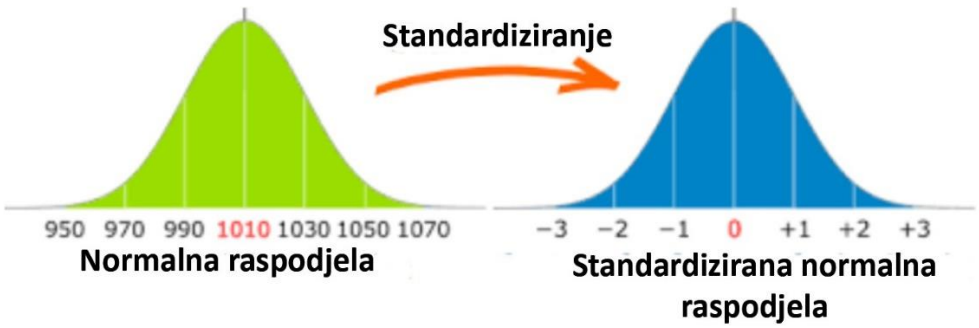
Slika 3.4. Prikaz asimetrične ili iskošene (engl. *skewed*) raspodjele podataka (a i c) te simetrične, normalne (b) raspodjele podataka.

Kod asimetrične raspodjele aritmetička sredina pomaknuta je prema „repu“, a nekoliko neobičnih vrijednosti povećava standardnu devijaciju. Zbog toga ta dva osnovna statistička parametra (tj. aritmetička sredina i SD) ne mogu opisati glavninu asimetrično raspodijeljenih podataka pa je u tom slučaju daleko bolje koristiti medijan i kvartile. Primjena parametarskih

testova na asimetrično raspodijeljenim podatcima također je upitna jer oni zahtijevaju simetričnost podataka. Stoga za asimetrično raspodijeljene podatke nije moguće dati stabilne procjene i zaključke ni pouzdano tumačenje parametarskim metodama. Ako se takva opažanja logaritmiraju (log baza 10) i ako te logaritmirane vrijednosti slijede normalnu raspodjelu, za takve se varijable kaže da su **lognormalne**.

U početcima razvoja statistike smatralo se da svi podatci iz prirode slijede simetričnu (zvonoliku) normalnu krivulju. Ako to nije bio slučaj, proces skupljanja podataka bio je izložen sumnji. Istina je da je univerzalnost normalne krivulje samo mit, a primjerima potpuno nenormalnih (asimetričnih ili iskošenih) raspodjela obiluju gotova sva znanstvena polja. Unatoč tomu normalna raspodjela igra glavnu ulogu u statistici, a postupci zaključivanja derivirani iz nje imaju široku primjenu i tvore temelj aktualnih metoda statističke analize (npr. Reimann i sur. 2008).

Bilo koja raspodjela podataka (sl. 3.5) može se u potpunosti odrediti s pomoću svoje aritmetičke sredine (μ; taj simbol označava aritmetičku sredinu populacije) i standardne devijacije (σ; taj simbol označava standardnu devijaciju populacije) koje se nalaze u formuli za funkciju vjerojatnosti gustoće podataka.

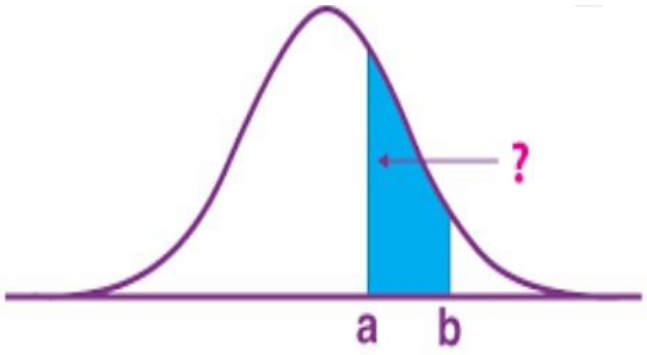


Slika 3.5. Slikoviti prikaz matematičke pretvorbe bilo koje normalne raspodjele podataka u standardizirani normalni oblik.

Krivulja normalne raspodjele simetrična je oko svoje μ koja predstavlja njezin vrh. Interval μ ± 1σ ima vjerojatnost 0,683 (tj. 68,3 %), interval μ ± 2σ ima vjerojatnost 0,954 (tj. 95,4 %), a onaj za μ ± 3σ iznosi 0,997 (tj. 99,7 %). To znači da ako je odstupanje pojedinačnoga mjerenja > 3σ, tada postoji svega 0,03 % vjerojatnosti da ono ne uđe u populaciju određenu N-raspodjelom.

Normalna raspodjela, koja ima $\mu = 0$, a $\sigma = 1$, naziva se **standardizirana normalna raspodjela** (sl. 3.5). Slučajna varijabla Z koja ima $\mu = 0$ i $\sigma = 1$ naziva se **standardizirana normalna varijabla**. Označava odstupanje nekoga podatka od aritmetičke sredine izraženo u jedinicama standardne devijacije $\mathbf{z = (x_i - \mu)/\sigma}$.

Za takvu normalnu raspodjelu s poznatom aritmetičkom sredinom μ i standardnom devijacijom σ točan udio vrijednosti koje leže unutar bilo kojega intervala mogu se pronaći u **tablici standardiziranih normalnih vjerojatnosti** (moguće ju je pronaći u svakom udžbeniku iz statistike te na internetu). U njoj su dane površine pod krivuljom (odn. udjeli vrijednosti) lijevo od neke specifične vrijednosti z tako da vrijedi $P [Z \leq z]$ = površina pod krivuljom lijevo od z. Stoga vrijedi da je vjerojatnost intervala [a, b] (sl. 3.6) $P [a \leq Z \leq b]$ = [površina lijevo od b] – [površina lijevo od a].



Slika 3.6. Površina intervala [a, b] pod standardiziranom normalnom krivuljom.

Sljedeća obilježja mogu se derivirati iz simetrije vjerojatnosti oko 0.

a) $P [Z \leq 0] = 0,5$

b) $P [Z \leq -z] = 1 – P [Z \leq z]$ ili $P [Z > z] = 1 – P [Z \leq z]$ ➔ $P [Z \leq -z] = P [Z > z]$

c) ako je $z > 0$ $P [Z \leq z] = 0,5 + P[0 < Z \leq z]$ ili $P [Z \leq -z] = 0,5 – P [0 < Z \leq z]$.

Primjerice, udio vrijednosti ispod (ili lijevo od) $z = 2$ iznosi $F(2) = 0,9772$, a udio vrijednosti ispod $z = -2$ iznosi $F(-2) = 0,0228$. Stoga točna vrijednost udjela mjerenja koja leže unutar 2 standardne devijacije oko aritmetičke sredine iznosi $0,9772 – 0,0228 = 0,9544$ (tj. 95,4 %).

4. Opisivanje pouzdanosti procjene

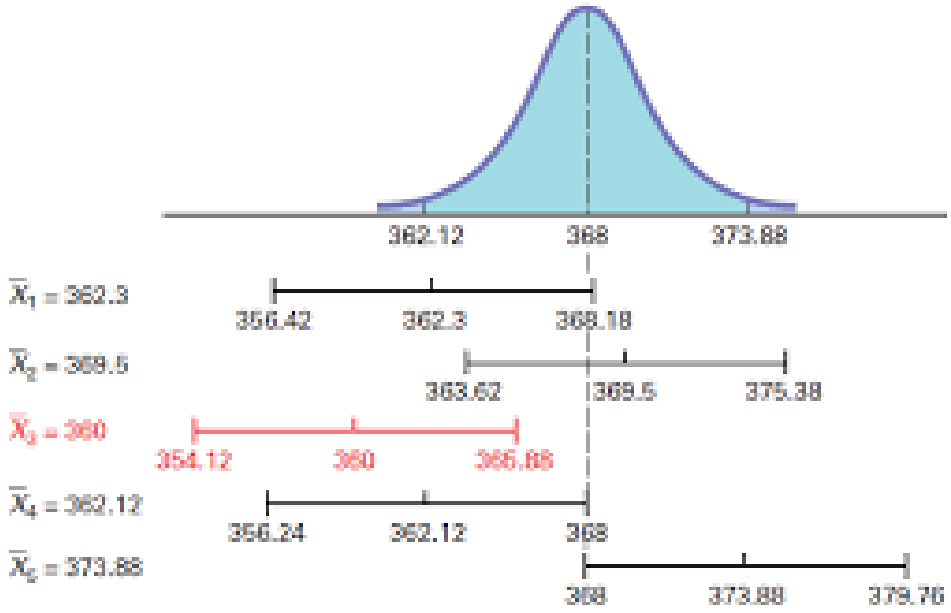
U ovom poglavlju opisuje se neizvjesnost odnosno pouzdanost osnovnih statističkih parametara kojima procjenjujemo mjere središnje tendencije. Umjesto prikazivanja pojedinačnih vrijednosti osnovnih statističkih parametara ovdje se pokazuje korisnost rada s rasponima vrijednosti poznatima kao **intervalne procjene** (engl. *interval estimates*). Ti intervali mogu poslužiti za testiranje značajnosti razlike nekoga parametra populacije u odnosu na neku prethodno definiranu vrijednost (npr. Helsel i sur. 2020).

Uzmimo za primjer središnju vrijednost koncentracije NO_3^-u plitkom vodonosniku ispod poljoprivrednoga tla koja izračunom iznosi 5,1 µg/L. Pitanje je koliko je pouzdana ova procjena, tj. znači li vrijednost dotičnoga parametra prekoračenje zdravstveno propisana ograničenja od 5 µg/L? Treba li taj vodonosnik tretirati drukčije nego neki drugi u kojem ta koncentracija iznosi 4,8 µg/L (Helsel i sur. 2020)?

4.1. Definicija intervalne procjene

S obzirom na to da su medijan i aritmetička sredina uzorka procjene odgovarajućih središnjih točaka populacije, dotične procjene nazivaju se još i **točkaste procjene populacije** (engl. *point estimates*). One ne prikazuju pouzdanost odnosno varijabilnost tih procjena. Primjerice, dva skupa podataka A i B (s istim brojem uzoraka) imaju srednju vrijednost 5. Međutim, podatci skupine B imaju vrijednosti vrlo slične broju 5, dok su podatci skupine A daleko varijabilniji (daleko različitiji od broja 5). Stoga je točkasta procjena 5 za skupinu A daleko manje pouzdana nego za grupu B zbog veće varijabilnosti A podataka.

Alternativa točkastim procjenama jesu **intervalne procjene** (engl. *interval estimates*), a to su intervali s vjerojatnošću da sadrže pravu (istinsku) vrijednost nekoga osnovnog statističkog parametra populacije (sl. 4.1). Oni su širi za nizove podataka s većom varijabilnošću. Stoga za interval 4,7 – 5,3 vrijedi 95 % vjerojatnost da sadrži (nepoznatu) pravu srednju vrijednost populacije B. Ista vjerojatnost zahtijevala bi daleko veći interval, recimo 2 – 8 za nalaženje prave srednje vrijednosti populacije A. Razlika u pouzdanosti dviju procjena srednjih vrijednosti populacija jasno se vidi iz veličina intervalnih procjena.



Slika 4.1. Intervali pouzdanosti (npr. 95 %) vrijednosti aritmetičke sredine populacije. Od njih pet jedan (označen crvenom bojom) neće sadržavati dotični parametar 95 % vremena.

Intervalne procjene mogu, za razliku od točkastih procjena, pružiti dvije informacije (Helsel i sur. 2020): 1) utvrđenu (navedenu) vjerojatnost da interval sadrži pravu srednju vrijednost populacije (njezina pouzdanost) – to su **intervali pouzdanosti** (engl. *confidence interval*); 2) utvrđenu vjerojatnost da neki (novi) podatak naznačene veličine pripada istraživanoj populaciji – to su **intervali predviđanja** (engl. *prediction interval*).

4.2. Tumačenje intervalnih procjena

Vjerojatnost da interval pouzdanosti doista sadrži pravu vrijednost populacije naziva se **razina pouzdanosti** (engl. *confidence level*) intervala (npr. 95 %). Vjerojatnost da taj interval ne sadrži pravu vrijednost populacije, poznat kao **alfa nivo (α)**, računa se prema formuli α = 1 – nivo pouzdanosti; (npr. 1 – 0,95 = 0,05, ili 100 – 95 = 5 %). Širina intervala pouzdanosti ovisi o obliku raspodjele podataka (simetrično vs asimetrično), veličini uzorka (n) i željenom nivou pouzdanosti (95 %-tni interval pouzdanosti širi je od onoga 90 %-tnog).

Najčešće se računaju simetrični intervali pouzdanosti srednje vrijednosti (aritmetička sredina ili medijan) populacije, pod pretpostavkom da podatci imaju normalnu raspodjelu. Ako su podatci asimetrični ili uzorak ima $n < 50$, simetrični intervali pouzdanosti neće sadržavati pravu srednju vrijednost populacije $(1 – \alpha)$ % vremena. Što je veća asimetrija, to broj uzoraka mora biti veći da bismo se mogli osloniti na simetrične intervale pouzdanosti. Alternativno je moguće izračunati asimetrične intervale pouzdanosti u uobičajenim situacijama asimetrično raspodijeljenih podataka.

4.3. Intervali pouzdanosti za srednju vrijednost

Prava srednja vrijednost populacije označava se s μ (srednja je vrijednost uzorka aritmetička sredina). Raspodjela srednjih vrijednosti uzoraka prilično se dobro aproksimira normalnom raspodjelom ako veličine uzoraka postaju sve veće unatoč tomu što podatci ne moraju biti striktno normalno raspodijeljeni, što je poznato kao središnji granični teorem. Međutim, u slučaju manjih uzoraka (maloga broja podataka) srednja vrijednost neće *a priori* podrazumijevati normalnu raspodjelu ako sami podatci nisu normalno raspodijeljeni. Što su podatci asimetričniji, to je potrebno više podataka da bi srednja vrijednost mogla biti aproksimirana kao podatak iz normalne raspodjele. Kod jako iskošenih raspodjela ili u slučaju ekstrema može se dogoditi da je potrebno više od 100 mjerenja kako bi srednja vrijednost ostala netaknuta najvećim (ekstremnim) vrijednostima i da bi se udovoljilo zahtjevu simetričnosti raspodjele podataka (npr. Davis 1986; Pentecost 1999; Helsel i sur. 2020). Kod standardizirane normalne raspodjele, tzv. z-raspodjele (sredina je 0, a standardna devijacija 1), vrijednost „z“ za različite intervale pouzdanosti 90, 95, 99 i 99,7 % iznose 1,65, 1,96, 2,58 odnosno 2,97.

5. Testiranje nulte hipoteze (H_0)

Znanstvenici skupljaju podatke da bi spoznali procese i sustave koje ti podatci predstavljaju. Oni često imaju prethodne ideje, zvane hipoteze, o tome kako se sustavi ponašaju. Jedna od glavnih svrha skupljanja podataka jest testiranje odnosno podupiranje tih hipoteza. Statistički su testovi kvantitativni načini utvrđivanja utemeljenosti tih hipoteza (npr. Davis 1986; Pentecost 1999; Helsel i sur. 2020).

Statistički testovi koji pretpostavljaju da podatci imaju neku određenu raspodjelu (obično normalnu) nazivaju se **parametarski** testovi. To je zbog toga što se informacija sadržana u podatcima može sažeti s pomoću osnovnih statističkih parametara (uglavnom su to aritmetička sredina i SD). Testovi koji ne zahtijevaju pretpostavku o određenoj raspodjeli podataka nazivaju se **neparametarski** testovi, pri čemu se informacija iz podataka izvlači usporedbom rangova odgovarajućih vrijednosti skupina podataka, a ne računanjem osnovnih statističkih parametara.

Varijable se dijele na ovisne i neovisne. Ovisna je ona čija se varijabilnost istražuje (npr. koncentracija Cu triju skupina tla), a neovisna varijabla (npr. vrsta tla, prisutnost antropogenih čimbenika itd.) služi za objašnjavanje zašto i kako se mijenja veličina ovisne varijable.

5.1. Struktura statističkih testova

Prvo je potrebno izabrati primjeren test. Prvi je kriterij odabira testa mjerna skala, drugi je cilj koji se želi postići testom, a treći je izbor između parametarskih i neparametarskih testova. Općenito vrijedi da neparametarski testovi nisu nikada lošiji od parametarskih po svojoj sposobnosti da otkriju odstupanje od početne (radne) nulte hipoteze (H_0) te mogu čak biti i daleko bolji u tome. Dakle, ako se radi o iskošenosti podataka i ekstremima, obilježjima uobičajenima u geoznanostima, neparametarski testovi mogu imati veću snagu od parametarskih. Najveća je snaga parametarskih metoda u modeliranju i procjenama kao što je regresijska analiza (npr. Helsel i sur. 2020).

Zatim je potrebno postaviti nultu hipotezu i alternativnu hipotezu. Dotične hipoteze trebaju biti postavljene prije skupljanja podataka, a one su sažet prikaz ciljeva istraživanja. **Nulta hipoteza (H_0)** jest ono što se pretpostavlja da vrijedi za istraživani sustav prije skupljanja

podataka, sve dotle dok oni ne pokažu suprotno. To obično znači nulto stanje u tri uobičajena slučaja kod većine istraživanja:

1) raspodjela varijable je normalna
2) nema razlike među skupinama podataka i
3) nema odnosa među varijablama unutar iste skupine podataka.

Alternativna hipoteza (H_1) situacija je za koju se očekuje da vrijedi ako podatci pokažu da istinitost nulte hipoteze ima malu vjerojatnost. Postoje dva opća testa: jednostrani i dvostrani. Jednostrani testovi koriste se kad bi odstupanje od nulte hipoteze samo u jednom smjeru dovelo do njezina odbacivanja te prihvaćanja alternativne hipoteze; npr. ako nulta hipoteza glasi „koncentracija Cu manja je ili jednaka zakonski propisanoj“, onda alternativna glasi „koncentracija Cu veća je od zakonski propisane“. Dvostrani testovi koriste se kad je očito da u bilo kojem smjeru od nulte hipoteze (veće ili manje, pozitivno ili negativno) dolazi do njezina odbacivanja i prihvata alternativne hipoteze (npr. ako imamo dokaz da je „koncentracija Cu veća od zakonski propisane“ ili „manja od zakonski propisane“, u oba slučaja imamo sumnju o točnosti nulte hipoteze).

Potom je potrebno odabrati prihvatljivu mjeru pogreške α. Vrijednost **α** ili **razina značajnosti** je vjerojatnost pogrešna odbacivanja nulte hipoteze kad je ona zapravo istinita, što se naziva „greška tipa I“. Obično se za α uzima nivo 0,05 (tj. 5 %).

Nakon obavljena statističkoga testa slijedi proučavanje rezultirajućih vrijednosti p, koje nisu ništa drugo nego **postignuta razina značajnosti** izračunata na temelju konkretnih podataka. Te vrijednosti predstavljaju vjerojatnost dobivanja rezultata statističkoga testa koji bi upućivali na to da vrijedi nulta hipoteza. Što je p manji, to je snažniji dokaz za odbacivanje nulte hipoteze i prihvata one alternativne.

Vrijednost α ne ovisi o podatcima, nego određuje rizik nastanka pogreške tipa I; to je kritična vrijednost za donošenje odluke „da/ne“. Vrijednost p daje potpuniju informaciju i to je u biti jačina (snaga, moć) znanstvenoga dokaza. Dakle, H_0 odbacujemo ako je $p < \alpha$ (npr. Helsel i sur. 2020).

5.2. Testiranje normalnosti raspodjele podataka

Glavni je razlog testiranja vrste raspodjele podataka utvrditi jesu li podatci normalno raspodijeljeni jer samo na takvima smijemo koristiti parametarske testove. Nulta hipoteza glasi: „podatci su normalno raspodijeljeni“. Preporučuje se korištenje većega α nivoa (npr. 0,10) da bi se lakše otkrila asimetričnost raspodjele, osobito u slučaju malih uzoraka.

Računalni programi koriste nekoliko metoda, od kojih se neke temelje na pravcima normalne vjerojatnosti, a najčešći je test Shapiro-Wilk uz uvjet ako je $p < 0,10$, odbacujemo nultu hipotezu.

Sada možemo testirati vrstu raspodjele za Al, Ca, Cu i Zn u trima skupinama tla (B, CP i K) s pomoću programa PAST. Označimo njihove stupce (podatci) → *Univariate* → *Normality tests* te potom promotrimo rezultate toga testa (p vrijednosti Shapiro-Wilk testa, tj. p (normal), sl. 5.1).

Tests for normal distribution

	Al(B)	Ca(B)	Cu(B)	Zn(B)	Al(CP)	Ca(CP)	Cu(CP)	Zn(CP)	Al(K)	Ca(K)	Cu(K)	Zn(K)
N	15	15	15	15	15	15	15	15	5	5	5	5
Shapiro-Wilk W	0.8758	0.9529	0.6945	0.9485	0.9048	0.6091	0.4975	0.8264	0.8516	0.8208	0.9758	0.8868
p(normal)	0.0411	0.5713	0.000217	0.501	0.1126	3.09E-05	3.356E-06	0.008183	0.1995	0.1185	0.9107	0.341
Anderson-Darling A	0.9837	0.2725	1.488	0.3016	0.7004	2.143	3.343	0.9271	0.5311	0.4376	0.1655	0.4061
p(normal)	0.009669	0.6159	0.0004701	0.5333	0.05315	9.373E-06	7.524E-09	0.01358	0.08562	0.1636	0.8724	0.2046
p(Monte Carlo)	0.0102	0.6553	0.0002	0.563	0.0543	0.0001	0.0001	0.0129	0.0871	0.1803	0.9454	0.2273
Lilliefors L	0.2367	0.1164	0.2346	0.1318	0.2403	0.3384	0.4234	0.2403	0.3552	0.2413	0.1555	0.3242
p(normal)	0.02333	0.8428	0.02565	0.6815	0.01998	0.0001	0.0001	0.01991	0.03802	0.4447	1.221	0.08465
p(Monte Carlo)	0.0244	0.8426	0.0219	0.6787	0.0204	0.0001	0.0001	0.0194	0.032	0.4471	0.9641	0.0807
Jarque-Bera JB	0.2642	0.3788	27.33	1.084	1.406	37.94	40.93	13.86	0.4083	0.638	0.3615	0.6148
p(normal)	0.8763	0.8275	1.16E-06	0.5815	0.4952	5.781E-09	1.294E-09	0.0009788	0.8153	0.7269	0.8347	0.7353
p(Monte Carlo)	0.864	0.7967	0.0005	0.3184	0.1914	0.0001	0.0001	0.0031	0.7109	0.3123	0.7867	0.3598

Slika 5.1. Rezultati statističkih testova vrste raspodjele podataka četiriju varijabli u trima grupama tla (oznake B, CP i K u zagradama).

Vidimo da su p vrijednosti Shapiro-Wilkova testa manje od 0,10 u slučaju Al (B), Cu (B) te Ca, Cu i Zn u skupini CP; za njih odbacujemo H_0 i prihvaćamo alternativnu hipotezu koja glasi da su njihove raspodjele asimetrične. U skupini K sve četiri varijable imaju p vrijednosti >0,10 te prihvaćamo H_0 koja glasi da su te varijable normalno raspodijeljene.

U ovom slučaju (tablica 5.1.) uputno je sve varijable triju skupina podataka obraditi neparametarskim metodama, ali u svrhu učenja koristimo i parametarske te uspoređujemo rezultate obiju metoda.

	B	CP	K
Al	A	S	S
K	A	A	A
Ca	S	A	S
Ti	S	S	S
V	S	S	S
Cr	A	S	S
Mn	A	A	S
Fe	S	A	S
Co	S	S	A
Ni	S	S	S
Cu	A	A	S
Zn	S	A	S
Pb	S	A	S
Ga	S	A	S
As	A	S	S
Rb	S	S	S
Sr	S	S	S
Y	A	S	S
Zr	S	S	S
OT	S	S	S

Tablica 5.1. Vrste raspodjele podataka za glavne, sporedne i elemente u tragovima te za udio organske tvari (OT) u trima skupinama tla (B, CP i K). S – simetrična raspodjela ($p > 0,10$), A – asimetrična raspodjela ($p < 0,10$).

Dakle, pri testiranju vrste raspodjele analiziranih varijabli u grupi B uočeno je jako odstupanje od simetrije u slučaju Cu ($p<0,001$), umjereno (p vrijednosti su bliže 0,05) u slučaju Al, K, Cr, Mn, As i Y, dok ostatak varijabli pokazuje simetričnu raspodjelu (npr. Zn, $p = 0,50$). Asimetrija u slučaju Cu upućuje na onečišćenje tla B bakrom koji vjerojatno potječe od industrijskih aktivnosti u kemijsko-prerađivačkoj tvornici B. Što se tiče skupine CP, najveće odstupanje od normalnosti ponovno je uočeno za Cu ($p < 0,001$), umjereno u slučaju Zn ($p = 0,008$), K, Ca, Mn, Fe, Pb i Ga, dok ostale varijable imaju normalnu raspodjelu. Ti se rezultati mogu pripisati prirodnim geokemijskim procesima koji se obično događaju u profilima tla izloženima trošenju u površinskim uvjetima, a u manjoj mjeri i onečišćenju bakrom koje je uglavnom antropogena porijekla. U skupini K, s izuzetkom kalija i kobalta, sve varijable imaju simetričnu raspodjelu, što upućuje na neporemećen geokemijski sastav dotičnoga tla u kojem one imaju slične, usko raspršene vrijednosti (npr. Medunić i sur. 2009).

6. Testiranje razlika dviju neovisnih skupina podataka

Sljedeći primjer opisuje dvije skupine podataka čija se povezanost ispituje. Izvori uzvodno i nizvodno od opasnoga odlagališta otpada uzorkovani su u svrhu ispitivanja je li koncentracija nekih toksičnih sastojaka veća u nizvodnom smjeru kao posljedica negativna utjecaja otpada. Ako jest, je li to na nivou značajnosti $\alpha < 0,01$ ili 0,05? Ako se te sumnje potvrde testom, podzemna će se voda proglasiti zagađenom. Dakle, ovdje se radi o usporedbi dviju neovisnih skupina podataka u smislu da ne postoji prirodna struktura u poretku opažanja kroz skupine, na primjer, najniži podatak iz skupine A ne može se spariti (tj. dovesti u neku vezu) s najnižim podatkom iz skupine B. Neparametarska metoda analize je Mann-Whitneyjev U test, a parametarska t-test.

S t-testom povezano je pet problema zbog kojih on ima slabiju snagu u odnosu na neparametarski test ako se primjenjuje na asimetrično raspodijeljenim podatcima iz prirode, a oni su sljedeći:

1) zadovoljava uvjet normalne razdiobe podataka

2) matematički algoritam temelji se na ovisnosti o aditivnom modelu (podrazumijeva da su rangirani podatci jedne skupine veći od istovjetno rangiranih one druge za isti pribrojnik)

3) nemogućnost primjene na cenzuriranim (ispod praga detekcije) podatcima

4) temelji se na pretpostavci da je sredina dobra mjera središnje tendencije podataka, što ne vrijedi za asimetrično raspodijeljene podatke te

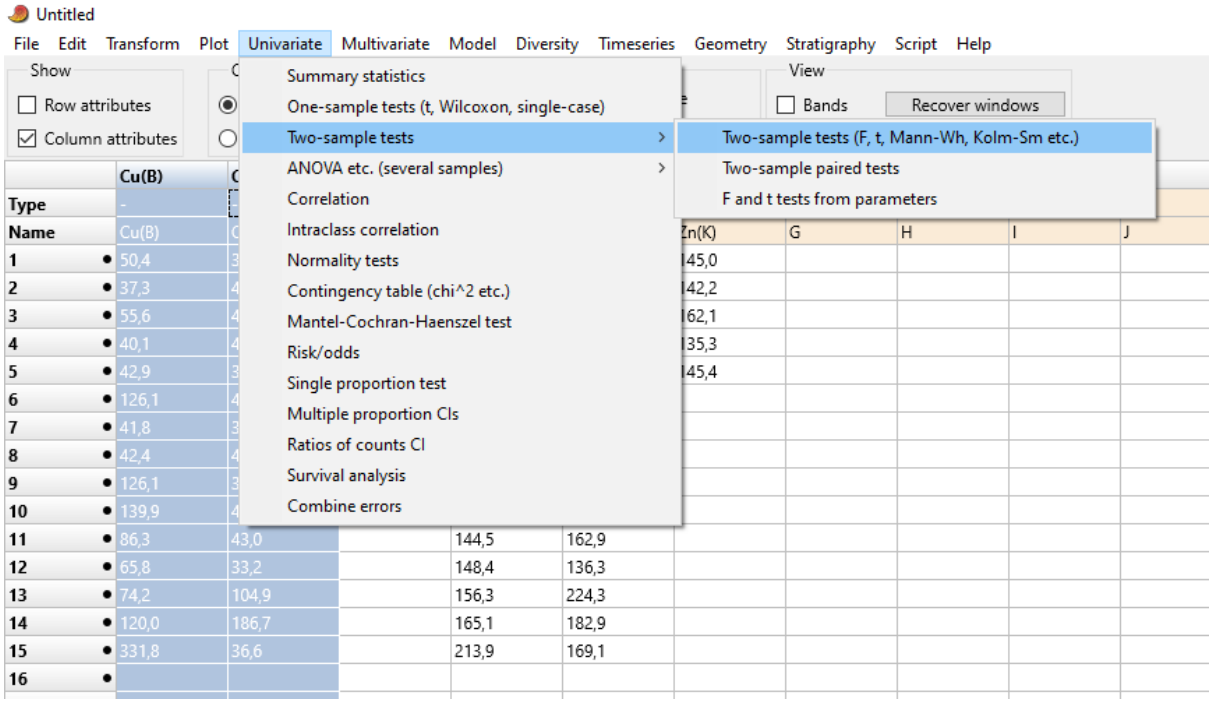
5) poteškoće u otkrivanju asimetrije podataka i nejednakosti njihovih varijanci u slučaju malobrojnih uzoraka (Helsel i sur. 2020).

Zbirno taj test podrazumijeva da su podatci normalno raspodijeljeni oko svojih aritmetičkih sredina i da imaju jednake varijance. Time bi se podatci dviju skupina trebali razlikovati jedino po svojim srednjim vrijednostima.

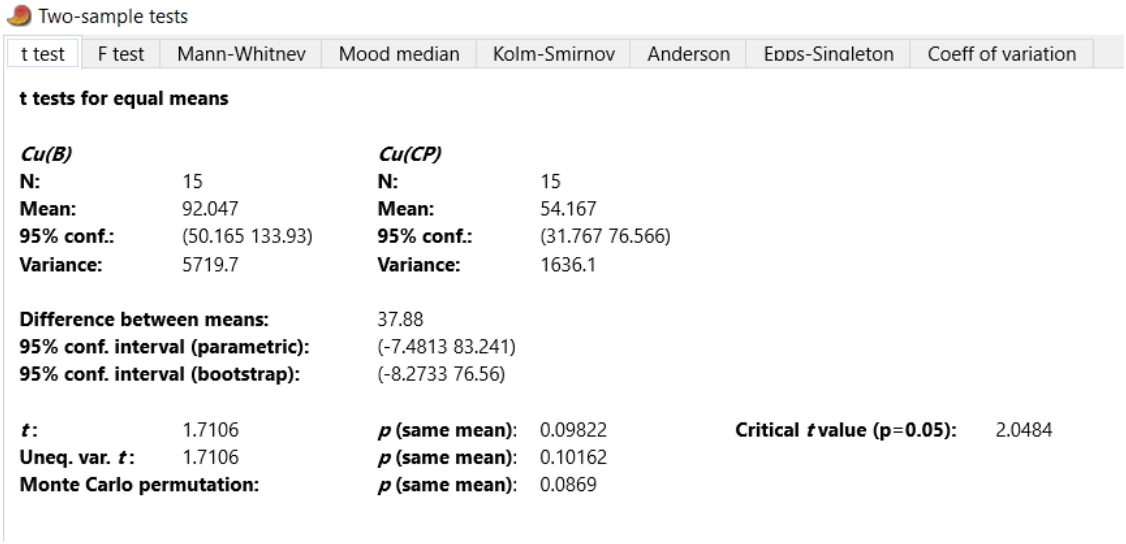
Kad se parametarski testovi primjenjuju na asimetrično raspodijeljenim podatcima, njihova moć otkrivanja razlika koje doista postoje među dvjema skupinama daleko je slabija od njihovih neparametarskih inačica. „Iskošenost“ podataka i ekstremi uvećat će standardnu devijaciju koja se koristi u t-testu, a glavnina asimetrično raspodijeljenih podataka može čak imati manju raspršenost od usporedivih simetričnih. Nadalje, za podatke ispod praga detekcije obvezno treba koristiti neparametarski test jer srednja vrijednost nije otporna na ekstreme, a ne mora biti niti blizu medijanu ili 50-om percentilu.

U slučaju malobrojnih uzoraka (n < 30), kakvi su česti u geološkim istraživanjima, uputno je obaviti oba testa ili samo onaj neparametarski. Grafički prikaz rezultata najbolje je obaviti usporednim kutijastim dijagramima, koji vrlo učinkovito i sažeto prikazuju osnovna obilježja skupina podataka kao i razlike među njima (npr. Kang i sur. 2022).

PRIMJER 6.1. Ispituju se razlike među parovima skupina (B – CP, CP – K te B – K) s obzirom na Cu i Zn s pomoću t-testa i Mann-Whitneyjeva testa u programu PAST. Napomena: uspoređuju se dvije skupine s obzirom na istu varijablu (npr. Cu), a ne npr. Cu u skupini B sa Zn u skupini CP. Prvo treba podatke dviju varijabli u svim trima skupinama unijeti u stupce (ukupno šest) jedan pokraj drugoga (sl. 6.1), označiti istu varijablu dviju skupina (npr. Cu u B i Cu u CP) → *Univariate* → *Two-sample tests* → *Two-sample tests* (F, t, Mann-Whitney, etc.) čime dobijemo rezultate dvaju navedenih testova (sl. 6.2).



Slika 6.1. Prikaz unosa podataka za Cu i Zn u trima skupinama uzoraka (B, CP i K) u svrhu testiranja razlika među dotičnim skupinama (bilo da je riječ o njih dvije bilo o tri i više).



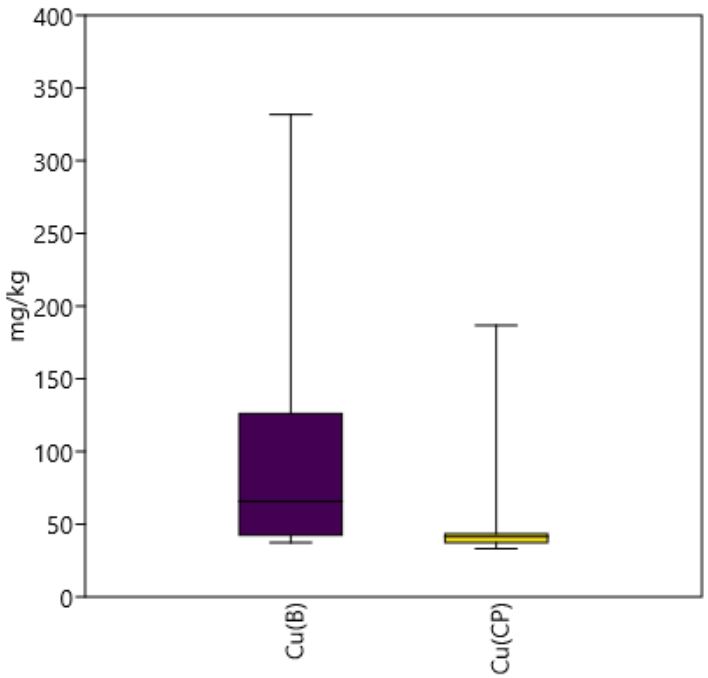
Slika 6.2. Rezultati t-testa za Cu u skupinama B i CP.

Rezultati t-testa (sl. 6.2) sastoje se od nekoliko brojeva gdje treba gledati samo vrijednost p (0,098) uz vrijednost t (1,71). Budući da je $p > 0,05$, prihvaća se H_0 i zaključuje da nema statistički značajne razlike između skupina B i CP glede Cu. Međutim, Cu ima asimetrično raspodijeljene podatke u tim skupinama (tablica 5.1) pa je u tom slučaju primjereniji Mann-Whitneyjev test (sl. 6.3).

Slika 6.3. Rezultati testa Mann-Whitney za Cu u skupinama B i CP.

Rezultati Mann-Whitneyjeva testa (sl. 6.3) također se sastoje od nekoliko brojeva gdje treba gledati samo vrijednost p (0,012) uz vrijednost z (2,51). Budući da je $p < 0,05$, odbacuje se H_0 i zaključuje da ima statistički značajne razlike između skupina B i CP glede Cu. Time je pokazano da t-test daje drukčiji rezultat u odnosu na svoju neparametarsku inačicu u slučaju

asimetrično raspodijeljene varijable. Usporedni kutijasti dijagrami (sl. 6.4) jasno pokazuju da nema znatnijega preklapanja vrijednosti Cu (misli se na podatke unutar „kutije“) u dvjema skupinama, što je sukladno rezultatu Mann-Whitneyjeva (sl. 6.3).



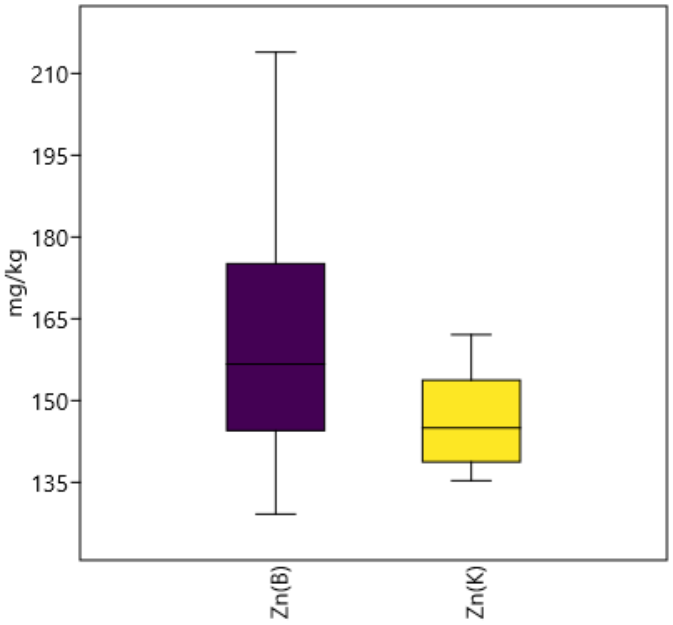
Slika 6.4. Usporedni kutijasti dijagrami za Cu u skupinama B i CP.

Rezultati obaju testova za obje varijable za tri para skupina prikazani su u tablici 6.1. Vidimo da se rezultati testova znatno razlikuju za Cu koji ima izrazito asimetričnu raspodjelu u skupinama B i CP, dok su rezultati testova za Zn gotovo identični jer mu je raspodjela u skupini B simetrična. To je još jedan dokaz naprijed navedene činjenice da se asimetrično raspodijeljeni podatci ne bi smjeli obrađivati parametarskim metodama. U slučaju Cu odbacuje se H_0 za sva tri para skupina (tj. uzimaju se u obzir rezultati Mann-Whitneyjeva testa) te zaključujemo da se tri skupine tala međusobno statistički značajno razlikuju po sadržaju Cu (tablica 6.1.). U slučaju Zn prihvaća se H_0 za sva tri para skupina i zaključuje da nema statistički značajne razlike u sadržaju Zn u trima skupinama tala (tablica 6.1.).

	Cu		Zn	
	T	MW	T	MW
B vs CP	0,098	**0,012**	0,92	0,97
CP vs K	0,30	**0,01**	0,16	0,07
B vs K	0,11	**0,002**	0,18	0,19

Tablica 6.1. Rezultati parametarskoga (t-test, T) i neparametarskoga (Mann-Whitney, MW) testa za Cu i Zn za tri para skupina (B, CP i K). Istaknute vrijednosti su statistički značajne ($p < 0,05$).

Usporedni kutijasti dijagrami (sl. 6.5) jasno pokazuju znatnije preklapanje vrijednosti Zn u dvjema skupinama (kutije), što je sukladno rezultatu obaju testova (tablica 6.1). To je bilo očekivano s obzirom na to da je Zn simetrično raspodijeljen u obje skupine (tablica 5.1), što se vidi i na sl. 6.5 (tzv. zalisci su podjednakih visina, a medijani približno u sredini kutije).



Slika 6.5. Usporedni kutijasti dijagrami za Zn u skupinama B i K.

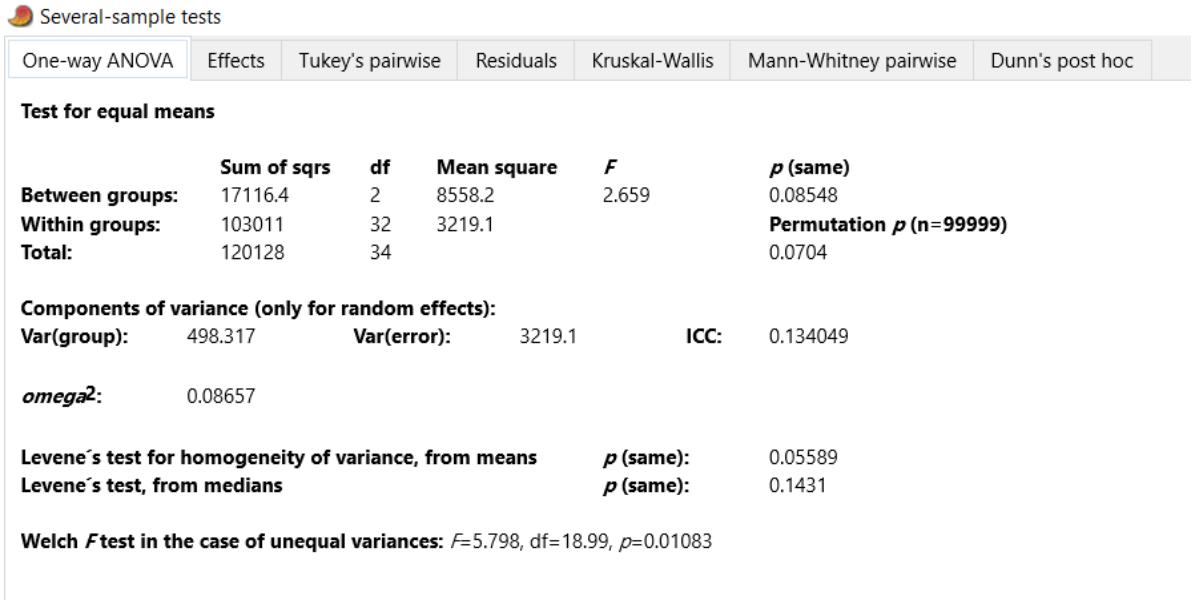
7. Testiranje razlika triju i više neovisnih skupina podataka

Ove metode primjenjive su na danim podatcima (Medunić i sur. 2009) jer postoje tri skupine uzoraka tla nad istom geološkom podlogom, ali različite namjene (urbano/industrijsko tlo i netaknuto šumsko tlo). Postavlja se pitanje razlikuju li se međusobno koncentracije mjerenih varijabli dotičnih triju skupina i ako se razlikuju, na koji način.

Testovima se utvrđuje imaju li sve skupine istu središnju vrijednost (medijan ili aritmetičku sredinu, ovisno o testu) ili se barem jedna od skupina razlikuje od ostalih. Ako su podatci unutar svake skupine normalno raspodijeljeni i imaju identične varijance, tada se koristi analiza varijance, tzv. ANOVA (engl. *analysis of variance*). Njome se utvrđuje jesu li aritmetičke sredine svih skupina identične. Ako spomenuti preduvjeti nisu ispunjeni, treba upotrijebiti neparametarske testove, kao što je Kruskal-Wallisov test kojim se uspoređuju medijani (npr. Pentecost 1999; Helsel i sur. 2020).

Nulta hipoteza za testove ANOVA i Kruskal-Wallis glasi da su aritmetičke sredine odnosno medijani identični, a alternativna je hipoteza da se barem jedno od njih razlikuje od ostalih. Cilj nije samo utvrditi razlikuju li se međusobno aritmetičke sredine ili medijani skupina, nego koje se skupine razlikuju. To je moguće obaviti testovima višestruke usporedbe, kojima se uspoređuju sve moguće kombinacije parova medijana ili aritmetičkih sredina tretiranih skupina, a izvode se tek nakon što nulta hipoteza bude odbačena.

PRIMJER 7.1. Usporedba triju skupina (B vs. CP vs. K) za Cu i Zn testovima ANOVA i Kruskal-Wallis u programu PAST. Podatke dviju varijabli potrebno je unijeti u stupce (ukupno šest) kao u primjeru 6.1. te označiti istu varijablu triju skupina (npr. Cu u B, CP i K) i → *Univariate* → ANOVA etc. (several tests) → *Several-sample tests* (ANOVA, Kruskal-Wallis) čime dobivamo rezultate dvaju navedenih testova (sl. 7.1).

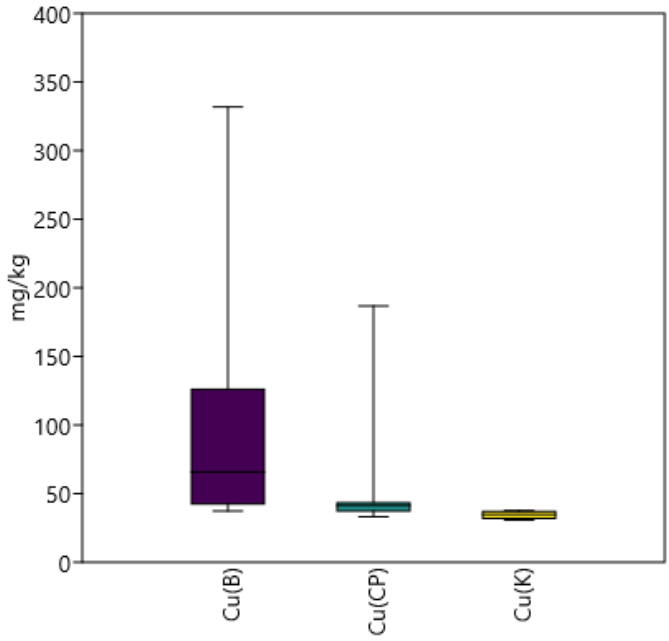


Slika 7.1. Rezultati testa ANOVA za Cu u trima skupinama (B, CP i K).

Rezultati testa ANOVA (sl. 7.1) sastoje se od nekoliko brojeva gdje treba gledati vrijednost p (0,085) uz vrijednost F (2,659). Budući da je p > 0,05, prihvatili bismo H_0 i zaključili da nema statistički značajne razlike među skupinama B, CP i K s obzirom na Cu. Budući da Cu ima asimetrično raspodijeljene podatke u skupinama B i CP (tablica 5.1), za tu varijablu treba primijeniti Kruskal-Wallisov test (sl. 7.2).

Slika 7.2. Rezultati Kruskal-Wallisova testa za Cu u skupinama B, CP i K.

Rezultat Kruskal-Wallisova testa (sl. 7.2), tj. p iznosi 0,0005 što je < 0,05 te odbacujemo H_0 i zaključujemo da ima statistički značajne razlike među skupinama B, CP i K s obzirom na Cu. Time smo se opet uvjerili (slično kao pri usporedbi dviju skupina s obzirom na Cu) da parametarska ANOVA daje drukčiji rezultat od onoga neparametarskog za asimetrično raspodijeljenu varijablu. Usporedni kutijasti dijagrami (sl. 7.3) jasno pokazuju da nema znatnijega preklapanja (površina „kutija“) vrijednosti Cu u trima skupinama, što je sukladno rezultatu Kruskal-Wallisova testa (sl. 7.2).



Slika 7.3. Usporedni kutijasti dijagrami za Cu u skupinama B, CP i K.

Sada treba vidjeti između kojih točno skupina postoji razlika s obzirom na Cu, a to je moguće s pomoću Mann-Whitneyjeva testa višestruke usporedbe, vrste *pairwise* 'u paru' (sl. 7.4).

Several-sample tests

One-way ANOVA | Effects | Tukey's pairwise | Residuals | Kruskal-Wallis | Mann-Whitney pairwise | Dunn's post hoc

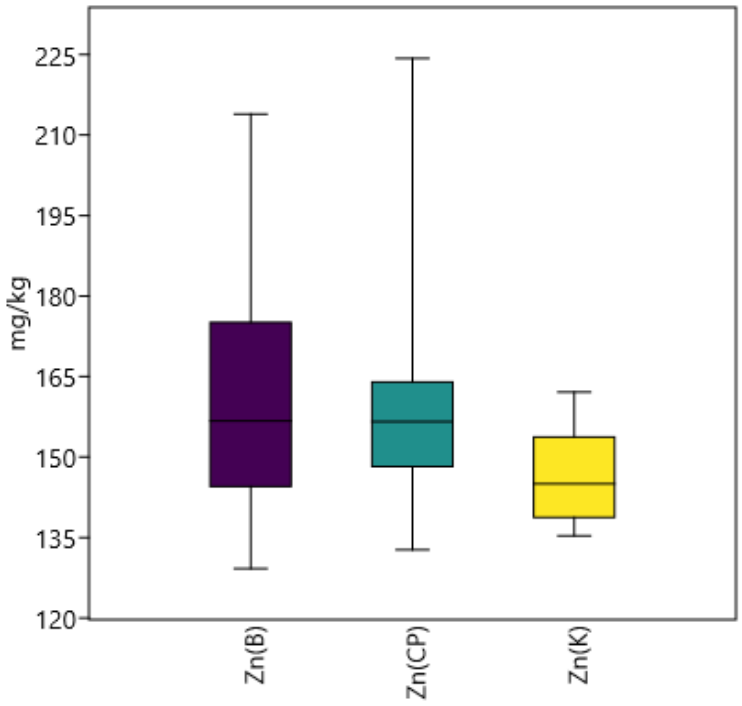
Raw p values, uncorrected significance

	Cu(B)	Cu(CP)	Cu(K)
Cu(B)		0.01207	0.001669
Cu(CP)	0.01207		0.009997
Cu(K)	0.001669	0.009997	

Slika 7.4. Rezultati testa Mann-Whitney *pairwise* (u parovima) za Cu u skupinama B, CP i K.

Slično grafičkom prikazu (sl. 7.3) test (sl. 7.4) pokazuje da su sve tri skupine uzoraka tala međusobno statistički značajno različite s obzirom na koncentracije Cu. Najveća je razlika između skupina B i K, na što upućuje najmanja vrijednost p (0,001).

Rezultati testova ANOVA i Kruskal-Wallis za Zn iznose 0,35 odnosno 0,22, što je u skladu s usporednim kutijastim dijagramima (sl. 7.5; znatna preklapanja vrijednosti podataka predstavljenih kutijama) te se prihvaća H_0 i zaključuje da nema statistički značajne razlike među trima skupinama s obzirom na Zn.



Slika 7.5. Usporedni kutijasti dijagrami za Zn u skupinama B, CP i K.

8. Korelacijska analiza

U brojnim situacijama treba izmjeriti jačinu povezanosti dviju kontinuiranih varijabli unutar iste skupine podataka. Kad se računaju koeficijenti korelacije, podatci se prikazuju dijagramom raspršenja (engl. *scatter*) jer različiti modeli mogu imati isti koeficijent korelacije i, suprotno tomu, slične jačine odnosa mogu dati različite koeficijente. Pritom se analizira povećava li se jedna varijabla s porastom druge, smanjuje li se s porastom druge ili je njihov model varijacije nepovezan. Time se mjeri opažena kovarijacija (međusobni odnos) dviju varijabli, no ne dokazuje se uzročni odnos među dvjema varijablama. Jedna može izazvati drugu, kao što oborine izazivaju otjecanje vode, no dokaz za uzročni odnos među varijablama ne može dati statistika već poznavanje uključenih geoloških procesa (npr. Helsel i sur. 2020; Fiket i sur. 2018).

Koeficijenti korelacije su bez dimenzije, a vrijednosti su im raspona od -1 do 1. Kad nema korelacije, koeficijent iznosi 0. Nulta hipoteza glasi da nema odnosa među varijablama (koeficijent korelacije teži nuli).

Podatci mogu biti korelirani na linearan ili nelinearan način. Kada se y općenito povećava ili smanjuje s porastom x, tada se za dvije varijable kaže da posjeduju monotonu korelaciju. Neparametarski Kendallov tau test mjeri jačinu monotonoga odnosa između x i y. On se temelji na postupku rangiranja zbog čega je otporan na učinak maloga broja neobičnih vrijednosti (ekstremi). Pogodan je i za asimetrično raspodijeljene varijable. Budući da je tau ovisan samo o rangovima podataka, a ne njihovim vrijednostima, može se koristiti i za cenzurirane podatke. Tau općenito ima niže vrijednosti koeficijenata u odnosu na parametarsku Pearsonovu vrijednost r, ali to ne znači da on ima manju osjetljivost, već je riječ o različitim ljestvicama korelacije. On je nepromjenjiv u odnosu na transformacije, tako da je identičan za bilo koji od omjera poput log (y)-log (x), y-log (x), ili y-x.

Parametarski Pearsonov r najčešće je korištena mjera linearne korelacije. Njime se mjeri linearna povezanost dviju varijabli. Tu je jako bitno proučiti dijagram raspršenja jer neznačajna vrijednost koeficijenta može biti posljedica kako nepovezanosti tako i „nagiba“ podataka te poglavito ekstrema. Pearsonov r nije otporan na ekstreme (kao tau) jer se računa s pomoću aritmetičke sredine i standardne devijacije, a pretpostavlja da podatci slijede bivarijatnu normalnu raspodjelu. To znači da varijable x i y moraju biti normalno raspodijeljene, ali i njihova zajednička varijabilnost mora imati specifičan, normalan oblik. Zbog toga nije

primjenjiv u slučaju kada asimetrično raspodijeljeni podatci imaju rastuću varijancu i ekstreme te ovaj koeficijent nije pogodan za netransformirane podatke.

PRIMJER 8.1. Korelacijska analiza nekoliko mjerenih varijabli u skupini B s pomoću obiju metoda u programu PAST. Podatke svih varijabli skupine B potrebno je unijeti u stupce te ih označiti → *Univariate* → *Correlation*, i time dobivamo Pearsonovu korelacijsku matricu (sl. 8.1), pri čemu na desnoj strani ekrana (dijalog s ponuđenim mogućnostima programa) možemo odabrati među ponuđenim koeficijentima onaj koji nas zanima.

Correlation

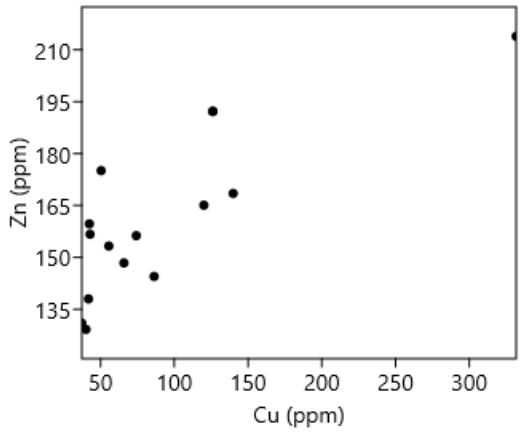
Table | Plot

	Al (%)	Fe (%)	Ni (ppm)	Cu (ppm)	Zn (ppm)	Pb (ppm)	As (ppm)	Rb (ppm)	OT (%)
Al (%)		0.00051665	0.61013	0.91035	0.12924	0.88589	0.070603	2.5842E-05	0.068157
Fe (%)	0.7857		0.14058	0.18502	0.58547	0.31487	0.27532	3.6716E-05	0.25179
Ni (ppm)	0.14341	0.39909		0.2326	0.73803	0.039358	0.56218	0.32532	0.33318
Cu (ppm)	-0.031826	-0.36189	-0.32804		0.0003498	0.023093	0.89244	0.43357	0.729
Zn (ppm)	0.40982	0.15329	-0.09435	0.79927		0.024353	0.17178	0.19918	0.05468
Pb (ppm)	-0.040559	-0.27848	-0.53619	0.58111	0.5769		0.10204	0.99404	0.067387
As (ppm)	0.47936	0.30119	-0.16277	-0.038215	0.37228	0.43851		0.018484	0.00012547
Rb (ppm)	0.8692	0.86152	0.27276	-0.2187	0.35129	-0.0021123	0.59823		0.04172
OT (%)	0.48305	0.31564	-0.26853	0.097718	0.50531	0.48423	0.83068	0.53091	

Slika 8.1. Pearsonova korelacijska matrica odnosa odabranih varijabli u skupini B.

Na sl. 8.1 prikazana je korelacijska matrica koja se sastoji od dviju skupina brojčanih vrijednosti: lijevo ispod dijagonale (prazna polja) nalaze se korelacijski koeficijenti, a desno iznad dijagonale nalaze se vrijednosti p na temelju kojih zaključujemo jesu li pojedini koeficijenti statistički značajni ($p < 0,05$) ili nisu ($p > 0,05$).

Primjerice, koeficijent odnosa Fe – Al iznosi 0,78, a njegova vrijednost p iznosi 0,0005, što je < 0,05 te se zaključuje da je pozitivna korelacija dviju varijabli statistički značajna. Naprotiv, odnos Al – Ni ($r = 0,14$) nije statistički značajan jer je vrijednost p (0,61) > 0,05. Pozitivni korelacijski koeficijent odnosa Cu – Zn iznosi 0,80 i on je statistički značajan jer je vrijednost p (0,0003) < 0,05 te odbacujemo H_0. Dijagram raspršenja odnosa Cu – Zn prikazan je na sl. 8.2.



Slika 8.2. Dijagram raspršenja pozitivne statistički značajne korelacije Cu – Zn.

Na sl. 8.3 prikazana je Kendallova tau korelacijska matrica koja funkcionira na isti, naprijed opisan način.

Correlation

Table | Plot

	Al (%)	Fe (%)	Ni (ppm)	Cu (ppm)	Zn (ppm)	Pb (ppm)	As (ppm)	Rb (ppm)	OT (%)
Al (%)		0.0021996	0.51454	0.74451	0.15789	0.66391	0.02125	0.00014348	0.063501
Fe (%)	0.58926		0.23887	0.384	0.23887	0.384	0.2146	0.00060133	0.27988
Ni (ppm)	0.12544	0.22667		0.1339	0.61734	0.016475	0.54491	0.36847	0.48214
Cu (ppm)	0.062718	-0.16754	-0.28846		0.0027195	0.012484	0.18962	1	0.1598
Zn (ppm)	0.27178	0.22667	-0.096154	0.57692		0.035866	0.086287	0.04566	0.012065
Pb (ppm)	-0.083624	-0.16754	-0.46154	0.48077	0.40385		0.069337	0.9204	0.087828
As (ppm)	0.44331	0.23883	-0.11651	0.25244	0.33011	0.34953		0.0087051	0.00082029
Rb (ppm)	0.73171	0.6603	0.17308	0	0.38462	-0.019231	0.50488		0.034978
OT (%)	0.35712	0.20796	-0.13527	0.27053	0.4831	0.32851	0.64391	0.4058	

Slika 8.3. Kendallova tau korelacijska matrica odnosa odabranih varijabli u skupini B.

U usporedbi s Pearsonovom korelacijskom matricom (sl. 8.1) Kendallovi tau korelacijski koeficijenti (sl. 8.3) imaju niže vrijednosti, što je naprijed u teorijskom dijelu objašnjeno. Inače, statistička značajnost korelacijskih koeficijenata izrazito ovisi o broju mjerenja. Ako je n malen broj (npr. 5–10), čak i visok koeficijent korelacije (bilo 0,99 ili -0,99) neće biti proglašen statistički značajnim. Vrijedi i obrnuto: ako je n velik broj (npr. 90–100), čak i niži korelacijski koeficijent (npr. 0,37 ili -0,37) može biti proglašen (na temelju vrijednosti p) statistički značajnim. Geološko tumačenje korelacijskih koeficijenata mjerenih varijabli u skupinama B, CP i K detaljno je predstavljeno u radu Medunić i sur. 2009.

9. Zaključak

Ovaj priručnik prikazuje preporučene korake statističke analize podataka dobivenih analizom geoloških uzoraka iz okoliša primjenom parametarskih i neparametarskih metoda. Uobičajeni je redoslijed koraka analize podataka sljedeći:

1) testiranje vrste raspodjele podataka (npr. s pomoću Shapiro-Wilkova testa)

2) određivanje osnovnih statističkih parametara

3) testiranje razlika među skupinama podataka (npr. za dvije se skupine koriste parametarski i neparametarski t-test odnosno Mann-Whitneyev test, a za tri i više skupina koriste se ANOVA odnosno Kruskal-Wallisov test)

4) korelacijska analiza varijabli unutar iste skupine (npr. parametarski i neparametarski Pearsonov odnosno Kendallov tau korelacijski koeficijent).

Nulta hipoteza koja se u analizi geoloških podataka najčešće postavlja glasi:

a) raspodjela podataka je normalna (simetrična)

b) nema razlike među skupinama podataka

c) nema odnosa među varijablama (unutar iste skupine podataka).

Nadalje, vrijednost p je postignuti nivo značajnosti; ako je p veći od zadanoga nivoa značajnosti, koji je obično 0,05 (0,10 u slučaju ispitivanja vrste raspodjele podataka), nulta se hipoteza prihvaća, a u protivnom ($p < 0,05$ ili $p < 0,10$) odbacuje. Ako na temelju Shapiro-Wilkova testa zaključimo da su podatci asimetrično raspodijeljeni, kao što je bio slučaj s elementima Cu i Zn u uzorcima tla oko dviju tvornica, potrebno ih je obraditi neparametarskim metodama. Od prikazanih grafičkih alata najveću važnost imaju dijagrami normalne vjerojatnosti, kutijasti dijagrami i dijagrami raspršenja. Dijagrami normalne vjerojatnosti prikazuju vrste raspodjele podataka analiziranih varijabli, a kutijasti dijagrami od velike su pomoći pri usporedbi skupina podataka, npr. Cu u tlu oko tvornice B u odnosu na Cu u šumskom tlu (geokemijski neporemećenom ljudskim aktivnostima) .

Zaključno, matematičke metode obrade geoloških podataka nužne su u geološkoj struci u 21. stoljeću, obilježenom trendovima posvemašnje digitalizacije svega oko nas. Ipak, studenti geologije ne smiju izgubiti iz vida činjenicu da se smislena (geo)statistička analiza podataka temelji na ispravno osmišljenoj shemi uzorkovanja, dobru poznavanju temeljnih geoloških principa i pozornu terenskom radu, što ima neprocjenjivu važnost u geološkim znanstvenim disciplinama.

10. Literatura

Barudžija, U.; Velić, J.; Malvić, T.; Trenc, N.; Matovinović Božinović, N. 2020. Morphometric Characteristics, Shapes and Provenance of Holocene Pebbles from the Sava River Gravels (Zagreb, Croatia). *Geosciences* 10/3. 92. doi:10.3390/geosciences10030092

Bhattacharyya, G. K.; Johnson, R. A. 1991. *Statistical Concepts and Methods*. John Wiley & Sons. New York. 639 str.

Davis, J. C. 1986. *Statistics and data analysis in geology*. John Wiley & Sons. New York. 646 str.

Fiket, Ž.; Medunić, G.; Furdek Turk, M.; Ivanić, M.; Kniewald, G. 2017. Influence of soil characteristics on rare earth fingerprints in mosses and mushrooms : example of a pristine temperate rainforest (Slavonia, Croatia). *Chemosphere* 179. 92–100. doi:10.1016/j.chemosphere.2017.03.089

Fiket, Ž.; Ivanić, M.; Furdek Turk, M.; Mikac, N.; Kniewald, G. 2018: Distribution of trace elements in waters of the Zrmanja River estuary (eastern Adriatic coast, Croatia). *Croatica chemica acta* 91/1. 29–41. doi:10.5562/cca3202

Fiket, Ž.; Fiket, T.; Ivanić, M.; Mikac, N.; Kniewald, G. 2019. Pore water geochemistry and diagenesis of estuary sediments – an example of the Zrmanja River estuary (Adriatic coast, Croatia). *Journal of soils and sediments* 19/4. 2048–2060. doi:10.1007/s11368-018-2179-9

Fiket, Ž.; Medunić, G.; Vidaković-Cifrek, Ž.; Jezidžić, P.; Cvjetko, P. 2020. Effect of coal mining activities and related industry on composition, cytotoxicity and genotoxicity of surrounding soils. *Environmental science and pollution research* 27/6. 6613–6627. doi:10.1007/s11356-019-07396-w

Hammer, Ø.; Harper, D. A. T.; Ryan, P. D. 2001. Past: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* 4/1. 9 str. 178kb. http://palaeo-electronica.org/2001_1/past/issue1_01.htm

Helsel, D. R.; Hirsch, R. M.; Ryberg, K. R.; Archfield, S. A.; Gilroy, E. J. 2020. Statistical Methods in Water Resources. USGS Publications Warehouse. 484 str. http://pubs.er.usgs.gov/publication/tm4A3

Ivšinović, J.; Malvić, T. 2020. Application of the radial basis function interpolation method in selected reservoirs of the Croatian part of the Pannonian Basin System. *Mining of mineral deposits* 14/3. 37–42. doi:10.33271/mining14.03.037

Kang, S.; Ivošević, T.; Medunić, G.; Dai, S. 2022. Coal-derived sulphur and selenium in marine sediment cores (Raša Bay, Croatia): recommended steps of analysing environmental earth data. *Mathematical methods and terminology in geology* 2022. Ur. Malvić, T.; Ivšinović, J. Rudarsko-geološko-naftni fakultet. Zagreb. 75–83.

Malvić, T. 2008. Primjena geostatistike u analizi geoloških podataka. INA-Industrija nafte d.d. Zagreb. Sveučilišni priručnik. 103 str.

Malvić, T.; Cvetković, M.; Balić, D. 2008. *Geomatematički rječnik*. Hrvatsko geološko društvo. Zagreb. 74 str.

Malvić, T.; Cvetković, M. 2013. Neuronski alati u geologiji ležišta ugljikovodika. Hrvatsko geološko društvo. Zagreb. 89 str.

Malvić, T.; Medunić, G. 2015. *Statistika u geologiji*. Rudarsko-geološko-naftni fakultet – Prirodoslovno-matematički fakultet. Zagreb. Sveučilišni udžbenik. 88 str.

Malvić, T.; Bošnjak, M.; Velić, J.; Sremac, J.; Ivšinović, J.; Pimenta Dinis, M. A.; Barudžija, U. 2020a. Recent Advances in Geomathematics in Croatia: Examples from Subsurface Geological Mapping and Biostatistics. *Geosciences* 10/5. 188. doi:10.3390/geosciences10050188

Malvić, T.; Ivšinović, J.; Velić, J.; Sremac, J.; Barudžija, U. 2020b. Application of the Modified Shepard's Method (MSM): A Case Study with the Interpolation of Neogene Reservoir Variables in Northern Croatia. *Stats* 3/1. 68–83, doi:10.3390/stats3010007

Medunić, G.; Tomašić, N.; Balen, D.; Oreščanin, V.; Prohić, E.; Kampić, Š.; Ivanišević, D. 2009. Distribution of copper and zinc in the soil of an industrial zone in the city of Garešnica, Croatia. *Geologia Croatica* 62/3. 179–187.

Medunić, G.; Ahel, M.; Božičević Mihalić, I.; Gaurina Srček, V.; Kopjar, N.; Fiket, Ž.; Bituh, T.; Mikac, I. 2016. Toxic airborne S, PAH, and trace element legacy of the superhigh-organic-sulphur Raša coal combustion: Cyto-toxicity and genotoxicity assessment of soil and ash. *Science of the Total Environment* 566. 306–319.

Medunić, G. 2022. Skripta za predmet *Osnove elementne i fazne analize* (neobjavljeno djelo, preddiplomski studij geologije, PMF).

Medunić, G.; Chakravarty, S.; Kundu, R. 2022. Computational skills in geosciences higher education system for the 21st century. *Mathematical methods and terminology in geology 2022*. Ur. Malvić, T.; Ivšinović, J. Sveučilište u Zagrebu, Rudarsko-geološko-naftni fakultet. Zagreb. 8 str.

Mesić Kiš, I.; Malvić, T. 2014. Zonal estimation and interpolation as simultaneous approaches in the case of small input data set (Šandrovac Field, Northern Croatia). *Rudarsko-geološko-naftni zbornik* 29/1. 9–16.

Mesić, I.; Medunić, G. 2014. Declustering of field's data located on northern margin of the Bjelovar subdepression. *Geomathematics – from theory to practice*. Ur. Cvetković, M.; Novak Zelenika, K.; Geiger, J. Hrvatsko-geološko društvo. Zagreb. 21–28.

Miller, J. N.; Miller, J. C. 2010. *Statistics and chemometrics for analytical chemistry*. Pearson Education Limited. Edinburgh. 278 str.

Motulsky, H. 1999. *Analyzing data with GraphPad Prism. A companion to GraphPad Prism version 3*. https://cdn.graphpad.com/faq/1283/file/AnalyzingDataPrism3.pdf

Motulsky, H. 2014. Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Archives of Pharmacology* 387. 1017–1023.

Pavlović, G.; Prohić, E.; Tibljaš, D. 2004. Statistical assessment of geochemical pattern in overbank sediments of the river Sava, Croatia. *Environmental Geology* 46/1. 132–143.

Pentecost, A. 1999. *Analysing environmental data*. Pearson Education Limited. Harlow. 214 str.

Reimann, C.; Filzmoser, P.; Garrett, R. G.; Dutter, R. 2008. *Statistical Data Analysis Explained*. Applied Environmental Statistics with R. John Wiley & Sons, Ltd. 343 str.

Swan, A. R. H.; Sandilands, M. 1995. *Introduction to geological data analysis*. Blackwell Science Ltd. Oxford. 446 str.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46 (Supplement). 234–240.

Zhang, Y. 2011. *Introduction to Geostatistics — Course Notes*. http://www.uwyo.edu/geolgeophys/people/faculty/yzhang/_files/docs/geosta1.pdf