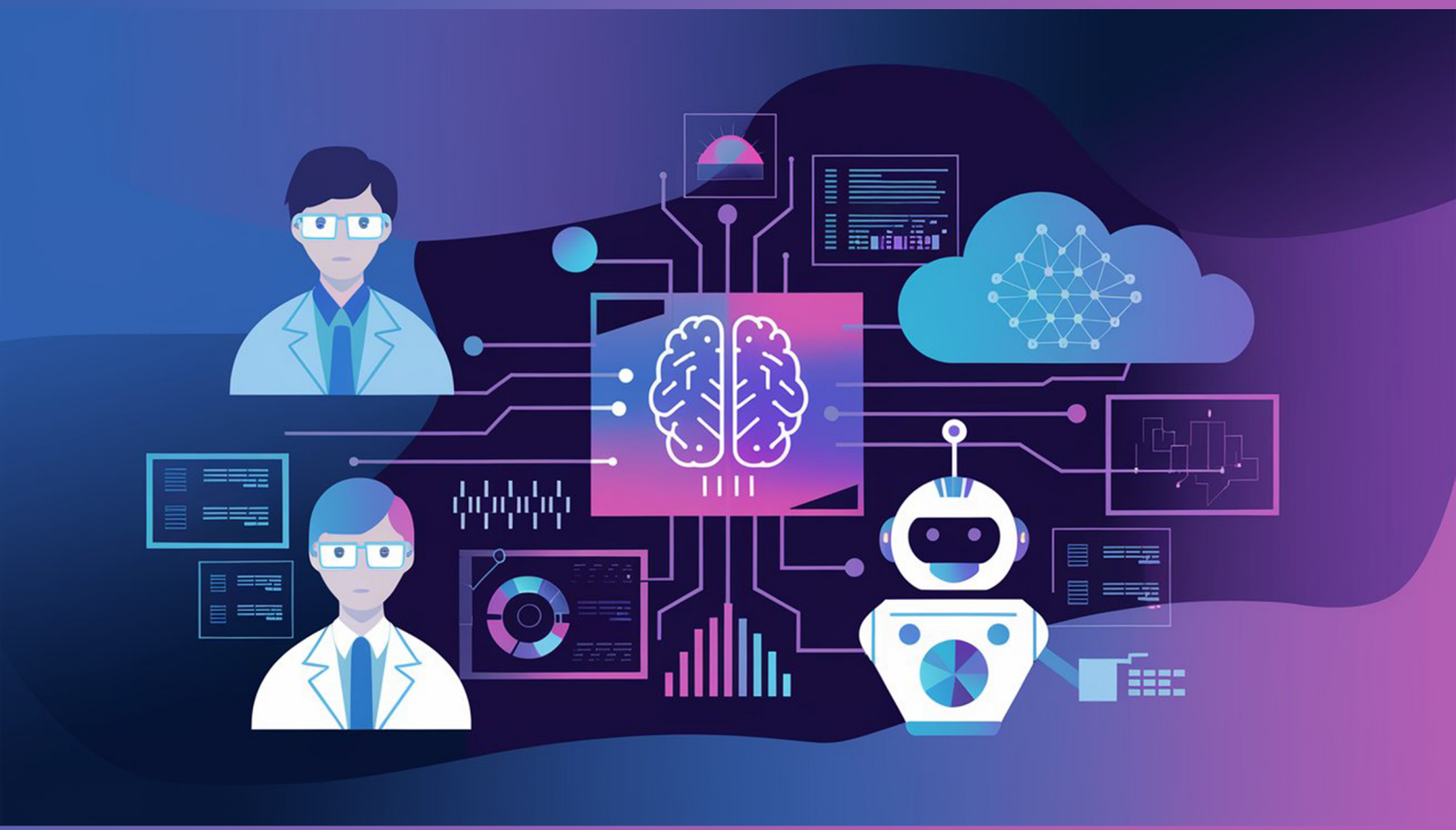
Oļegs Užga-Rebrovs
Pēteris Grabusts

DATU PIRMSAPSTRĀDES METODES DATU ANALĪZES UZDEVUMOS



Oļegs UŽGA-REBROVS, Pēteris GRABUSTS. *Datu pirmsapstrādes metodes datu analīzes uzdevumos*. Rēzekne: Rēzeknes Tehnoloģiju akadēmija. 2024. 182 lpp.

Recenzenti:
- Dr. **Sergejs PARŠUTINS** (RTU)
- Dr. **Ilmārs KANGRO** (RTA)

Zinātniskā monogrāfija sagatavota un izdota ar Rēzeknes Tehnoloģiju akadēmijas finansiālo atbalstu.

Monogrāfijā tiek apskatītas izplatītākās datu pirmsapstrādes metodes veiksmīgai datu analīzei un doti datu analīzes pamatjēdzieni. Detalizēti aplūkotas darbības ar iztrūkstošiem datiem, pievēršot uzmanību dažāda veida datu anomāliju identificēšanai, datu atribūtu vērtību diskretizācijai un normalizācijai. Sagatavošanas procesā bieži ir nepieciešams pārveidot datus citos formātos ērtākai to turpmākajai analīzei.
Materiāls ir papildināts ar pielikumiem, kuros iztirzāti jēdzieni, kas nepieciešami satura pilnvērtīgākai izpratnei. Datu analīzes metožu formālie apraksti papildināti ar detalizētiem piemēriem, kas izskaidro katras metodes būtību un darbības principus.
Dotais materiāls paredzēts dažāda līmeņa datorzinātņu un informācijas tehnoloģiju studiju programmu studējošajiem un pētniekiem, kam nepieciešams veikt datu analīzi un sagatavošanu to turpmākajai izmantošanai dažādos statistikas vai mākslīgā intelekta metožu pielietojumos.

Redaktore: **Vita Ansone**

ISBN 978-9984-44-287-7

SATURA RĀDĪTĀJS

1. IEVADS

1. Dati, informācija un zināšanas

Šajā grāmatā ir aplūkotas datu pirmapstrādes metodes turpmākās analīzes nolūkos. Tāpēc ir nepieciešams ieviest formālu datu jēdziena definīciju, kā arī ar datiem cieši saistītos informācijas un zināšanu jēdzienus. Jāatzīmē, ka datu jēdziens ir saprotams vienkāršā, intuitīvā līmenī. Tomēr, runājot par informācijas un zināšanu jēdzieniem, viss nav tik vienkārši. Interesentiem, kuri vēlas iegūt sīkāku informāciju par šiem svarīgajiem jēdzieniem un to savstarpējām attiecībām, ieteicams atsaukties uz darbiem [Burgin M., 2004; Zins Ch ., 2007; Sanders J. D., 2016].

Autors definē *datus* kā formālu objektu, faktu un notikumu attēlojumu (aprakstu) reālajā pasaulē. Šie objekti, fakti un notikumi var pastāvēt neatkarīgi vai atkarīgi no cilvēkiem. Gaisa temperatūru, vēja stiprumu un virzienu nosaka tikai dabas faktori, un tie nav atkarīgi no cilvēkiem. Savukārt preču klāsts veikalā un to izvietojums plauktos ir atkarīgs tikai un vienīgi no cilvēku rīcības.

Šīs grāmatas kontekstā ar datiem saprot novērojumu, mērījumu, eksperimentu, iedzīvotāju aptauju rezultātus, dažāda veida tekstus, audio un video ierakstus un citas ārpasaules realitātes.

Datu prezentācijas formas var būt ļoti dažādas. Vienkāršākā un ļoti izplatītā forma ir datu tabula (matrica), kur katrā rindā tiek parādīts objekts, subjekts vai cita entītija un atribūtu vērtību (iezīmju) vektors, kas raksturo šo entītiju.

Jebkura pētījuma mērķis nav paši dati, bet gan datos ietvertā informācija.

Iespējams, zinātnē un praktiskajās jomās nav neviena cita jēdziena, kam būtu tik daudz dažādu definīciju kā informācijas jēdzienam. Diemžēl šim jēdzienam nav vienotas vispārpieņemtas definīcijas. Kibernētikas tēvs Norberts Vīners informāciju definē kā uztveres satura apzīmējumu, ko saņemam no ārējās vides. Statistikas informācijas teorijā informācija ir saistīta ar nenoteiktību, kas ietverta noteiktā ziņojumā.

Informācijas jēdziena definīciju dažādība ir saistīta ar kontekstu dažādību, kādos šis jēdziens tiek lietots. Datu analīzes literatūrā piedāvātas daudzas informācijas jēdziena definīcijas. Darbā [Zins Ch., 2006] ir dotas vairākas šādas definīcijas.

Autors par pamatu ņem šādu vispārīgu definīciju: *informācija liecina par struktūru, attiecībām, atkarībām un citām specifiskām iezīmēm, kas pastāv datos slēptā formā.*

Bieži vien datu vākšana ir īpaši izstrādāta, lai iegūtu atbilstošu informāciju. Informācijas iegūšana, pamatojoties uz socioloģiskām aptaujām, ir ļoti izplatīta socioloģiskajos un politiskajos pētījumos. Datu vākšanu var organizēt, lai apstiprinātu vai noraidītu hipotēzi, precizētu un padziļinātu iepriekš pabeigtu pētījumu rezultātus, un daudziem citiem mērķiem.

Pēc savas būtības no datiem iegūtā informācija var būt precīza, aptuvena, neskaidra vai nepatiesa. Var runāt par precīzu informāciju, ja šāda informācija ir *apriori* ietverta datos un tā ir pareizi iegūta. Piemēram, dati satur precīzus divu mainīgo lielumu mērījumus, no kuriem viens ir funkcionāli atkarīgs no otra. Pamatojoties uz šādu datu apstrādi, var iegūt precīzu attiecību grafisku vai analītisko izteiksmi.

Lielākajā daļā gadījumu precīza informācija, kas iegūta no datiem, ir izņēmums, nevis likums. Parasti visi novērojumi un mērījumi tiek veikti ar lielākām vai mazākām kļūdām. Ja iepriekš dotajā piemērā mainīgo vērtību mērījumi tika veikti ar dažām kļūdām, tad iegūto datu apstrādes rezultātā var iegūt tikai aptuvenu vēlamās attiecības attēlojumu. Esošo informācijas precizitāti var modelēt, piemēram, izmantojot ticamības intervālus, kuros atrodas atkarīgā mainīgā faktiskās vērtības.

No datiem iegūtās informācijas nenoteiktībai ir divi galvenie iemesli: (1) datu iekšējā nenoteiktība; (2) neskaidru, subjektīvu datu izmantošana. Pieņemsim, ka dati atspoguļo gadījuma lieluma realizācijas. Apstrādājot šos datus, var izveidot šī nejaušā lieluma sadalījuma funkciju un/vai blīvuma funkciju un novērtēt iegūtās funkcijas parametrus. Šāda informācija atspoguļo sākotnējo datu iekšējo nenoteiktību.

Pieņem, ka dati atspoguļo iedzīvotāju aptaujas rezultātus par kādu konkrētu tēmu. Var sagaidīt, ka dažas atbildes uz uzdotajiem jautājumiem var būt neobjektīvas, nejaušas vai apzināti nepatiesas.

Skaidrs, ka šādu datu apstrādes rezultātā tiks iegūta vairāk vai mazāk neskaidra informācija. Tas ir jāpatur prātā, izmantojot šādu informāciju.

Bieži, ja nav iespējams iegūt objektīvus datus, tiek izmantoti subjektīvi ekspertu vērtējumi. Eksperti kalpo kā "mērinstrumenti". Protams, šādu subjektīvu datu apstrādes rezultātā iegūtā informācija būs vairāk vai mazāk neskaidra. Ja ekspertiem ir grūti veikt nepārprotamu subjektīvu vērtējumu, viņi var izteikt savus spriedumus nenoteiktā formā: punktu aplēšu vietā dot intervālu vai izplūdušus skaitļus. Bet šāda datu nenoteiktība radīs būtisku nenoteiktību informācijā, kas iegūta, apstrādājot šos datus.

Mūsdienās ir izstrādātas spēcīgas un efektīvas metodes neprecīzu un nenoteiktu datu apstrādei, taču šīs metodes netiek apskatītas šīs grāmatas kontekstā.

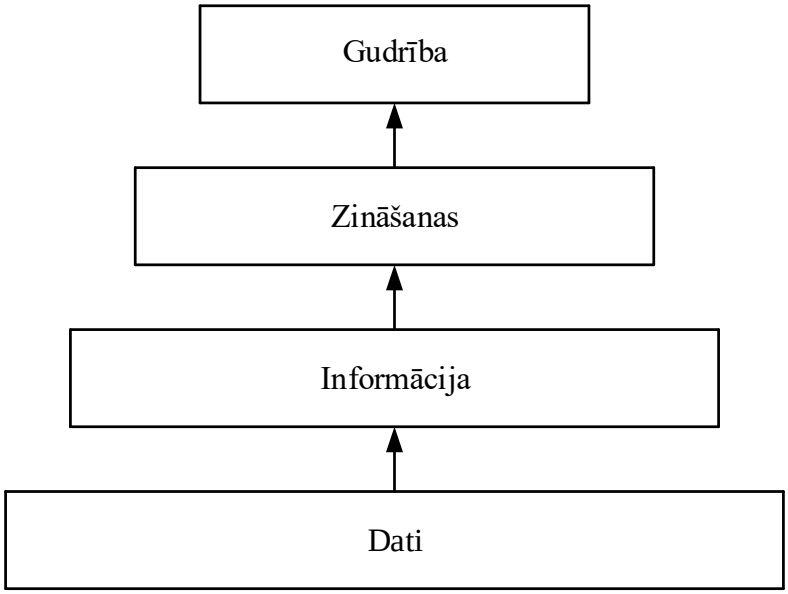
Var runāt par nepatiesu informāciju, ja šī informācija iegūta no dažāda veida viltus avotiem (datiem). Nepatiesas informācijas izmantošana var radīt ļoti nepatīkamas sekas. Liekot likmes uz konkrētu spēlētāju, lai uzvarētu derībās, pamatojoties uz nepatiesu informāciju, derību slēdzējs vienkārši zaudēs naudu. Tomēr neveiksmīgi rezultāti, kas iegūti nepatiesas informācijas izmantošanas dēļ, var būt daudz nopietnāki. Ja militārā operācija tiek plānota, pamatojoties uz nepatiesu informāciju par iespējamajiem ienaidnieka spēkiem un darbībām, tas var radīt ievērojamus zaudējumus.

Acīmredzot attiecībā uz nepatiesu informāciju vienīgā rīcība, kas varētu būt piemērota, ir atpazīt šīs informācijas nepatiesību un nekādā gadījumā to neizmantot.

Zināšanu jēdziens ir ļoti plaša filozofiska kategorija. Datu analīzes kontekstā autors definē zināšanas kā datu apstrādes un analīzes rezultātā iegūtās informācijas izpratnes un sintēzes rezultātu.

Zināšanas var klasificēt pēc informācijas, uz kuras pamata tās iegūtas: precīzas, aptuvenas, neskaidras un nepatiesas zināšanas.

Ir skaidrs, ka datu analīzes kontekstā datu informācijas un zināšanu jēdzieni ir savstarpēji saistīti jēdzieni. Darbā [Rowley J., 2007] šie jēdzieni tiek interpretēti, izmantojot hierarhisko struktūru (sk. 1.1. attēlu).



1.1. attēls. **Hierarhiskā struktūra, kas savieno datu, informācijas, zināšanu un gudrības jēdzienus**

Šo struktūru literatūrā bieži sauc par DIKW diagrammu (angļu valodā – *Data - Information - Knowledge - Wisdom*). Struktūra skaidri atspoguļo hierarhiskās attiecības starp jēdzieniem – dati ir informācijas avots, zināšanas balstās uz informāciju. Gudrības jēdziens šajā hierarhijā raksturo indivīda kognitīvās spējas, jo viņa rīcībā ir noteiktas zināšanas. Autors neiedziļināsies šajā koncepcijā, kurai ir izteikta filozofiska nozīme.

1.2. Datu veidi

Vārds "dati" ir cēlies no latīņu valodas, kas nozīmē "kaut kas, kas mums tiek dots". Datus var klasificēt dažādos veidos atkarībā no klasifikācijas pamatā esošajiem raksturlielumiem. Vārda plašākajā nozīmē datus var klasificēt atkarībā no to iegūšanas un izmantošanas jomas. Šajā ziņā var runāt par bioloģiskajiem, medicīniskajiem, socioloģiskajiem un citiem datiem.

Zinātniskajos pētījumos datus bieži klasificē pēc tā, kādā veidā tie iegūti:

- *novērojumu dati*. Tie var būt vizuālo novērojumu rezultāti, dažāda veida mērījumu dati, datu ieraksti, sensoru dati un tamlīdzīgi;
- *eksperimentālie dati*. Tie ir dati, kas iegūti iepriekš plānotu eksperimentu rezultātā;
- *simulācijas dati*. Tādi dati tiek iegūti pētnieku interesējošā procesa datorsimulācijas rezultātā. Simulāciju bieži izmanto gadījumos, kad ir neiespējami vai ļoti dārgi tiešā veidā izpētīt reālu procesu;
- *kompilētie dati*. Šāda veida dati ir datu, kas iegūti no dažādiem avotiem, apvienošanas rezultāts. Šajā gadījumā bieži tiek izmantota atsevišķu datu pārveidošana. Rezultātā tiks iegūti viendabīgi dati.

Datu klasifikācija ir balstīta uz to atribūtu vērtību veidiem. Atribūtu vērtības var izmērīt dažādās skalās, kas ir datu klasifikācijas avots. Lai vienkāršotu turpmāko izklāstu, tiek pieņemts, ka ir darīšana ar viendimensiju datiem, t. i., datiem, ko raksturo viena atribūta vērtības.

Sīkāk tiks apskatīta atribūtu vērtību mērīšanas skalu klasifikācija un uz tā pamata izveidoto datu klasifikācija.

Zemāko līmeni skalu hierarhijā ieņem *nominālā skala*. Šī skala mēra atribūtus, kuru kategorijas var tikai iezīmēt vai nosaukt. Nosaukšanas skalu izmantošanas piemēri ir priekšmetu nosaukumi, cilvēku vārdi, kategoriskas pazīmes, piemēram, cilvēku dzimums, politisko partiju nosaukumi, konkursa uzvarētāju saņemto medaļu kategorijas. Dažkārt tiek lietoti nominālās skalas kategoriju ciparu vai simboliskie kodējumi, piem., 1 – "vīrietis", 2 – "sieviete". Bet tie ir tikai citi dzimuma apzīmējumi, nekas vairāk.

Ja atribūtu vērtības mēra nominālā skalā, ar šīm vērtībām nevar veikt nekādas darbības – vērtības nevar salīdzināt, pievienot, atņemt utt. Vienīgais, ko var izdarīt ar atribūtu vērtībām, kas izteiktas nominālā skalā, ir noteikt šo vērtību režīmu.

Dažreiz, ņemot vērā nominālās skalas īpašības, to sauc arī par *nominālvērtības skalu*. Šis nosaukums labāk atspoguļo skalas būtību, taču literatūrā parasti tiek lietots jēdziens *nosaukuma skala*.

Ordinālā skala ir nominālās skalas paplašinājums tādā nozīmē, ka var veikt relatīvus objektu salīdzinājumus. Tipisks mērījumu piemērs ordinālajā skalā ir studentu rezultāti. Ja viens students eksāmenā ir ieguvis 7 punktus, bet otrs students – 9, var teikt, ka otram studentam eksāmenā veicās labāk. Bet mēģināt novērtēt, cik labāk vai cik reižu labāk, ir bezjēdzīgi.

Citi ordinālās skalas mērījumu piemēri ir subjektīvi temperatūras mērījumi: auksti, silti, karsti; subjektīvi sāpju novērtējumi: vieglas, vidēji smagas, smagas un Likerta skalas mērījumi. IQ mērījumus klasificē arī pēc ordinālās skalas, lai gan diezgan bieži šie mērījumi tiek kļūdaini klasificēti kā intervālu skala.

Bieži vien literatūrā nominālā un ordinālā skala tiek apvienotas zem *kategoriskās skalas* vispārīgā nosaukuma.

Ja atribūtu vērtību mērījumus veic nominālā vai ordinālā skalā, tad datus sauc par *kvalitatīvajiem datiem*.

Augstāku līmeni ieņem *intervālu skala*. Tipisks šīs skalas mērījumu piemērs ir temperatūras mērīšana. Skalas raksturīga īpašība ir tāda, ka var salīdzināt mērījumu rezultātus un secināt, cik viens rezultāts ir lielāks vai mazāks par citu. Piemēram, ja ir divi temperatūras mērījumi: $t_1 = 15^0$ C, $t_2 = 20^0$ C, tad var pareizi norādīt, ka otrā vērtība ir par 5 vienībām lielāka nekā pirmā vai pirmā vērtība ir par 5 vienībām mazāka nekā otrā. Bet apgalvojumam, ka 20^0 C temperatūra ir divreiz lielāka par 10^0 C, šajā skalā nav nozīmes. Šī skalas iezīme ir saistīta ar to, ka intervālu skalai nav nosacīta sākuma punkta. Tādējādi temperatūra 0^0 C tiek uzskatīta par ūdens sasalšanas temperatūru. Ar tādiem pašiem

panākumiem jebkuru citu temperatūru var uzskatīt par skalas nulles punktu (Fārenheita skalā ūdens sasalšanas punkts ir 32^0 F).

Intervāla skala parāda izmērīto vērtību intervālu attiecību. Tādējādi var apgalvot, ka temperatūras intervāls (20^0 C – 40^0 C) ir divreiz lielāks par intervālu (10^0 C–20^0 C).

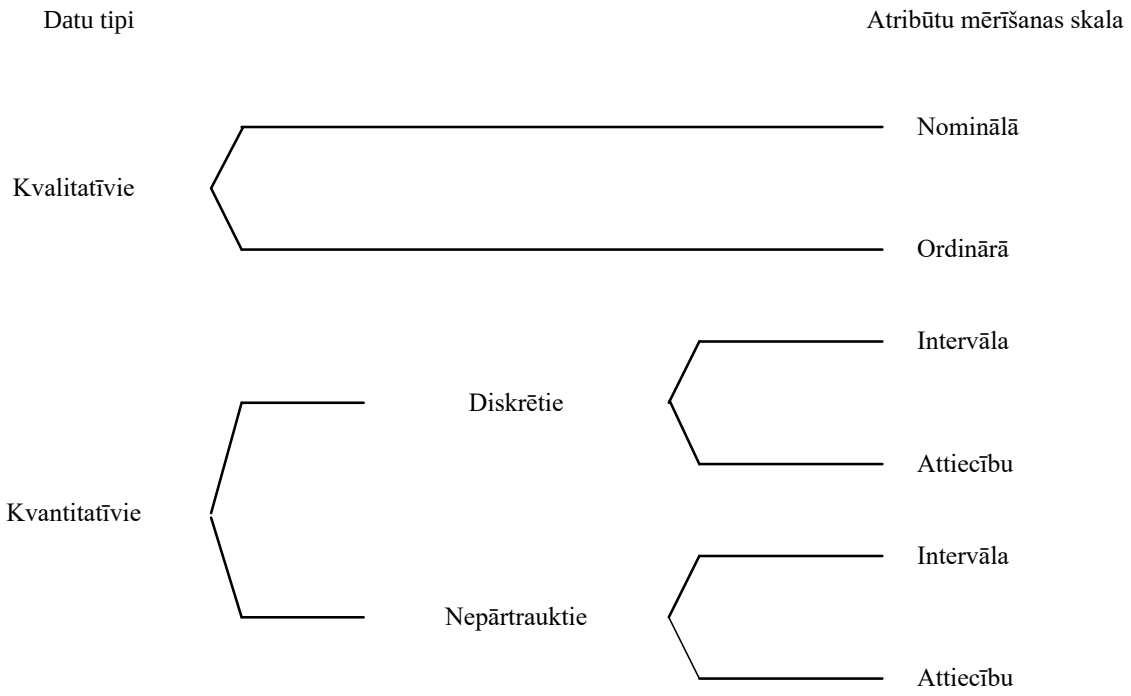
Mērījumu skalu hierarhijas augstākais līmenis ir *attiecību skala*. Tipiski mērījumu piemēri attiecību skalā ir auguma, svara vai attāluma mērījumi. Ja atribūtu vērtības mēra attiecību skalā, tad var norādīt, par cik viena vērtība ir pārāka par citu. Šādām atribūtu vērtībām var aprēķināt režīmu, mediānu un vidējo.

Ja atribūtu vērtības mēra intervālu skalā vai attiecību skalā, tad šādus datus sauc par *kvantitatīviem datiem*.

Vēl viena plaši izmantota vienfaktoru datu klasifikācija ir balstīta uz atribūtu vērtību būtību:

- *nepārtraukti dati*. Atribūts var iegūt jebkuru vērtību noteiktā intervālā. Nepārtrauktu datu piemēri ir personas augums un svars, viņu mājsaimniecības mantas vērtība vai algas;
- *diskrēti dati*. Atribūtam noteiktā intervālā var būt tikai noteiktas vērtības. Piemērs varētu būt studentu skaits lekcijā.

Vispārinot iepriekš ieviestos jēdzienus un definīcijas, 1.2. attēlā ir sniegta vispārīga vienfaktoru datu klasifikācija.



1.2. attēls. **Datu tipu grafiskais attēlojums**

Iepriekš apskatītās datu tipu klasifikācijas tika uzrādītas vienfaktoru datu gadījumiem. Ja visu daudzfaktoru datu atribūtu vērtības tiek mērītas vienā skalā, visi iepriekš minētie jēdzieni un definīcijas ir pilnībā piemērojami šādiem datiem.

Parasti reālos daudzfaktoru datos atribūtu vērtības tiek mērītas dažādās skalās. Šādos gadījumos ir runa par *jauktajiem datiem*. Šāda situācija ievērojami apgrūtina datu pirmsapstrādi un to turpmāko analīzi.

1.1. tabulā parādīti aritmētisko darbību veidi, ko var veikt ar dažāda veida datiem.

1.1. tabula

Aritmētisko darbību saraksts dažādiem datu tipiem

Datu tipi un skalas	***Aritmētiskās darbības***			
Kvalitatīvie	Saskaitīšana	Atņemšana	Reizināšana	Dalīšana
Nominālā	Nav	Nav	Nav	Nav
Ordinārā	Nav	Nav	Nav	Nav
Kvantitatīvie	Saskaitīšana	Atņemšana	Reizināšana	Dalīšana
Intervāla	Ir	Ir	Nav	Nav
Attiecību	Ir	Ir	Ir	Ir

Ir arī citas pieejas datu klasifikācijai. Tā kā šīs pieejas ir ierobežoti izmantojamas, tās šajā sadaļā nav izklāstītas.

1.3. Ievads datu analīzē

Tā kā datu pirmapstrāde, kas ir šīs grāmatas galvenā tēma, ir paredzēta datu turpmākajai analīzei, šī sadaļa iepazīstina ar datu analīzes pamatjēdzieniem. Tālāk sniegtais materiāls ir balstīts uz datiem, kas sniegti darbā [Uzhga-Rebrov O., 2021].

Statistikā jau sen ir izstrādātas spēcīgas metodes hipotēžu un teoriju patiesuma pārbaudei. Kā norādīts darbā [De Mast J., Kemper B. P. H ., 2009], eksperimentālo metožu teorijas, hipotēžu pārbaude un modeļu veidošana ir daži no svarīgākajiem statistikas ieguldījumiem. Hipotēze tiek pārbaudīta šādi – tiek apkopoti eksperimenti, novērojumi vai attiecīgie dati; dati tiek analizēti un, pamatojoties uz iegūtajiem rezultātiem, tiek izdarīti secinājumi par to, vai izvirzītā hipotēze ir patiesa vai noraidāma.

Šādu datu analīzi sauc par *apstiprinošo datu analīzi* (turpmāk ADA) (angliski *Confirmatory Data Analysis)*. Faktiski šāda analīze ir deduktīvs secinājums. Autors [Haig B.B, 2005] atzīmē, ka 150 gadus dabaszinātnēs izvēlētā metode ir bijusi hipotētiskā indukcija. Hipotētiski deduktīvā metode parasti tiek izteikta minimālos terminos: pētniekam ir jāierosina hipotēze vai teorija un tā ir vispārīgi jāpārbauda, iegūstot vienu vai vairākas prognozes. Šīs prognozes tiek pakļautas tiešai empīriskai pārbaudei. Ja prognozes apstiprina dati, tad rezultāts tiek pieņemts kā hipotēzes vai teorijas apstiprinājums. Ja prognozes nesakrīt ar datiem, tad šis fakts tiek pieņemts kā hipotēzes vai teorijas noliegums.

ADA ir ļoti nozīmīga loma zinātnisko, tehnisko, ekonomisko, sociālo un citu pētījumu rezultātā iegūto secinājumu apstiprināšanā vai noliegšanai.

Savukārt 20. gadsimta otrajā pusē, strauji attīstoties informācijas tehnoloģijām, sāka uzkrāties milzīgi dažāda veida datu apjomi. Paši dati ir pārskats par situāciju kādā jomā. Lietotājus var ļoti interesēt šajos datos slēptās zināšanas. Piemēram, uz datiem virzītajā mārketingā vēlams iegūt zināšanas par pārdošanas procesa īpatnībām, lai pieņemtu pārdomātus lēmumus par pārdošanas procesa efektivitātes uzlabošanu.

Vienkārši izsakoties, ir nepieciešama datu analīze, kas ļauj identificēt datos netiešā veidā attēlotās datu struktūras, modeļus un/vai atkarības atbilstošās iezīmes. Šāda veida analīzi sauc par *skaidrojošo datu analīzi* (turpmāk SDA) (angliski *Exploratory Data Analysis*).

SDA paveids ir *paredzamā datu analīze* (turpmāk PDA) (angliski *Predictive Data Analysis*). Šīs analīzes būtība ir identificēt nozīmīgas iezīmes ierobežotos datu apgabalos, lai turpmāk tās varētu attiecināt uz plašāku datu apgabalu. PDA plaši izmanto medicīnā. Piemēram, izmantojot datus, kas saistīti ar noteiktas slimības diagnostiku un ārstēšanu, tiek iegūti secinājumi un prognozes, ko var veiksmīgi izmantot citos medicīniskajos pētījumos.

SDA rašanās un attīstība galvenokārt ir saistīta ar amerikāņu pētnieka Džona Tukeja un viņa kolēģu vārdu. No daudzajiem šī autora darbiem jāmin vispilnīgākais darbs [Tukey J.W, 1977], kur faktiski noteikti SDA formālie pamati un izskatīti dažādi šīs analīzes aspekti.

Kā var formāli definēt SDA? Šī darba autors sniedz divas šī jēdziena definīcijas alternatīvas.

Tukeja darbā [Tukey J.W, 1977] SDA raksturota kā (1) filozofija vai attieksme, nevis noteiktu, formālu procedūru kopums; (2) datu visaptverošas izpratnes process, lai no tiem iegūtu informāciju; (3) vienkāršu aprakstošu novērtējumu izmantošana datu apkopošanai un atkārtotai izteikšanai; (4) uzsvērts datu grafiskais attēlojums; (5) elastība, piemērojot analīzi datu struktūrām un attiecīgi arī neatpazītiem modeļiem; (6) koncentrēšanās uz eksperimentālu modeļu veidošanu un hipotēžu ģenerēšanu.

Behrena darbā [Behrens J.T., 1997] sniegta cita SDA definīcija, kas šeit dota citāta veidā no šī darba: "Šajā rakstā tiek uzskatīts, ka SDA ir īpaša datu analīzes tradīcija, kas izriet no Džona Tukeja un viņa domubiedru darba. Šo SDA tradīciju var raksturot ar: (a) neatkarīgas izpratnes saglabāšanu par datiem, kas pievēršas plašajam jautājumam "Kas šeit notiek?"; b) datu grafisko attēlojumu uzturēšanu; c) koncentrēšanos uz eksperimentālu modeļu veidošanu un hipotēžu ģenerēšanu, izmantojot iteratīvu modeļu specifikācijas, modeļa atkārtotas specifikācijas procesu; d) stabilu aplēšu izmantošanu, atkārtotu izteiksmi un apakškopu analīzi; (e) skepticisma attieksmi, elastību attiecībā uz izmantotajām metodēm".

Jebkuras skaidrojošās analīzes mērķis ir atpazīt un izskaidrot sākotnējā datu kopā slēptās pazīmes. Literatūrā nav vispārpieņemta "raksturīgo pazīmju" jēdziena. Angļu valodas literatūrā tos bieži sauc par "paraugiem" vai "parādībām". Modeļu piemēri var būt standarta vai specifiski atribūtu vērtību sadalījumi, atsevišķu atribūtu vērtību korelācija, klasteru klātbūtne datu kopā un daudzas citas iespējas.

Kādi ir SDA galvenie mērķi? Darbā [De Mast J., Kempers B.P.H., 2009] autori apgalvo, ka SDA mērķis nav izdarīt secinājumus par iepriekš definētajiem pētījuma jautājumiem (vai tā ir modeļa konstruēšana, parametru novērtēšana vai hipotēzes apstiprināšana vai noraidīšana), patiesībā SDA problēmu risināšana bieži tiek izmantota savāktajiem datiem bez precīzi definētām hipotēzēm.

Vissvarīgākais par SDA ir tas, ka analīzi var veikt, sākotnēji nemaz neformulējot hipotēzes, kuras varētu secināt no datiem. Tas nenozīmē, ka SDA tiek veikta "tukšā vietā", t. i., uzsākot analīzi, nav nojausmas, ko šīs analīzes rezultātā var iegūt. Parasti tiek formulēti pētījuma jautājumi, uz kuriem var atbildēt ar analīzi. SDA mērķis ir apstrādāt un parādīt sākotnējo datu kopu tā, lai varētu identificēt konkrētus datu modeļus. Pamatojoties uz iegūtajiem paraugiem, var izvirzīt hipotēzi par to, kādas zināšanas ir ietvertas datos. Vēl viens labi izstrādāta SDA mērķis ir identificēt iekšējos faktorus datos, kas izraisa iegūtos modeļus.

Kopumā SDA ietver šādu vispārinātu procedūru kopumu [De Mast J., Kempers B.P.H., 2009]:

1) datu attēlošanu;
2) būtisku pazīmju identificēšanu;
3) pamanāmu pazīmju interpretāciju.

SDA izmantotie neapstrādātie dati ir atbilstoši jāapstrādā un jānorāda piemērotā (bieži vien grafiskā) veidā, lai ļautu identificēt datos raksturīgās pazīmes.

SDA pēc būtības ir induktīvā pieeja, t. i., analīzes process tiek veikts saskaņā ar shēmu "no konkrētā uz vispārīgo". Kad tiek identificēts datu modelis, tas ir jāinterpretē (jāpaskaidro). Šī ir abduktīvo secinājumu darbības joma. Autors iepazīstinās ar abduktīvo secinājumu definīcijām, ko ierosinājuši labi pazīstami pētnieki SDA jomā.

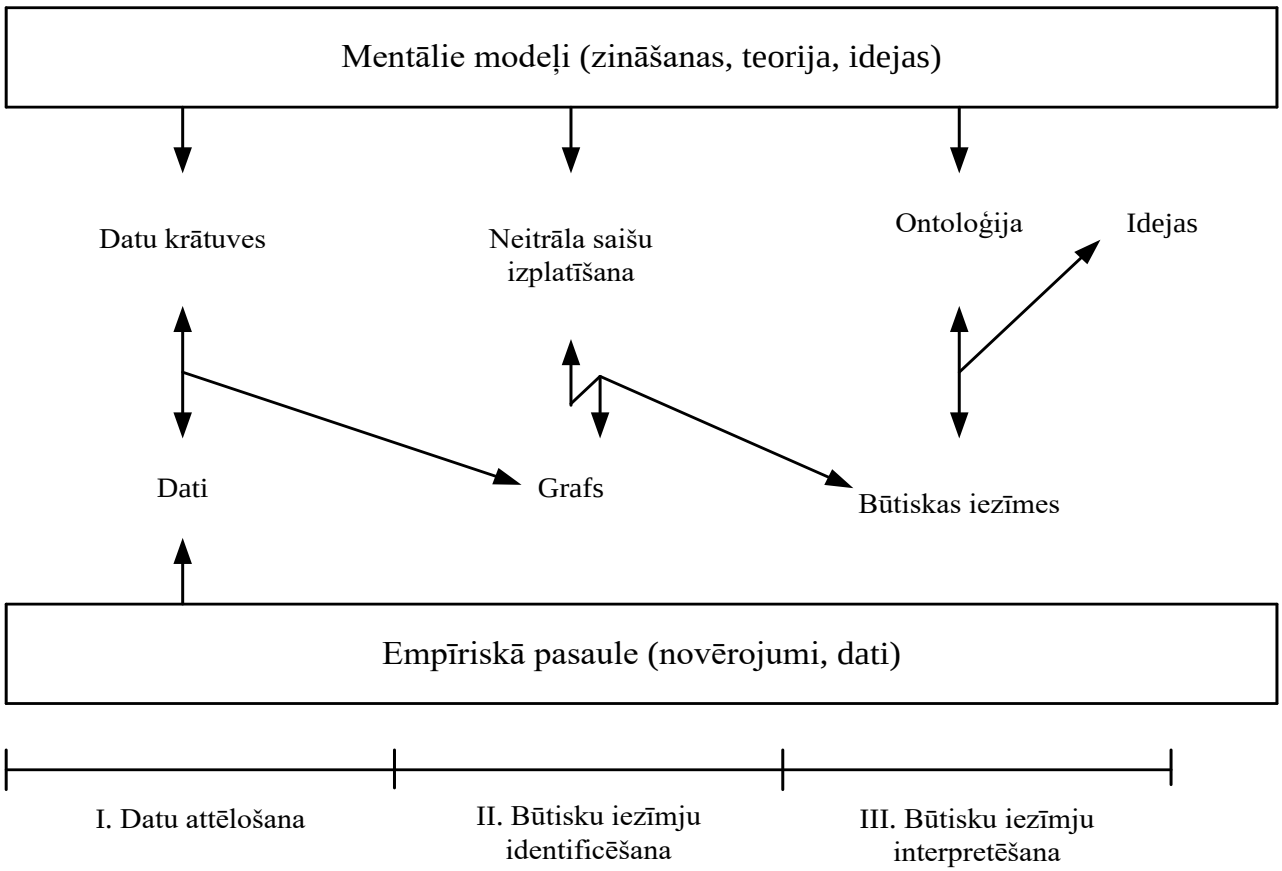
Secinājumu veidu, kad, apvienojot jēdzienus un analoģijas, daļas ir potenciāli saistītas un no tā izriet pamanāmas iezīmes, bieži sauc par abduktīvu [De Mast J., Kempers B.P.H., 2009].

Datu analītiķim ir grūti sākt un izskaidrot datus bez jebkādas atsauces. Tradicionāli pētnieki ir klasificējuši spriešanas veidus pētījumos kā indukciju (vadoties pēc pierādījumiem) un dedukciju (vadoties pēc teorijas vai hipotēzes). Patiesībā ir trešais veids – abdukcija. Abduktīva spriešana ne vienmēr sākas ar pilnībā izstrādātiem modeļiem vai vispār bez modeļiem... Abduktīvajā testēšanā SDA būtība ir nevis atteikšanās no modelēšanas un sākotnējām koncepcijām, bet gan izpēte un iedrošinājums nesākt analīzi tikai ar "stingru" sākotnējo koncepciju... Abdukcija – tā nav simboliska loģika, bet gan kritiskā domāšana [Cheng H.Y., 1994].

Abduktīva secinājuma pamatideja ir meklējama amerikāņu filozofa un zinātnieka Pīrsa darbos. Turpmākā filozofijas un mākslīgā intelekta attīstība ir balstīta uz Pīrsa idejām, lai sniegtu pētniekiem būtisku izpratni par abduktīvo spriešanu [Haig B.B., 2005].

Apkopojot iepriekš minētās definīcijas, var apgalvot, ka abduktīvie secinājumi ir saistīti ar hipotēžu meklēšanu, kas vislabāk izskaidro datos atpazītos modeļus. Lai to izdarītu, pētniekam ir jāpatur prātā vai jāizmanto ārējie avoti, lai analizētu visus iespējamos identificēto modeļu skaidrojumus. Abduktīvo secinājumu procesus var attēlot diagrammas veidā, kas parādīts 1.3. attēlā [De Mast J., Kempers B.P.H., 2009]. Jāpiebilst, ka šīs diagrammas autori pieņem, ka visus mentālos modeļus var attēlot grafiku veidā. Realitātē potenciālo mentālo modeļu reprezentācijas iespējas ir daudz plašākas. Grafiki ir tikai viens alternatīvs veids, kā attēlot šādus modeļus.

Darbā [Haig B.B., 2005], atsaucoties uz citu avotu, ir parādīts, ka, ģenerējot hipotēzes, tiek nošķirta eksistenciālā un analogā abdukcija. *Eksistenciālā abdukcija* postulē iepriekš nezināmu paraugu esamību, bet *abdukcija pēc analoģijas* balstās uz iepriekšējiem hipotēžu formulēšanas gadījumiem, lai atrastu analogu, piemērotu pašreizējam SDA uzdevumam. Kā norādīts darbā [Haig B.B., 2005], eksistenciālā abdukcija ir abdukcijas veids, ko plaši izmanto faktorus aprakstošo hipotēžu ģenerēšanai. Skaidrojošās faktoru analīzes mērķis ir atvieglot slēpto mainīgo noteikšanu. Lai gan skaidrojošās faktoru analīzes būtība un mērķis tiek diskutējams, to var saprast kā abduktīvu metodi hipotēžu ģenerēšanai.



1.3. attēls. **Abduktīvo secinājumu procesu shematisks attēlojums**

Standarta SDA darbojas saskaņā ar šādu shēmu:

Uzdevums → Dati → Analīze → Modelis → Secinājums.

Šajā sistēmā secinājumu rezultātā tiek formulēta hipotēze, kas ticami izskaidro datu modeļus.

Pēdējos gados standarta SDA procedūras ir papildinātas ar procedūrām, kas ļauj pārbaudīt hipotēzes, t. i., ir iekļauti arī ADA elementi.

Kā norādīts darbā [Cheng H.Y, 2010], SDA atbalsta mainīgos lielumus, konstrukcijas un hipotēzes, kas tiek uzskatītas par vērtīgām, bet ADA veic nākamo soli, lai to apstiprinātu. Tomēr, izmantojot atkārtotu paraugu ņemšanu *(resampling),* skaidrojošo datu analīze var vienlaikus atbalstīt un apstiprināt modeli.

Atkārtota paraugu ņemšana skaidrojošās datu analīzes kontekstā attiecas uz vispārināšanu starp paraugiem, nevis rezultātu, kas iegūts no viena parauga. Lai iegūtu pareizus vispārinātus rezultātus, tiek izmantota šķērsvalidācija (*crossvalidation).*

Šeit ir citāts no [Cheng HY, 2010], kurā autors atklāj mūsdienu uzskatus par SDA: "Viena no problēmām ar standarta pieejām SDA ir tā, ka SDA raksturlielumi ir saistīti gan ar datu atribūtiem (sadalījums, mainīgums, linearitāte, novirzes, mērījumu skalas utt.), gan ar galīgajiem mērķiem (klasteru atpazīšana, mainīgo un slēptu modeļu un sarežģītu saišu atpazīšana). Faktiski datu atribūtu apstrāde ir vidējā posma procedūra, nevis beigu. Mūsdienu pieejas SDA pamatā ir koncentrēšanās uz mērķiem, nevis vērtībām. Tāpēc dažas modernas pieejas var apstiprināt analīzes rezultātus, un tos var salīdzināt ar SDA - ADA kombināciju." Lai ilustrētu šo apgalvojumu, 1.3. tabulā [Liu B. Q., 2014] ir apkopoti tradicionālās SDA un mūsdienu SDA salīdzināšanas rezultāti.

1.3. tabula

Tradicionālās un mūsdienu SDA salīdzinājums

	Tradicionālā SDA	Mūsdienu SDA
Tips	Orientēta uz vērtībām	Orientēta uz mērķi
Forma	Vizualizēta, raksturojoša analīze	Vispārīga, raksturojoša analīze
Līdzekļi	Vienkārša aritmētika. Vienkāršoti attēlojumi	Aprakstošā statistika. Attīstītas datu vizualizācijas metodes. Datu iegūšanas metodes
Galvenās iezīmes	Nedod noslēguma atbildes	Var nodrošināt noslēguma atbildes

Datu iegūšana (*Data Mining)* ir kļuvusi plaši izplatīta. Lai iegūtu detalizētu ievadu šajā zinātniskajā un lietišķajā jomā, var atsaukties, piemēram, uz [Han J., Kamber M., Pei J., 2012; Zaki M.J., Wagner M.Jr., 2014; Xu G., Zong Y., Yang Zh ., 2013]. Papildus šiem avotiem plaša informācija par datu iegūšanu ir atrodama arī citās monogrāfijās un daudzos rakstos.

Kas ir datu iegūšana? Tas ir dažādu informācijas tehnoloģiju pielietojums lielām datu kopām, lai tās apstrādātu un identificētu interesējošos paraugus. Jāatzīmē, ka termins "datu iegūšana", šķiet, nav pārāk labs veids, kā aprakstīt šo procesu. Tāpēc literatūrā dažreiz tiek izmantots alternatīvs nosaukums: zināšanu atklāšana no datiem (*knowledge discovery from data).* Tomēr termins "datu iegūšana" ir daudz izplatītāks nekā otrais termins, tāpēc tā lietošana ir kļuvusi par tradīciju visā pasaulē.

Kādas ir līdzības un atšķirības starp SDA un datu iegūšanu? Būtībā šīs ir ļoti līdzīgas pieejas. Abas pieejas apstrādā un parāda datus īpašos veidos, lai identificētu modeļus, t. i., nozīmīgas datu pazīmes. Abās pieejās mērķis ir izskaidrot radušos modeļus. Abas pieejas izmanto ADA, lai pārbaudītu iegūtos rezultātus.

Starp abām pieejām ir atšķirība. Datu iegūšanā plaši tiek izmantotas specifiskas metodes, kas paredzētas tieši liela apjoma datu apstrādei automātiskajā režīmā, piemēram, iztrūkstošo atribūtu vērtību atjaunošana, atribūtu skaita samazināšana, noņemot neefektīvās atribūtu vērtības un citas. Jebkurš datu iegūšanas process kā neatņemama sastāvdaļa ietver vienu vai otru datu analīzes veidu.

1.4. Sākotnējo datu pirmapstrādes procedūras

Kvalitatīvu analīzi var veikt tikai kvalitatīviem datiem. Datu kvalitātes sastāvdaļas ir to precizitāte, pilnīgums un saskaņotība. Datu precizitāte ir saistīta ar atribūtu vērtību precizitāti un pareizību. Datu pilnīgums attiecas uz datu paraugu un atribūtu vērtību pieejamību. Dati ir pilnīgi, ja datos ir visi nepieciešamie un pietiekamie modeļi un atribūtu vērtības. Datu saskaņotība attiecas uz atbilstošo datu modeļu un atribūtu klātbūtni, pareizu kodēšanu un pareizu datu prezentācijas formātu.

Reālās situācijās avota dati var būt neprecīzi, nepilnīgi un nesakarīgi. Tam ir daudz iemeslu: neprecīzi datu ieraksti, datu vākšanas rīku bojājumi, personāla kļūdas, ievadot datus datorā, un daudzi

citi, bieži subjektīvi iemesli. Tāpēc tieša neapstrādātu datu analīze var sniegt neapmierinošus rezultātus.

Lai veiktu kvalitatīvu datu analīzi, dati ir iepriekš jāapstrādā tā, lai tie būtu pieņemamā kvalitātē turpmākai analīzei. Datu pirmapstrāde prasa daudz laika un pūļu. Darbā [Pyle D., 1993] atzīmēts, ka datu pirmapstrādei var būt nepieciešami līdz 60 % no kopējā laika, lai veiktu pilnīgu datu analīzi. Tas arī parāda datu pirmapstrādes nozīmi.

Datu pirmapstrādi var definēt kā dažādu veidu procedūru secības veikšanu, lai oriģinālos datus iegūtu tālākai analīzei piemērotā formā. Visu attiecīgo procedūru kopumu var iedalīt četrās galvenajās kategorijās [Acuna E., 2011; Alasadi S., 2017].

1. *Datu attīrīšana*. Šis ir pirmais datu pirmapstrādes solis. Tas ietver iztrūkstošo atribūtu vērtību aizpildīšanu vai izņemšanu, trokšņainu datu izlīdzināšanu, anomāliju identificēšanu un izņemšanu no datiem, kā arī datu neatbilstību novēršanu.

2. *Datu integrācija*. Ja analīzei tiek izmantoti dati no dažādiem avotiem, šie dati ir jāintegrē vienotā veselumā. Tas var radīt nepieciešamību pielāgot atsevišķas datu sadaļas, lai tām būtu vienāds formāts, viena kodēšanas sistēma un uzdošana.

3. *Datu transformācija*. Daudzi datu analīzes algoritmi darbojas labāk, ja atribūtu vērtības tiek normalizētas vai mērogotas noteiktā veidā. Normalizācijas veikšana ir nepieciešama, lai izmantotu uz attālumu balstītus algoritmus, jo attālumu rezultāti starp atribūtiem ar lielākām vērtībām pārsvērs attālumu mērīšanas rezultātus starp atribūtiem ar mazākām vērtībām.

Citas datu pārveidošanas metodes ietver datu vispārināšanu. Šādas metodes rada jaunus atribūtus, apvienojot esošos atribūtus vai izmantojot augstāka līmeņa koncepcijas.

4. *Datu samazināšana*. Datu samazināšanas galvenā ideja ir izveidot jaunu datu izlasi, kas ir mazāka par sākotnējo datu kopu, bet saglabā visas savas pamatīpašības. Darbā [Han J., Kamber M., Pei J., 2012] tiek prezentētas šādas datu samazināšanas stratēģijas:

- *dimensijas samazināšana,* kas saistīta ar neatbilstošu vai lieku atribūtu noņemšanu;
- *datu saspiešana*, kad tiek izmantoti kodēšanas mehānismi, lai izveidotu samazinātu vai saspiestu sākotnējo datu attēlojumu;
- *skaita samazināšana*, kur dati tiek aizpildīti vai tiek novērtēti, izmantojot alternatīvus mazākus to attēlojumus, piemēram, parametriskus modeļus vai neparametriskas metodes, piemēram, klasterizāciju;
- *jēdzienu hierarhijas diskretizācija un ģenerēšana*, kur neapstrādātu datu vērtības tiek aizstātas ar rangiem (intervāliem) vai augstākiem konceptuālajiem līmeņiem;
- *datu elementu atlase*, kur tiek atlasīta labāko datu vienību (paraugu) apakškopa no pilnas kopas. Šī pieeja ir efektīva laika un izmaksu ziņā.

Lielākā daļa šajā sadaļā aprakstīto datu pirmapstrādes procedūru tiks detalizēti apskatītas nākamajās nodaļās.

2. DARBĪBAS AR IZTRŪKSTOŠĀM ATRIBŪTU VĒRTĪBĀM

2.1. Problēmas formulējums

Datu analīzē iztrūkstoši dati parasti ir kā likums, nevis izņēmuma gadījums. Datu iztrūkumam ir daudz iemeslu: respondentu nevēlēšanās atbildēt uz dažiem socioloģisko un politisko pētījumu jautājumiem, datu savākšanai izmantoto sensoru darbības traucējumi, datu daļas zaudējumi, personāla kļūdas datu ievades procesā utt.

Datu iztrūkums rada nopietnas problēmas datu tālākai analīzei [Kang H., 2013]:

- samazinās datu analīzes metožu statistiskais nozīmīgums;
- var rasties neobjektīvi vērtējumi un secinājumi;
- jūtami samazinās datu izlases reprezentativitāte;
- datu analīze kļūst daudz sarežģītāka.

Kā atzīmēts darbā [Dong Y., Peng Ch.-Y.J., 2013], iztrūkstošie dati var parādīties divos līmeņos: datu vienumu līmenī (objekti, priekšmeti utt.) un datu struktūrā (atribūtos). Datu iztrūkums vienumu līmenī rodas, ja netiek apkopota vai nav pilnīga informācija par šiem vienumiem jeb elementiem. Datu iztrūkums atribūtu līmenī nozīmē, ka datos trūkst atsevišķu atribūtu vērtību.

Kas attiecas uz datu iztrūkumu vienumu līmenī, šo problēmu var atrisināt, vispirms organizējot visu nepieciešamo datu savākšanu. Tā drīzāk ir organizatoriska, nevis konceptuāla problēma. Daudz grūtāk ir tikt galā ar iztrūkstošajām atribūtu vērtībām. Tagad ir vispāratzīts, ka vienīgais veids, kā atrisināt šo problēmu, ir iztrūkstošo vērtību imputācija (*imputation*), t. i., aizstāšana ar citām atbilstošām vērtībām. Statistikā imputācija ir iztrūkstošo datu aizstāšanas process ar aizvietotām vērtībām. Pašlaik ir izstrādāts liels skaits metožu iztrūkstošo atribūtu vērtību imputācijai. Šo metožu analīze arī ir šīs nodaļas mērķis.

Teorētiskie un praktiskie imputācijas procesu pamati darbībām ar iztrūkstošām atribūtu vērtībām tika noteikti darbos [Rubin D.B., 1976; Little R.J.A., Rubin D.B., 2002; Schafer J.L., 1997]. Tieši šie darbi bija par pamatu efektīvu mūsdienu metožu izstrādei un izmantošanai, lai risinātu iztrūkstošo datu problēmu.

Trīs pamata faktori, kas ietekmē piemērotu datu imputācijas metodes izvēli, ir [Dong Y., Peng Ch.-Y.J., 2013]:

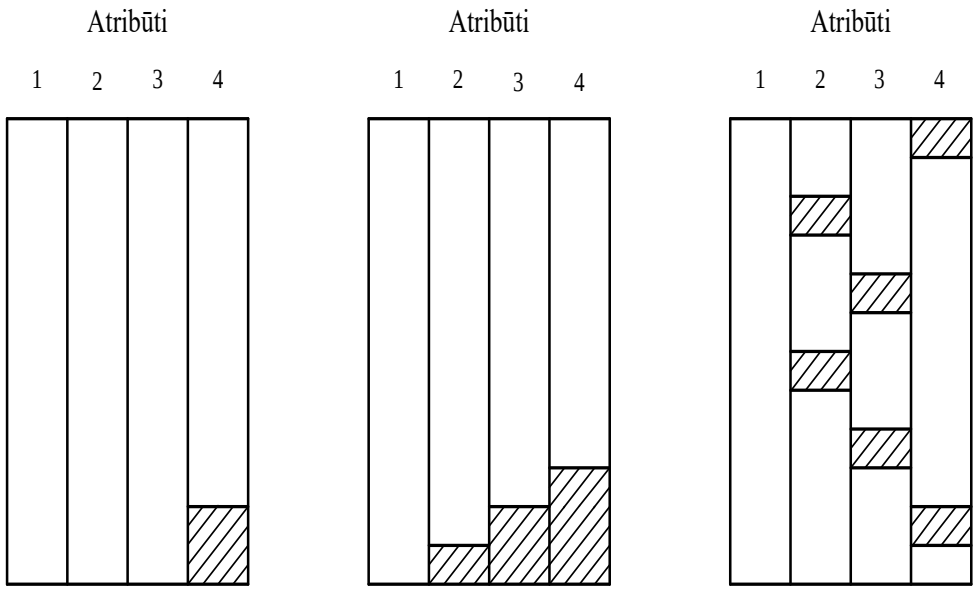
- iztrūkstošo datu īpatsvars;
- iztrūkstošo datu paraugi;
- iztrūkstošo datu veidošanās modelis.

Iztrūkstošo datu īpatsvars ir tieši saistīts ar šo datu analīzes kvalitāti. Diemžēl nav pamatots iztrūkstošo datu skaita ierobežojums, pie kura analīze var tikt veikta bez būtiskiem rezultātu izkropļojumiem. Literatūrā tiek pieminēta 5 % iztrūkstošo datu robeža, taču tas ir tikai empīrisks novērtējums bez jebkāda teorētiskā pamatojuma.

Iztrūkstošo datu paraugi tiek klasificēti šādi: vienveidīgie, monotonie un patvaļīgie.

Vienveidīgā parauga modelis rodas, ja datos trūkst vērtību tikai vienam atribūtam. Monotonā parauga modeli raksturo fakts, ka, ja attiecīgā veidā tiks sakārtoti atribūti, tad, sākot ar atribūtu *j*, visiem nākamajiem atribūtiem būs iztrūkstošas vērtības. Patvaļīgā parauga modelī dažiem vai visiem atribūtiem iztrūkstošie dati ir patvaļīgā veidā.

Ilustrācijas nolūkā iztrūkstošo datu paraugi ir parādīti 2.1. attēlā [Sim J., et al., 2015].



a) Vienveidīgs paraugs b) Monotons paraugs c) Patvaļīgs paraugs

2.1. attēls. **Datu paraugu grafiskais attēlojums**

Vienveidīgā parauga modelī (skat. 2.1.a attēlu) vērtību nav tikai ceturtajā atribūtā. Monotonā parauga modelī (skat. 2.1.b attēlu) vērtību nav 2., 3. un 4. atribūtam. Patvaļīgā parauga modelī (skat. 2.1.c attēlu) vērtību nav 2.–4.atribūtā.

Vienveidīgā parauga modelis sastopams retos gadījumos. Par monotonā parauga modeļa piemēru var minēt aptauju rezultātus, kuros respondents, neatbildot uz kādu aptaujas jautājumu, neatbild arī uz citiem jautājumiem, kas kaut kādā veidā saistīti ar neatbildēto. Patvaļīgā parauga modelis ir visbiežāk izplatītais datu analīzes uzdevumos.

Kā jau iepriekš tika atzīmēts, pastāv daudz iemeslu, kāpēc iztrūkst datu atribūtu vērtības. Ja neņem vērā konkrēto datu iztrūkuma iemeslu, ir iespējams definēt dažus vispārīgus datu iztrūkuma (atribūtu vērtības) rašanās un apstrādes paņēmienus. Šādi paņēmieni pirmo reizi tika ieviesti darbā [Rubin DB, 1976]. Šie paņēmieni tagad ir vispārpieņemti, un atribūtu vērtību imputācija vienmēr balstās uz šiem paņēmieniem.

Tālāk tiks definēti un detalizēti apskatīti trīs datu iztrūkuma paņēmieni, kas balstās uz atribūtu vērtību sadalījumu. Ja pieņem, ka dati satur *m* datu elementus (datu matricas rindas) un *n* atribūtus (datu matricas kolonnas), tad datu matrica A būs šādā veidā (skat. 2.1. tabulu).

2.1. tabula

Datu matrica A

Datu elementi	Atribūti					
	a_1	a_2	a_3	a_4	....	a_{1n}
o_1	a_{11}	?	a_{13}	a_{14}	....	a_{2n}
o_2	a_{21}	a_{22}	?	a_{24}	....	?
o_3	?	a_{32}	a_{33}	?	....	a_{3n}
o_4	a_{41}	a_{42}	a_{41}	?	....	?
o_5	?	?	a_{51}	a_{54}	....	a_{5n}
....	....	....	....	....	....	....
o_m	a_{m1}	a_{m2}	?	a_{m4}	....	a_{mn}

Matricas rindas satur datu elementus $o_1, o_2, ..., o_m$, kolonnās tiek parādīti atribūti $a_1, a_2, ..., a_n$. Vērtība a_{ij}, $i = 1,...,\mathrm{m}$, , $j = 1,...,\mathrm{n}$ $i-tās$ rindas un $j-tās$ kolonnas krustpunktā parāda atribūta a_j vērtību elementam o_i. Jautājuma zīmes matricas šūnās norāda, ka trūkst atbilstošo atribūtu vērtības.

Vispārīgie pieņēmumi par datiem ir tādi, ka visi datu elementi ir neatkarīgi un tiem ir identisks atribūtu vērtību sadalījums. Šis pieņēmums ir formulēts darbā [Schafer J.L., 1997].

Saskaņā ar šo pieņēmumu atribūtu vērtību daudzdimensiju sadalījuma funkciju sākotnējā datu kopā var attēlot formā:

$$P(\mathbf{A}/\theta)=\prod_{j=1}^{m} f\left(\mathbf{a}_j/\theta\right), \tag{2.1}$$

kur θ – j atribūta vērtību sadalījuma parametrs.

Parasti parametra θ vērtība nav zināma. Par sadalījuma funkcijas formu tiek izteikti šādi pieņēmumi:

- daudzfaktoru normālais sadalījums kvantitatīvo atribūtu vērtību gadījumam;
- multinominālais modelis kvalitatīvu atribūtu vērtību gadījumam;
- jauktais modelis kvalitatīvu un kvantitatīvu atribūtu vērtību kombināciju gadījumam.

Pieņēmums par kvantitatīvo atribūtu vērtību daudzfaktoru normālu sadalījumu ir diezgan ierobežojošs iztrūkstošo atribūtu vērtību imputācijas kontekstā. Bet literatūrā ir parādīts, ka šī pieņēmuma pārkāpums nopietni neietekmē imputācijas rezultātus.

Lai korekti noteiktu datu iztrūkuma paņēmienus, iepriekš definētajā matricā A, kuras šūnās vērtība 1 nosaka atribūta vērtības esamību, bet 0 – nav atribūta vērtības. Tādējādi iegūst jaunu matricu B (skat. 2.2. tabulu).

2.2. tabula

Datu matrica B

Datu elementi	Atribūti					
	a_1	a_2	a_3	a_4	....	a_n
o_1	1	0	1	1	....	1
o_2	1	1	0	1	....	0
o_3	0	1	1	0	....	1
o_4	1	1	1	0	....	0
o_5	0	0	1	1	....	1
....	....	....	....	....	....	....
o_m	1	1	0	1	....	1

Sākotnējo datu kopu var korekti attēlot veidā $\mathbf{A}=\left(\mathbf{A}_{obs},\mathbf{A}_{mis}\right)$, kur $\mathbf{A}_{obs}$ ir faktisko (novērojamo) atribūtu vērtību izlase un $\mathbf{A}_{mis}$ ir iztrūkstošo atribūtu vērtību izlase.

Iztrūkstošo atribūtu vērtību reprezentē sadalījumu veidā $P\left(\mathbf{B}/\mathbf{A},\xi\right)$, kur ξ – neesamības jeb prombūtnes parametrs. Vispārīgā veidā šo priekšstatu var paplašināt šādi:

$$P\left(\mathbf{B}/\mathbf{A},\xi\right)=P\left(\mathbf{B}/\mathbf{A}_{obs},\mathbf{A}_{mis},\xi\right). \tag{2.2}$$

Izmantojot iepriekš sniegtos apzīmējumus un vienādojumus, var definēt iztrūkstošo datu priekšstatus.

1. *Nejaušs iztrūkums* (*Missing at random* (*MAR)*). Šajā gadījumā iztrūkstošo vērtību varbūtību sadalījums ir definēts kā

$$P\left(\mathbf{B}/\mathbf{A}_{obs},\mathbf{A}_{mis},\xi\right)=P\left(\mathbf{B}/\mathbf{A}_{obs},\xi\right). \tag{2.3}$$

Vienādojums (2.3) var tikt interpretēts šādā veidā: iztrūkstošo vērtību varbūtība ir atkarīga tikai no novērojamo atribūtu vērtību izlases $\mathbf{A}_{obs}$ un prombūtnes parametra ξ.

Nejauša iztrūkuma gadījumā var apgalvot, ka attiecībā uz konkrētām novērotajām atribūtu vērtībām no izlases $\mathbf{A}_{obs}$ atlikušo vērtību sadalījums starp esošajām vērtībām un iztrūkstošajām vērtībām ir vienāds.

Lai demonstrētu atribūtu nejauša iztrūkuma piemēru, pieņem, ka cilvēkiem, kuri vēlas mainīt profesiju, tiek organizēti karjeras pārorientēšanās kursi. Kursa sākumā dalībniekiem tika lūgts atbildēt

uz testa jautājumiem, lai novērtētu viņu zināšanu un prasmju līmeni. Pieņem, ka daži no dalībniekiem atbildēja tikai uz daļu no jautājumiem, kas liecināja par viņu zināšanu un pieredzes zemo līmeni.

Mācību procesā daļa no dalībniekiem ar zemu novērtētu līmeni pameta kursus. Pēc kursu pabeigšanas dalībniekiem tika lūgts atbildēt uz cita testa jautājumiem, lai novērtētu viņu profesionālās pārkvalifikācijas rezultātus. Acīmredzami, ka jauniegūtajos rezultātos nebūs kursus pametušo darbinieku datu.

2. *Pilnīgi nejaušs iztrūkums* (*Missing completely at random* (*MCAR*)). Šis iztrūkuma mehānisms formāli tiek izteikts kā

$$P\left(\mathbf{B} / \mathbf{A}_{obs}, \mathbf{A}_{mis}\xi\right) = P\left(\mathbf{B}, \xi\right). \tag{2.4}$$

Šis vienādojums izsaka to, ka iztrūkstošo vērtību sadalījums ir tāds pats kā pašreizējo atribūtu vērtību sadalījums. Tāpēc pieejamās vērtības var uzskatīt par paraugu no vērtību izlases, kurā ir visas vērtības, ieskaitot tās, kas atrodas iztrūkstošo vērtību vietā. Taču iztrūkstošās vērtības nevar uzskatīt par paraugu no pilnas vērtību izlases.

Pilnīgi nejauši iztrūkstošo atribūtu vērtību piemērs – organizācija nolēma veikt savu darbinieku medicīnisko pārbaudi, lai novērtētu viņu veselību. Lai to izdarītu, darbiniekiem bija jāziedo asinis analīzei, jāveic fluorogrāfija un elektrokardiogramma. Pirms asins analīžu veikšanas laborants nejauši sasita vairākas ampulas ar asinīm. Veicot elektrokardiogrammu, sabojājās aprīkojums un par dažiem darbiniekiem rezultāti netika iegūti. Par fluorogrāfijas veikšanu atbildīgais speciālists nejauši pazaudēja dažus rezultātus.

3. *Nav nejaušs iztrūkums* (*Not missing at random (NMAR*)) . Formāli tas nozīmē, ka iztrūkstošo vērtību sadalījums ir atkarīgs gan no novērotajām vērtībām, gan no iztrūkstošajām vērtībām:

$$P\left(\mathbf{B} / \mathbf{A}_{obs}, \mathbf{A}_{mis}, \xi\right). \tag{2.5}$$

Šis ir visnelabvēlīgākais iztrūkstošo datu gadījums, jo šāda veida iztrūkstošo atribūtu vērtību imputācija nav piemērota plaši izmantotām metodēm.

Spilgts piemērs šim gadījumam var būt aptaujas rezultāti, kur viens no jautājumiem ir personu ienākumu līmenis. Prakse rāda, ka personas ar zemu un vidēju ienākumu līmeni faktiski norāda savus reālos ienākumus. Savukārt privātpersonas ar augstu ienākumu līmeni labprātāk nenorāda savus reālos ienākumus.

Ja pieņem, ka *MAR* datu iztrūkuma mehānismā vērtību sadalījuma parametrs $\left(\theta\right)$ un iztrūkuma parametrs $\left(\xi\right)$ ir neatkarīgi, tad tiek uzskatīts, ka iztrūkstošo datu mehānisms ir ignorējams [Little R., J., A. , Rubin D. B., 2002]. Tas pats attiecas uz *NMAR* mehānismu, jo *NMAR* ir īpašs *MAR* gadījums.

Lielākā daļa pašreizējo jaudīgo metožu iztrūkstošo atribūtu vērtību piedēvēšanai pieņem, ka tām trūkst *MAR* vai *NMAR* iespēju. Tomēr ļoti bieži reālās datu analīzes problēmās ir lielākas vai mazākas novirzes no šīm metodēm. Praktiskie pētījumi ir parādījuši, ka šo iespēju neizmantošana būtiski neietekmē imputācijas rezultātus. Tāpēc iztrūkstošo atribūtu vērtību imputācija dažreiz tiek veikta, nepārbaudot faktiskos iztrūkstošo datu iespēju metodes.

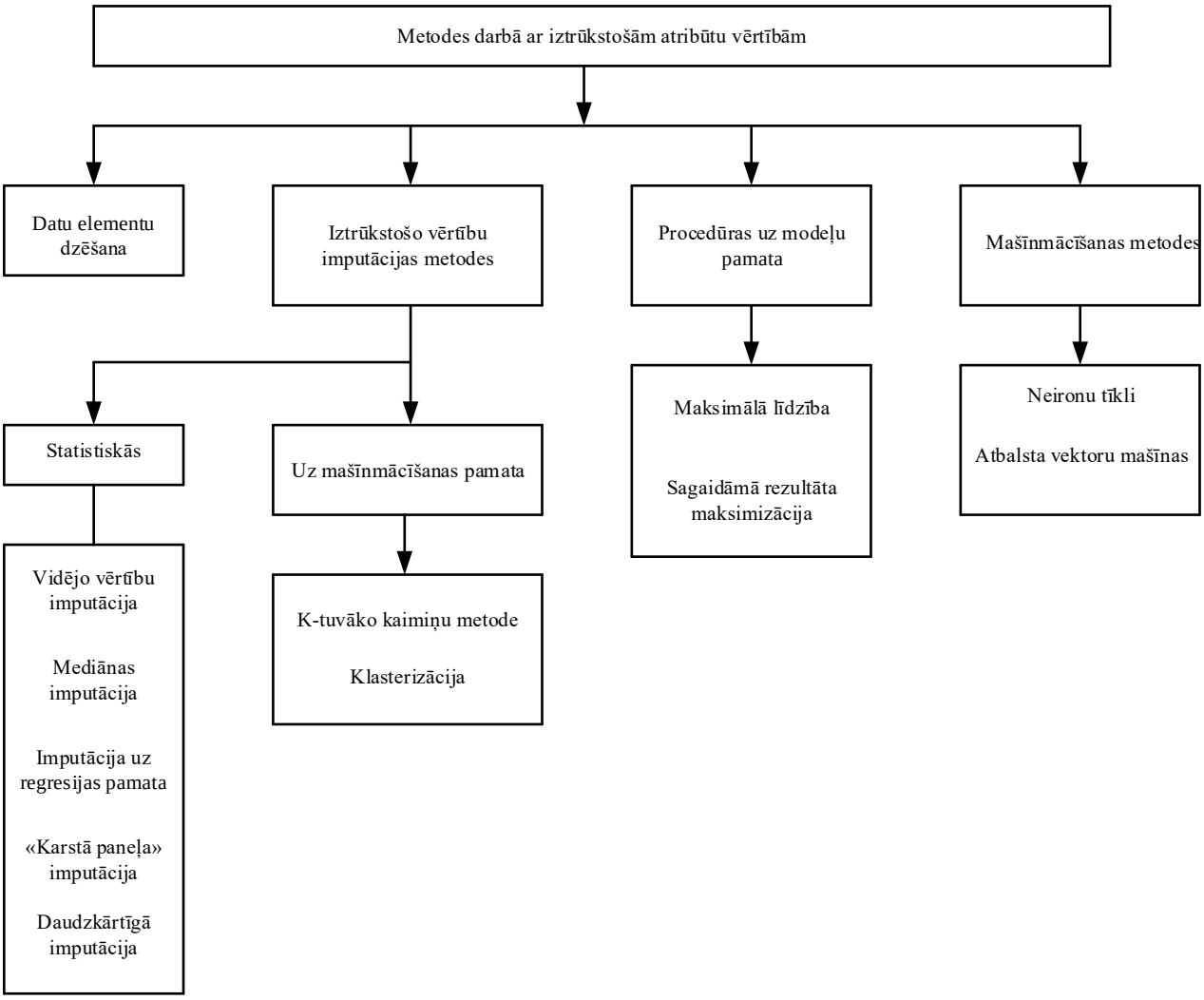
Šīs sadaļas noslēgumā tiek piedāvātas pamata metodes, kā rīkoties ar iztrūkstošām atribūtu vērtībām, kā arī dažādas šo metožu klasifikācijas iespējas. 2.2. attēlā sniegts klasifikācijas variants, ko var uzskatīt par dažādos literatūras avotos piedāvāto variantu vispārinājumu.

Datu elementu dzēšana ietver visu elementu izdzēšanu no sākotnējiem datiem, kuriem trūkst vismaz vienas atribūta vērtības. Lai gan šī metode nav saistīta ar atribūtu vērtību imputāciju, ir lietderīgi to izmantot, ja ir liels datu apjoms un ļoti maz iztrūkstošo atribūtu vērtību (apmēram 1–3 %).

Kopumā iztrūkstošo atribūtu vērtību imputāciju var iedalīt divās lielās klasēs:

- *viena imputācija*, kad tiek aprēķinātas atsevišķas iztrūkstošās vērtības;
- *daudzkārtīgā imputācija*, kad vienlaicīgi tiek imputētas iztrūkstošo vērtību grupas.

Imputācijas metodes var iedalīt divās grupās: *statistikas metodes* un *uz mašīnmācīšanos balstītas metodes*.



2.2. attēls. **Metožu klasifikācijas grafiskais attēlojums iztrūkstošo atribūtu vērtību apstrādē**

Pirmajā metožu grupā ietilpst vidējo vērtību imputācija, uz regresiju balstīta imputācija, mediānas imputācija, "karstā paneļa" (*Hot Deck*) imputācija un daudzkārtīgā imputācija. Visas šīs grupas metodes veic vienu imputāciju, izņemot vairākkārtēju imputāciju, kā to arī norāda pats metodes nosaukums. Šī metode dažkārt tiek klasificēta kā uz modeli balstīta procedūra, jo metode izmanto noteiktu datu modeli.

Otrā metožu grupa veic vienu imputāciju, pamatojoties vai nu uz k-tuvākā (*k-nearest*) kaimiņa metodi, vai datu elementu klasterizāciju.

Uz modeļiem balstītas metodes ir vairākas imputācijas metodes, un tās ietver atbilstošu skaitļošanas procedūru veikšanu. Paredzamās vērtības palielināšanas metode dod labus rezultātus, taču tai ir nepieciešami specializēti skaitļošanas rīki.

Pēdējā grupa sastāv no metodēm, kuru pamatā ir mašīnmācīšanās pieejas. Viena no metodēm izmanto mākslīgos neironu tīklus (*ANN*), otra – atbalsta vektoru mašīnas (*SVM*) pieeju. Šīs ir visjaudīgākās mūsdienu metodes iztrūkstošo atribūtu vērtību daudzkārtējai imputācijai, taču to pielietošanai ir nepieciešami sarežģīti skaitļošanas līdzekļi.

2.2. Vienkāršākās imputācijas metodes

1. Imputācija uz izvietojuma parametru pamata

Šīs metodes ideja ir ļoti vienkārša. Ja katra atribūtu vērtību izlase tiek uzskatīta par šo vērtību sadalījumu, tad iztrūkstošo atribūtu vērtību vietā tiek izmantota sadalījuma izvietojuma parametra vērtība.

Izvietojuma parametra veids ir atkarīgs no skalas veida, kurā tiek mērītas atribūtu vērtības (sk. pielikumu P1).

Ja atribūtu vērtības mēra nominālā skalā, tad imputācijai var izmantot tikai atribūtu vērtību sadalījuma veidu.

Pieņem, ka atribūtu vērtības a_j izmērītas nominālā skalā

$$a_j^T = (1,1,2,3,?,1,3,1,2,2,3,1,3,?,1,2).$$

(Šeit tiek izmantots atribūtu vērtību transponētais attēlojums, jo datu tabulā atribūtu vērtības tiek attēlotas šīs tabulas kolonnās).

Skaitļi 1, 2, 3 apzīmē atbilstošo nominālo atribūtu kategoriju kodus.

Iepriekš norādītajā atribūtu vērtību kopā a_j jautājuma zīmes "?" norāda uz šī atribūta iztrūkstošajām vērtībām.

Visbiežāk sastopamā atribūta vērtība tiek ņemta par sadalījuma modu. Šajā piemērā visizplatītākā kategorija ir nominālā kategorija 1, tātad tā ir sadalījuma moda un šī vērtība tiek ielikta iztrūkstošo vērtību vietā (jautājuma zīmju vietā). Rezultātā ir šāds pilns atribūtu vērtību kopums:

$$a_j^T = (1,1,2,3,1,3,1,2,23,1,3,1,1,2).$$

Ja atribūtu vērtības mēra ordinālajā skalā, tad par šo vērtību izlases (sadalījuma) parametru var izmantot modu vai mediānu.

Tiek uzdota atribūtu a_j vērtību kopa, kas mērīta ordinālajā skalā

$$a_j^T = ... \begin{pmatrix} 1.....2.....3.....4 \\ 21(2)...14(1)...16(1)...18(2) \end{pmatrix}.$$

Šeit skaitļi 1, 2, 3, 4 augšējā rindā apzīmē atbilstošo atribūta kategoriju kodus. Apakšējā rindā esošie skaitļi norāda katras kategorijas piemēru skaitu. Skaitļi iekavās apakšējā rindā norāda iztrūkstošo atribūta vērtību skaitu katrā kategorijā.

Dotās izlases (sadalījuma) moda ir ordinālā kategorija 1, kurai pieder lielākais piemēru skaits. Tāpēc šai kategorijai jāpievieno visi piemēri ar iztrūkstošajām atribūtu vērtībām. Rezultātā ir šāds pilns atribūtu vērtību kopums:

$$a_j^T = \begin{pmatrix} 1.....2.....3.....4 \\ 27.....14.....16.....18 \end{pmatrix}.$$

Tā kā piemērā ordinālo kategoriju skaits ir $n = 4$, sadalījuma mediāna būs kategorija, kas atrodas 2. pozīcijā, t. i., 2. kategorija. Rezultātā ir šāda pilna atribūtu vērtību izlase:

$$a_j^T = \begin{pmatrix} 1.....2.....3.....4 \\ 21.....20.....16.....18 \end{pmatrix}.$$

Ja atribūtu vērtības tiek mērītas intervāla vai attiecību skalā, izvietojuma parametra varianti var būt moda, mediāna vai vidējā vērtība.

Uzdodot atribūtu vērtību kopu a_j, kas izmērīta skaitliskajā skalā (intervāla skalā vai attiecību skalā),

$$a_j^T = (10,2,4,?,6,8,?,9,4,7,3,11).$$

Šīs vērtību izlases (sadalījuma) moda ir 4. Aizstājot iztrūkstošās vērtības ar modas vērtību, ir šāds pilns atribūtu vērtību kopums:

$$a_j^T = (10,2,4,4,6,8,4,9,4,7,3,11).$$

Lai noteiktu mediānu, esošās atribūtu vērtības sakārto augošā secībā:

$$a_j^T = (2,3,4,4,6,7,8,9,10,11).$$

Šīs virknes mediāna ir skaitlis pozīcijā $n/2 = 10/2 = 5$, t. i., skaitlis 6. Aizstājot iztrūkstošās vērtības ar mediānas vērtību, iegūst šādu pilnu vērtību kopu:

$$a_j^T = (10,2,4,6,6,8,6,9,4,7,3,11).$$

Atribūta a_j vidējo vērtība ir:

$$\overline{a}_j = \frac{10+2+4+6+8+9+4+7+3+11}{10} = \frac{64}{10} = 6.4.$$

Tā kā atribūtu a_j vērtības ir dotas veselos skaitļos, rezultātu noapaļo: $\overline{a}_j \approx 6$. Aizstājot iztrūkstošās vērtības ar vidējo vērtību, ir šāds pilns atribūtu vērtību kopums:

$$a_j^T = (10,2,4,6,6,8,6,9,4,7,3,11).$$

Aplūkotās iztrūkstošo atribūtu vērtību imputācijas metodes vienīgā priekšrocība ir tās vienkāršība un pilnīga atribūtu vērtību kopuma saglabāšana.

Metodei nav loģiska pamatojuma, aprēķinātās vērtības ir diezgan patvaļīgas. Aprēķinu kvalitāte lielā mērā ir atkarīga no tā, kuras faktiskās vērtības iztrūkst. Ja šīs vērtības ir tuvu izvietojuma parametra vērtībai, tad imputācijas rezultāti var būt veiksmīgi. Ja iztrūkstošās faktiskās vērtības ir tuvu atribūta maksimālajām vai minimālajām vērtībām, imputācijas rezultāti var būt ļoti slikti.

Vispārīgi sakot, imputācijas rezultātu pēc izvietojuma parametra vērtības nosaka atribūta sākotnējo vērtību izkliedes pakāpe un šo vērtību skaits.

Iepriekš aprakstītā metode iztrūkstošo atribūtu vērtību imputēšanai bija viena no pirmajām šāda veida metodēm. Pašlaik šī metode tiek izmantota ārkārtīgi reti.

Ja sākotnējie dati ir apmācības izlase, tad katram datu piemēram tiek piešķirta atbilstošās klases iezīme. Tad iepriekš minēto metodi var piemērot katrai piemēru klasei atsevišķi.

2. Imputācija uz "karstā paneļa metodes" pamata

Metodes pamati tika likti ASV pagājušā gadsimta 40.–50. gados. Statistiskās izpētes nolūkos iedzīvotājiem bieži tika lūgts veikt dažāda veida testus vai atbildēt uz aptaujās dotajiem jautājumiem. Dažādu iemeslu dēļ cilvēki bieži neatbildēja uz dažiem testa jautājumiem vai neaizpildīja dažus obligātos laukus aptaujās. Lai risinātu šāda veida iztrūkstošo datu problēmu, tika izmantota šāda pieeja. Piemēriem testā vai pārskatā ar iztrūkstošām vērtībām tika atlasītas citas piemēru versijas, kurās atlikušie dati atbilst piemēru datiem ar iztrūkstošo vērtību. Tad iztrūkstošā vērtība tika uzskatīta par atbilstošo vērtību no cita piemēra. Šo metodi var attiecināt arī uz gadījumiem, kad iztrūkst vairākas vērtības.

Šīs metodes acīmredzamais trūkums ir tāds, ka iztrūkstošās vērtības imputācijas pamatā ir viens līdzīgs atribūtu vērtību vektors. Visas pārējās datu izlases vispārīgās īpašības tiek ignorētas.

Cits šīs metodes variants ir "aukstā paneļa metode" (*cold deck*), kad atribūtu vērtību vektors, kas ir līdzīgs vektoram ar iztrūkstošo vērtību, tiek meklēts citā datu izlasē, nevis izmantotajā datu izlasē.

"Karstā paneļa metode" praktiski netiek izmantota mūsdienu pieejās datu pirmapstrādei. Tomēr šī metode literatūrā bieži tiek minēta kā pirmā šāda veida metode.

3. Imputācija uz regresijas atkarības pamata

Šo metodi var izmantot kvantitatīvām atribūtu vērtībām. Teorētiski to var attiecināt arī uz kvalitatīvām atribūtu vērtībām. Taču rodas šādas problēmas:

- loģistiskās regresijas modelim (sk. pielikumu P1) ir nepieciešamas sarežģītas skaitļošanas procedūras;
- aprēķinu rezultāti tiek interpretēti kā iztrūkstošās atribūta vērtības, kas pieder pie dažādām kvalitatīvām kategorijām, varbūtības.

Aplūkojamās imputācijas metodes būtība ir šāda: ja atribūtu vērtību izlases a_j, a_l korelē savā starpā, tad var izveidot regresijas attiecības starp šīm vērtību kopām. Pamatojoties uz iegūto atkarību, ir iespējams noteikt otrā atribūta imputētās vērtības.

Vienkāršs, ilustratīvs piemērs: dotas divas atribūtu vērtību izlases:

$$a_j^T = (5,...,4,...,6,...,7,...,6,...,8,...,9,...,3,...,7,...,5,...,7)$$
$$a_l^T = (7,...,6,...,7,...,?,...,13,..,14,.,18,.,5,...,12,..,6,...,12).$$

Izmantojot pielikumā P1.3 norādītās skaitļošanas procedūras vai piemērotus skaitļošanas rīkus, tiek noteikts regresijas vienādojums, kas apvieno atribūtu vērtību izlases a_j, a_l:

$$a_l = -6.8 + 2.8a_j .$$

Tā kā atribūta a_l vērtība iztrūkst pie $a_j = 7$, aprēķina imputēto vērtību

$$a_j = -6.8 + 2.8 * 7 = 12.8 \approx 13 .$$

(aprēķināto atribūta vērtību a_j tika noapaļota, jo visas sākotnējās atribūta vērtības ir vesela skaitļa formā).

Rezultātā ir šādas pilnīgas atribūtu vērtību izlases:

$$a_j^T = (5,...,4,...,6,...,7,...,6,...,8,...,9,...,3,...,7,...,5,...,7)$$
$$a_l^T = (7,...,6,...,7,...,13,...,13,..,14,.,18,.,5,...,12,..,6,...,12).$$

Lietojot apskatīto metodi praksē, vispirms jānovērtē atribūtu vērtību izlases korelācijas (ar iztrūkstošajām vērtībām) ar visām pārējām atribūtu vērtību kopām. Par pieņemamu datu izlasi tiek uzskatīta datu izlase, kas visvairāk korelē ar mūs interesējošo datu izlasi.

Lai gan šī imputācijas metode šķiet loģiska, tās būtiskais trūkums ir netiešais pieņēmums par lineāru saistību starp atribūtu vērtību kopām. Tā kā šis pieņēmums ne vienmēr ir spēkā, imputācija var neizdoties.

2.3. K-tuvāko kaimiņu metode

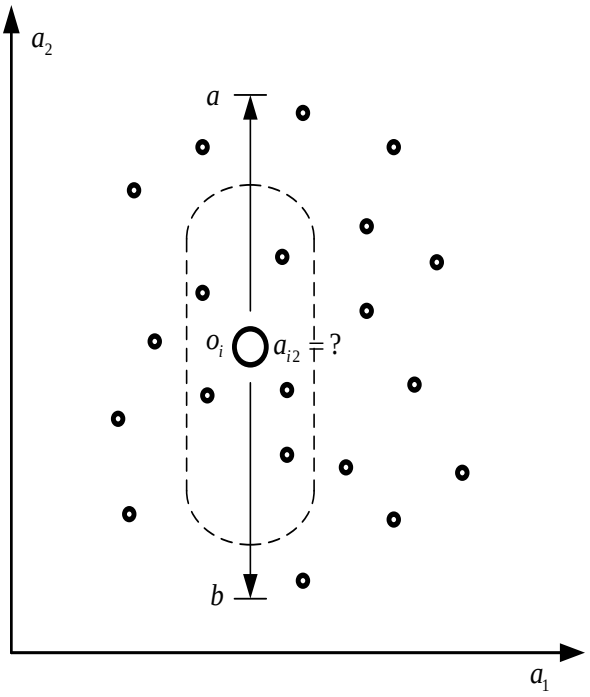
K-tuvāko kaimiņu metode (k-nearest neighbour method) tiek plaši izmantota iztrūkstošo atribūtu vērtību imputēšanai tās vienkāršības, lietderīguma un diezgan augstās rezultātu ticamības dēļ. Metode nav parametriska, jo tā izmanto tikai sākotnējos datus un vairāk neko citu.

Metodes pielietojums būtiski atšķiras kategoriskiem un skaitliskiem datiem. Šīs metodes būtība iztrūkstošo skaitlisko atribūtu vērtību imputācijai parādīta 2.3. attēlā.

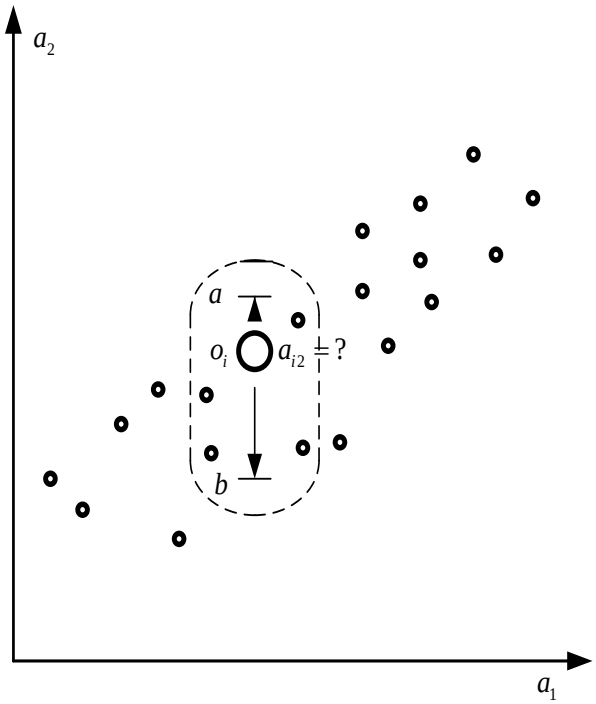
Kāda varētu būt iztrūkstošā atribūta a_2 vērtība piemērā o_i, a_{i2}. Atsaucoties uz 2.3. attēlu, ir viegli pārbaudīt, vai tas var būt robežās $[a,b]$. Lai noteiktu šīs iztrūkstošās vērtības meklēšanas virzienu, par pamatu ņem piemēra tuvākos kaimiņus o_i. Šeit tuvums tiek novērtēts pēc tuvāko kaimiņu atribūtu vērtībām a_1. Pieņemot, ka ir k tuvākie kaimiņi, tāpēc arī ir tāds metodes nosaukums, var aprēķināt vidējo atribūta vērtību a_2 šiem kaimiņiem un ņemt to par iztrūkstošo vērtību a_{i2}.

Sprieduma loģika šeit ir ļoti vienkārša. Ja piemērs ar iztrūkstošu atribūta vērtību a_2 ir tuvs kaimiņiem pēc atribūta vērtības a_1, tiek pieņemts, ka tā iztrūkstošā vērtība ir tuva tā tuvāko kaimiņu vidējai vērtībai a_2.

Iztrūkstošās atribūta vērtības imputācijas rezultātu, izmantojot k-tuvāko kaimiņu metodi, var būtiski ietekmēt atribūtu vērtību a_1 un a_2 korelācija. No 2.3. attēla ir skaidrs, ka atribūtu vērtības šajā gadījumā nav savstarpēji saistītas (nekorelē savā starpā).



2.3. attēls. **Datu izlases shematisks attēlojums divu atribūtu a_1, a_2 telpā** (piemērā o_i nav atribūta vērtības a_{i2})



2.4. attēls. **Piemēra o_i shematisks attēlojums ar iztrūkstošo atribūta a_{i2} vērtību datu kopā ar korelētu atribūtu a_1, a_2 vērtībām**

No 2.4. attēla var izdarīt nenoliedzamu secinājumu, ka atribūtu vērtības a_1, a_2 ir pozitīvi korelētas. Tas sašaurina piemēra o_i tuvākos kaimiņu loku. Tāpēc, izmantojot piemēra o_i tuvāko kaimiņu izlases vidējo atribūta vērtību a_2 kā imputācijas vērtību a_{i2}, tiks iegūts ticamāks imputācijas rezultāts.

Galvenie uzdevumi, kas ir jāatrisina tuvāko k-kaimiņu metodes praktiskajā pielietošanā, ir:

- vērtības k izvēle;
- sākotnējo atribūtu vērtību normalizēšana;
- metrikas izvēle attālumu noteikšanai no piemēra ar iztrūkstošu atribūta vērtību līdz tuvākajiem kaimiņiem.

Tuvāko kaimiņu skaita izvēle ir neformalizējama problēma. Šī izvēle ir atkarīga no risināmās problēmas rakstura un ievades datu apjoma. Ļoti liela ievades datu skaita gadījumā to samazināšanai var izmantot dažādas pieejas: ierobežotu datu apgabalu definēšana, piemēru ignorēšana, kas potenciāli var nebūt potenciālie kaimiņi un tamlīdzīgi.

Lai noteiktu atbilstošo tuvāko kaimiņu skaitu, daži literatūras avoti iesaka veikt nepieciešamo imputāciju dažādām k vērtībām un par pamatu ņemt rezultātu, kas dod mazāko novērtējuma kļūdu. Tomēr šī pieeja prasa lielu skaitu aprēķinu, tāpēc to var ieteikt tikai gadījumos, kad iztrūkstošo vērtību skaits ir mazs, bet ir nepieciešama augsta imputācijas ticamība.

Atribūtu skaitlisko vērtību normalizācijas mērķis ir samazināt dažādu atribūtu vērtību amplitūdu ietekmi un nodrošināt attāluma aprēķinu pareizību, jo visu atribūtu vērtības pēc normalizācijas tiks uzrādītas vienotā veidā. Atribūtu vērtību normalizēšanas metodes detalizēti tiks aplūkotas 5. nodaļā. Šajā nodaļā kā normalizējošie rādītāji tiek izmantotas vidējās atribūtu vērtības.

Eiklīda attālumu (*Euclidean distance*) parasti izmanto, lai novērtētu attālumus starp piemēru ar iztrūkstošo atribūta vērtību un citiem piemēriem:

$$d\left(o_i, o_k\right) = \sqrt{\sum_j \left(a_{ij} - a_{kj}\right)^2}\,, \tag{2.6}$$

kur $j = 1, 2, \ldots$ – atribūtu skaits, pēc kura tiek novērtēts attālums;

$a_{ij} - j$ atribūta vērtība piemērā ar iztrūkstošu vērtību (atribūts ar iztrūkstošu vērtību nav iekļauts atribūtu sarakstā attiecīgo attālumu noteikšanai);

$a_{kj} - j$ atribūta vērtība k piemēram.

Manhetenas attālumu (*Manhattan distance*) var izmantot kā alternatīvu attāluma mēru:

$$d\left(o_i - o_k\right) = \sum_j \left|a_{ij} - a_{kj}\right|. \tag{2.7}$$

Tomēr šis attāluma novērtējums tiek izmantots daudz retāk. 2.3. tabulā ir parādīta sākotnējā datu izlase.

2.3. tabula

Sākotnējā datu izlase

Piemēri (objekti)	Atribūti		
	a_1	a_2	a_3
o_1	7	9	12
o_2	6	8	11
o_3	5	8	14
o_4	8	10	13
o_5	6	8	15
o_6	? (5)	7	14
o_7	7	10	? (12)
o_8	5	8	11
o_9	8	10	12
o_{10}	6	7	14
Kolonnu summas	58	85	116
Vidējā vērtība	6.444	8500	12.889

Šajos datos piemēram o_6 iztrūkst atribūta vērtības a_1 un piemēram o_7 iztrūkst atribūta a_3 vērtības. Iztrūkstošās vērtības tiek apzīmētas ar jautājuma zīmi "?". Šo atribūtu faktiskās vērtības ir parādītas blakus iekavās. Tas tika darīts, lai salīdzinātu un analizētu imputācijas rezultātus (ir jāsaprot, ka šīs vērtības faktiski nav zināmas).

Sākotnējās atribūtu vērtības tiek normalizētas, izmantojot vidējās atribūtu vērtības, kas norādītas 2.3. tabulas pēdējā rindā. Atribūtu normalizētās vērtības dotas 2.4. tabulā.

Eiklīda attālumus no piemēra o_6 ar iztrūkstošu atribūta vērtību a_{61}^n līdz citiem piemēriem aprēķina, izmantojot atribūtu vērtības a_2^n, a_3^n:

$$d\left(o_6 - o_1\right) = \sqrt{\left(a_{62}^n - a_{12}^n\right)^2 + \left(a_{63}^n - a_{13}^n\right)^2} = \sqrt{\left(0.082 - 0.106\right)^2 + \left(0.121 - 0.103\right)^2} =$$
$$= \sqrt{0.00058 + 0.00032} = \sqrt{0.00090} = 0.0300 .$$

2.4. tabula

Normalizētās atribūtu vērtības datiem no 2.3. tabulas

Piemēri (objekti)	Normalizētās atribūtu vērtības		
	a_1^n	a_2^n	a_3^n
o_1	1.086	1.059	0.931
o_2	0.931	0.941	0.853
o_3	0.776	0.941	1.086
o_4	1.241	1.176	1.009
o_5	0.776	0.941	1.164
o_6	?	0.824	1.086
o_7	1.086	1.176	?
o_8	0.776	0.941	0.853
o_9	1.241	1.176	0.931
o_{10}	0.931	0.824	1.086

Pārējie aprēķini tiek veikti pēc analoģijas. Aprēķinu rezultāti ir parādīti 2.5. tabulā.

2.5. tabula

Aprēķinātās attālumu vērtības no piemēra o_6 līdz citiem piemēriem, pamatojoties uz atribūtu vērtībām o_2^n, o_3^n

	$o_6 - o_1$	$o_6 - o_2$	$o_6 - o_3$	$o_6 - o_4$	$o_6 - o_5$	$o_6 - o_7$	$o_6 - o_8$	$o_6 - o_9$	$o_6 - o_{10}$
$d\left(\cdot - \cdot\right)$	0.2814	0.2608	0.1170	0.3603	0.1407	-	0.2608	0.3544	0

Attālums $d\left(o_6 - o_7\right)$ netiek aprēķināts, jo piemērā o_7 nav atribūta vērtības a_3. Piemēri tiek sakārtoti augošā secībā pēc to attāluma līdz piemēram o_6:

$$o_{10},\ o_3,\ o_5,\ o_2,\ o_8,\ o_1,\ o_9,\ o_4 .$$

Pieņemot, ka $k = 3$, imputētā atribūta vērtība $a_{61(i)}$ ir vienāda ar atribūtu $o_{10,1}$, o_{31}, o_{51} vidējo vērtību:

$$a_{61(i)} = \frac{a_{10,1} + a_{31} + a_{51}}{3} = \frac{6 + 5 + 6}{3} = 5.67 \approx 6 .$$

Ja $k = 4$, tad:

$$a_{61(i)} = \frac{a_{10,1} + a_{31} + a_{51} + a_{21}}{4} = \frac{6+5+6+6}{4} = 5.75 \approx 6 .$$

Abos gadījumos tika iegūti labi imputācijas rezultāti. Aprēķinātā vērtība 6 būtiski neatšķiras no faktiskās vērtības 5. Tā kā reālā vērtība a_{61} nav zināma, tās aprēķinātā vērtība sniedz labu reālās vērtības tuvinājumu.

Tālāk tiek noteikta atribūta imputētā vērtība a_{73}. Eiklīda attālumi no piemēra o_7 līdz citiem piemēriem tiek aprēķināti, izmantojot atribūtu a_1^n un a_2^n vērtības. Aprēķinu rezultāti ir parādīti 2.6. tabulā.

2.6. tabula

Aprēķinātās attālumu vērtības no piemēra o_7 līdz citiem piemēriem, pamatojoties uz atribūtu a_1^n, a_2^n vērtībām

	$o_7 - o_1$	$o_7 - o_2$	$o_7 - o_3$	$o_7 - o_4$	$o_7 - o_5$	$o_7 - o_6$	$o_7 - o_8$	$o_7 - o_9$	$o_7 - o_{10}$
$d(\cdot - \cdot)$	0.1170	0.2793	0.3895	0.1550	0.3390	-	0.3895	0.1550	0.3846

Attālums $d(o_7 - o_6)$ netika aprēķināts. jo piemērā o_6 trūkst atribūta vērtības a_1.

Piemēri tiek sakārtoti augošā secībā pēc to attāluma no piemēra o_7:

$$o_1,\ o_4,\ o_9,\ o_2,\ o_5,\ o_{10},\ o_3,\ o_8 .$$

Pieņem, ka $k = 3$:

$$a_{73(i)} = \frac{a_{13} + a_{43} + a_{93}}{3} = \frac{12+13+12}{3} = 12.33 \approx 12 .$$

Pieņem, ka $k = 4$:

$$a_{73(i)} = \frac{a_{13} + a_{43} + a_{93} + a_{23}}{4} = \frac{12+13+12+11}{3} = 12 .$$

Abām k vērtībām tika iegūti ļoti labi imputācijas rezultāti.

Iepriekš minētajā piemērā visiem attālumiem no piemēriem ar trūkstošām atribūtu vērtībām tika piešķirta vienāda nozīme. Šajā sakarā rodas ideja, ka attālumiem piešķirt dažādus svarus pēc principa – jo mazāks attālums, jo lielāks tā svars.

Attiecīgo attālumu svēršanai ir ierosināts ļoti liels skaits dažādu pieeju. Šajā nodaļā tiks aplūkota vienkārša un efektīva pieeja šādā veidā – svara vērtība noteiktam attālumam tiek definēta kā šī attāluma kvadrāta apgrieztā vērtība:

$$w(\cdot - \cdot) = \frac{1}{\left(d(\cdot - \cdot)\right)^2} . \tag{2.8}$$

Tad svērtā attāluma vērtību aprēķina kā:

$$d^w(\cdot - \cdot) = w(\cdot - \cdot)\, d(\cdot - \cdot) . \tag{2.9}$$

Aprēķina svaru un svērto attālumu vērtības 2.5. tabulā aprēķinātajām vērtībām. Aprēķinu rezultāti ir parādīti 2.7. tabulā.

Tā kā ir darīšana ar svērtajiem attālumiem, lielāka svērtā attāluma vērtība atbilst lielākai tuvības pakāpei starp piemēriem.

Piemēri tiek sarindoti svērtā attāluma palielināšanās attālumā no piemēra o_6:

$$o_{10},\ o_3,\ o_5,\ o_2,\ o_8,\ o_1,\ o_9,\ o_4 .$$

2.7. tabula

Aprēķinātās svaru vērtības un svērtie attālumi no piemēra o_6 līdz citiem piemēriem saskaņā ar 2.5. tabulu

	$o_6 - o_1$	$o_6 - o_2$	$o_6 - o_3$	$o_6 - o_4$	$o_6 - o_5$	$o_6 - o_7$	$o_6 - o_8$	$o_6 - o_9$	$o_6 - o_{10}$
$d(\cdot - \cdot)$	0.2814	0.2608	0.1170	0.3603	0.1407	-	0.2608	0.3544	0
$(d(\cdot - \cdot))^2$	0.0792	0.0680	0.0137	0.1298	0.0198	-	0.0680	0.1256	0
$w(\cdot - \cdot)$	12.6263	14.7016	73.0460	7.7023	50.5050	-	14.7016	7.9618	0
$d^w(\cdot - \cdot)$	3.5530	3.8342	8.5464	2.2775	7/1060	-	3.8342	2.8216	0

Tika iegūta tādu pašu piemēru secība, ja izmantotu nesvērtos attālumus, tāpēc tuvāko k kaimiņu izvēle un aprēķinātās imputētās atribūtu vērtības $o_{61(i)}$ būs vienādas.

Svērto attālumu aprēķins 2.6. tabulā tiek veikts saskaņā ar to pašu shēmu. Kopumā var būt nelielas atšķirības starp piemēru apstrādi pēc tiešā attāluma vērtībām un svērtajām attāluma vērtībām. Bet šīs iespējamās atšķirības būtiski neietekmē k tuvāko kaimiņu izvēli.

Iztrūkstošo atribūtu vērtību imputācijas vispārīgie principi pie dažādām k vērtībām, pamatojoties uz k tuvāko kaimiņu metodi, ir šādi:

- mazas k vērtības palielina "trokšņa" ietekmi datos un rada zemu rezultātu vispārināmību;
- lielākas k vērtības "nogludina" lokālos efektus un rada lielāku rezultātu vispārināmību.

Jāņem vērā: ja vienam un tam pašam atribūtam iztrūkst vairākas vērtības, attiecīgie aprēķini jāveic katrai iztrūkstošajai vērtībai atsevišķi.

Lieliem ievades datu apjomiem visi attiecīgie aprēķini tiek veikti, izmantojot piemērotus programmatūras rīkus. Lai formalizētu skaitļošanas procesus, bieži izmanto fiktīvu mainīgo (*dummy variable*). Šī mainīgā lieluma vērtības ir parādītas tāda paša izmēra tabulā (matricā) kā sākotnējo datu tabula. Ja datu matricas šūnā ij ir ierakstīta atribūta vērtība a_{ij}, tad fiktīvā mainīgā vērtībām attiecīgajā matricas šūnā tiek ierakstīta vērtība 1. Ja atribūta vērtības nav, tad fiktīvajam mainīgajam ieraksta vērtību 0. Piemēram, šī matrica attēlo fiktīvo mainīgo vērtības D sākotnējo datu 2.4. tabulai:

$$D = \begin{pmatrix} 1...1...1 \\ 1...1...1 \\ 1...1...1 \\ 1...1...1 \\ 1...1...1 \\ 0...1...1 \\ 1...1...0 \\ 1...1...1 \\ 1...1...1 \\ 1...1...1 \end{pmatrix}.$$

Pēc tam tiek veikts attiecīgo attālumu aprēķins gadījumiem, kad matricas D rindās ir tikai vērtības 1 to atribūtu pozīcijās, kas ir iekļauti attāluma aprēķinos.

Dažkārt šis noteikums tiek formalizēts, ieviešot tā saucamo indikatora funkciju, kuras vērtības ir vienādas ar 1 pieejamām atribūtu vērtībām un vienādas ar 0 iztrūkstošajām atribūtu vērtībām.

Tādā gadījumā tiek aprēķināts attālums starp piemēra o_i ar iztrūkstošu atribūta vērtību un piemēru o_k, fiktīvo mainīgo vērtību matricā tiek pārbaudītas paraugam atbilstošās šūnas o_k. Ja visās šajās šūnās ir 1 (t. i., indikatora funkcijas vērtības ir vienādas ar 1), tad tiek aprēķināts attiecīgais

attālums. Ja vismaz vienā no šīm šūnām ir 0 (indikatora funkcijas vērtība ir 0), attiecīgais attālums netiek aprēķināts.

Ja atribūtu vērtības ir norādītas kategoriskā (nominālā vai ordinārā) skalā, iztrūkstošo atribūtu vērtību imputācijas process kļūst ievērojami sarežģītāks. Tas ir saistīts ar tuvības mēra starp piemēriem noteikšanas problēmu. Skaitlisku atribūtu vērtību gadījumā attāluma aprēķins starp diviem piemēriem ir balstīts uz katra atribūta vērtību atšķirību summu šiem piemēriem. Gadījumā, ja atribūtu vērtības mēra nominālā vai ordinārā skalā, atribūtu vērtību skaitliskās atšķirības jēdzienam nav jēgas. Tāpēc ir nepieciešami citi tuvības mērījumu (attālumu) aprēķini starp attiecīgajiem piemēriem.

Diemžēl ir maz zinātnisko darbu *k*-tuvāko kaimiņu metodes izmantošanas kategorisko atribūtu novērtējumos. Darbā [Jönsson P., Volins C., 2004] piedāvāta metode iztrūkstošo atribūtu vērtību imputācijai gadījumam, kad visu atribūtu vērtības mēra Likerta skalā. Attālumi starp piemēriem tiek aprēķināti, apstrādājot Likerta skalas punktus kā reālus skaitļus. Šī pieeja šķiet nepamatota, jo šādas darbības ar Likerta skalas punktiem nav korekta.

Darbā [Schwenger H., Icstad K., 2008] tiek piedāvāta korektāka pieeja attālumu novērtēšanai starp piemēriem, pamatojoties uz kategoriskām atribūtu vērtībām. Diemžēl šī pieeja ir piemērojama tikai gadījumos, kad visiem atribūtiem ir vienāds vērtēšanas kategoriju skaits.

Darbs [Tutz G., Ramzan Sh., 2014] ir saistīts ar jaunu piemēru klasifikāciju, novērtējot attālumus starp šo piemēru atribūtu kategoriskajiem novērtējumiem un dotajām klasēm piederošo piemēru atribūtu kategoriskajiem novērtējumiem. Piedāvātā attālumu novērtēšanas metode tika modificēta un izmantota, lai imputētu iztrūkstošās atribūtu vērtības [Faisal Sh., Tutz G., 2016].

Šajā sadaļā tiek piedāvāta vienkāršākā tiešā metode, lai novērtētu attālumus starp šajā darbā piedāvātajiem attiecīgajiem piemēriem.

Pieņem, ka dati uzdoti taisnstūra izmēra tabulas veidā $m \times n$, kurā katra rinda atbilst piemēram (objekts, indivīds) o_i, $i = 1,...,m$ un katra kolonna atbilst atribūtam a_j, $j = 1,...,n$. Katra atribūta a_j vērtība ir novērtēta ar nominālo vai ordinālo kategoriju r_j, $j = 1,...,r_j$. Atribūta a_{ij} vērtība tabulas i rindas un j kolonas krustojumā tiek parādīta kategoriskā formā kā $p_{ij}^{r_j}$. Tad attālumu starp piemēriem o_i un o_l var novērtēt ar vienkārša atbilstības koeficienta (*simple matching coefficient*) palīdzību:

$$d\left(o_i - o_l\right) = \sum_{j=1}^{n} I\left(p_{ij}^{r_j} - p_{lj}^{r_j}\right), \tag{2.10}$$

kur $p_{ij}^{r_j}$ – piemēra o_i atribūtu kategorijas vērtība a_j;

$p_{lj}^{r_j}$ – piemēra o_l atribūtu kategorijas vērtība a_j;

I – indikatora funkcijas vērtība, kas definēta šādā veidā:

$I = 0$, ja $p_{ij}^{r_j} = p_{lj}^{r_j}$;

$I = 1$, ja $p_{ij}^{r_j} \neq p_{lj}^{r_j}$. (2.11)

Citiem vārdiem sakot, par attāluma starp piemēriem o_i un o_l novērtējumu tiek ņemta nuļļu un vieninieku summa, kas noteikta ar vienādojumiem (2.10), (2.11) .

Ja piemērā o_i atribūtam a_{ij} ir iztrūkstošas atribūta vērtības (kategorijas). Tad, izmantojot vienādojumus (2.10), (2.11), var aprēķināt attālumus no šī piemēra līdz visiem pārējiem piemēriem, izmantojot visu atribūtu vērtības, izņemot atribūtu ar iztrūkstošo vērtību.

Pamatojoties uz iegūtajiem attāluma aprēķiniem, tiek atlasīti piemēri, kuriem ir vismazākie attālumi līdz piemēram ar iztrūkstošo vērtību. Vienkārša piemēru sakārtošana kategorisko atribūtu novērtējumos vairumā gadījumu nav iespējama, kā tas tiks parādīts nākamajā piemērā. Tāpēc arī bieži vien nav iespējams iestatīt vēlamo tuvāko kaimiņu vērtību k. Rezultātā tuvākie kaimiņi tiek izvēlēti pēc principa - izmantot visus iespējamos.

Par imputēto vērtību tiek pieņemta atribūta vērtību mediāna ar iztrūkstošo šī atribūta vērtību, salīdzinājumā ar tuvāko kaimiņu kopumu.

Ja atlasīto atribūtu vērtību izlasei ir vairāk nekā viena mediāna (t. i., ir vienāds atribūta biežāko kategoriju skaits), tad imputētās vērtības izvēle starp tikpat bieži sastopamajām vērtībām tiek veikta nejauši vai pamatojoties uz kādu papildu nosacījumu.

Ilustratīvs piemērs: ir pieejams daudz datu ar 10 piemēriem (personas). Šīs personas raksturo trīs atribūtu vērtības: a_1 – dzimums, a_2 – acu krāsa, a_3 – matu krāsa. Atribūtam a_1 ir divas nominālās kategorijas: p_1^1 – vīrietis, p_1^2 – sieviete. Atribūtam a_2 ir trīs nominālās kategorijas: p_2^1 – brūnas acis, p_2^2 – pelēkas acis, p_2^3 – zilas acis. Atribūtam a_3 ir trīs nominālās kategorijas: p_3^1 – blondīne/blondīns, p_3^2 – brūni mati, p_3^3 – brunete/brunets.

Sākotnējie dati ir parādīti 2.8. tabulā.

2.8. tabula

Sākotnējā datu izlase

Piemēri	Atribūtu vērtību kategorijas		
	$(a_1)\quad p_1^{r_1}$	$(a_2)\quad p_2^{r_2}$	$(a_2)\quad p_3^{r_3}$
o_1	p_1^1	p_2^2	p_3^3
o_2	p_1^2	p_2^2	p_3^1
o_3	p_1^2	p_2^1	p_3^1
o_4	p_1^1	?	p_3^3
o_5	p_1^2	p_2^3	p_3^2
o_6	p_1^1	p_2^1	p_3^3
o_7	p_1^2	p_2^2	p_3^1
o_8	p_1^1	p_2^1	p_3^1
o_9	p_1^2	p_2^1	p_3^2
o_{10}	p_1^1	p_2^1	?

Šajā tabulā ar "?" apzīmētas iztrūkstošās atribūtu vērtības a_2 un a_3.

Izmantojot vienādojumus (2.10), (2.11), aprēķina attālumus no piemēra o_4 līdz citiem piemēriem, pamatojoties uz atribūtu vērtībām a_1, a_3:

$$d(o_4 - o_1) = 0 + 0 = 0;$$
$$d(o_4 - o_2) = 1 + 1 = 2;$$
$$d(o_4 - o_3) = 1 + 1 = 2;$$
$$d(o_4 - o_5) = 1 + 1 = 2;$$
$$d(o_4 - o_6) = 0 + 0 = 0;$$
$$d(o_4 - o_7) = 1 + 1 = 2;$$
$$d(o_4 - o_8) = 0 + 1 = 1;$$
$$d(o_4 - o_9) = 1 + 1 = 2.$$

Attālums $d(o_4 - o_{10})$ netiek aprēķināts, jo piemērā o_{10} nav atribūta vērtības $a_{10,3}$.

No piemēra o_4 izdala piemērus ar mazākajiem attālumiem Tie ir piemēri o_1, o_6, o_8, kuros ir šādas atribūtu a_2 kategorijas : p_2^2 p_2^1, p_2^2. Šīs vērtību izlases mediāna atbilst kategorijai p_2^2. Tāpēc

šī kategorija ir jāuzskata par iztrūkstošo atribūta vērtību a_{42}. Citiem vārdiem sakot, personai o_4 ir jābūt pelēkām acīm.

Veic aprēķinus, lai atrastu iztrūkstošo atribūta vērtību $a_{10,3}$:

$$d(o_{10} - o_1) = 1+1 = 2;$$
$$d(o_{10} - o_2) = 0+1 = 1;$$
$$d(o_{10} - o_3) = 0+1 = 1.$$

Attālums $d(o_{10} - o_4)$ netiek aprēķināts, jo piemērā o_4 nav atribūta vērtības a_{42}.

$$d(o_{10} - o_5) = 0+1 = 1;$$
$$d(o_{10} - o_6) = 1+0 = 1;$$
$$d(o_{10} - o_7) = 0+1 = 1;$$
$$d(o_{10} - o_8) = 1+0 = 1;$$
$$d(o_{10} - o_9) = 0+1 = 1.$$

Šajā piemērā rodas problēma ar k-tuvāko kaimiņu izvēli. Tā kā lielākā daļa attālumu ir vienādi ar 1, šīs vērtības nevar sarindot un pamatoti noteikt tuvāko kaimiņu skaitu k. Tāpēc visus piemērus, kuriem ir attālumi līdz piemēram o_{10}, uzskata par piemēra o_{10} tuvākajiem kaimiņiem. Šajos piemēros ir šādas atribūta a_3 vērtības: p_3^1, p_3^1, p_3^2, p_3^3, p_3^1, p_3^1, p_3^2. Šo vērtību izlases mediāna ir kategorija p_3^1, tātad šī kategorija tiek uzskatīta par iztrūkstošā atribūta $a_{10,3}$ vērtību. Citiem vārdiem sakot, sievietei, kas avota datu tabulā norādīta kā o_{10}, jābūt blondīnei.

Darbā [Tutz G., Ramzan Sh., 2014] tiek piedāvāta alternatīva pieeja attālumu noteikšanai starp piemēriem, kas tiek novērtēti pēc kategorisku atribūtu vērtībām. Šī pieeja izmanto sākotnējo datu tabulu un īpašu atribūtu vērtējuma kategoriju – bināro kodējumu. Šajā nodaļā šī pieeja netiek apskatīta. Interesenti var iepazīties ar šo pieeju norādītajā darbā.

2.4. Imputācija uz sagaidāmā rezultāta maksimizācijas pamata

Šīs nodaļas iepriekšējās sadaļās apskatītās metodes iztrūkstošo atribūtu vērtību imputēšanai neņēma vērā paņēmienus, kādā veidā datos parādās iztrūkstošie dati (sk. 2.1. sadaļu). Šajā sadaļā aprakstītā metode ņem vērā iztrūkstošos datu veidošanās paņēmienus: *MAR* un *MCAR*. *NMAR* gadījumā šī metode nav piemērojama.

Ir daudzas praktiskas pieejas, lai noteiktu faktisko iztrūkstošo datu veidošanās paņēmienu. Lielākā daļa statistikas programmatūras pakotņu, tostarp *IBM SPSS Statistica*, izmanto sagaidāmā rezultāta maksimizēšanas metodi, lai pēc noklusējuma noteiktu iztrūkstošo datu veidošanās paņēmienus.

Jāņem vērā, ka turpmāk aprakstītā metode ir piemērojama tikai skaitliskiem datiem, kuros atribūtu vērtības tiek mērītas intervālu vai attiecību skalās. Lai gan ir bijuši mēģinājumi šo metodi attiecināt arī uz jauktu datu gadījumiem (kvantitatīvās un kategoriskās atribūtu vērtības), tomēr šie mēģinājumi nav guvuši plašu praktisku pielietojumu.

Sagaidāmā rezultāta maksimizēšanas princips ir formulēts dažādos veidos un izmantots dažādos pētījumos kopš 20. gadsimta 40. gadiem. Tikai darbā [Dempster A. et al., 1977] tika sniegts stingrs matemātiskais pamatojums iztrūkstošo atribūtu vērtību imputācijai un tika pierādīta rezultātu iegūšanas konverģence. Tāpat šīs metodes pielietojamība imputācijas problēmām ir izklāstīta darbā [Little P.J.A., Rubin D., 1987]. Pēdējās desmitgadēs sagaidāmās vērtības maksimizēšanas metode ir radusi plašu pielietojumu dažādās datu iegūšanas un analīzes jomās.

Sagaidāmā rezultāta vērtības maksimizēšanas metode ir balstīta uz maksimālās līdzības principa. Kāda ir šī principa būtība? Lai to izskaidrotu, var izmantot vienkāršu piemēru – metot

monētu, ir jānovērtē ģerboņa izkrišanas varbūtība. Monētas priekšpuse tiek saukta par "aversu", mugurpuse - par "reversu", bet mala par "jostu". Monētas bieži mēdz izmantot kā vienkāršotu metamo kauliņu. Monētas aversu monētas mešanā mēdz dēvēt par "ciparu", savukārt reversu – par "ērgli" vai "ģerboni".

Lai novērtētu monētas izkrišanas varbūtību, tiek veikta monētas mešanas sērija. Pieņem, ka ir bijuši 100 metieni, ģerbonis izkritis 60 reizes un cipars – 40. Ģerboņa izkrišanas varbūtību apzīmē ar θ: p (ģerbonis) = θ, tad p (cipars) = $1-\theta$. Kā novērtēt vērtību θ? Ir skaidrs, ka

$$\hat{\theta} = \frac{h}{n} = \frac{\text{Ģerboņa izkrišanas reižu skaits}}{\text{Kopējais metienu skaits}} = \frac{60}{100} = 0.60.$$

Formāli vērtības θ sadalījums ir saskaņā ar binomiālo likumu:

$$p(\theta) = \binom{h}{n} \theta^h (1-\theta)^{n-h}.$$

Šajā vienkāršajā piemērā parametra $\hat{\theta}$ novērtējums ir acīmredzams un neapšaubāms.

Pieņem, ka ir gadījumu skaitļu vērtību izlase, kuru sadalījums ir atbilstošs normālajam sadalījumam. Šajā gadījumā var viegli aprēķināt maksimālās līdzības novērtējumu vidējai vērtībai μ un dispersijai σ^2.

Reālajā dzīvē viss nav tik vienkārši, jo var būt tādu normālu sadalījumu izlases, kurām ir jānovērtē daudz parametru. Iespējams, ka nejaušā lieluma vērtību izlasei ir noteikts sadalījums, tāpēc ir grūti tieši novērtēt tā parametrus.

Lai sniegtu priekšstatu par šāda veida problēmām, var apskatīt šādu piemēru [Do Ch.B., Batzoglou S., 2008]. Pieņem, ka ir divas monētas A un B un tika veiktas trīs monētas A mešanas sērijas un divas monētas B mešanas sērijas. Katra sērija sastāvēja no 10 attiecīgās monētas metieniem. Katrā metienu sērijā tika uzskaitīts izkritušo ciparu un ģerboņu skaits. Aprēķinu rezultāti ir parādīti 2.5.a attēlā). Pamatojoties uz iegūtajiem rezultātiem, tika aprēķinātas varbūtības izkrist ģerbonim katras monētas mešanas laikā: $\hat{\theta}_A = 0.80$, $\hat{\theta}_B = 0.45$. Šīs aplēses tiek iegūtas, pamatojoties uz maksimālās līdzības principu.

Mainot nosacījumus uzdevuma risināšanai, metienu sērijas tiek izpildītas tieši tādā pašā veidā un tiek iegūti tie paši metienu rezultāti. Taču tagad *nav zināms,* kura monēta tika izmantota katrā metienu sērijā. Tāpat kā iepriekš, uzdevums ir novērtēt ģerboņa izkrišanas varbūtību, nejauši metot monētas A un B.

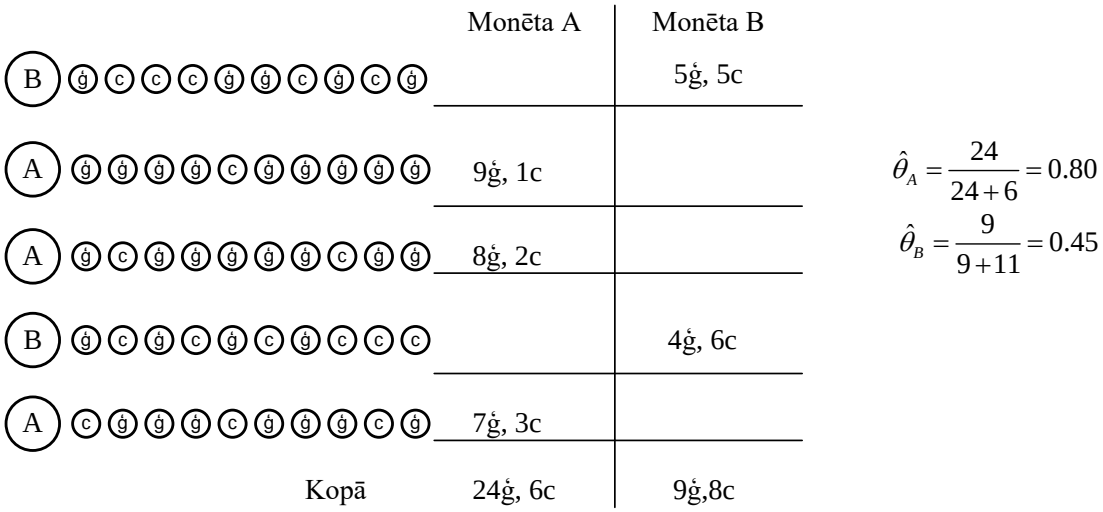
Ar ko šis eksperiments atšķiras no iepriekšējā eksperimenta? Galvenā atšķirība ir tā, ka eksperiments tagad tiek veikts *nepilnīgu datu apstākļos* – nav zināms, kura monēta tiek izmantota katrā metienu sērijā. Acīmredzot tiešs veids, kā aprēķināt interesējošās varbūtības, tagad nav iespējams. Ir nepieciešams cits variants, kā atrisināt šo problēmu.

Lai to izdarītu, ievieš jaunu mainīgo lielumu $z = (z_1, z_2)$, kur z_1 ir varbūtība, ka monēta A tiek izmantota konkrētajā metienu sērijā, un z_2 – varbūtība, ka monēta B tiek izmantota konkrētajā metienu sērijā. Šāda veida mainīgo sauc par *latento mainīgo*.

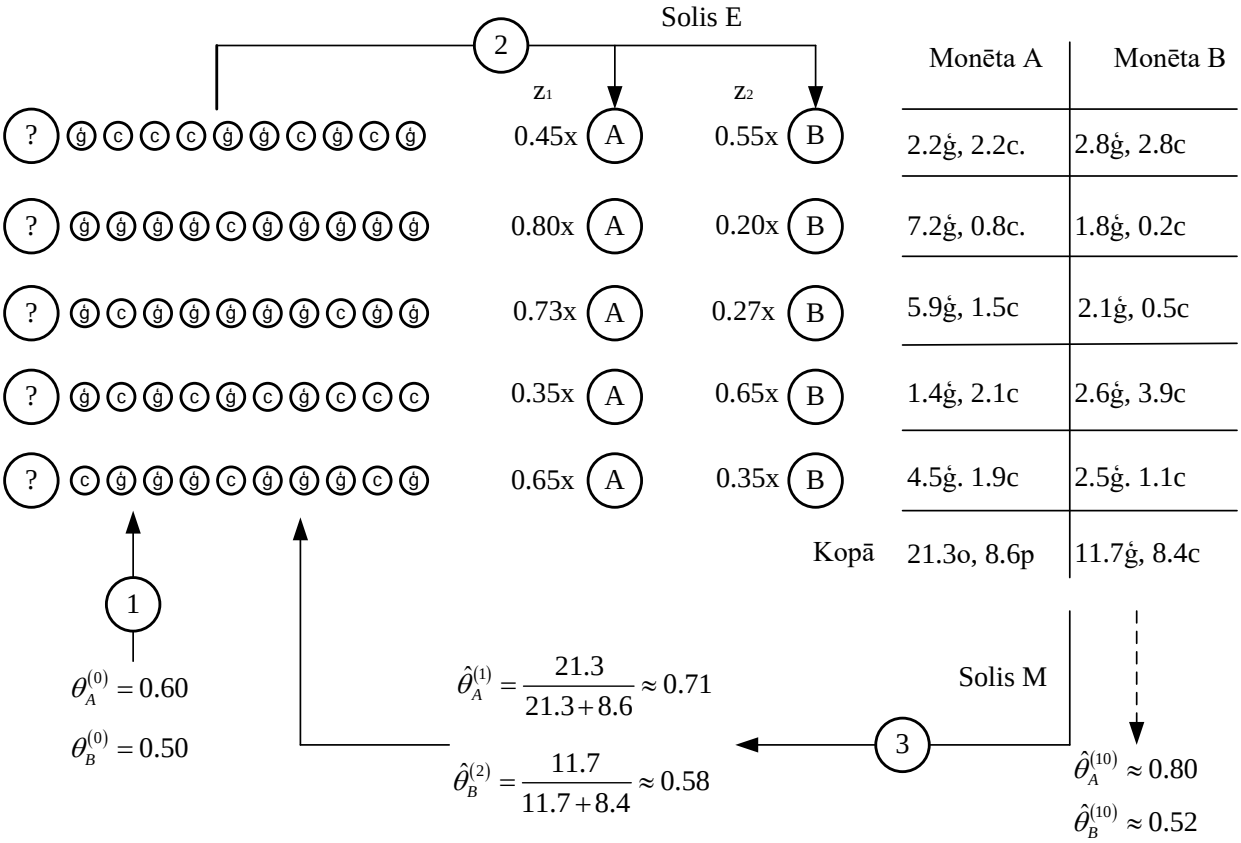
Ņemot vērā sākotnējos nosacījumus, problēmu var atrisināt, izmantojot sagaidāmās vērtības maksimizācijas metodi. Aprēķinu procedūru diagramma ir parādīta 2.5.b attēlā.

Aplūkojamās problēmas kontekstā runā par sagaidāmo parametru aplēšu (varbūtību $\hat{\theta}_A$, $\hat{\theta}_B$) maksimizāciju, pamatojoties uz visu pieejamo informāciju. Lai inicializētu algoritmu, ir jāuzdod interesējošo varbūtību sākotnējās vērtības $\hat{\theta}_A^0$, $\hat{\theta}_B^0$. Šīs vērtības var izvēlēties patvaļīgi vai pamatojoties uz dažiem papildu apsvērumiem. Šajā piemērā ir norādītas šādas vērtības: $\hat{\theta}_A^0 = 0.60$, $\hat{\theta}_B^0 = 0.50$. Pēc tam tiek veikts iepriekš aprakstītais eksperiments un tiek iegūti tādi rezultāti, kādi parādīti 2.5.a attēlā, bet tagad šos rezultātus var attiecināt gan uz monētu A, gan uz monētu B.

a) maksimālā līdzība



b) sagaidāmās vērtības maksimizācija



2.5. attēls. **Shēmas, lai iegūtu aplēses par ģerboņa un cipara izkrišanas varbūtību, metot monētas A un B,**
a) pamatojoties uz maksimālās līdzības principa;
b) pamatojoties uz sagaidāmās vērtības maksimizēšanas principa

Pamatojoties uz pašreizējām parametru vērtībām, var novērtēt varbūtības z_1 un z_2, t. i., varbūtības, vai konkrētā metienu sērijā izmantotā monēta ir monēta A vai B. Pirmajai metienu sērijai $z_1 = 0.45$, $z_2 = 0.55$.

Katras metienu sērijas rezultāti tiek reizināti ar atbilstošajām varbūtību vērtībām z_1, z_2. Tādā veidā iegūst atbilstošās ģerboņu un ciparu proporcijas pie nosacījuma, ka izmestā monēta ir vai nu monēta A, vai monēta B. Faktiski tiek pārdalīts ģerboņu un ciparu skaits starp abām monētām, pamatojoties uz varbūtību z_1 un z_2 vērtībām.

Visas iepriekš aprakstītās procedūras veido algoritma E soli. Šis apzīmējums cēlies no angļu vārda "*expectation*" – cerība.

Pamatojoties uz iegūtajiem rezultātiem, tiek aprēķinātas jaunās varbūtību vērtības: $\hat{\theta}_A^{(1)} = 0.71$, $\hat{\theta}_B^{(1)} = 0.58$. Šo soli apzīmē ar M – no angļu valodas vārda *"maximization"* (maksimizācija).

Izmantojot jaunās varbūtības aprēķinus $\hat{\theta}_A^{(1)}$, $\hat{\theta}_B^{(1)}$, tiek veikta nākamā algoritma soļu E un M iterācija. Rezultātā iegūst aplēses $\hat{\theta}_A^{(2)}$, $\hat{\theta}_B^{(2)}$. Process tiek atkārtots cikliski, līdz aplēšu izmaiņas soļos (t) un $(t+1)$ kļūst mazākas par uzdoto robežvērtību. Iepriekšējā piemērā pēc desmitās algoritma iterācijas tiek iegūti šādi rezultāti: $\hat{\theta}_A^{(10)} = 0.80$, $\hat{\theta}_B^{(10)} = 0.52$. Šie rezultāti labi saskan ar rezultātiem, kas iegūti uz maksimālās līdzības pamata ar pilniem ieejas datiem, ņemot vērā, ka šajā gadījumā rezultāti ir iegūti, pamatojoties uz nepilnīgiem datiem.

Analizējot iepriekš dotā piemēra rezultātus, var izdarīt neapstrīdamu secinājumu, ka varbūtības aplēses, kas iegūtas, pamatojoties uz sagaidāmās vērtības maksimizēšanas metodes pamata, tajā pašā laikā ir arī maksimālās līdzības aplēses, kas iegūtas ar nepietiekamiem sākotnējiem datiem.

Formālas definīcijas, kas saistītas ar sagaidāmās vērtības maksimizācijas metodes izmantošanu iztrūkstošo atribūtu vērtību imputēšanai, ieviestas darbā [Garsia S., et al., 2015].

Ja mērķis ir modelēt atkarīgos gadījuma mainīgos kā novēroto mainīgo a un latento mainīgo b, kas ģenerē a, par pamatu ņem to, ka nezināmu parametru izlase Q nosaka varbūtību $p_\theta(a)$, $p_\theta(b)$ sadalījumus. Šo nezināmo parametru novērtēšanas process sastāv no diviem posmiem, kas tiek atkārtoti secīgi, līdz tiek panākta rezultātu konverģence.

1. solī E – tiek aprēķināta sagaidāmā vērtība:

$$Q(\theta,\theta')' = \sum p_{\theta'}(b/a)\log p_\theta(b,a), \qquad (2.12)$$

kur: $p_{\theta'}$ – nosacītā varbūtība b (atkarībā no a), pie jaunām θ' parametru vērtībām;

$p_\theta(b,a)$ – kopējā varbūtība a un b pie vecajām parametru θ vērtībām.

Kāpēc vienādojumā (2.11) tiek lietots termins $\log p_\theta(a,b)$? Lieta tāda, ka varbūtību vērtības, kas parādās vienādojumā (2.11), ir ļoti mazas un šo varbūtību reizināšanas rezultāts var nebūt precīzs. Pāreja no varbūtību vērtībām uz to logaritmiem ļauj atrisināt šo problēmu. Tā kā visas varbūtības ir mazākas par 1, ir svarīgas tādas parametru izlases θ logaritmu vērtības, kas ir tuvas 0.

2. solī M – tiek maksimizētas iepriekš aprēķinātās θ vērtības. Kā tiek panākta šī maksimizācija? Tiek apskatītas mainīgo A un B nosacītās sagaidāmās vērtības, kas iegūtas no sadalījumiem $p(a)$, $p(b)$, lai izpildītu soli M. Nosacītie sadalījumi tiek definēti kā

$$p(b/a) = \frac{p(b,a)}{p(a)}.$$

Tad gadījuma mainīgā B nosacītā sagaidāmā vērtība tiek uzdota kā

$$E(B) = \sum_b p(b)b.$$

Nosacītā sagaidāmā vērtība $\log p_\theta(b,a)$ pie dotajām a un θ' vērtībām var tikt aprēķināta šādā veidā:

$$E\left(\log p(b,a/\theta)/a,\theta'\right) = \sum_b p(b/a,\theta')\log p(b,a/\theta) =$$
$$= \sum_b p_{\theta'}(b,a)\log p_\theta(b,a). \qquad (2.13)$$

Acīmredzot, ja $\sum_b p_{\theta'}(b/a)\log p_{\theta'}(b,a) > \sum_b p_{\theta'}(b/a)\log p_{\theta}(b,a)$, tad $p_{\theta'}(a) > p_{\theta}(a)$. Šī arī ir sagaidāmās vērtības maksimizācijas metodes galvenā ideja: secīgi uzlabojot varbūtību sadalījuma parametru θ vērtības, var secīgi palielināt šī sadalījuma sagaidāmo vērtību. Iepriekš šajā sadaļā tika demonstrēta šī ideja ar vienkāršu ilustratīvu piemēru. Tagad šī ideja tika formāli pamatota.

Kā sagaidāmās vērtības maksimizēšanas metode darbojas iztrūkstošo atribūtu vērtību imputācijas kontekstā? Lai izmantotu šo metodi, vispirms ir jānosaka dažas sākotnējās parametru vērtības atribūtu vērtību sadalījumiem. Sagaidāmās vērtības maksimizēšanas metodes algoritms labi darbojas gadījumos, ja atribūtu vērtību izlases ir sadalītas pēc normāliem vai tiem tuviem sadalījumiem.

Katrā iterācijā sagaidāmās vērtības maksimizēšanas metodes algoritms novērtē atribūtu $\hat{\mu}_j$ vidējās vērtības un variācijas/kovariācijas vērtības visām atribūtu vērtību kopām. Variācijas/kovariācijas vērtības tiek parādītas atbilstoša izmēra matricas $\hat{\Sigma}$ veidā. Pamatojoties uz iegūtajiem datiem, tiek aprēķināti regresijas parametri B, pēc kuriem tiek aprēķinātas atribūtu pašreizējās vērtības. Šīs pašreizējās vērtības tiek noteiktas pēc vienādojuma:

$$A_{mis} = \mu_{mis} + (A_{obs} - \mu_{obs})B + e, \tag{2.14}$$

kur A_{mis} – iztrūkstošo atribūtu vērtību izlase;

A_{obs} – esošo atribūtu vērtību izlase;

μ_{mis} – iztrūkstošo atribūtu vērtību vidējās vērtības;

μ_{obs} – esošo atribūtu vērtību vidējās vērtības;

B – regresijas parametri;

e – kļūdas rādītājs, kas tiek uzdots kā nejaušs vektors ar vidējās vērtību 0 un ar nezināmu variāciju matricu.

Iepriekš aprakstītās skaitļošanas procedūras atspoguļo sagaidāmās vērtības maksimizācijas metodes algoritma soļa E būtību iztrūkstošo atribūtu vērtību imputācijā. Vienkāršoti izsakoties, šī soļa darbības mērķis ir atkārtoti novērtēt iztrūkstošās atribūtu vērtības, pamatojoties uz novērtētajiem regresijas parametriem B.

Algoritma M solī tiek atkārtoti novērtēti atribūtu vērtību kopu sadalījumu parametri, izmantojot gan esošajā, gan iepriekšējā iterācijā aprēķinātās atribūtu vērtības.

$$\hat{\mu}^{(t+1)}(a_j) = \frac{\sum_{i=1}^{m} a_{ij}^{(t)}}{m}, \quad j = 1,...,n. \tag{2.15}$$

$$\hat{\Sigma}^{(t+1)} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\Sigma}^{(t)} - \left(\hat{\mu}^{(t+1)}\hat{\mu}^{(t+1)}\right)\right). \tag{2.16}$$

Izmantojot pārvērtēšanas rezultātus $\hat{\mu}^{(t+1)}(a_j)$ un $\hat{\Sigma}^{(t+1)}$, $j = 1,...,n$ var pārvērtēt regresijas parametru vērtības $\hat{B}^{(t+1)}$. Pēc tam aprēķinātās atribūtu vērtības tiek atkārtoti novērtētas un process tiek atkārtots, kamēr tiks sasniegta rezultātu konverģence.

Būtisks sagaidāmās vērtības maksimizēšanas metodes trūkums ir tas, ka faktiski to izmanto, lai meklētu sadalījuma parametru maksimumu lokālās vērtības. Šie lokālie maksimumi var nebūt globālie maksimumi. Lai izvairītos no konverģences uz lokālajiem maksimumiem, ieteicams atkārtot metodes izpildi pie dažādiem sākuma nosacījumiem. Iztrūkstošo atribūtu vērtību imputācijas kontekstā tas nozīmē, ka sākotnējo atribūtu vidējo vērtību vietā iztrūkstošās vērtības tiek aizstātas ar skaitļiem, kas nav vidējās vērtības. Tādējādi metodes soļi tiek veikti pie jauniem sākuma nosacījumiem. Ja abos gadījumos iegūtie rezultāti sakrīt, tas norāda, ka ir sasniegts globālais maksimums, ja rezultāti atšķiras, tad par galarezultātu var ņemt labāko no diviem iegūtajiem.

Sagaidāmās vērtības maksimizēšanas metodes praktiskai izmantošanai ir nepieciešams ļoti liels skaitļošanas procedūru skaits. Pat ar ļoti ierobežotu ievaddatu skaitu nepieciešamo aprēķinu veikšana manuāli nav reāla, tāpēc attiecīgie aprēķini jāveic, izmantojot piemērotus programmatūras rīkus.

2.5. Daudzkārtīgā imputācija

Iepriekšējā sadaļā tika apspriesta sagaidāmās vērtības maksimizācijas metode iztrūkstošo atribūtu vērtību imputēšanai. Šajā sadaļā īsumā tiks atkārtota šīs metodes būtība, lai to salīdzinātu ar daudzkārtīgās imputācijas metodi.

Sagaidāmās vērtības maksimizācijas metodes pamatā ir interesējošo atribūtu vērtību sadalījuma parametru novērtējums, pamatojoties uz to sagaidāmo vērtību maksimizēšanu. Maksimizēšanas process tiek veikts iteratīvi. Katrā iterācijā notiek sadalījuma parametru secīga tuvināšana to faktiskajām vērtībām. Pamatojoties uz atkārtoti novērtētajām parametru vērtībām, tiek veikts nākamais iztrūkstošo atribūtu vērtību imputācijas posms. Jaunā atribūtu vērtību komplektā (novērotās + imputētās vērtības) sadalījuma parametri tiek atkārtoti novērtēti un process cikliski atkārtots, līdz tiek sasniegti galīgie rezultāti.

Daudzkārtīgās imputācijas pamatā ir cits imputācijas princips. *Apriori* tiek pieņemts, ka atribūtu vērtību izlases tiek sadalītas saskaņā ar kādu sadalījuma likumu. Parasti tiek pieņemts, ka šis sadalījums ir daudzfaktoru normālais sadalījums.

Izmantojot iegūtos rezultātus, katram sadalījumam tiek atlasītas (modelētas) tā aprēķinātās vērtības no atbilstošā sadalījuma. Pēc tam, izmantojot Beijesa formulu, pamatojoties uz sākotnējo iepriekšējo sadalījumu un aprēķinātajām atribūtu vērtībām, tiek noteikts atribūtu vērtību *aposteriori* sadalījums, ņemot vērā to imputētās vērtības.

Tad tiek izveidota jauna imputēto atribūtu vērtību izlase un process tiek atkārtots, līdz rezultāti konverģē (rezultātu konverģences jēdziena definīcija metodē tiks dota vēlāk šajā sadaļā).

Tagad var veikt abu metožu salīdzinājumu. Abās metodēs tiek pieņemts, ka atribūtu vērtību izlases ir sadalītas saskaņā ar kādu varbūtisko likumu. Bet sagaidāmās vērtības maksimizācijas metodē, pamatojoties uz šo pieņēmumu, šo sadalījumu parametri tiek secīgi pārvērtēti. Katrā iterācijā ar aprēķinātajām parametru vērtībām tiek veikta nākamā imputēto atribūtu vērtību pārvērtēšana. Citiem vārdiem sakot, katrs nākamais aprēķins tiek veikts, pamatojoties uz iepriekšējo atribūta vērtību parametru pārvērtēšanu.

Izmantojot daudzkārtīgās imputācijas metodi, katrā iterācijā no šo vērtību *aposteriori* sadalījumiem tiek veidota nejauša imputēto atribūtu vērtību izlase.

Daudzkārtīgās imputācijas metode pirmo reizi tika piedāvāta darbā [Rubin D.B., 1977] un prezentēta pilnā formā darbā [Rubin D.B., 1987]. Šīs metodes attīstību izraisīja nepieciešamība apstrādāt ļoti lielu skaitu apskatu (dažādu veidu aptauju rezultāti). Kā minēts iepriekš, šāda veida apskatos dažādu iemeslu dēļ trūkst daudz atbilžu uz uzdotajiem jautājumiem. Lai atrisinātu šo problēmu, arī tika izstrādāta daudzkārtīgās imputācijas metode.

Sagaidāmās vērtības maksimizācijas metode un citas imputācijas metodes parasti ir saistītas ar konkrētiem datu apstrādes un analīzes mērķiem un uzdevumiem. Daudzkārtīgās imputācijas metode nav vērsta uz konkrētiem mērķiem un uzdevumiem. Tā ir universāla tādā nozīmē, ka iztrūkstošo atribūtu vērtību imputācija tiek veikta datu vākšanas posmā un iegūtie rezultāti tiek izmantoti citu šo datu analīzes problēmu risināšanai. Citiem vārdiem sakot, izmantojot sagaidāmās vērtības maksimizācijas metodi, iztrūkstošo atribūtu vērtību imputāciju un turpmāko datu analīzi veic viens lietotājs konkrēta uzdevuma kontekstā. Savukārt, izmantojot daudzkārtīgās imputācijas metodi, imputācija tiek veikta, neņemot vērā datu analīzes mērķa uzdevumus, un iegūtos rezultātus izmanto citi lietotāji konkrētu uzdevumu kontekstā.

Daudzkārtējās imputācijas metodes īpatnība ir tāda, ka iztrūkstošo atribūtu vērtību imputācija tiek veikta nevis vienu reizi, kas ir raksturīgi visām pārējām metodēm, bet gan noteiktu skaitu m. Jebkuru vai visus iegūtos rezultātus var izmantot turpmākajos pētījumos.

Daudzkārtējās imputācijas metodes izpilde ir balstīta uz trim pieņēmumiem [Sinharay S., et al., 2001].

1. Datu modelis

Pirmais un vissvarīgākais solis daudzkārtējās imputācijas metodes izmantošanā ir pieņemt varbūtisko modeli, kas saista visu datu izlasi A, kas ir novēroto vērtību A_{obs} un iztrūkstošo vērtību A_{mis} kombinācija, ar šo vērtību parametru sadalījumu izlasi. Izmantojot šo varbūtisko modeli un

apriori sadalījumu pēc parametriem, var noteikt sagaidāmo sadalījumu $p(A_{mis} / A_{obs})$ iztrūkstošām vērtībām atkarībā no novērotajām vērtībām. Izmantojot šo sagaidāmo sadalījumu, var ģenerēt imputētās atribūtu vērtības.

Vispiemērotākais modelis nepārtrauktām atribūtu vērtībām ir pieņemt daudzfaktoru normālo sadalījumu visām atribūtu vērtībām. Ja dažu atribūtu vērtības ir sadalītas saskaņā ar likumu, kas būtiski atšķiras no normālā sadalījuma, tad šīs vērtības var pārveidot tā, lai transformācijas rezultāti tiktu sadalīti saskaņā ar normālo sadalījumu (atribūtu vērtību transformācijas problēmas ir apskatītas 6. nodaļā). Imputētās transformētās atribūtu vērtības var tikt apgrieztas, lai iegūtu to sākotnējās vērtības.

2. Atribūtu vērtību kopu apriori sadalījums

Tā kā uz modeli balstītā daudzkārtējās imputācijas metode pēc būtības ir balstīta uz Beijesa teorēmas pamata (*Bayes rule*), ir jānosaka vispārējs *apriori* sadalījums, no kura var secināt *aposteriori* sagaidāmo sadalījumu $p(A_{mis} / A_{obs})$. Parasti ērtības labad tiek izmantots kāds neinformatīvs *apriori* sadalījums, piemēram, $p(\theta) \propto \frac{1}{\sigma^2}$.

3. Iztrūkstošo datu mehānismi

Uz modeli balstītā daudzkārtējās imputācijas metodē pieņem to, ka iztrūkstošo datu mehānisms ir vai nu pilnīgi nejaušs iztrūkums (*MCAR*), vai nejaušs iztrūkums (*MAR*).

Kopumā daudzkārtējās imputācijas metodes procesus var definēt kā Montekarlo simulāciju Markova ķēdēs. Tipisks Montekarlo simulācijas piemērs ir imputēto atribūtu vērtību nejauša atlase no to *apriori* sadalījumiem. *Apriori* paredzamo sadalījumu secīga pārvērtēšana pēc būtības veido Markova ķēdi.

Aplūkojot daudzkārtējās imputācijas metodes matemātiskos pamatus, turpmākais izklāsts ietver definīcijas, kas sniegtas darbā [Sinharay S., et., 2001], kas savukārt ir balstīts uz [Schafer J.L., 1997] darbu. Tiek pieņemts, ka atribūtu vērtību izlases ir nepārtraukti nejauši mainīgie. Ja atribūtu vērtību izlases ir diskrēti gadījuma mainīgie, tad visas turpmākās definīcijas ir derīgas, aizstājot integrēšanas darbības ar summēšanas operācijām.

Sākotnējo datu izlase tiek apzīmēta ar A. Šī izlase sastāv no divām daļām: $A = (A_{obs}, A_{mis})$, kur A_{obs} ir novēroto datu izlase (atribūtu vērtības) un A_{mis} – iztrūkstošo datu izlase (atribūtu vērtības). Tālāk pieņem, ka visu atribūtu vērtības ir sadalītas saskaņā ar likumu $p(A/\theta)$, kur θ ir visu modeļa parametru izlase (vektors).

Prognozējamo sadalījumu $p(A_{mis} / A_{obs})$ var attēlot kā

$$p(A_{mis} / A_{obs}) = \int p(A_{mis}, \theta / A_{obs})\, d\theta = \\ = \int p(A_{mis} / A_{obs}, \theta) p(\theta / A_{obs})\, d\theta . \qquad (2.17)$$

Iztrūkstošo atribūtu vērtību imputēšana no izlases A_{mis} tiek veikta divos posmos. Pirmais posms ietver parametru vērtību modelēšanu no *aposteriori* sadalījuma $p(\theta / A_{obs})$. Otrais posms ir modelēt iztrūkstošo atribūtu vērtību vektoru no nosacītā *aposteriori* sadalījuma $p(A_{mis} / A_{obs}, \theta)$, izmantojot pirmajā posmā ģenerētās θ vērtības.

Iztrūkstošo atribūtu vērtību imputācijas tiek veiktas, izmantojot pirmo no vienādojumiem (2.17), proti,

$$p(A_{mis} / A_{obs}) = \int p(A_{mis}, \theta / A_{obs})\, d\theta .$$

Šajā gadījumā tiek izmantots datu pieauguma algoritms (*data augmentation algorithm*), kas tika piedāvāts darbā [Tanner M.A., Wong W.H., 1987]. Būtībā šis algoritms ir Markova ķēdes simulācija. Algoritma k iterācijā imputētās atribūtu vērtības tiek atlasītas no nosacītā paredzamā sadalījuma $A_{mis}^{(k)} \sim p(A_{mis} / A_{obs}, \theta^{(k-1)})$, bet nākamās parametru vērtības tiek atlasītas no *aposteriori*

sadalījuma visai datu izlasei $\theta^{(k)} \sim p\left(\theta / A_{obs} / A_{mis}\right)^{(k)}$. Sākot no noteiktām sākotnējām vērtībām, šādu procedūru secība veido Markova ķēdi $\left\{A_{mis}^{(k)}, \theta^{(k)}, k = 1, 2, ...\right\}$.

Kā var noteikt datu pieauguma algoritma rezultātu konverģenci? Kādos apstākļos (ar kādu k vērtību) jāpārtrauc algoritma izpildes procedūru secība? Aplūkojot sagaidāmās vērtības maksimizācijas metodi iepriekšējā sadaļā, algoritma konverģence tika noteikta pēc nosacījuma, ka $\left(k+1\right)$ algoritma solī iegūtie rezultāti maz atšķiras no rezultātiem, kas iegūti k solī. Citiem vārdiem sakot, rezultātu atšķirības nepārsniedz noteiktu slieksni. Datu pieauguma algoritmā šī algoritma konverģence tiek noteikta sarežģītākā veidā. Fakts ir tāds, ka šajā algoritmā parametru sadalījums nemainās, veicot secīgas darbības, bet mainās nejauši izvēlētās parametru vērtības. Darbā [Schafer J.L., Olson M.K., 1998] tiek piedāvāta šāda definīcija: datu pieauguma algoritms konverģē pēc k iterācijām, ja jebkura parametra vērtība iterācijā $t \in 1, 2, ...$ ir statistiski neatkarīga no šī parametra vērtības iterācijā $t+k$.

Pēc vienas pilnas algoritma izpildes tiek iegūta parametru vērtību izlase un imputēto atribūtu vērtību izlase. Bet, kā saka pats metodes nosaukums, lieta nebeidzas ar vienu algoritma izpildi. Algoritms tiek izpildīts noteiktu skaitu m reižu. Ņemot vērā aprēķināto atribūtu vērtību izvēles nejaušību un konverģences nosacījumu, iegūst m aprēķināto atribūtu vērtību izlases. Lai veiktu nepieciešamo datu analīzi, var izmantot jebkuru no pilnajām datu izlasēm, ieskaitot novērotās un aprēķinātās atribūtu vērtības.

Kādam vajadzētu būt daudzkārtējās imputāciju skaitam m? Darbā [Rubin D.B., 1987] tiek apgalvots, ka pietiek ar 3 līdz 5 imputācijām. Tika iegūta šāda sakarība starp daudzkārtējo imputāciju skaitu m un novērtējuma efektivitāti:

$$\text{Novērtējuma efektivitāte} \approx \left(1+\frac{\gamma}{m}\right)^{-1},$$

kur γ – iztrūkstošo atribūtu vērtību proporcija sākotnējā datu izlasē.

Ja datos iztrūkst 30 % atribūtu vērtību, 94 % efektivitāti var sasniegt ar 5 galīgām aizpildītām datu izlasēm. Ja $m=10$, tad efektivitāte ir 97 %, kas dod tikai nelielu efektivitātes pieaugumu par dubultām skaitļošanas izmaksām.

Saistībā ar daudzkārtējām imputācijām rodas jautājums: kāpēc ģenerēt m imputēto atribūtu vērtību izlases, ja tikai viena no šīm izlasēm tiek izmantota turpmākai datu analīzei? Lieta tāda, ka, izmantojot daudzkārtīgās imputācijas rezultātus, var tikt novērtēta kopējo imputācijas rezultātu nenoteiktības pakāpe.

Darbā [Rubin D.B., 1987] tika formulēti šādi noteikumi, lai apvienotu interesējošo parametru vērtības, kas iegūtas no katras m imputācijas.

Interesējošo parametru apzīmē ar R. Tas varētu būt atribūtu vērtību izlases vidējais lielums, regresijas koeficients starp divām atribūtu vērtību izlasēm vai jebkurš cits atbilstošs parametrs.

Ar parametra $\hat{R}_i$ vērtību apzīmē interesējošo sadalījumu i iterācijā. Tad šī parametra apkopoto novērtējumu visām m imputācijām var aprēķināt ar šāda vienādojuma palīdzību:

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m}\hat{R}_i . \qquad (2.18)$$

Novērtēšanas variācija ietver divus komponentus: variācijas katrā iegūtajā datu izlasē un variācijas starp iegūtajām pilnajām datu izlasēm.

Iekšējā variācija tiek novērtēta pēc vienādojuma:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}U_i . \qquad (2.19)$$

Acīmredzami šī variācija skaitliski ir vienāda ar vidējo variāciju visās pilnajās datu izlasēs.

Variācija starp iegūtajām pilnajām datu izlasēm aprēķina, izmantojot vienādojumu:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{R}_i - \bar{R}\right)^2 . \quad (2.20)$$

Kopējā variācija T ir šo divu komponentu koriģētā summa:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B . \quad (2.21)$$

Kvadrātsakne no T ir standarta kļūdas aplēse, kas saistīta ar vidējo novērtējumu $\bar{R}$.

Viena no daudzkārtējās imputācijas metodes priekšrocībām ir spēja novērtēt nenoteiktības parametru aplēsēs, kas saistītas ar iztrūkstošo atribūtu vērtību vairākkārtēju imputāciju. Neviena cita imputācijas metode nevar novērtēt rezultātu nenoteiktību.

Novērtējam daudzkārtējās imputācijas metodes galvenās priekšrocības un trūkumus. Šī analīze ir balstīta uz apkopotajiem datiem, kas sniegti darbā [Allison P.D., 2000].

Ērtības labad autors piedāvā vispārinātu daudzkārtējās imputācijas procedūru secību:

1) imputējiet iztrūkstošās atribūtu vērtības, izmantojot piemērotu modeli, kas ietver nejaušas variācijas;
2) atkārtojiet imputācijas procesu m reizes (parasti 3–5 reizes);
3) veiciet vēlamo analīzi katram (vai jebkuram) sadalījumam no iegūtās datu izlases, izmantojot piemērotas statistikas metodes;
4) izskaitļojiet vidējās parametru vērtības visās m izlasēs;
5) aprēķiniet standartkļūdas: (a) aprēķinot standarta vidējo vērtību kļūdas; b) m parametru aplēšu variāciju aprēķināšana visās datu izlasēs; c) abu iegūto aplēšu apvienošana.

Daudzkārtējai imputācijai ir šādas vēlamās īpašības:

- piemērotu nejaušu novērtējumu ieviešana imputācijas procesā padara par iespējamu iegūt objektīvus visu vēlamo parametru aprēķinus;
- atkārtota imputācija nodrošina ticamus standarta kļūdas aprēķinus. Vienas imputācijas metodes nevar apstrādāt papildu kļūdas imputācijas procesa laikā;
- daudzkārtējo imputāciju var pielietot jebkura veida datiem, izmantojot piemērotu programmatūra.

Praktiski pielietojot daudzkārtējās imputācijas metodi, ir jāievēro noteiktas prasības. Galvenās pamatprasības [Rubin D. B., 1987, 1996] ir:

- iztrūkstošo datu mehānismam ir jābūt *MCAR* vai *MAR*;
- modelim, kas izmantots imputēto vērtību ģenerēšanai, kaut kādā ziņā jābūt "korektam";
- datu analīzei izmantotajam modelim ir jāatbilst modelim, ko izmanto atribūtu vērtību imputēšanai.

Iespējamie daudzkārtējās imputācijas metodes neveiksmīgi lietojumi ir saistīti tikai ar vienas vai vairāku šo prasību neievērošanas.

Daudzkārtējās imputācijas metodes praktiskā izmantošana prasa veikt daudzas sarežģītas skaitļošanas procedūras, tāpēc jāizmanto specializēti skaitļošanas rīki.

Pastāv arī citas pieejas iztrūkstošo atribūtu vērtību imputēšanai. Starp šādām pieejām var minēt imputāciju, kas balstīta uz komponentu analīzi, un imputāciju, kuras pamatā ir neironu tīkli.

3. ANOMĀLIJU IDENTIFICĒŠANA DATOS

Šajā un nākamajās divās nodaļās ir aplūkotas datu attīrīšanas (*cleaning*) problēmas. Datu attīrīšana ir saistīta ar kļūdu un neatbilstību identificēšanu un novēršanu datos, lai uzlabotu šo datu kvalitāti. Kvalitātes problēmas datu kopās rodas nepareizas objektu vai atribūtu definīcijas, iztrūkstošu atribūtu vērtību, neinformatīvu atribūtu izmantošanas un citu iemeslu dēļ. Nepieciešamība attīrīt datus ievērojami palielinās, ja dati tiek integrēti no dažādiem avotiem.

Šajā nodaļā tiks aplūkota viens datu attīrīšanas aspekts, proti, anomāliju identificēšana datos.

3.1. Kas ir anomālijas datos?

Zinātniskajā literatūrā par datu analīzi ir vairāki termini, ar kuru palīdzību tiek raksturotas anomālijas datos: "*anomaly detection*", "*outlier detection*" vai "*novelty detection*". Turpmākajā izklāstā tiks izmantots termins "anomāliju noteikšana jeb identificēšana".

Darbā [Hawkins D.M., 1980] dota šāda definīcija: anomālija datos ir "novērojums, kas tik ļoti atšķiras no citiem novērojumiem, ka ir aizdomas, ka to ģenerēja cits mehānisms". Autori [Moore Y.D., McCabe G., 1999] anomāliju definē kā datus, kas neatbilst nekādiem nosacītajiem sadalījumiem. Darbā [Barnett V., Lewis T., 1984] sniegta šāda definīcija: "Anomālija datu kopā ir novērojums, kas nav saderīgs ar doto datu kopu". Dažkārt literatūrā par datu iegūšanu un analīzi anomālijas sauc par "novirzēm", "pretrunām" vai "novirzēm no normas".

Apkopojot visas šīs definīcijas, šajā nodaļā anomāli dati tiek definēti kā datu paraugi (objekti), kas būtiski atšķiras no citiem datu paraugiem atribūtu vērtībās.

Anomāliju ignorēšana datos var izraisīt nopietnas negatīvas sekas [Salgato C. et al., 2016]: (1) kļūdu dispersijas palielināšanās un statistiskās nozīmes samazināšanās; (2) samazināta normalitāte gadījumos, kad anomālija neatbilst normāla sadalījuma likumiem.

Anomāliju jēdzienu nevajadzētu jaukt ar trokšņa jēdzienu datos. Troksnis datos ir saistīts ar nelielām atribūtu vērtību novirzēm no to patiesajām vērtībām. Tajā pašā laikā anomālijām ir raksturīgas ievērojamas novirzes atribūtu vērtībās, salīdzinot ar šo atribūtu vērtībām galvenajā datu kopā.

Tālāk ir dots ieskats pielietojumos, kas izmanto anomāliju atpazīšanu un analīzi [Hedge V.J., Austin J., 2004]:

- krāpšanas atpazīšana kredītkaršu izmantošanā;
- problemātisku klientu noteikšana kredīta izlietošanā;
- nesankcionētu darbību datortīklos identificēšana;
- aktivitāšu monitorings: krāpšanās ar mobilajiem tālruņiem identificēšana;
- tīkla veiktspējas uzraudzība, piemēram, lai identificētu tīkla vājās vietas;
- bojājumu diagnostika – procesa uzraudzība, lai noteiktu bojājumus iekārtās, piemēram, kosmosa kuģos;
- strukturālo bojājumu atpazīšana – ražošanas līniju monitorings;
- satelītattēlu analīze – jaunu vai vēl neklasificētu pazīmju noteikšana;
- jaunumu atpazīšana paraugos – jaunas robotu vai novērošanas sistēmu reakcijas;
- kustību segmentācija – attēla iezīmju atpazīšana, kas kustas neatkarīgi no pamata attēla;
- medicīnisko stāvokļu uzraudzība, piemēram, sirdsdarbības;
- farmaceitiskie pētījumi – jaunu molekulāro struktūru noteikšana;
- neparedzētu šūnu atpazīšana datu bāzēs, lai identificētu kļūdas, viltojumus, negaidītas šūnas;
- nemarķētu datu atpazīšana apmācības datu kopās.

Anomāliju parādīšanās datos iemesli var būt cilvēka kļūdas, mērinstrumentu darbības traucējumi, dabiskas novirzes datu unikālo objektu atribūtu vērtībās, krāpnieciska rīcība, izmaiņas tehniskās sistēmas darbībā, iekārtu elementu bojājumi u. c.

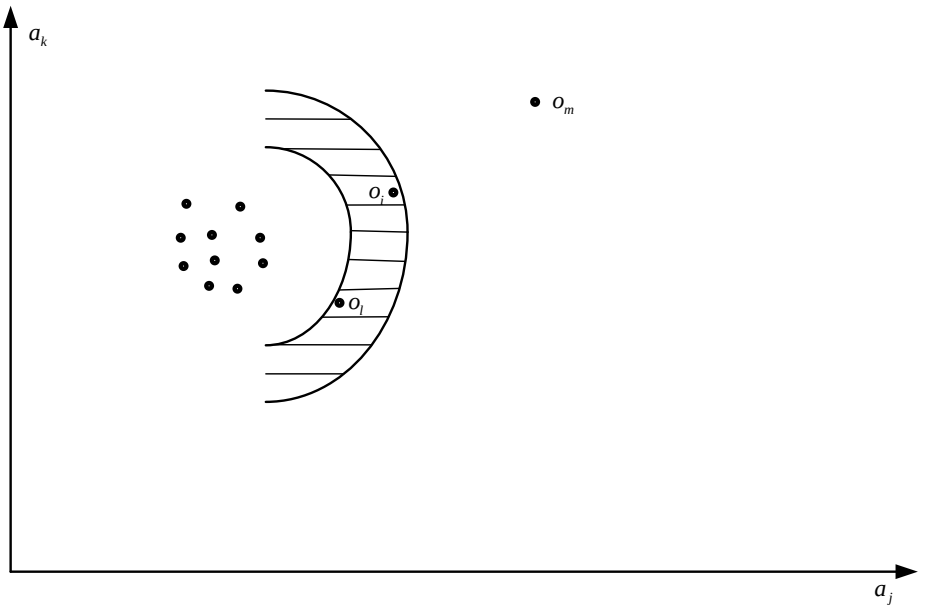
Ir jāsaprot, kādi ir galvenie mērķi anomāliju identificēšanai datos. Piemēram, formatējot tekstu datorā, pareizrakstības programma automātiski atpazīst kļūdas, signalizē par to un piedāvā pareizu

attiecīgā vārda pareizrakstību. Uzdevums šeit ir identificēt anomāliju un savlaicīgi to novērst. Citiem vārdiem sakot, šajā kontekstā mērķis ir pilnībā attīrīt datus no anomālijām un iegūt normālus datus, t. i., tekstu bez gramatiskām kļūdām. Var teikt, ka svarīgi ir tikai normāli dati un anomālijas nemaz neinteresē.

Cits piemērs. Lai identificētu pacienta iespējamo slimību, ārsts aicina pacientu veikt nepieciešamās pārbaudes un veikt izmeklējumus. Analizējot šo izmeklējumu rezultātus, ārsts pievērš uzmanību tiem pacienta ķermeņa darbības parametriem, kas pārsniedz normu, t. i., meklē anomālijas. Tādējādi ārsts izmanto anomālijas, lai identificētu pacienta iespējamo slimību. Šajā gadījumā ārstu neinteresē normālie rādījumi, bet anomālijas kā svarīgas informācijas avoti.

Neatkarīgi no anomāliju identificēšanas mērķa, to noteikšanas procesi ir līdzīgi.

Anomāliju identifikāciju datos būtiski sarežģī trokšņu klātbūtne šajos datos un nepārprotamas robežas starp troksņiem un anomālijām neiespējamība. 3.1. attēlā objekti ir attēloti ar punktiem divu atribūtu telpā: a_j, a_k . Lielākā daļa datu punktu veido kompaktu kopu un tos var korekti interpretēt kā normālus datus. Datu objekta punkts o_m ļoti atšķiras no visiem citiem datiem, tāpēc objektu o_m var identificēt kā anomāliju.

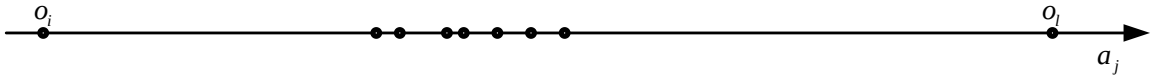


3.1. attēls. **Trokšņu un anomāliju grafisks attēlojums**

Problēmas rodas ar datu punktiem o_i un o_l . Šie punkti atrodas tā sauktajā "pelēkajā zonā", kas ir izplūdusi robeža starp normāliem datiem un anomālijām.

Objektus o_i un o_l datos var interpretēt kā trokšņus. Bet objekts o_l ir tuvāk normālu datu apgabalam, bet objekts o_i ir tuvāk anomāliju apgabalam. Tas, vai objekts o_l pieder normālu datu kopai un vai objekts o_i ir anomāls, ir jautājums, uz kuriem var atbildēt, tikai rūpīgi analizējot šos datu punktus. Objektus, kuru punkti atrodas pelēkajā zonā, sauc par iespējamām anomālijām, bet objektus, kuru punkti atrodas ārpus pelēkās zonas augšējās robežas, par varbūtējām anomālijām [Tukey J., 1977]. Acīmredzot galveno problēmu rada objekti, kas, iespējams, ir anomālija.

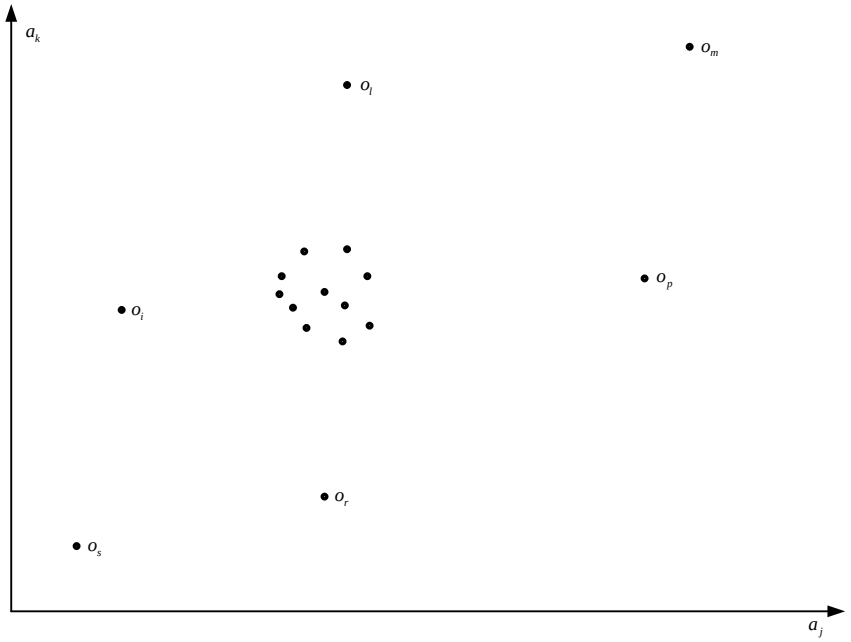
Tālāk tiks apskatīta vēl viena problēma, kas saistīta ar anomāliju identificēšanu datos. 3.2. attēlā punkti parāda objektus, kurus raksturo tikai viena atribūta vērtības a_j .



3.2. attēls. **Objektu grafiskais attēlojums atribūtu vērtību telpā** a_j

Lielākā daļa objektu punktu atrodas ierobežotā atribūtu vērtību diapazonā a_j. Acīmredzot tie ir normāli dati. Objektus o_i un o_l var uzskatīt par potenciālām anomālijām. Objekta o_i atribūta vērtība a_j ir ievērojami mazāka nekā šī atribūta vērtības lielākajai daļai datu, objekta o_l atribūta vērtība a_j ir ievērojami lielāka nekā šī atribūta vērtības lielākajai daļai datu.

3.3. attēlā punkti apzīmē datu objektus divu atribūtu telpā a_j, a_k.



3.3. attēls. **Objektu grafiskais attēlojums atribūtu telpā a_j, a_k**

Lielākā daļa datu tiek parādīta skaidri noteiktas datu kopas veidā.

Apskatot objektus o_i, o_p, var redzēt, ka šie objekti ir potenciālas anomālijas pēc atribūta a_j vērtībām, bet nav anomālija pēc atribūta a_k vērtībām, jo šī atribūta vērtības šiem objektiem atbilst atribūtu vērtībām lielākajai daļai datu. Tāpat objekti o_l, o_r ir potenciālas anomālijas pēc atribūta a_k vērtībām, bet ne pēc atribūta a_j vērtībām. Tikai objektus o_m, o_s var uzskatīt par potenciālām anomālijām, pamatojoties uz abu atribūtu vērtībām.

Situācija var būt vēl sarežģītāka, ja objekti tiek novērtēti pēc atribūtu vērtībām $n > 2$. Šādos apstākļos dažus objektus var uzskatīt par potenciālām anomālijām, pamatojoties uz viena vai vairāku atribūtu vērtībām, bet, pamatojoties uz citu atribūtu vērtībām, tie var atbilst lielākajai daļai datu. Tas ļauj secināt šo: lai praktiski identificētu anomālijas, kas novērtētas pēc noteiktas atribūtu kopas vērtībām, ir nepieciešams identificēt potenciālās anomālijas pēc katra atribūta vērtību kopas.

Vispārīgā gadījumā anomālija var piederēt kādam no šiem veidiem.

1. Globālā anomālija

Objektu var uzskatīt par globālu anomāliju, ja tas būtiski atšķiras no citiem objektiem datu kopā pēc visām atribūtu vērtībām. Lai identificētu globālo anomāliju, ir jāizmanto piemērota dispersijas metrika.

2. Kontekstuālā anomālija

Spilgts piemērs atribūtam, kas var izraisīt kontekstuālas anomālijas, ir temperatūra. Pieņem, ka objekti atrodas dažādās apsildāmās telpās un tiek izmērīta šo objektu temperatūra. Ja telpas temperatūra ir robežās no 18 ^{0}C–22 ^{0}C, šī temperatūra atbilst objekta normālajam stāvoklim. Ja temperatūra ir 12 ^{0}C vai 29 ^{0}C, tad objekts neatrodas normālos apstākļos un to var identificēt kā kontekstuālu anomāliju.

3. Kolektīvā anomālija

Ja noteikta objektu apakškopa būtiski atšķiras no objektu lielākās daļas, tad šādus objektus var raksturot kā kolektīvas anomālijas.

Ir trīs pamata pieejas anomāliju identificēšanai [Hodge V.J, Austin J., 2004].

1. veids. Nosaka anomālijas, iepriekš nezinot informāciju par datiem. Būtībā šī ir apmācības pieeja, kas ir analoģiska neuzraudzītai klasterizācijai (*unsupervised clusterization*). Algoritms klasterizē datus, nosaka attālākos punktus un atzīmē tos kā potenciālo anomāliju. Šāda veida identifikācija paredz, ka kļūdainie dati ir atdalīti no lielākās daļas datu un tāpēc tiek pieņemti kā anomālijas.

Praksē tiek izmantotas dažādas šīs pieejas variācijas. Diagnostikas pieeja identificē potenciālās anomālijas un izslēdz tās no turpmākās datu apstrādes. Daudzas diagnostikas pieejas iteratīvi izslēdz anomālijas. Process tiek atkārtots, līdz pamatdatos vairs nav anomāliju.

Vēl viena pieeja realizē veidu, kas ģenerētajā datu modelī iekļauj anomālijas un izmanto dažādas metodes šo datu tālākai apstrādei. Šādas metodes var dod iespēju ignorēt anomālijas un nosaka robežu ap lielāko normālo daļu datu.

2. veids. Nosaka modeļa normalitāti. Šī pieeja ir līdzīga uzraudzītajai klasifikācijai (*supervised classification*) un prasa iepriekšēju normālu un anomālu datu iezīmēšanu. Visi dati ārpus normālo datu apgabala tiek traktēti kā anomāli.

Šo pieeju var izmantot kā klasifikatoru, kurš tiek apmācīts par klasifikācijas modeli un pēc tam klasificē jaunus piemērus (objektus). Ja jauns piemērs atrodas normalitātes apgabalā, tas tiek klasificēts kā normāls, pretējā gadījumā piemērs tiek atzīmēts kā anomālija.

3. veids. Modelē tikai normalitāti. Šī pieeja parasti tiek izmantota, lai atpazītu jaunus papildu datus. Tā ir līdzīga daļēji uzraudzītai atpazīšanai (*semi-supervised recognition*). Tiek pieņemts, ka ir normālu datu kopa, un algoritms meklē anomālijas, kas varētu interesēt pētnieku.

Lielāko daļu anomāliju identificēšanas metožu var iedalīt šādās lielās klasēs:

- statistiskās metodes;
- metodes, kuru pamatā ir objektu tuvības mērs;
- metodes anomāliju identificēšanai regresijas atkarībās.

Nākamajās sadaļās tiks apspriestas visbiežāk izmantojamās metodes katrā grupā.

3.2. Statistiskās metodes anomāliju identificēšanai

Visās šāda veida metodēs ir pieņemts, ka atribūtu vērtības tiek sadalītas saskaņā ar noteiktu (visbiežāk normālu) sadalījuma likumu. Tomēr daudzās reālās datu kopās esošie sadalījumi var būt nezināmi vai sarežģīti. Tāpēc, lai izmantotu statistikas metodes, ir rūpīgi jānosaka sadalījuma robežas, kurās atribūtu vērtības var attiecināt uz anomālijām.

Apskatam tiek piedāvātas dažas plaši izmantotas statistikas metodes anomāliju identificēšanai.

1. z iezīmju metode.

Šīs metodes būtība ir tāda, ka atribūtu faktiskās vērtības tiek aizstātas ar tā sauktajām z iezīmēm. Atribūta a_{ij} vērtībai tās z iezīmi aprēķina šādi:

$$z_{ij} = \frac{a_{ij} - \bar{a}_j}{s_j},\ i = 1,...,m,\ j = 1,...,n, \tag{3.1}$$

kur $\bar{a}_j -$ j atribūta vidējā vērtība;

s_j – j atribūta standartnovirzes vērtība.

Atribūtu vērtību pārveidošana atbilstošās z iezīmēs ļauj novērtēt attālumu no katras atribūta vērtības līdz vidējai atribūtu vērtībai. Pastāv problēma ar iezīmes robežvērtības z_{jg} definēšanu, kas atdala normālo atribūtu vērtības un šīm vērtībām atbilstošus trokšņus no atribūtu vērtībām, kas atbilst anomālijām. Ja atribūtu a_j vērtības tiek sadalītas saskaņā ar normālo sadalījuma likumu, tad atribūtu

iezīmes $|z_{ij}| \geq 3$ tiek uzskatītas par anomālijām. Citiem atribūtu vērtību sadalījumiem var iestatīt citas robežvērtības, piemēram, $|z_{ij}| \geq 2.5$ vai $|z_{ij}| \geq 2.7$ un tamlīdzīgi.

Vienkāršs ilustratīvs piemērs. 3.1. tabulā ir parādītas divu atribūtu vērtības: a_1 un a_2, kas uzdotas transponētā veidā. Mērķis ir pārbaudīt tabulas datus attiecībā uz anomāliju klātbūtni. Šim nolūkam atribūtu vērtības pārveido atbilstošās z iezīmēs. Aprēķinu rezultāti parādīti 3.2. un 3.3. tabulā.

3.1. tabula

Sākotnējās atribūtu a_1, a_2 vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
a_1	7	6	8	6	20	7	7	6	5	8
a_2	9	11	10	11	9	8	10	30	12	10

3.2. tabula

Atribūta a_1 vērtību transformācijas z iezīmēs rezultāti datiem no 3.1. tabulas

Objekti	a_{i1}	$a_{i1} - \bar{a}_1$	$(a_{i1} - \bar{a}_1)^2$	z_{i1}
o_1	7	-1	1	-0.244
o_2	6	-2	4	-0.488
o_3	8	0	0	0
o_4	6	-2	4	-0.488
o_5	20	12	144	2.930
o_6	7	-1	1	-0.244
o_7	7	-1	1	-0.244
o_8	6	-2	4	-0.488
o_9	5	-3	9	-0.732
o_{10}	8	0	0	0
	$\sum = 80$		$\sum = 168$	$\max z_{i1} = 2.930$
	$\bar{a}_1 = 8.000$		$s_1\sqrt{\frac{168}{10}} = 4.100$	

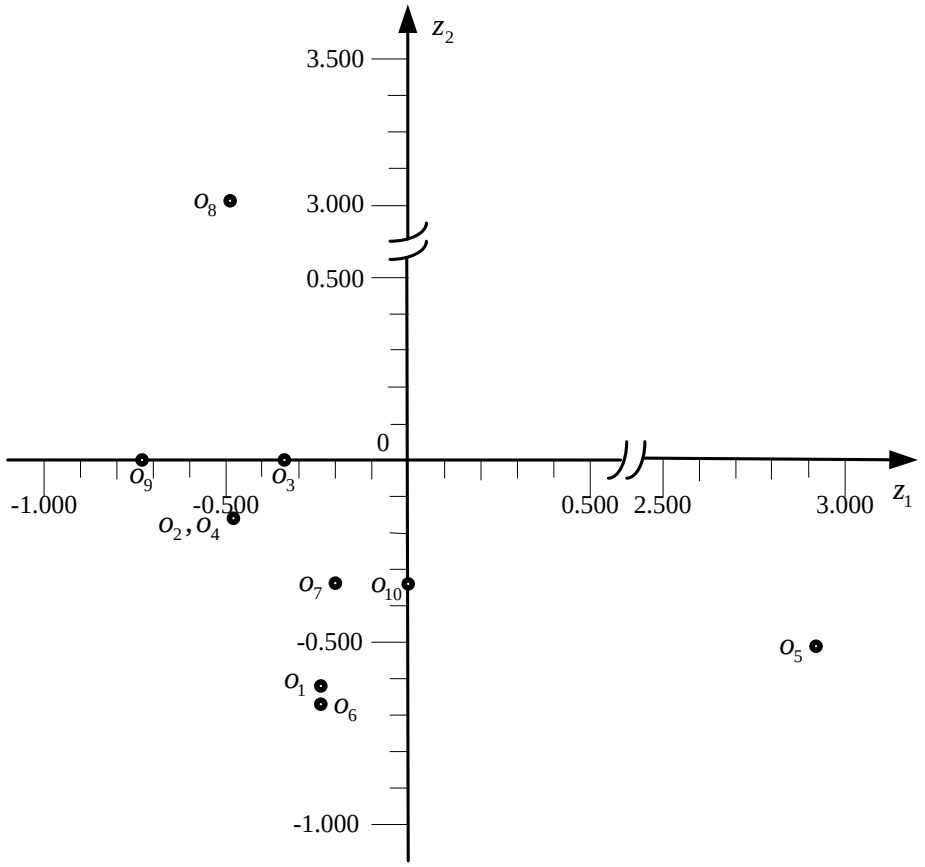
3.3. tabula

Atribūta a_2 vērtību transformācijas z iezīmēs rezultāti datiem no 3.1. tabulas

Objekti	a_{i2}	$a_{i2} - \bar{a}_2$	$(a_{i2} - \bar{a}_2)^2$	z_{i2}
o_1	9	- 3	9	-0.512
o_2	11	- 1	1	-0.171
o_3	10	-2	4	-0.342
o_4	11	- 1	1	-0.171
o_5	9	-3	4	-0.512
o_6	8	- 4	16	-0.6 83
o_7	10	- 2	4	-0.342
o_8	30	18	324	3.073

o_9	12	0	0	0
o_{10}	10	-2	4	-0.342
	$\sum = 120$		$\sum = 343$	$\max z_{i2} = 3.073$
	$\bar{a}_2 = 12.000$		$s_2 = \sqrt{\frac{343}{10}} = 5.857$	

Uzskatāmības labad 3.4. attēlā objektu punkti $o_1 - o_{10}$ tiek grafiski attēloti z_1, z_2 iezīmju telpā.



3.4. attēls. **Objektu punktu grafiskais attēlojums datiem no 3.1. tabulas z_1, z_2 iezīmju telpā**

Ja $z_{1g} = z_{2g} = 3.000$ iezīmju vērtības ņem kā robežvērtības, tad objektu a_1 ar iezīmes vērtību $z_{51} = 2.930$, kas tikai nedaudz atšķiras no robežvērtības, var traktēt kā anomāliju pēc atribūta o_5 vērtības. Atbilstoši iezīmes z_2 vērtībai anomālija noteikti ir objekts o_8, kuram $z_{82} = 3.073 > 3.000$.

Būtisks šīs metodes trūkums ir novērtējumu $\bar{a}_j$ un s_j atkarība no anomālijām datos. Jāatzīmē, ka šis trūkums ir raksturīgs visām pieejām, kur vidējās vērtības tiek izmantotas kā sadalījuma parametri un standarta novirzes tiek izmantotas kā diapazona parametri.

Šajā sakarā tiek piedāvāta modificētu z iezīmju metode.

2. Modificēto z iezīmju metode.

Ar šo metodi [Iglowitz B., Hoaglin D.C., 1993] atribūtu vērtību iezīmes tiek aprēķinātas, izmantojot vienādojumu:

$$z_{mij} = \frac{0.6745\left(a_{ij} - med\left(a_j\right)\right)}{mad\left(a_j\right)},\ i = 1,...,m,\ j = 1,...,n, \qquad (3.2)$$

kur $med\left(a_j\right)$ – j atribūta mediānas vērtība;

$mad\left(a_j\right) = med\left\{\left|a_{ij} - med\left(a_{ij}\right)\right|\right\}$ – j atribūta noviržu mediāna no vērtību mediānas.

Vērtības $med\left(a_j\right)$ un $mad\left(a_j\right)$ neietekmē anomāliju klātbūtne datos.

Ja uzdota vērtību $\mathbf{X} = \left\{x_1, x_2, ..., x_n\right\}$ kopa, tad, lai aprēķinātu mediānu $med\left(X\right)$, vērtības $x_i \in X$ jāsakārto nedilstošā secībā:

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n-1)} \le x_{(n)},$$

kur apakšindeksi iekavās apzīmē vērtību $x_{(i)}$ numurus sakārtotajā secībā.

Mediānas vērtību aprēķina šādi:

$$med\left(\mathbf{X}\right) = x_{(n+1)/2},\ n \text{ - nepāra;} \qquad (3.3.a)$$

$$med\left(\mathbf{X}\right) = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right),\ n = 2k \text{ - pāra.} \qquad (3.3.b)$$

Vērtību $z_{mijg} > 3.5$ ieteicams izmantot kā robežvērtību starp normālo datu kopu un anomāliju kopu.

Vienkāršs ilustratīvs piemērs. Par pamatu ņem atribūtu a_1 un a_2 sākotnējās vērtības no 3.1. tabulas. Aprēķina katra atribūta vērtību mediānu. Atribūta a_1 vērtības sakārto nedilstošā secībā:

$$\left\{a_{i1}\right\} = 5, 6, 6, 6, 7, 7, 7, 8, 8, 20.$$

Tā kā šajā sarakstā ir 10 vērtības $\left(n = 10\right)$, mediānas vērtību aprēķina, izmantojot vienādojumu (3.3.b):

$$med\left(a_1\right) = \frac{1}{2}\left(7 + 7\right) = 7.$$

Atribūta a_{i2} vērtības sakārto nedilstošā secībā:

$$\left\{a_{i2}\right\} = 8, 9, 9, 10, 10, 10, 11, 11, 12, 30.$$

$$med\left(a_2\right) = \frac{1}{2}\left(10 + 10\right) = 10.$$

Aprēķina absolūtās novirzes no atribūtu mediānas. Aprēķinu rezultāti parādīti 3.4. tabulā.

3.4. tabula

Atribūtu a_1, a_2 vērtību absolūtās novirzes no mediānas $med\left(a_1\right)$, $med\left(a_2\right)$

Objekti	a_{i1}	$\left\|a_{i1} - med\left(a_1\right)\right\|$	a_{i2}	$\left\|a_{i2} - med\left(a_2\right)\right\|$
o_1	7	0	9	1
o_2	6	1	11	1
o_3	8	1	10	0
o_4	6	1	11	1
o_5	20	13	9	1
o_6	7	0	8	2
o_7	7	0	10	0
o_8	6	1	30	20
o_9	5	2	12	2
o_{10}	8	1	10	0

Noviržu absolūtās vērtības no mediānas $med\left(a_1\right)$ sakārto nedilstošā secībā:

$$\left\{\left|a_{i1}-med\left(a_1\right)\right|\right\}=0,0,0,1,1,1,1,1,2,13.$$

Aprēķina vērtību $mad\left(a_1\right)$:

$$mad\left(a_1\right)=\frac{1}{2}(1+1)=1.$$

Noviržu absolūtās vērtības no mediānas $med\left(a_2\right)$ sakārto nedilstošā secībā:

$$\left\{\left|a_{i2}-med\left(a_1\right)\right|\right\}=0,0,0,1,1,1,1,2,2,20.$$

Aprēķina vērtību $mad\left(a_2\right)$:

$$mad\left(a_2\right)=\frac{1}{2}(1+1)=1.$$

Ar vienādojuma (3.2) palīdzību aprēķina modificēto iezīmju z_{mi1}, z_{mi2} vērtības. Aprēķinu rezultāti parādīti 3.5. tabulā.

3.5. tabula

Modificēto iezīmju z_{mi1}, z_{mi2} aprēķinātās vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
z_{mi1}	0	0.6745	0.6745	0.6745	8.7685	0	0	0.6745	1.3490	0.6745
z_{mi2}	0.6745	0.645	0	0.6745	0.6745	1.3490	0	13.4900	1.3490	0

Saskaņā ar 3.5. tabulu var izdarīt pamatotu secinājumu, ka objekts o_5 ir anomālija pēc atribūta a_1 vērtības un objekts o_8 ir anomālija pēc atribūta a_2 vērtības. Acīmredzami, ka anomāliju identificēšana datos, pamatojoties uz modificētām z iezīmēm, šķiet saprātīgāka un pareizāka nekā identificēšana, pamatojoties uz z iezīmēm. Šīs metodes galvenā priekšrocība ir tā, ka modificēto iezīmju aprēķinu rezultātus neietekmē ekstremālas atribūtu vērtības, kas var gadīties z iezīmju gadījumā.

3. Metode, kas balstīta uz starpkvartiļu rangiem.

Lai izskaidrotu šīs metodes būtību, jāaplūko daži statistikas jēdzieni. Par pamatu ņem sakārtoto atribūtu a_1 vērtības no 3.1. tabulas. Šīs vērtības sauc par kārtas statistikām.

Šeit ir kopsavilkums par dažiem svarīgiem jēdzieniem darbā ar kārtas statistikām.

Rangs augšup – tas ir i kārtas statistikas vērtības skaitlis, ja numerācija sākas no mazākās vērtības.

A	Kārtas statistikas numurs	1	2	3	4	5	6	7	8	9	10
B	Kārtas statistikas	5	6	6	6	7	7	7	8	8	20
C	Rangs augšup	1	2	3	4	5	6	7	8	9	10
D	Rangs lejup	10	9	8	7	6	5	4	3	2	1
E	Dziļums	1	2	3	4	5	5	4	3	2	1

Ranga augšup vērtības parādītas C rindā.

Rangs lejup – tas ir i kārtas statistikas vērtības skaitlis, ja numerācija sākas no lielākās vērtības (D rinda).

Kārtas statistikas a_{i1} rangu augšup apzīmē ar $r\uparrow\left(a_{i1}\right)$, bet ar $r\downarrow\left(a_{i1}\right)$ – šīs kārtas statistikas rangu lejup. Acīmredzot jebkurai kārtas statistikai a_{i1} ir spēkā attiecība:

$$r\uparrow\left(a_{i1}\right)+r\downarrow\left(a_{i1}\right)=n+1,$$

kur n – kopējais atribūtu a_1 vērtību skaits.

Ja saskaita divas vērtības no katra vērtību pāra rindās C un D, iegūst vērtību 11, kas apstiprina iepriekš minēto attiecību.

Sakārtoto vērtību mediānu aprēķina pēc vienādojumiem (3.3.a,b). Mediānu var definēt kā kārtas statistiku, kurai ir vislielākais dziļums. Iepriekš dotajos datos kārtas statistikām 5 un 6 vislielākais dziļums ir 5. Tāpēc mediānas vērtība tiek aprēķināta kā šīs kārtas statistikas vidējā vērtība:

$$med\left(a_1\right)=\frac{1}{2}\left(7+7\right)=7.$$

Formāli definē sakārtotas atribūtu a_1 vērtību kvartiles jēdzienu. Kvartile ir kārtas statistika, kuras dziļumu nosaka vienādojums:

$$g\left(quartile\right)=\left(\left\lfloor g\left(median\right)\right\rfloor+1\right)/2. \qquad (3.4)$$

Šajā vienādojumā $g\left(median\right)$ – mediānas dziļums. Apzīmējums $\lfloor x\rfloor$ nozīmē lielāko veselo skaitli, kas nepārsniedz x. Piemēram, $\lfloor 4.6\rfloor=4$.

Kvartiles dziļumu sakārtotām atribūtu a_1 vērtībām aprēķina, izmantojot vienādojumu (3.4):

$$g\left(quartile\right)=\left(5+1\right)/2=3.$$

Tas nozīmē, ka apakšējās kvartiles Q_1 vērtība ir kārtas statistika dziļumā 3, t. i., $Q_1=6$. Augšējās kvartiles vērtība ir kārtas statistika dziļumā 3, t. i., $Q_3=8$. Starpkvartiļu rangs ir vienāds ar starpību:

$$IQR=Q_3-Q_1=8-6=2.$$

Nosaka kvartiles un starpkvartiļu rangu sakārtotām atribūtu a_2 vērtībām no 3.1. tabulas.

A	Kārtas statistikas numurs	1	2	3	4	5	6	7	8	9	10
B	Kārtas statistikas	8	9	9	10	10	10	11	11	12	30
C	Rangs augšup	1	2	3	4	5	6	7	8	9	10
D	Rangs lejup	10	9	8	7	6	5	4	3	2	1
E	Dziļums	1	2	3	4	5	5	4	3	2	1

Kārtas statistikām 5 un 6 ir vienāds dziļums, kas vienāds ar 5. Tāpēc:

$$med\left(a_2\right)=\frac{1}{2}\left(10+10\right)=10.$$

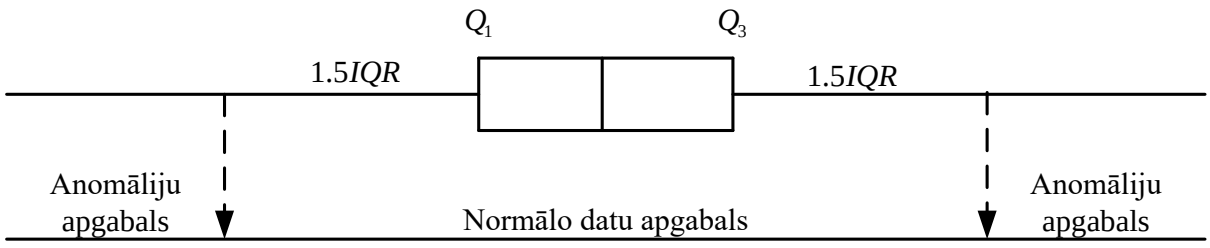
Kvartiļu dziļumi ir vienādi ar

$$g\left(quartile\right)=\left(5+1\right)/2=3.$$

Iegūst $Q_1=9$, $Q_2=11$.

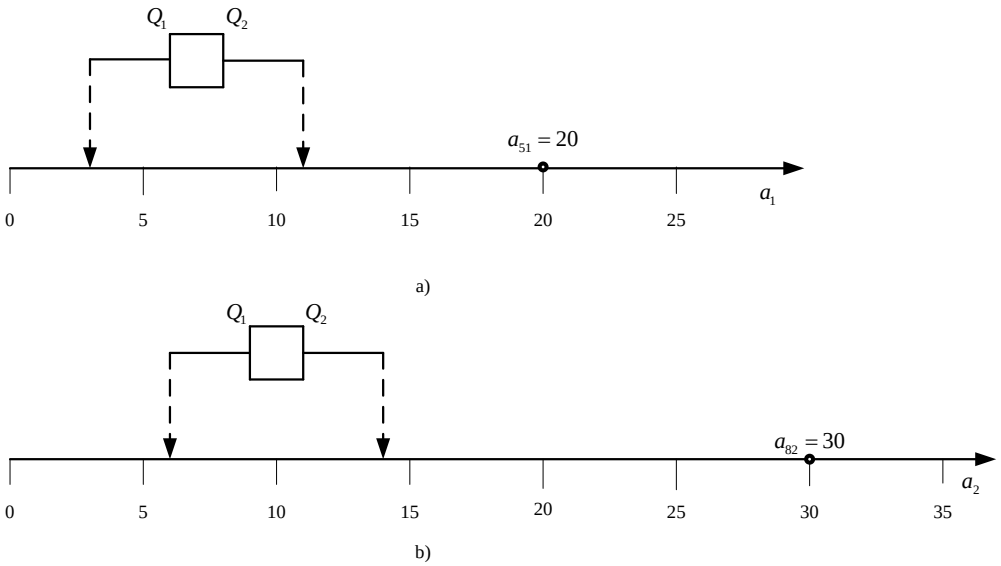
Starpkvartiļu rangs ir $IQR=Q_3-Q_1=11-9=2$.

Kā starpkvartiļu rangus var izmantot anomāliju identificēšanai datos? Robežas starp normāliem datiem un anomālijām nosaka, atņemot vērtību $1.5IQR$ no apakšējās kvartiles Q_1 vērtības un pievienojot vērtību $1.5IQR$ augšējai kvartilei Q_3. Uz šīs metodes balstītā anomāliju identifikācijas shēma ir parādīta 3.5. attēlā.



3.5. attēls. **Shēma robežu noteikšanai starp normāliem datiem un anomālijām, izmantojot starpkvartiļu rangus**

Nosaka normālo atribūtu a_1, a_2 vērtību apgabalu robežas datiem no 3.1. tabulas (sk. 3.6. attēlu).



3.6. attēls. **Normālo atribūtu a_1, a_2 robežu grafiskais attēlojums**

Normālām atribūtu a_1 vērtībām jāatrodas intervālā $[3,11]$. Atribūta $a_{51} = 20$ vērtība ir acīmredzama anomālija. Normālām atribūtu vērtībām a_2 jāatrodas intervālā $[7,14]$. Vērtība $a_{82} = 30$ ir acīmredzama anomālija.

Šīs metodes priekšrocība ir tās spēcīgā ietekme attiecībā uz atribūtu ekstrēmajām vērtībām. Mediānas un kvartiles vērtības tiek noteiktas pēc atribūta atrašanās vietas sakārtotā atribūtu secībā, tāpēc šī atribūta ekstremālās vērtības neietekmē šo parametru noteikšanu.

3.3. Anomāliju identificēšanas metodes uz tuvības mēra starp objektiem pamata

Statistiskās metodes anomāliju identificēšanai, kas aplūkotas 3.2. sadaļā, pieņem, ka atribūtu vērtības tiek sadalītas saskaņā ar noteiktiem (visbiežāk normāliem) sadalījuma likumiem. Statistikas metožu būtība ir novērtēt atribūtu vērtību sadalījumu parametrus. Anomāliju identifikācija tiek veikta, pamatojoties uz robežu noteikšanu starp normāliem datiem un anomālijām. Šīs robežas tiek noteiktas atkarībā no datu izvietojuma parametriem.

Reālās datu pirmapstrādes un analīzes problēmās atribūtu vērtību sadalījumi un šo sadalījumu parametri ir zināmi tikai diezgan retos gadījumos. Tas nozīmē, ka ir jāizmanto metodes, lai identificētu anomālijas, kas strādā ar visu datu kopumu, neņemot vērā faktisko atribūtu vērtību sadalījumu. Tā kā šie sadalījumi nav zināmi, tad nav zināms, kā to parametri tiek izmantoti par atskaites punktiem attāluma līdz objektiem novērtēšanai. Ņemot to vērā, ir nepieciešami citi principi tuvības mēru noteikšanai starp objektiem, lai šos tuvības mērus varētu izmantot kā pamatu anomāliju identificēšanai.

Ir piedāvātas daudzas metodes tuvības noteikšanai objektu kopai. Labi pazīstama šāda veida metode ir DB(p, D) – attāluma metode, kas tika piedāvāta darbā [Knorr E.M,. Ng R.T., 1998, 1999]. Tā ir balstīta uz attāluma mērīšanu starp visiem objektiem sākotnējā datu kopā. Pamatojoties uz konkrētu šo attālumu analīzi, datos var identificēt anomālijas.

Vēl viena izplatīta metode ir k-tuvāko (*k-nearest*) kaimiņu metode, kuras pamatā ir attālumi starp objektiem.

Anomāliju identificēšanas metodes, kuru pamatā ir objektu grupēšana datu kopā, ir kļuvušas ļoti izplatītas. Klasterizācija attiecas uz objektu grupēšanu klasteros, pamatojoties uz šo objektu līdzību. Jāņem vērā viens svarīgs apstāklis: grupējot tiek runāts par objektiem, kas nav iezīmēti jeb marķēti. Pretējā gadījumā visi objekti nav atšķirami tādā nozīmē, ka tie nepieder noteiktām klasēm vai kaut kādā veidā atšķiras.

Jebkuras klasterizācijas metodes galvenā īpašība ir spēja apmācīties, izmantojot sākotnējo datu kopu. Šajā ziņā klasterizāciju mašīnmācības jomā sauc par neuzraudzīto apmācību.

Lai novērtētu objektu līdzības pakāpi, atribūtu vērtību kopai ir nepieciešama piemērota metrika. Parasti šim nolūkam izmanto Eiklīda attālumu jeb distanci. Dažās īpašās situācijās tiek izmantots Mahalanobisa attālums (dažādu attālumu metrikas datu telpā sniegtas pielikumā P2).

Ideja par anomāliju identificēšanu, pamatojoties uz objektu klasterizāciju, ir šāda. Vispārīgā gadījumā starp atsevišķiem objektiem pastāv lielākas vai mazākas līdzības. Uz tā pamata objektus var apvienot atsevišķās datu kopās. No otras puses, potenciālās anomālijas var būt atsevišķi objekti, kas atrodas lielos attālumos no visām izveidotajām kopām vai arī anomālijas veido mazas kopas, kas atrodas attālu no lielajiem klasteriem. Klasterizācija ļauj strukturēt sākotnējos datus tā, lai lielas normālu objektu kopas krasi atšķirtos no atsevišķām anomālijām vai ļoti mazām anomāliju kopām. Šī normālu datu klasteru nošķiršana no potenciālajām anomālijām ir pamats šo anomāliju identificēšanai.

Šajā sadaļā tiks parādītas divas galvenās datu klasterizācijas metodes anomāliju identificēšanai:

- k-vidējā metode (*k-means*);
- k-medoīdu metode (*k-medoids*).

1. K-vidējā metode

K-vidējā metode ir paredzēta tikai skaitliskām atribūtu vērtībām. Tā kā atribūtu vērtībām ir dažādas mērvienības, lai piemērotu k-vidējā metodi, šīs vērtības ir jāpārveido bezdimensiju formā. Parasti pārveidošana notiek pārvēršot sākotnējās a_{ij} atribūtu vērtības par z iezīmēm z_{ij} pēc vienādojuma:

$$z_{ij} = \frac{a_{ij} - \overline{a}_j}{s_j}.\ i = 1,...,m.\ \ j = 1,...,n. \tag{3.5}$$

kur $\overline{a}_j$ – atribūta a_j vidējā vērtība;

s_j – atribūta a_j vērtības standarta novirze.

(Jāatzīmē, ka sākotnējo atribūtu vērtību z transformācija tika izmantota 3.2. sadaļā).

Šajā sadaļā objektu klasterizācijas, kuras pamatā ir k-vidējais, teorētiskie pamati ir izklāstīti darbā [Bishop Ch., 2006]. Detalizētu informāciju par dažādām klasterizācijas metodēm var atrast darbos [Wu J., 2012, Alekseeva L., Borisov A., Uzhga-Rebrov O., 2015].

Pieņem, ka ir m objekti, piemēri vai citas entītijas (datu vienības), kas novērtētas pēc n atribūtu vērtībām. Citiem vārdiem sakot, datu punkti atrodas n dimensiju atribūtu vērtību telpā.

Pirmā k-vidējā algoritma procedūra ir klasteru skaita noteikšana. Šo uzdevumu var atrisināt, pamatojoties uz iepriekšēju datu analīzi vai izmantojot šādas palīgmetodes:

- elkoņa metode (*elbow method*) [Torndike R.L., 1953].
- silueta metode (*silhouette method*) [Rousseeuw P.J., 1987].

Pieņem, ka ir noteikts klasteru skaits k. Izvēles veidā katram klasterim var piešķirt tā centrālo vērtību (centroīdu) $\boldsymbol{\mu}_k$. $k = 1,...,k$. Dažreiz vērtību $\boldsymbol{\mu}_k$ sauc par prototipu, kas saistīts ar k klasteri, $k = 1,...,k$. Var pieņemt, ka vektori $\boldsymbol{\mu}_k$ ir atbilstošo klasteru centri. Tad klasterizācijas problēmu, kas balstīta uz k-vidējā metodi, var formulēt šādi: jānosaka vektoru $\boldsymbol{\mu}_k$ kopa un visus datu objektus (punktus) piešķirt klasteriem tā, lai attālumu kvadrātu summa no jebkura objekta k klasterī līdz tā centram $\boldsymbol{\mu}_k$ būtu minimāla.

Lai formāli attēlotu objekta o_i saistību ar k klasteri, ievieš mainīgo $r_{ik} \in \{0,1\}$ ... $i \in \{1,...,m\}$, $k \in \{1,...,k\}$ kā bināro indikatoru kopu. Ja objekts o_i pieder klasterim k, tad $r_{ik} = 1$ un $r_{ik} = 0$ visām pārējām k vērtībām.

Pamatojoties uz šo formalizāciju, var definēt mērķa funkciju šādā veidā:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{k} r_{ik} \left\| \mathbf{a}_i - \boldsymbol{\mu}_k \right\|^2. \tag{3.6}$$

kur $\mathbf{a}_i$ – atribūtu vērtību vektors i objektam;

$\boldsymbol{\mu}_k$ – k klastera centrs;

r_{ik} – iepriekš definētais binārais indikatora mainīgais.

Vienādojums (3.6) attēlo attālumu kvadrātu summu no objektiem, kas pieder k klasterim līdz šī klastera centram $\boldsymbol{\mu}_k$. Uzdevums ir atrast r_{ik} vērtības $i=1,...,m$, $k=1,...,k$ un $\{\boldsymbol{\mu}_k\}$ tā, lai samazinātu mērķa funkciju (3.6). To var izdarīt ar iteratīvas procedūras palīdzību, kurā katra iterācija ietver divus secīgus soļus, kas atbilst r_{ik} un $\boldsymbol{\mu}_k$ vērtību secīgai optimizācijai.

Sākotnēji tiek atlasītas $\boldsymbol{\mu}_k$, $k=1,...,k$ vērtības. Pēc tam mērķa funkcija J tiek minimizēta attiecībā pret r_{ik} vērtībām, turklāt $\boldsymbol{\mu}_k$ vērtības ir fiksētas. Pēc tam $\boldsymbol{\mu}_k$ vērtības tiek pārrēķinātas, pamatojoties uz klasterim k attiecināmo objektu atribūtu vērtībām. Tad abas darbības tiek atkārtotas, līdz rezultāti konverģē ($\boldsymbol{\mu}_k$ vērtības ir vienādas ar iepriekšējā iterācijā iegūtajām vērtībām).

Var apgalvot, ka vērtību r_{ik} izmaiņas atbilst paredzamās vērtības maksimizēšanas algoritma solim E. (Šis algoritms tika parādīts 2.4. sadaļā iztrūkstošo atribūtu vērtību imputācijas kontekstā).

Vērtību r_{ik} optimizēšana nesastāda būtiskas problēmas. Tā kā mērķa funkcijas J vērtība (vienādojums (3.6)) ir lineāri atkarīga no r_{ik} vērtībām un r_{ik} vērtības pie dažādiem i ir neatkarīgas, var veikt optimizāciju katrai atsevišķai i vērtībai, izvēloties vērtību r_{ik} =1 jebkuram k, kas dod minimālo vērtību $\|\mathbf{a}_i-\boldsymbol{\mu}_k\|^2$. Faktiski objektu o_i vienkārši attiecina tuvākajam klasterim.

Formāli šo apgalvojumu var izteikt kā:

$$\begin{array}{l} 1.\ \text{Ja } k=\operatorname{argmin}\|\mathbf{a}_i-\boldsymbol{\mu}_k\|^2; \\ 0.\ \text{citādi.} \end{array} \qquad (3.7)$$

Tagad ievieš $\boldsymbol{\mu}_k$ optimizācijas principu pie fiksētām r_{ik} vērtībām. Mērķa funkcija J ir kvadrātiskā funkcija no $\boldsymbol{\mu}_k$ vērtībām, un to var minimizēt, iestatot tās pirmo atvasinājumu vienādu ar nulli attiecībā pret $\boldsymbol{\mu}_k$:

$$2\sum_{i=1}^{m} r_{ik}\left(\mathbf{a}_i-\boldsymbol{\mu}_k\right)=0.$$

Atrisinot šo vienādojumu attiecībā pret $\boldsymbol{\mu}_k$, iegūst:

$$\boldsymbol{\mu}_k=\frac{\sum_i r_{ik}\mathbf{a}_i}{\sum_i r_{ik}}.\ k=1,...,k. \qquad (3.8)$$

Šīs izteiksmes saucējs parāda klasterim attiecināmo datu punktu (objektu) skaitu k. Var secināt, ka vērtība μ_k, $k=1,...,k$ katrā iterācijā tiek aprēķināta kā klasterim attiecināto objektu atribūtu vektoru vidējā vērtība k. Līdz ar to arī ir šāds metodes nosaukums: klasterizācija uz k-vidējā pamata.

k-vidējā algoritma procedūru saraksts:

1) jānosaka klasteru skaits k;
2) jāatlasa k punktus atribūtu vērtību telpā un jāuzdod tos kā klasteru sākotnējos centroīdus;
3) jāizskaitļo attālumus no katra datu punkta (objekta) līdz katram no dotajiem centroīdiem, izmantojot Eiklīda attālumu;
4) jāattiecina katru objektu klasterim, kuram Eiklīda attālums starp objektu un dotā klastera centroīdu ir minimāls;
5) pārrēķina jaunās centroīdu koordinātes un izdara 4. punktā dotās darbības;
6) jāatkārto 3.–5. darbība, līdz centroīdu koordinātes vairs nemainās.

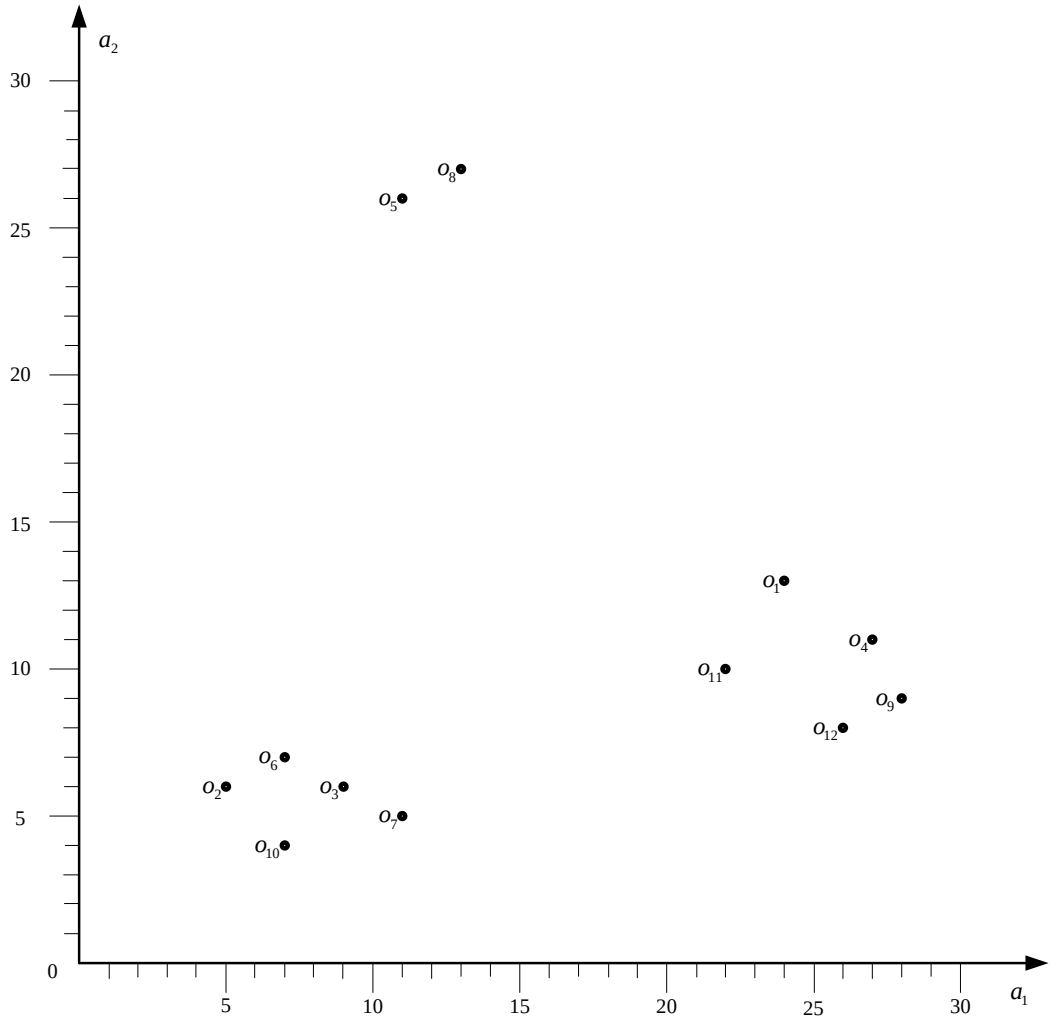
Vienkāršs ilustratīvs piemērs. Objektus o_1-o_{12} raksturo skaitlisko atribūtu vērtības a_1, a_2 (sk. 3.6. tabulu).

3.6. tabula

Objektu $o_1 - o_{12}$ sākotnējās atribūtu a_1, a_2 vērtības

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}
a_1	24	5	9	27	11	7	11	13	28	7	22	26
a_2	13	6	6	11	26	7	5	27	9	4	10	8

Uzskatāmības labad dati no 3.6. tabulas grafiski parādīti 3.7. attēlā.



3.7. attēls. **Sākotnējo datu no 3.6. tabulas grafiskais attēlojums**

Mērķis ir identificēt anomālijas sākotnējā datu kopā ar k-vidējā klasterizācijas algoritma palīdzību.

Sākotnējās atribūtu vērtības pārveido z iezīmju formā. Aprēķina atribūtu a_1 un a_2 vidējās vērtības un standartnovirzes. Iegūst:

$\bar{a}_1 = 15.83$. $s_1 = 8.44$;

$\bar{a}_2 = 11.00$. $s_2 = 7.36$.

Izmantojot vienādojumu (3.1), pārveido atribūtu a_1, a_2 sākotnējās vērtības. Pārveidošanas rezultāti ir parādīti 3.7. tabulā.

3.7. tabula

Pārveidotās atribūtu vērtības datiem no 3.6. tabulas

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}
z_1	0.97	-1.28	-0.81	1.32	-0.57	-1.05	-0.57	-0.33	1.44	-1.05	0.73	1.20
z_2	0.27	-0.68	-0.68	0	2.04	-0.54	-0.82	2.17	-0.27	-0.95	-0.14	-0.41

Uzskatāmības labad pārveidotās atribūtu z_1, z_2 vērtības ir grafiski parādītas 3.8. attēlā. Pamatojoties uz pārveidoto datu struktūru, iestata vērtību $k = 3$.

Uzdod klasteru centroīdu sākotnējās vērtības. Tā kā šajā piemērā datu struktūra ir tāda, ka objekti veido trīs grupas novērtēšanas telpā z_1, z_2, izvēlas centroīdu sākotnējās koordinātes aptuveni katras grupas centrā.

$c_1^{(o)}$: (-0.90. -0.80);

$c_2^{(0)}$: (1.10. -0.20);

$c_3^{(0)}$: (-0.50. 2.10).

3.8. attēlā sākotnējie centroīdi ir parādīti ar krustiņiem.

Aprēķina Eiklīda attālumus no katra objekta līdz katram centroīdam. Aprēķinu rezultāti ir parādīti 3.8. tabulā.

3.8. tabula

Eiklīda attāluma vērtības no objektiem līdz sākotnējiem klasteru centroīdiem

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}
c_1	4.67	0.15	0.02	5.57	1.65	0.09	0.11	9.14	5.75	0.04	3.10	4.5 5
c_2	0.24	5.89	3.88	0.09	7.81	4.74	3.17	7.66	0.13	5.28	0.17	0.05
c_3	5.51	2.63	7.83	7.72	0.01	7.27	8.53	0.03	9.38	9.60	6.53	9.19
Klasteris objektam	c_2	c_1	c_1	c_2	c_3	c_1	c_1	c_3	c_2	c_1	c_2	c_2

Rezultātā ir šādas kopas:

c_1: $\{o_2, o_3, o_6, o_7, o_{10}\}$;

c_2: $\{o_1, o_4, o_9, o_{11}, o_{12}\}$;

c_3: $\{o_5, o_8\}$.

Kā arī bija sagaidāms, klasteros nokļuva objekti, kas tika sagrupēti attiecībā pret uzdotajiem centroīdiem. Šajā gadījumā tas izskaidrojams ar vienkāršo datu struktūru šajā piemērā. Reālās datu priekšapstrādes problēmās objekti var pārvietoties no viena klastera citā, kad centroīdu koordinātes izmainās katrā algoritma iterācijā.

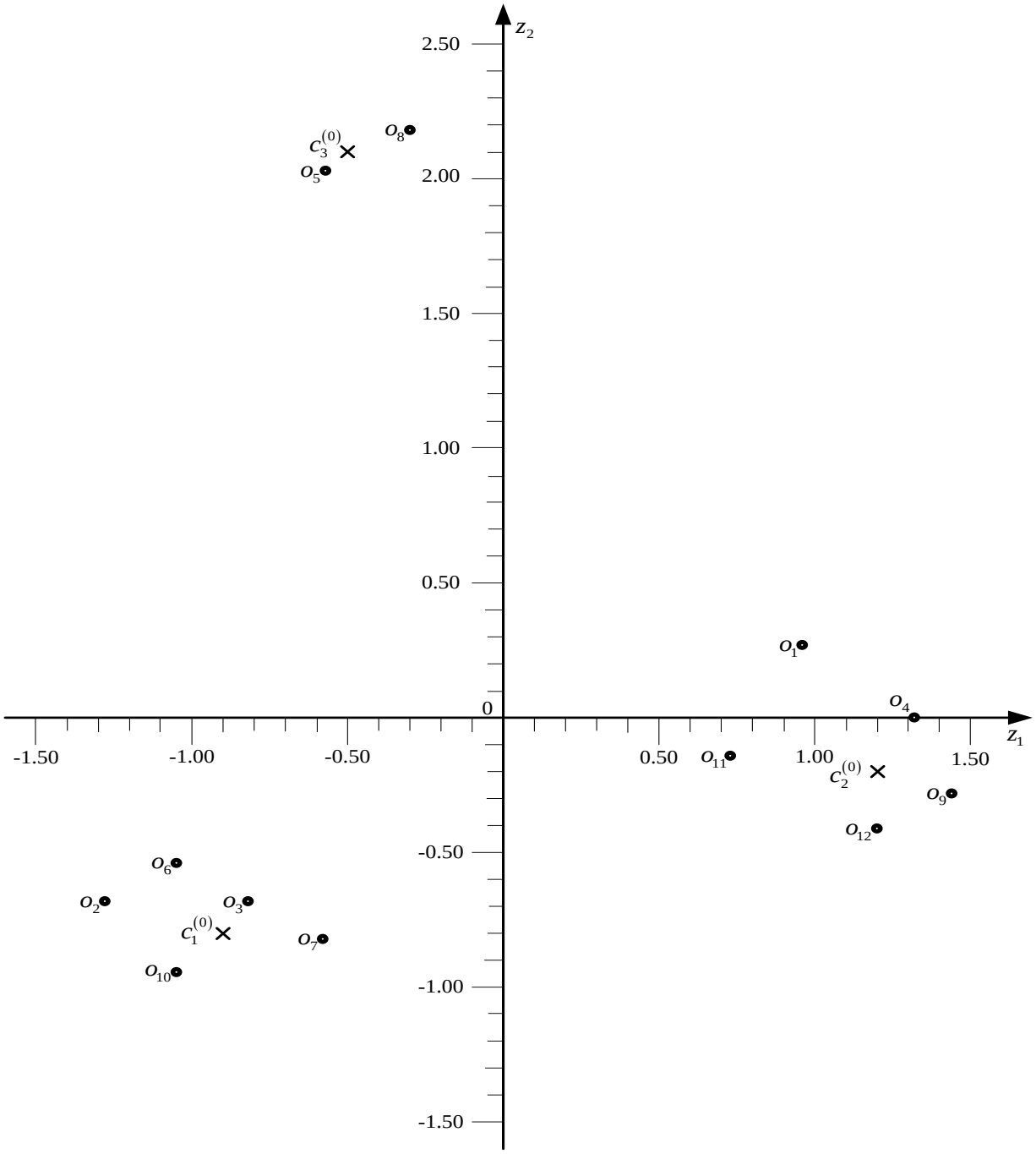
Atkārtoti pārvērtē klasteru centroīdu koordinātes kā atbilstošajos klasteros iekļauto objektu atribūtu transformēto vērtību vidējās vērtības:

- klasteris c_1:

$$\bar{z}_1\left(c_1^{(1)}\right) = \frac{-1.28 + (-0.81) + (-1.05) + (-0.57) + (-1.05)}{5} = \frac{-4.76}{5} = -0.95;$$

$$\bar{z}_2\left(c_1^{(1)}\right) = \frac{-0.68 + (-0.68) + (-0.54) + (-0.82) + (-0.95)}{5} = \frac{-3.67}{5} = -0.73.$$

$$c_1^{(1)} = (-0.95, -0.73).$$



3.8. attēls 3.3.2. **Datu kopas grafiskais attēlojums pārveidoto iezīmju telpā** z_1, z_2

- klasteris c_2:

  $\overline{z}_1\left(c_2^{(1)}\right)=\frac{0.97+1.32+1.44+0.73+1.20}{5}=\frac{5.66}{5}=1.13;$

  $\overline{z}_2\left(c_2^{(1)}\right)=\frac{0.27+0+(-0.27)+(-0.14)+(-0.41)}{5}=\frac{-0.55}{5}=-0.11.$

  $c_2^{(1)}=(1.13,-0.11).$

- klasteris c_3:

  $\overline{z}_1\left(c_3^{(1)}\right)=\frac{-0.57+(-0.33)}{2}=\frac{-0.90}{2}=-0.45;$

$$\bar{z}_2\left(c_3^{(1)}\right)=\frac{2.04+2.17}{2}=\frac{4.21}{2}=2.10.$$

$$c_3^{(1)}=\left(-0.45, 2.10\right).$$

Aprēķinātās jaunās $c_k^{(1)}$ centroīdu koordinātes maz atšķiras no sākotnējām $c_k^{(0)}$ koordinātēm. Nākamā algoritma iterācija netiks veikta tālāk norādītā iemesla dēļ. Ja aprēķinās attālumus no objektiem līdz jaunajiem centroīdiem $c_k^{(2)}$, tad šīs koordinātes sakritīs ar koordinātēm $c_k^{(1)}$. Tas nozīmē, ka ir panākta rezultātu konverģence. Reālās problēmās, lai panāktu konverģenci, var būt nepieciešams diezgan liels iterāciju skaits.

Analizējot iegūtos rezultātus, ir skaidrs, ka klasteros c_1 un c_2 ir normāli dati. Klasterī c_3 ir tikai divi objekti, kas atrodas attālu no citiem objektiem. Tāpēc ir nopietni iemesli apgalvot, ka objekti o_5, o_8 šajā datu kopā ir anomālijas.

Principā šādu secinājumu var izdarīt, pamatojoties uz 3.7. attēla grafiskā attēlojuma vizuālu analīzi. Formāli problēma tika atrisināta ar anomāliju identificēšanu šajos datos, lai sniegtu priekšstatu par klasterizācijas algoritma k-vidējais darbības principu.

Jāatzīmē šādas aplūkotā algoritma priekšrocības:

- algoritms ir vienkārši izpildāms;
- tas labi darbojas lielās datu kopās;
- labi pielāgojas jauniem objektiem.

Algoritma trūkumi ir:

- jūtība pret anomālijām;
- neveiksmīga klasteru skaita izvēle var novest līdz neapmierinošiem rezultātiem.

2. k-medoīdu metode

k-medoīdu metode ir klasiska klasterizācijas metode, kuras pamatā ir sākotnējās objektu kopas sadalīšana *apriori* noteiktā skaitā grupu (klasteru). Tas ir līdzīgs k-vidējā metodei ar dažām īpašām iezīmēm. k-vidējā metodē nejauši izvēlēti punkti atribūtu vērtību telpā tiek ņemti kā klasteru sākotnējie centroīdi. k-medoīdu metodē kā sākotnējie klasteru centri tiek atlasīti konkrēti objekti datu kopā (medoīdi). Klasteru skaits k ir noteikts *apriori*. Objekti tiek attiecināti uz klasteriem, pamatojoties uz aprēķiniem par attālumiem no medoīdiem līdz visiem objektiem datu kopā. Šī procedūra tiek veikta tieši tādā pašā veidā kā k-vidējā metodē.

Būtiskākā atšķirība starp abām metodēm ir centroīdu un medoīdu pārvērtēšana. Izmantojot k-vidējā metodi, jaunās centra koordinātes tiek aprēķinātas kā visu atribūtu vidējā vērtība visiem objektiem, kas iekļauti katrā klasterī. Izmantojot k-medoīdu metodi, katrā iterācijā par pašreizējo medoīdu tiek izvēlēts cits objekts un tiek aprēķināti attālumi no visiem objektiem klasterī līdz jaunajam medoīdam. Ja attālumu summa līdz jaunajam medoīdam ir mazāka nekā attālumu summa līdz iepriekšējam medoīdam, jaunais medoīds tiek uzskatīts par pašreizējo medoīdu. Ja attālumu summa līdz iepriekšējam medoīdam ir mazāka nekā attālumu summa līdz jaunajam medoīdam, iepriekšējo medoīdu uzskata par pašreizējo darba medoīdu. Process tiek atkārtots iteratīvi, līdz tiek identificēts medoīds, kuram ir vismazākā attālumu summa līdz objektiem noteiktā klasterī salīdzinājumā ar visiem citiem potenciālajiem medoīdiem.

k-medoīdu metode pirmo reizi tika piedāvāta darbā [Kaufman L., Rousseeuw P., 1987] kā k-vidējā metodes uzlabota versija PAM (*Partitioning Around Medoid*). Vēlāk tika piedāvātas uzlabotas šī algoritma versijas, piemēram, CLARA algoritms.

Formāli k-medoīdu algoritma mērķa funkcija tiek izteikta formā:

$$s_k=\arg\min\sum_{a_{ij}\in c_k}\left(a_{ij}-m_k\right)^2, \qquad (3.9)$$

kur c_k, $k=1,...,k-$ k klasteris;

m_k- k klastera medoīds.

k -medoīdu algoritms ietver šādu procedūru secību:

1) jānosaka klasteru skaits k;
2) jāatlasa punkti atribūtu vērtību telpā un jāuzdod tie kā klasteru sākotnējie centroīdi;
3) jāizskaitļo attālumi no katra datu punkta (objekta) līdz katram no dotajiem centroīdiem, izmantojot Eiklīda attālumu;
4) jāattiecina katrs objekts klasterim, kuram Eiklīda attālums starp objektu un dotā klastera centroīdu ir minimāls;
5) pārrēķina jaunās centroīdu koordinātes un izdara 4. punktā dotās darbības;
6) jāatkārto 3.–5. darbība, līdz centroīdu koordinātes vairs nemainās.

Piemērs. Par pamatu ņem sākotnējos normalizētos datus no 3.7. tabulas. Izvēlas klasteru skaitu $k = 3$ un sākotnējos medoīdus klasteriem c_1, c_2, c_3: $m_1^{(0)} = o_3$, $m_2^{(0)} = o_4$, $m_3^{(0)} = o_5$.

Objektu piederību klasteriem nenosaka. Šī procedūra tika veikta iepriekšējā piemērā. Datu struktūra šajā piemērā ir tāda, ka, nosakot objektu piederību klasteriem, pamatojoties uz jebkuriem sākotnējiem medoīdiem objektu grupās (3.8. attēls), iegūst tieši tādu pašu objektu sadalījumu klasteros, kas tika iegūti iepriekšējā piemērā (šis apgalvojums, protams, attiecas tikai uz šo piemēru).

Aprēķina attālumus no objektiem kopās līdz sākotnējiem medoīdiem un šo attālumu summas. Aprēķinu rezultāti ir parādīti 3.9. tabulā.

3.9. tabula

Eiklīda attālumi no objektiem līdz sākotnējiem medoīdiem $m_1^{(0)}$ un $m_2^{(0)}$ un šo attālumu summas

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}	s_k
$m_1^{(0)} = 0_3$		0.22	0			0.08	0.08			0.13			0.51
$m_2^{(0)} = 0_4$	0.20			0					0.08		0.37	0.19	0.84
$m_3^{(0)} = 0_5$					0			0.08					0.08

Tabulas pēdējā kolonnā ir norādītas mērķa funkciju vērtības kā attālumu summas pa attiecīgajām tabulas rindām.

Medoīdus uzdod šādi:

$c_1 : m_1^{(1)} = o_6 \ (-1.05, -0.54)$;

$c_2 : m_2^{(1)} = o_1 \ (0.97, 0.27)$.

Klasterim c_3 nav jēgas veikt medoīdu maiņu. Šajā klasterī ir tikai divi objekti. Par sākotnējo medoīdu tika izvēlēts objekts o_5. Ja tagad par pašreizējo medoīdu izvēlas objektu o_8, iegūst tādu pašu attālumu no objekta o_5 līdz medoīdai o_8, kas tika iegūts iepriekš.

Formāli izpildot algoritmu pēc jaunu medoīdu piešķiršanas, ir nepieciešams no jauna noteikt attālumus no visiem objektiem līdz šiem medoīdiem. Šī procedūra ir nepieciešama, jo medoīdu maiņa var izraisīt dažu objektu attiecināšanu uz citiem klasteriem. Šī procedūra netiek veikta, jo datu struktūra šajā piemērā ir tāda, ka, ņemot jebkuru objektu klasteros c_1, c_2 par medoīdu, iepriekš definētie klasteri netiks izmainīti.

Attālumu aprēķinu rezultāti no objektiem klasteros c_1, c_2 līdz jaunajiem medoīdiem $m_1^{(1)}$, $m_2^{(1)}$ ir parādīti 3.10. tabulā.

3.10. tabula

Eiklīda attāluma vērtības no objektiem klasteros c_1, c_2 līdz pašreizējiem medoīdiem $m_1^{(1)}$, $m_2^{(1)}$

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}	s_k
$m_1^{(1)} = o_6$		0.07	0.08		-	0	0.31	-		0.17			0.63
$m_2^{(1)} = o_1$	0			0.19	-			-	0.22		0.23	0.51	1.15

Var secināt, ka iepriekšējie rezultāti (3.9. tabulā) ir labāki par pašreizējiem rezultātiem (3.10. tabulā) tādā ziņā, ka tiem ir mazākas mērķa funkciju $s_k, k = 1,2$ vērtības. Klasterī c_3 izvēloties kādu no objektiem o_5 vai o_8 par medoīdu, tiks iegūti tādi paši rezultāti.

Iterācijas c_1 netiks atkārtotas. No 3.8. attēla vizuālas analīzes var secināt, ka jebkura objekta o_2, o_7 vai o_{10} izvēle par medoīdu dod lielus attālumus no objektiem klasterī līdz potenciālajam medoīdam. Tāpēc šīs iterācijas kļūst bezjēdzīgas.

Nākamo iterāciju veic klasterim c_2. Par pašreizējo medoīdu izvēlas objektu o_9. Aprēķinu rezultāti ir parādīti 3.11. tabulā.

3.11. tabula

Eiklīda attālumu vērtības no objektiem klasterī c_2 līdz pašreizējai medoīdai $m_2^{(2)} = o_9$

Objekti	o_1	o_4	o_9	o_{11}	o_{12}	s_2
$m_2^{(2)} = o_9$	0.22	0.08	0	0.52	0.08	0.90

Turpmākās iterācijas klasterī c_2 netiks veiktas. 3.8. attēla vizuālā analīze parāda, ka, izvēloties objektus o_{11} vai o_{12} par pašreizējo medoīdu, iegūst attālumu summu, kas ir lielāka nekā medoīdam $m_2^{(0)}$.

Apkopojot iegūst šādus rezultātus:

- klasteris c_1 ar objektu o_2 kā medoīdu;
- klasteris c_2 ar objektu o_4 kā medoīdu;
- klasteris c_3 ar objektiem o_5, o_5 vai o_8 kā medoīdu.

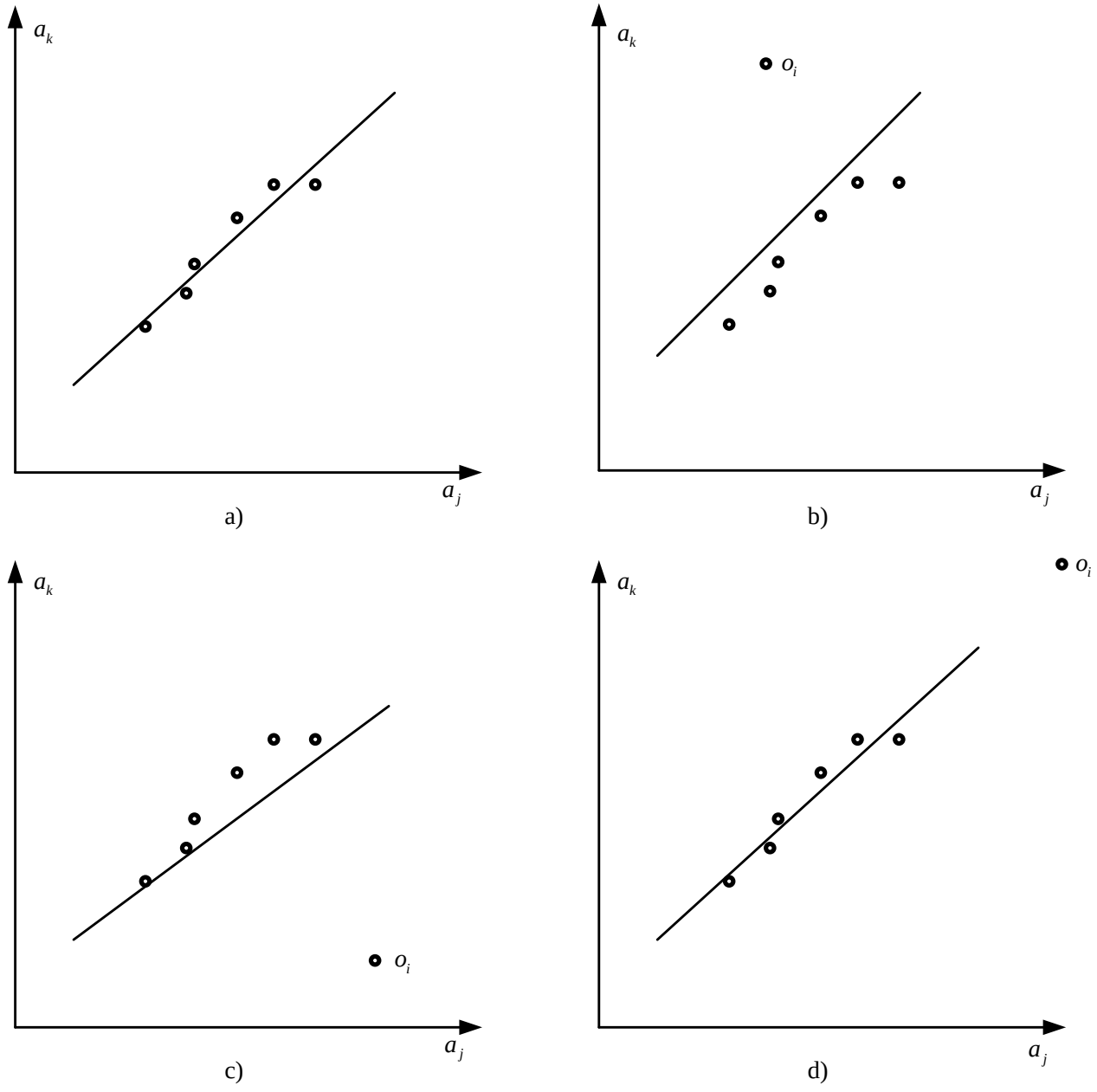
Anomāliju identifikācija, pamatojoties uz k-medoīdu klasterizācijas metodi, tiek veikta tieši tāpat kā anomāliju identifikācija ar k-vidējā klasterizācijas metodi. Šajā piemērā anomālijas ir objekti o_5, o_8.

3.4. Anomāliju identificēšana lineārajā regresijā

Pieņem, ka ir divas atribūtu a_j, a_k vērtību kopas un uzdevums ir izveidot regresijas līkni, kas parāda attiecības starp atribūtu vērtībām. Tas ir triviāls uzdevums normālu datu gadījumā, t. i., datiem bez anomālijām. Bet, ja datos ir anomālijas, tās var būtiski ietekmēt regresijas attiecību raksturu.

3.9.a attēlā ir parādīta regresijas līkne, kas konstruēta uz datiem bez anomālijām. 3.9.b attēlā ar tiem pašiem datiem papildus ir punkts o_i, kuram ir liela atribūta a_k vērtība. Šis punkts var būt

potenciāla anomālija. Šī punkta dēļ ir mainījusies regresijas līkne, tā atrodas virs datu punktiem un tāpēc neatbilst normālo datu daļai. 3.9.c attēlā pie tiem pašiem datiem papildus ir punkts o_i, ko literatūrā mēdz dēvēt par sviras vai līdzsvara punktu. Šim punktam ir augsta atribūta a_j vērtība, un tas arī varētu būt potenciāla anomālija. Šī iemesla dēļ regresijas līkne atrodas zem galvenajiem datu punktiem un īsti neatbilst normāliem datiem. 3.9.d attēlā anomālija ir objekts o_i, kuram ir lielas abu atribūtu vērtības. Neņemot vērā to, ka objekta punkts o_i atrodas uz regresijas līknes, šis objekts ir anomālija, jo tas atrodas tālu no lielākās datu daļas.



3.9. attēls. **Anomāliju ietekmes uz regresijas atkarību grafisks attēlojums**

Ir piedāvātas daudzas metodes anomāliju identificēšanai lineārajā regresijā. Tās var iedalīt statistiskajās un robustajās (*robust*) metodēs.

1. Statistikas metodes

Lai lineārajā regresijā varētu identificēt anomālijas, ir nepieciešams noteikt regresijas vienādojumu sākotnējiem datiem.

Detalizēts regresijas līknes noteikšanas un konstruēšanas procedūru apraksts ir sniegts pielikumā P1.3. Regresijas vienādojums ir dots formā:

$$y = b_0 + b_1 x, \tag{3.10}$$

kur x – faktoriālā pazīme (faktoriālais mainīgais);

y – rezultatīvā pazīme (rezultatīvais mainīgais).

Regresijas koeficientu vērtības aprēķina, izmantojot šādus vienādojumus:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.11}$$

kur $\bar{x}$ – faktoriālā mainīgā vidējā vērtība x;

$\bar{y}$ – rezultatīvā mainīgā vidējā vērtība y;

$$b_0 = \bar{y} - b_1 \bar{x}. \tag{3.12}$$

Koeficienta b_1 vērtība nosaka regresijas līknes slīpumu attiecībā pret horizontālo asi x. Parametra b_0 vērtība nosaka regresijas līknes krustošanās punktu ar vertikālo y asi.

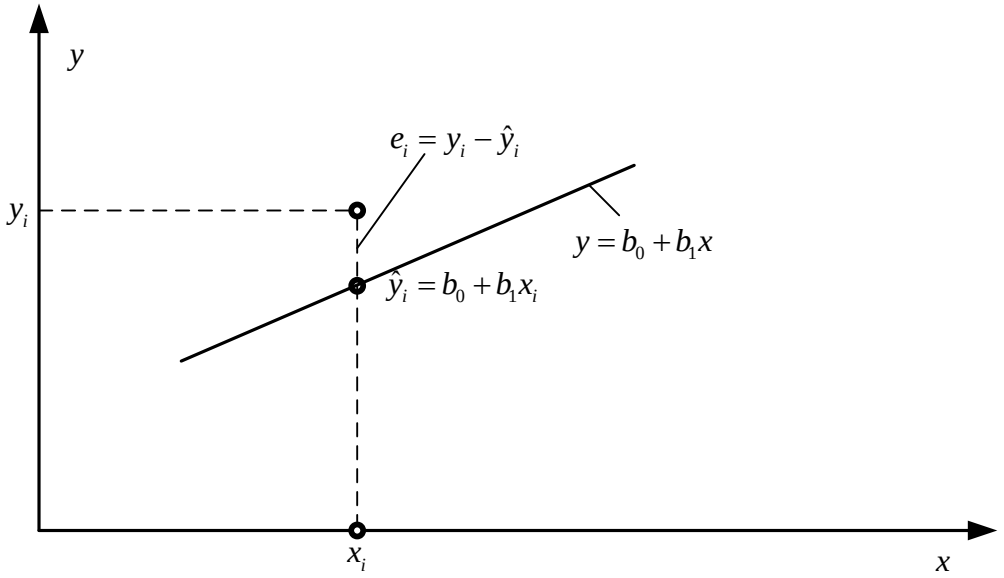
Uzdodot divas mainīgā x vērtības, tiek aprēķinātas atbilstošās rezultatīvā mainīgā y vērtības, izmantojot vienādojumu (3.10), tādējādi var izveidot regresijas līkni.

3.10. attēlā shematiski attēloti datu punkti (x_i, y_i) un konstruētā regresijas līkne $y = b_0 + b_1 x$.

Mainīgajiem x un y būs lielākas vai mazākas novirzes no regresijas līknes. Punkts $\hat{y}_i$ šajā attēlā atbilst vērtībai y, kas aprēķināta ar regresijas vienādojumu pie vērtības $x = x_i$. Atšķirība

$$e_i = (y_i - \hat{y}_i),\ i = 1,...,n \tag{3.13}$$

parāda datu punktu y_i novirzes no regresijas līknes (kļūdas).



3.10. attēls. **Datu punktu un regresijas līknes shematisks attēlojums**

Izmantojot sākotnējo datu punktu novirzes no regresijas līknes, pilnu regresijas modeli var attēlot formā:

$$y = b_0 + b_1 x + \varepsilon, \tag{3.14}$$

kur ε – kļūdas lielums, kas parāda datu punktu novirzes no regresijas līknes.

Standarta statistika izdara šādus pieņēmumus par kļūdas lielumu:

1) ε ir nulles vidējais rādītājs;
2) ε ir nemainīga variācija σ^2;
3) individuālās novirzes (kļūdas) ir nekorelēti gadījuma lielumi;
4) novirzes (kļūdas) ε_i atbilst normālajam sadalījumam.

Lai regresijā identificētu anomālijas, ir jādefinē regresijas vienādojums, jākonstruē regresijas līkne un jānosaka datu punktu novirzes ε_i no regresijas līknes.

Viena no statistiskajām metodēm anomāliju identificēšanai regresijā ir šāda. Tiek aprēķinātas standartnoviržu vērtības:

$$d_i = \frac{e_i}{\sqrt{MS_{RES}}}.\ i = 1,...,n, \tag{3.15}$$

kur e_i aprēķina pēc vienādojuma (3.13);

MS_{res} – aprēķinātā noviržu kvadrātu vidējā vērtība pēc vienādojuma:

$$MS_{RES} = \frac{\sum_{i=1}^{n} e_i^2}{n}. \tag{3.16}$$

Ja kādam objektam o_i $d_i \geq 3$, šo objektu var identificēt kā potenciālu anomāliju.

Ilustratīvs piemērs. 3.12. tabulā ir parādītas objektu $o_1 - o_{10}$ atribūtu vērtības a_1, a_2.

3.12. tabula

Objektu $o_1 - o_{10}$ sākotnējās atribūtu a_1, a_2 vērtības

Objekti	a_1	a_2
o_1	4.00	4.50
o_2	7.00	6.50
o_3	10.00	8.50
o_4	11.00	20.00
o_5	15.00	10.00
o_6	6.00	5.50
o_7	14.00	10.00
o_8	8.00	6.50
o_9	12.00	8.00
o_{10}	13.00	8.50
	$\sum = 100.00$, $\bar{a}_1 = 10.00$	$\sum = 88.00$, $\bar{a}_2 = 8.80$

Skaidrības labad dati no 3.12. tabulas ir grafiski parādīti 3.11. attēlā. Šis datu kopums ir jāanalizē, lai tajā identificētu anomālijas.

3.12. tabulas pēdējā rindā ir norādītas atribūtu a_1, a_2 vērtību summas attiecīgajās kolonnās un vidējās atribūtu vērtības: $\bar{a}_1 = 10.00$, $\bar{a}_2 = 8.80$.

Veic starpaprēķinus, lai noteiktu regresijas koeficientus b_o un b_1. Aprēķinu rezultāti parādīti 3.13. tabulā.

Izmantojot vienādojumu (3.11), aprēķina:

$$b_1 = \frac{70.00}{120.00} = 0.58.$$

Pēc vienādojuma (3.7):

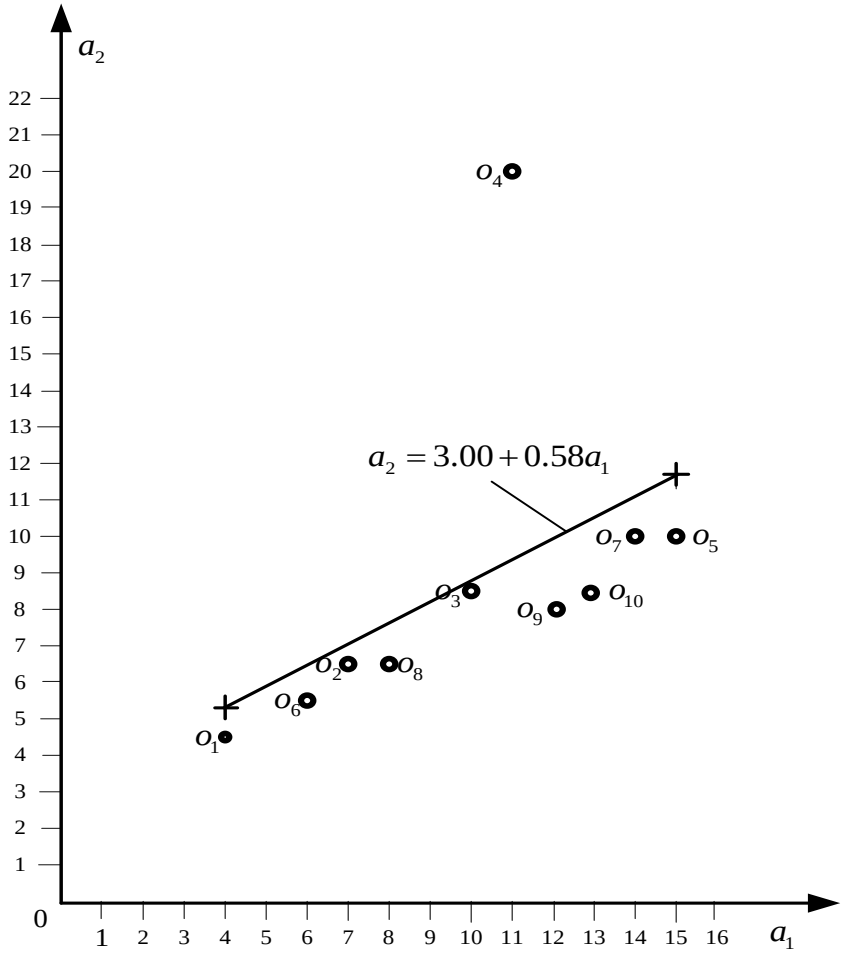
$b_0 = 8.80 - 0.58 * 10 = 8.80 - 5.80 = 3.00$.

Pieņem, ka $a_{11} = 4.00$.

$\hat{a}_{12} = 3.00 + 0.58 * 4.00 = 5.32$.

Pieņem, ka $a_{51} = 15.00$.

$\hat{a}_{52} = 3.00 + 0.58 * 15.00 = 11.70$.



3.11. attēls. **3.12. tabulas datu grafiskais attēlojums**

3.13. tabula

Regresijas koeficientu b_0, b_1 noteikšanas rezultāti

Objekti	$a_{i1} - \bar{a}_1$	$(a_{i1} - \bar{a}_1)^2$	$a_{i2} - \bar{a}_2$	$(a_{i1} - \bar{a}_1) * (a_{i2} - \bar{a}_2)$
o_1	-6.00	36.00	-4.30	25.00
o_2	-3.00	9.00	-2.30	6.90
o_3	0.00	0.00	-0.30	0.00
o_4	1.00	1.00	11.20	11.20
o_5	5.00	25.00	1.20	6.00
o_6	-4.00	16.00	-3.30	13.20
o_7	4.00	16.00	1.20	4.80
o_8	-2.00	4.00	-2.30	4.60
o_9	2.00	4.00	-0.80	-1.60
o_{10}	3.00	9.00	-0.30	-0.90
		$\sum = 120.00$		$\sum = 70.00$

Izmantojot izskaitļotās punktu vērtības, konstruē regresijas līkni. Šī līkne ir parādīta 3.11. attēlā. Redzams, ka šī līkne slikti atbilst sākotnējiem datiem, jo tā atrodas virs lielākās datu daļas.

Aprēķina datu punktu standartizētās novirzes no konstruētās regresijas līknes. Starpaprēķinu rezultāti ir parādīti 3.14. tabulā.

3.14. tabula

Starprezultātu aprēķinu vērtības standartizētas novirzes d_{i2} noteikšanai

Objekti	a_{i2}	$\hat{a}_{i2}$	$e_{i2} = a_{i2} - \hat{a}_{i2}$	e_{i2}^2
o_1	4.50	5 . 32	-0.82	0.672
o_2	6.50	7.06	-0.56	0.314
o_3	8.50	8.80	-0.30	0.09 0
o_4	20.00	9.38	10.62	112.784
o_5	10.00	11.70	-1.70	2.890
o_6	5.50	6.48	-0.98	0.960
o_7	10.00	11.12	-1.12	1.254
o_8	6.50	7.64	-1.14	1300
o_9	8.00	9.96	-1.96	3.842
o_{10}	8.50	10.54	-2.04	4.162
				$\sum = 128.268$

Pēc vienādojuma (3.16):

$MS_{RES} = \frac{128.268}{10} = 12.827$.

$\sqrt{MS_{RES}} = \sqrt{12.827} = 3.581$.

Izmantojot vienādojumu (3.15) nosaka standartizētās novirzes vērtības. Aprēķinu rezultāti parādīti 3.15. tabulā.

3.15. tabula

Standartizētās novirzes vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
d_{i2}	-0.230	-0.156	-0.083	2.966	-0.475	-0.274	-0.313	-0.318	-0.577	-0.570

Vērtība $d_{42} = 2.966 \approx 3$ ļauj secināt, ka objekts o_4 ir anomālija.

Lai parādītu anomālijas ietekmi uz regresijas kvalitāti, šo anomāliju (objektu o_4) izņem no sākotnējās datu kopas. Samazinātā datu kopa ir parādīta 3.16. tabulā.

3.16. tabula

Samazināta datu kopa no 3.12. tabulas

Objekti	a_1	a_2
o_1	4.00	4.50
o_2	7.00	6.50
o_3	10.00	8.50
o_5	15.00	10.00
o_6	6.00	5.50
o_7	14.00	10.00
o_8	8.00	6.50
o_9	12.00	8.00
o_{10}	13.00	8.50
	$\sum = 89.00$, $\bar{a}_1 = 9.89$	$\sum = 68.00$, $\bar{a}_2 = 7.56$

3.17. tabulā parādīti starpaprēķinu rezultāti regresijas koeficientu b_0, b_1 noteikšanai.

Pēc vienādojuma (3.11):

$b_1 = \frac{57.44}{118.87} = 0.48$.

Pēc vienādojuma (3.12):

$b_0 = 7.56 - 0.48 * 9.89 = 2.81$.

Tādējādi ir šāds regresijas vienādojums:

$a_2 = 2.81 + 0.48 * a_1$.

Pieņem, ka $a_{11} = 4.00$.

$a_{12} = 2.81 + 0.48 * 4.00 = 4.73$.

Pieņem, ka $a_{52} = 15.00$.

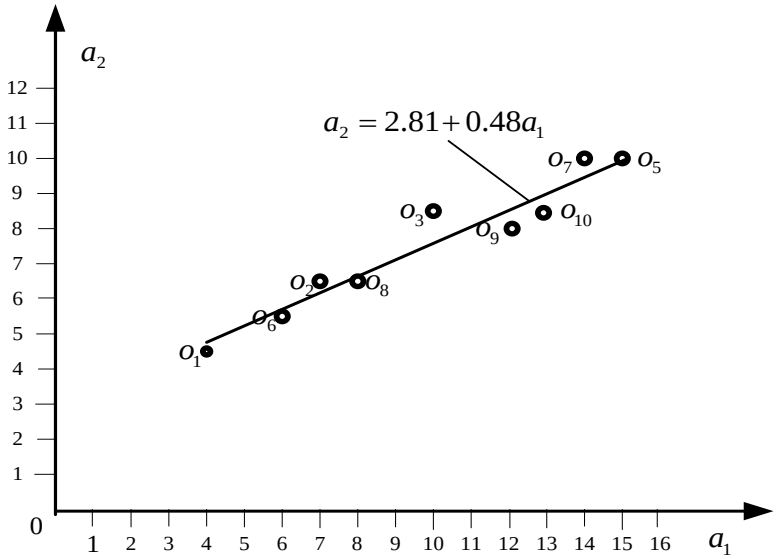
$a_{52} = 2.81 + 0.48 * 15.00 = 10.01$.

3.17. tabula

Regresijas koeficientu b_0, b_1 noteikšanas starpaprēķinu rezultāti

Objekti	$a_{i1} - \bar{a}_1$	$(a_{i1} - \bar{a}_1)^2$	$a_{i2} - \bar{a}_2$	$(a_{i1} - \bar{a}_1) * (a_{i2} - \bar{a}_2)$
o_1	-5.89	34.69	-3.05	18.02
o_2	-2.89	8.35	-1.06	3.06
o_3	0.11	0.01	0.94	0.10
o_5	5.11	26.11	2.44	12.47
o_6	-3.89	15.13	-2.06	8.01
o_7	4.11	16.89	2.41	10.03
o_8	-1.89	3.57	-1.06	2.00
o_9	2.11	4.45	0.44	0.93
o_{10}	3.11	9.67	0.94	2.02
		$\sum = 118.87$		$\sum = 57.54$

3.12. attēls grafiski attēlo sākotnējo samazināto datu kopu un regresijas līkni, kas izveidota uz šīs kopas pamata.



3.12. attēls. **Samazinātās datu kopas grafiskais attēlojums un konstruētā regresijas līkne**

Vienkārša vizuālā analīze parāda, ka regresijas līkne 3.12. attēlā labi atbilst datiem atšķirībā no regresijas līknes 3.11. attēlā, kas slikti atbilst datiem, jo tad datos bija anomālija.

Var redzēt, ka regresijas līknes 3.11. un 3.12. attēlā ir līdzīgas regresijas līknēm 3.9.a,b attēlā, kas shematiski attēlo regresijas līknes bez anomālijām datos un ar anomālijas punktu.

Alternatīva anomāliju identificēšanai regresijā ir izmantot studentizētās novirzes vērtības (*studentized residual*). Pieņem, tāpat kā iepriekš, $e_i = y_i - \hat{y}_i$, kur y_i ir rezultatīvā vērtība mainīgā x_i i vērtībai, $\hat{y}_i$ – rezultatīvā vērtība, kas noteikta no regresijas vienādojuma.

Ievieš mainīgo h_i, ko aprēķina pēc vienādojuma:

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.\ i = 1,...,n, \tag{3.17}$$

kur n – faktoriālā mainīgā vērtību skaits;

x_i – faktoriālā mainīgā i vērtība;

$\bar{x}$ – faktoriālā mainīgā vidējā vērtība.

Faktiski vērtība h_i ir attāluma novērtējums no vērtības x_i līdz vidējai vērtībai $\bar{x}$.

Ievieš vēl vienu mainīgo S_E:

$$S_E = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}, \tag{3.18}$$

kur $e_i = y_i - \hat{y}_i$ – punkta y_i novirze no regresijas līknes;

n – rezultatīvā mainīgā y vērtību skaits;

k – faktoriālo mainīgo skaits (šajā gadījumā $k = 1$).

Faktiski S_E ir summētā novirze visā noviržu kopumā $\{e_i\}$ no regresijas līknes.

Studentizētās i novirzes vērtību aprēķina pēc vienādojuma:

$$r_i = \frac{e_i}{S_E\sqrt{1-h_i}}, \tag{3.19}$$

kur h_i aprēķina pēc vienādojuma (3.17);

S_E aprēķina, izmantojot vienādojumu (3.18).

Rodas jautājums par robežvērtības r noteikšanu. Dažādi literatūras avoti piedāvā dažādas robežvērtības. Šajā sadaļā nosaka robežvērtību $r = 2$. Tas nozīmē, ka objektu o_i, kuram $r_i \geq 2$ var uzskatīt par potenciālu anomāliju datos.

Aprēķina h_i iepriekšējā piemēra datiem. Par pamatu ņem $\left(a_{i1} - \bar{a}\right)^2$ un $\sum_{i=1}^{10}\left(a_{i1} - \bar{a}_1\right)^2$ vērtības 3.13. tabulas trešās kolonnas. Šīs vērtības ir parādītas 3.18. tabulas 2. rindā.

Izmantojot vienādojumu (3.17), aprēķina h_i vērtības. Šīs vērtības ir parādītas 3.18. tabulas 3. rindā.

3.18. tabula

$\left(a_{i1} - \bar{a}\right)^2$ un $\sum_{i=1}^{10}\left(a_{i1} - \bar{a}_1\right)^2$ vērtības no 3.18. tabulas un aprēķinātās h_i vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	$\sum_{i=1}^{10}\left(a_{i1} - \bar{a}_1\right)^2$
$a_{i1} - \bar{a}_1$	36.00	9.00	0.00	1.00	25.00	16.00	16.00	4.00	4.00	9.00	120.00
h_i	0.400	0.175	0.00	0.108	0.308	0.233	0.233	0.133	0.133	0.175	

Lai aprēķinātu S_E vērtības, izmanto starpības $a_{i2} - \bar{a}_2$ no 3.13. tabulas 4. kolonnas. Šo starpību vērtības ir parādītas 3.19. tabulas 2. rindā. Trešajā rindā ir parādītas šo starpību kvadrātā vērtības. Tabulas apakšējā labajā šūnā ir parādīta šo starpību summa.

3.19. tabula

Aprēķinu rezultāti S_E vērtības noteikšanai

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	$\sum_{i=1}^{10}\left(a_{i2}-\overline{a}_2\right)^2$
$a_{i2}-\overline{a}_2$	-4.30	-2.30	-0.30	11.20	1.20	-3.30	1.20	-2.30	-0.80	-0.30	
$\left(a_{i2}-\overline{a}_2\right)^2$	18.49	5.29	0.09	125.44	1.44	10.89	1.44	5.29	0.64	0.09	169.10

Izmantojot vienādojumu (3.18), aprēķina:

$$S_E=\sqrt{\frac{169.100}{10-1-1}}=\sqrt{\frac{169.100}{8}}=\sqrt{21.124}=4.598.$$

3.20. tabulas 2. rindā ir parādītas novirzes e_{i2} vērtības no 3.14. tabulas 4. kolonnas. Tabulas 3. rindā parādītas studentizētās noviržu vērtības, kas aprēķinātas. izmantojot vienādojumu (3.19).

3.20. tabula

Noviržu e_{i2} un studentizēto noviržu r_i vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
e_{i2}	-0.82	-0.56	-0.30	10.62	-1.70	-0.98	-1.12	-1.14	-1.96	-2.04
r_i	-0.224	-0.314	-0.065	2.445	-0.444	-0.243	-0.278	-0.266	-0.458	-0.489

Būtiskākā ir studentizētā novirze $r_4=2.445$, tāpēc objektu o_4 var pamatoti uzskatīt par anomāliju dotajā datu kopā.

Studentizēto noviržu vērtību svarīga iezīme ir tā, ka to aprēķinos tiek izmantotas faktoriālā mainīgā lieluma attālums no tā vidējās vērtības.

Lai novērtētu rezultatīvā mainīgā vērtību noviržu no regresijas līknes ietekmes pakāpi uz pašu regresiju, tiek izmantoti dažādi skaitliskie rādītāji. Divi plaši izmantojamie šāda veida rādītāji ir *DFFITS* un Kuka attālums.

1. DFFITS

Šis rādītājs i rezultatīvajam mainīgajam tiek aprēķināts šādā veidā:

$$DFFITS_i=r_i\sqrt{\frac{h_i}{1-h_i}}, \qquad (3.20)$$

kur r_i – studentizētās novirzes vērtība faktoriālā mainīgā i vērtībai (pēc vienādojuma (3.19));

h_i – aprēķina, izmantojot vienādojumu (3.17).

3.21. tabulas 2. rindā attēlotas studentizēto noviržu r_i vērtības, kas ņemtas no 3.19. tabulas 3. rindas. Tabulas 3. rindā attēlotas h_i vērtības, kas ņemtas no 3.18. tabulas 3. rindas. Tabulas 4. rindā ir parādītas *DFFITS* vērtības, kas aprēķinātas pēc vienādojuma (3.20).

3.21. tabula

r_i, h_i un aprēķinātās *DFITS* vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
r_i	-0.224	-0.314	- 0.065	2.445	-0.444	-0.243	-0.278	-0.266	-0.458	-0.489
h_i	0.400	0.175	0000	0.108	0.308	0.233	0.233	0.123	0.133	0.175
$DFFITS_{\text{i}}$	-0.099	-0.067	0000	1.085	-0.245	-0.134	-0.153	-0.032	-0.179	-0.225

Izmantojot *DFFITS* rādītāju, tiek pieņemts, ka objekts o_i, kam $DFFITS_i > 1$, ir anomālija. Pamatojoties uz 3.21. tabulas pēdējā rindā dotajiem datiem, var izdarīt acīmredzamu secinājumu, ka objekts o_4 ir anomālija. Šis rezultāts sakrīt arī ar iepriekš iegūto rezultātu.

2. Kuka attālums (Cook's distance)

Šis rādītājs atspoguļo regresijas koeficienta vērtību atšķirību, kas aprēķināta datu kopā esošā i objekta klātbūtnē vai bez tā. Kuka attālumu aprēķina, izmantojot vienādojumu:

$$D_i = \frac{e_i^2}{S_E^2 p}\left(\frac{h_i}{\left(1-h_i\right)^2}\right), \tag{3.21}$$

kur $e_i = y_i - \hat{y}_{\bar{i}}$ – punkta y_i novirze no regresijas līknes;

n – rezultatīvā mainīgā y vērtību skaits;

k – faktoriālo mainīgo skaits (šajā gadījumā $k = 1$);

h_i – aprēķina, izmantojot vienādojumu (3.17):

$$S_E^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}. \tag{3.22}$$

Aprēķina Kuka attāluma vērtību iepriekš dotajiem datiem. 3.22. tabulas 2. rindā ir parādītas noviržu e_{i2} vērtības no 3.14. tabulas 4. kolonnas. Tabulas 3. rindā ir norādītas h_i vērtības no 3.21. tabulas 3. rindas. Tabulas 4. rindā ir norādītas aprēķinātās D_i vērtības.

Vērtības D_i nenosaka robežvērtību starp normālo datu apgabalu un anomālijām. Tikai objekts ar vislielāko D_i vērtību var tikt identificēts kā potenciāla anomālija. 3.22. tabulā objektam o_4 ir vislielākā D_i vērtība: $D_4 = 0.034$, tātad šis objekts ir potenciāla anomālija.

3.22. tabula

Vērtības e_i , h_i un aprēķinātās D_i vērtības

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
e_i	-0.82	-0.56	-0.30	10.62	-1.70	-0.98	-1.12	-1.14	-1.96	-2.04
h_i	0.400	0.175	0000	0.108	0.308	0.233	0.233	0.133	0.133	0.175
D_i	-0.021	-0.003	0000	0.034	-0.026	-0.090	-0.10	-0..005	-0.008	-0.012

Šajā sadaļā tika aplūkotas statistiskās metodes anomāliju identificēšanai lineārajā regresijā. Iegūtos rezultātus var attiecināt arī uz daudzfaktoru lineārās regresijas gadījumiem, kad ievērojami palielinās aprēķinu sarežģītība. Tā vietā, lai darbotos ar diskrētām skaitliskām vērtībām, daudzfaktoru lineārā regresija veic vektoru un matricu aprēķinus. Šāda veida aprēķinus var veikt, tikai izmantojot piemērotus skaitļošanas rīkus.

2. Robustas metodes anomāliju identificēšanai regresijā

Statistiskajām metodēm anomāliju noteikšanai regresijā ir viens būtisks trūkums: regresijas atkarības definīcija ir balstīta uz visiem sākotnējiem datiem, ieskaitot anomālijas. Tā rezultātā var rasties regresijas līkne, kas neatbilst lielākajai datu daļai. Spilgts piemērs tam ir regresijas līknes 3.8. un 3.9. attēlā. Pirmā līkne ir izveidota, izmantojot visu sākotnējo datu kopu, tostarp anomāliju objektu o_4. Otrā regresijas līkne tiek veidota uz datu kopas pamata, no kuras ir izslēgts objekts o_4. Acīmredzami otrā līkne labāk atbilst lielākajai datu daļai.

Lai izslēgtu anomāliju ietekmi uz konstruētās regresijas līkni, tiek izmantotas robustas metodes regresijas atkarības noteikšanai. Ir daudz spēcīgu šāda veida metožu. Šajā sadaļā tiks aplūkota vienkāršākā un visbiežāk izmantotā metode: M - novērtēšana. Regresijas M - novērtēšanas metodes

teorētiskie pamati tika likti darbā [Huber P.J., 1981]. Līdzīgi kā iepriekš, x ir faktoriālais mainīgais un y ir rezultatīvais mainīgais. Standarta lineārajā regresijā, lai noteiktu regresijas sakarību starp faktoriālo mainīgo un rezultatīvo mainīgo, mērķis ir minimizēt rezultatīvā mainīgā vērtību kvadrātu noviržu summu no regresijas līknes:

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} \left(y_i - x_i' b \right)^2 , \tag{3.23}$$

kur b – korelācijas koeficients, kas atspoguļo regresijas līknes slīpumu attiecībā pret horizontālo asi x.

Robustajā M - novērtēšanā vienādojuma (3.23) kvadrātiskās novirzes tiek aizstātas ar funkcijas $\rho(u_i)$ vērtībām, kur:

$$u_i = \frac{e_i}{S_m} . \tag{3.24}$$

Vienādojumā (3.24) e_i ir rezultatīvā i mainīgā vērtības novirze no regresijas līknes.

$$S_m = \frac{med\left| e_i - med\left\{ e_i \right\} \right|}{0.6475} = \frac{mad\left\{ e_i \right\}}{0.6475} , \tag{3.25}$$

kur $mad(e_i)$ – vērtību e_i noviržu mediāna no kopas mediānas vērtībām $\{e_i\}$.

Robustās M - novērtēšanas mērķa funkcija uzdota šādā veidā:

$$\sum_{i=1}^{n} \rho(u_i) = \sum_{i=1}^{n} \rho\left(\frac{e_i}{S_m} \right) = \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i'\beta}{S_m} \right). \tag{3.26}$$

Funkcijai $\rho(u_i)$ jāatbilst šādām prasībām:

- $\rho(u_i) \geq 0$;
- $\rho(0) = 0$;
- $\rho(u_i) = \rho(-u_i)$;
- $\rho(u_j) \geq \rho(u_i)$ priekš $|u_j| \geq |u_i|$.

Robustās regresijas atkarības M - novērtēšanas mērķis ir minimizēt mērķa funkciju (3.26):

$$\min \sum_{i=1}^{n} \rho(u_i) = \min \sum_{i=1}^{n} \rho\left(\frac{e_i}{S_m} \right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i'\beta}{S_m} \right). \tag{3.27}$$

Robustā M - novērtēšana ir balstīta uz maksimālās līdzības principa.

Ir piedāvātas dažādas funkcijas $\rho(u_i)$ definēšanas iespējas. Izplatīta pieeja ir izteikt funkciju $\rho(u_i)$ ar svērtās *Tukey* (*Tukey method*) izteiksmes palīdzību:

$$\frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4} . \ |u_i| \geq c , \quad \rho(u_i) = \frac{c^2}{6} , \qquad |u_i| > c , \tag{3.28}$$

kur $c = 4.685$.

Lai noteiktu β vērtību, ir jāatrisina vienādojums (3.26). Lai to izdarītu, ir jāizmanto funkcijas $\rho(u_i)$ atvasinājums $\Psi = \rho'(u_i)$ un vērtības $x_{ij} - i$ novērojuma j parametrs (atribūts). Tad:

$$\sum_{i=1}^{n} x_{ij} \Psi \left(\frac{y_i - \sum_{i=1}^{n} x_{ij} \beta_j}{S_m} \right) = 0 . \tag{3.29}$$

Lai nodrošinātu vienādojuma (3.29) atrisinājumu, svara funkciju $w(e_i)$ definē šādi:

$$w(e_i) = \frac{\Psi\left(\frac{y_i - x_i\beta}{S_m}\right)}{\left(\frac{y_i - x_i\beta}{S_m}\right)}. \tag{3.30}$$

Tā kā $u_i = \frac{e_i}{S_m}$, tad vienādojumu (3.30) var attēlot veidā:

$$\left(1 - \left(\frac{u_i}{c}\right)^2\right)^2, \ |u_i| \le c, \quad w_i = w(e_i) = 0, \ |u_i| > c, \tag{3.31}$$

kur $c = 4.685$.

Tagad vienādojums (3.29) iegūst šādu veidu:

$$\sum_{i=1}^{n} x_{ij} w_i \left(y_i - x_i'\beta\right) = 0. \tag{3.32}$$

Matricas formā vienādojumu (3.32) var attēlot šādā veidā:

$$\beta_{l+1} = \mathbf{X}^T\mathbf{W}_0 \times \boldsymbol{\beta} = \mathbf{X}^T\mathbf{W}_0\mathbf{Y}, \tag{3.33}$$

kur $\mathbf{W}_0$ – diagonālā svara matrica izmērā $n \times n$;

$\mathbf{X}$ – faktoriālo vērtību matrica $\left(n \times \left(p+1\right)\right)$. (Viendimensiju lineārajai regresijai $p = 1$).

Vērtības $\boldsymbol{\beta}$ aprēķināšanas procedūras tiek iteratīvi atkārtotas atbilstoši vienādojumam:

$$\beta_{l+1} = \left(\mathbf{X}^T\mathbf{W}_l\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}_l\mathbf{Y}\right). \tag{3.34}$$

Vienādojumā (3.33) $\mathbf{W}_0$ ir svara matrica, kurā i diagonāles elements ir vienāds ar w_{i0}. Šo vienādojumu sauc par svērto mazāko kvadrātu vienādojumu.

Rodas jautājums: kā var noteikt regresijas koeficienta sākotnējo vērtību $\hat{\beta}^{(0)}$? Parasti pirmais solis ir regresijas koeficienta noteikšana, izmantojot standarta lineārās regresijas metodi:

1) novērtēt regresijas koeficienta $\hat{\beta}^{(0)}$ sākotnējo vērtību, izmantojot standarta lineārās regresijas metodi;
2) noteikt novirzes vērtības $e_i = y_i - \hat{y}_i$;
3) aprēķināt S_m vērtību, izmantojot vienādojumu (3.25);
4) aprēķināt $u_i = \frac{e_i}{S_m}$ vērtības, izmantojot vienādojumu (3.24);
5) aprēķināt svaru $w(e_i)$ vērtības, izmantojot vienādojumu (3.30);
6) aprēķināt $\hat{\beta}^{(1)}$ vērtību, izmantojot mazāko kvadrātu svērto metodi, izmantojot svarus $w(e_i)$, kas aprēķināti pēc iepriekšējā procedūras;
7) atkārtot 2.– 6. procedūru, izmantojot iepriekšējā iterācijā iegūto $\hat{\beta}^{(l)}$ vērtību.

Par algoritma apstāšanās kritēriju kalpo starpība starp secīgām $\hat{\beta}^{(l+1)} - \hat{\beta}^{(l)}$ vērtībām. Ja šī starpība nepārsniedz *apriori* noteikto robežvērtību, algoritms pārstāj darboties.

4. NEPĀRTRAUKTU ATRIBŪTU VĒRTĪBU DISKRETIZĀCIJA

4.1. Kas ir diskretizācija?

Lai izprastu nepārtrauktu atribūtu vērtību diskretizācijas nozīmi, mērķus un uzdevumus, sadaļas sākumā par šo tēmu tiek sniegti citāti no diviem avotiem.

"Dati parasti parādās jauktā formātā: nominālā, diskrētā un/vai nepārtrauktā formātā. Diskrētie un nepārtrauktie dati ir kārtas datu tipi ar attiecībām starp to vērtībām, un nominālās vērtības tiem neuzliek nekādas atkarības. Diskrētās vērtības ir intervāli nepārtrauktā vērtību spektrā. Lai gan atribūta nepārtraukto vērtību skaits var būt bezgalīgi liels, diskrēto vērtību skaits bieži ir ierobežots un mazs. Divu veidu vērtības rada atšķirības apmācības un klasifikācijas kokos. Viens lēmumu koka izvades piemērs var vēl vairāk ilustrēt atšķirību starp diviem datu veidiem. Kad tiek veidots lēmumu koks, tiek atlasīta viena pazīme, lai izveidotu šķautni tās vērtībām. Ja ir saderīgas, nepārtrauktas un atsevišķas funkcijas, ir normāli izvēlēties nepārtrauktu pazīmi, jo tai ir vairāk iezīmju... Izvēloties nepārtrauktu atribūtu nākamajā koka līmenī, var ātri sasniegt "tīru stāvokli" ar visiem piemēriem, kas pieder vienai klasei... Tas izraisa sliktu klasifikatora veiktspēju. Tāpēc nav prātīgi izmantot nepārtrauktas vērtības, lai sadalītu lēmumu koka mezglu. Nepārtrauktas pazīmes ir nepieciešams diskretizēt vai nu pirms lēmumu koka izveides vai koka konstruēšanas procesā... Diskrētu vērtību izmantošanai salīdzinājumā ar nepārtrauktām vērtībām ir daudz citu priekšrocību. Diskrētās pazīmes ir tuvāk zināšanu reprezentācijas līmenim nekā nepārtrauktas. Datus var arī samazināt un vienkāršot, izmantojot sākotnējo izlasi. Gan lietotājiem, gan ekspertiem atsevišķas funkcijas ir vieglāk saprotamas, lietojamas un izskaidrojamas. Veiksmīga paraugu ņemšana var ievērojami paplašināt daudzu apmācības algoritmu robežas" [Liu H., et. al., 2002].

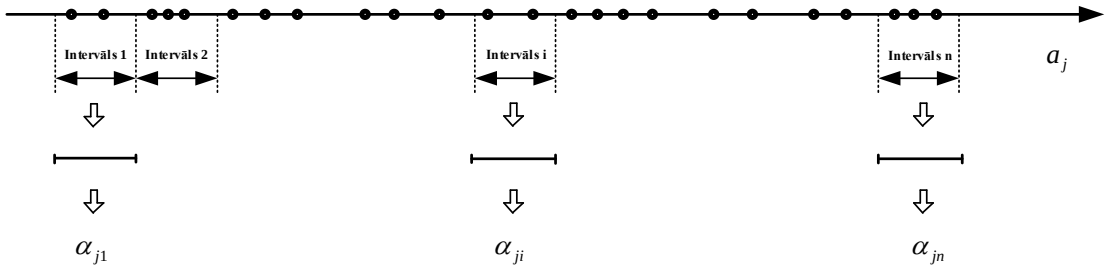
"...lielu skaitu statistikas un mašīnmācīšanās metožu var izmantot datu kopām, kas pilnībā sastāv no nominālvērtībām. Tomēr ļoti liela daļa reālo datu kopu ietver nepārtrauktus mainīgos... Viens problēmas risinājums ir sadalīt skaitliskos mainīgos vairākos intervālos un uzskatīt katru intervālu kā kategoriju. Šo nepārtraukto mainīgo sadalīšanas procesu kategorijās parasti sauc par diskretizāciju.

Nepārtraukta atribūtu izlase ir svarīga datu iegūšanas un mašīnmācīšanās algoritmu priekšapstrādes pieeja. Efektīva izlases metode var ne tikai samazināt prasības sistēmas modelim un uzlabot mašīnmācīšanās algoritmu efektivitāti, bet arī padarīt no diskretizētās datu kopas iegūtās zināšanas kompaktākas, vieglāk saprotamas un lietojamas. Diskretizācijas rezultāts ir ne tikai saistīts ar pašu algoritmu, bet arī ar datu sadalījumu un sadalīto punktu skaitu. Ja dažādām datu kopām tiek lietots viens un tas pats algoritms, var iegūt atšķirīgus rezultātus. Paraugu ņemšanas metodes efektivitāti var uzzināt no pēcapstrādes rezultātiem. Tas, vai izlases metode ir laba, ir saistīts arī ar vēlāk izmantoto secinājumu algoritmu.

Diskrētu vērtību izmantošanai ir daudz priekšrocību salīdzinājumā ar nepārtrauktajām vērtībām: (1) diskretizācija samazina nepārtraukto atribūtu vērtību skaitu, kā rezultātā samazinās sistēmas atmiņas prasības; (2) diskrētās pazīmes ir tuvāk zināšanu līmeņa atspoguļošanai nekā nepārtrauktas; (3) gan lietotājiem, gan ekspertiem atsevišķas funkcijas ir vieglāk saprotamas, lietojamas un izskaidrojamas; (4) diskretizācija padara apmācību precīzāku un ātrāku" [Hemada B., Lakshmi K.S.V., 2013].

Šie citāti parāda nepārtrauktu atribūtu vērtību diskretizācijas metožu būtību. Nosacītā atribūta a_j nepārtraukto vērtību diskretizācijas process shematiski parādīts 4.1. attēlā.

Viss nepārtrauktu vērtību diapazons, atribūtu a_j vērtības, ir sadalīts noteiktā intervālu skaitā n. Ir piedāvāti dažādi principi nepārtraukta lieluma vērtību sadalīšanai intervālos. Daži no pamatprincipiem tiks apspriesti vēlāk šajā nodaļā. Katrs intervāls tiek kodēts noteiktā veidā. Mašīnmācīšanā parasti ir ierasts kodēt intervālus, izmantojot dažus nominālos nosaukumus (kategorijas). Šīs kategorijas ir tikai atbilstošo intervālu apzīmējumi un nesatur nekādu citu informāciju.



4.1. attēls. **Nepārtrauktas atribūtu a_j vērtības diskretizācijas procesa shematisks attēlojums**

4.1. attēlā rezultējošie intervāli ir apzīmēti ar α_{ji}, $i = 1,...,n$. Jebkura nepārtraukta lieluma vērtība, kas ietilpst intervālā α_{ji}, tiek attēlota ar šo iezīmi un visas attiecīgās turpmākās datu apstrādes procedūras ir attēlotas šo intervālu kontekstā.

Statistiskās analīzes problēmās arī tiek izmantots nepārtrauktu lielumu dalījums intervālos. Piemēram, nepārtraukts mainīgais attēlo izlasē iekļauto indivīdu vecumu un pētījuma mērķis ir iegūt sadalījuma funkciju pa indivīdiem atbilstoši viņu vecumam. Lai to izdarītu, vecuma vērtību diapazons tiek sadalīts piemērotā intervālu skaitā un šajos intervālos tiek veiktas turpmākas datu analīzes procedūras. Bet jāpatur prātā, ka intervālu kopai ir šāda secība: jebkuram intervālam to sakārtotajā secībā katra nepārtraukta lieluma vērtība intervālos, kas atrodas pa kreisi no šī intervāla, būs mazāka par šī intervāla kreisās puses robežas vērtību. Tāpat katra nepārtraukta lieluma vērtība intervālos pa labi no noteiktā intervāla būs lielāka par šī intervāla labo robežvērtību.

Statistiskās izpētes kontekstā iegūto intervālu secībai var būt liela nozīme. Mašīnmācīšanās uzdevumu kontekstā (klasifikācija, lēmumu izvade utt.) iegūto intervālu secībai nav nozīmes. Piemēram, ir zināms no apmācības kopas, ka septiņi objekti, kuriem atribūtu vērtības α_j atrodas intervālā α_{ji}, pieder klasei c_k, un viens objekts ar atribūta vērtību α_j no šī intervāla pieder klasei c_l. Pieņem, ka klasificētajam objektam ir atribūta vērtība α_j, kas atbilst intervālam α_{ji}. No šiem datiem var secināt, ka šis objekts, iespējams, ar varbūtību $^7\!/_8$ pieder klasei c_k un, iespējams, ar varbūtību $^1\!/_8$ – klasei c_l.

No tā izriet svarīgs secinājums: lai izveidotu klasifikatora modeli, nav svarīgas atribūtu vērtības α_j kā tādas, bet gan attiecības starp atribūtu vērtībām noteiktā intervālā un objektu piederību klasēm. Citiem vārdiem sakot, ir svarīga sakarība: "intervāla iezīme – klases iezīme". Tāpēc intervālu kodēšana klasifikācijas un citos mašīnmācīšanās uzdevumos tiek veikta nominālajās skalās.

Var dot šādu formālu definīciju: diskretizācija ir atribūta nepārtraukto vērtību diapazona sadalīšana noteiktā intervālu skaitā. Visas atribūtu vērtības šajā intervālā tiek traktētas vienā šī intervāla nominālajā kategorijā. Turpmākās datu apstrādes un analīzes darbības tiek veiktas, izmantojot iegūtās nominālo intervālu iezīmes.

Nepārtrauktu atribūtu vērtību diapazona sadalīšana intervālos nav triviāls uzdevums. Kā norādīts darbā [Kotsiansis S., Kanellopoulos D., 2006], ir jāatrod kompromiss starp informācijas kvalitāti (viendabīgi intervāli attiecībā uz prognozēto atribūtu) un statistisko kvalitāti (pietiekams izlases lielums katrā intervālā, lai nodrošinātu vispārinājumus). Hī-kvadrāta testā koncentrējas uz statistisko priekšstatu, bet uz entropiju balstītajā testā – uz informācijas priekšstatu. Citi kritēriji mēģina atrast kompromisu starp informācijas un statistikas īpašībām.

4.2. Diskretizācijas metožu klasifikācija

Diskretizācijas metodes var klasificēt pēc daudzām pazīmēm. Pazīme šajā kontekstā ir principi, uz kuriem balstās diskretizācijas metodes. Pašlaik vispārpieņemtas ir piecas pazīmes, kas ir dažādu diskretizācijas metožu pamatā.

1. Neuzraudzīta - uzraudzīta

Neuzraudzītās metodes diskretizācijas veikšanai neizmanto informāciju par klasēm, kurām objekti pieder. Diskretizācija tiek veikta, pamatojoties tikai uz dotajām atribūtu vērtību kopām. Tipiski šīs metožu grupas pārstāvji ir diskretizācija pēc vienāda platuma intervāliem un vienāda biežuma intervāliem. *Uzraudzītās* metodes nepārtrauktu atribūtu vērtību diapazonu sadalīšanai izmanto informāciju par objektu piederību noteiktām klasēm.

2. *Globāla - lokāla*

Globālās metodes veic diskretizāciju pilnām attiecīgo atribūtu nepārtrauktu vērtību kopām. Citiem vārdiem sakot, diskretizācija tiek veikta pilnam objektu kopumam (piemēriem).

Lokālās metodes veic diskretizāciju atsevišķām attiecīgo atribūtu apakškopām, t. i., diskretizācija tiek veikta noteiktai objektu apakškopai (piemēriem).

3. *Diskretizācijas virziens*

Pēc šīs pazīmes diskretizācijas metodes iedalās *lejupejošās* metodēs un *augšupejošās* metodēs.

Lejupejošās metodes sākas ar pilnu nepārtrauktu atribūtu vērtību kopu. Pirmais diskretizācijas algoritma solis ir sadalīt visu atribūtu vērtību diapazonu divos intervālos. Intervālu dalīšanas process turpinās secīgi, līdz tiek sasniegts algoritma apstāšanās kritērijs.

Šīs grupas metodēm problēma ir griezuma punktu noteikšana, kas nosaka robežu starp diviem jaunizveidotajiem intervāliem. Uzmanība šim jautājumam tiks pievērsta vēlāk šajā nodaļā.

Augšupejošās metodes sākas ar pieņēmumu, ka katrai nepārtrauktai atribūta vērtībai ir atsevišķs intervāls. Algoritma sākotnējā solī tiek apvienotas atsevišķas vērtības un tiek veidoti daži sākotnējie intervāli. Intervālu sapludināšanas process turpinās, līdz tiek sasniegts algoritma apstāšanās kritērijs.

Šīs grupas metodēm problēma ir izvēlēties piemērotu nosacījumu, lai atlasītu pašreizējos intervālus kā sapludināšanas kandidātus. Uzmanība šim jautājumam tiks pievērsta vēlāk šajā nodaļā.

4. *Statiska - dinamiska*

Statiskās metodes veic nepārtrauktu atribūtu vērtību diskretizāciju pirms datu analīzes algoritma piemērošanas. Citiem vārdiem sakot, uz šīm metodēm balstīta datu izlase ir neatņemama datu pirmapstrādes sastāvdaļa.

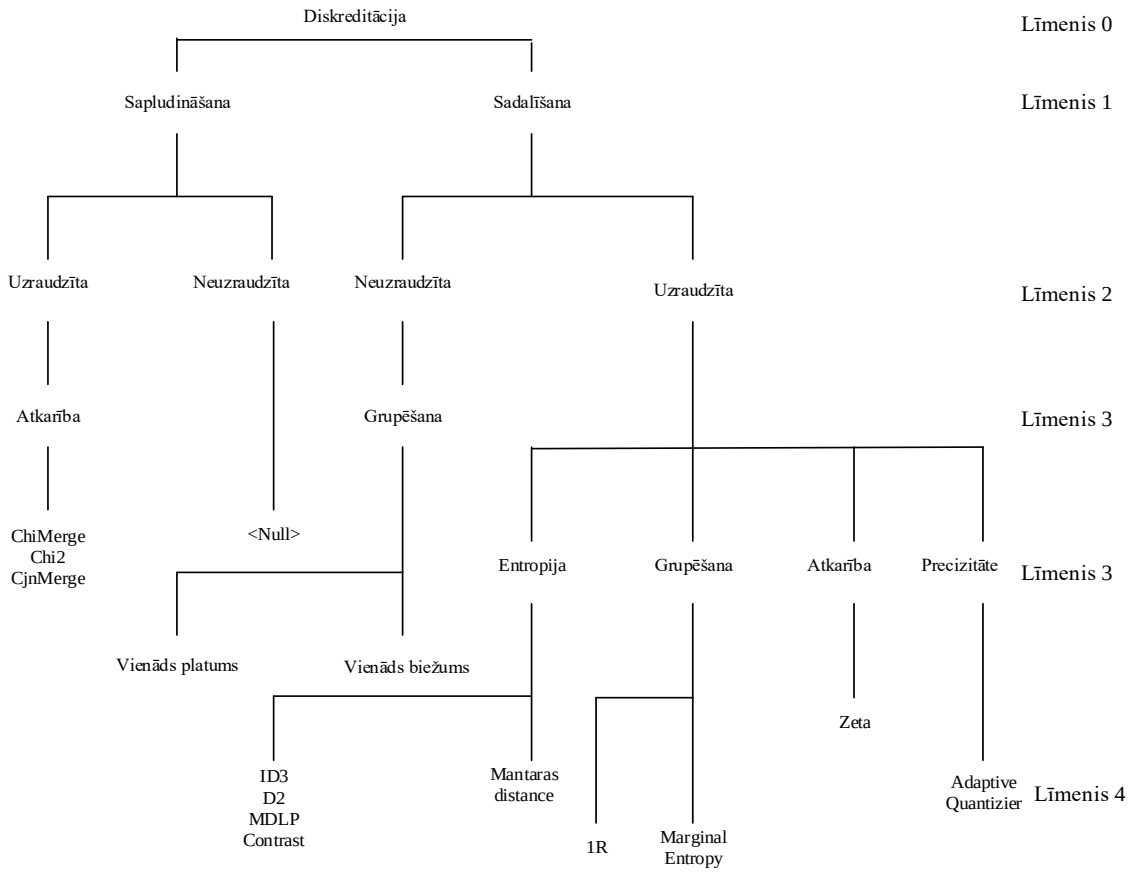
Dinamiskās metodes veic nepārtrauktu atribūtu vērtību diskretizāciju datu apstrādes laikā. Piemēram, izmantojot populāro C4.5 klasifikācijas algoritmu, attiecīgo atribūtu nepārtrauktās vērtības tiek diskretizētas klasifikatora modeļa veidošanas laikā.

5. *Tieša - pieaugoša*

Tiešās metodes sadala nepārtrauktu atribūtu vērtību diapazonus *apriori* noteiktā intervālu skaitā. Diskretizācija notiek, pamatojoties uz vienāda platuma vai vienāda biežuma intervāliem.

Pieaugošajām metodēm ir raksturīgs tas, ka diskretizācija sākas no kāda attiecīgā atribūta vērtību sākuma stāvokļa, piemēram, no viena intervāla ar *lejupejošām* metodēm vai no vairākiem intervāliem ar vienu atribūta vērtībām *augšupejošās* metodēs. Diskretizācijas procesa laikā pašreizējie intervāli tiek sadalīti ar *lejupejošām* metodēm vai apvienoti intervālos ar *augšupejošām* metodēm. Abos gadījumos ir nepieciešams apstāšanās kritērijs, lai noteiktu, kad diskreditācijas procesam jābeidzas.

Pamatojoties uz iepriekš dotajām pazīmēm un klasifikācijas shēmu, kas piedāvāta darbā [Liu H. et al., 2002], var izveidot vispārīgu klasifikācijas shēmu konkrētām diskretizācijas metodēm. Šī diagramma ar dažām izmaiņām ir parādīta 4.2. attēlā.



4.2. attēls. **Diskretizācijas metožu hierarhiskā klasifikācijas shēma**

Šī hierarhiskā diagramma satur 4 līmeņus. 1. līmenī diskretizācijas metodes tiek iedalītas divās lielās grupās: *sapludināšana* un *sadalīšana* atkarībā no tā, kurš no intervālu veidošanas principiem ir konkrētās metodes pamatā. 2. līmenī gan sapludināšanas, gan sadalīšanas metodes tiek iedalītas uzraudzītajās un neuzraudzītajās. 3. līmenī ir norādīti konkrētie intervālu veidošanas principi.

Struktūras galīgā pozīcija (4. līmenis) atspoguļo konkrētās izmantotās diskretizācijas metodes.

Lai noteiktu klasifikācijas pazīmes, kas raksturo konkrēto diskretizācijas metodi, ir nepieciešams detalizētāk apskatīt dotās hierarhiskās struktūras elementus.

4.3. Diskretizācijas process

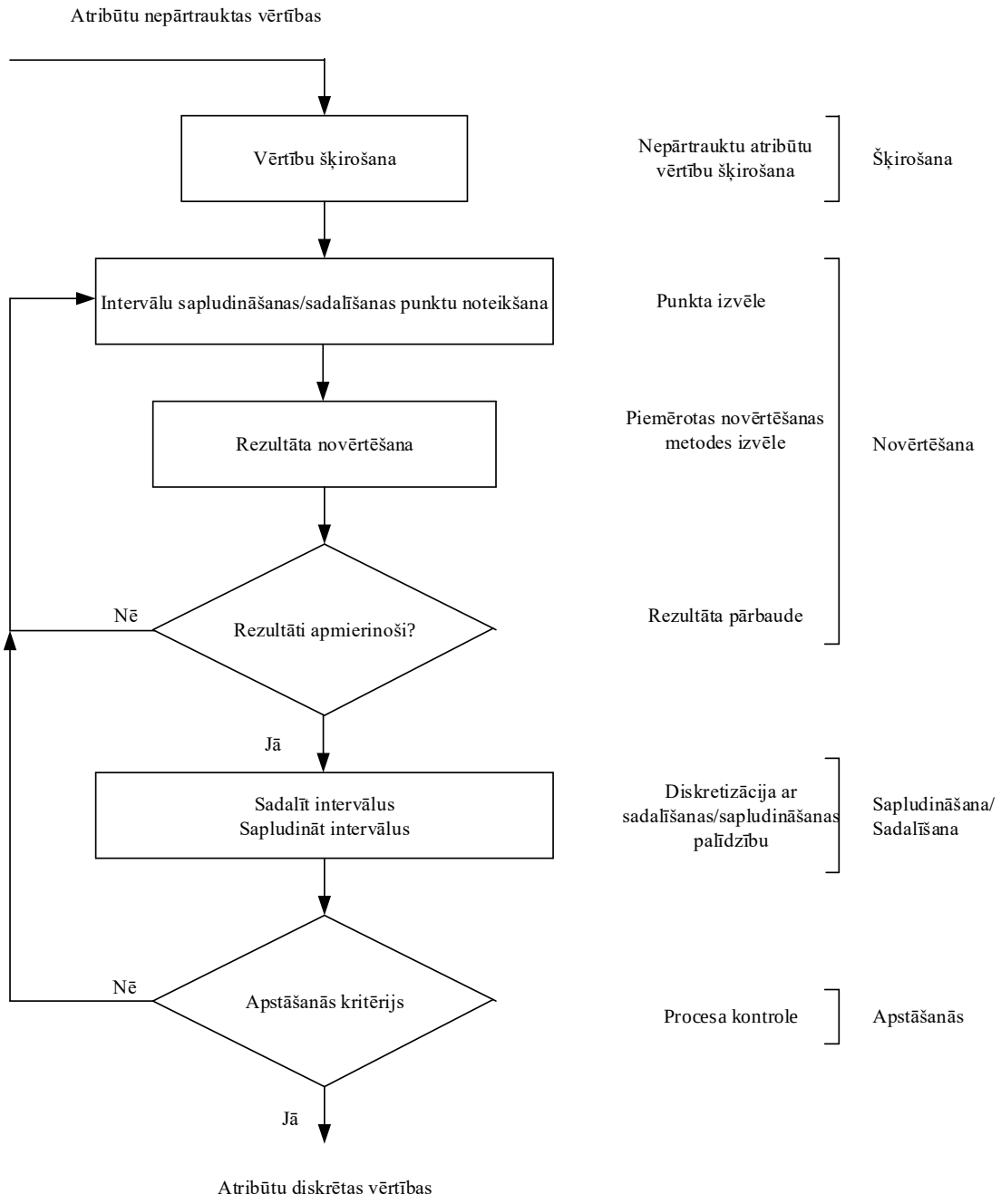
Nepārtraukto atribūtu vērtību diskretizācijas procedūru secība shematiski parādīta 4.3. attēlā [Liu H. et al., 2002].

Attiecīgo procedūru apraksts un raksturojums.

1. Šķirošana. Pirms diskretizācijas procedūru veikšanas nepārtrauktas atribūtu vērtības jāsakārto augošā vai dilstošā secībā.

2. Intervālu sapludināšanas vai sadalīšanas punktu noteikšana. Ja tiek izmantotas lejupejošas diskretizācijas metodes, tad nepārtrauktu atribūtu vērtību diapazonā vai vienā no iepriekš izveidotajiem intervāliem ir jānosaka griezuma punkta kandidāts. Ja tiek izmantota augšupejoša diskretizācijas metode, ir jāidentificē sapludināšanai blakus esošu intervālu kandidātu pāris.

3. Rezultātu novērtēšana. Sapludināšanas/sadalīšanas rezultāts ir jāpārbauda saskaņā ar piemērotu kritēriju. (Daži biežāk izmantojamie intervālu sapludināšanas/sadalīšanas kritēriji tiks parādīti turpmākajās nodaļas sadaļās.) Ja pašreizējais intervālu sadalījuma novērtēšanas rezultāts ir pozitīvs, griezuma punkts tiek fiksēts un iegūtie intervāli tiek pakļauti turpmākai sadalīšanai. Ja sadalīšanas rezultāts neatbilst izmantotajam kritērijam, tiek izvēlēts cits griezuma punkts un process tiek atkārtots, līdz tiek sasniegts pozitīvs rezultāts.



4.3. attēls. **Nepārtrauktu atribūtu diskretizācijas procedūru vispārīgā shēma**

Sapludinot intervālus, tiek pārbaudīta arī jaunizveidotā intervāla atbilstība kvalitātes kritērijam. Ja pārbaudes rezultāts ir pozitīvs, sapludināšanas process turpinās. Ja sapludināšanas rezultāts ir neapmierinošs, apvienošanai tiek atlasīts cits blakus esošo intervālu pāris. Process tiek atkārtots, līdz tiek identificēts piemērots intervālu pāris sapludināšanai.

4. ***Procesa apturēšana.*** Intervālu sapludināšanas/sadalīšanas process apstājas, kad tiek sasniegts apstāšanās kritērijs. Kopumā apstāšanās kritērija formulēšana ir neformāls uzdevums. Šis kritērijs ir noteikts konkrētā problēmā, pamatojoties uz kompromisu starp nepārtraukto vērtību skaitu intervālos un intervāla iezīmes un klases iezīmes atbilstību iegūtajiem intervāliem.

Atbilstoša apstāšanās kritērija noteikšana, it īpaši, ja katram atribūtam ir liels skaits nepārtrauktu vērtību, ir sarežģīts uzdevums. Šīs problēmas risināšana ir iespējama, tikai pamatojoties uz esošo datu rūpīgu analīzi. Vienīgais, ko var droši apgalvot, ka problēmai nav globāli optimāla risinājuma. No tā izriet, ka ir jāpanāk piemērots kompromiss starp pretrunīgajām prasībām gan attiecībā uz diskretizācijas rezultātiem, gan attiecībā uz pašas problēmas risināšanas rezultātiem.

Iepriekš aprakstītās diskretizācijas procedūras tiek veiktas katram atribūtam ar nepārtrauktām vērtībām. Jāatzīmē, ka patlaban ir izstrādātas metodes, kas ļauj vienlaicīgi veikt visu nepārtrauktu atribūtu vērtību diskretizāciju. Bet šīs metodes ir diezgan sarežģītas un ir ārpus šīs nodaļas tēmas.

4.4. Vienkāršākās diskretizācijas metodes

Šajā sadaļā ir apskatītas populārākās uzraudzītās/neuzraudzītās metodes nepārtrauktu atribūtu vērtību diskretizācijai.

1. Diskretizācija ar vienāda platuma intervāliem

Šīs metodes ideja ir sadalīt nepārtraukto atribūtu a_j vērtību diapazonu *apriori* uzdotā vienāda platuma intervālu skaitā. Uzdotam k intervālu skaitam katra intervāla platums tiek aprēķināts kā:

$$\text{Intervāla platums} = \frac{a_{j\max} - a_{j\min}}{k}, \tag{4.1}$$

kur $a_{j\max}$ – atribūta a_j maksimālā vērtība;

$a_{j\min}$ – atribūta a_j minimālā vērtība.

Sākotnējā atribūtu vērtību diapazona griezuma punkti ir definēti kā

$$\text{Griezuma punkts } i = a_{i\min} + i \text{ (Intervāla platums).} \tag{4.2}$$

Vienkāršs ilustratīvs piemērs. 4.1. tabulā ir parādīta sākotnējā nepārtraukto atribūtu a_j vērtību kopa.

4.1. tabula

Nepārtraukto atribūtu a_j vērtību sākotnējā kopa

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
a_{ij}	22	26	10	16	18	28	12	24

Šīs nepārtrauktās vērtības ir nepieciešams diskretizēt, izmantojot četrus $(k = 4)$ vienāda platuma intervālus.

Vispirms sakārto atribūtu a_j vērtības augošā secībā. Sakārtotās vērtības ir parādītas 4.2. tabulā.

4.2. tabula

Sakārtotas atribūtu a_j vērtības datiem no 4.1. tabulas

Objekti	o_3	o_8	o_4	o_5	o_1	o_8	o_2	o_6
a_{ij}	10	12	16	18	22	24	26	28

No šīs tabulas iegūst: $a_{j\min} = 10$, $a_{j\max} = 28$. Intervālu platumu aprēķina, izmantojot vienādojumu (4.1).

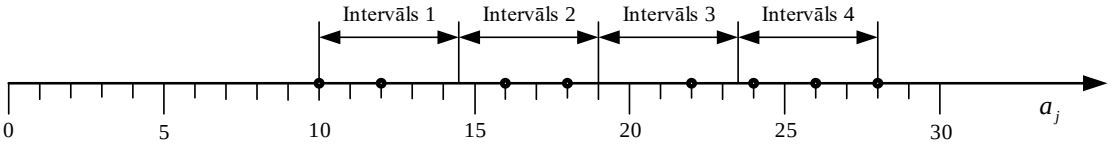
$$\text{Intervāla platums} = \frac{28-10}{4} = \frac{18}{4} = 4.5.$$

Tā kā diskretizācijai tika izmantoti četri intervāli, atribūtu a_j vērtību diapazonā būs trīs griezuma punkti. Šo punktu vērtības nosaka, izmantojot vienādojumu (4.2):

- griezuma punkts 1 = 10+1*4,5=14,5;
- griezuma punkts 2 = 10+2*4,5=19;
- griezuma punkts 3 = 10+3*4,5=23,5.

Šai metodei nav nepieciešams īpašs apstāšanās kritērijs, jo diskretizācijai intervālu skaits ir noteikts *apriori*.

Skaidrības labad iegūtie diskretizācijas rezultāti ir grafiski parādīti 4.4. attēlā.



4.4. attēls. **Diskretizācijas rezultātu grafisks attēlojums nepārtrauktu atribūtu a_j vērtībām no 4.1. tabulas**

Vienāda platuma intervālos var nebūt vienāds datu punktu skaits, kas ietilpst katrā intervālā. 4.4. attēlā ir redzams, ka 1. un 2. intervālā ir divi punkti, 3. intervālā – viens punkts, 4. intervālā – trīs punkti.

Būtisks vienāda platuma intervālu diskretizācijas metodes trūkums ir tas, ka intervālu veidošanas procesu var negatīvi ietekmēt anomāliju klātbūtne datos. Lai parādītu šī apgalvojuma pareizību, pieņem, ka atribūta vērtība $a_{6j} = 28$ ir kļūdaini uzrakstīta – $a_{6j} = 88$. No šiem datiem:

- intervāla platums = $\frac{88-10}{4} = 19.5$;
- griezuma punkts 1 = 10 +1 * 19,5 = 29,5;
- griezuma punkts 2 = 10 + 2 * 19,5 = 49;
- griezuma punkts 3 = 10 + 3 * 19,5 = 68,5.

Tagad visi datu punkti, izņemot punktu $a_{6j} = 88$, ietilpst 1. intervālā, 2. un 3. intervālā nav punktu, bet 4. intervālā ir viens datu punkts. Acīmredzot šādai diskretizācijai nav praktiskas nozīmes.

2. Diskretizācija ar vienāda biežuma intervāliem

Pieņem, ka ir n nepārtrauktu atribūtu a_j vērtības un intervālu skaits tiek norādīts *apriori*. Tad katrā intervālā ir jāietver aptuveni n/k atribūtu vērtības. Šeit jēdziens "aptuveni" ir saistīts ar to, ka pie dotajām n un k vērtībām ne vienmēr ir iespējams iegūt veselu dalījuma vērtību.

Par pamatu ņem sakārtotās atribūtu a_j vērtības no tabulas 4.2. Šī vērtību kopa tika diskretizēta, izmantojot četrus vienāda biežuma intervālus. Katrā intervālā ir jābūt $\frac{8}{4} = 2$ vērtībām. Tādējādi ir šādi intervāli: $[10,12]$, $[16,18]$, $[22,24]$, $[26,28]$.

Atribūtu vērtības katrā no intervāliem ir diezgan tuvu viena otrai. Tagad pieņem, ka vērtības vietā $a_{6j} = 28$ kļūdaini ir uzrakstīta vērtība $a_{6j} = 88$. Ko radīs šī kļūda? Acīmredzot, ka pirmajos trīs intervālos nekādas izmaiņas nenotiks. Bet ceturtais intervāls tagad satur vērtības $[26,88]$. Šķiet, ka tik liela atšķirība starp vērtībām neatbilst citiem datu punktiem. Tāpēc anomālijas datos rada arī nepamatotus diskretizācijas rezultātus. Lai gan anomāliju ietekmes sekas šajā gadījumā nešķiet tik katastrofālas kā izlases gadījumā ar vienāda platuma intervāliem, tomēr tās var ļoti negatīvi ietekmēt turpmākās datu analīzes rezultātus.

Jāņem vērā, ka vienāds datu punktu skaits katrā intervālā nenozīmē vienādu intervālu platumu. Šai metodei nav nepieciešams arī īpašs apstāšanās kritērijs, jo intervālu skaits ir iestatīts *apriori*.

3. 1R klasifikācijas metode

Vienāda platuma un vienāda biežuma intervālu metodes ir tipiskas neuzraudzītas diskretizācijas metožu pārstāvji. Pat ja objekti sākotnējā datu kopā ir saistīti ar dotajām klasēm, izmantojot šīs metodes informācija par objektu klasēm tiek pilnībā ignorēta.

Lai novērstu šo trūkumu [Holte R.C., 1993], ierosināja iepriekš aprakstīto metožu modifikāciju, kas tika nosaukta par 1R metodi. Šīs metodes būtība ir šāda. Ignorējot informāciju par objektu klasēm, tiek diskretizētas nepārtrauktu atribūtu a_j sākotnējās vērtības ar kādu no minētajām divām metodēm.

Tad tiek pielāgotas robežas starp intervāliem (griezuma punkti), lai izvairītos no situācijām, kad objekti abos intervālos, kas atrodas blakus robežai starp šiem intervāliem, pieder vienai un tai pašai klasei. Minimālajam datu punktu skaitam jebkurā no intervāliem ir jābūt vismaz 6, izņemot galējo labo intervālu, kurā var būt patvaļīgs datu punktu skaits. Šī prasība nosaka arī algoritma apstāšanās kritēriju.

Vienkāršs ilustratīvs piemērs. 4.3. tabulā ir parādītas nepārtrauktu atribūtu a_j vērtības. Tabulas pēdējā rindā ir norādītas to klašu iezīmes, kurām objekti pieder. Šīs nepārtrauktās vērtības ir nepieciešams diskretizēt ar 1R metodi.

4.3. tabula

Sākotnējā atribūtu a_j vērtību kopa

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}	o_{12}	o_{13}	o_{14}	o_{15}	o_{16}
a_{ij}	13	5	24	8	17	7	25	20	11	15	19	22	10	12	14	18
Klases	c_2	c_1	c_2	c_1	c_2	c_2	c_2	c_2	c_2	c_2	c_1	c_1	c_1	c_1	c_2	c_1

Atribūtu a_j vērtības sakārto augošā secībā, rezultāti ir doti 4.4.tabulā.

4.4. tabula

Sakārtotas atribūtu a_j vērtības no 4.3. tabulas

Objekti	o_2	o_6	o_4	o_{13}	o_9	o_{14}	o_1	o_{15}	o_{10}	o_5	o_{16}	o_{11}	o_8	o_{12}	o_3	o_7
a_{ij}	5	7	8	10	11	12	13	14	15	17	18	19	20	22	24	25
Klases	c_1	c_2	c_1	c_1	c_2	c_1	c_2	c_2	c_2	c_2	c_1	c_1	c_2	c_1	c_2	c_2

Izmantojot 4.4. tabulas datus, datu punktus (atribūtu a_j vērtības un to atbilstošos objektus) sadala divos vienāda biežuma intervālos. Kreisajā intervālā ir objekti o_2, o_6, o_4, o_{13}, o_9, o_{14}, o_1, o_{15} ar atribūtu a_j vērtībām no 5 līdz 14, labajā intervālā – objekti o_{10}, o_5, o_{16}, o_{11}, o_8, o_{12}, o_3, o_7 ar atribūtu a_j vērtībām no 15 līdz 25.

Analizējot robežu starp intervāliem, var redzēt, ka kreisā intervāla labajā pusē ir divi objekti (o_1 un o_{15}), kas pieder klasei c_2, labā intervāla kreisajā pusē ir objekti o_{10} un o_5, kas arī pieder klasei c_2. Saskaņā ar 1R metodes prasībām, ir jāpielāgo robeža starp intervāliem. Jāņem vērā, ka lielākā daļa objektu kreisajā intervālā pieder klasei c_1 un lielākā daļa objektu labajā intervālā – klasei c_2. Tāpēc šķiet loģiski pārvietot robežu starp intervāliem divas pozīcijas pa kreisi. Pēc tam jaunināta kreisajā intervālā būs objekti o_2, o_6, o_4, o_{13}, o_9, o_{14} ar atribūtu a_j vērtībām no 5 līdz 12, bet labajā intervālā – objekti o_1, o_{15}, o_{10}, o_5, o_{16}, o_{11}, o_8, o_{12}, o_3, o_7 ar atribūtu a_j vērtībām no 13 līdz 25.

Nekādi citi robežu pielāgojumi starp intervāliem netiks veikti, jo labajā pusē tagad ir tikai seši objekti ar atbilstošām atribūtu a_j vērtībām. Apstāšanās kritērijs ir sasniegts.

4.5. Diskretizācija uz entropijas novērtējuma pamata

Datu klasifikācijas kontekstā entropija ir informācijas apjoma novērtējums, uz kura pamata tiek konstruēta klasifikatora shēma. Plaši zināmajā ID3 datu klasifikācijas algoritmā [Quinlan J.R., 1986] entropijas aplēses tiek izmantotas, lai noteiktu zarus konstruējamā klasifikācijas koka mezglos.

Nepārtraukto atribūtu vērtību diskretizācijas sakarā autori [Fayyad U., Irani K., 1993] ierosināja izmantot entropijas aplēses, lai veidotu diskretizējamos intervālus.

Parasti datu kopas S entropijas aprēķins tiek definēts kā

$$H(S) = \sum_{i=1}^{m} p_i \log_2 p_i , \tag{4.3}$$

kur p_i – varbūtība, ka dotais objekts no kopas S pieder klasei c_i, $i = 1,...,m$.

Ar n apzīmē objektu skaitu kopā S un ar n_i, $i = 1,...,m$ – objektu apakškopas no S, kas pieder klasei c_i. Tad varbūtības p_i vērtību vienādojumā (4.5.1) aprēķina kā

$$p_i = \frac{n_i}{n}, \; i = 1,...,m . \tag{4.4}$$

Vienādojumā (4.3) $\log_2$ ir logaritms ar bāzi 2. Šādu logaritmu izmantošana ir saistīta ar faktu, ka informācija tiek kodēta bitos.

Entropijas aplēses var atrasties intervālā $[0, \log_2 m]$, kur m ir klašu skaits datu kopā.

Minimālā vērtība $H = 0$ tiek sasniegta, ja visi objekti sākotnējā datu kopā pieder vienai un tai pašai klasei i. Tad $p_i = 1$ un $\log_2 1 = 0$. Šādā hipotētiskā situācijā nav nepieciešams izveidot klasifikācijas modeli, jo nav vajadzīga objektu klasifikācija.

Maksimālā entropijas vērtība $H = \log_2 m$ tiek sasniegta hipotētiskā situācijā, kad visi sākotnējās datu kopas objekti ir vienmērīgi sadalīti pa visām m klasēm. Šī ir datu kopas maksimālā iespējamā nenoteiktība, pie kuras nav iespējams pareizs klasifikatora modelis.

Apskatāmā diskretizācijas algoritma ideja ir šāda. Nepārtraukto atribūtu a_j vērtības tiek sakārtotas augošā vai dilstošā secībā. Tādējādi tiek izveidota jauna datu kopas objektu secība, kas atbilst sakārtotajām atribūtu vērtībām. Iegūtajā objektu secībā to sākotnējā sadalīšana tiek veikta tā, lai kreisajā intervālā būtu viens objekts, kas atbilst mazākajai atribūta a_j vērtībai, bet labajā intervālā ir visi pārējie objekti.

Pēc tam kreisajam intervālam secīgi tiek pievienots viens objekts. Objektu skaits attiecīgajā labajā intervālā tiek samazināts par vienu. Galīgajā sadalījumā būs viens objekts ar vismazāko atribūta a_j vērtību labajā intervālā un visi pārējie objekti kreisajā intervālā.

Faktiski katrā iterācijā griezuma punkts nobīdās pa labi par vienu datu punktu.

Katram sadalījuma variantam tās entropijas novērtējums tiek aprēķināts, izmantojot vienādojumu:

$$H(l) = -p_{left} \sum_{i=1}^{m} p_{i,left} \log_2 p_{i,left} - p_{right} \sum_{i=1}^{m} p_{i,right} \log_2 p_{i,right} . \tag{4.5}$$

Šajā vienādojumā:

$l = 1,...,n-1$ – pašreizējā griezuma punkta numurs (iterācijas numurs);

p_{left} – varbūtība, ka l iterācijas reizē objekts nokļūst kreisajā intervālā. Praksē vērtību p_{left} aprēķina kā objektu skaita kreisajā intervālā attiecību pret kopējo objektu skaitu;

p_{right} – varbūtība, ka l iterācijas reizē objekts nokļūst labajā intervālā. Praksē vērtību p_{right} aprēķina kā objektu skaita labajā intervālā attiecību pret kopējo objektu skaitu;

$p_{i,left}$ – varbūtība, ka objekts kreisajā intervālā pieder klasei c_i. Šīs varbūtības vērtību aprēķina kā objektu skaita piederību klasei c_i kreisajā intervālā attiecībā pret kopējo objektu skaitu šajā intervālā;

$p_{i,right}$ – varbūtība, ka objekts labajā intervālā pieder klasei c_i. Šīs varbūtības vērtību aprēķina kā objektu skaita piederību klasei c_i labajā intervālā attiecībā pret kopējo objektu skaitu šajā intervālā.

Aprēķini, izmantojot vienādojumu (4.5), tiek veikti visiem alternatīvajiem griezuma punktiem. Efektīvais griezuma punkts ir punkts l^*, kas dod viszemāko entropijas aprēķinu $H\left(l^*\right)$ l^* iterācijā.

Algoritma pirmā pilnā soļa izpildes rezultātā sākotnējā objektu kopa tiek sadalīta divos intervālos. Ja nepieciešams, diskretizācija turpinās iegūtajos intervālos un intervālos, kas iegūti algoritma otrās un turpmākās darbības rezultātā.

Šī algoritma apstāšanās kritēriju var norādīt divos alternatīvos veidos:

1) *apriori* ir norādīts galīgs intervālu skaits;
2) kā apstāšanās kritērijs tiek izmantots minimālā apraksta garuma princips (*minimum description length principle*).

Minimālā apraksta garuma princips netiks neizmantots tā matemātiskās sarežģītības dēļ. Sīkāka informācija par šo principu atrodama darbos [Grünwald P.D., 2007; Rissanens J., 2007].

Vienkāršs ilustratīvs piemērs. 4.5. tabulā ir parādīta sākotnējā datu kopa: objektu kopa un šo objektu atribūtu a_j vērtību kopa. Tabulas 3. rindā ir norādītas klases, kurām pieder attiecīgie objekti.

4.5. tabula

Sākotnējo datu kopa

Objekti	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
a_{ij}	7	21	10	12	8	14	20	18
Objektu klases	c_1	c_2	c_1	c_1	c_2	c_1	c_2	c_1

Nepārtraukto atribūtu a_j vērtības ir nepieciešams diskretizēt divos intervālos, izmantojot entropijas aplēses, lai noteiktu efektīvo griezuma punktu.

Atribūtu a_j vērtības sakārto augošā secībā. Datu kopa ar sakārtotām atribūtu vērtībām ir parādīta 4.6. tabulā.

4.6. tabula

Datu kopa no 4.5. tabulas ar sakārtotām atribūtu a_j vērtībām

Objekti	o_1	o_5	o_3	o_4	o_6	o_8	o_7	o_2
a_{ij}	7	8	10	12	14	18	20	21
Objektu klases	c_1	c_2	c_1	c_1	c_1	c_1	c_2	c_2

Izpilda diskretizācijas algoritma darbības. Sākotnējo objektu kopu sadala divos intervālos.

1. iterācija:

- kreisais intervāls: objekts o_1;
- labais intervāls: objekti o_5, o_3, o_4, o_6, o_8, o_7, o_2.

(Intervālos tika iekļauti objekti ar atbilstošām atribūtu vērtībām, nevis atribūtu a_j vērtības. Lieta tāda, ka, diskretizējot nepārtrauktas atribūtu vērtības ar šeit piedāvāto metodi, interesē tikai objekti un to piederība attiecīgajām klasēm. Atribūtu a_j vērtības kalpo tikai datu sakārtošanai. Visi turpmākie aprēķini ietver objektu skaitu iegūtajos intervālos un objektu skaitu katrā no klasēm. Pēc diskretizācijas procesa pabeigšanas vienkārši jāpāriet no objektiem uz atbilstošo atribūtu vērtībām).

Tā kā kreisajā intervālā ir tikai viens objekts o_1, ir skaidrs, ka $p_{left} = \frac{1}{8} = 0.125$. Tā kā objekts o_1 pieder klasei c_1, $p_{1,left} = 1$, $\log_2 1 = 0$, un izteiksmes (4.5) labās puses kreisā summa ir vienāda ar 0.

Labajā intervālā ir 7 objekti, tāpēc $p_{right} = \frac{7}{8} = 0.875$. Četri objekti šajā intervālā pieder klasei c_1 un 3 objekti pieder klasei c_2. No šejienes $p_{1,right} = \frac{4}{7} = 0.571$. $p_{2,right} = \frac{3}{7} = 0.429$.

Pilnu aprēķinu veic, izmantojot vienādojumu (4.5).

$$H(1) = -0 - 0.875 * (0.571 * \log_2 0.571 + 0.429 * \log_2 0.429) =$$
$$= -0.875 * (0.571 * (-0.8084) + 0.429 * (-1.2209)) =$$
$$= -0,875 * (-0.4616 - 0.5238) = -0.875 * (-0.9854) = 0.8622.$$

2. iterācija:

- kreisais intervāls: objekti o_1, o_5;
- labais intervāls: objekti o_3, o_4, o_6, o_8, o_7, o_2.

Tagad kreisajā intervālā ir 2 objekti, bet labajā – 6 objekti. Tāpēc

$p_{left} = \frac{2}{8} = 0.250$; $p_{right} = \frac{6}{8} = 0.750$.

Viens objekts kreisajā intervālā pieder klasei c_1 un viens objekts pieder klasei c_2. Tāpēc

$p_{1,left} = \frac{1}{2} = 0.500$; $p_{2,left} = \frac{1}{2} = 0.500$.

Četri objekti labajā intervālā pieder klasei c_1 un divi objekti pieder klasei c_2. Tāpēc

$p_{1,right} = \frac{4}{6} = 0.667$; $p_{2,right} = \frac{2}{6} = 0.333$.

Sadalījuma entropijas novērtējumu aprēķina, izmantojot vienādojumu (4.5).

$$H(2) = -0.250 * (0.500 \log_2 0.500 + 0.500 \log_2 0.500) -$$
$$-0.750 * (0.667 \log_2 0.667 + 0.333 \log_2 0.333) =$$
$$= -0.250 * (0.500 * (-1) + 0.500 * (-1)) - 0.750 * (0.667 * (-0.5842) + 0.333 * (-1.5864)) =$$
$$= -0.250 * (-0.500 - 0.500) - 0.750 * (-0.3896 - 0.5827) =$$
$$= -0.250 * (-1) - 0.750 * (-0.9723) = 0.2500 + 0.7299 = 0.9799.$$

Pārējie aprēķini tiek veikti pēc analoģijas. Aprēķinu rezultāti ir apkopoti 4.7. tabulā.

4.7. tabula

Entropijas aprēķinu rezultāti potenciālajiem griezuma punktiem

Iterācijas numurs (l)	1	2	3	4	5	6	7
H(l)	0.8622	0.9799	0.9502	0.9056	0.7946	0.9371	0.8622

Kā izriet no šīs tabulas datiem, zemākais entropijas novērtējums tika sasniegts algoritma 5. iterācijā. Tāpēc ir iegūts šāds objektu sadalījums:

- kreisais intervāls: objekti o_1, o_5, o_3, o_4, o_6;
- labais intervāls: objekti o_8, o_7, o_2.

Šī diskretizācija jāuzskata par veiksmīgu, jo kreisajā intervālā tikai objekts o_5 pieder klasei c_2, visi pārējie objekti pieder klasei c_1. Labajā intervālā tikai objekts o_8 pieder klasei c_1, visi pārējie objekti pieder klasei c_2.

Galvenais mērķis šajā piemērā ir diskretizēt nepārtraukto atribūtu a_j vērtības. Izmantojot iepriekš minēto objektu sadalījumu intervālos un datus no 4.6. tabulas, katrā no intervāliem ir šādas atribūtu a_j vērtības:

- kreisais intervāls: 7, 8, 10, 12,14 $\Rightarrow \alpha_{j1} = [7,14]$;
- labais intervāls: 18, 20, 21 $\Rightarrow \alpha_{j2} = [18,21]$.

Ja ir jauns objekts ar atribūta vērtību $a_j = 11$, tad ar lielu varbūtības pakāpi var teikt, ka šis objekts pieder klasei c_1.

4.6. Uz χ^2 statistiku balstītas metodes

χ^2 - tas ir statistiskais novērtējums, ar kura palīdzību tiek veikts nozīmības tests attiecībām "atribūtu vērtības - objektu piederība klasēm". Uz χ^2 statistiku balstīto diskretizācijas metožu būtība ir tāda, ka blakus intervālos, kas veidojas kādā algoritma iterācijā, tiek noteikts, vai divas kategorijas var tikt apvienotas vai nē, pamatojoties uz to biežumu sadalījumu datu kopā. Ja atribūtu vērtības divos blakus intervālos ir neatkarīgas no pazīmju klases iezīmēm, tad šādus intervālus var apvienot vienā kopējā intervālā. Pretējā gadījumā intervālu apvienošana nav iespējama.

Kā piemēru nepārtrauktu atribūtu vērtību diskretizācijai, pamatojoties uz χ^2 statistiku, var apskatīt *ChiMerge* metodi [Kerber R., 1992]. To var raksturot kā augšupejošu uzraudzītu metodi.

Kā šī metode darbojas? Dota datu kopa, kas sastāv no n objektiem, no kuriem katrs tiek novērtēts ar atbilstošo atribūta a_j vērtību. Turklāt visiem objektiem ir norādītas klašu iezīmes.

Sākotnēji tiek pieņemts, ka katrs datu punkts veido vienu intervālu, t. i., sākotnējais intervālu skaits ir vienāds ar objektu skaitu datu kopā.

Vispirms tiek veidoti primārie intervāli ar objektiem, kas pieder tikai vienai klasei.

Tad pirmajā iterācijā katram blakus esošu intervālu pārim aprēķina χ^2 vērtību:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{p} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{4.6}$$

kur p – klašu skaits;

A_{ij} – dažādu klašu iezīmju i intervālam un j klasei;

Lai noteiktu vērtību E_{ij} vienādojumā (4.6), ievieš šādus apzīmējumus:

R_i – objektu skaits i intervālā, $R_i = \sum_{i=1}^{p} A_{ij}$;

C_j – objektu skaits j klasē, $C_j = \sum_{j=1}^{m} A_{ij}$;

N – kopējais objektu skaits, $N = \sum_{i=1}^{p} C_j = \sum_{i=1}^{m} R_i$.

Tad E_{ij} vērtība tiek interpretēta kā sagaidāmais biežums:

$$E_{ij} = \frac{R_i C_j}{N}. \tag{4.7}$$

χ^2 vērtības tiek aprēķinātas katram blakus esošu intervālu pārim, kas iegūti sākotnējā algoritma iterācijā. Pēc tam iegūtās aplēses tiek salīdzinātas savā starpā. Blakus esošo objektu pāris, kuriem ir viszemākais χ^2 vērtējums, tiek sapludināts. Tādējādi tiek izveidots jauns intervālu kopums.

Procesu iteratīvi atkārto, līdz tiek sasniegts apstāšanās kritērijs.

Apstāšanās kritēriju var norādīt divos alternatīvos veidos:

1) diskretizācijas intervālu skaits ir *apriori* zināms;
2) tiek noteikta χ^2 robežvērtība (ieteicams uzdot šo robežvērtību intervālā $[0.90, 0.99]$).

Ilustratīvs piemērs. Par pamatu ņem 4.5. tabulas datus. 4.6. tabulā sakārtotie dati ir atkārtoti 4.8. tabulā.

4.8. tabula

Sakārtotie dati no 4.6. tabulas

Objekti	o_1	o_5	o_3	o_4	o_6	o_8	o_7	o_2
a_{ij}	7	8	10	12	14	18	20	21
Klases	c_1	c_2	c_1	c_1	c_1	c_1	c_2	c_2

Izmantojot *ChiMerge* metodi, nepārtraukto atribūtu a_j vērtības ir jādiskretizē divos intervālos.

Sākumā pieņem, ka katrs intervāls ietver vienu datu objektu. Tādējādi ir sākotnējais 8 intervālu kopums. Veic objektu sadalīšanu. Definē primāros intervālus tā, lai katrā intervālā būtu objekti, kas pieder vienai klasei.

Intervāls1: objekts o_1;

Intervāls 2: objekts o_5:

Intervāls 3: objekti o_3, o_4, o_6;

Intervāls 4: objekti o_8, o_7;

Intervāls 5: objekts o_2.

(Tāpat kā iepriekšējā sadaļā, veicot attiecīgos aprēķinus, jādarbojas ar intervāliem, kas satur objektus, nevis atbilstošās atribūtu a_j vērtības. Tas tiek darīts, lai uzskatāmāk atspoguļotu aprēķinu procesus un jaunu intervālu veidošanos. Pāreja uz atribūtu vērtību intervāliem var viegli tikt izpildīta pēc visu algoritma iterāciju pabeigšanas).

Lai noteiktu intervālu pāri – sapludināšanas kandidātus, aprēķina χ^2 katram intervālu pārim.

1. iterācija.

- Intervālu pāris $Int.1, Int.2$.

Lai atvieglotu aprēķinus, tika izveidota 4.9. tabula.

4.9. tabula

Darba tabula aprēķinu veikšanai

	1. klase $(j=1)$	2. klase $(j=2)$	Σ pa rindām
$Int.1 (i=1)$	A_{11}	A_{12}	R_1
$Int.2 (i=2)$	A_{21}	A_{22}	R_2
Σ pa kolonnām	C_1	C_2	N

Tabulā ir divas kolonnas, jo sākotnējā datu kopā ir divas objektu klases. Ja ir lielāks klašu skaits, kolonnu skaitam jāatbilst objektu klašu skaitam.

Šajā tabulā:

A_{11} – c_1 klasei piederošo objektu skaits 1. intervālā;

A_{12} – c_2 klasei piederošo objektu skaits 1. intervālā;

A_{21} – c_1 klasei piederošo objektu skaits 2. intervālā;

A_{22} – c_2 klasei piederošo objektu skaits 2. intervālā;

R_1, R_2 – skaitļu summa tabulas rindās;

C_1, C_2 – skaitļu summas tabulas ailēs;

N – visu skaitļu summa tabulas rindās vai kolonnās.

4.10. tabula

Darba tabula aprēķinu veikšanai

	Klase c_1 $(j=1)$	Klase c_2 $(j=2)$	Σ pa rindām
$Int.1(i=1)$	$A_{11}=1$	$A_{12}=0$	$R_1=1$
$Int.2(i=2)$	$A_{21}=0$	$A_{22}=1$	$R_2=1$
Σ pa kolonnām	$C_1=1$	$C_2=1$	$N=2$

Izmantojot vienādojumu (4.7), aprēķina E_{ij} vērtības:

$$E_{11}=\frac{R_1C_1}{N}=\frac{1*1}{2}=0.500;\ E_{12}=\frac{R_1C_2}{N}=\frac{1*1}{2}=0.500;$$

$$E_{21}=\frac{R_2C_1}{N}=\frac{1*1}{2}=0.500;\ E_{12}=\frac{R_2C_2}{N}=\frac{1*1}{2}=0.500.$$

Izmantojot vienādojumu (4.6), aprēķina $\chi^2_{1,2}(1)$ vērtību. Šeit apakšindekss 1,2 nozīmē, ka ir runa par 1. un 2. intervālu sākotnējā intervālu secībā. Skaitlis (1) nozīmē 1. iterāciju.

$$\chi^2_{1,2}(1)=\frac{\left(A_{11}-E_{11}\right)^2}{E_{11}}+\frac{\left(A_{12}-E_{12}\right)}{E_{12}}+\frac{\left(A_{21}-E_{12}\right)}{E_{12}}+\frac{\left(A_{22}-E_{22}\right)}{E_{22}}=$$

$$=\frac{\left(1-0.500\right)^2}{0.500}+\frac{\left(0-0.500\right)^2}{0.500}+\frac{\left(0-0.500\right)^2}{0.500}+\frac{\left(1-0.500\right)^2}{0.500}=$$

$$=0.500+0.500+0.500+0.500=2.000.$$

- Intervālu pāris $Int.2, Int.3$.

4.11. tabula

Darba tabula aprēķinu veikšanai

	c_1 $(j=1)$	c_2 $(j=2)$	Σ
$Int.2(i=1)$	0	1	1
$Int.3(i=2)$	3	0	3
Σ	3	1	4

(indekss $i=1$ tika izmantots $Int.2$ un $i=2$ – $Int.3$, lai veiktu aprēķinus līdzīgā veidā).

$$E_{11}=\frac{1*3}{4}=0.750;\ E_{12}=\frac{1*1}{4}=0.250;\ E_{21}=\frac{3*3}{4}=2.250;\ E_{22}=\frac{3*1}{4}=0.750.$$

$$\chi^2_{2,3}(1)=\frac{\left(0-0.750\right)^2}{0.750}+\frac{\left(1-0.250\right)^2}{0.250}+\frac{\left(3-2.250\right)^2}{2.250}+\frac{\left(0-0.750\right)^2}{0.750}=$$

$$=0.750+0.250+2.250+0.750=4.000.$$

- Intervālu pāris $Int.3, Int.4$.

4.12. tabula

Darba tabula aprēķinu veikšanai

	c_1 $(j=1)$	c_2 $(j=2)$	Σ
$Int.3(i=1)$	3	0	3
$Int.4(i=2)$	0	2	2
Σ	3	2	5

$$E_{11}=\frac{3*3}{5}=1.800;\ E_{12}=\frac{3*2}{5}=1.200;\ E_{21}=\frac{2*3}{5}=1.200;\ E_{22}=\frac{2*2}{5}=0.800.$$

$$\chi^2_{3,4}(1)=\frac{(3-1.800)^2}{1.800}+\frac{(0-1.200)^2}{1.200}+\frac{(0-1.200)^2}{1.200}+\frac{(2-0.800)^2}{0.800}=$$
$$=0.800+1.200+1.200+1.800=5.000.$$

- Intervālu pāris $Int.4, Int.5$.

4.13. tabula

Darba tabula aprēķinu veikšanai

	$c_1\ (j=1)$	$c_2\ (j=2)$	Σ
$Int.4(i=1)$	0	2	2
$Int.5(i=2)$	1	0	1
Σ	1	2	4

$$E_{11}=\frac{2*1}{3}=0.667;\ E_{12}=\frac{2*2}{3}=1.333;\ E_{21}=\frac{1*1}{3}=0.333;\ E_{22}=\frac{1*2}{3}=0.667.$$

$$\chi^2_{4,5}(1)=\frac{(0-0.667)^2}{0.667}+\frac{(2-1.333)^2}{1.333}+\frac{(1-0.333)^2}{0.333}+\frac{(0-0.667)^2}{0.667}=$$
$$=0.667+0.333+1.336+0.667=3.000.$$

Vismazākā no visām aprēķinātajām aplēsēm ir $\chi^2_{1,2}(1)=2.000$. Tāpēc ir nepieciešams sapludināt pašreizējos intervālus $Int.1$, $Int.2$. Rezultātā tika iegūti starpintervāli.

2. iterācija.

$Int.1(2)$: objekti o_1, o_5;

$Int.2(2)$: objekti o_3, o_4, o_5;

$Int.3(2)$: objekti o_8. o_7;

$Int.4(2)$: objekts o_2.

Šo intervālu pāru χ^2 aprēķini tiek veikti pēc analoģijas ar iepriekš sniegtajiem aprēķiniem. Aprēķinu rezultāti:

$Int.1(2), Int.2(2)$: $\chi^2_{1,2}(2)=1.875$;

$Int.2(2), Int.3(2)$: $\chi^2_{2,3}(2)=5.000$;

$Int.3(2), Int.4(2)$: $\chi^2_{3,4}(2)=2.940$.

Mazākā no šīm vērtībām ir $\chi^2_{1,2}(2)=1.875$. Tāpēc ir nepieciešams sapludināt intervālus. Rezultātā ir trīs nākamie starpintervāli.

3. iterācija.

$Int.1(3)$: objekti o_1, o_5, o_3, o_4, o_6;

$Int.2(3)$: objekti o_8, o_7;

$Int.3(3)$: objekts o_2.

$Int.1(3), Int.2(3)$: $\chi^2_{1,2}(3)=3.750$;

$Int.2(3), Int.3(3)$: $\chi^2_{2,3}(2)=5.000$.

No tā izriet, ka $Int.1(3), Int.2(3)$ ir jābūt sapludinātiem. Rezultātā ir šādi divi intervāli:

$Int.1$: objekti o_1, o_5, o_3, o_4, o_6;

$Int.2$: objekti o_8, o_7, o_2.

Tā kā tika iegūti divi rezultējošie intervāli, tika sasniegts apstāšanās kritērijs un algoritma darbība beidzas.

Pārejot no objektiem uz atbilstošajām nepārtraukto atribūtu a_j vērtībām, ir

$Int.1$ (kreisais intervāls): 7, 8, 10. 12, 14 $\Rightarrow \alpha_1 = [7,14]$;

$Int.2$ (labais intervāls): 18, 20. 21 $\Rightarrow \alpha_2 = [18,21]$.

Jāatzīmē, ka šie intervāli ir tādi paši kā intervāli, kas iegūti no tiem pašiem sākuma datiem iepriekšējās sadaļas piemērā, izmantojot pilnīgi citu diskretizācijas metodi. Šī sakritība norāda uz abu metožu efektivitāti.

5. NORMALIZĀCIJA UN ATRIBŪTU VĒRTĪBU STANDARTIZĀCIJA

5.1. Definīcijas un piezīmes

Jebkuras datu pirmapstrādes procedūras mērķis ir sagatavot datus un pārveidot tos formā, kas ir vispiemērotākā šo datu tālākai analīzei.

Sākotnējos datos atribūtu vērtības var izmērīt dažādās skalās un to vērtības dažādiem atribūtiem var būt ļoti atšķirīgas. Piemēram, starp dažādām iedzīvotāju grupām tiek veikti socioloģiski pētījumi. Izmantotie atribūti ietver personu vecumu un gada ienākumus. Indivīdu vecums grupās ir ierobežots intervālā [25–75] gadi. Savukārt indivīdu gada ienākumi var svārstīties ļoti plašās robežās gan katras grupas ietvaros, gan starp grupām.

Turpmāk pieņem, ka pētījuma nolūkos tiek mēģināts identificēt indivīdu apakšgrupas (klasterus), pamatojoties uz abu atribūtu vērtībām. Klasterizāciju var veikt ar k-tuvāko kaimiņu metodi vai k-vidējā metodi. Izmantojot katru no šīm metodēm, ir jānovērtē attālumi starp punktiem (indivīdiem) datu telpā. Visbiežāk šim nolūkam tiek izmantots Eiklīda attālums. Ja tiek izmantotas sākotnējās atribūtu vērtības, tad attiecīgo attālumu aprēķinos lielā mērā dominēs gada ienākumu vērtības, jo šīs vērtības vienmēr ir daudz lielākas nekā atbilstošās indivīdu vecuma vērtības.

Arī daudzās citās datu analīzes metodēs, mašīnmācīšanā un neironu tīklos nevar tiešā veidā izmantot datu sākotnējās atribūtu vērtības.

Kritiski svarīga ir kritēriju vērtību normalizācija lēmumu pieņemšanas problēmās. Fakts ir tāds, ka daudzas metodes daudzkritēriju lēmumu izvēlei izmanto lineāras kritēriju vērtību kombinācijas katram no alternatīvajiem risinājumiem. Plaši pazīstamā TOPSIS metode izmanto aprēķinus par attālumiem no katra alternatīvā risinājuma līdz "ideāli pozitīvajam" un "ideāli negatīvajam". Abos gadījumos iegūtajos attāluma aprēķinos dominēs lielas kritēriju vērtības.

Lai risinātu problēmas, kas saistītas ar neapstrādātām atribūtu vai kritēriju vērtībām, šīs vērtības ir jāpārvērš piemērotā formā.

Ir trīs galvenās pieejas oriģinālo atribūtu vērtību pārveidošanai kompaktā bezdimensiju formā, proti, normalizēšana, standartizācija un transformācija. Diemžēl šie jēdzieni literatūrā bieži tiek lietoti kā sinonīmi. Tas rada neskaidrības un šo svarīgo jēdzienu nepareizu lietošanu. Šajā darbā tiek izmantotas šādas definīcijas:

- *normalizēšana* – atribūta sākotnējo vērtību pārveidošana bezdimensiju formā, izmantojot īpašas paša atribūta vērtības vai šo vērtību kombinācijas, vai kādu no attiecīgā atribūta vērtību sadalījuma parametriem normalizēšanai;
- *standartizācija* – sākotnējo atribūtu vērtību pārveidošana bezdimensiju standarta formā, izmantojot tikai attiecīgo atribūtu vērtību sadalījuma parametrus;
- *transformācija* – sākotnējo atribūtu vērtību pārveidošana citā formā, kuras pamatā ir kāda matemātiska funkcija.

Dažreiz literatūrā tiek izmantots vispārināts atribūtu vērtību *mērogošanas jēdziens*, kas tiek saprasts kā atribūtu vērtību pārveidošana vēlamajā bezdimensiju formā.

Šajā nodaļā tiks parādītas labi zināmās metodes sākotnējo atribūtu vērtību normalizācijai un standartizēšanai. 6. nodaļā tiks apskatītas metodes datu atribūtu vērtību transformācijai.

5.2. Sākotnējo atribūtu vērtību normalizācijas metodes

Normalizācijas metodes var sagrupēt, ņemot vērā normalizējošā faktora noteikšanas principus.

1. Metožu grupa ar vienu normalizējošā faktora vērtību

- *Normalizācija attiecībā pret maksimālo atribūta vērtību:*

$$a_{ij}^{n} = \frac{a_{ij}}{a_{j\max}}, \tag{5.1}$$

kur a_{ij} – objekta o_i sākotnējā atribūta a_j vērtība;

$a_{j\max}$ – maksimālā atribūta a_j vērtība.

Normalizācija ar vienādojuma (5.1) palīdzību ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Piemērs. 5.1. tabulā ir parādītas atribūtu a_j sākotnējās vērtības.

5.1. tabula

Sākotnējās atribūtu vērtības a_j

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9

Atribūta maksimālā vērtība $a_{j\max} = 9$. Normalizējot visas sākotnējās vērtības pret šo vērtību, iegūst atribūtu vērtības, kas parādītas 5.2. tabulā.

5.2. tabula

Normalizētas atribūtu vērtības a_j no 5.1. tabulas

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}^n	0.333	0.556	0.778	0.444	0.889	1000

Šo normalizācijas metodi raksturo fakts, ka visas normalizētās vērtības ir mazākas par 1. Maksimālajai atribūta vērtībai normalizētā vērtība ir 1.

- *Normalizācija pret vidējo vērtību:*

$$a_{ij}^n = \frac{a_{ij}}{\bar{a}_j}, \tag{5.2}$$

kur a_{ij} – objekta o_i sākotnējā atribūta a_j vērtība;

$\bar{a}_j$ – atribūta a_j vidējā vērtība.

Normalizācija ar vienādojuma (5.2) palīdzību ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Piemērs. Par pamatu ņem sākotnējās atribūtu a_j vērtības no 5.1. tabulas. Vidējā atribūta a_j vērtība ir:

$$\bar{a}_j = \frac{3+5+7+4+8+9}{6} = \frac{36}{6} = 6.$$

Sākotnējās atribūtu a_j vērtības, kas normalizētas ar vienādojumu (5.2.), ir parādītas 5.3. tabulā.

5.3. tabula

Atribūtu vērtību a_j normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.2)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
a_{ij}^n	0.500	0.833	1.167	0.667	1.333	1500

Jāatzīmē, ka, normalizējot ar vienādojumu (5.2), normalizētās vērtības sākotnējām vērtībām, kas ir mazākas par tās vidējo vērtību, ir mazākas par 1 un sākotnējām vērtībām, kas lielākas par vidējo vērtību, ir lielākas par 1.

- *Normalizācija pret atribūtu vērtību standarta novirzi:*

$$a_{ij}^n = \frac{a_{ij}}{s}, \tag{5.3}$$

kur a_{ij} – objekta o_i sākotnējā atribūta a_j vērtība;

s_j – atribūtu vērtību standarta novirze a_j no tās vidējās vērtības:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n}\left(a_{ij} - \bar{a}_j\right)^2}{n-1}}.$$

Normalizācija ar vienādojuma (5.3) palīdzību ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Piemērs. Par pamatu ņem sākotnējās atribūtu vērtības no 5.1. tabulas. Šī atribūta vidējā vērtība ir $\bar{a}_j = 6$. Standartnovirzes vērtība ir:

$$s_j = \sqrt{\frac{(3-6)^2 + (5-6)^2 + (7-6)^2 + (4-6)^2 + (8-6)^2 + (9-6)^2}{5}} =$$

$$= \sqrt{\frac{9+1+1+4+4+9}{5}} = \sqrt{\frac{28}{5}} = \sqrt{5.600} = 2.366.$$

Atribūta a_j sākotnējās vērtības un tā normalizētās vērtības ir parādītas 5.4. tabulā.

5.4. tabula

Atribūtu a_j vērtību normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.3)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
a_{ij}^n	1.268	2.113	2.958	1.691	3.381	3.804

2. Metožu grupa, kas izmanto normalizēšanu attiecībā pret atribūtu vērtību starpībām

- *Normalizēšana attiecībā pret vērtību starpību* $a_{j\max} - a_{j\min}$.

Šai metodei ir divas variācijas:

$$a_{ij}^n = \frac{a_{ij}}{a_{j\max} - a_{j\min}}; \tag{5.4}$$

$$a_{ij}^n = \frac{\left(a_{ij} - a_{j\min}\right)}{\left(a_{j\max} - a_{j\min}\right)}; \tag{5.5}$$

Vienādojumā (5.4) un (5.5):

a_{ij} – objekta o_i atribūta a_j sākotnējā vērtība;

$a_{j\min}$ – minimālā atribūta a_j vērtība;

$a_{j\max}$ – maksimālā atribūta a_j vērtība.

Normalizācija ar vienādojuma (5.4) palīdzību ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Normalizācija ar vienādojuma (5.5) palīdzību ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā vai intervālu skalā. Normalizācijas rezultāti tiek izteikti intervālu skalā.

Piemērs. Par pamatu ņem sākotnējās atribūtu vērtības no 5.1. tabulas.

No šīs tabulas $a_{j\min} = 3$, $a_{j\max} = 9$, $a_{j\max} - a_{j\min} = 9 - 3 = 6$. Sākotnējo atribūtu vērtību normalizēšanas rezultāti pēc vienādojuma (5.4) ir parādīti 5.5. tabulā.

5.5. tabula

Atribūtu a_j vērtību normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.4)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
a_{ij}^n	0.500	0.833	1.167	0.667	1.333	1500

Jāatzīmē, ka normalizēšanas rezultāti 5.5. tabulā sakrīt ar normalizēšanas rezultātiem attiecībā pret vidējo vērtību 5.3. tabulā. Šajā gadījumā tā ir sakritība, jo vērtību atšķirība $a_{j\max} - a_{j\min} = 6$ sakrīt ar atribūta vidējo vērtību $\bar{a}_j = 6$.

Atribūta sākotnējās vērtības normalizē, izmantojot vienādojumu (5.5). Normalizācijas rezultāti ir parādīti 5.6. tabulā.

5.6. tabula

Atribūtu a_j vērtību normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.5)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
$a_{ij} - a_{j\min}$	0	2	4	1	5	6
a_{ij}^n	0	0.333	0.667	0.167	0.833	1000

Normalizācijas vienādojuma (5.5) raksturīga iezīme ir tāda, ka atribūta normalizētās vērtības vienmēr atrodas intervālā no 0 līdz 1. Pie vērtības $a_{ij} = a_{j\min}$ vienādojuma (5.5) skaitītājs ir vienāds ar 0 un normalizācijas rezultāts arī ir vienāds ar 0. Pie vērtības $a_{ij} = a_{j\max}$ vienādojuma (5.5) skaitītājs un saucējs ir vienādi un normalizācijas rezultāts ir 1.

3. Normalizācijas metožu grupa attiecībā pret atribūtu vērtību summām

Pirmā metode šajā grupā ietver sākotnējo atribūtu a_j vērtību normalizēšanu attiecībā pret visu atribūtu vērtību summu:

$$a_{ij}^n = \frac{a_{in}}{\sum_{i=1}^{m} a_{ij}}. \qquad (5.6)$$

Izmantojot šo metodi, atribūtu a_j vērtības jāmēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Piemērs. Par pamatu ņem sākotnējās atribūtu a_j vērtības no 5.1. tabulas. No šīs tabulas:

$\sum_{i=1}^{6} a_{ij} = 3+5+7+4+8+9 = 36$.

Normalizācijas rezultāti ir parādīti 5.7. tabulā.

5.7. tabula

Atribūtu vērtību a_j normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.6)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
a_{ij}^n	0.083	0.139	0.194	0.111	0.222	0.250

Otrā metode no šīs grupas ietver sākotnējo atribūtu vērtību normalizēšanu attiecībā pret kvadrātsakni no atribūtu vērtību kvadrātu summas:

$$a_{ij}^n = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} a_{ij}^2}}. \qquad (5.7)$$

Normalizācija ar vienādojumu (5.7) ir iespējama, ja atribūtu a_j vērtības mēra attiecību skalā. Normalizācijas rezultātus izsaka arī attiecību skalā.

Piemērs. Par pamatu ņem sākotnējās atribūtu a_j vērtības no 5.1. tabulas. No šīs tabulas:

$$\sum_{i=1}^{6} a_{ij}^2 = 3^2 + 5^2 + 7^2 + 4^2 + 8^2 + 9^2 = 9 + 25 + 49 + 16 + 64 + 81 = 244.$$

$$\sqrt{\sum_{i=1}^{6} a_{ij}} = \sqrt{244} = 15.620.$$

Normalizācijas rezultāti ir parādīti 5.8. tabulā.

5.8. tabula

Atribūtu vērtību a_j normalizācijas rezultāti 5.1. tabulas datiem pēc vienādojuma (5.7)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
a_{ij}^n	0.192	0.320	0.448	0.256	0.512	0.576

Pastāv arī citas metodes atribūtu vērtību normalizēšanai. Normalizējot kritēriju vērtības lēmumu pieņemšanas problēmās, ir jāņem vērā tas, ka daži kritēriji atspoguļo ieguvumus, bet citi – izmaksas (zaudējumus), kas saistītas ar alternatīvu lēmumu rezultātiem.

Šajā nodaļā šīs metodes netiek aplūkotas, jo galvenā uzmanība tiek pievērsta datu pirmapstrādes metodēm. Interesentiem ieteicams meklēt darbu [Vafaei N. et al., 2018], lai iegūtu detalizētāku informāciju par vērtēšanas kritēriju normalizēšanas metodēm lēmumu pieņemšanas problēmās.

5.3. Sākotnējo atribūtu vērtību standartizācijas metodes

Visizplatītākā ir standartizācijas metode, kuras pamatā ir atribūtu vērtību sadalījuma parametri: vidējā vērtība un standartnovirze:

$$a_{ij}^s = \frac{a_{ij} - \bar{a}_j}{s_j}, \qquad (5.8)$$

kur a_{ij} – objekta o_i sākotnējā atribūta a_j vērtība;

$\bar{a}_j$ – atribūta a_j vidējā vērtība;

s_j – atribūtu vērtību a_j standarta novirze no tās vidējās vērtības.

Standartizāciju ar vienādojumu (5.8) var veikt, ja atribūtu a_j vērtības mēra attiecību skalā vai intervālu skalā. Standartizācijas rezultāti tiek izteikti intervālu skalā. Standartizācijas ar vienādojumu (5.8) raksturīga iezīme ir tāda, ka standartizēto atribūtu vērtību sagaidāmā vērtība ir 0 un to standartnovirze ir 1.

Piemērs. Par pamatu ņem sākotnējās atribūtu a_j vērtības no 5.1. tabulas. Iepriekšējā sadaļā tika veikti šādi tā vērtību sadalījuma parametru aprēķini:

$\bar{a}_j = 6$; $s_j = 2.366$.

Atribūtu a_j vērtības standartizē pēc vienādojuma (5.8). Standartizācijas rezultāti ir parādīti 5.9. tabulā.

5.9. tabula

Atribūtu vērtību a_j standartizācijas rezultāti 5.1. tabulas datiem tabulas pēc vienādojuma (5.8)

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
$a_{ij} - \overline{a}_j$	-3	-1	1	-2	2	3
a_{ij}^s	-1.268	-0.423	0.423	-0.845	0.845	1.268

Alternatīva metode atribūtu vērtību standartizēšanai ir standartizācija, kuras pamatā ir atribūtu vērtību sadalījuma mediāna un novirze no mediānas (Vēbera standartizācija):

$$a_{ij}^s = \frac{\left(a_{ij} - med\left(a_j\right)\right)}{1.4826 mad\left(a_j\right)}, \tag{5.9}$$

kur a_{ij} – objekta o_i sākotnējā atribūta a_j vērtība;

$med\left(a_j\right)$ – atribūtu vērtību a_j sadalījuma mediāna;

$mad\left(a_j\right)$ – atribūtu vērtību a_j noviržu kopas mediāna no mediānas $med\left(a_j\right)$.

Standartizāciju pēc vienādojuma (5.9) var veikt, ja atribūtu a_j vērtības mēra attiecību skalā vai intervālu skalā. Standartizācijas rezultāti tiek izteikti intervālu skalā.

Vērtību $med\left(a_j\right)$ aprēķina šādi: vispirms n atribūtu vērtības a_{ij} jāsakārto un jāindeksē augošā secībā: $a_{(1)j}$, $a_{(2)j}$,..., $a_{(n)j}$. Vērtība $med\left(a_j\right)$ tiek aprēķināta, izmantojot vienādojumus:

$$med\left(a_j\right) = a_{(n+1)j/2}, \; n \text{ - nepāra;} \tag{5.9a}$$

$$med\left(a_j\right) = \frac{1}{2}\left(a_{(k)j} + a_{(k+1)j}\right), \; n = 2k \text{ - pāra.} \tag{5.9b}$$

Ja n ir nepāra skaitlis, tad par $med\left(a_j\right)$ vērtību tiek ņemta atribūta a_j vērtība, kas atrodas $\left(n+1\right)/2$ pozīcijā tā vērtību sakārtotajā secībā. Ja n ir pāra skaitlis, tad vērtība $med\left(a_j\right)$ tiek ņemta par vidējo vērtību no atribūta a_j vērtībām, kas atrodas k un $\left(k+1\right)$ pozīcijā vērtību sakārtotajā secībā.

Pēc vērtības $med\left(a_j\right)$ aprēķināšanas, izmantojot vienādojumus (5.9a) vai (5.9b), tiek aprēķinātas atribūtu vērtību a_j novirzes no vērtības $med\left(a_j\right)$. Vērtība $mad\left(a_j\right)$ tiek aprēķināta, izmantojot vienādojumus (5.9a) vai (5.9b), bet atribūta vērtību a_j vietā tagad tiek izmantotas šī atribūta novirzes vērtības no mediānas $med\left(a_j\right)$.

Piemērs. Par pamatu ņem atribūtu a_j vērtības no 5.9. tabulas. Lai aprēķinātu $med\left(a_j\right)$, šīs vērtības sakārto augošā secībā: 3, 4, 5, 7, 8, 9.

Tā kā ir pāra skaits atribūtu vērtību – 6, aprēķina vērtību $med\left(a_j\right)$, izmantojot vienādojumu (5.9b):

$$med\left(a_j\right) = \frac{1}{2}\left(a_{(3)j} + a\left(a_{(4)j}\right)\right) = \frac{1}{2}\left(5+7\right) = 6.$$

Aprēķini par atribūtu a_j vērtību novirzēm no mediānas $med\left(a_j\right)$ ir parādīti 5.10. tabulā.

5.10. tabula

Atribūtu vērtību a_j noviržu no mediānas $med\left(a_j\right)$ aprēķini 5.9. tabulas datiem

a_{ij}	3	4	5	7	8	9
$a_{ij} - med\left(a_j\right)$	3	2	1	1	2	3

Novirzes vērtības sakārto augošā secībā: 1, 1, 2, 2, 3, 3.

Vērtību $mad\left(a_j\right)$ aprēķina, izmantojot vienādojumu (5.9b):

$$mad\left(a_j\right) = \frac{1}{2}\left(a_{(3)j} - med\left(a_j\right) + a_{(4)j} - med\left(a_j\right)\right) = \frac{1}{2}\left(2+2\right) = 2.$$

Vienādojuma (5.9) saucējs ir vienāds ar $1.4826 * 2 = 2.9652$.

Atribūtu a_j vērtības standartizē pēc vienādojuma (5.9). Starpaprēķinu rezultāti un standartizētās atribūtu vērtības ir parādītas 5.11. tabulā.

5.11. tabula

Starpaprēķinu rezultāti un standartizētās atribūtu a_{ij} vērtības 5.9. tabulas datiem

Objekti	o_1	o_2	o_3	o_4	o_5	o_6
a_{ij}	3	5	7	4	8	9
$a_{ij} - med\left(a_{ij}\right)$	-3	-1	1	-2	2	3
a_{ij}^s	-1.012	-0.337	0.337	-0.674	0.674	1.012

6. ATRIBŪTU VĒRTĪBU TRANSFORMĀCIJA

6.1. Kad un kāpēc tiek izmantotas transformācijas?

Datu analīzes veikšana, izmantojot sākotnējās atribūtu vērtības, dažkārt var būt problemātiska dažādās īpašās situācijās. Pieņemsim, ka atribūtu a_j vērtības atrodas intervālā $[10,10000]$. Acīmredzot darboties ar šādām vērtībām ir ļoti neērti. Risinājums var būt šī atribūta vērtību transformēšana citā piemērotā mērogā. Ja sākotnējo atribūtu a_{ij} vērtību vietā izmantosim vērtības $\lg(a_{ij})$, tad jaunajā skalā atribūtu vērtības atradīsies intervālā $[1,4]$. Ja decimāllogaritmu vietā mēs izmantojam naturālos logaritmus, tad jaunās atribūtu vērtības atradīsies intervālā $[2.303,9.210]$. Jebkurā gadījumā darbība ar transformētām atribūtu vērtībām būs vienkāršāka un ērtāka nekā darbība ar sākotnējām vērtībām.

Šāda veida atribūtu vērtību transformēšana bieži tiek izmantota praksē. Piemēram, ūdeņraža jonu koncentrācija (pH) vienmēr tiek izteikta logaritmiskā skalā tieši šī parametra sākotnējo vērtību lielās izkliedes dēļ.

Bieži ikdienā mēs izmantojam dažādas transformācijas, pat nedomājot par to. Ja mēs mainām noteiktu eiro summu uz Polijas zlotiem, mēs formāli veicam vienā skalā mērītas naudas summas (eiro) lineāru pārveidošanu līdzvērtīgā citā skalā (zlotos).

Datu pirmapstrādes un analīzes procesos tiek izmantotas sarežģītākas atribūtu vērtību nelineāras transformācijas. Kādi ir šādu pārveidojumu mērķi? Problēmas ieskatam der apskatīt dažus ilustratīvus piemērus.

Statistisko parametru datu analīze balstās uz *apriori* pieņēmumiem par datu struktūru un īpašībām. Viens no šādiem pieņēmumiem ir tāds, ka attiecīgo atribūtu vērtības ir normāli sadalītas. Ja tas tā patiešām ir, tad daudzas statistiskās analīzes procedūras var veikt standarta veidā: analizēt atribūtu vidējās vērtības, novērtēt atribūtu vērtību izplatību attiecībā pret to vidējām vērtībām, konstruēt ticamības intervālus, analizēt variācijas starp paraugiem utt.

6.1. attēlā parādītas trīs nosacītās atribūtu a_j, a_k, a_l vērtību sadalījuma histogrammas.

Histogramma 6.1.a attēlā ataino to, ka atribūtu a_j vērtības ir sadalītas saskaņā ar likumu, kas ir tuvs normālam sadalījumam. Šīm vērtībām var piemērot jebkuru interesējošo parametru statistiskās analīzes procedūru.

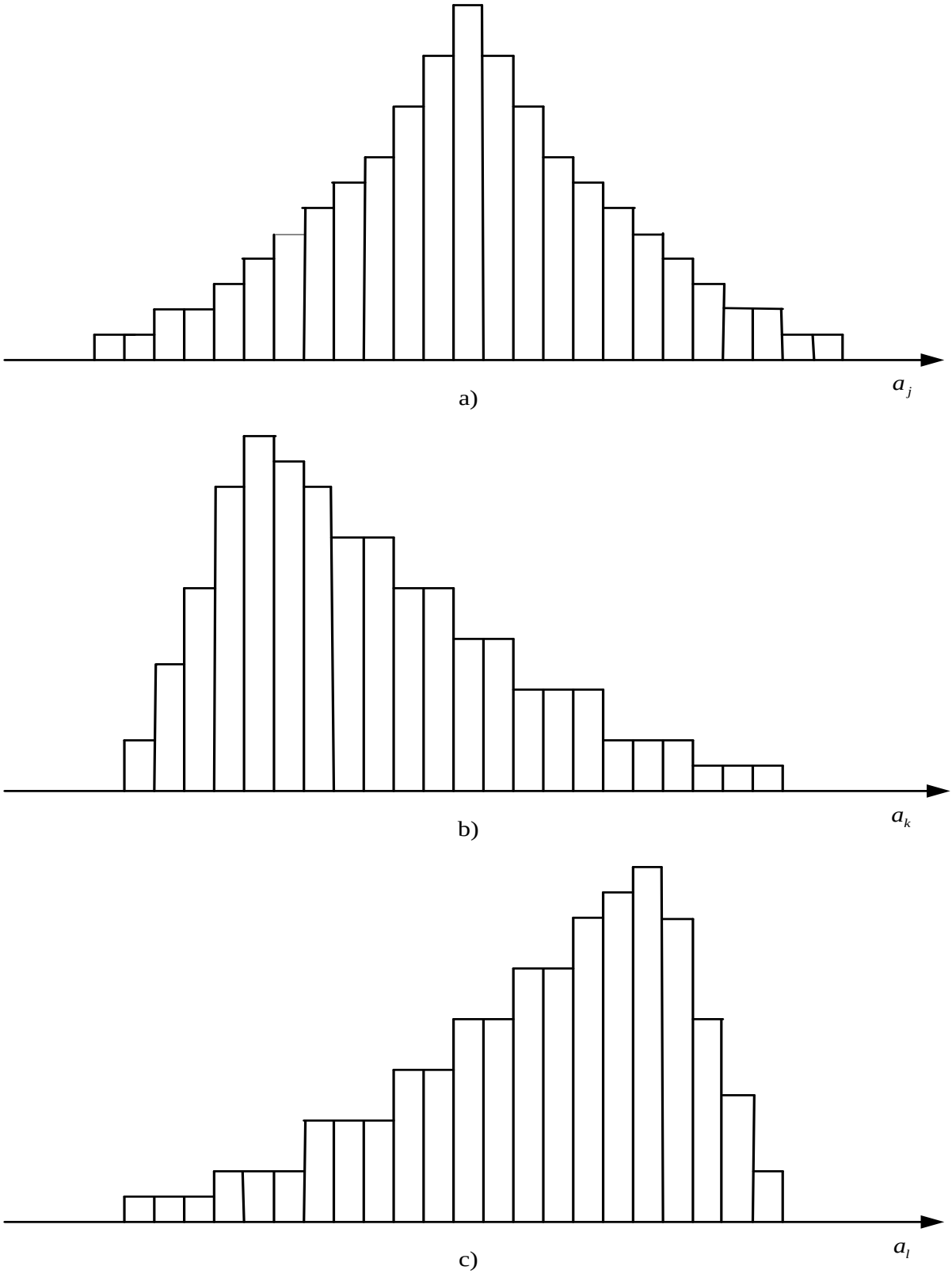
Histogramma 6.1.b attēlā norāda uz to, ka, palielinoties atribūtu a_k vērtībām, to parādīšanās biežums sākotnēji strauji palielinās. Pēc maksimālās vērtības sasniegšanas notiek pakāpeniska frekvenču samazināšanās. Rezultātā tiek iegūts asimetrisks slīpums, uz labo pusi vērsts atribūtu a_k vērtību sadalījums.

Grafiks 6.1.c attēlā norāda uz to, ka sākumā atribūtu a_l vērtību parādīšanās biežums lēnām palielinās. Pēc maksimālās vērtības sasniegšanas notiek strauja frekvenču samazināšanās. Rezultātā tiek iegūts asimetrisks slīpums, uz kreiso pusi vērsts sadalījums.

Ir acīmredzams, ka atribūtu a_k, a_l vērtību sadalījums ir tālu no normālā sadalījuma un tiem nevar piemērot statistiskās analīzes procedūras, kas ir derīgas normāliem sadalījumiem.

Ja histogrammas ir līdzīgas tām, kas parādītas 6.1. attēlā, tad, veicot šo histogrammu vizuālo analīzi, var izdarīt aptuvenus secinājumus par šo sadalījumu būtību. Reālās situācijās attiecīgajiem atribūtiem var būt ļoti liela vērtību kopa, kas attēlota datu tabulas (matricas) kolonnās.

Kā šādus atribūtu vērtību sadalījumus var pārbaudīt, vai tie ir normāli? Pielikumā P3 ir parādītas izmantojamās parametriskās un neparametriskās metodes, lai pārbaudītu sadalījumu atbilstību normālajam sadalījumam.



6.1. attēls. **Atribūtu a_j, a_k, a_l sadalījumu histogrammas**

Kādi rīcības varianti ir iespējami gadījumos, kad attiecīgā atribūtu vērtību sadalījums ir līdzīgs 6.1.b vai 6.1.c attēlā redzamajiem? Darbā [Sakia R., 1992] tiek piedāvātas šādas iespējamās darbības:

1) ignorēt pieņēmumu pārkāpumus un veikt analīzi tā, it kā visi pieņēmumi būtu izpildīti;
2) izlemt, kāds ir pareizais pieņēmums pārkāptā vietā un izmantot procedūru, kurā tiek ņemts vērā jaunais pieņēmums;

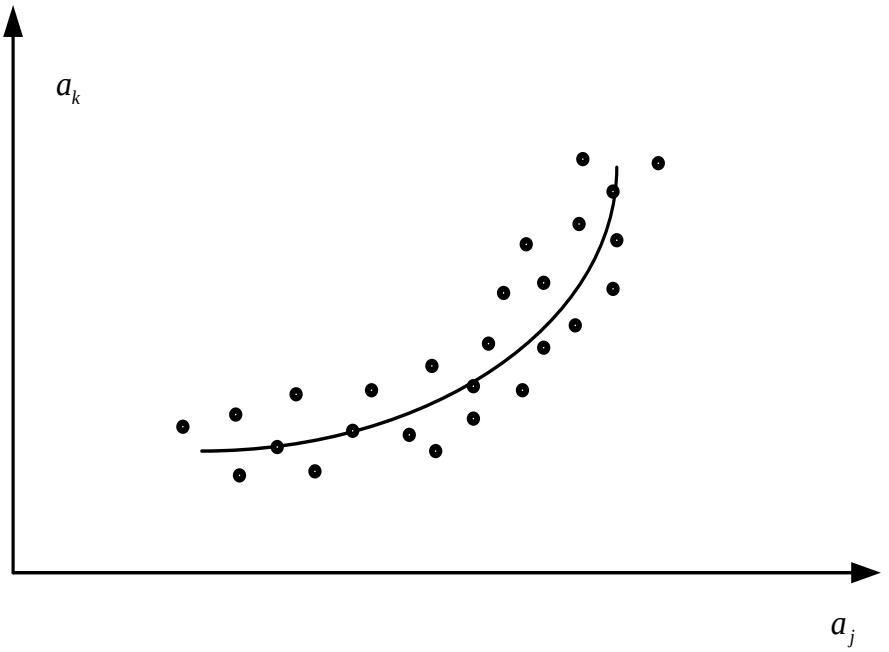
3) izstrādāt jaunu modeli, kam ir svarīgi sākotnējā modeļa aspekti un kas atbilst pieņēmumiem, piemēram, izmantojot efektīvu datu transformāciju, vai filtrēt dažus aizdomīgus punktus, kurus var uzskatīt par anomālijām.

Acīmredzot 1. iespēja jebkurā gadījumā nav pieņemama. Atribūtu vērtību statistiskās analīzes veikšana, pieņemot, ka tās ir normāli sadalītas, lai gan patiesībā nav, var dot kļūdainus analīzes rezultātus. 2. variants ir diezgan pieņemams, ja attiecīgā atribūta vērtību sadalījums precīzi vai aptuveni atbilst kādam likumam, kas nav normāls sadalījums.

Šķiet, ka 2. variants ir visdaudzsološākais. Ja ir iespējams pārveidot atribūtu a_k , a_l vērtības no 6.1.a,b attēla tādā veidā, ka to transformētās vērtības tiks aptuveni sadalītas saskaņā ar normālo sadalījuma likumu, tad problēmu var uzskatīt par atrisinātu. Nepieciešamo analīzi var veikt transformētajām atribūtu vērtībām. Ja nepieciešams, analīzes rezultātus var apgriezti pārveidot sākotnējo atribūtu vērtību telpā. Šī vajadzība bieži ir saistīta ar analītisko rezultātu interpretāciju. Lietotājam var būt ļoti grūti interpretēt analīzes rezultātus transformēto atribūtu vērtību telpā. Apgrieztā transformācija ļauj atrisināt šo problēmu.

Tādējādi ir iezīmēts viens atribūtu vērtību transformācijas mērķis. Atribūtu vērtību sadalījumam, kas ir vērsts pa kreisi vai pa labi, vajag pārveidot šīs vērtības tā, lai transformētajām vērtībām sadalījums būtu pēc iespējas tuvāks normālajam sadalījumam.

Cits virziens, kurā plaši tiek izmantotas atribūtu vērtību transformācijas, redzams 6.2. attēlā.



6.2. attēls. **Datu punktu grafiskais attēlojums atribūtu a_j un a_k vērtību telpā un regresijas sakarība starp šiem punktiem**

Kā izriet no 6.2. attēla, atribūtu a_j , a_k vērtības ir savstarpēji saistītas (starp to vērtību kopām pastāv korelācija). Lai analizētu šādus datus, ir šādas iespējas:

- mēģināt modelēt atkarības starp atribūtu vērtībām, izmantojot regresijas vienādojumus;
- transformēt sākotnējās atribūtu vērtības tā, lai atkarības starp to pārveidotajām vērtībām varētu modelēt, izmantojot lineāro regresiju.

Daudzās reālās datu apstrādes un analīzes problēmās priekšroka dodama 2. variantam.

Jāpiezīmē, ka gan sākotnējo sadalījumu normalizēšanas gadījumā, gan regresijas atkarības linearizācijas gadījumā sākotnējo atribūtu vērtību transformācija nav brīnumlīdzeklis, kas ļauj sasniegt vēlamos rezultātus visos gadījumos. Sākotnējo atribūtu vērtību transformācijas var un ir jāveic, ja iegūtie rezultāti ir piemēroti turpmākai datu analīzei. Ja plaši pazīstamu datu transformācijas

metožu izmantošana neļauj iegūt vēlamos rezultātus, jāizmanto citas (t. sk. no parametriem neatkarīgas) datu apstrādes un analīzes metodes.

Skaidrības labad jāievieš formāla definīcija: "Datu transformācijas ir matemātisko modifikāciju pielietošana mainīgo vērtībām. Ir daudz dažādu datu transformāciju, sākot no konstantu pievienošanas līdz reizināšanai, celšana kvadrātā vai citā pakāpē, konvertēšana logaritmiskajās skalās, invertēšanai un attēlošanai, kā arī vērtību kvadrātsaknes vilkšanu" [Osborne J.W, 2007].

Pieminēšanas vērti ir divi citāti no darba [Lee D.K., 2020], kas akcentē atribūtu vērtību transformācijas vietu un lomu datu pirmapstrādes kontekstā.

"Dažādi pieņēmumi tādi kā normalitāte, lineārās attiecības bieži ir nepieciešami parametriskās analīzes metodēs. Dati, kas savākti no eksperimentiem, bieži pārkāpj šos pieņēmumus. Mainīgo lielumu transformācija sniedz iespēju bez sistemātiskām kļūdām padarīt datus pieņemamus parametriskai statistiskai analīzei. Mainīgo lielumu transformācijas uzdevums ir ļaut veikt parametrisku statistisko analīzi, kuras galvenais mērķis ir laba rezultātu interpretācija ar transformētajiem mainīgajiem. Mainīgo lielumu transformācija parasti maina to vērtību sākotnējos raksturlielumus. Apgrieztā transformācija ir būtiska rezultātu interpretācijai.

... Lielākajai daļai parametriskās statistiskās analīzes metožu ir nepieciešami pieņēmumi par normālo sadalījumu. Ja tie tiek pārkāpti, statistikas rezultāti no nenormāliem sadalījumiem var izraisīt nopietnas kļūdas... Lai gan centrālā robežas teorēma var aptvert nenormalitāti, ja izlases lielums ir pietiekams, daudzi eksperimentālie dati neatbilst pieņēmumam par normālo sadalījumu, neskatoties uz to, ka ir salīdzinoši liels datu izlases lielums. Par laimi, vienkārša statistikas pieeja, mainīgo transformācija, nodrošina metodi datu sadalījuma pārveidošanai no nenormāla uz normālu. Turklāt mainīgo lielumu transformācija var veidot lineāras attiecības starp mainīgajiem no nelineārās attiecības un var stabilizēt aprēķinātās lineārās modelēšanas variācijas".

"Atbilstoši datu sadalījuma raksturlielumiem var izmantot dažādas transformācijas metodes, lai panāktu normalitātes testa izpildi. Šāda veida transformācijas var padarīt datus simetriski sadalītus. Šīs transformācijas metodes var izmantot, lai nodrošinātu lineāras attiecības starp mainīgajiem. Ir labi zināms, ka daudzas statistiskās modelēšanas metodes balstās uz lineārām sakarībām un lineāras attiecības veidošana starp transformētajiem mainīgajiem, kas atvieglo modeļa statistisko novērtēšanu. Tipisks piemērs ir *logit* transformācija, ko izmanto binomiālajai loģistikas regresijai. *Logit* transformācija pārvērš notikumu varbūtības koeficientu logaritmos, ļaujot veikt regresijas analīzi starp rezultāta mainīgo un neatkarīgo mainīgo, kas darbojas kā lineārs prognozētājs".

Darbā [McCune B., Grace J. 2002] tiek apgalvots, ka lielākajai daļai datu to transformācijām ir priekšrocības – tiek uzlaboti pieņēmumi par normālo sadalījumu, linearitāti, variācijām utt.; atribūtu vērtības kļūst salīdzināmas, kad tās tiek mērītas dažādos mērogos.

6.2. Izplatītākie transformāciju veidi

1. Logaritmisko transformāciju saime

Jebkuras logaritmiskās transformācijas būtība ir aizstāt sākotnējās atribūtu a_j vērtības ar šo vērtību logaritmiem:

$$a_{ij}^t = \log\left(a_{ij}\right). \tag{6.1}$$

Šādas transformācijas ļoti bieži izmanto dažādos bioloģiskos un medicīnas pētījumos. Kādi ir tam iemesli? Pirmkārt, komponentu koncentrācijas dažādos bioloģiskajos paraugos var atšķirties daudzkārtīgi. Darbība ar šādiem datiem ir ļoti sarežģīta un neērta. Šādu datu pārveidošana logaritmiskā formā ļauj ievērojami samazināt transformēto vērtību izmaiņu diapazonu, kas ievērojami vienkāršo datu apstrādi un analīzi. Otrkārt, daudzu bioloģisko rādītāju sākotnējo vērtību variācijas ir proporcionālas to paredzamo vērtību pakāpēm. Šādu rādītāju logaritmiskās transformācijas ir optimālas tādā ziņā, ka ļauj iegūt transformēto vērtību normālus sadalījumus vai tuvu normālam sadalījumam.

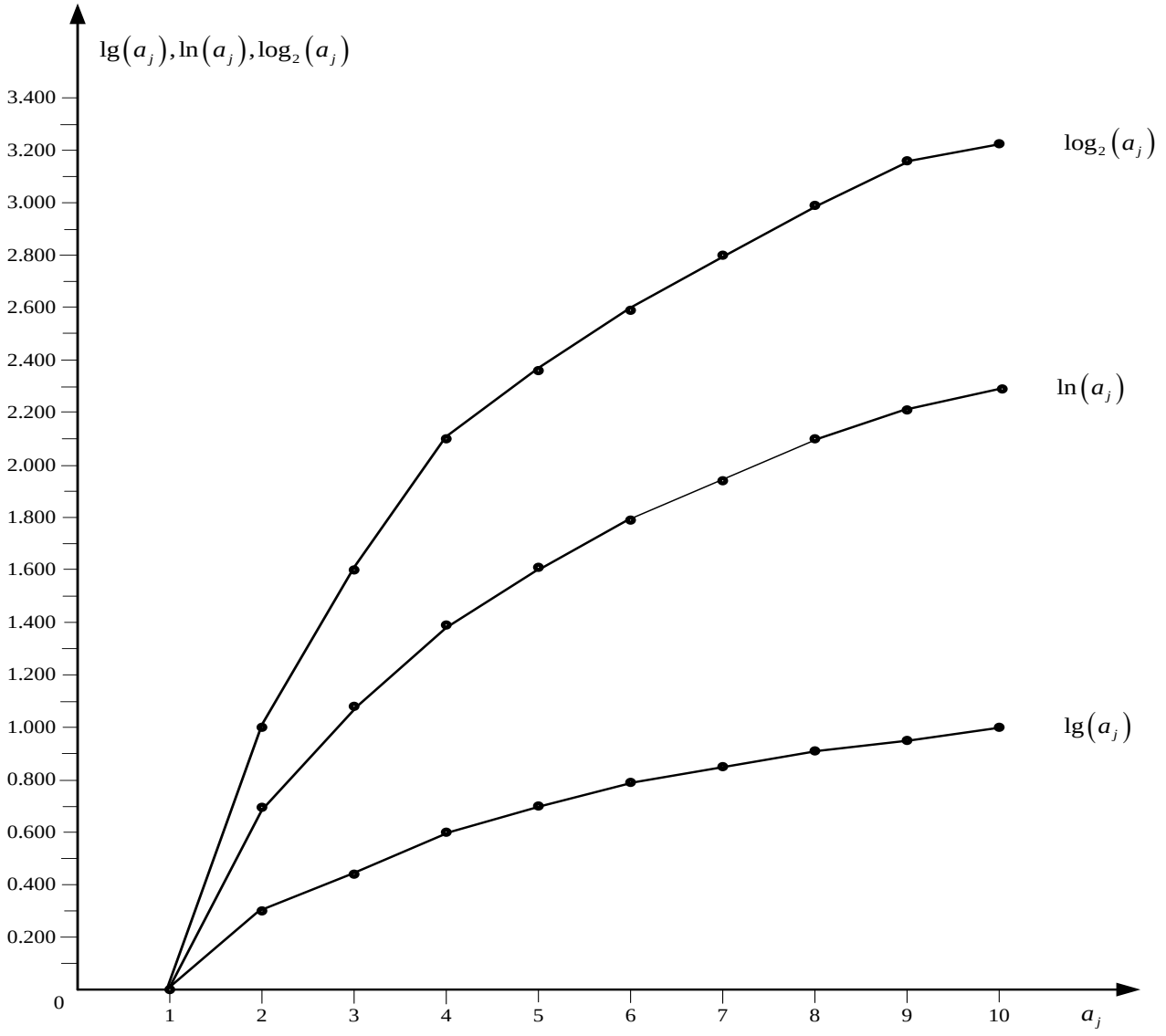
Šeit ir runa par logaritmisko pārveidojumu saimi, jo šādas transformācijas var veikt, izmantojot logaritmus ar dažādām bāzēm: logaritmus ar bāzi 10 (decimāllogaritmi), logaritmus ar bāzi $e \approx 2.718$ (naturālie logaritmi) un logaritmus ar bāzi 2 (binārie logaritmi).

Lai uzskatāmi nodemonstrētu pārveidoto vērtību atkarību no izmantotajām logaritmu bāzēm, 6.1. tabulā ir parādītas sākotnējās atribūtu a_j vērtības diapazonā $[1,10]$ un to transformētās vērtības, izmantojot logaritmus ar dažādām bāzēm. 6.3. attēlā transformāciju rezultāti tiek attēloti grafiskā veidā.

6.1. tabula

Sākotnējās atribūtu a_j vērtības un to transformētās vērtības ar dažādām logaritmu bāzēm

a_j	1	2	3	4	5	6	7	8	9	10
$\lg(a_{ij})$	0	0.301	0.477	0.602	0.699	0.778	0.845	0.903	0.954	1000
$\ln(a_{ij})$	0	0.693	1.097	1.386	1.609	1.792	1.946	2.079	2.197	2.302
$\log_2(a_{ij})$	0	1000	1.585	2000	2.322	2.585	2.807	3000	3.169	3.222



6.3. attēls. **Sākotnējo un logaritmiski transformēto atribūtu a_j vērtību datiem no 6.1. tabulas grafisks attēlojums**

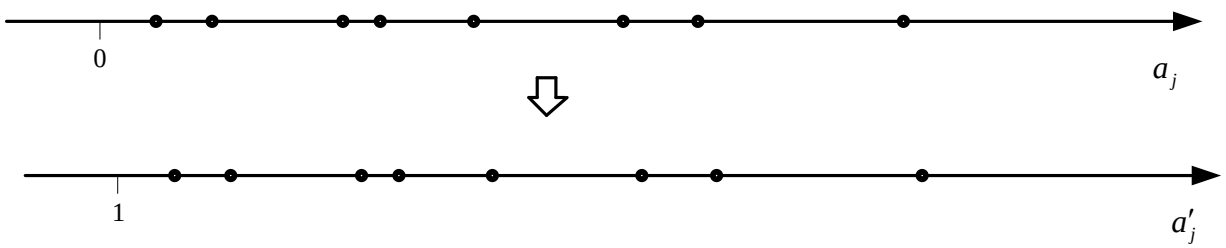
Kādus secinājumus var izdarīt, analizējot 6.3. attēlu? Acīmredzot visas transformācijas saglabā tādu pašu atribūta a_j transformēto vērtību secību, kāda pastāv tā sākotnējo vērtību kopā. Ja $a_{ij} > a_{kj}$, tad $\lg(a_{ij}) > \lg(a_{kj})$, $\ln(a_{ij}) > \ln(a_{kj})$ un $\log_2(a_{ij}) > \log_2(a_{kj})$. Tomēr sakarības starp sākotnējo un transformēto atribūtu vērtībām nav lineāras.

Atšķirības starp atribūta transformētajām vērtībām nav tieši proporcionālas atšķirībām starp šī atribūta atbilstošajām sākotnējām vērtībām. Palielinoties atribūta sākotnējām vērtībām a_j, atšķirības starp tā transformētajām vērtībām mēdz samazināties. Tieši šī logaritmisko transformāciju īpašība tiek izmantota, lai pārveidotu sākotnējos atribūtu vērtību sadalījumu, kas ir asimetrisks, uz labo pusi vērsts sadalījums, formā, kas ir tuvu normālajam sadalījumam.

Ja visas atribūtu a_j vērtības ir lielākas par 1, tad šo vērtību transformācija logaritmiskā formā nav problēma. Ja visas atribūtu a_j vērtības ir stingri lielākas par 0, tad logaritmiski transformētās vērtības sākotnējām atribūtu vērtībām, kas ir mazākas par 1, būs negatīvas. Lai izvairītos no negatīvām transformētajām vērtībām, var piemērot šādu sākotnējo atribūtu vērtību mērogošanu:

$$a'_{ij} = a_{ij} + 1. \tag{6.2}$$

Šī mērogošanas maiņa shematiski parādīta 6.4. attēlā.



6.4. attēls. **Sākotnējās atribūtu a_j vērtības mērogošanas procesa shematisks attēlojums pēc vienādojuma (6.2)**

Ja nav vēlamas ne tikai negatīvas, bet arī nulles transformētās atribūtu a_j vērtības, tad var izmantot šādu sākotnējo atribūtu vērtību mērogošanu:

$$a''_{ij} = a_{ij} + 1 + \varepsilon, \tag{6.3}$$

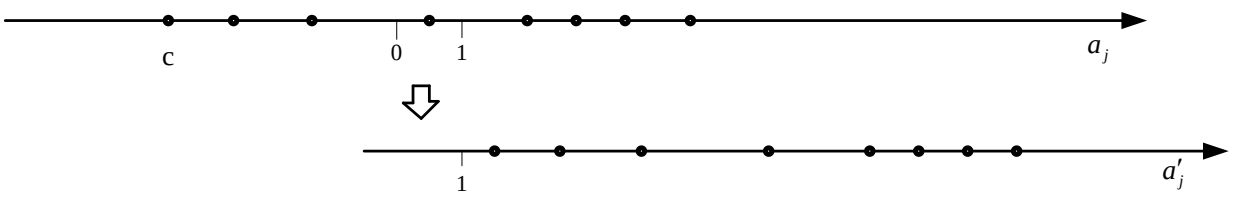
kur ε – kļūdas lielums (mazs pozitīvs skaitlis, piemēram, $\varepsilon = 0.01$ vai $\varepsilon = 0.001$).

Problēmas rodas gadījumos, ja dažas no atribūta a_j sākotnējām vērtībām ir negatīvas. Lai nodrošinātu to, ka visas transformētās atribūtu vērtības nav negatīvas, var izmantot šādu sākotnējo atribūtu vērtību mērogošanu:

$$a'_{ij} = a_{ij} + c + 1, \tag{6.4}$$

kur c – atribūta a_j lielākā negatīvā vērtība.

Šādas mērogošanas maiņas shēma ir parādīta 6.5. attēlā.



6.5. attēls. **Sākotnējās atribūtu a_j vērtības mērogošanas procesa shematisks attēlojums pēc vienādojuma (6.4)**

Šajā attēlā simbols c apzīmē atribūta a_j lielāko negatīvo vērtību. Pievienojot šai vērtībai summu $c+1$, panāk to, ka jaunajā skalā vērtība c atbildīs 1 un visas atribūta vērtības būs pozitīvas.

Ja transformētās atribūtu a_j nulles vērtības nav vēlamas, var izmantot šādu mērogošanu:

$$a''_{ij} = a_{ij} + c + 1 + \varepsilon, \tag{6.5}$$

kur parametra ε nozīme ir izskaidrota vienādojumā (6.3).

Jebkuras logaritmiskās transformācijas svarīga iezīme ir ievērojams transformēto atribūtu vērtību diapazona samazinājums, salīdzinot ar tā sākotnējo vērtību diapazonu.

Transformāciju, izmantojot decimāllogaritmus, var pielietot atribūtiem ar ekstremāliem vērtību diapazoniem. Tad transformēto vērtību diapazons ļoti būtiski mainās. Tomēr jāņem vērā, ka arī transformēto vērtību kopas mērogs ievērojami samazinās. Piemēram, ir divas sākotnējo atribūtu a_j vērtības: 1500 un 1600. Transformējot šīs vērtības ar decimāllogaritmiem, iegūst vērtības: 3,176 un 3,204. Atšķirība starp šīm vērtībām ir ļoti maza, kas var radīt problēmas datu analīzes laikā.

Nelieliem sākotnējo atribūtu vērtību diapazoniem ir vēlama transformācija, izmantojot naturālos logaritmus vai logaritmus ar bāzi 2. Lietojot šādus logaritmus, iegūst arī ievērojamu samazinājumu transformēto atribūtu vērtību diapazonos, savukārt relatīvi tuvu transformēto vērtību mērogs nesamazinās tik daudz kā decimāllogaritmu lietošanas gadījumā.

2. Pakāpes transformāciju saime

Vispārīgā gadījumā transformētās atribūtu a_j vērtības var iegūt formā:

$$a_{ij}^t = \left(a_{ij}\right)^\alpha, \ \alpha \neq 0. \tag{6.6}$$

Pie dažādām α vērtībām iegūst dažādus transformācijas rezultātus:

- $\alpha = 3$ – kubiskā transformācija;
- $\alpha = 2$ – kvadrātiskā transformācija;
- $\alpha = 1$ – transformētās atribūtu vērtības sakrīt ar sākotnējām vērtībām (acīmredzot šādai transformācijai nav jēgas);
- $\alpha = 1/2$ – kvadrātsaknes transformācija.

Pie vērtībām $\alpha < 1$ notiek lielo atribūtu vērtību "saspiešana". Šo pakāpes transformācijas īpašību var izmantot, lai normalizētu asimetriskus, uz labo pusi vērstus sadalījumus.

6.2. tabulā parādīti sākotnējo atribūtu a_j vērtību pakāpes transformācijas pie dažādām α vērtībām.

6.2. tabula

Transformētās atribūtu a_j vērtības pie dažādiem α

a_j	1	2	3	4	5	6	7	8	9	10
$\alpha = 2$	1	4000	9000	16 000	25 000	36 000	49 000	64 000	81 000	100 000
$\alpha = 1.5$	1	2.828	5.196	8000	11.180	14.697	18.520	22.627	27 000	31.623
$\alpha = 0.5$	1	1.414	1.732	2000	2.236	2.449	2.645	2.829	3000	3.162
$\alpha = 0.25$	1	1.189	1.136	1.414	1.495	1.565	1.625	1.682	1.732	1.778

Tabulā var redzēt ka, pie $\alpha > 1$ palielinoties atribūta a_j sākotnējām vērtībām, to transformētās vērtības strauji palielinās. Šajā gadījumā atribūtu vērtībām $a_{ij} > 1$ transformētās atribūtu vērtības pārsniedz sākotnējās vērtības.

Pie $\alpha < 1$, palielinoties atribūta a_j sākotnējām vērtībām, palielinās arī to transformētās vērtības. Bet šis pieaugums notiek lēni un visas transformētās vērtības (izņemot $a_{ij} = 1$) ir mazākas par atbilstošām sākotnējām atribūtu vērtībām.

3. Box-Cox transformāciju saime

Šis pakāpes transformācijas variants pirmo reizi tika piedāvāts darbā [Box G.E.P., Cox D.R., 1964] tikai pozitīvām sākotnējo atribūtu vērtībām. Vēlāk to attiecināja arī uz negatīvām atribūtu vērtībām. Sākotnējo atribūtu vērtību transformācijas var veikt ar vienādojumu palīdzību:

- visas atribūtu a_j vērtības ir pozitīvas:

$$a_{ij}^t(\lambda) = \frac{\left(a_{ij}\right)^{\lambda} - 1}{\lambda},\ \lambda \neq 0;$$
$$= \log\left(a_{ij}\right),\ \lambda = 0, \tag{6.7}$$

kur λ – transformācijas parametrs. Vērtības λ var variēt diapazonā $[-5,5]$.

- atribūtu a_j vērtības ir negatīvas:

$$a_{ij}^t(\lambda) = \frac{\left(a_{ij} + \lambda_2\right)^{\lambda_1} - 1}{\lambda_1},\ \lambda_1 \neq 0;$$
$$\log\left(a_{ij} + \lambda_2\right), \tag{6.8}$$

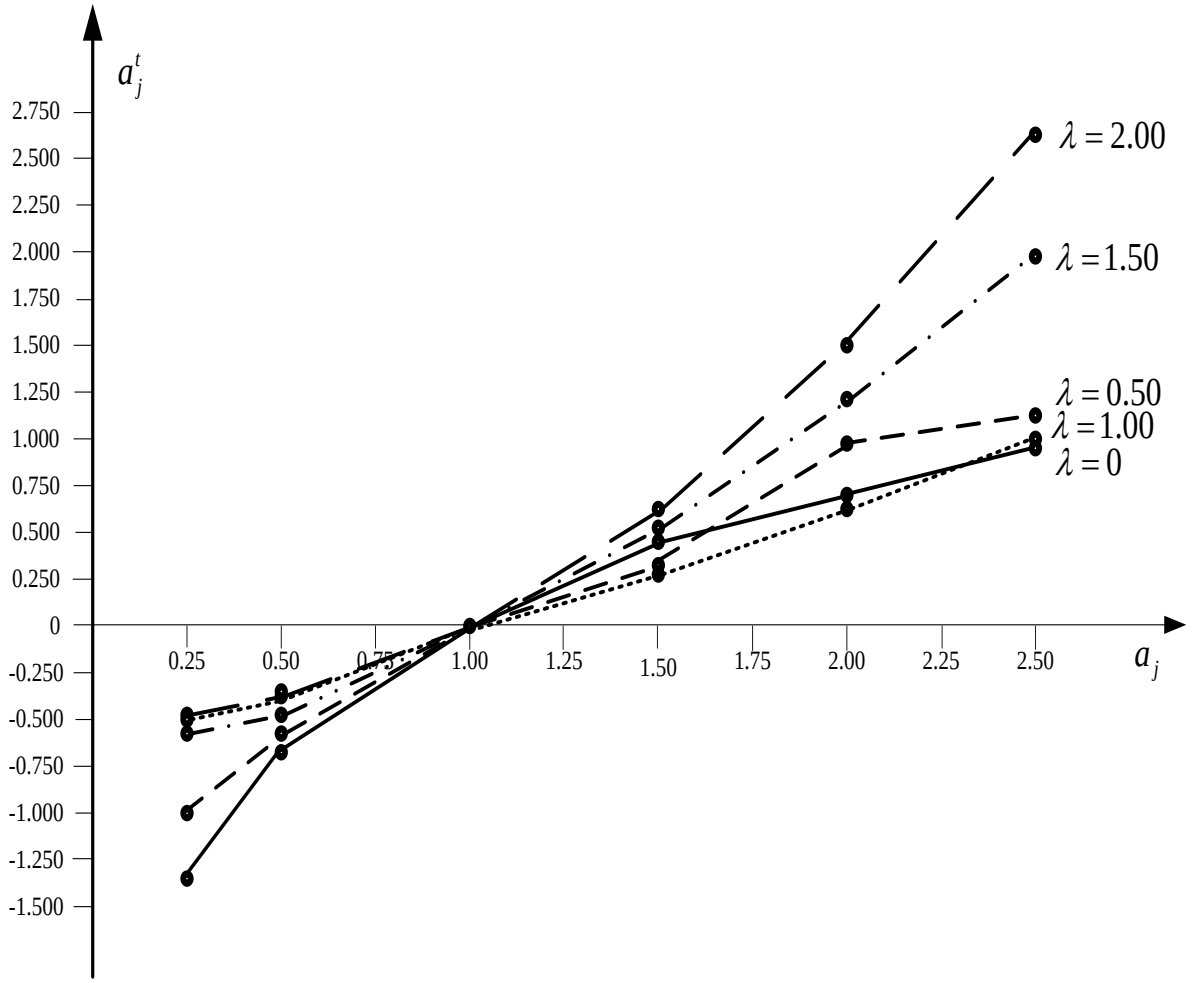
kur λ_2 – parametrs, ko izmanto sākotnējo atribūtu vērtību skalas izmaiņai, lai visas vērtības būtu lielākas par 1.

6.3. tabulā parādītas sākotnējās atribūtu a_j vērtības un to transformētās vērtības pie dažādām λ vērtībām. Transformētās atribūtu a_j vērtības ir grafiski attēlotas 6.6. attēlā. No 6.3. tabulas un 6.6. attēla var izdarīt pamatotu secinājumu, ka, norādot dažādas λ vērtības, ir iespējams nodrošināt ļoti plašas sākotnējo atribūtu a_j vērtību transformēšanas iespējas, izmantojot *Box-Cox* transformācijas.

6.3. tabula

Sākotnējo atribūtu a_j vērtību transformācijas rezultāti saskaņā ar vienādojumu (6.7) dažādām λ vērtībām

a_j	0.25	0500	1.00	1.50	2.00	2.50
$\lambda = 2.00$	-0.469	-0.375	0	0.625	1500	2.625
$\lambda = 1.50$	-0.583	-0.431	0	0.558	1.219	1.969
$\lambda = 1.00$	-0.500	-0.333	0	0.333	0.666	1000
$\lambda = 0.50$	-1000	-0.586	0	0.449	0.828	1.162
$\lambda = 0$	-1.386	-0.693	0	0.405	0.693	0.961



6.6. attēls. **Sākotnējo un transformēto atribūtu a_j vērtību grafisks attēlojums datiem no 6.3. tabulas**

6.3. Atribūtu vērtību transformācija normālā sadalījuma sasniegšanai

Ir piedāvātas dažādas pieejas šādu transformāciju veikšanai. Uz labo pusi vērstiem sadalījumiem bieži izmanto logaritmiskās transformācijas. Piemēra pēc tiks izmantots logaritmiskās transformācijas atribūts a_k vērtībām no P3.2.2. tabulas (pielikums P3.2). Šis sadalījums ir tipisks uz labo pusi vērsta sadalījuma pārstāvis. P3.2.2. un 6.1. tabulā sniegto datu kopsavilkums ir parādīts 6.4. tabulā.

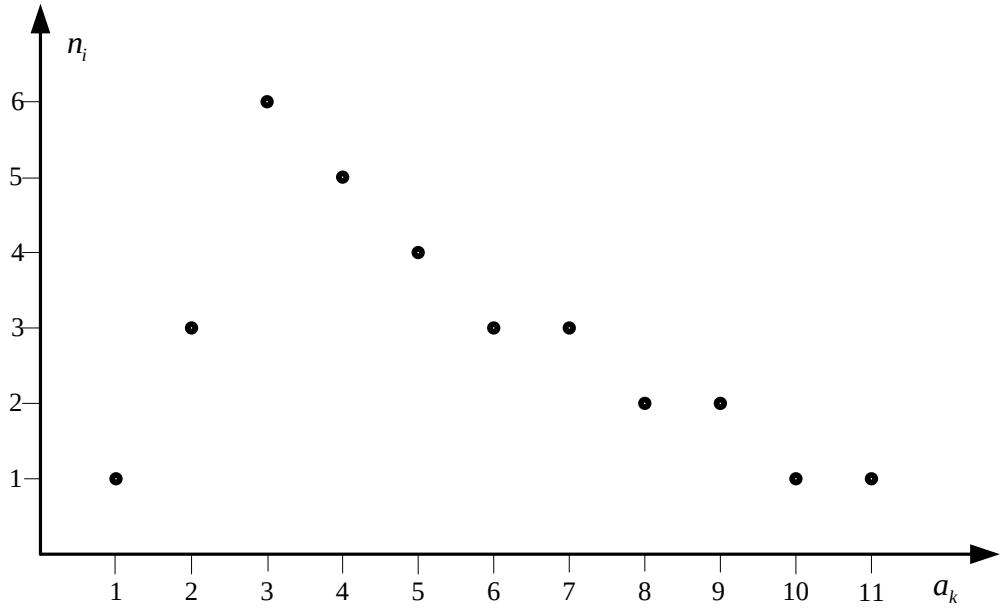
6.4. tabula

Vispārināts datu kopsavilkums no P3.2.2. un 6.1. tabulas

a_k	1	2	3	4	5	6	7	8	9	10	11
$\lg(a_k)$	0	0.301	0.477	0.602	0.699	0.778	0.845	0.903	0.954	1000	1.041
$\ln(a_k)$	0	0.693	1.097	1.386	1.609	1.792	1.946	2.079	2.197	2.302	2.398
$\log_2(a_k)$	0	1000	1.585	2000	2.322	2.585	2.807	3000	3.169	3.222	3.458
n_i	1	3	6	5	4	3	3	2	2	1	1

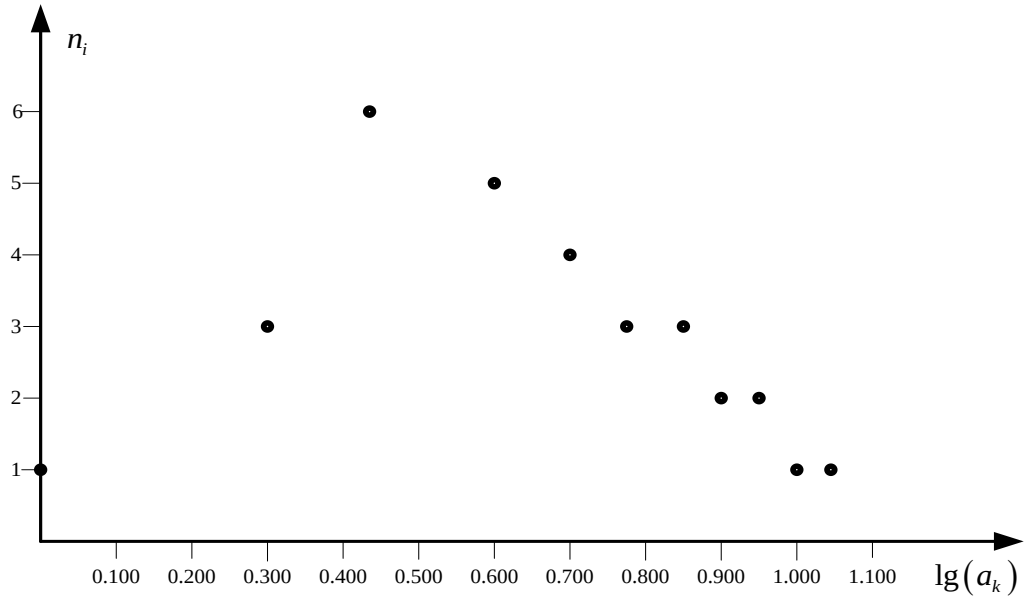
No P3.2.2. tabulas ir ņemtas atribūta sākotnējās vērtības a_k un tā vērtību atkārtošanās gadījumu skaits n_i noteiktos intervālos.

Vērtības $\lg(a_k)$, $\ln(a_k)$, $\log_2(a_k)$ ir ņemtas no 6.1. tabulas un pievienotas vērtības $a_k = 11$. P3.2.2. attēls parāda atribūtu vērtību a_k histogrammu datiem no P3.2.2. tabulas. Iegūto rezultātu tālākas salīdzināšanas nolūkos 6.7. attēlā parādīta atribūtu a_k vērtību sadalījuma histogramma atbilstoši 6.4. tabulas datiem.



6.7. attēls. Atribūtu a_k vērtību sadalījuma histogramma datiem no 6.4. tabulas

6.8. attēlā parādīta vērtību $\lg(a_k)$ sadalījuma histogramma.



6.8. attēls. **Vērtību $\lg(a_k)$ sadalījuma histogramma datiem no 6.4. tabulas**

6.5. tabulā sniegti statistikas dati par vērtību $\lg(a_k)$ sadalījumiem datiem no 6.4. tabulas. Turpmākajos aprēķinos tiks izmantoti pakotnes *IBM SPSS Statistica* aprēķini (detalizētāk skatīt pielikumā P3).

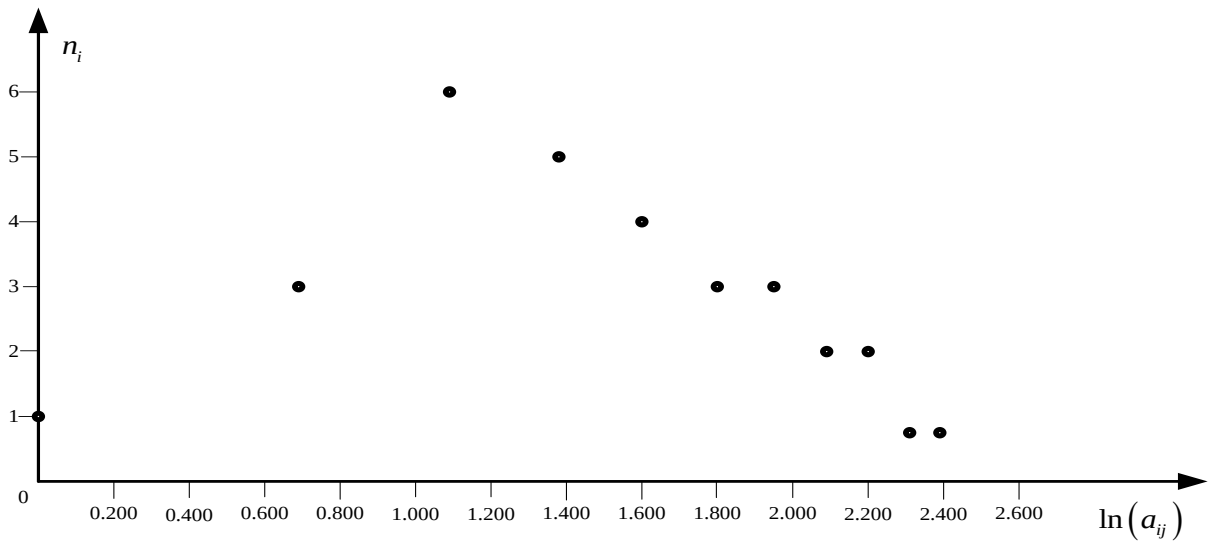
6.5. tabula

Vērtību $\lg(a_k)$ sadalījuma statistikas kopsavilkums datiem no 6.4. tabulas

Descriptives

		Statistic	Std. Error
Mean		,64887	,042686
95% Confidence Interval for Mean	Lower Bound	,56169	
	Upper Bound	,73605	
5% Trimmed Mean		,65801	
Median		,69900	
Variance		,056	
Std. Deviation		,237668	
Minimum		,000	
Maximum		1,041	
Range		1,041	
Interquartile Range		,368	
Skewness		-,603	,421
Kurtosis		,361	,821

6.9. attēlā parādīta histogramma vērtību $\ln(a_k)$ sadalījumam datiem no 6.4. tabulas.



6.9. attēls. **Vērtību $\ln(a_k)$ sadalījuma histogramma datiem no 6.4. tabulas**

6.6. tabulā sniegti statistikas dati par vērtību $\ln(a_k)$ sadalījumu 6.4. tabulas datiem.

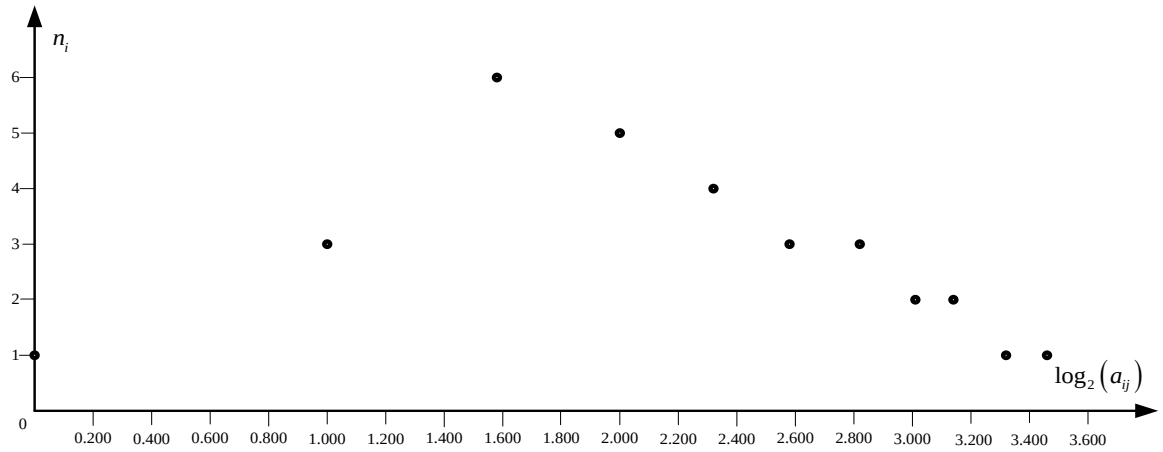
6.6. tabula

Vērtību $\ln(a_k)$ sadalījuma statistikas kopsavilkums datiem no 6.4. tabulas

Descriptives

		Statistic	Std. Error
Mean		1,49977	,099650
95% Confidence Interval for Mean	Lower Bound	1,29626	
	Upper Bound	1,70329	
5% Trimmed Mean		1,52142	
Median		1,60900	
Variance		,308	
Std. Deviation		,554826	
Minimum		,000	
Maximum		2,398	
Range		2,398	
Interquartile Range		,849	
Skewness		-,562	,421
Kurtosis		,282	,821

6.10. attēlā parādīta histogramma vērtību $\log_2(a_k)$ sadalījumam datiem no 6.4. tabulas.



6.10. attēls. **Vērtību $\log_2(a_k)$ sadalījuma histogramma datiem no 6.4. tabulas**

6.7. tabulā sniegti statistikas dati par vērtību $\log_2(a_k)$ sadalījumu datiem no 6.4. tabulas.

6.7. tabula

Vērtību $\log_2(a_k)$ sadalījuma statistikas kopsavilkums datiem no 6.4. tabulas

Descriptives

		Statistic	Std. Error
Mean		2,16103	,142847
95% Confidence Interval for Mean	Lower Bound	1,86930	
	Upper Bound	2,45276	
5% Trimmed Mean		2,19397	
Median		2,32200	
Variance		,633	
Std. Deviation		,795337	
Minimum		,000	
Maximum		3,458	
Range		3,458	
Interquartile Range		1,222	
Skewness		-,588	,421
Kurtosis		,308	,821

Analizējot iegūtos rezultātus, 6.8. tabulā ir parādītas sākotnējās un transformētās atribūtu a_k vērtību sadalījuma asimetrijas rādītāja (*skewness*) un ekscesa rādītāja (*kurtosis*) vērtības. Šo sadalījuma parametru definīcijas ir sniegtas pielikumā P3.3. Sākotnējo atribūtu vērtību asimetrijas un ekscesa vērtības ir ņemtas no tabulas pielikumā P3.4.

6.8. tabula

Asimetrijas un ekscesa vērtības oriģinālajiem un transformētajiem a_k atribūtiem

	a_k	$\lg(a_k)$	$\ln(a_k)$	$\log_2(a_k)$
Asimetrija	0.581	0.361	0.282	0.308
Ekscess	-0.427	-0.603	-0.562	-0.588

Visiem transformēto atribūtu a_k vērtību sadalījumiem asimetrijas vērtības ir mazākas nekā sākotnējo atribūtu vērtību sadalījuma vērtības. Transformējot ar binārajiem logaritmiem, tika iegūta vismazākā asimetrijas vērtība. Tas nozīmē, ka asimetrijas korekcijas ziņā šī transformācija ir vislabākā.

Kas attiecas uz ekscesu, visi transformēto atribūtu a_k vērtību sadalījumi ir tālu no normālā sadalījuma nekā tā sākotnējo vērtību sadalījums.

Jāpiebilst, ka izmantojot logaritmiskās transformācijas, mēs sasniedzām savu galveno mērķi – samazinājām sākotnējā sadalījuma asimetriju.

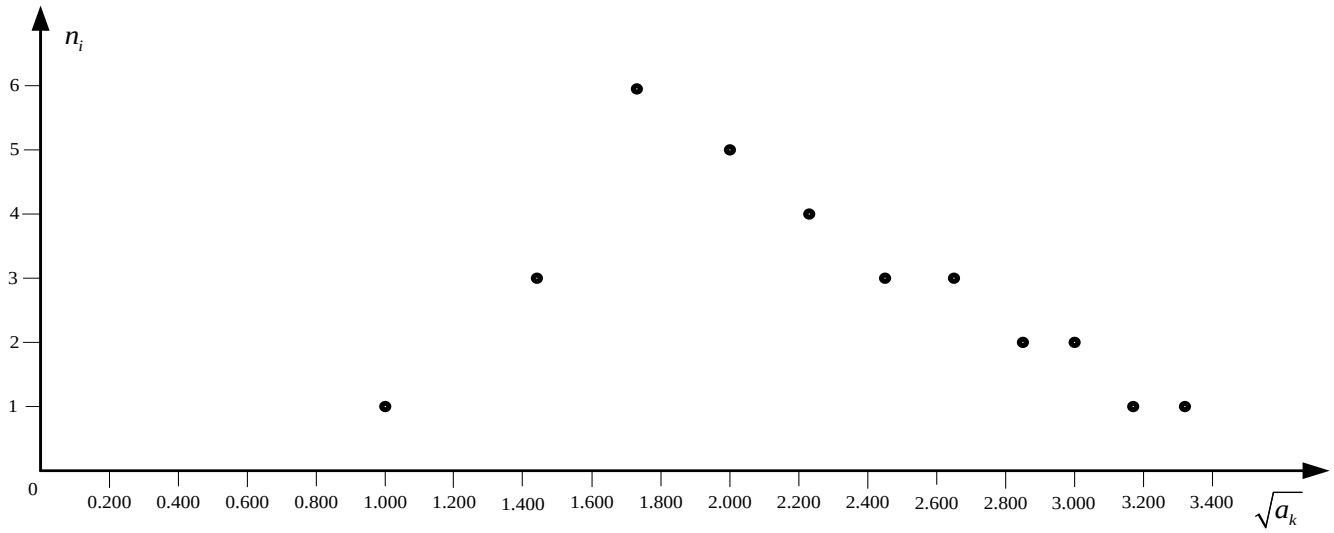
Uz labo pusi vērstajiem sadalījumiem ieteicams izmantot arī transformācijas ar pakāpes funkcijām ar pakāpēm, kas <1. Piemēram, 6.9. tabulā ir parādīti sākotnējo atribūtu a_k vērtību pakāpju transformāciju rezultāti ar pakāpēm $1/2$ un $1/3$. Pretējā gadījumā transformētās atribūtu a_k vērtības tiek aprēķinātas, ņemot kvadrātsakni vai kubsakni no sākotnējām vērtībām.

6.9. tabula

Sākotnējo atribūtu a_k vērtību transformācijas rezultāti, izmantojot kvadrātsakni un kubsakni

a_k	1	2	3	4	5	6	7	8	9	10	11
$\sqrt{a_k}$	1	1.414	1.732	2000	2.236	2.449	2.646	2.829	3000	3.162	3.317
$\sqrt[3]{a_k}$	1	1.260	1.442	1.589	1.709	1.816	1.912	2000	2.078	2.153	2.222
n_i	1	3	6	5	4	3	3	2	2	1	1

Transformēto atribūtu $\sqrt{a_k}$ vērtību histogramma parādīta 6.11. attēlā.



6.11. attēls. **Vērtību $\sqrt{a_k}$ sadalījuma histogramma datiem no 6.9. tabulas**

6.10. tabulā sniegti statistikas dati par vērtību $\sqrt{a_k}$ sadalījumu.

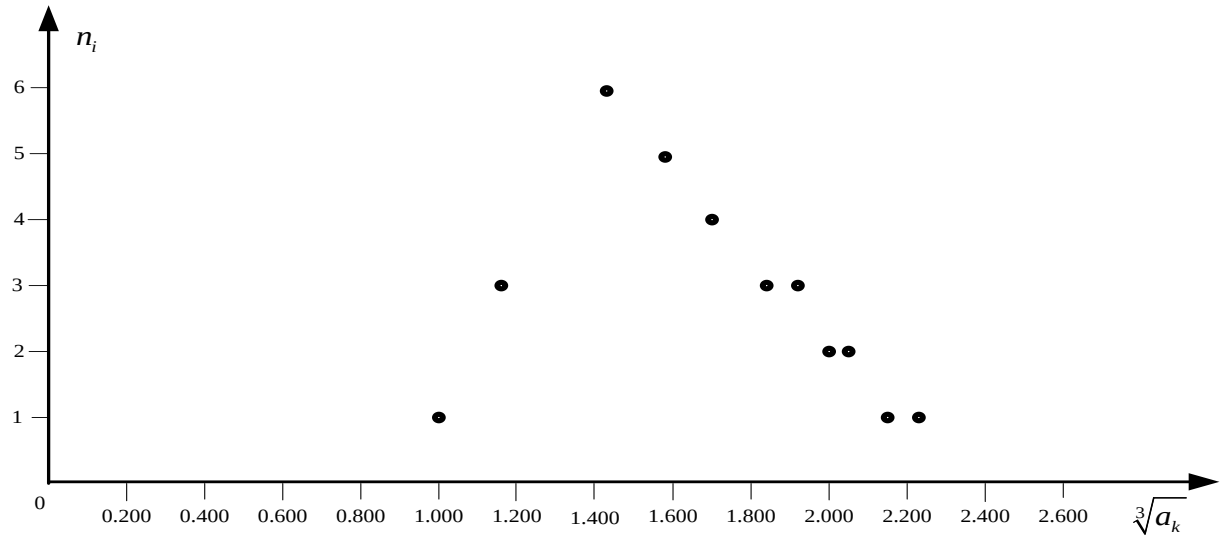
6.10. tabula

Vērtību $\sqrt{a_k}$ sadalījuma statistikas kopsavilkums datiem no 6.9. tabulas

Descriptives

		Statistic	Std. Error
Mean		2,19355	,102855
95% Confidence Interval for Mean	Lower Bound	1,98349	
	Upper Bound	2,40361	
5% Trimmed Mean		2,19234	
Median		2,23600	
Variance		,328	
Std. Deviation		,572672	
Minimum		1,000	
Maximum		3,317	
Range		2,317	
Interquartile Range		,914	
Skewness		,089	,421
Kurtosis		-,599	,821

Transformēto atribūtu $\sqrt[3]{a_k}$ vērtību histogramma parādīta 6.12. attēlā.



6.12. attēls. **Vērtību $\sqrt[3]{a_k}$ sadalījuma histogramma datiem no 6.9. tabulas**

6.11. tabulā sniegti statistikas dati par vērtību $\sqrt[3]{a_k}$ sadalījumu.

6.11. tabula

Vērtību $\sqrt[3]{a_k}$ sadalījuma statistikas kopsavilkums datiem no 6.9. tabulas

Descriptives

		Statistic	Std. Error
Mean		1,67510	,053194
95% Confidence Interval for Mean	Lower Bound	1,56646	
	Upper Bound	1,78373	
5% Trimmed Mean		1,67845	
Median		1,70900	
Variance		,088	
Std. Deviation		,296173	
Minimum		1,000	
Maximum		2,222	
Range		1,222	
Interquartile Range		,470	
Skewness		-,111	,421
Kurtosis		-,457	,821

Analizējot iegūtos rezultātus, 6.12. tabulā ir parādītas asimetrijas un ekscesa vērtības atribūta sākotnējo vērtību a_k sadalījumam un transformētajām vērtībām $\sqrt{a_k}$, $\sqrt[3]{a_k}$.

6.12. tabula

Asimetrijas un ekscesa vērtības sākotnējam un transformētajam atribūtu a_k sadalījumiem

	a_k	$\sqrt{a_k}$	$\sqrt[3]{a_k}$
Ekscess	-0. 427	-0. 599	-0. 457
Asimetrija	-0.581	0.086	-0.11

Transformēto atribūtu sadalījuma asimetrijas vērtība ievērojami samazinājās, salīdzinot ar sākotnējo sadalījuma asimetrijas vērtību. Transformācijai $\sqrt[3]{a_k}$ nozīmē to, ka asimetrijas vērtība pamainīja savu zīmi un kļuva pozitīva, t. i., vērtību sadalījums kļuva vērsts uz kreiso pusi.

Transformēto atribūtu sadalījumiem ir lielas ekscesa negatīvas vērtības.

Pamatojoties uz šiem un iepriekš iegūtajiem rezultātiem, var droši teikt, ka gan logaritmiskās, gan pakāpes transformācijas ievērojami samazina sadalījumu asimetriju, bet mazākā mērā ietekmē šo sadalījumu ekscesu. Tā kā šādu transformāciju primārais mērķis ir samazināt sākotnējo sadalījumu asimetriju, tad abu veidu transformācijas var sasniegt šo mērķi.

Uz kreiso pusi vērsto sadalījumu transformācija ir sarežģītāka. Šādos gadījumos ieteicams izmantot pakāpes transformācijas ar pakāpju rādītājiem 2 un 3.

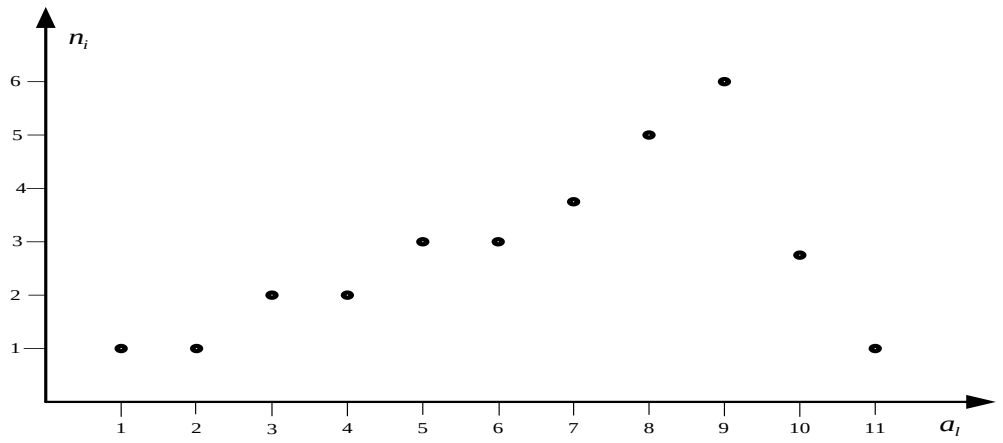
6.13. tabulā ir parādītas sākotnējās atribūtu a_l vērtības no P3.2.3. tabulas (pielikums P3.2), transformētās vērtības $(a_l)^2$ un $(a_l)^3$, un atbilstošo atribūtu vērtību biežums.

6.13. attēlā iegūto rezultātu salīdzināšanas un analīzes nolūkos parādīta sākotnējo atribūtu a_l vērtību sadalījuma histogramma.

6.13. tabula

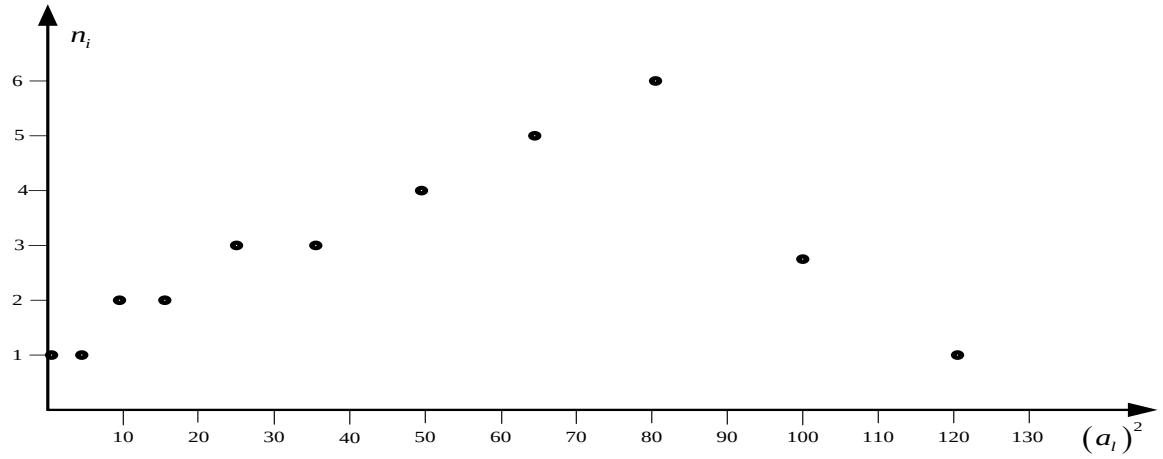
Sākotnējās atribūtu a_l vērtības, transformētās vērtības $(a_l)^2$, $(a_l)^3$ un biežums n_i

a_l	1	2	3	4	5	6	7	8	9	10	11
$(a_l)^2$	1	4	9	16	25	36	49	64	81	100	121
$(a_l)^3$	1	8	27	64	125	216	343	512	729	1000	1331
n_i	1	1	2	2	3	3	4	5	6	3	1



6.13. attēls. **Sākotnējā atribūtu a_l vērtību sadalījuma histogramma datiem no 6.13. tabulas**

6.14. attēlā parādīta transformēto vērtību $(a_l)^2$ sadalījuma histogramma.



6.14. attēls. **Transformēto vērtību $(a_l)^2$ sadalījuma histogramma datiem no 6.13. tabulas**

6.14. tabulā sniegti statistikas dati par vērtību $(a_l)^2$ sadalījumu.

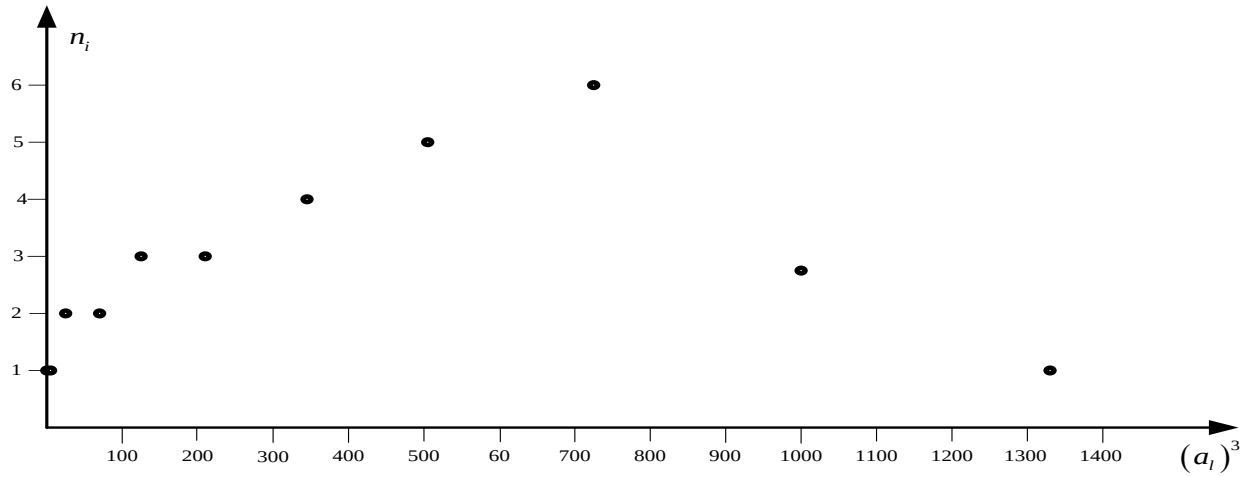
6.15. attēlā parādīta transformēto vērtību $(a_l)^3$ sadalījuma histogramma.

6.14. tabula

Vērtību $(a_l)^2$ sadalījuma statistikas kopsavilkums datiem no 6.13. tabulas

Descriptives

		Statistic	Std. Error
Mean		53,58	5,824
95% Confidence Interval for Mean	Lower Bound	41,69	
	Upper Bound	65,48	
5% Trimmed Mean		53,11	
Median		49,00	
Variance		1051,585	
Std. Deviation		32,428	
Minimum		1	
Maximum		121	
Range		120	
Interquartile Range		56	
Skewness		,103	,421
Kurtosis		-,898	,821



6.15. attēls. **Transformēto vērtību $(a_l)^3$ sadalījuma histogramma datiem no 6.16. tabulas**

6.15. tabulā sniegti statistikas dati par vērtību $(a_l)^3$ sadalījumu.

Analizējot iegūtos rezultātus, 6.16. tabulā ir parādītas asimetrijas un ekscesa vērtības sākotnējo atribūtu a_l vērtību sadalījumam un transformēto vērtību $(a_l)^2$, $(a_l)^3$ sadalījumiem. Parametru vērtības sākotnējo atribūtu vērtību sadalījumam ir ņemtas no tabulas pielikumā P3.4.

6.15. tabula

Vērtību $(a_l)^3$ sadalījuma statistikas kopsavilkums datiem no 6.13. tabulas

Descriptives

		Statistic	Std. Error
Mean		446,81	63,495
95% Confidence Interval for Mean	Lower Bound	317,13	
	Upper Bound	576,48	
5% Trimmed Mean		428,84	
Median		343,00	
Variance		124978,828	
Std. Deviation		353,523	
Minimum		1	
Maximum		1331	
Range		1330	
Interquartile Range		604	
Skewness		,614	,421
Kurtosis		-,319	,821

6.16. tabula

Asimetrijas un ekscesa vērtības sākotnējo un transformēto vērtību a_l sadalījumiem

	a_l	$(a_l)^2$	$(a_l)^3$
Ekscess	-0.427	-0.898	-0.319
Asimetrija	-0.581	0.103	-0.111

Transformēto atribūtu sadalījuma asimetrijas vērtības a_l ir mazākas par asimetrijas vērtībām priekš sākotnējo vērtību sadalījuma. Transformēto vērtību $(a_l)^2$ sadalījums tagad ir ar nelielu

asimetriju uz labo pusi. Tā kā praktiskās iespējas labot uz kreiso pusi vērsto sadalījumu ir ierobežotas, literatūrā tiek ieteikta šāda pieeja. Sākotnējo atribūtu a_{il} vērtību vietā iesaka izmantot vērtības $c - a_{il}$, kur c ir konstante. Šīs konstantes vērtība ir izvēlēta tā, lai varētu korekti aptvert visas sākotnējās atribūta a_l vērtības. Tādu transformāciju sauc par sākotnējā sadalījuma *attēlojumu.*

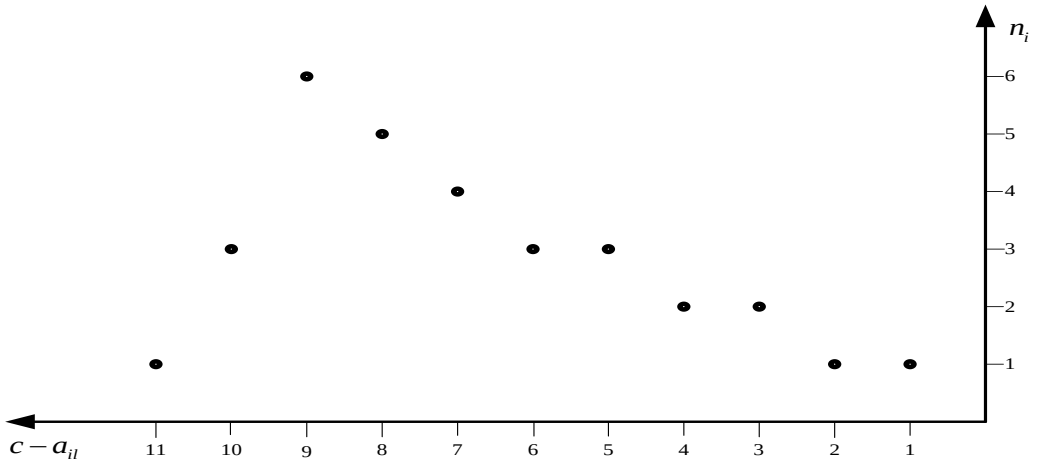
Sadalījuma attēlošanas procesu var demonstrēt ar piemēru. 6.13. tabulā ir parādītas sākotnējās atribūtu a_l vērtības. Šīs vērtības ir uzdotas 6.17. tabulas pirmajā rindā. Izvēlas konstantes vērtību $c = 12$ un no šī skaitļa atņem sākotnējās atribūtu vērtības. Transformētās vērtības $c - a_{il}$ ir parādītas 6.17. tabulas otrajā rindā. Biežuma vērtības tabulas pēdējā rindā attiecas uz transformētajām atribūtu vērtībām.

6.17. tabula

Sākotnējās un transformētās atribūtu a_l vērtības

a_l	1	2	3	4	5	6	7	8	9	10	11
$c - a_{il}$	11	10	9	8	7	6	5	4	3	2	1
n_i	1	3	6	5	4	3	3	2	2	1	1

Transformēto atribūtu vērtību histogramma ir parādīta 6.16. attēlā.



6.16. attēls. **Vērtību $c - a_{il}$ sadalījuma histogramma datiem no 6.17. tabulas**

Salīdzinot histogrammu 6.16. attēlā ar sākotnējo atribūtu a_l vērtību histogrammu 6.13. attēlā, var izdarīt acīmredzamu secinājumu, ka histogramma 6.16. attēlā ir 6.13. attēla histogrammas spoguļattēls. Bet vērtību $c - a_{il}$ sadalījums ir uz labo pusi vērsts sadalījums, un šim sadalījumam var izmantot iepriekš minētās un citas piemērotas normalizācijas metodes.

Izmantojot sākotnējā sadalījuma attēlošanas pieeju, ir jāpārliecinās, ka visi rezultāti attiecas tieši uz pārveidoto sadalījumu. Tāpēc ir nepieciešama iegūto rezultātu korekcija, lai tos piemērotu sākotnējam sadalījumam.

Apkopojot rezultātus, sākotnējo sadalījumu transformācijas, lai tos tuvinātu normāliem sadalījumiem, tiek veiktas dažādos veidos sadalījumiem, kas ir vērsti uz kreiso vai labo pusi.

Logaritmiskās transformācijas tiek plaši izmantotas, lai pārveidotu uz labo pusi vērstos sadalījumus. Šīs transformācijas ir īpaši piemērotas šādos gadījumos:

- iespējama eksponenciāla komponente datos;
- datu vērtības atšķiras pēc izkliedes pakāpes;
- galvenie efekti datos ir multiplikatīvi, bet ne aditīvi.

Uz labo pusi vērstu sadalījumu transformācijas ar kvadrātsaknes palīdzību ir piemērotas šādos gadījumos:

- dati var būt aprēķinu rezultāti vai biežumu vērtības;

- datos ir daudz nulles vērtību un ārkārtīgi mazas vērtības.

Uz labo pusi vērstiem sadalījumiem *Box-Cox* transformācijas dod labus rezultātus, taču optimālās λ vērtības izvēle ir diezgan sarežģīta.

Uz kreiso pusi vērstiem sadalījumu transformācijām plaši izmanto kvadrātiskās un kubiskās transformācijas. Var izmantot arī sākotnējā sadalījuma attēlojumu, kas dod labus rezultātus. Bet tas prasa transformācijas rezultātu pielāgošanu.

Izmantojot pakāpes transformācijas un *Box-Cox* transformācijas, rodas problēma izvēlēties transformācijas parametru optimālās vērtības. Lai sasniegtu šādu optimizāciju, var ieteikt iestatīt dažādas attiecīgā parametra vērtības, veikt katrai izvēlētajai vērtībai transformāciju un salīdzināt rezultātus. Pēdējais variants ir tas, kas nodrošina vislabāko sākotnējā sadalījuma tuvinājumu normālajam sadalījumam. Kā variants ir salīdzināt asimetrijas un ekscesa aplēses visām transformācijas iespējām. Tā kā nepieciešamās transformācijas un rezultātu analīzi var veikt, izmantojot piemērotus programmatūras rīkus, daudzkārtīgu transformāciju veikšana un iegūto rezultātu analīze nav problēma.

Kā minēts iepriekš, asimetrisko sākotnējo sadalījumu transformāciju mērķis ir pārveidot šos sadalījumus gandrīz normālā formā.

Transformētajam atribūtu vērtību a_j sadalījumam var iegūt interesējošos šī sadalījuma statistiskos raksturlielumus: sagaidāmo vērtību, standartnovirzi, ticamības intervālu paredzamajai vērtībai un tamlīdzīgu. Tās ir absolūti korektas darbības. Iepriekš šajā sadaļā tika sniegti dažādu transformētu sadalījumu statistikas datu kopsavilkumi. Dati tika iegūti, izmantojot SPSS programmatūras pakotni. Iegūtos raksturlielumus var izmantot transformāciju rezultātu analīzei.

Problēma ar sākotnējo sadalījumu transformēšanu ir saistīta ar citu šādu transformāciju aspektu. Lieta tāda, ka jebkura transformēta sadalījuma analīze nav tieši saistīta ar sākotnējo sadalījumu. Šāda analīze, kaut arī absolūti pareiza, neko nevar pateikt par sākotnējo sadalījumu, lai gan tas ir viens no analīzes galvenajiem mērķiem.

Lai atrisinātu šo problēmu, ir nepieciešams pārvērst transformētā sadalījuma statistisko analīzi atpakaļ sākotnējā sadalījumā. Kā to var panākt?

Transformētā sadalījuma sagaidāmo vērtību var pārvērst, izmantojot apgriezto transformāciju, kas tika piemērota sākotnējam sadalījumam. Tādējādi, ja tika izmantota transformācija ar decimāllogaritmiem, tad transformēto atribūtu vērtību kopas sagaidāmā vērtība $\bar{a}_j^t$ tiek pārvērsta sākotnējā datu telpā kā $\bar{a}_j = 10^{\bar{a}_j^t}$. Ja tika izmantota transformācija ar naturālajiem logaritmiem, tad $\bar{a}_j = e^{\bar{a}_j^t}$, ja tika izmantota transformācija ar binārajiem logaritmiem, tad $\bar{a}_j = 2^{\bar{a}_j^t}$.

Ja tika izmantota transformācija, kuras pamatā ir pakāpes funkcija, tad transformētā sadalījuma parametru vērtību apgrieztā transformēšana tiek veikta, pamatojoties, piemēram, uz apgriezto pakāpes funkciju:

- kvadrātiskajai transformācijai – pamatojoties uz kvadrātisko funkciju;
- kvadrātsaknes transformācijai – uz kvadrātsaknes funkciju;
- patvaļīgai x pakāpei – uz pakāpes funkcijas ar pakāpi $\frac{1}{x}$.

Jāņem vērā, ka transformēto sadalījumu standartnovirzes un standartkļūdas netiek pārvērstas atpakaļ sākotnējo datu telpā, jo pēc transformēšanas šie raksturlielumi zaudē savu jēgu.

6.4. Atribūtu vērtību transformācija regresijas linearitātes sasniegšanai

Bez atribūtu vērtību transformācijas, lai normalizētu šo vērtību sākotnējos sadalījumus, datu pirmapstrādes jomā ir plaši pielietotas transformācijas lineāru atkarību starp atribūtu vērtību kopām sasniegšanai.

Standarta lineārās regresijas teorētiskais pamats ir dots pielikumā P1.3. Pieņem, ka ir dotas divas atribūtu vērtību kopas a_j un a_k. Ja starp šīm vērtību kopām pastāv lineāra sakarība, tad šīs

sakarības var aprakstīt ar lineārās regresijas vienādojumu (lineārā regresija jau tika apskatīta 3. nodaļas 4. sadaļā):

$$a_k = b_0 + b_1\left(a_j\right) + \varepsilon, \tag{6.9}$$

kur b_0 – atribūta a_k vērtība regresijas līknes krustpunktā ar vertikālo asi, kas attēlo atribūtu vērtības a_k;

b_1 – parametrs, kas nosaka regresijas līknes slīpumu pret horizontālo asi, kas attēlo atribūtu a_j vērtības;

ε – kļūdas lielums, kas raksturo datu punktu novirzi no regresijas taisnes.

Ja atribūtu a_j un a_k sākotnējās vērtības ir tādas, ka attiecības starp tām var raksturot ar atbilstošu precizitātes pakāpi, izmantojot regresijas vienādojumu (6.9), tad šo vērtību transformācijas nav nepieciešamas. Problēmas rodas, ja attiecības starp atribūtu vērtību kopām nav lineāras un tās nevar aprakstīt ar regresijas vienādojumu.

Šādā situācijā ir iespējamas šādas darbības:

- modelēt esošo atkarību ar piemērotu nelineārās regresijas vienādojumu un izmantot šo modeli turpmākajā datu apstrādes un analīzes procesā;
- transformēt atribūta a_k vai a_j vērtības (jeb abas), lai iegūtu vēlamo lineāro attiecību.

Šajā sadaļā tiks aplūkota otrā iespēja, proti, viena atribūta vai abu atribūtu vērtību transformācija.

Visbiežāk, lai panāktu attiecīgu lineāru sakarību, izmanto logaritmiskās transformācijas. Pēc noklusējuma transformācijām tiek izmantoti naturālie logaritmi. Ir trīs šādu transformāciju veidi:

- *lineāri logaritmiskā transformācija:*

$$a_k = b_{01} + b_{11}\ln\left(a_j\right) + \varepsilon, \tag{6.10}$$

kur b_{01} un b_{11} – attiecīgi krustošanās ar līkni un slīpuma parametri;

- *logaritmiski lineārā transformācija:*

$$\ln\left(a_k\right) = b_{o2} + b_{12}a_j + \varepsilon; \tag{6.11}$$

- *logaritmiski logaritmiskā transformācija:*

$$\ln\left(a_k\right) = b_{03} + b_{13}\ln\left(a_j\right) + \varepsilon. \tag{6.12}$$

Nav speciālu ierobežojumu par to, kuras no iepriekš minētajām transformācijām piemērot konkrētajā datu priekšapstrādes uzdevumā. Izvēlei tiek pielietota izmēģinājumu un kļūdu metode. Sākotnējiem datiem tiek piemēroti visi trīs transformāciju veidi. Par galīgo tiek uzskatīts tas variants, kas dod vislabāko transformācijas rezultātu (mazākais kļūdas ε lielums).

Tālāk ir doti ilustratīvi piemēri atribūtu vērtību transformācijām, lai panāktu regresijas linearitāti.

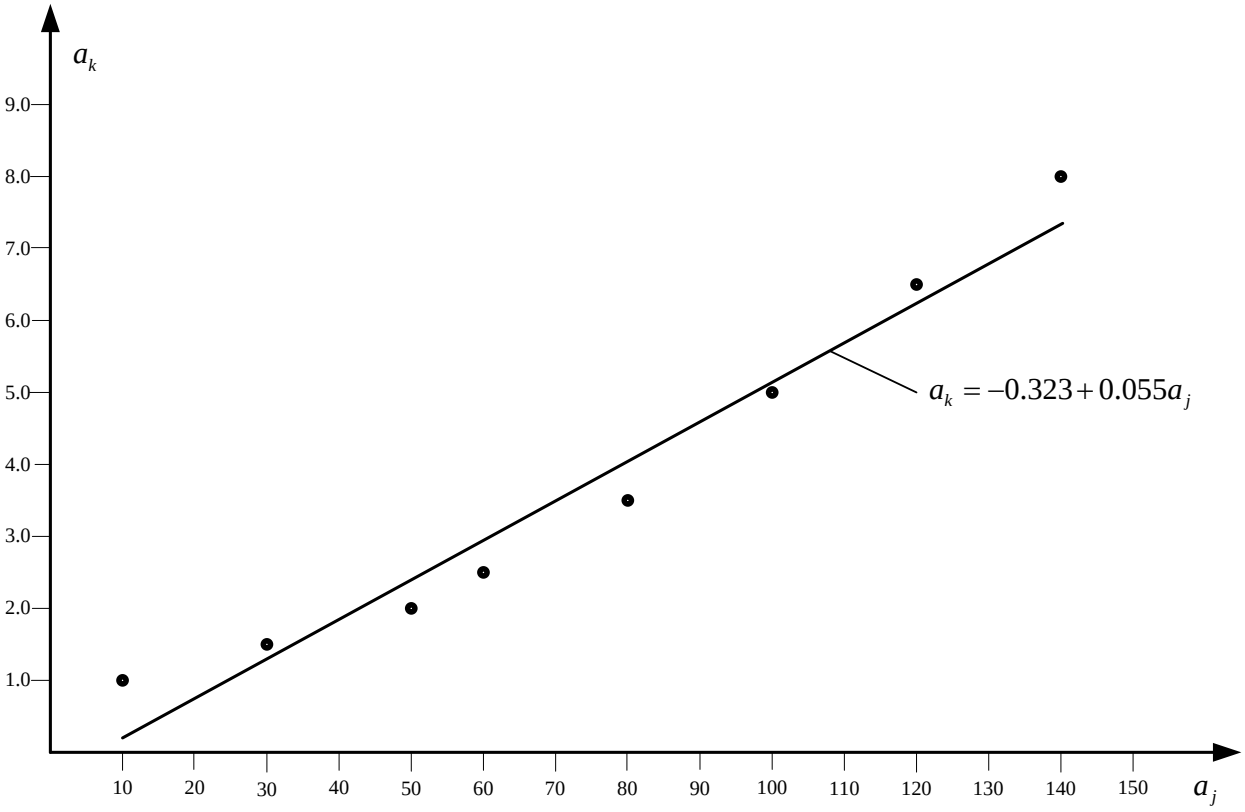
6.1. piemērs. 6.18. tabulas pirmajās divās rindās ir dotas atribūtu a_j, a_k sākotnējās vērtības, trešajā rindā ir parādītas transformētās vērtības $\ln\left(a_j\right)$, ceturtajā rindā – transformētās vērtības $\ln\left(a_k\right)$.

6.18. tabula

Sākotnējās un transformētās atribūtu a_j, a_k vērtības

a_j	10	30	50	60	80	100	120	140
a_k	1.0	1.5	2.0	2.5	3.5	5.0	6.5	8.0
$\ln\left(a_j\right)$	2.302	3.401	3.952	4.094	4.382	4.605	4.787	4.942
$\ln\left(a_k\right)$	0	0.405	0.693	0.916	1.253	1.609	1.872	2.079

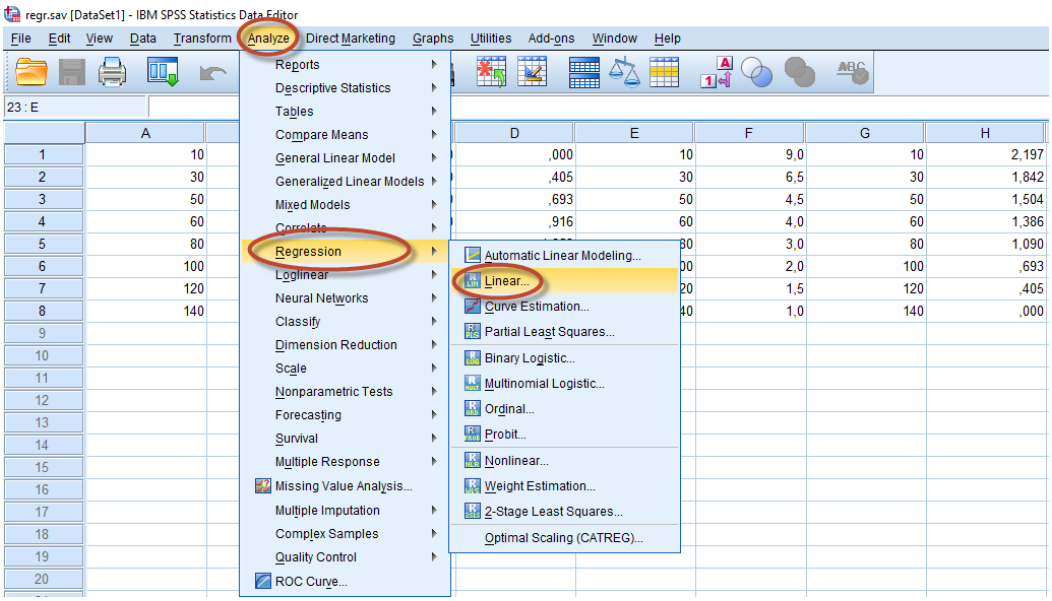
Sākotnējās atribūtu vērtības ir grafiski parādītas 6.17. attēlā.



6.17. attēls. **Sākotnējo atribūtu a_j, a_k vērtību grafiskais attēlojums datiem no 6.18. tabulas**

Tā kā atribūta a_k vērtības palielinās, palielinoties atribūta a_j vērtībām, tad šeit ir pozitīva regresijas sakarība.

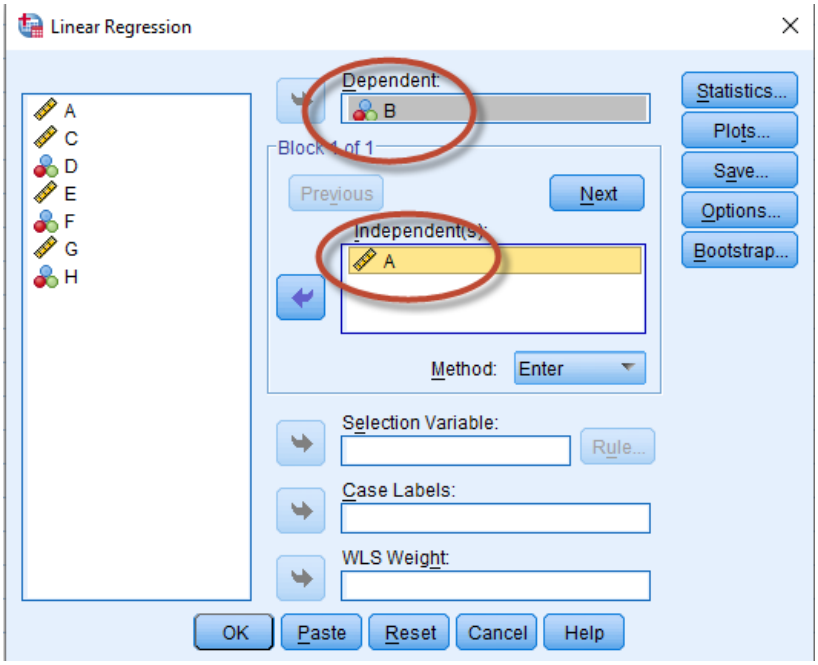
6.17. attēlā datu vizuālā analīze parāda, ka attiecības starp atribūtu a_j un a_k vērtībām nav lineāras. Lai aproksimētu šo sakarību, izmantojot lineārās regresijas vienādojumu, jāveic aprēķini, izmantojot SPSS programmatūras pakotni. Aprēķini sākas, izvēloties procedūru soļus **Analyze – Regression – Linear**.



Pēc tam jāievada dati. Dotajā gadījumā tiek ievadīti šādi dati: $A = a_j$, $B = a_k$.

Lai novērtētu regresiju starp vērtību kopām A un B, šie dati jāievada attiecīgajos laukos.

AB



Pēc aprēķinu veikšanas ir šādi rezultāti:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,979[a]	,959	,952	,5532

a. Predictors: (Constant), A
b. Dependent Variable: B

AB

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,323	,397		-,815	,446	-1,293	,647
	A	,055	,005	,979	11,806	,000	,044	,067

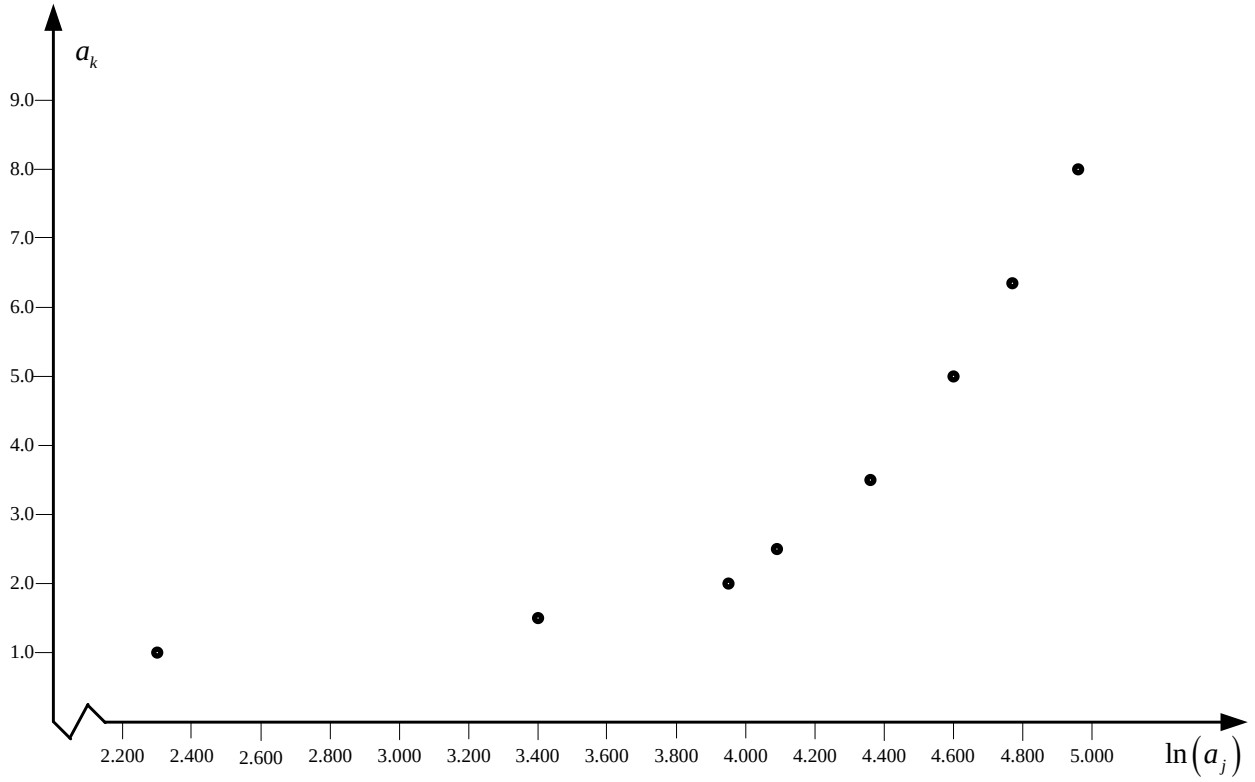
a. Dependent Variable: B

Izmantojot pēdējās tabulas B ailē aprēķinātos datus, iegūst regresijas vienādojumu:
$a_k = -0.323 + 0.055 a_j$.

Izmantojot iegūto vienādojumu, izveido regresijas līkni, kas parādīta 6.17. attēlā.

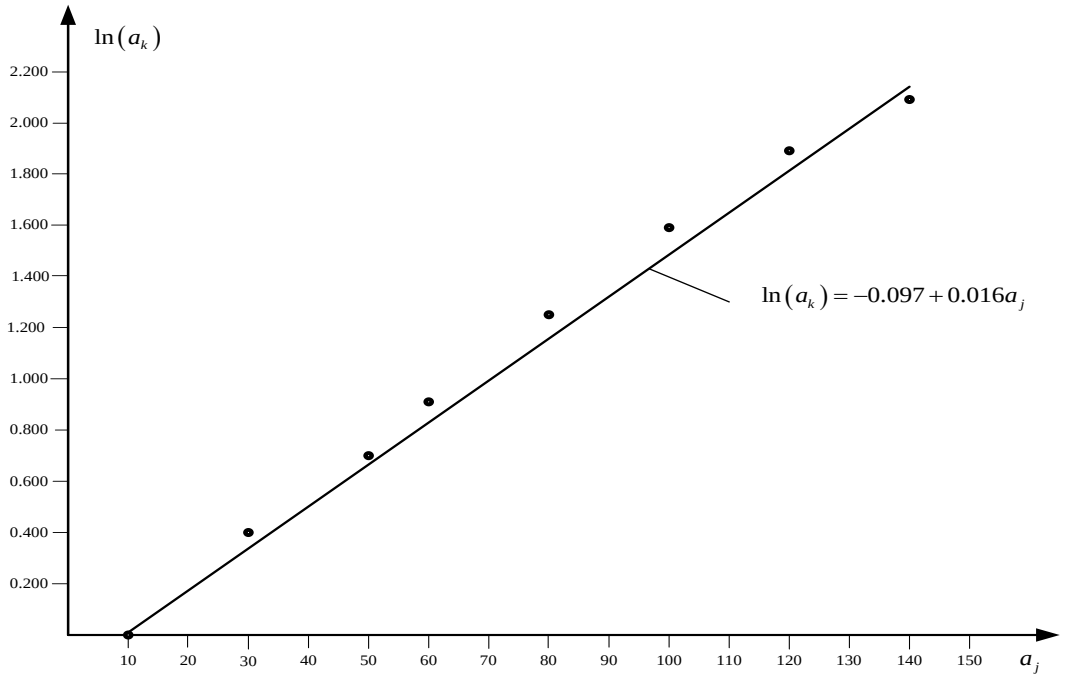
Atribūtu a_j vērtības transformē logaritmiskā formā. Vērtību $\ln(a_j)$ un a_k grafiskais attēlojums parādīts 6.18. attēlā.

Acīmredzams, ka atribūtu a_j vērtību pārveidošana logaritmiskā formā neizveidoja lineāru sakarību starp vērtībām $\ln(a_j)$ un a_k. Šis secinājums tieši izriet no 6.18. attēla vizuālās analīzes un neprasa īpašus komentārus.



6.18. attēls. **Vērtību $\ln(a_j)$, a_k grafisks attēlojums datiem no 6.18. tabulas**

Atribūtu a_k vērtības pārveido logaritmiskā formā. Vērtības a_j un $\ln(a_k)$ grafiski parādītas 6.19. attēlā.



6.19. attēls. **Vērtību a_j, $\ln(a_k)$ grafisks attēlojums datiem no 6.18. tabulas**

Lai noteiktu attiecību izteiksmi starp vērtībām a_j, $\ln(a_k)$, veic nepieciešamos aprēķinus ar SPSS, ieviešot šādus apzīmējumus: $C = a_j$, $D = \ln(a_j)$. Aprēķinu rezultāti ir parādīti divās nākamajās tabulās:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,997[a]	,993	,992	,065367

a. Predictors: (Constant), C

b. Dependent Variable: D

CD

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-,097	,047		-2,064	,085	-,211	,018
	C	,016	,001	,997	29,442	,000	,015	,018

a. Dependent Variable: D

Tika iegūts šāds regresijas vienādojums:

$\ln(a_k) = -0.097 + 0.016a_j$.

Izmantojot šo vienādojumu, izveido regresijas līkni, kas parādīta 6.19. attēlā.

Analizējot iegūtos rezultātus, regresijas "kvalitātes" rādītāji ir koeficienta R^2 vērtība un standartkļūdas vērtība, kas atbilst kļūdas lielumam ε lineārās regresijas vienādojumos (6.4) – (6.8). Šo abu regresiju rādītāju kopsavilkums ir parādīts 6.19. tabulā.

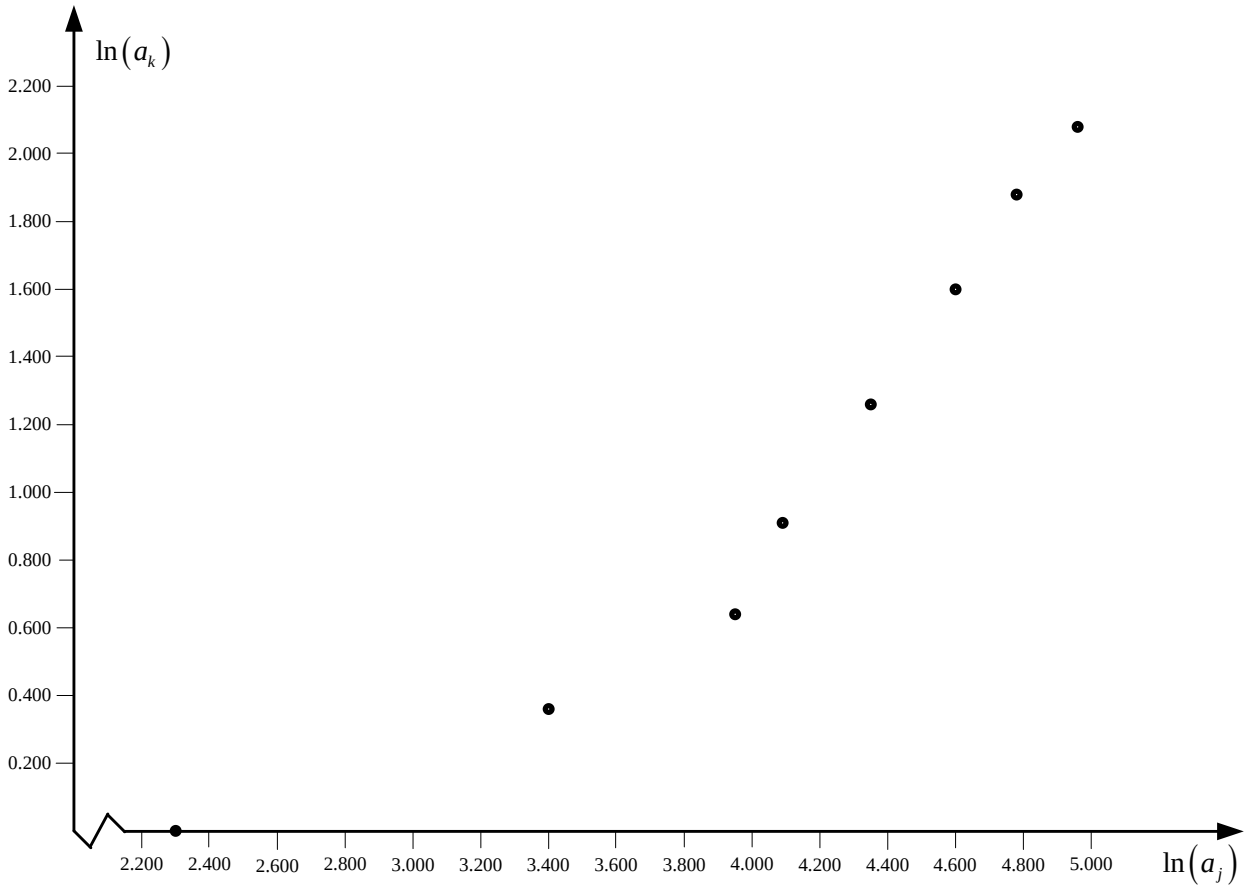
6.19. tabula

Regresijas attiecību $a_j - a_k$, $a_j - \ln(a_k)$ rezultējošo rādītāju kopsavilkums no 6.1. piemēra

Regresija	R^2	ε
$a_j - a_k$	0.959	0.5532
$a_j - \ln(a_k)$	0.993	0.0654

Regresijas attiecības $a_j - \ln(a_k)$ gadījumā vērtība $R^2 = 0.993$ ir lielāka par vērtību $R^2 = 0.959$ attiecībai $a_j - a_k$, un vērtība ir $\varepsilon = 0.0654$ ievērojami mazāka par vērtību $\varepsilon = 0.5532$. Tas nozīmē, ka atribūtu vērtību a_k logaritmiskā transformācija ļāva sasniegt ievērojami precīzāku regresijas $a_j - \ln(a_k)$ attiecību nekā $a_j - a_k$.

6.20. attēlā vērtības $\ln(a_j), \ln(a_k)$ tiek parādītas grafiskā veidā.



6.20. attēls. **Vērtību** $\ln(a_j)$, $\ln(a_k)$ **grafisks attēlojums datiem no 6.18. tabulas**

6.20. attēlā redzamo datu vizuālā analīze parāda, ka sakarība starp vērtībām $\ln(a_j)$, $\ln(a_k)$ nav lineāra. Tādējādi sākotnējām atribūtu a_j, a_k vērtībām no 6.18. tabulas tikai atribūtu vērtību logaritmiskā transformācija a_k nodrošina lineāru sakarību starp vērtībām a_j un $\ln(a_k)$.

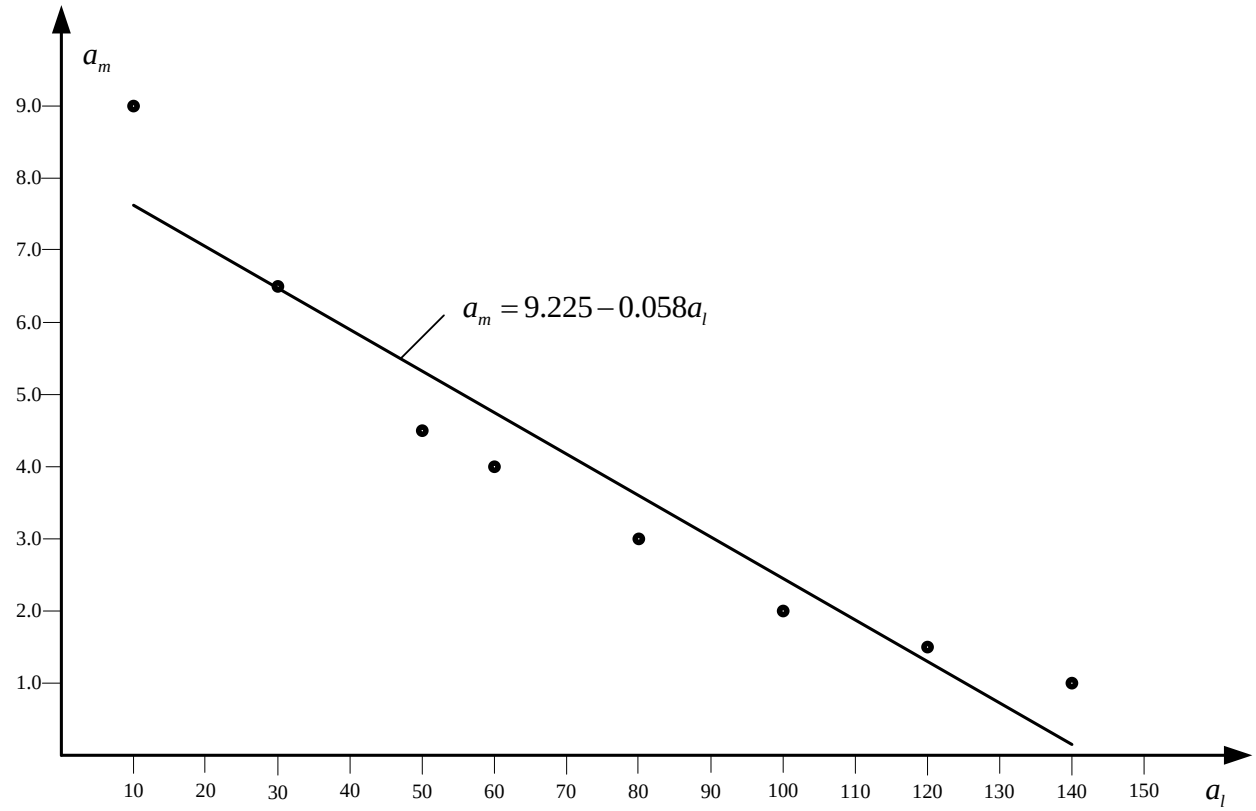
6.2. piemērs. 6.20. tabulā parādītas atribūtu a_l, a_m vērtības un transformētās vērtības $\ln(a_l)$, $\ln(a_m)$.

6.20. tabula

Sākotnējās un transformētās atribūtu a_l, a_m **vērtības**

a_l	10	30	50	60	80	100	120	140
a_m	9.0	6.5	4.5	4.0	3.0	2.0	1.5	1.0
$\ln(a_l)$	2.302	3.401	3.952	4.094	4.382	4.605	4.787	4.942
$\ln(a_m)$	2.197	1.872	1.504	1.386	1.090	0.693	0.405	0

Tā kā atribūtu a_m vērtības samazinās, palielinoties atribūtu a_l vērtībām, tad šeit ir negatīva regresijas sakarība. 6.21. attēls grafiski attēlo atribūtu a_l, a_m vērtības datiem no 6.20. tabulas.



6.21. attēls. **Atribūtu a_l, a_m vērtību grafisks attēlojums datiem no 6.20. tabulas**

6.21. attēlā redzamo datu vizuālā analīze parāda, ka sakarības starp atribūtu a_l un a_m vērtībām nav lineāras. Šo sakarību aproksimēt, izmantojot lineārās regresijas vienādojumu. Aprēķini tika veikti, izmantojot SPSS un apzīmējot: $E = a_l$, $F = a_m$. Pēc attiecīgo aprēķinu veikšanas ir šādi rezultāti:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,956[a]	,914	,900	,8612

a. Predictors: (Constant), E
b. Dependent Variable: F

EF

Coefficients[a]

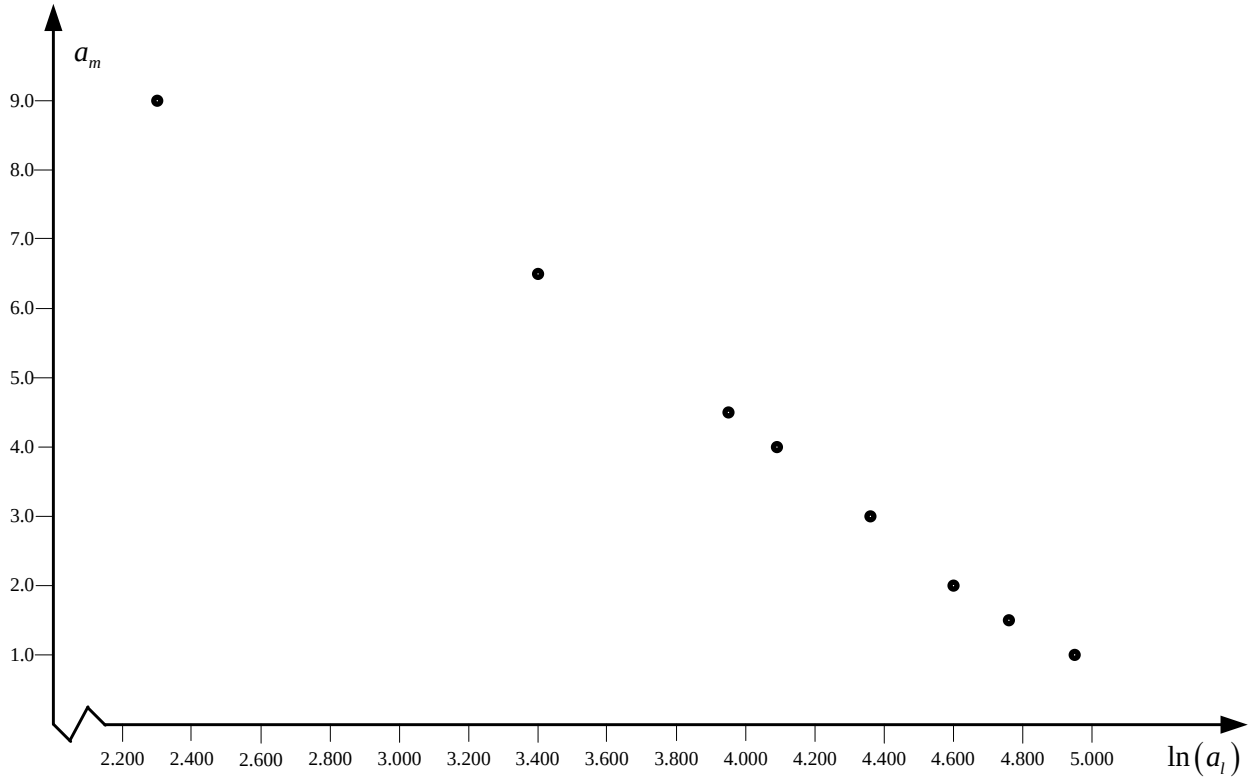
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	8,225	,617		13,323	,000	6,714	9,735
	E	-,058	,007	-,956	-7,983	,000	-,076	-,040

a. Dependent Variable: F

Tika iegūts šāds regresijas vienādojums:
$a_m = 8.225 - 0.058a_l$.

Izmantojot šo vienādojumu, izveido regresijas līkni, kas parādīta 6.21. attēlā.

Atribūtu a_l vērtības pārveido logaritmiskā formā. Vērtību $\ln(a_l)$ grafiskais attēlojums ir a_m parādīts 6.22. attēlā.



6.22.attēls. **Vērtību** $\ln(a_l)$, a_m **grafisks attēlojums datiem no 6.19. tabulas**

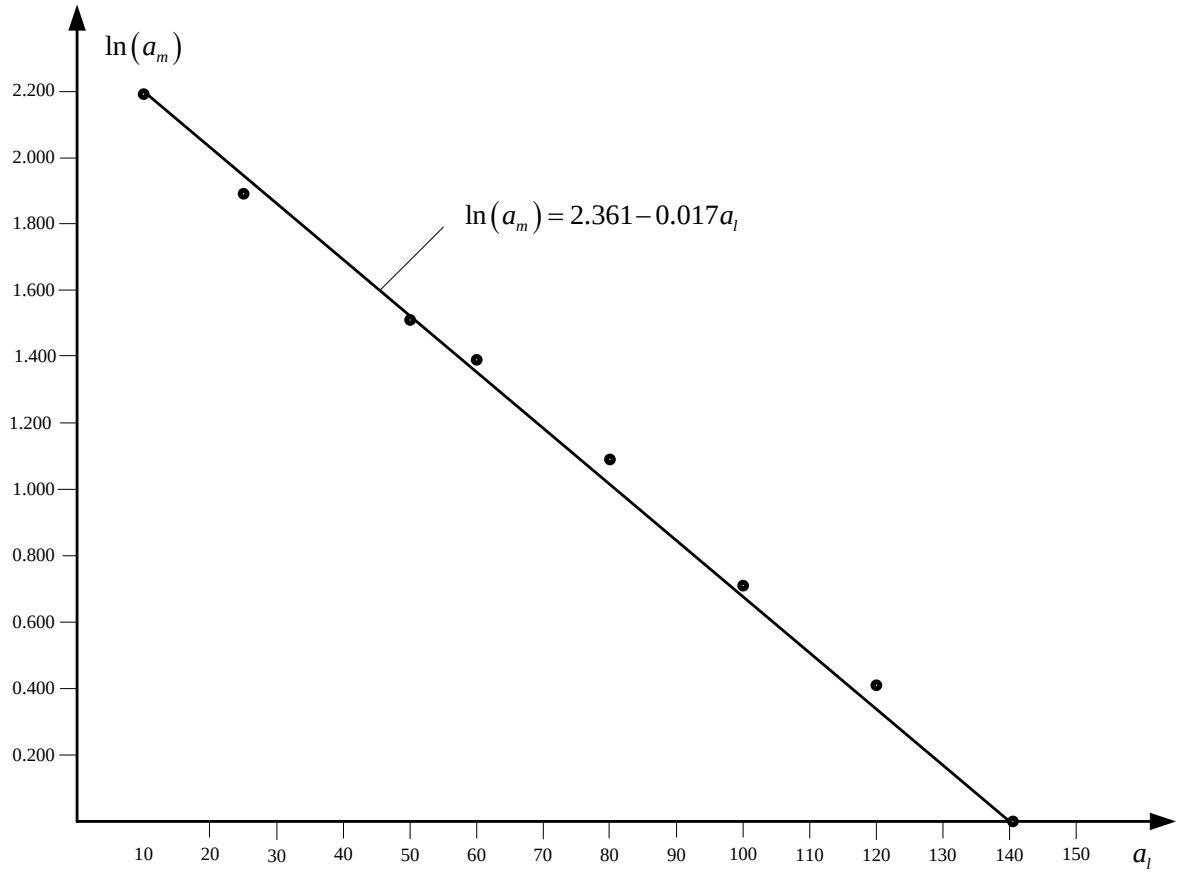
Acīmredzams, ka sakarība starp vērtībām $\ln(a_l)$ un a_m ir "lineārāka" nekā sakarība starp atribūtu sākotnējām vērtībām a_l un a_m. Tomēr šī sakarība nav stingri lineāri atkarīga.

Atribūtu a_m vērtības logaritmiski transformē. a_l un $\ln(a_m)$ vērtības grafiski parādītas 6.23. attēlā. Tajā redzamo datu vizuālā analīze ļauj secināt, ka sakarība starp vērtībām a_l un $\ln(a_m)$ ir "aptuveni" lineāra. Ieviešot apzīmējumu: $G = a_l$, $H = \ln(a_m)$, ar SPSS palīdzību nosaka lineārās regresijas vienādojumu starp šīm vērtībām un analizē sakarību statistiskos raksturlielumus.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,999[a]	,998	,998	,035376

a. Predictors: (Constant), G
b. Dependent Variable: H



6.23. attēls. **Vērtību** a_l, $\ln(a_m)$ **grafisks attēlojums datiem no 6.20. tabulas**

G.H.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,361	,025		93,113	,000	2,299	2,423
	G	-,017	,000	-,999	-55,377	,000	-,017	-,016

a. Dependent Variable: H

Iegūts regresijas vienādojums:

$\ln(a_m) = 2.361 - 0.017a_l$.

Izmantojot šo vienādojumu, izveido regresijas līkni, kas parādīta 6.23. attēlā.
Abu regresiju statistikas kopsavilkums ir parādīts 6.21. tabulā.

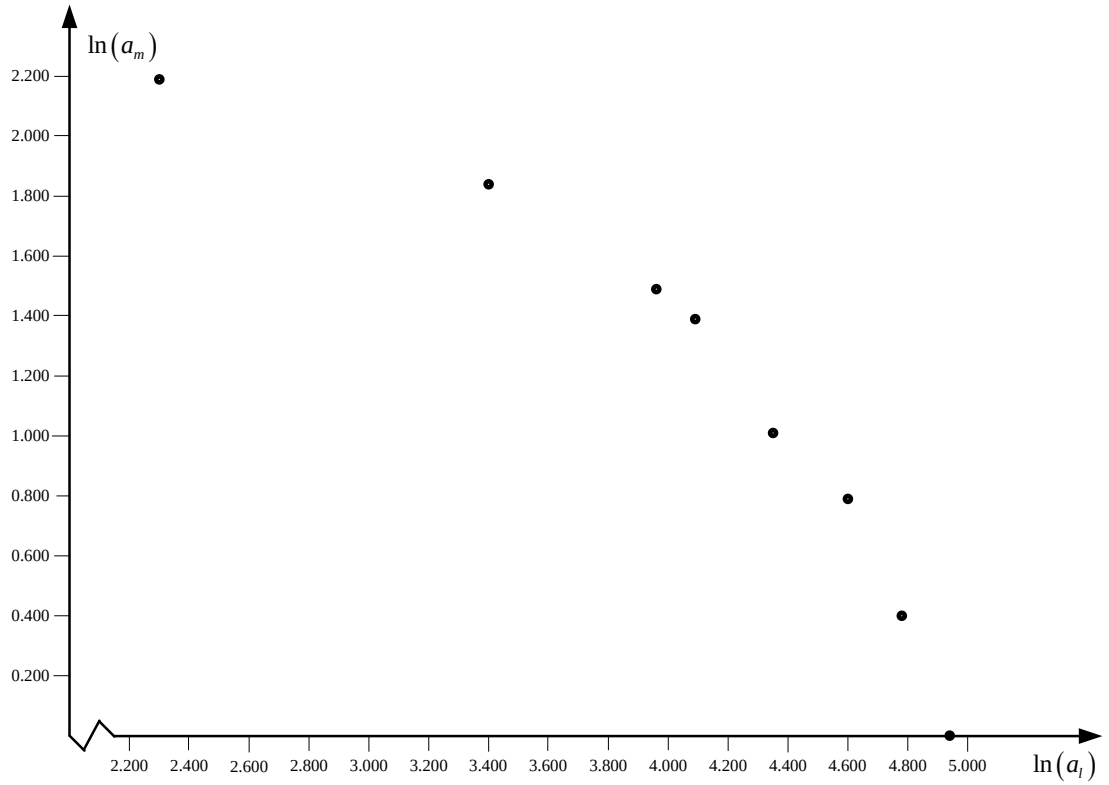
6.21. tabula

Regresijas sakarību $a_l - a_m$, $a_l - \ln(a_m)$ **rezultātu rādītāju kopsavilkums 6.2. piemēram**

Regresija	R^2	ε
$a_l - a_m$	0.914	0.8612
$a_l - \ln(a_m)$	0.998	0.0354

Ir pilnīgi skaidrs, ka atribūtu a_m vērtību logaritmiskā transformācija izraisīja stingru lineāru sakarību starp vērtībām a_l un $\ln(a_m)$.

6.24. attēlā grafiskā veidā tiek parādītas $\ln(a_l), \ln(a_m)$ vērtības.



6.24. attēls. **Vērtību $\ln(a_l)$, $\ln(a_m)$ grafisks attēlojums datiem no 6.19. tabulas**

6.24. attēlā redzamo datu vizuālā analīze parāda, ka sakarība starp vērtībām $\ln(a_l)$un $\ln(a_m)$ nav lineāra.

Abos iepriekšminētajos piemēros tika panākta regresijas linearitāte starp netransformētajām atribūtu a_j un logaritmiski transformētām vērtībām $\ln(a_k)$, $(\ln(a_m))$.

Kā minēts iepriekš, nav vispārīgu ieteikumu par to, kuras atribūtu vērtību transformācijas novedīs pie lineāras regresijas sakarības. Literatūrā bieži tiek izmantots abu atribūtu vērtību logaritmisko transformāciju "klasiskais" piemērs, lai panāktu lineāras regresijas sakarību starp to transformētajām vērtībām. Piemēram, viens no atribūtiem var būt ātri augošas priedes stumbra diametrs, ko mēra centimetros. Otrs atribūts ir priedes stumbra tilpums, ko mēra kubikcentimetros. Tikai ar abu atribūtu vērtību logaritmisku transformēšanu ir iespējams panākt lineāras regresijas sakarības starp šīm transformētajām vērtībām.

Ir jāapskata lineārās regresijas koeficienta b_1 interpretācija, izmantojot dažādas attiecīgo atribūtu vērtību transformācijas.

Ja izmanto sākotnējo atribūtu a_j, a_k vērtību standarta lineāro regresiju (vienādojums 6.1), tad koeficienta b_1 vērtība parāda atribūta a_k vērtības izmaiņas, kad atribūta a_j vērtība mainās par 1 vienību. Acīmredzot ar pozitīvu korelāciju starp atribūtu a_j, a_k vērtībām, b_1 ir pozitīvs un, palielinot b_1 vērtību par 1 vienību, atribūta a_k vērtība palielinās par b_1 vienībām. Ja starp atribūtu vērtībām a_j, a_k ir negatīva korelācija, tad, atribūta a_j vērtībai palielinoties par 1 vienību, atribūta a_k vērtība samazinās par b_1 vienībām.

Koeficienta b_1 vērtības interpretācija standarta lineārā regresijā starp sākotnējām atribūtu vērtībām a_j, a_k ir saprotama un neprasa nekādus papildu paskaidrojumus.

Problēmas ar koeficienta b_1 interpretāciju rodas, izmantojot transformētās atribūtu a_j, a_k vērtības. Tāpēc ir vairākas šī koeficienta interpretācijas dažāda veida transformācijām.

1. Lineāri logaritmiskais modelis (vienādojums 6.10).

Vērtības $\ln(a_j)$ palielināšana par 1 vienību izraisa atribūta a_k vērtības izmaiņas (palielinās vai samazinās atkarībā no koeficienta zīmes) par b_1 vienībām.

Kādā veidā var interpretēt vērtības $\ln(a_j)$ izmaiņas par 1 vienību saistībā ar atribūta a_j sākotnējām vērtībām? Lai to izdarītu, izmanto šādu sakarību:

$$\ln(a_j)+1=\ln(a_j)+\ln(1)=\ln(e*(a_j)).$$

No tās izriet, ka vērtības $\ln(a_j)$ palielināšana par 1 vienību ir līdzvērtīga naturālā logaritma izteiksmei $\ln(e*(a_j))$.

Alternatīvs veids, kā interpretēt operāciju vērtības a_j reizināšanai ar e, ir izteikt a_j vērtības procentuālo pieaugumu, izmantojot šādu sakarību:

$$100*(2.72-1)=172\%.$$

No tās izriet, ka koeficients b_1 atspoguļo atribūta a_j vērtības pieaugumu, kad vērtība $\ln(a_j)$ palielinās par 172 %.

2. Logaritmiski lineārais modelis (vienādojums 6.11).

Atribūta a_j vērtības palielināšana par 1 vienību izraisa $\ln(a_k)$ izmaiņas (palielinājumu vai samazinājumu) par b_1 vienībām. Sākotnējo atribūtu a_k vērtību izteiksmē tas nozīmē, ka atbilstošā atribūta a_k vērtība tiek reizināta ar e^{b_1}. Citiem vārdiem sakot, atribūta a_j vērtības palielināšana par 1 vienību izraisa a_k vērtības izmaiņas par e^{b_1} vienībām.

Cits atribūta a_j vērtības palielinājums par kādu citu skaitli, piemēram, c vienībām, izraisa sākotnējās atribūta a_k vērtības izmaiņas par e^{cb_1} vienībām.

3. Logaritmiski logaritmiskā transformācija (vienādojums 6.12).

Šajā gadījumā atribūta a_j vērtības reizināšana ar e atbilst atribūta a_k vērtības reizināšanai ar e^{b_1}.

Izmantojot šo modeli, tiek izmantotas proporcionālas atribūta a_k vērtības izmaiņas (palielinājums vai samazinājums), kad atribūta a_j vērtība palielinās par noteiktu skaitu, piemēram, c procentiem. Lai to izdarītu, vispirms jāaprēķina:

$$d=\ln(100+c)/100.$$

Šai d vērtībai atribūta a_k vērtība mainās par e^{db_1} procentiem.

Papildu logaritmiskajām transformācijām, lai panāktu lineāras regresijas sakarību, tiek izmantotas kvadrātsaknes transformācijas un *Box-Cox* transformācijas. Tomēr pēdējā gadījumā transformācijas parametra λ optimālās vērtības noteikšana ir sarežģīta matemātiska problēma.

IZMANTOTĀ LITERATŪRA

Acuna E., Rodrigues C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. In book "Classification, Clustering and Data Mining Applications, pp. 639 – 647.

Alasadi S. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences, 12 (16), pp. 4102 – 4107.

Acuna E. (2011). Preprocessing in Data Mining. In Lovric M. (Ed.). International Encyclopedia of Ststistical Science, Springer, doi: 10.1007/978-3-642-04898-2_51.

Aljuaid T., Sasi S. (2016). Intelligent Imputation Technique for Missing Values. Int. Conference on Advances in Computing, Communications and Informatics, September 21 – 24, 2016, Jaipur, India, pp. 2441 – 2445.

Agresti A., Franklin Ch. (2013). Statistics. The Art and Science of Learning from data. (third edition). Pearson Education, Inc., 834 p.

Anath C.V., Kleinbaum D.G. (1997). Regression Model for Ordinal Responses: A Review of Methods and Applications. Information Journal of Epidemiology, Vol. 26, (6), pp. 1323 – 1333.

Agresti A. (2010). Modelling Ordinal Categorical data. www.stat.ufl.edu/~aa/ordinal/ord.html.

Agresti A. (2002). Categorical Data Analysis (second edition). John Wiley & Sons, Inc., Hoboken, New Jersey, 394 p. DOI:10.1002/0471249688.

Agresti A. (2007). An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, 394 p., DOI:10.1002/0470114754.

Alison P.D. (2005). Imputation of categorical variables with proc mi. Proceedings of the SUGI 30, 113 (30), pp. 1-14.

Allison P.D. (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research, 28, pp. 301-309.

Aggarwal Ch.C. (2016). Outlier Analysis (Second Edition). http://rd.springer.com.book/10.1007/978-3319-47578-3.

Aşikgil B., Erar A. (2009). Research into Multiple Outliers in Linear Regression Analysis. Hacetpere Journal of Mathematics and Statistics, 28 (2), pp. 185 – 198.

Alma Ö.G. (2011). Comparison of Robust Regression Methods in Linear Regression. Int. J. Contemp. Math. Science, 6 (9), pp. 409 – 421.

Anitha S., Metilda M., (2016). A Survey on Cluster Based Outlier Detection Techniques In Data Stream. Int. Journal of Data Mining Techniques and Applications, 5, (1), pp. 96 – 101.

Batista G.E., Monard M.-C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence, Vol. 17, Issue 5 – 6, pp. 519 – 533.

Batista G., Monard M.C. (2003). Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data. Technical Report, University of Sao Paulo.

Batista G., Monard M.C. (2002). A study of K-nearest neighbor as an imputation method. In Abraham A. et al. (Eds.) Hybrid Intell. Systems, IOS Press, pp. 251 – 260.

Barnard J., Rubin D.B. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86 (4), pp. 948 – 955, doi: 10.1093/biomet/86.4.948.

Burgin M. (2004). Data, Information, and Knowledge. Information, 7 (1), pp. 47 – 57.

Behrens J.T. (1997). Principles and Procedures of Exploratory Data Analysis. Psychological Methods, Vol. 2, No. 2, pp. 131 – 160.

Batista G.E., Monard M.C. (2001). A Study of k-Nearest Neighbour as a Model-Based Method to Treat Missing Data. In Proceedings of the Argentine Symposium on Artificial Intelligence, Vol. 30, Buenos-Aires, Argentine, pp. 1 – 9.

Bürkner P.-Ch., Vuorre M. (2019). Ordinal Regression Models in Psychology: A Tutorial. Advances in Methods and Practices in Psychological Science, Vol. 2 (1), pp. 77-101. DOI:10.1177/2515245918823199.

Beretta L., Santaniello A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. Medical Informatics and Decision Making, 16, pp. 197-208. DOI:10.1186/s12911-016-0318-z.

Ben-Gal I. (2005). Outlier Detection. In Maimon O and Rockash (Eds.). Data Mining and

Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researches, Kluwer Academic Publishers, 16 p.
Barnett C., Lewis T. (1994). Outliers in Statistical Data. (3rd edition). John Wiley & Sons, Chichester, 584 p.
Bogashaw G.B., Yohannes Y.B. (2020). Review of Outlier Detection and Identifying Using Robust Regression Model. Int Journal of Systems Science and Applied Mathematics, 5 (1), pp. 4 – 11.
Breunig M.M., Kriegel H.-P., Ng R.T., Sender J. (2000). LOF: identifying density-based local outliers. ACM sigmod record, 29 (2), pp. 93 – 104.
Bishop Ch.M. (2006). Pattern Recognition and Machine Learning. Springer, 758 p.
Boulle M. (2004). Khiops: A Statistical Discretization Method of Continuous Attributes. Machine Learning, 55, pp. 53 – 69.
Box G.E.P., Cox D.R. (1964). An Analysis of Transformations. Journal of the Roual Statistical Society, Series B (Methodological), 26 (2), pp. 211 – 252.
Cios K.J., Pedrycz W., Swiniarski R.W., Kurgan L.A. (2007). Data Mining. A Knowledge Discovery Approach. Springer, 600 p.
Cheng H.Y. (1994*). Abduction? Deduction? Induction? Is There a Logic of Exploratory Data Analysis?* Annual Meeting of the American Educational Research Association, New Orleans, LA, April 4 8, 28 p.
Cheng H.Y. (2010). Exploratory Data Analysis in the Context of Data Mining and Resampling. Int. Journal of Psychological Research, 3, (1), pp. 9 – 22.
Chen H.-Ch., Wang N.-Sh. (2014). The Assignment of Scores Procedure for Ordinal Categorical Data. The Scientific World Journal, Vol. 2014, 7 p, http://dx.doi.org/101155/2014/304213.
Chen J., Shao J. (2000). Nearest neighbor imputation for survey data. Journal of official Statistics, 16,
Cousineneau D. (2010). Outliers detection and treatment: a review. Int. Journal of Psychological Research, 5 (1), pp. 58 – 67.
Çetin M., Toka O. (2011). The Comparing of S-estimators and M-estimators in Linear Regression. Journal of Science, Gazi University, 24 (4), pp. 747 – 752.
Christy A.G., Gandhi M., Vaithyasubramanian (2015). Cluster Based Outlier Detection for Healthcare Data. Procedia Computer Science, 50, pp. 209 – 215.
Chmilewski M.R., Grzymala-Busse J.W. (1996). Global Continuous Attributes as Preprocessing for Machine Learning. Int. Journal of Approximate Reasoning, 15, pp. 319 - 331.
Ching J.Y., Wong A.K.C. (1995). Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, (7), pp. 641 – 651.
Dong Y., Peng Ch.-Y. J. (2013). Principled missing data methods for researcher. Springer Plus, 2, 17 p.
Dempster A.P., Laird N.M., Rubin D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 9 (1), pp. 1 – 38.
De Mast J., Kemper B.P.H. (2009). Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case? Quality Engineering, 21, pp. 366 – 375.
Dubani S.A. (1976). Transactions. Systems Man Cybernetics, 6 (4), pp. 325 – 327.
Do Ch.B., Batzoglou S. (2008). What is the expectation maximization algorithm. Primer. Nature Biotechnology, 26 (8). pp. 897 – 899.
Dempster A.P., Laird N.M., Rubin D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm. Journal Royal Statistical Society, 39 (1), pp. 1 – 38.
Dan E.D., Ijeona O.A. (2013). Statistical analysis/methods of detecting outliers in a univariate data in a regression analysis model. Int. Journal of Education and Research, 1 (5), 24 p.
Davies L., Gather U. (1993). The identification of multiple outliers. Journal of the American Statistical Association, 88 (423), pp. 782 – 792.
Dan E.D., Ijeoma O.A. (2013). Statistical Analysis/Methods 0f Detecting Outliers in a Univariate Regression Analysis Model. Int. Journal of Education and Research, 1, 24 p.

Dimitrova E.S., Licona M.P.V., McGee J., Laubenbacher R. (2010). Discretization of Time Series Data. Journal of Computational Biology, 17, (6), pp. 853 – 868. DOI:10.1089/cmb.2008.0023.
Dash R., Dash R. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. Int. Journal of Advances in Science and Technilogey, 2 (3), pp. 29 – 37.
Dougherty J., Kohavi R., Sahami M. (1995). Supervised and unsupervised discretization of continuous features. Proceedings of the 12th Int. Conference on Machine Learning. San Francisco, pp. 194 – 202.
Dęlkowska K., Jarocka M. (2013). The impact of the methods of the data normalization on the result of linear ordering. Acta Universitatis Lodziensis, Polia Oeconomica, 286, pp. 181 – 188.
Forhangfar A., Kurgan L.A., Dy J. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41, pp. 3693 – 3705.
Feng H., Guoshun C., Yang B., Chen Y. (2005). A SVM regression based approach to filling in missing values. In Khosla R., Howlett R.J., Jain L.C. (Eds.) ″KES (3)″, Vol. 3683 of lecture notes in computer science, Derlin, Springer, pp. 581 – 587.
Farhangfar A., Kurgan L.A., Pedrycz W. (20007). A novel framework for imputation of missing values in databases. IEEE Transaction on Systems Man Cybernetics, Part A, 37 (5), pp. 692 – 709.
Faisal Sh., Tutz G. (2016). Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes. Proceedings of the 31st International Workshop on Statistical Modeling, Vol. 1, 28 p.
Feng Ch., Wang H., Lu N., Chen T., He H., Lu Y., Tu X.M. (2014). Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry, 26 (2), pp. 105 – 109.
Fayyad U.M., Irani B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th Int. Joint Conference Artificial Intelligence, pp. 1022 – 1027.
Feng Ch., Wang H., Lu N., Chen T., He H., Lu Y., Tu X.M. (2014). Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry, 26 (2), pp. 105 – 109.
Grzymala-Buse J.W., Grzymala-Buse W.J. (2005). Handling missing attribute values. In Maimon O., Rokash L. (Eds.). Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, pp. 37 – 57.
Grzymala-Buse J.W., Ming Hu (2000). A comparison of Several Approaches to Missing Attribute Values in Data Mining. Proceedings of Int. Conference on Rough Sets and Current Trends in Computing, pp. 378 – 388.
Garśia S., Luengo J., Herrera F. (2015). Data Preprocessing in Data Mining. Springer, 327 p.
Grzymala-Buse J., Goodwin L., Grzymala-Buse W., Zhang X. (2005). Handling missing attribute values in preterm birth data sets. Proceedings of 10th Int. Conference of rough sets and fuzzy sets and data mining and granular computing, pp. 342 – 352.
Geler Z., Kurbalija V., Radovanović M., Ivanović M. (2016). Comparison of different weighting schemes for the kNN classifier on time-series data. Knowledge and Information Systems, 48, pp. 331 – 378. htpps://doi.org/10.1007/s10115-015-0881-0.
Gökmen N., Erar A. (2018). Robust Estimators In Linear Regression: A Simulation Study And An Application On Health Care. 11 International Statistics Days Conference, 3 – 7 October 2018, pp. 874 – 885.
Gan G., Ng K.-P. (2017). k-means clustering with outlier removal. Pattern Recognition Letters, 90, pp. 8 – 14.
Garsia S., Luengo J., Saez J.A., Lopez V. Herrera F. (2011). A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. IEEE Transaction On Knowledge and Data Engineering. pp.
Grzymala-Buse J.W. (2004). Three strategies to rule induction from data with numerical attributes. Lecture Notes in Computer Science 3135, pp. 54 – 62.
Grűnwald P.D. (2007). The minimum description length principle. The MIT Press, Cambridge, London, 703 p.
Howell D.C. (2007).. The analysis of missing data. In Outwaite W. & Turner S. Handbook of Social Science Methodology, London, Sage.

Hestie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition, Springer, 764 p.
Haig B.B. (2005). An Abductive Theory of Scientific Method. Psychological Methods, Vol. 10, No. 4, pp. 371 – 388.
Han J., Kamber M., Pei J. (2012). *Data Mining Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 703 p.
Howell D. (2007). The analysis of missing data. SAGE Publication, London.
Hechenbichler K., Schliep K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Ludwig-Maximilians-Universität Műnchen. Sonderforschungsbereich 386, Paper 399, 16 p.
Hedge V.J. Augustin J. (2004). A Survey of Outlier Detection Methodologies. Sydney, Kluwer Academic Publishers.
Hadi A.S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society, Series B, 54, pp. 761 – 771.
Huber P.J. (1981). Robust Statistics. John Wiley & Sons, New York.
Hawkings D.M. (1980). Identification of OUtliers. Chapman and Hall-Science, 188 p.
https:/towardsdatascience.com/11-dimensionality-reduction-techniques--you-should-know-in-2021-dcb9500d388b.
Hemada B., Lakshmi K.S.V. (2013). A Study On Discretization Techniques. Int Journal of Engineering Research & Technology, 2, (8), pp. 1887 – 1892.
Ho K.M., Scott P.D. (1997). Zeta: A Global Method for Discretization of Continuous Variables. III Int. Conference on Knowledge Discovery and Data Mining, pp. 191 – 194.
Holte R.C. (1993). Very simple classification rules perform well on most commonly used data sets. Machine Learning, 11, pp. 63 – 90.
Iglewitz B., Hoaglin D.C. (1993). How to detect and handle outliers. ADQC Quality Press.
Jović W A., Brkić K., Bogunović N. (2015). A review of feature selection methods with applications. Proceedings of the 38 Int. Convention on Information and Communication Technology, Electronics and Microelectronics, 6 p.
Jain S., Yain K., Chodhary N (2016). A survey paper on missing data in data mining. Int. Journal of Innovations in Engineering Research and Technology, 3 (12), pp. 45 – 50.
Jönsson P., Wohlin C. (2004). An Evaluation of k-nearest Neighbour Imputation Using Likert Data. Proceedings of the 10th International Symposium pn Software Metrics, pp. 108 – 118.
Jolliffe I.T. (2002). *Principal Component Analysis* (Second Edition). Springer-Verlag, New York, Berlin Heidelberg, 487 p.
Jackson J.D. (1991). *A User's Guide to Principal Component Analysis*. John Willey & Sons, Inc.
Jahan A., Edwards K.L. (2015). A state0of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. Materials and Design, 65, pp. 335 – 342.
Krzanowski W. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data Computer & . Mathematics Applications, 12 (2, part A), pp. 179 – 185.
Kang H. (2013). The prevention and handling of the missing data. Korean J. Anesthesiol, 64 (5), pp. 402 406. doi: 10.4097/kjal.2013.64.5.402.
Kuppusamy V., Paramasivam I. (2016). A Study of Impact on Missing Categorical Data - A Qualitative Review. Indian Journal of Science and Technology, ((32), 6 p., doi:10.17485/ijst/2016/v9i32/83088.
Kwak S.K., Kim J.H. (2017). Statistical data preparation management of missing values and outliers. Korean Journal of Anesthesiology, 79 (4), pp. 407 – 411.
Knorr E.M., Ng R.T. (1999). Finding Intensional Knowledge of Distance-Based Outliers. Proceedings of the 25th ULDB Conference, Edinburg, Scotland, 12 p.
Knorr E.M., Ng R.T. (1998). Algorithms for Mining Distance-Based Outlier in Large Datasets. Proceedings of the 24th WLDB Conference, New York, USA, pp. 392 – 403.
Kumar V., Kumar S., Singh K. (2013). Outlier Detection: A Clustering-Based Approach. Int Journal of Science and Modern Engineering.
Kaufman L., Rousseeuw P.J. (1987). Clustering by means of Medoids. In (Y.Dodge, ed.). Statistical Data Analysis Based on the Norm and Related Methods, North-Holland, pp. 405 – 416.

Kotsiantis S., Kanellopou D. (2006). Discretization Techniques: A recent survey. GESTS Int. Transactions on Computer Science and Engineering, 32, (1), pp. 47 58.
Kurgan L.A., Krzysztof J.C. (2004). CAIM Discretization Algorithm. IEEE Transactions On Knowledge and Data Engineering, 16, (2), .
Kerber R. (1992). Chimerge: Discretization of numeric attributes. Proceedings of the Ninth National Conference Artifical Intelligence, pp. 123 – 128.
Laencina P.J.G., Sancho-Gomez J.L., Figneras-Vidal A.R. (2010). Pattern classification with missing data: A review. Neural Computing and Applications, 19, (2), pp. 263 – 282.
Little R.J.A., Rubin D.B. (2002). Statistical analysis with missing data. 2nd Edition, Willey, New Jersey.
Liu Ch.-H., Tsai Ch.-F., Sue K.-L., Huang M.-W. (2020). The Feature Selection Effect On Missing Value Imputation on Medical Datasets. Applied Sciences, 10, 12 p.
Li D., Deogun J., Spaulding W., Shuart B. (2004). Towards missing data imputation: a study of fuzzy k-means clustering method. Proceedings of 4th Conference of
Luengo J., Garcia S., Herrera F. (2011). On the choice the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 32, pp. 77 – 108.
Liu B.Q. (2014). The application of exploratory data analysis in auditing. A dissertation submitted to the Graduate School-Newark, The State University of New Jersey for the degree Doctor of Philosophy.
Little R.J.A., Schluchter M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika, 72, pp. 497 – 512.
Liddell T., Krushke J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology, 79, pp. 328 – 348.
Larose D.T. (2005). Discovering Knowledge in data: An Introduction to Data Mining. Wiley.
Landerman L.R., Kenneth C.L., Pieper C.F. (1979). An Empirical Evaluation of the Predictive Mean Matching Method for Imputation Missing Values. Sociological Methods & Research, 26, pp. 3 – 23.
LaLonde S.M. (2012). Transforming Variables for Normality and Linearity – When, How, Why and Why Not's. SAS Global Forum 2012. Statistics and Data Analysis, 8 p.
Lee D.K. (2020). Data transformation: a focus on the interpretation. Korean Journal Anesthesiol, 73 (6), pp. 503 – 508.
LakshmiM.S. (2018). An Overview Study on Data Cleaning, its Types and its Methods for Data Mining. Int. Journal of Pure and Applied Mathematics, 119 (12), pp. 16837 - 16848.
Loureiro A., Togo L. Soares C. (2004). Outlier detection using clustering methods: a data cleaning application. Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector, Bonn, Germany.
Lin T.Y. (2002). Attribute Transformations for Data Mining I: Theoretical Explanations. Int. Journal of Intelligent Systems, 17, pp. 213 – 222.
Liu H., Hussain F., Lim C., Dash M. (2002). Discretization. An Enabling Technique. Data Mining and Knowledge Discovery, 6 (4), pp.393 – 423.
Lu D.K. (2020). Data transformation: a focus on the interpretation. Korean Journal Anesthesiol, 73 (6). pp. 503 – 508.
LaLonde S.M. (2012). Transforming Variables for Normality and Linearity – When, How, Why, and Why Not's. SBS Global Forum 2012, Paper 430-2012. 8 p.
McGullgh P. (1980). Regression model for ordinal data. Journal of the Royal Statistical Society, series B, 42, pp. 109 – 142.
Moon T.K. (1996). The expectation-maximization algorithm. IEEE Signal Proceeding Magazine, December 1996, pp. 47 – 60.
Mielke A. (2016). Robust Statistics. Trufa Science Inside № 16 5p.
Mofanavalli S.S., Poornima N.S. (2018). Outlier Detection using Clustering Techniques. Int. Journal of Engineering & Technology, 7, pp. 813 – 818.
Moore Y.D., MCCabe G. (1999). The Practice of Statistics, New York, Freeman.
Mohanavalli S.S., Sriprija N., Poornima S. (2018). Outlier Detection using Clustering Techniques. Int. Journal of Engineering & Technology, 7, pp. 813 – 818.
Muhlenbach F., Rakotomalala R. (2005). Discretization of Continuous Attributes.

John Wang Encyclopedia of Data Warehousing and Mining, Idea Group Reference, pp. 397 – 402.
Nügroho W.H., Wardhani N.W.S., Fernandes A.R., Solimun (2020). Robust Regression Analysis Study for Data with Outliers at Some Significance Ltvels. Mathematics and Statistics, 8 (4), pp. 373 – 381.
Nayak S.C., Misra B.B., Behera H.S. (2014). Impact of Data Normalization on Stock Index Forecasting. Int. Journal of Computer Information Systems and Industrial Management Applications, 6, pp. 257 – 269.
Oba S., Sato M., Takamasa I., Monden M., Matsubara K. (2003). A bayesian missing value estimation method for gene expression profile data. Biometrics, 19 (16), pp. 2088 – 2096.
Ozdemir O. (2016). A comparison Study of Data Transformation Methods to Achieve Normality. Int. Journal of Mathematical and Computational Methods, 1, pp. 382 – 383.
Osborne J.W. (2002). Notes on the Use of Data Transformation. Practical Assessment, Research & Evaluation, 8 (6), 7 p.
Pyle D. (1999). Data preparation for data mining. Morgan Kaufman, San Francisco, 560 p.
Pitush K.A., Stevens J.P. (2016). Applied Multivariate Statistics for the Social Sciences. Analyses with SAS and IBM's SPSS (sixth edition). Taylor & Francis, 814 p.
Peng L., Stuart E.A., Allison D.B. (2015). Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA, 314 (18). pp. 1966 – 1967.
Penny K.I., Jollife I.T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. The Statistician, 50 (3), pp. 295 – 308.
Porkodi F.A. (2014). A comparison of filter based feature selection algorithms: An overview. Int. Journal of Innovative Research in Technology & Science, 2, (2), pp. 1957 – 2000.
Palczewski K., Salabun W. (2019). Influence of various normalization methods in PROMETHEE: an empirical study on the selection of the airport location. Procedia Computer Science, 159, pp. 2051 – 2060.
Pandey A., Jain A. (2017). Analysis of KNN Algorithm using Various Normalization Techniques. Int Journal Computer Network and Information Security, 11, pp. 36 – 42.
Quinlan J.R. (1986). Induction of Decision Trees. Machine Learning, 1 (1), pp. 81 – 106.
Rubin D.B. (1976). Inference and missing data. Biometrika, 63 (3), pp. 581 – 592.
Rubin D.B. (1996). Multiple Imputation after 18+ Years. JASA, 91, pp. 473 – 478, doi: 10.1080/01621459.1996.10476908.
Rubin D.B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc., New York.
Rowley J. (2007). The wisdom hierarchy: representation of the DIKW hierarchy. Journal of Information Science, 33 (2), pp. 163 – 180, /doi: 10.1177/0165551506070706.
Rubin D.B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91 (434), pp. 473 – 489.
Rubin D.B. (1976). Inference and Missing Data. Biometrika, 63, pp. 581 – 592.
Rubin D.B. (1977). Formalizing Subjective Notion about The Effect of Nonrespondents in Sample Surveys. Journal of the American Statistical Association, 72, pp. 538 – 543.
Rubin D.b. (1987). Multiple Imputation for Nonresponse in Surveys. New York, John Wiley & Sons.
Rousseuw P.J, Leory A. (1987). Robust Regression and Outlier Detection. Wiley Series In Probability and Statistics.
Rousseuw P.J., van Zomeren B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85, pp. 633 – 639.
Rousseeuw P.J., Hubert M. (2011). Robust statistics for outlier detection. WIREs Data Mining Knowledge Discovery, 1, pp. 73 – 79.
Rousseeuw P.J. (1987). Silhouettes: A Practical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20, pp. 53 – 65.
Rissanen J. (2007). Information and complexity in statistical modeling. New York, Springer, 142 p.
Schafer J.L. (1997). Analysis of incomplete multivariate data. Chapman & Hall/CRC, London.
Schafer J.L. (1999). Multiple imputation: a primer. Stat. Methods in Medicine. 8 (1), pp. 3 – 15, doi: 10.1177/096228029900800102.
Schafer J.L., Olsen M.K. (1998). Multiple Imputation for Multivariate Missing-Data Problems.

A Data Analyst's Perspective. Multivar. Behav. Res., 33 (4), pp. 545 – 571, doi: 10.1207/s1532790mbr2204_5.
Schafer J.L., Graham J.W. (2002). Missing data: Our view of the state of art. Psychological Methods., 7 (2), pp. 147 – 177, doi: org/10.1037/1082-989X.7.2.147.
Schneider T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate, 14, pp. 853 – 871.
Sanders J.D. (2016). Definition Terms: Data, information and Knowledge. SAI Computing Conference, 6 p. doi: 10.1109/SAI.2016.7555986.
Stonier T. (1993). The wealth of information. London: Thames/Methuen.
Stonier T. (1997). Information and meaning – An Evolutionary perspective. Berlin, Springer.
Safarinejadan B., Menhaj M., Karrari M.B. (2010). A distributed EM algorithm to estimate the parameters of a finite mixture of components. Knowl. Inf. Syst., pp. 267 – 292.
Shylaja B., Kumar R.S. (2018). Traditional versus modern missing data handling techniques: an overview. Int. Journal of Pure and Applied Mathematics, 118 (14), pp. 77 – 83.
Sim J., Lee J.S., Kwon O. (2015). Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications. Mathematical Problems in Engineering, Vol. 2015, 14 p., doi: 10.1155/2015/538613.
Schmitt P., Mandel J., Guedj M., (2015). A comparison of Six Methods for Missing Data Imputation. Journal Biomet. Biostat., 6 (1), pp. 1 – 6.
Schwenger H., Ickstadt K. (2008). Imputing missing genotypes with weighted k nearest neighbor. Technical Report< No. 2008,03, Technische Univesität Dortmund, 15 p.
Sinharay S., Stern H.S., Russel D. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. Psychological Methods, 6 (4), pp. 317 – 329.
Schafer J.L. (1997). Analysis of incomplete multivariate data. Chapman & Hall, New York.
Schafer J.L., Olsen M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analysis perspective. Multivariate Behavioral Research, 33, pp. 545 – 571.
Stephen R.S., Santhamarai K.K., (2017). Detection of Outliers in Regression Model for Medical Data. Int. Journal of Medical Research & Health Sciences, 6 (7), pp. 50 – 56.
Salgato C., Arevedo C., Proença H.M., Vieira S. (2016). Noise Versus Outliers. In Secondary Analysis of Electronic Health Records. Springer, Cham, htpps://doi.org/10.1007/978-3-319-43742-2_2.
Susanti Y., Prativi H., Sri Sulistijowati H., Liana T. (2014). M Estimation, S Estimation, and MM Estimation in Robust Regression. Int. Journal of Pure and Applied Mathematics, 91 (3), pp. 349 – 360.
Soni K.G., Patel A. (2017). Comparative Analysis of k-means and k-medoids Algorithm on IRIS Data. Int. Journal of Computational Intelligence Research, 13, (5), pp. 899 – 906.
Suarez-Alvarez M.N., Pham D.-T., Prostov M.Y., Prostov J.I. (2012). Statistical approach to normalization of feature vectors and clustering of mixed databases. Proceedings of the Royal Society, 468, pp. 2630 – 2651, htpps://doi:10.1098/rspa.2011.0704.
Sakia R. (1992). The Box-Cox Transformation Technique: A Review. Journal of the Royal Statistical Society. Series D (Statistician), pp. 169 – 178.
Tabachnik B.G., Fidell L.s. (2012). Using multivariate statistics. 6th Edition, Allyn & Becon, Needhaum, Heights, MA.
Tukay J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.
Tutz G., Ramzan Sh. (2014). Improved Methods for the Imputation of Missing Data by Nearest Neighbor Methods. Technical Report Number 172, Department of Statistics, University of Münich, 22 p.
Trojanskaja O., Cantor M., Sherlok G., Brown P., Hastie T., Tilshirani R., Botstein D., Altman R.B. (2007). Missing value estimation method for DNA microarrays. Bioinformatics, 17, pp. 1 – 6.
Tanner M.A. Wong W.H. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82, pp. 528 – 550.
Tukey J. (1997). Exploratory data analysis. Pearson.
Tharwad A. (2017). *Principal component analysis – a tutorial*. https://www.academia.edu/36699625/Principal-Component-Analysis.
Thorndike R.L. (1953). Who Belongs in the Family. Psychometrika, 18, (4), pp. 267 – 271.

Tukey J.W. (1957). The comparative anatomy of transformations. Annals of Mathematical Statistics, 28, pp. 602 – 632.
Vafaei N., Ribeiro R.A., Camarinha-Matos L.M. (2018). Data normalization techniques in decision making: case study with TOPSIS method. Int Journal Information and Decision Sciences, 10 (1).
Wong A.K.C., Chin D.K.Y. (1987). Synthesizing statistical knowledge from incomplete mixed-mode data. IEEE Trans. Pattern Anal. Mach. Intell., 9 (6), pp. 796 – 805. doi: 10.1109/tpami.1987.4767986.
Winship Ch., Mare R.D. (1984). Regression Models with Ordinal Variables. American Sociological Review, 49 (4): 512, pp. 511 – 525, DOI:10.2307/2095465.
Wu J. (2012). Advances in k-means Clustering: A Data Mining Thinking, Springer, 194 p.
Xu G., Zong Yu, Yang Zh. (2013). Applied Data Mining. Taylor & Francis Group.
Yamini C., Kowsalya (2015). Survey of classification of incomplete data handling technologies. Int. Journal of Advance Research in Science and Engineering, Vol 4, No. 4, Special Issue (01), pp. 111 – 118.
Zins Ch. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. Journal of the American Society for Information Science and Technology, 58 (4), pp. 479 -493.
Zaki M.J., Wagner M. Jr. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 607 p.
Ужга-Ребров О.И. (2021). Введение в анализ данных. Резекне.

PIELIKUMI

P1. REGRESIJAS ATRIBŪTU VĒRTĪBU SADALĪJUMA PARAMETRU APRĒĶINI

Atribūtu vērtību kopu var uzskatīt par šo vērtību sadalījumu. Datu priekšapstrādes un analīzes problēmās bieži vien ir nepieciešams raksturot atribūtu vērtību sadalījumu. Raksturošanu var saistīt ar sadalījuma formas noteikšanu, izvietojuma un diapazona parametru aprēķināšanu un attiecību novērtēšanu starp atsevišķām atribūtu vērtību kopām. Šajā pielikumā ir sniegtas pieejas izvietojuma parametru vērtību un atribūtu vērtību sadalījuma diapazona aprēķināšanai un regresijas noteikšanai starp atsevišķu atribūtu vērtību kopām.

Sadalījumu un regresijas parametru tipi ir atkarīgi no skalu veida, kurā tiek mērītas atribūtu vērtības. Turpmāk tiks parādītas aprēķinātās izteiksmes katram skalu tipam vispārpieņemtajā šo skalu hierarhijā.

P1.1. Nominālās skalas

Ja atribūtu vērtības mēra nominālā skalā, tas nozīmē tikai to, ka katrs no atbilstošajiem objektiem vai subjektiem pieder kādai no šīs skalas kategorijām. Nominālās skalas piemēri var ietvert subjektu dzimumu, politisko partiju piederību, subjektu matu krāsu un tamlīdzīgus elementus.

Nominālā skala ir pati vienkāršākā no visām mērījumu skalām. Tā sniedz ļoti ierobežotu informāciju par lietu būtību. Tomēr diezgan bieži tieši šajā skalā tiek mērītas atribūtu vērtības.

Kādi parametri var raksturot atribūtu vērtību sadalījumu, kas mērīts nominālā skalā? Vienīgais šāda sadalījuma parametrs ir *moda*. Moda ir statistisks raksturlielums, kas nosaka skaitli, kam ir vislielākais biežums kādā datu kopā.

Pieņem, ka j atribūts raksturo subjektu dzimumu. Šī atribūta vērtības kodē šādi: 1 – vīrietis, 2 – sieviete. Pieņem, ka ir dota šāda atribūtu a_j vērtību kopa:

$a_j^T = 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1.$

(Šajā piemērā attēlo atribūtu a_j vērtības formā a_j^T, jo datu tabulās atribūtu vērtības ir attēlotas šīs tabulas kolonnās, un dotajā piemērā šīs vērtības ir attēlotas rindas veidā.)

No dotā piemēra ir skaidrs, ka dati ietver sevī 9 vīriešus un 6 sievietes, tātad sadalījuma moda ir vienāda ar $\mathrm{mod}(a_j) = 1$.

Citiem vārdiem sakot, sadalījuma parametrs ir iestatīts vērtībai 1 – vīrieši.

Jādefinē regresijas jēdziens starp diviem mainīgajiem (divām atribūtu vērtību kopām), kas mērīti nominālajā skalā. Tā kā standarta lineārās regresijas modelis šim gadījumam nav piemērots (skat. 11. nodaļu darbā [Pitush K.A., Stewens I.R., 2016] un 13.6. sadaļu darbā [Agresti A., Franklin Ch., 2013]), tiek izmantots *loģistiskās regresijas modelis*. Šī modeļa būtība ir šāda. Dota faktoriālā pazīme (mainīgais) – $\mathbf{X}$ un rezultējošā pazīme (mainīgais) – $\mathbf{Y}$. Mainīgās vērtības tiek izteiktas nominālajās skalās. Pieņem, ka mainīgie $\mathbf{X}$ un $\mathbf{Y}$ ietver divas nominālās kategorijas. (Vēlāk šis ierobežojums tiks paplašināts līdz patvaļīgam skaitam mainīgā lieluma $\mathbf{X}$ nominālo kategoriju).

Loģistiskās regresijas būtību var skaidrot ar konkrētu piemēru. Pirms gaidāmajām pilsētas vadības vēlēšanām tika veikta iedzīvotāju aptauja: kurš no diviem kandidātiem A vai B ir piemērotākais pilsētas mēra amatam. Aptaujā kopumā piedalījās 100 respondenti, no tiem 70 bija vīrieši un 30 sievietes. Respondentu viedokļi dalījās šādi: vīrieši: par A – 50, par B – 20; sievietes: par A – 10, par B – 20.

Uzdevums ir noteikt attiecības starp katra kandidāta uzvaras varbūtību un respondentu viedokļiem katrā nominālajā kategorijā.

Sākotnējos datus kodē šādi:

$y = 1$ – kandidāta A uzvara vēlēšanās;

$y = 2$ – kandidāta B uzvara vēlēšanās;

$x = 1$ – aptaujas dalībnieku vīriešu kategorija;

$x = 2$ – aptaujas dalībnieku sieviešu kategorija.

Sākotnējos datus uzdod P1.1.1. tabulas veidā.

P1.1.1. tabula

Iedzīvotāju aptaujas rezultāti par diviem kandidātiem

Respondentu grupas	Kandidāti		
	A $(y=1)$	*B* $(y=2)$	Summa rindās
Vīrieši $(x=1)$	50	10	60
Sievietes $(x=2)$	20	20	40

Aptaujas rezultātus no P1.1.1. tabulas pārveido ar biežuma (varbūtību) aplēsēm (P1.1.2. tabula).

P1.1.2. tabula

Kandidāta uzvaras varbūtību aplēses, pamatojoties uz iedzīvotāju aptaujas rezultātiem

Respondentu grupas	Kandidāti		
	A $(y=1)$	*B* $(y=2)$	Summa rindās
Vīrieši $(x=1)$	0.833	0.167	1. 000
Sievietes $(x=2)$	0.500	0.500	1.000

Šajā tabulā varbūtību vērtības ir aprēķinātas kā vīriešu īpatsvars, kuri attiecīgi deva priekšroku kandidātiem A un B, kā arī sieviešu īpatsvars, kas deva priekšroku kandidātiem A un B.

Kandidātu izredzes uzvarēt aprēķina, pamatojoties uz P1.1.2. tabulu, šādi:

$$\mathrm{Ch}(y=j)/(x=i)=\frac{p(y=j)/(x=i)}{1-p(y=j)/(x=i)},\ i,j=1,2 \qquad \text{(P.1.1.1)}$$

kur $p(y=j)/(x=i)$ – kandidāta j uzvaras iespējamība, pamatojoties uz respondentu grupas i aptaujas rezultātiem.

Tika iegūti šādi rezultāti:

$$\mathrm{Ch}(y=1)/(x=1)=\frac{0.873}{1-0.873}=\frac{0.873}{0.167}=4.988;$$

$$\mathrm{Ch}(y=1)/(x=2)=\frac{0.500}{1-0.500}=\frac{0.500}{0.500}=1.000;$$

$$\mathrm{Ch}(y=2)/(x=1)=\frac{0.167}{1-0.167}=\frac{0.167}{0.833}=0.200;$$

$$\mathrm{Ch}(y=2)/(x=2)=\frac{0.500}{1-0.500}=\frac{0.500}{0.500}=1.000;$$

Iegūtie rezultāti ir apkopoti P.1.1.3. tabulā.

P1.1.3. tabula

Kandidātu izredžu aplēses, pamatojoties uz aptaujas rezultātiem

Respondentu grupas	Kandidāti	
	A $(y=1)$	*B* $(y=2)$
Vīrieši $(x=1)$	4.988	1.000
Sievietes $(x=2)$	0.200	1.000

Ja ir zināmas kandidātu iespējas uzvarēt noteiktos apstākļos, tad uz tā pamata var aprēķināt viņu uzvaras varbūtību. Piemēram, $x = 1 \text{ un } y = 1$

$$p(y=1)/(x=1) = \frac{p(y=1)/(x=1)}{1+p(y=1)/(x=1)}. \qquad \text{(P1.1.2)}$$

Ievietojot skaitliskās vērtības, iegūst

$$p(y=1)/(x=1) = \frac{4.988}{1+4.988} = \frac{4.988}{5.988} = 0.714.$$

Aprēķinātā varbūtības vērtība sakrīt ar tās vērtību P1.1.2. tabulā. Pārējās varbūtības tiek aprēķinātas pēc analoģijas.

Jāņem vērā, ka varbūtības aprēķiniem, izmantojot vienādojumu (P1.1.2) dotajā piemērā nav jēgas, jo pašas varbūtības vērtības tiek aprēķinātas, pamatojoties uz aplēstajām varbūtības vērtībām. Tomēr šo procedūru var izmantot praktiskos loģistiskās regresijas pielietojumos.

Tāpēc jāievieš *izredžu attiecības* jēdziens:

$$\xi(y=1) = \frac{Ch(y=1)/(x=1)}{Ch(y=1)/(x=2)}; \qquad \text{(P1.1.3.a)}$$

$$\xi(y=2) = \frac{Ch(y=2)/(x=1)}{Ch(y=2)/(x=2)}. \qquad \text{(P1.1.3.b)}$$

Aizstājot skaitliskās vērtības, ir

$$\xi(y=1) = \frac{4.988}{1.000} = 4.988; \xi(y=2) = \frac{0.200}{1.000} = 0.200$$

Būtībā izredžu attiecība parāda, cik reižu izredzes uzvarēt kandidātam A vai B pēc vīriešu vērtējuma ir lielākas vai mazākas nekā šo kandidātu izredzes uzvarēt, pamatojoties uz sieviešu vērtējumiem.

Kā minēts iepriekš, loģistiskās regresijas mērķis ir noteikt rezultatīvā mainīgā lieluma kategoriju varbūtības (dotajā piemērā: kandidāts A vai kandidāts B uzvarēs gaidāmajās vēlēšanās) atkarībā no tiem dotajām priekšrocībām (šajā piemērā: vīrieši vai sievietes). Aprēķinātās varbūtības P1.1.2. tabulā un izredžu aplēšu P1.1.3. tabulā nav adekvātas aplēses, lai noteiktu attiecīgo regresiju (sīkāku informāciju skatīt darba 11. nodaļā [Pitush K.A., Stewens I.R., 2016] un 13.6. sadaļā [Agresti A., Franklin Ch., 2013]). Lineārās regresijas modelis var būt piemērots gadījumos, kad rezultatīvā mainīgā vērtības ir sadalītas pēc normālā sadalījuma likuma. Pretējā gadījumā vērtībām jāatrodas simetriski attiecībā pret centrālo vērtību un jābūt neierobežotām (t. i., jāatrodas diapazonā $-\infty, +\infty$).

Ja izredžu aplēses piemērotā veidā var transformēt, lai iegūtu šo transformēto aprēķinu sadalījumu, kas tuvinās normālajam sadalījumam, tad var izveidot lineārās regresijas modeli.

Transformācija, kas var atrisināt šo problēmu, ir izredžu aprēķinu transformācija, izmantojot naturālos logaritmus:

$$e^{\ln Ch} = Ch. \qquad \text{(P1.1.4)}$$

Izmantojot šādu transformāciju, var izveidot lineāru sakarību starp faktoriālā mainīgā kategorijām un rezultatīvā mainīgā lieluma ln kategoriju izredzēm. Piemēram, iepriekš minētajam piemēram un $y = 1$:

$$\ln(Ch(y=1)) = \beta_0 + \beta_1 x, \qquad \text{(P1.1.5)}$$

kur β_0, β_1 – regresijas koeficienti.

Aprēķinot vērtību $\ln(Ch(y=j))$, izmantojot vienādojumu (P1.1.4), var aprēķināt iespēju vērtību $Ch(y=1)$ un ar izteiksmi (P1.1.2) aprēķināt vēlamās varbūtības vērtību $p(y=j)$, $j = 1, 2$.

Iepriekš dotajā piemērā darbojās ar nominālām aptaujāto iedzīvotāju kategorijām: vīriešiem un sievietēm. Bet patiesībā skaitļi P1.1.1. tabulā ir norādīti nepārtrauktā reālo skaitļu x skalā. Var aprēķināt varbūtību, ka kandidāts A uzvarēs, dodot priekšroku 50, 100, 200 vai jebkuram citam vīriešu skaitam, vai patvaļīgam sieviešu skaitam, vai jebkuram patvaļīgam vīriešu un sieviešu

skaitam. Pretējā gadījumā ir vajadzīgas vispārīgas attiecības starp kandidātam A piešķirto priekšrocību skaitu katrā respondentu kategorijā vai abās kategorijās.

Tas pats pamatojums attiecas uz aplēsēm par iespējamību, ka kandidāts B uzvarēs vēlēšanās.

No tā izriet, ka jāizmanto mainīgais x, kura vērtības atspoguļo interesējošo iedzīvotāju skaitu gan katrai nominālajai kategorijai "vīrieši", "sievietes", gan to kopējo vērtību. Šāds mainīgais dotajā kontekstā bieži tiek saukts par fiktīvo mainīgo (*dummy variable*).

Vispārīgā gadījumā loģistiskās regresijas grafikam ir viena no S-veida formām, kas parādīta P1.2.1. attēlā. P1.2.1.a attēls atbilst gadījumam $\beta_1 > 0$, P1.2.1.b attēls – gadījumam $\beta_1 < 0$.

Iepriekš minētajā piemērā faktoriālais mainīgais un rezultatīvais mainīgais bija bināri nominālie mainīgie. Bet loģistiskās regresijas jēdzienu var attiecināt arī uz gadījumu, kad rezultatīvais mainīgais ir binārs nominālais mainīgais un faktoriālajam mainīgajam ir m kategorijas ($m > 2$). Šajā gadījumā loģistiskās regresijas vienādojums iegūst formu:

$$\ln\left(\mathrm{Ch}(y=1)\right) = \beta_0 + \beta_1 X_1 + ... + \beta_m X_m, \qquad \text{(P1.1.6)}$$

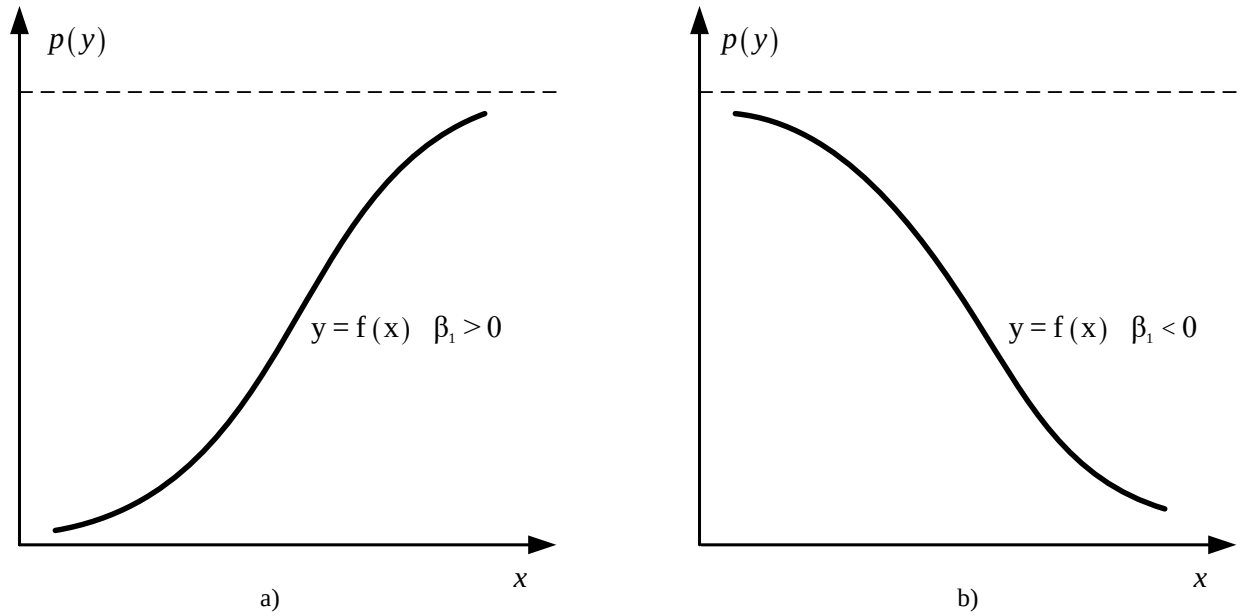
kur β_i, $i = 0,1,...,m$ – regresijas koeficienti.

Ja ir zināmas konstantes β_0 un regresijas koeficienta vērtības β_1 regresijas vienādojumā (P1.2.5), tad noteiktai fiktīvā mainīgā vērtībai x atbilstošās nominālās kategorijas varbūtības vērtību y_j var tieši aprēķināt, izmantojot šādu vienādojumu:

$$p(y_j) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}. \qquad \text{(P1.2.7)}$$

Ja faktoriālajam mainīgajam X ir m vērtības, tad atbilstošās nominālās kategorijas varbūtības vērtību y_j var tieši aprēķināt no vienādojuma:

$$p(y_j) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}}. \qquad \text{(P1.2.8)}$$



P2.1.1. attēls. **Loģistiskās regresijas diagrammas a)** $\beta_1 > 0$; **b)** $\beta_1 < 0$

Kādā veidā var tikt aprēķināti konstantu un regresijas koeficientu vērtības vienādojumos (P1.2.7), (P1.2.8)? Šī ir sarežģīta skaitļošanas problēma, kuras pamatā ir maksimālās varbūtības princips. Praksē šī problēma tiek atrisināta, izmantojot atbilstošu programmatūru.

P1.2. Ordinālās skalas

Ordinālās skalas kopā ar nominālajām skalām veido kategorisko (kategorijas) skalu klasi. Atšķirībā no nominālajām skalām ordinālās skalas ļauj sakārtot kategorijas datus atbilstoši noteiktam atribūtam.

Tipisks ordinālās piemērs ir Likerta skala. Šo skalu ļoti bieži izmanto dažāda veida aptauju organizēšanai, kuru mērķis ir noskaidrot respondentu viedokļus par kādu konkrētu priekšlikumu, plānu, reformu u. c. Piemēram, varētu organizēt aptauju, lai noskaidrotu iedzīvotāju attieksmi pret pašvaldības ierosinājumu mainīt pilsētas sabiedriskā transporta kustības sarakstus. Šīs skalas kategorijas var būt:

1 – kategoriski nepiekrītu;
2 – nepiekrītu;
3 – neitrāls;
4 – piekrītu;
5 – pilnībā piekrītu.

Pilsētas iedzīvotājs, kas piedalās aptaujā, vienkārši atzīmē kategoriju, kas atbilst viņa attieksmei pret vērtējamo piedāvājumu. Ja aptauja tiks veikta noteiktu iedzīvotāju kategoriju vidū (piemēram, dažādu pilsētas rajonu iedzīvotāji), tad būs daudz atbilžu kategoriju daudzām iedzīvotāju kategorijām. Visus var kompakti attēlot tabulas veidā, kurā katra rinda atbilst kādai no aptaujāto iedzīvotāju kategorijām, bet katra kolonna atbilst kādai no atbilžu kategorijām. Skaitlis rindas i un kolonas j krustpunktā apzīmē to iedzīvotāju skaitu no i kategorijas, kuri kā atbildi izvēlējušies kategoriju j no visu vērtēšanas kategoriju saraksta.

Vēl viens mērīšanas piemērs ordinālajā skalā ir respondentu apmierinātības vērtējums par viņu dalību kādā pasākumā, izstādē, teātra vai kino apmeklējumos u. tml. Šādas skalas kategorijas var būt:

1 – pilnīgi neapmierināts;
2 – neapmierināts;
3 – vienaldzīgs;
4 – apmierināts;
5 – pilnībā apmierināts.

Vēl viens mērīšanas piemērs ordinālajā skalā. Atkarībā no personas vecuma viņu var iedalīt vienā no šādām kategorijām:

1 – bērns;
2 – pusaudzis;
3 – jaunietis;
4 – pusmūža cilvēks;
5 – vecāka gadagājuma cilvēks.

Visiem iepriekš sniegtajiem kārtas skalu piemēriem raksturīga iezīme ir tā, ka visas kategorijas var sakārtot pēc sakritības ar pašvaldības priekšlikumu, pēc apmierinātības pakāpes un pēc vecuma. Bet par šādām skalām nav iespējams pateikt, *cik ļoti* vai *cik daudz* kategorijas atšķiras. Tādējādi nevar konstatēt, kādā mērā atšķirība starp bērnu un pusaudzi vai pusaudzi un jaunieti ir lielāka vai mazāka.

Ja nosaka skaitliskas robežas starp vecuma kategorijām un sadala noteiktas izlases indivīdus (piemēram, novada iedzīvotājus) iegūtajās kategorijās, atkal iegūs mērījumu pēc ordinālās skalas. Lieta tāda, ka personas tiek iedalītas vecuma kategorijās, pamatojoties uz viņu faktisko vecumu. Tādējādi tiek izveidoti ordināli dati, un šie dati ir jāapstrādā saskaņā ar ordinālo skalu noteikumiem.

(Nedrīkst jaukt indivīdu vecuma salīdzināšanu ar viņu vecuma kategoriju salīdzināšanu. Pieņem, ka ir divi indivīdi A un B. Indivīda A vecums ir 13 gadi un 6 mēneši, indivīda B vecums ir 18 gadi un 3 mēneši. Var droši apgalvot, ka indivīds B ir 4 gadus un 9 mēnešus vecāks par indivīdu A. Bet, ja zina tikai to, ka indivīds A pieder kategorijai "Pusaudži" un indivīds B pieder kategorijai "Jaunieši", tad vienīgais, ko var teikt ir tas, ka indivīds B ir vecāks par indivīdu A).

Lai atvieglotu datu analīzi, to ordinālās kategorijas var atbilstoši kodēt. Iepriekš minētajos piemēros kategoriju numurus var izmantot kā šo kategoriju kodus.

Tipisks ordinālās skalas kategoriju kodēšanas piemērs ir studentu snieguma rādītāji:
2 – ļoti vājš;
3 – vājš;
4 – gandrīz apmierinoši;
5 – apmierinoši;
6 – gandrīz labi;
7 – labi;
8 – ļoti labi;
9 – teicami;
10 – lieliski.

Skaitlisko snieguma novērtējumus var interpretēt kā atbilstošo verbālo un kategorisko kategoriju kodus, tāpēc mēģinājums novērtēt studenta vidējo sniegumu, aprēķinot viņa punktu vidējo vērtību, ir nekorekti.

Kā var parādīt un vizualizēt datus, kas mērīti ordinālajā skalā? Šīs problēmas risinājumu var apskatīt uz piemēra pamata.

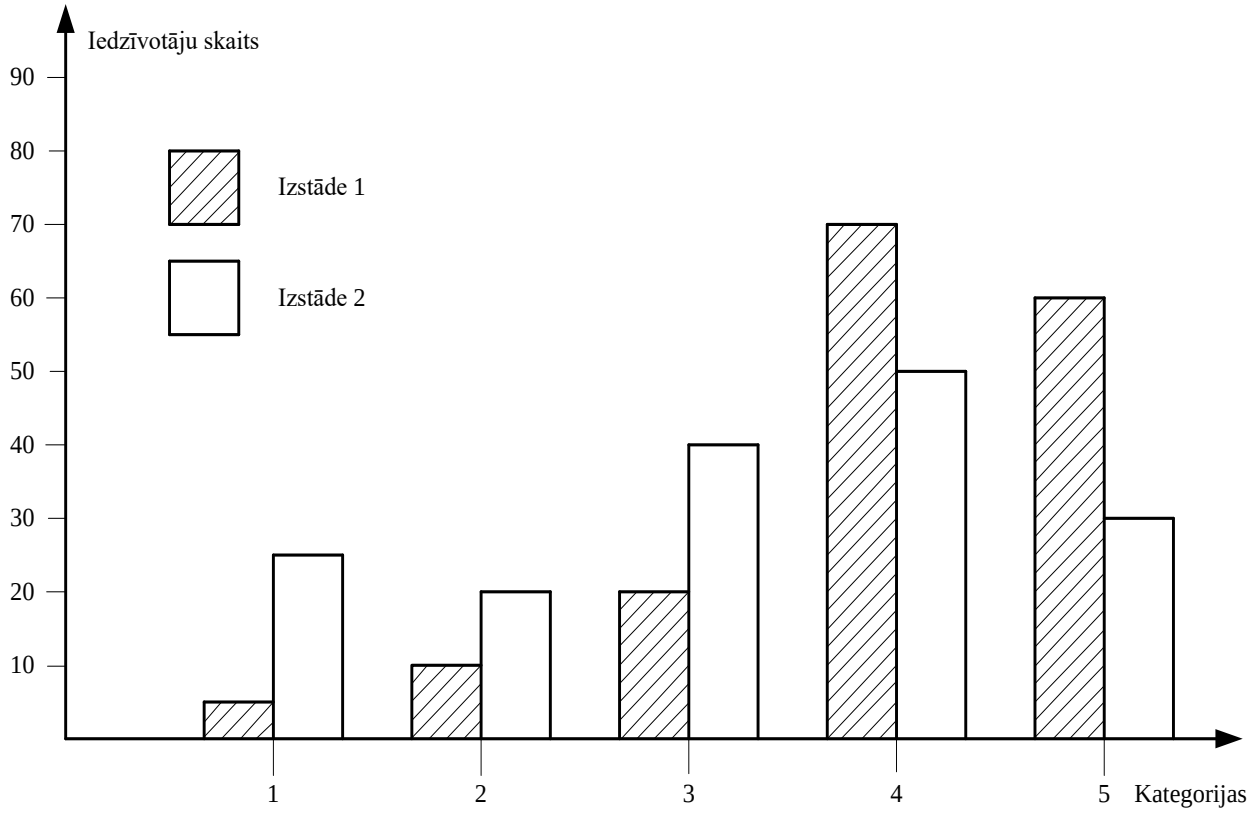
Pieņem, ka pilsētā tiek rīkotas divas izstādes vienlaikus. Izstādes apmeklētāji tika lūgti novērtēt savu apmierinātības līmeni ar izstādēm, atzīmējot vienu no kategorijām: "1 – pilnīgi neapmierināts..., 5 – pilnīgi apmierināts", kas tika parādītas iepriekš šajā sadaļā. Apkopotie rezultāti ir doti P1.2.1. tabulā.

P1.2.1. tabula

Apmierinātības rezultāti izstādes apmeklētājiem

Izstādes	Apmierinātības kategorijas					Summa pa rindām
	1	2	3	4	5	
1	5	10	20	70	60	165
2	25	20	40	50	30	165

Šie dati ir grafiski vizualizēti P1.2.1. attēlā.



P1.2.1. attēls. **Datu grafiskā vizualizācija datiem no P1.2.1. tabulas**

Vienkārša datu analīze P1.2.1. tabulā un P1.2.1. attēlā ļauj izdarīt acīmredzamu secinājumu, ka 1. izstādes apmeklētāji ir apmierinātāki nekā 2. izstādes apmeklētāji.

Ordinālo datu analīzē var būt lietderīgi aprēķināt un vizualizēt rezultātu proporcijas (biežumu) katrā kategorijā. Iepriekš minētajā piemērā vērtējumu daļas biežumi ir parādīti P1.2.2. tabulā.

P1.2.2. tabula

Respondentu vērtējumu proporcijas (biežums) pa kategorijām saskaņā ar datiem no P1.2.1. tabulas

Izstādes	Apmierinātības kategorijas					Summa pa rindām
	1	2	3	4	5	
1	0.030	0.061	0.121	0.424	0.364	1000
2	0.152	0.121	0.242	0.303	0.182	1000

Lai vizuāli attēlotu novērtējumu biežumu pa kategorijām, var aprēķināt uzkrātās biežumu vērtības. Šīs vērtības P1.2.2. tabulas datiem ir parādītas P1.2.3. tabulā.

P1.2.3. tabula

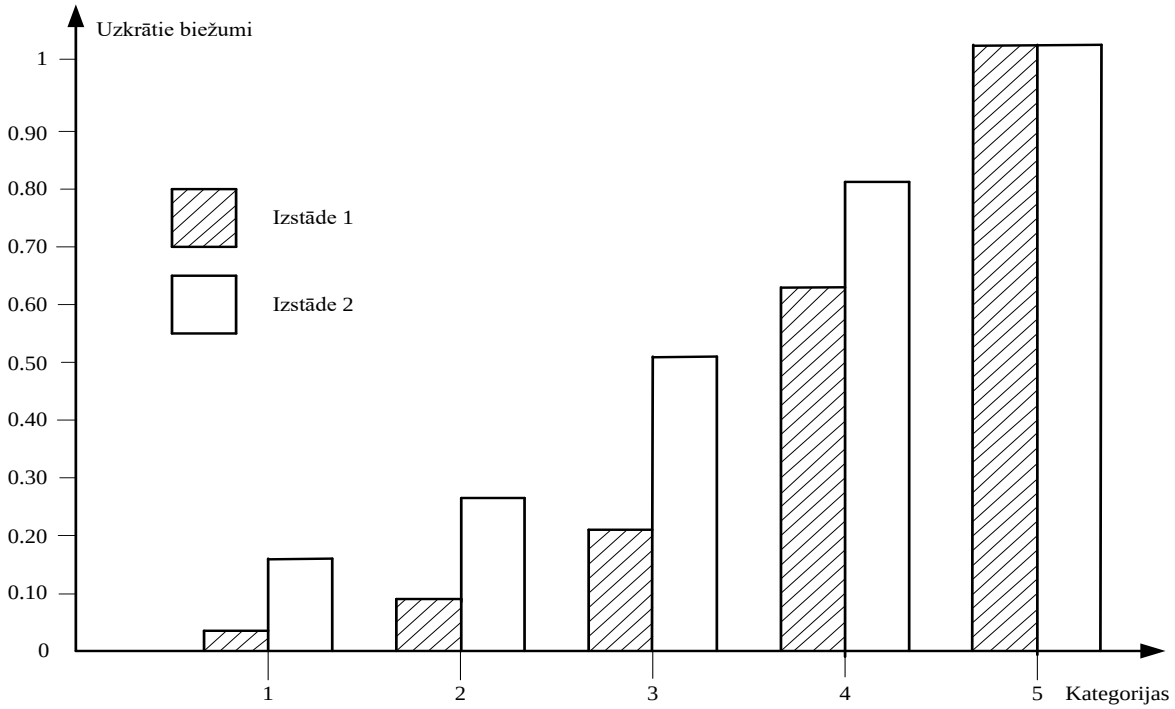
Uzkrātās biežumu vērtības datiem no P1.2.2. tabulas

Izstādes	Apmierinātības kategorijas				
	1	2	3	4	5
1	0.030	0.091	0.212	0.636	1000
2	0.152	0.273	0.512	0.818	1000

Dati no P1.2.3. tabulas ir grafiski vizualizēti P1.2.2. attēlā.

Uzkrāto biežumu grafiki šajā attēlā skaidri parāda, ka 1. izstādes apmeklētājiem uzkrātie biežumi lēnām pieaug mazo kategoriju numuru zonā un strauji palielinās 4. un 5. kategorijā. Tas ir mazā neapmierināto apmeklētāju skaita sekas un lielais apmierināto apmeklētāju skaits 1. izstādē. 2. izstādes apmeklētājiem ir cita aina. Uzkrātie biežumi strauji pieaug mazo kategoriju numuru zonā un lēnām palielinās 4. un 5. kategorijā. Tas ir lielā neapmierināto apmeklētāju skaita sekas.

Kā tiek noteikts izvietojuma parametrs ordinālās skalas kategoriju sadalījumam? Pirmā iespēja ir izmantot modu kā kategoriju ar lielāko vērtību skaitu. Apmierinātības kategoriju sadalījumam starp izstādes apmeklētājiem moda ir 4. kategorija abām izstādēm.



P1.2.2. attēls. **Uzkrāto biežumu grafiskā vizualizācija datiem no P1.2.3. tabulas**

Otra iespēja ir izmantot sadalījuma mediānu. Mediānas vērtību nosaka pēc vienādojumiem (n ir kategoriju skaits):

$$med = \frac{n}{2},\ n \text{ - pāra;} \qquad \text{(P1.2.1.a)}$$

$$med = \frac{n+2}{2},\ n \text{ - nepāra.} \qquad \text{(P 1.2.1.b)}$$

Dotajā piemērā ar izstādēm sadalījuma mediāna ir 3. kategorija, jo $n = 5$.

Ordinālo kategoriju sadalījumam principā var noteikt diapazona parametrus: rangu un procentuālo diapazonu. Bet šie parametri pārāk informatīvi, tāpēc praksē tos izmanto reti.

Ordinālo kategoriju sadalījumu analīze, pamatojoties uz šo sadalījumu parametriem, šķiet problemātiska, jo šie parametri sniedz maz būtiskas informācijas. Tāpēc šādu sadalījumu analīzei tiek plaši izmantotas tiešās metodes, tostarp metodes, kuru pamatā ir sadalījumu vizualizācija (sk. P1.2.1, P1.2.2. attēlu).

Jāatzīmē, ka praksē ordinālo skalu dati bieži tiek analizēti, izmantojot metodes, kas paredzētas kvantitatīviem datiem, piemēram, aprēķinot šādu datu vidējo vērtību un variācijas. Tomēr tas novest pie nelabvēlīgiem rezultātiem.

Šādu metožu trūkumi ir detalizēti analizēti darbā [Liddell T., Krushke J. K., 2018]. Pirmkārt, analizējot ordinālos datus statistiskajiem modeļiem, kas paredzēti metriskajiem mainīgajiem, piemēram, T-testiem un dispersijas analīzei, var rasties zems pareizas atpazīšanas līmenis vai ietekmes lieluma aplēšu kļūdas. Otrkārt, ordinālo mainīgo sadalījums var būt neparasts, jo īpaši, ja dominē ļoti mazas vai ļoti lielas vērtības. Treškārt, variācijas mainīgajos var atšķirties dažādās grupās, apstākļos, laika diapazonos utt. Tas ļauj nepārprotami secināt, ka ordinālie mainīgie jāanalizē tikai ar metodēm, kas paredzētas tieši šādiem mainīgajiem.

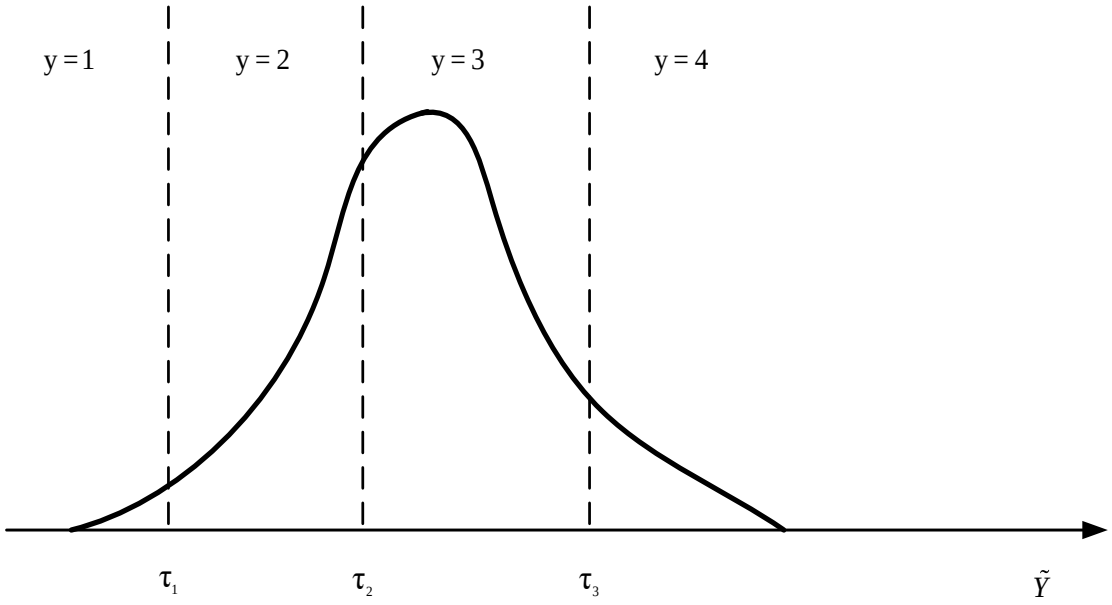
Kā var definēt regresiju ordinālajiem mainīgajiem? P1.1 sadaļā ir definēta loģistiskā regresija nominālajiem mainīgajiem. Tā kā ordinālā skala kopējā skalu hierarhijā ieņem augstāku vietu nekā nominālās skala, tad ordinālajiem mainīgajiem ir iespējams definēt plašāku regresijas modeļu klasi.

Kopumā var definēt 6 regresijas modeļus ordinālajiem mainīgajiem [Ananth G.V., Kleinbaum D.G., 1997]. Šajā sadaļā tiks prezentētas trīs vispārinātas šādu modeļu klases [Bükner P.-Ch., Vuorre M., 2019].

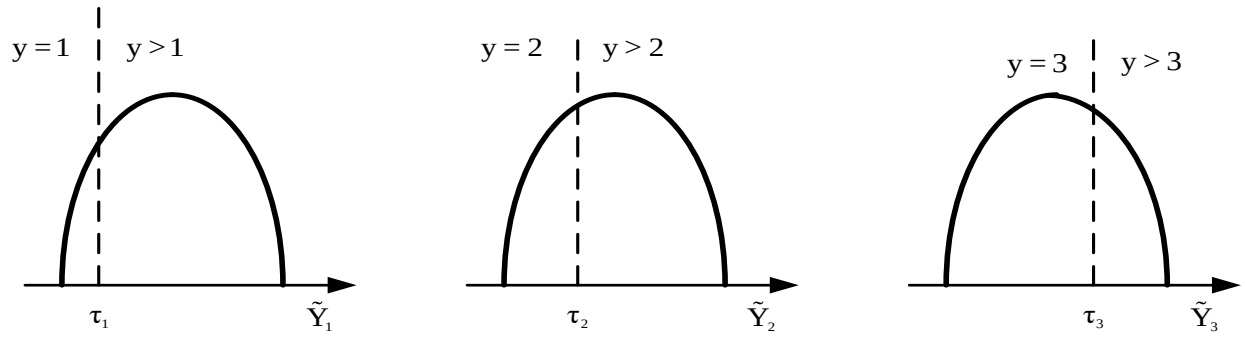
Dota faktoriālā pazīme (mainīgais) **X** un rezultējošā pazīme (mainīgais) **Y**. Ievieš slēpto (latento) mainīgo $\tilde{Y}$, kura vērtības tiek iegūtas, atbilstoši transformējot rezultējošo mainīgo **Y**, un šīs transformētās vērtības tiek sadalītas saskaņā ar normālā sadalījuma likumu. Tad, ja mainīgā lieluma $\tilde{Y}$ sadalījuma blīvuma funkcijai ir tāda forma, kā parādīts P1.2.3. attēlā, ir darīšana ar kumulatīvo regresijas modeli (*cumulative logistic regression models*).

Šajā attēlā y_j, $j = 1,2,3,4$ ir rezultējošā mainīgā **Y** ordinālā mainīgā kategorijas, τ_j, $j = 1,2,3$ ir robežas starp kategorijām. Šajā gadījumā kategorijas y_j tiek uzdotas tā, lai tām atbilstošā latentā mainīgā $\tilde{Y}$ vērtības tiktu sadalītas saskaņā ar normālo sadalījuma likumu.

Citā modelī pieņem, ka ir tāds rezultējošais mainīgais **Y**, lai katrai ordinālajai kategorijai varētu izveidot latento mainīgo $\tilde{Y}$, kura vērtības tiek sadalītas saskaņā ar normālā sadalījuma likumu. Šī situācija ir grafiski parādīta P1.2.4. attēlā. Šo modeli sauc par secīgas regresijas modeli (*sequential regression model*).

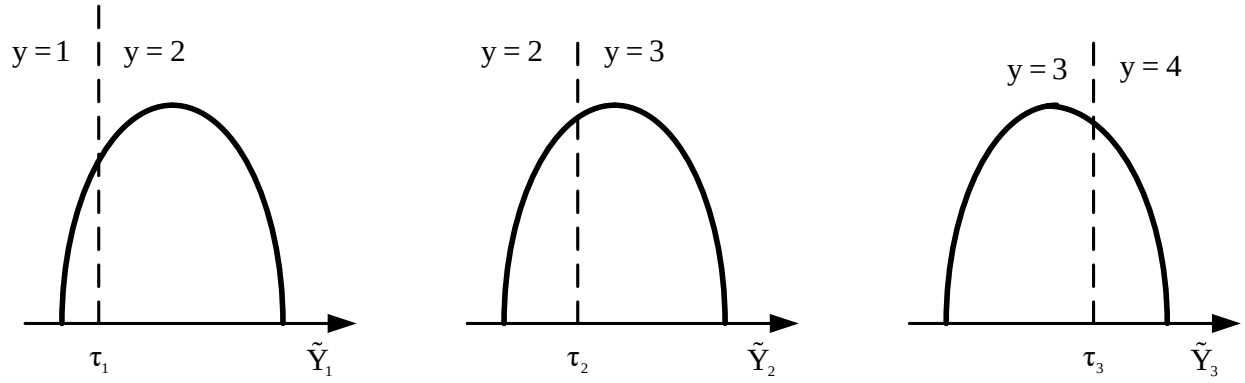


P1.2.3. attēls. **Latentā mainīgā $\tilde{Y}$ vērtību blīvuma sadalījuma grafiks kumulatīvajā regresijas modelī**



P1.2.4. attēls. **Latentā mainīgā $\tilde{Y}$ vērtību blīvuma sadalījuma grafiks secīgas regresijas modelī**

Trešo regresijas modeli sauc par konjugētās kategorijas modeli (*adjacent-category model*). Tā atšķirība no iepriekšējiem modeļiem ir tāda, ka latentie mainīgie $\tilde{Y}_j$ tiek veidoti dažādām rezultējošā mainīgā **Y** ordinālo mainīgo kategorijām. Latento mainīgo sadalījuma funkciju grafiki $\tilde{Y}_j$ šādam regresijas modelim ir parādīti P1.2.5. attēlā.



P1.2.5. attēls. **Latentā mainīgā $\tilde{Y}$ vērtību blīvuma sadalījuma grafiks konjugētās kategorijas modelī**

Interesentiem, kuri vēlas iegūt sīkāku informāciju par apspriestajiem jautājumiem, ieteicams izmantot darbu [Bükner P.-Ch., Vuorre M., 2019].

Savukārt kumulatīvās regresijas modeli var konstruēt dažādos veidos (sīkāk darbā [Ananth C.V., Kleinbaum D.G., 1997]. Šajā sadala mēs iepazīstināsim ar loģistiskā modeļa paplašinājumu, proti, kumulatīvo logit modeli (*Cumulative Logit Model*), jeb precīzāk, vienu no tā variantiem, ko sauc par proporcionālo izredžu modeli. Tālāk sniegtais materiāls ir balstīts uz datiem, kas sniegti darbā [Agresti A., 2002; Agresti A., 2007; Agresti A., 2010]. Iepazīties ar šī regresijas modeļa pamatprocedūrām var, izmantojot konkrētu piemēru.

Pieņem, ka pirms pašvaldību vēlēšanām tika aptaujātas trīs pilsētas iedzīvotāju grupas, lai noskaidrotu, kā viņi vērtē domes darbu iepriekšējā periodā.

Pilsētas iedzīvotājus kodē atspoguļojošā faktoriālā mainīgā **X** ordinālās skalas kategorijas šādi:

x_1 – jaunieši;

x_2 – pusmūža cilvēki;

x_3 – gados veci cilvēki.

Mainīgais **Y** ir saistīts ar domes darba kvalitātes novērtējumu ordinālām kategorijām. Šīs kategorijas kodē šādi:

y_1 – labi;

y_2 – apmierinoši;

y_3 – slikti.

Apkopotie iedzīvotāju aptaujas rezultāti atspoguļoti P1.2.4. tabulā.

P1.2.4. tabula

Pilsētas iedzīvotāju aptaujas rezultāti par iepriekšējās domes darba kvalitāti

Iedzīvotāju kategorijas	Domes darba kvalitātes novērtējuma kategorijas			Summa pa rindām
	y_1	y_2	y_3	
x_1	50	40	10	100
x_2	40	40	20	100
x_3	30	40	30	100

Kā interpretēt skaitļus šajā tabulā? Iedzīvotāju kategoriju x_i skaita apzīmēšanai izmanto indeksu i, $i=1,2,3$ un ar indeksu j, $j=1,2,3$ apzīmē novērtēšanas kategoriju y_j numurus. Tad cipars $n_{11}=50$ pirmās rindas un pirmās kolonas krustpunktā parāda to iedzīvotāju skaitu no kategorijas x_1 "Jaunieši", kuri pilsētas domes darbu novērtējuši ar kategoriju y_1 "Labi". Skaitlis $n_{33}=30$ trešās rindas un trešās kolonas krustpunktā parāda to iedzīvotāju skaitu no kategorijas x_3 "Gados veci cilvēki", kuri pilsētas domes darbu novērtējuši kā y_3 "Vāji".

Kopumā skaitlis n_{ij} atspoguļo iedzīvotāju skaitu no kategorijas x_i, $i=1,2,3$, kuri pilsētas domes darba kvalitāti novērtēja ar kategoriju y_j, $j=1,2,3$.

(Turpmākajā izklāstā pilno kategoriju kodu vietā tiks izmantoti to indeksi i un j. Tas nesagādās grūtības, apgūstot materiālu).

P1.2.4. tabulas pēdējā kolonnā ir norādītas skaitļu summas tabulas rindās – n_i, $i=1,2,3$. Katrs skaitlis apzīmē kopējo attiecīgās kategorijas iedzīvotāju skaitu, kas piedalījās aptaujā. Dotajā gadījumā visi šie skaitļi ir vienādi ar 100.

Aprēķina katras vērtējuma kategorijas nosacītās varbūtības katrai respondentu grupai:

$$p_{j/i}\frac{n_{ij}}{n_i},\ i,j=1,2,3. \qquad \text{(P1.2.2)}$$

Aprēķinu rezultāti P1.2.4. tabulas datiem ir parādīti P1.2.5. tabulā.

P1.2.5. tabula

Domes darba kvalitātes novērtējumu kategoriju nosacītās varbūtības pēc respondentu kategorijas

Iedzīvotāju kategorijas	Domes darba kvalitātes novērtējuma kategorijas			Summa pa rindām
	y_1	y_2	y_3	
x_1	0.500	0.400	0.100	1.000
x_2	0.400	0.400	0.200	1.000
x_3	0.300	0.400	0.300	1.000

Aprēķina nosacīto varbūtību $F_{j/i}$ kumulatīvās vērtības katrai iedzīvotāju kategorijai. Lai to izdarītu, izmanto datus no P1.2.5. tabulas.

$F_{1/1} = p_{1/1} = 0.500$; $F_{2/1} = p_{1/1} + p_{2/1} = 0.500 + 0.400 = 0.900$;

$F_{3/1} = p_{1/1} + p_{2/1} + p_{3/1} = 0.500 + 0.400 + 0.100 = 1.000$.

$F_{1/2} = p_{1/2} = 0.400$; $F_{2/2} = p_{1/2} + p_{2/2} = 0.400 + 0.400 = 0.800$;

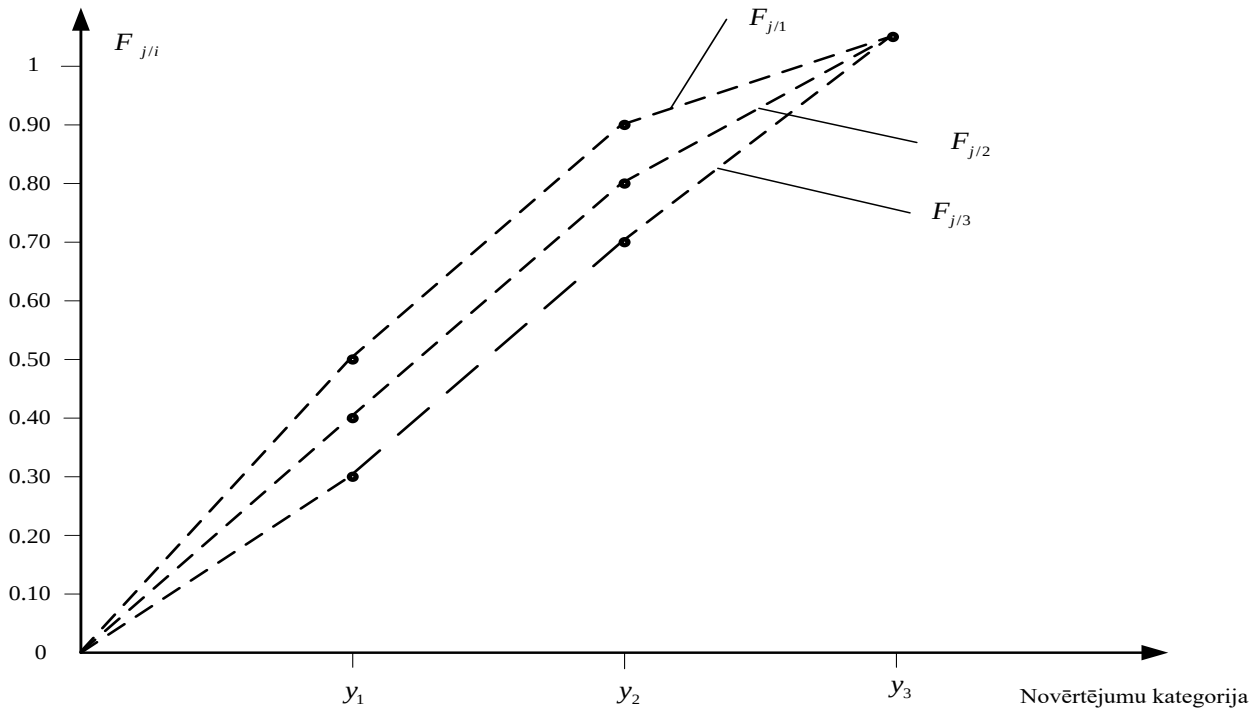
$F_{1/2} = p_{1/2} + p_{2/2} + p_{3/2} = 0.400 + 0.400 + 0.200 = 1.000$.

$F_{1/3} = p_{1/3} = 0.300$; $F_{2/3} = p_{1/3} + p_{2/3} = 0.300 + 0.400 = 0.700$;

$F_{3/3} = p_{1/3} + p_{2/3} + p_{3/3} = 0.300 + 0.400 + 0.300 = 1.000$.

Nosacīto varbūtību kumulatīvās vērtības ir apkopotas P1.2.6. tabulā.

Skaidrības labad iegūtie rezultāti ir grafiski parādīti P1.2.6. attēlā.



P1.2.6. attēls. Nosacīto varbūtību kumulatīvo vērtību $F_{j/i}$ grafiki datiem no P.1.2.6. tabulas

P1.2.6. tabula

Nosacīto varbūtību kumulatīvās vērtības datiem no P1.2.5. tabulas

Iedzīvotāju kategorijas	Domes darba kvalitātes novērtējuma kategorijas		
	y_1	y_2	y_3
x_1	0.500	0.900	1000
x_2	0.400	0.800	1000
x_3	0.300	0.700	1000

Beznosacījumu varbūtību vērtības p_{ij} var aprēķināt, izmantojot vienādojumu:

$$p_{ij} = \frac{n_{ij}}{n},\ i, j = 1, 2, 3, \qquad \text{(P1.2.3)}$$

kur $n = \sum_i n_i = 300$ – kopējais iedzīvotāju skaits, kas uzrādīts P1.2.4. tabulā.

Aprēķinu rezultāti ir parādīti P1.2.7. tabulā.

P1.2.7. tabula

Beznosacījumu varbūtības p_{ij} vērtības datiem no P1.2.4. tabulas

Iedzīvotāju kategorijas	Domes darba kvalitātes novērtējuma kategorijas			Summa pa rindām
	y_1	y_2	y_3	
x_1	0.167	0.133	0.033	0.333
x_2	0.133	0.133	0.067	0.333
x_3	0.100	0.133	0.133	0.333

Ir jāievieš izredžu attiecības jēdziens. Atšķirībā no loģistiskās regresijas, kas aplūkota P1.1. sadaļā, logitregresijas modelī, izmantojot P1.2.4. tabulā esošos datus, var uzreiz aprēķināt izredžu attiecību [Agresti A., 2010]. Tabulas ar izmēriem $r \times c$ datiem var aprēķināt $(r-1)(c-1)$ izredžu attiecību. P1.2.4. tabulai var aprēķināt $2 \times 2 = 4$ izredžu attiecību.

1. Lokālā izredžu attiecība:

$$\xi_{ij}^{L} = \frac{n_{ij} n_{i+1,\,j+1}}{n_{i,\,j+1} n_{i+1,\,j}}. \qquad \text{(P1.2.3)}$$

Aprēķina lokālo izredžu attiecību P1.2.4. tabulas datiem, izmantojot vienādojumu (P1.2.3):

$$\xi_{11}^{L} = \frac{n_{11} n_{22}}{n_{12} n_{21}} = \frac{50 * 40}{40 * 40} = 1.250;$$

$$\xi_{12}^{L} = \frac{n_{12} n_{23}}{n_{13} n_{22}} = \frac{40 * 20}{10 * 40} = 2.000;$$

$$\xi_{21}^{L} = \frac{n_{21} n_{32}}{n_{22} n_{31}} = \frac{40 * 40}{40 * 30} = 1.333;$$

$$\xi_{22}^{L} = \frac{n_{22} n_{33}}{n_{23} n_{32}} = \frac{40 * 30}{20 * 40} = 1.500.$$

2. Globālā izredžu attiecība:

$$\xi_{ij}^{G} = \frac{\left(\sum_{a<i} \sum_{b \le j} n_{ab}\right)\left(\sum_{a>i} \sum_{b>j} n_{ab}\right)}{\left(\sum_{a \le i} \sum_{b>j} n_{ab}\right)\left(\sum_{a>i} \sum_{b \le j} n_{ab}\right)}. \qquad \text{(P1.2.4)}$$

Aprēķina globālo izredžu attiecību P1.2.4. tabulas datiem, izmantojot vienādojumu (P1.2.4):

$$\xi_{11}^{G}=\frac{n_{11}\left(n_{22}+n_{23}+n_{32}+n_{33}\right)}{\left(n_{12}+n_{13}\right)\left(n_{21}+n_{31}\right)}=\frac{50*\left(40+20+40+30\right)}{\left(40+10\right)*\left(40+30\right)}=\frac{50*130}{50*70}=\frac{6500}{3500}=1.857;$$

$$\xi_{12}^{G}=\frac{\left(n_{11}+n_{12}\right)\left(n_{23}+n_{33}\right)}{n_{13}\left(n_{21}+n_{22}+n_{31}+n_{32}\right)}=\frac{\left(50+40\right)*\left(20+30\right)}{10*\left(40+40+30+10\right)}=\frac{90*50}{10*120}=\frac{3750}{1200}=3.750;$$

$$\xi_{21}^{G}=\frac{\left(n_{11}+n_{21}\right)\left(n_{32}+n_{33}\right)}{\left(n_{12}+n_{13}+n_{22}+n_{23}\right)\left(n_{31}+n_{32}\right)}=\frac{\left(50+40\right)*\left(40+30\right)}{\left(40+10+40+20\right)*\left(30+40\right)}=\frac{90*70}{110*70}=\frac{6300}{7700}=0.818;$$

$$\xi_{22}^{G}=\frac{\left(n_{11}+n_{12}+n_{21}+n_{22}\right)n_{33}}{\left(n_{13}+n_{23}\right)\left(n_{31}+n_{32}\right)}=\frac{\left(50+40+40+40\right)*30}{\left(10+20\right)*\left(30+40\right)}=\frac{170*30}{30*70}=\frac{5100}{2100}=2.429.$$

3 . Kumulatīvā izredžu attiecība:

$$\xi_{ij}^{C}=\frac{\left(\sum_{b\le j}n_{ib}\right)\left(\sum_{b>j}n_{i+1,b}\right)}{\left(\sum_{b>j}n_{ib}\right)\left(\sum_{b\le j}n_{i+1,b}\right)}. \qquad \text{(P1.2.5)}$$

Aprēķina kumulatīvo izredžu attiecību P1.2.4. tabulas datiem, izmantojot vienādojumu (P1.2.5):

$$\xi_{11}^{C}=\frac{n_{11}\left(n_{22}+n_{23}\right)}{\left(n_{12}+n_{13}\right)n_{21}}=\frac{50*\left(40+20\right)}{\left(40+10\right)*40}=\frac{50*60}{50*40}=\frac{3000}{2000}=1.500;$$

$$\xi_{12}^{C}=\frac{\left(n_{11}+n_{12}\right)n_{23}}{n_{13}\left(n_{21}+n_{22}\right)}=\frac{\left(50+40\right)*20}{10*\left(40+40\right)}=\frac{90*20}{10*80}=\frac{1800}{800}=2.250;$$

$$\xi_{21}^{C}=\frac{n_{21}\left(n_{31}+n_{32}\right)}{\left(n_{22}+n_{23}\right)n_{31}}=\frac{40*\left(30+40\right)}{\left(40+20\right)*30}=\frac{40*70}{60*30}=\frac{2400}{1400}=1.715.$$

Kumulatīvo izredžu attiecību var arī aprēķināt, izmantojot kumulatīvās nosacīto varbūtību vērtības $F_{j/i}$:

$$\xi_{ij}^{C}=\frac{F_{j/i}/\left(1-F_{j/i}\right)}{F_{j/i+1}/\left(1-F_{j/i+1}\right)}. \qquad \text{(P1.2.6)}$$

Aprēķina kumulatīvo izredžu attiecību P1.2.6. tabulas datiem, izmantojot vienādojumu (P1.2.6):

$$\xi_{11}^{C}=\frac{F_{1/1}/\left(1-F_{1/1}\right)}{F_{1/2}/\left(1-f_{1/2}\right)}=\frac{0.500/\left(1-0.500\right)}{0.400/\left(1-0.400\right)}=\frac{1.000}{0.667}=1.500;$$

$$\xi_{12}^{C}=\frac{F_{2/1}/\left(1-F_{2/1}\right)}{F_{2/2}/\left(1-f_{2/2}\right)}=\frac{0.900/\left(1-0.900\right)}{0.800/\left(1-0.800\right)}=\frac{9.000}{4.000}=2.250;$$

$$\xi_{21}^{C}=\frac{F_{1/2}/\left(1-F_{1/2}\right)}{F_{1/3}/\left(1-f_{1/3}\right)}=\frac{0.400/\left(1-0.400\right)}{0.300/\left(1-0.300\right)}=\frac{0.667}{0.429}=1.556;$$

$$\xi_{22}^{C}=\frac{F_{2/2}/\left(1-F_{2/2}\right)}{F_{2/3}/\left(1-f_{2/3}\right)}=\frac{0.800/\left(1-0.800\right)}{0.700/\left(1-0.700\right)}=\frac{4.000}{2.233}=1.715.$$

Kumulatīvo izredžu attiecību vērtības, kas aprēķinātas, izmantojot vienādojumus (P1.2.5; P1.2.6), sakrīt.

Ir jāievieš kumulatīvais regresijas logitmodelis ar proporcionālajām izredzēm.

Pieņem, ka ir rezultējošais mainīgais $\mathbf{Y}$, kas satur c ordinālās kategorijas $y_1, y_2, ..., y_c$. Turpmāk apzīmējumā y_j simbols y tiks izlaists un aprēķinu izteiksmēs vienkārši izmantots indekss

j. Tātad tā vietā, lai apzīmētu nosacīto varbūtību $p\left(Y \le y_j\right)$, vienkārši tiks izmantots vienkāršots apzīmējums $p\left(Y \le j\right)$. Tas vienkāršos aprēķinu izteiksmes un neradīs nekādus pārpratumus.

Sekojot [Agresti A., 2010], katras ordinālās kategorijas nosacītās varbūtības $p_{j/i}$ apzīmē ar π_j. Kategorijas j kumulatīvā nosacītā varbūtība tiek aprēķināta kā visu iepriekšējo kategoriju nosacīto varbūtību summa:

$$p\left(Y \le j\right) = \pi_1 + ... + \pi_j, \ j = 1,...,c.$$

Šīs varbūtību vērtības veido sakārtotu secību:

$$\pi_1 \le \pi_2 \le ... \le \pi_c.$$

Ir jāievieš logitu (*logits*) jēdziens uzkrātajām nosacītajām varbūtībām:

$$\log it\left(p\left(Y \le j\right)\right) = \ln\left(\frac{p\left(Y \le j\right)}{1 - p\left(Y \le j\right)}\right) = \ln\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_c}, \ j = 1,..,c-1. \qquad \text{(P1.2.7)}$$

Vērtības $\log it\left(p\left(Y \le j\right)\right)$ sauc par *kumulatīviem logitiem*. Piemēram, priekš $c = 3$

$\log it\left(p\left(Y \le 1\right)\right) = \ln\frac{\pi_1}{\pi_2 + \pi_3}$; $logit\left(p\left(Y \le 2\right)\right) = \ln\frac{\pi_1}{\pi_2 + \pi_3}$.

Apskatāmā kumulatīvā logita modeļa vispārējā ideja ar proporcionālajām izredzēm ir šāda. Katrai mainīgā j kategorijai var izmantot šādu regresijas vienādojumu:

$$\log it\left(p\left(Y \le j\right)\right) = \alpha_j + \beta x, \ j = 1,...,c-1. \qquad \text{(P1.2.8)}$$

Vienādojumā (P1.2.8) regresijas koeficientu β vērtības ir vienādas visām kategorijām y_j. Tā ir apskatāmā regresijas logitmodeļa raksturīga iezīme. Tomēr konstantu α_j nozīme visām kategorijām y_j ir atšķirīga. Tas nozīmē, ka, veidojot regresijas līknes visām kategorijām y_j, šīm līknēm būs vienāda forma, bet tās tiks nobīdītas viena pret otru proporcionāli α_j vērtībām.

Var iegūt šādu izredžu attiecību j mainīgā kategorijai diviem mainīgajiem x_1 un x_2:

$$\frac{p\left(Y \le j / x_2\right) / p\left(Y > j / x_2\right)}{p\left(Y \le j / x_1\right) / p\left(Y > j / x_1\right)}.$$

Šīs attiecības naturālais logaritms ir starpība starp kumulatīvo logitu vērtībām x_1 un x_2. Šī starpība ir vienāda ar $\beta\left(x_2 - x_1\right)$ un proporcionāla attālumam starp vērtībām x_1 un x_2.

Šim izredžu attiecību logaritmam tas pats proporcionalitātes koeficients β attiecas uz visām uzkrātajām varbūtībām. No šejienes arī ir cēlies nosaukums *uzkrāto izredžu modelis*.

Pieņem, ka, izmantojot iepriekš uzdoto modeli, ir aprēķinātas visu uzkrāto varbūtību $p\left(Y \le j\right)$ vērtības. Tad visu kategoriju y_j *neuzkrātās varbūtības* vērtību aprēķina kā:

$$p\left(y_j\right) = p\left(Y \le j\right) - p\left(Y \le j-1\right). \qquad \text{(P1.2.9)}$$

Izmantojot vienādojumu (P1.2.9), var aprēķināt visu interesējošo kategoriju varbūtības.

Izmantojot konstantes α_j un koeficienta β vērtību vienādojumā (P1.2.8), uzkrāto varbūtību vērtības var aprēķināt ar vienādojumu:

$$p\left(Y \le j\right) = \frac{e^{\alpha_j + \beta x}}{1 + e^{\alpha_j + \beta x}}. \qquad \text{(P1.2.10)}$$

Faktoriālais mainīgais kumulatīvajā logitmodelī var būt nepārtraukts vai kategorisks. (Iepriekš sniegtajā piemērā par domes darba kvalitātes novērtēšanu mainīgais ir ordināls mainīgais, kura kategorijas ir iedzīvotāju vecuma gradācijas).

Nepārtrauktu faktoriālā mainīgā vērtību gadījumā formulā (P1.2.10) tiek iekļauta x vērtība un veikti attiecīgie aprēķini. Ja faktoriālais mainīgais ir kategoriskais mainīgais, tad x vērtības vienkārši tiek aizstātas ar kategoriju numuriem un tiek veikti attiecīgie aprēķini.

Ja viena faktoriālā mainīgā $\mathbf{X}$ vietā tiek izmantota mainīgo $\mathbf{X}_r$ kopa r, $r = 1,...,c$, tad vienādojums (P1.2.10) iegūst šādu formu:

$$p(Y \le j) = \frac{e^{\alpha_j + \beta_1 x_1 + ... + \beta_r x_r}}{1 + e^{\alpha_j + \beta_1 x + ... + \beta_r x_r}}. \qquad \text{(P1.2.11)}$$

No skaitļošanas viedokļa visgrūtākais ir konstantes α_j un regresijas koeficienta β (vai regresijas koeficientu β_r izteiksmē (P1.2.11) aprēķināšanas uzdevums. To var atrisināt tikai ar piemērotu programmatūras rīku palīdzību.

P1.3. Intervālu skalas

Intervālu skalas ir kvantitatīvās vai skaitliskās skalas. Šo skalu raksturīga iezīme ir spēja noteikt, par *cik* viena vērtība ir lielāka vai mazāka par citu vērtību. Piemēram, veicot divus temperatūras mērījumus, 10 ^{0}C un 15 ^{0}C, var droši apgalvot, ka otrā vērtība ir par 5 ^{0}C lielāka nekā pirmā vai pirmā vērtība ir par 5 ^{0}C mazāka nekā otrā vērtība. Vēl viena intervālu skalu iezīme ir tāda, ka vienādiem intervāliem starp vērtībām ir tāda pati nozīme. Piemēram, temperatūras diapazonam 10 ^{0}C–20 ^{0}C ir tāda pati nozīme kā intervālam ir 35 ^{0}C–45 ^{0}C jeb intervālam 96 ^{0}C–106 ^{0}C.

Vēl viens mērījumu piemērs intervālu skalā ir indivīdu vecuma mērīšana. Lai gan šos mērījumus var interpretēt kā attiecību skalas mērījumus, indivīdu vecuma salīdzināšana ir tipisks intervālu skalas mērījumu piemērs. Dažreiz Likerta skala tiek interpretēta kā intervālu skala. Bet tā ir nekorekta interpretācija, jo nevar skaitliski izteikt atšķirības starp skalas kategorijām. Tāpēc Likerta skala ir tipiska ordināla skala. Tas pats attiecas uz IQ mērīšanas skalu. Tā ir tipiska ordināla skala, nevis intervālu skala.

Intervālu skalām nav noteikta nulles punkta. Piemēram, Celsija skalā ūdens sasalšanas punkts tiek uzskatīts par nulli, bet Fārenheita skalā ūdens sasalšanas punkts atbilst 32 ^{0}F. Vērtības intervālu skalā var saskaitīt un atņemt, bet ne reizināt un dalīt. Par parametriem intervālu skalā var izmantot vidējo vērtību, mediānu un modu.

Piemērs. Tika izmērītas šādas temperatūru vērtības:

$X = \{12,16,14,17,14,18,14,10,19,16\}$.

Vidējo vērtību aprēķina šādi:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad \text{(P1.3.1)}$$

Dotajai datu kopai:

$$\overline{x} = \frac{12+16+14+17+14+18+14+10+19+16}{10} = \frac{150}{10} = 15.$$

Lai aprēķinātu mediānas vērtību, sākotnējās vērtības sakārto nedilstošā secībā:

$X = \{10,12,14,14,14,16,16,17,18,19\}$.

Tā kā vērtību skaits ir pāra ($n = 10$), tad par mediānas vērtību pieņem skaitli, kas atrodas pozīcijā $n/2 = 10/2 = 5$. Šī vērtība ir 14, tāpēc $med(X) = 14$.

Par modu pieņem visbiežāk izmantoto vērtību šajā kopā. Šajā piemērā visizplatītākā vērtība ir 14, tāpēc $mod(X) = 14$.

Visbiežāk izmantotie diapazona parametri datiem, ko mēra intervālu skalā, ir dispersija un standarta novirze.

Vērtību kopas dispersija tiek aprēķināta, izmantojot vienādojumu:

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}, \qquad \text{(P1.3.2)}$$

kur $\bar{x}$ – vidējā vērtība, kas aprēķināta, izmantojot vienādojumu (P1.3.1).

Aprēķina temperatūras vērtību kopas dispersiju. Sākotnējie dati un aprēķinu rezultāti ir parādīti P1.3.1. tabulā.

P1.3.1. tabula

Temperatūras vērtību kopas dispersijas aprēķināšanai

x_i	12	16	14	17	14	18	14	10	19	16	
$x_i - \bar{x}$	-3	1	-1	2	-1	3	-1	-5	4	1	
$\left(x_i - \bar{x}\right)^2$	9	1	1	4	1	9	1	25	16	1	$\sum = 67$

No vienādojuma (P1.3.1) iegūst:

$$S^2 = \frac{67}{9} = 7.444 .$$

Standartnovirzes vērtību aprēķina, izmantojot vienādojumu:

$$s = \sqrt{S^2} . \qquad \text{(P1.3.3)}$$

Šajā piemērā:

$s = \sqrt{7.444} = 2.728$.

Kā definēt lineāru regresiju starp divām vērtību kopām, no kurām katra tiek mērīta intervālu skalā? Lineārās regresijas modeli var izteikt šādā lineārā formā:

$$E\left(Y / x\right) = \beta_0 + \beta_1 x . \qquad \text{(P1.3.4)}$$

Vienādojumā (P1.3.4) β_0 ir konstante jeb brīvais loceklis, β_1 – slīpuma koeficients. Skaitliski β_1 nosaka, par cik vienībām izmainīsies Y, ja X tiks palielināts par vienu vienību.

Līknei, kas atbilst vienādojumam (P1.3.4), jābūt tādai, lai sākotnējo vērtību novirzes no šīs līknes būtu minimālas. To var panākt, pielāgojot koeficientus β_0 un β_1.

Pastāv divas konceptuālas pieejas lineārās regresijas risināšanai: maksimālās līdzības metode un mazāko kvadrātu metode. Nākamās aprēķinu izteiksmes ir balstītas uz mazāko kvadrātu metodes principa.

Vienādojumu (P1.3.4) uzdod uzskatāmākā formā. Regresijas koeficienti b_0 un b_1 tiek aprēķināti, pamatojoties uz sākotnējiem datiem:

$$y = b_0 + b_1 x . \qquad \text{(P1.3.5)}$$

Koeficienta b_1 vērtības tiek aprēķinātas, izmantojot vienādojumu:

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} . \qquad \text{(P1.3.6)}$$

Šajā vienādojumā $\bar{x}$, $\bar{y}$ ir vērtību kopu X un Y vidējās vērtības.

Koeficienta b_0 vērtību aprēķina, izmantojot vienādojumu:

$$b_0 = \bar{y} - b_1 \bar{x} . \qquad \text{(P1.3.7)}$$

Vienkāršs ilustratīvs piemērs. Karstā ūdens temperatūra mājās ar centrālo apkuri ir atkarīga no ārējā gaisa temperatūras. Temperatūras mērījumu dati ir parādīti P1.3.2. tabulā.

P1.3.2. tabula

Āra temperatūras un ūdens temperatūras mērījumu dati

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
	-10	-15	-7	-20	-18	-12	-10	-9	-16	-10
Y	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
	72	74	70	80	78	74	71	68	77	74

Ir nepieciešams izveidot regresijas modeli, kas parāda sakarību starp vērtību kopām X un Y. Aprēķina vidējās temperatūras.

$$\overline{x} = \frac{(-10)+(-15)+(-7)+(-20)+(-18)+(-12)+(-10)+(-9)+(-16)+(-10)}{10} = \frac{-127}{10} \approx -13.$$

$$\overline{y} = \frac{72+74+70+80+78+74+71+68+77+74}{10} = \frac{738}{10} \approx 74.$$

(Aprēķinu rezultātus noapaļo, lai vienkāršotu turpmākos aprēķinus).

Veic starpaprēķinus, lai noteiktu regresijas koeficientus b_0 un b_1. Aprēķinu rezultāti ir apkopoti P1.3.3. tabulā.

P1.3.3. tabula

Starpaprēķinu rezultāti, lai noteiktu nepieciešamo regresijas atkarība saskaņā ar P1.3.2. tabulu

Mērījumi	1	2	3	4	5	6	7	8	9	10	$\sum$
$x_i - \overline{x}$	3	-2	6	-7	-5	1	3	5	-3	3	
$y_i - \overline{y}$	-1	0	-4	6	4	0	-3	-6	3	0	
$(x_i - \overline{x}) * * (y_i - \overline{y})$	-3	0	-24	-42	-20	0	-9	- 30	-9	0	-137
$(x_i - \overline{x})^2$	9	4	36	49	25	1	9	25	9	9	176

Izmantojot starpaprēķinu rezultātus un vienādojumu (P1.3.6), aprēķina koeficienta b_1 vērtību.

$$b_1 = \frac{-137}{176} = -0.778.$$

Izmantojot vienādojumu (P1.3.7), aprēķina koeficienta b_0 vērtību.

$$b_0 = 74 - (-0.778) * (-13) = 74 - 10.150 = 63.850.$$

Galīgais regresijas vienādojums izskatās šādi:

$$y = 63.850 - 0.778x.$$

Lai izveidotu regresijas līkni, aprēķina y vērtības divām uzdotajām x vērtībām.

(Faktiski vienai no vērtībām vajadzētu būt vērtībai $x = 0$, pie kuras $y = b_0 = 63.850$. Taču šī vērtība ir problemātiska grafiskajam attēlojumam (sk. P1.3.1. attēlu). Tāpēc izmanto tādas x vērtības, kuras var attēlot grafiski).

$$x = -10$$

$$y = 63.850 - 0.778 * (-10) = 63.850 + 7.780 = 71.630.$$

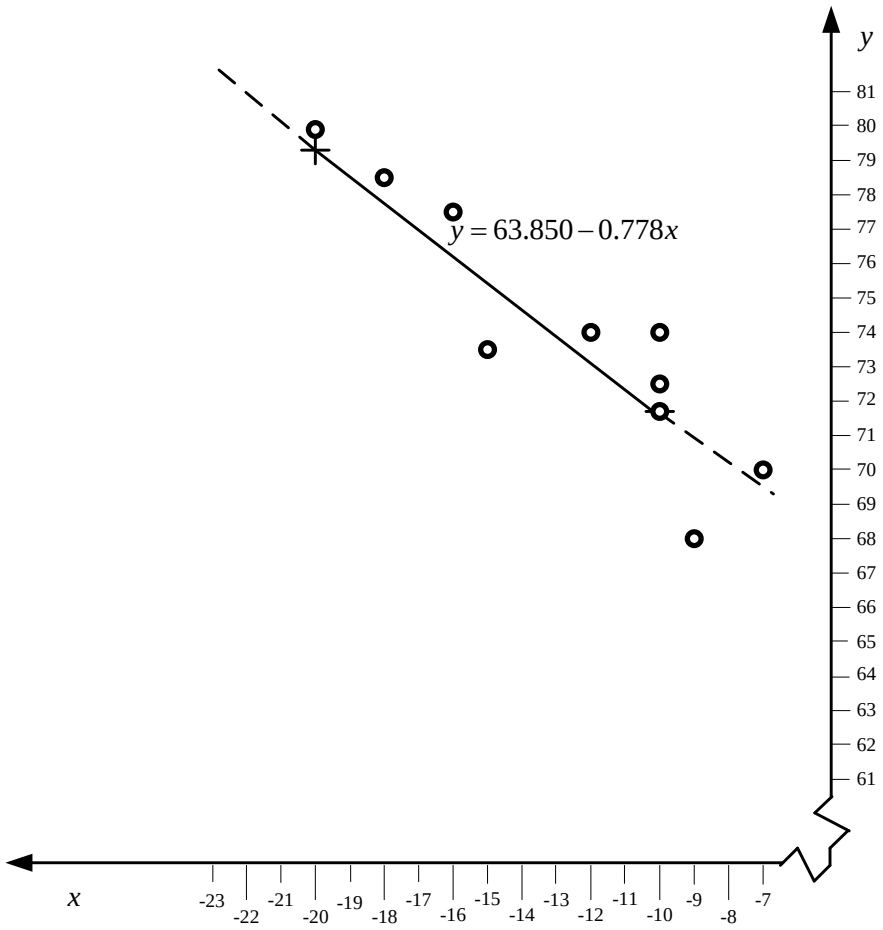
$$x = -20$$

$$y = 63.850 - 0.778 * (-20) = 63.850 + 15.560 = 79.410.$$

P1.3.1. attēlā parādīta konstruētā regresijas līkne. Punkti, kas atbilst vērtībām $x=-10$, $y=71.630$ un $x=-20$, $y=79.410$, šajā attēlā ir apzīmēti ar krustiņiem.

Šis grafiks parāda izteikti negatīvu korelāciju starp temperatūras vērtībām. Palielinoties ārējās gaisa temperatūras negatīvajām vērtībām, paaugstinās ūdens temperatūras vērtība.

Reālās centrālās apkures sistēmās, mainoties āra gaisa temperatūrai, istabas ūdens temperatūra svārstās daudz plašākā diapazonā nekā iepriekš dotajā piemērā. Bet šis piemērs ir sniegts tikai ar mērķi ilustrēt lineārās regresijas attiecības starp diviem mainīgajiem konstruēšanas principus. Tāpēc piemērā tiek izmantotas mākslīgas temperatūras vērtības.



P1.3.1. attēls. **Regresijas grafiks starp ūdens temperatūru un āra temperatūru**

P1.4. Attiecību skalas

Šīs skalas ieņem augstāko vietu vispārējā mērījumu skalu hierarhijā. Attiecību skalas ir kvantitatīvās (skaitliskās) skalas ar definētu nulles punktu. Ar vērtībām attiecību skalā var veikt saskaitīšanu, atņemšanu, reizināšanu un dalīšanu. Ņemot divas vērtības šajā skalā, ir iespējams noteikt *par cik* viena vērtība ir pārāka par otru.

Ar attiecību skalas vērtībām var noteikt vidējo vērtību, mediānu un modu. Tāpat var izskaitļot dispersiju un standarta novirzi, var izveidot lineāru regresiju, kas attiecas uz abām vērtību kopām.

Šeit netiek sniegtas aprēķinu izteiksmes un ilustratīvi piemēri, jo ir spēkā viss, kas tika parādīts iepriekšējā sadaļā par vērtībām, kas mērītas intervālu skalās.

P2. METRIKAS DATU TELPĀ

P2.1. Kas ir metrika?

Šajā pielikumā esošais materiāls ir balstīts uz materiāliem, kas sniegti darbā [Užga-Rebrovs O.I., 2021].

Datu analīzē bieži vien ir jānovērtē, cik tuvu vai tālu atrodas datu elementi viens no otra. Kā var novērtēt interesējošos attālumus starp punktiem vai vektoriem sākotnējā datu kopā? Divus datu elementus (objektus, subjektus vai citas entītijas) apzīmē ar a un b. Attālumu starp šiem elementiem apzīmē ar $d(a,b)$.

Tagad ir jāformulē prasības, kurām jāatbilst jebkuram pareizam attāluma novērtējumam datu telpā:

1) *negativitāte*: jebkurš rezultāts $d(a,b)$ nav negatīvs:

$$d(a,b) \geq 0;$$

2) *identitāte*: $d(a,b) = 0$ tad un tikai tad, ja a un b ir vienādi;

3) *simetrija*: attālumam starp a un b jābūt vienādam ar attālumu starp b un a:

$$d(a,b) = d(b,a);$$

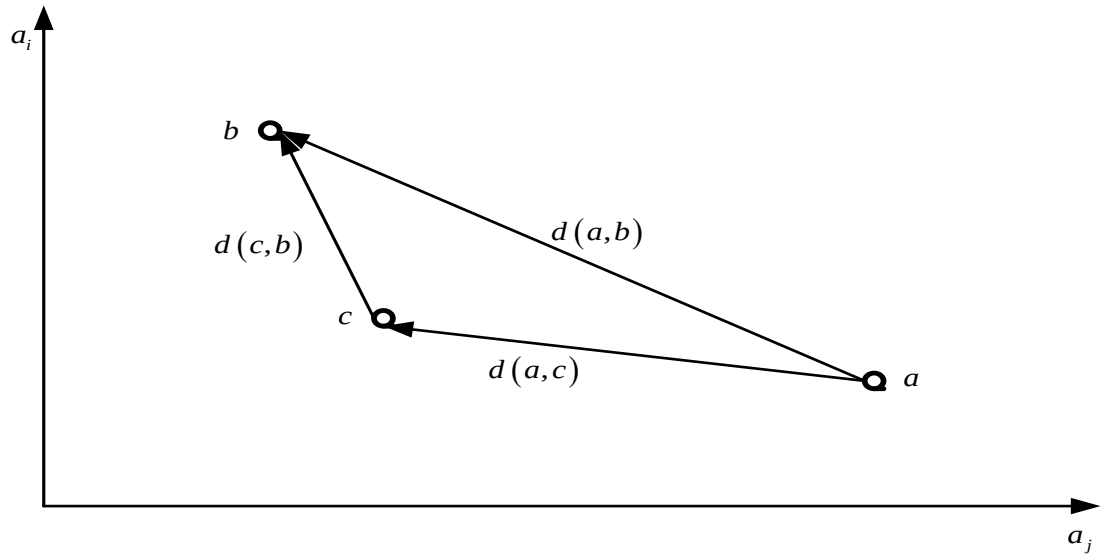
4) *trijstūra nevienādība:*

$$d(a,b) \leq d(a,c) + d(c,b).$$

Pēdējā prasība ir grafiski interpretēta P2.1.1. attēlā divu atribūtu vērtību telpā a_j, a_k.

Jebkuru attālumu mērīšanas metodi datu telpā, kas atbilst 1.–4. prasībai, sauc par metriku jeb distanci.

Attālumu mērīšanas metodi, kas neatbilst dažām no šīm prasībām, sauc par semimetriku. Metodi, kas neatbilst 2. prasībai, sauc par pseidometriku, un metodi, kas neatbilst 3. prasībai, sauc par kvazimetriku.



P2.1.1. attēls. **Trīsstūra noteikuma grafiskā interpretācija divu atribūtu vērtību telpā**

Nākamajās sadaļās tiks aplūkotas izplatītās metodes attālumu (metrikas) mērīšanai datu telpās.

P2.2. Eiklīda attālums

Pieņem, ka sākotnējie dati ir uzdoti tabulas veidā. Katrā tabulas rindā tiek parādīts viens datu elements un tam atbilstošais atribūtu vērtību vektors. Katrā kolonnā tiek parādīta atbilstošo atribūtu vērtību kopa visiem datu vienumiem.

Jāapskata divi datu elementi o_i, o_k. Šos elementus raksturo atribūtu vērtību vektori:

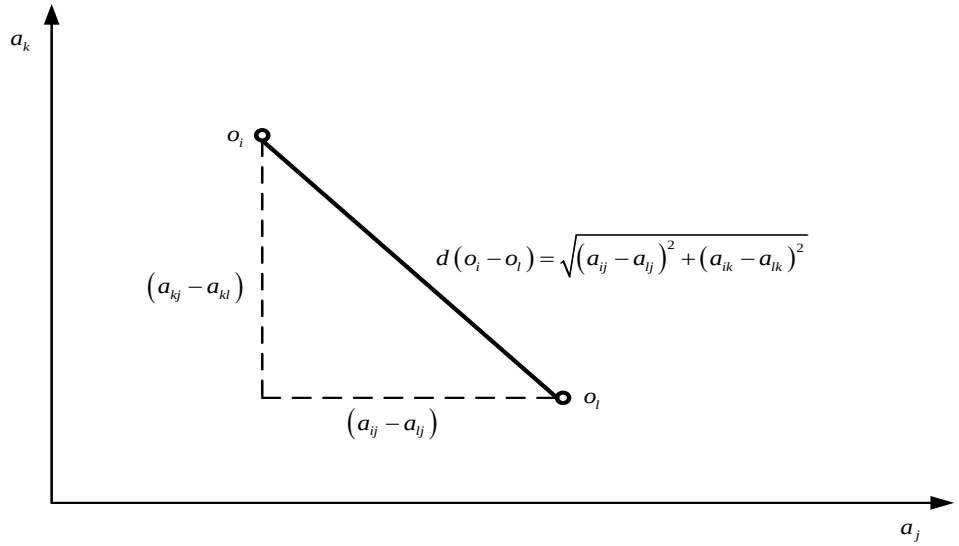
$o_i : \left(a_{i1}, a_{i2}, ..., a_{ij}, ..., a_{in-1}, a_{in}\right)$;

$o_l : \left(a_{l1}, a_{l2}, ..., a_{lj}, ..., a_{ln-1}, a_{ln}\right)$.

Eiklīda attālumu starp šiem elementiem aprēķina pēc vienādojuma:

$$d\left(o_i, o_l\right) = \sqrt{\sum_{j=1}^{n}\left(a_{ij} - a_{lj}\right)^2}\ . \qquad \text{(P2.2.1)}$$

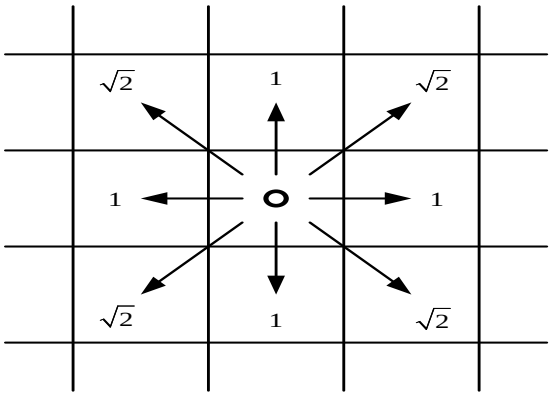
P2.2.1. attēlā parādīta grafiska interpretācija Eiklīda attālumam starp diviem datu elementiem divu atribūtu vērtību telpā.



P2.2.1. attēls. **Eiklīda attāluma grafiskā interpretācija divu atribūtu telpā**

Acīmredzams, ka Eiklīda attālums ir mazākais attālums starp elementiem o_i, o_k.

Lai vizualizētu aprēķinātās izteiksmes attālumiem datu telpā, dažkārt tiek izmantota šaha galdiņa interpretācija. Šī Eiklīda attāluma vizualizācija ir parādīta P2.2.2. attēlā.



P2.2.2. attēls. **Eiklīda attāluma vizualizācija ar šaha galdiņu**

Tiek pieņemts, ka attālums starp blakus esošajām šūnām horizontāli un vertikāli ir vienāds ar 1. Šie attālumi ir parādīti P2.2.2. attēlā ar skaitli 1. Lai pārietu no centrālās šūnas uz jebkuru no diagonālajām šūnām, jāveic viens solis horizontāli un viens solis vertikāli. Saskaņā ar vienādojumu (P2.2.1) attālums starp centrālo šūnu un jebkuru no diagonālajām šūnām ir vienāds ar $\sqrt{2}$.

P2.2.1. piemērs – Eiklīda attālumu aprēķināšanai starp elementu (objektu) pāriem. Sākotnējā datu kopa ir parādīta P2.2.1. tabulā. Ir nepieciešams aprēķināt Eiklīda attālumus starp visiem objektu pāriem.

P2.2.1. tabula

Sākotnējie dati P2.2.1. piemērā

	a_{i1}	a_{i2}	a_{i3}	a_{i4}
o_1	7	3	2	10
o_2	9	2	5	6
o_3	12	5	8	8
o_4	5	4	6	9

Risinājums. Izmantojot vienādojumu (P2.2.1), aprēķina attālumus no objekta o_1 līdz citiem objektiem.

$$d(o_1-o_2)=\sqrt{(7-9)^2+(3-2)^2+(2-5)^2+(10-6)^2}=\sqrt{(-2)^2+1^2+(-3)^2+4^2}=$$
$$=\sqrt{4+1+9+16}=\sqrt{30}=5.48;$$
$$d(o_1-o_3)=\sqrt{(7-12)^2+(3-5)^2+(2-8)^2+(10-8)^2}=\sqrt{(-5)^2+(-2)^2+(-6)^2+2^2}=$$
$$=\sqrt{25+4+36+4}=\sqrt{69}=8.31;$$
$$d(o_1-o_4)=\sqrt{(7-5)^2+(3-4)^2+(2-6)^2+(10-9)^2}=\sqrt{2^2+(-1)^2+(-4)^2+1^2}=$$
$$=\sqrt{4+1+16+1}=\sqrt{22}=4.69.$$

Nav nepieciešams aprēķināt attālumus $d(o_2-o_1)$, $d(o_3-o_1)$, $d(o_4-o_1)$, jo, pamatojoties uz simetrijas prasību $d(o_2-o_1)=d(o_1-o_2)$, $d(o_3-o_1)=d(o_1-o_3)$, $d(o_4-o_1)=d(o_1-o_4)$.

Aprēķina attālumus no objekta o_2 līdz objektiem o_3, o_4

$$d(o_2-o_3)=\sqrt{(9-12)^2+(2-5)^2+(5-8)^2+(6-8)^2}=\sqrt{(-3)^2+(-3)^2+(-3)^2+(-2)^2}=$$
$$=\sqrt{9+9+9+4}=\sqrt{31}=5.57;$$
$$d(o_2-o_4)=\sqrt{(9-5)^2+(2-4)^2+(5-6)^2+(6-9)^2}=\sqrt{4^2+(-2)^2+(-1)^2+(-3)^2}=$$
$$=\sqrt{16+4+1+9}=\sqrt{30}=5.48.$$

Aprēķina attālumu no objekta o_3 līdz objektam o_4

$$d(o_3-o_4)=\sqrt{(12-5)^2+(5-4)^2+(8-6)^2+(8-9)^2}=\sqrt{7^2+1^2+2^2+(-1)^2}=$$
$$=\sqrt{49+1+4+1}=\sqrt{55}=7.42.$$

Visi nepieciešamie aprēķini ir veikti. Citas vērtības tiek noteiktas, pamatojoties uz simetrijas principu. Aprēķinu rezultāti ir parādīti P2.2.2. tabulā.

P2.2.2. tabula

Attālumu starp objektiem aprēķinu rezultāti P2.2.1. piemērā

	o_1	o_2	o_3	o_4
o_1	0	5.48	8.31	4.69
o_2	5.48	0	5.57	5.48
o_3	8.31	5.57	0	7.42
o_4	4.69	5.48	7.42	0

Šūnās gar tabulas diagonāli ir nulles, jo šīs šūnas parāda objektu attālumus pašas uz sevi. Pamatojoties uz simetrijas principu, vērtības tabulas šūnās ir simetriskas attiecībā pret diagonāli.

P2.2.1. piemērā ir aprēķināti P2.2.1. tabulā uzrādīto objektu attālumi telpā. Tomēr attālumus atribūtu vērtību telpā var aprēķināt, izmantojot to pašu principu. To var parādīt ar piemēru.

P2.2.2. piemērs. Izmantojot sākotnējos datus no P2.2.1. tabulas, ir nepieciešams aprēķināt attālumus starp atribūtiem, izmantojot vienādojumu (P2.2.1).

Risinājums:

$$d(a_1-a_2)=\sqrt{(7-3)^2+(9-2)^2+(12-5)^2+(5-4)^2}=\sqrt{4^2+7^2+7^2+1^2}=$$
$$=\sqrt{16+49+49+1}=\sqrt{115}=10.72;$$
$$d(a_1-a_3)=\sqrt{(7-2)^2+(9-5)^2+(12-8)^2+(5-6)^2}=\sqrt{5^2+4^2+4^2+(-1)^2}=$$
$$=\sqrt{25+16+16+1}=\sqrt{58}=7.62;$$
$$d(a_1-a_4)=\sqrt{(7-10)^2+(9-6)^2+(12-8)^2+(5-9)^2}=\sqrt{(-3)^2+3^2+4^2+(-4)^2}=$$
$$=\sqrt{9+9+16+16}=\sqrt{50}=7.07;$$
$$d(a_2-a_3)=\sqrt{(3-2)^2+(2-5)^2+(5-8)^2+(4-6)^2}=\sqrt{1^2+(-3)^2+(-3)^2+(-2)^2}=$$
$$=\sqrt{1+9+9+4}=\sqrt{23}=4.80;$$
$$d(a_2-a_4)=\sqrt{(3-10)^2+(2-6)^2+(5-8)^2+(4-9)^2}=\sqrt{(-7)^2+(-4)^2+(-3)^2+(-5)^2}=$$
$$=\sqrt{49+16+9+25}=\sqrt{99}=9.95;$$
$$d(a_3-a_4)=\sqrt{(2-10)^2+(5-6)^2+(8-8)^2+(6-9)^2}=\sqrt{(-8)^2+(-1)^2+0+(-3)^2}=$$
$$=\sqrt{64+1+0+9}=\sqrt{74}=8.60.$$

Aprēķinu rezultāti ir parādīti P2.2.3. tabulā.

P2.2.3. tabula

Attālumu starp atribūtiem aprēķinu rezultāti P2.2.2. piemērā

	a_1	a_2	a_3	a_4
a_1	0	10.72	7.62	7.07
a_2	10.72	0	4.80	9.95
a_3	7.62	4.80	0	8.60
a_4	7.07	9.95	8.60	0

Eiklīda attāluma divkāršā mērogošana.

Tās būtība ir šāda. Datu tabulas j kolonnai tiek aprēķināta starpība:

$$\chi_j = \left(a_j^{\max} - a_j^{\min}\right),\ j = 1,...,n,$$

kur $a_j^{\max}$, $a_j^{\min}$ – atribūta a_j maksimālās un minimālās vērtības.

Tiek aprēķināta mērogotā Eiklīda attāluma vērtība pēc vienādojuma:

$$d^*\left(o_i - o_l\right) = \sqrt{\sum_{j=1}^{n}\left(\frac{\left(a_{ij} - a_{lj}\right)^2}{\chi_j}\right)},\ i,l = 1,...,m, \qquad \text{(P2.2.2)}$$

Visbeidzot, divkāršā mēroga Eiklīda attāluma vērtība tiek aprēķināta pēc vienādojuma:

$$d^{**}\left(o_i - o_l\right) = \frac{\sqrt{\sum_{j=1}^{n}\left(\frac{\left(a_{ij} - a_{lj}\right)^2}{\chi_j}\right)}}{\sqrt{n}},\ i,l = 1,...,m,\ i \neq l, \qquad \text{(P2.2.3)}$$

kur n – atribūtu (kolonnu) skaits sākotnējo datu tabulā.

P2.2.3. piemērs. Nepieciešams aprēķināt mērogotos un divkārši mērogotos Eiklīda attālumus starp objektiem no P2.2.1. tabulas.

Risinājums. Izmantojot sākotnējos datus no P2.2.1. tabulas, aprēķina vērtības χ_j, $j = 1,2,3,4$.

$\chi_1 = (12-5) = 7$, $\chi_2 = (5-2) = 3$, $\chi_3 = (8-2) = 6$; $\chi_4 = (10-6) = 4$.

Izmantojot vienādojumu (P2.2.2), aprēķina mērogotos Eiklīda attālumus.

$$d^*\left(o_1 - o_2\right) = \sqrt{\frac{(7-9)^2}{7} + \frac{(3-2)^2}{3} + \frac{(2-5)^2}{6} + \frac{(10-6)^2}{4}} = \sqrt{\frac{4}{7} + \frac{1}{3} + \frac{9}{6} + \frac{16}{4} =}$$

$$= \sqrt{0.571 + 0.333 + 1.500 + 4.000} = \sqrt{6.404} = 2.53;$$

$$d^*\left(o_1 - o_3\right) = \sqrt{\frac{(7-12)^2}{7} + \frac{(3-5)^2}{3} + \frac{(2-8)^2}{6} + \frac{(10-8)^2}{4}} = \sqrt{\frac{25}{7} + \frac{4}{3} + \frac{36}{6} + \frac{4}{4} =}$$

$$= \sqrt{3.571 + 1.333 + 6.000 + 1.000} = \sqrt{11.904} = 3.45;$$

$$d^*\left(o_1 - o_4\right) = \sqrt{\frac{(7-5)^2}{7} + \frac{(3-4)^2}{3} + \frac{(2-6)^2}{6} + \frac{(10-9)^2}{4}} = \sqrt{\frac{4}{7} + \frac{1}{3} + \frac{16}{6} + \frac{1}{4} =}$$

$$= \sqrt{0.571 + 0.333 + 2.6678 + 0.250} = \sqrt{3.821} = 1.95;$$

$$d^*\left(o_2 - o_3\right) = \sqrt{\frac{(9-12)^2}{7} + \frac{(2-5)^2}{3} + \frac{(5-8)^2}{6} + \frac{(6-8)^2}{4}} = \sqrt{\frac{9}{7} + \frac{9}{3} + \frac{9}{6} + \frac{4}{4} =}$$

$$= \sqrt{1.286 + 3.000 + 1.500 + 1.000} = \sqrt{6.786} = 2.60;$$

$$d^*\left(o_2 - o_4\right) = \sqrt{\frac{(9-5)^2}{7} + \frac{(2-4)^2}{3} + \frac{(5-6)^2}{6} + \frac{(6-9)^2}{4}} = \sqrt{\frac{16}{7} + \frac{4}{3} + \frac{1}{6} + \frac{9}{4} =}$$

$$= \sqrt{2.286 + 1.333 + 0.167 + 2.250} = \sqrt{6.036} = 2.46;$$

$$d^*\left(o_3 - o_4\right) = \sqrt{\frac{(12-5)^2}{7} + \frac{(5-4)^2}{3} + \frac{(8-6)^2}{6} + \frac{(8-9)^2}{4}} = \sqrt{\frac{49}{7} + \frac{1}{3} + \frac{4}{6} + \frac{1}{4} =}$$

$$= \sqrt{7.000 + 0.333 + 0.667 + 0.250} = \sqrt{8.250} = 2.87.$$

Iegūtie rezultāti ir apkopoti P2.2.4. tabulā.

P4.2.4. tabula

Mērogoto Eiklīda attālumu vērtības P2.2.3. piemērā

	o_1	o_2	o_3	o_4
o_1	0	2.53	3.45	1.95
o_2	2.53	0	2.60	2.45
o_3	3.45	2.60	0	2.87
o_4	1.95	2.46	2.87	0

Izmantojot vienādojumu (P2.2.3), aprēķina divkārši mērogotos Eiklīda attālumus. Tā kā šajā piemērā atribūtu skaits $n=4$, $\sqrt{n}=\sqrt{4}=2$, lai aprēķinātu nepieciešamos attālumus, pietiek dalīt ar 2 attāluma vērtības no P2.2.4. tabulas. Galīgie rezultāti ir parādīti P2.2.5. tabulā.

P2.2.5. tabula

Divkārši mērogoto Eiklīda attālumu vērtības P2.2.3. piemērā

	o_1	o_2	o_3	o_4
o_1	0	1.27	1.72	0.98
o_2	1.27	0	1.30	1.23
o_3	1.72	1.30	0	1.43
o_4	0.98	1.23	1.43	0

Lai noteiktu attālumus starp atribūtiem, var izmantot relatīvos, mērogotos un divkārši mērogotos Eiklīda attālumus.

Noslēdzot šo sadaļu, jāatzīmē, ka Eiklīda attālums ir Minkovska attāluma īpašs gadījums:

$$d\left(o_i - o_l\right) = \sum_{j=1}^{n} \left(\left(a_{ij} - a_{lj}\right)^p \right)^{\frac{1}{p}}. \qquad \text{(P2.2.5)}$$

Pie $p=2$ Minkovska attālumā tiek iegūta standarta Eiklīda distance.

P2.3. Manhetenas attālums

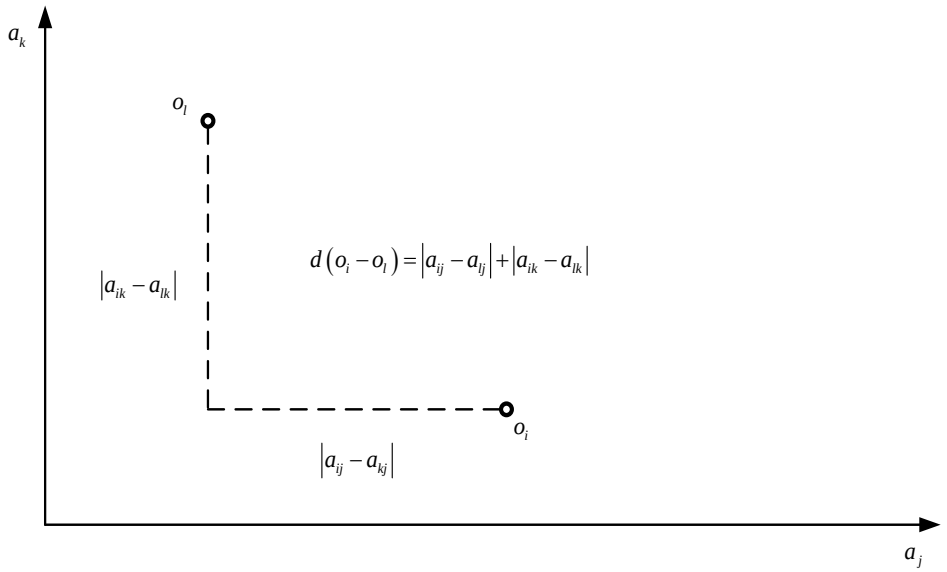
Pieņem, ka dati attēloti tabulas veidā, kurā katra rinda attēlo objektu un tam atbilstošu atribūtu vērtību vektoru $\left(a_{ij}\right)$. Manhetenas attālums starp objektiem a_i, a_k tiek aprēķināts kā

$$d\left(o_i - o_l\right) = \sum_{j=1}^{n} \left| a_{ij} - a_{lj} \right|, \qquad \text{(P2.3.1)}$$

kur n – atribūtu skaits.

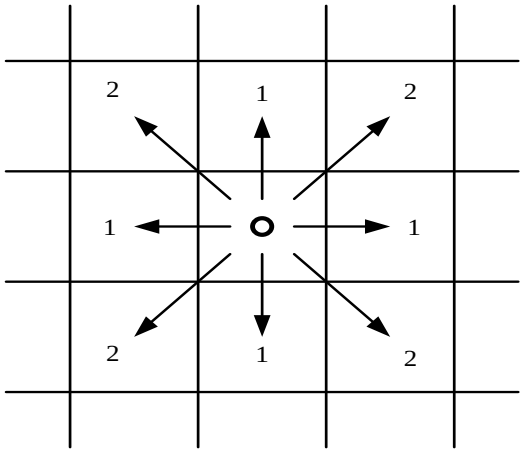
Šo attālumu var interpretēt kā ceļa garumu, kas jāiet pa pilsētas ielām (jo īpaši Ņujorkas Manhetenas rajonā), lai no punkta a_i sasniegtu punktu a_k. Līdz ar to radies šī attāluma nosaukums.

P2.3.1. attēls attēlo Manhetenas attāluma grafisku interpretāciju divu atribūtu telpā.



P2.3.1. attēls. **Manhetenas attāluma grafiskā interpretācija divu atribūtu telpā**

Manhetenas attāluma vizualizācija, izmantojot šaha galdiņu, ir parādīta P2.3.2. attēlā. Lai nokļūtu no centrālās šūnas uz jebkuru blakus esošo šūnu horizontāli vai vertikāli, ir jāveic viens solis. Lai nokļūtu no centrālās šūnas uz jebkuru pa diagonāli blakus esošo šūnu, ir jāveic divas darbības: viena horizontāli un viena vertikāli.



P2.3.2. attēls. **Manhetenas attāluma vizualizācija ar šaha galdiņu**

Tā kā pēc definīcijas soļa garums ir vienāds ar 1, summējot divu soļu garumus saskaņā ar vienādojumu (P2.3.1), iegūst, ka visi šie attālumi ir vienādi ar 2. Tādējādi Manhetenas attālums ir metrika.

P2.3.1. piemērs – Manhetenas attālumu aprēķini objektu telpā. Nepieciešams aprēķināt Manhetenas attālumus starp P2.2.1. tabulā dotajiem objektiem.

Risinājums. Ērtības labad sākotnējie dati no P2.2.1. tabulas ir reproducēti P2.3.1. tabulā.

P2.3.1. tabula

Sākotnējie dati P2.3.1. piemērā

	a_{i1}	a_{i2}	a_{i3}	a_{i4}
o_1	7	3	2	10
o_2	9	2	5	6
o_3	12	5	8	8
o_4	5	4	6	9

Aprēķina nepieciešamos attālumus, izmantojot vienādojumu (P2.3.1).

$d(o_1 - o_2) = |7-9| + |3-2| + |2-5| + |10-6| = 2+1+3+4 = 10$;

$d(o_1 - o_3) = |7-12| + |3-5| + |2-8| + |10-8| = 5+2+6+2 = 15$;

$d(o_1 - o_4) = |7-5| + |3-4| + |2-6| + |10-9| = 2+1+4+1 = 8$;

$d(o_2 - o_3) = |9-12| + |2-5| + |5-8| + |6-8| = 3+3+3+2 = 11$;

$d(o_2 - o_4) = |9-5| + |2-4| + |5-6| + |6-9| = 4+2+1+3 = 10$;

$d(o_3 - o_4) = |12-5| + |5-4| + |8-6| + |8-9| = 7+1+2+1 = 11$.

Aprēķinu rezultāti ir parādīti P2.3.2. tabulā.

P2.3.2. tabula

Attālumu starp objektiem aprēķinu rezultāti P2.3.1. piemērā

	o_1	o_2	o_3	o_4
o_1	0	10	15	8
o_2	10	0	11	10
o_3	15	11	0	11
o_4	8	10	11	0

Manhetenas attālumus var arī aprēķināt, izmantojot normalizētas atribūtu vērtības.

P2.4. Alternatīvi attāluma mēri starp objektiem

Eiklīda attālums un Manhetenas attālums ir populāri attāluma mēri, kurus plaši izmanto datu analīzē. Tomēr papildus tiek izmantoti arī citi attāluma mēri. Šajā sadaļā ir aplūkoti daži no šiem alternatīvajiem mēriem.

1. Breja-Kurtisa koeficients (Bray-Curtis coefficient):

$$d(o_i - o_l) = 1 - \frac{2\sum_{j=1}^{n} \min(a_{ij}, a_{lj})}{\sum_{j=1}^{n} a_{ij} + \sum_{j=1}^{n} a_{lj}},\ i,l = 1,...,m,\ i \neq l, \qquad \text{(P2.4.1)}$$

kur *min* – minimuma operators.

P2.4.1. piemērs. Ir nepieciešams aprēķināt Breja-Kurtisa koeficientu vērtības pēc vienādojuma (P2.4.1) objektu kopai P2.2.1. piemērā (P2.2.1. tabula).

Risinājums. P2.4.1. tabulā tiek atjaunoti dati no P2.2.1. tabulas. Tabulas pēdējā kolonna parāda atribūtu vērtību summas visās tabulas rindās.

P2.4.1. tabula

Sākotnējie dati P2.4.1. piemērā

	a_{i1}	a_{i2}	a_{i3}	a_{i4}	$\sum_{j=1}^{4} a_{ij}$
o_1	7	3	2	10	22
o_2	9	2	5	6	22
o_3	12	5	8	8	33
o_4	5	4	6	9	24

Veic nepieciešamos aprēķinus, izmantojot vienādojumu (P2.4.1).

$$d(o_1 - o_2) = 1 - \frac{2*(7+2+2+6)}{22+22} = 1 - \frac{34}{44} = 1 - 0.77 = 0.23;$$

$$d(o_1 - o_3) = 1 - \frac{2*(7+3+2+8)}{22+33} = 1 - \frac{40}{55} = 1 - 0.73 = 0.27;$$

$$d(o_1 - o_4) = 1 - \frac{2*(5+3+2+9)}{22+24} = 1 - \frac{38}{46} = 1 - 0.83 = 0.17;$$

$$d(o_2 - o_3) = 1 - \frac{2*(9+2+5+6)}{22+33} = 1 - \frac{44}{55} = 1 - 0.80 = 0.20;$$

$$d(o_2 - o_4) = 1 - \frac{2*(5+2+5+6)}{22+24} = 1 - \frac{36}{46} = 1 - 0.78 = 0.22;$$

$$d(o_3 - o_4) = 1 - \frac{2*(5+4+6+8)}{33+24} = 1 - \frac{46}{57} = 1 - 0.81 = 0.19.$$

Aprēķinu rezultāti ir apkopoti P2.4.2. tabulā.

P2.4.2. tabula

Attālumu starp objektiem aprēķinu rezultāti P2.4.1. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.23	0.27	0.17
o_2	0.23	0	0.20	0.22
o_3	0.27	0.20	0	0.19
o_4	0.17	0.22	0.19	0

2. Relatīvais Sorensena attālums (Sorensen Distance).

$$d(o_i - o_l) = 1 - \sum_{j=1}^{n} \min\left(\frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}, \frac{a_{lj}}{\sum_{j=1}^{n} a_{lj}} \right),\ i,l = 1,...,m,\ i \neq l, \qquad \text{(P2.4.2)}$$

P2.4.2. piemērs. Izmantojot sākotnējos datus no P2.4.1. tabulas, ir nepieciešams aprēķināt relatīvos Sorensena attālumus starp objektiem, izmantojot vienādojumu (P2.4.2).

Risinājums. Veic starpposma aprēķinus, pamatojoties uz P2.4.1. tabulas datiem. Šo aprēķinu rezultāti ir parādīti P2.4.3. tabulā.

P2.4.3 tabula

Starpposma aprēķinu rezultāti P2.4.2. piemēram

	$\frac{a_{i1}}{\sum_{j=1}^{4} a_{ji}}$	$\frac{a_{i2}}{\sum_{j=1}^{4} a_{ji}}$	$\frac{a_{i3}}{\sum_{j=1}^{4} a_{ji}}$	$\frac{a_{i4}}{\sum_{j=1}^{4} a_{ji}}$
o_1	0.32	0.14	0.09	0.45
o_2	0.41	0.09	0.23	0.27
o_3	0.36	0.15	0.24	0.24
o_4	0.21	0.17	0.25	0.37

Veic aprēķinus, izmantojot vienādojumu (P2.4.2).

$d(o_1 - o_2) = 1 - (0.32 + 0.09 + 0.09 + 0.27) = 1 - 0.77 = 0.23;$

$d(o_1 - o_3) = 1 - (0.32 + 0.14 + 0.09 + 0.24) = 1 - 0.79 = 0.21;$

$d(o_1 - o_4) = 1 - (0.21 + 0.14 + 0.09 + 0.37) = 1 - 0.81 = 0.19;$

$d(o_2 - o_3) = 1 - (0.36 + 0.09 + 0.23 + 0.24) = 1 - 0.92 = 0.08;$

$d(o_2 - o_4) = 1 - (0.21 + 0.09 + 0.23 + 0.27) = 1 - 0.80 = 0.20;$

$d(o_3 - o_4) = 1 - (0.21 + 0.15 + 0.24 + 0.24) = 1 - 0.84 = 0.16.$

Aprēķinu rezultāti ir parādīti P2.4.4. tabulā.

P2.4.4. tabula

Attālumu starp objektiem aprēķinu rezultāti P2.4.2. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.23	0.21	0.19
o_2	0.23	0	0.08	0.20
o_3	0.21	0.08	0	0.16
o_4	0.19	0.20	0.16	0

Relatīvais Sorensena attālums ir semimetrisks.

Vienādojuma (P2.4.2) variants, izmantojot starpību absolūtās vērtības operatora min vietā, ir vienādojums:

$$d(o_i - o_l) = \frac{1}{2}\sum_{j=1}^{n}\left|\frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}} - \frac{a_{lj}}{\sum_{j=1}^{n} a_{lj}}\right| \ldots\ i,l = 1,\ldots,m,\ \ i \neq l \qquad \text{(P2.4.3)}$$

P2.4.3. piemērs. Izmantojot starpposma aprēķinu rezultātus P2.4.3. tabulā, ir jāaprēķina attālumu vērtības starp objektiem, izmantojot vienādojumu (P2.4.3).

Risinājums. Veic aprēķinus, izmantojot vienādojumu (P2.4.3).

$$d(o_1 - o_2) = \frac{1}{2} * \left(|0.32 - 0.41| + |0.14 - 0.09| + |0.09 - 0.23| + |0.45 - 0.27|\right) =$$

$$= \frac{1}{2} * (0.09 + 0.05 + 0.14 + 0.18) = \frac{1}{2} * 0.46 = 0.23;$$

$$d(o_1 - o_3) = \frac{1}{2} * \left(|0.32 - 0.36| + |0.14 - 0.15| + |0.09 - 0.24| + |0.45 - 0.24|\right) =$$

$$= \frac{1}{2} * (0.04 + 0.01 + 0.15 + 0.21) = \frac{1}{2} * 0.41 = 0.21.$$

Pārējie aprēķini tiek veikti līdzīgi. Aprēķinu rezultāti ir parādīti P2.4.5. tabulā.

P2.4.5. tabula

Attālumu starp objektiem aprēķinu rezultāti P2.4.3. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.23	0.21	0.19
o_2	0.23	0	0.08	0.20
o_3	0.21	0.08	0	0.16
o_4	0.19	0.20	0.16	0

P2.5. χ^2 attālums

Atgriežoties pie Eiklīda attāluma, kas bija aplūkots P2.2. sadaļā, objekta o_i atribūtu vērtību vektoru apzīmē ar $\mathbf{a}_i$ un objekta o_l atribūtu vērtību vektoru – ar $\mathbf{a}_l$. Vispārīgā veidā kvadrātiskā Eiklīda attāluma aprēķināšanai starp atribūtu vērtību vektoriem $(\mathbf{a}_i)$, $(\mathbf{a}_l)$ (objektiem o_i, o_l) var uzdot vektoru reizinājuma formā:

$$d^2 d^2 (o_i, o_l) = d^2 (\mathbf{a}_i, \mathbf{a}_l) = (\mathbf{a}_{ij} - \mathbf{a}_{lj})^T (\mathbf{a}_{ij} - \mathbf{a}_{lj}) = \sum_{j=1}^{n} (a_{ij} - a_{lj})^2,$$

kur T – vektora transponēšanas simbols.

(Jāņem vērā, ka šeit ir runa par kvadrātisko Eiklīda attālumu. Šis attālums tiek izmantots arī dažādos praktiskos pielietojumos. Lai iegūtu standarta Eiklīda attāluma aprēķinu, vienkārši jāņem kvadrātsakne no Eiklīda kvadrātiskā attāluma aprēķina).

Aprēķinot kvadrātisko (un standarta) Eiklīda attālumu, visām atribūtu vērtībām ir vienāds svars. Tomēr var ieviest paplašināto Eiklīda attāluma versiju, kurā atribūtu vērtību vektoriem datu tabulas kolonnās tiek piešķirti dažāds svars. Tādā veidā var iegūt kvadrātiskā attāluma χ_q^2 aprēķinus. Vispārīgā veidā kvadrātisko attālumu χ_q^2 starp atribūtu $\mathbf{a}_i$, $\mathbf{a}_l$ vērtību vektoriem (objektiem o_i, o_i) var definēt kā:

$$\chi_q^2 (o_i - o_l)\, \chi_q^2 (\mathbf{a}_i, \mathbf{a}_l) = (\mathbf{a}_{ij} - \mathbf{a}_{lj})^T \mathbf{W} (\mathbf{a}_{ij} - \mathbf{a}_{lj}). \qquad \text{(P2.5.1)}$$

Vienādojumā (P2.5.1) simbols $\mathbf{W}$ apzīmē atsevišķiem atribūtiem piešķirto svaru matricu.

Šajā sadaļā tiks apskatītas divas svara matricas $\mathbf{W}$ formēšanas iespējas.

Būtiska attāluma χ_q^2 aprēķinu iezīme ir tā, ka tie balstās nevis uz faktiskajām atribūtu vērtībām, bet gan uz šo vērtību profiliem. Lai pārietu no atribūtu vērtībām uz to profilu vērtībām, katrai datu tabulas rindai tiek aprēķināta atribūtu vērtību summa:

$$a_i^* = \sum_{j=1}^{n} a_{ij}\ ,\ i = 1, ..., m.$$

Profila vērtības tiek aprēķinātas kā faktisko atribūtu vērtību attiecība pret to summām a_i^*:

$$p_{ij} = \frac{a_{ij}}{a_i^*} \ldots\ i = 1, ..., m \_\ j = 1, ..., n.$$

Pirmajā variantā svara matricas $\mathbf{W}$ komponentu vērtības tiek aprēķinātas šādi. Katrā datu tabulas kolonnā tiek aprēķināta atribūtu vērtību summa:

$$a_j^* = \sum_{i=1}^{m} a_{ij}\ ,\ \ j = 1,...,n\,.$$

Par svara vērtībām tiek pieņemtas vērtības , kas ir a_{ij}^* apgrieztās vērtības:

$$w_j = \frac{1}{a_j^*}\,.$$

Svara matrica $\mathbf{W}$ ir diagonāla matrica:

$$\mathbf{W} = \begin{bmatrix} \frac{1}{a_1^*}...0...0...0...0 \\ ..... \\ 0...0...\frac{1}{a_j^*}...0...0 \\ ..... \\ 0....0...0...0...\frac{1}{a_n^*} \end{bmatrix}.$$

Formāli pirmajam variantam var formulēt šādu aprēķina vienādojumu:

$$\chi_1^2\left(o_i - o_l\right) = \sum_{j=1}^{n} \frac{1}{a_j^*}\left(\frac{a_{ij}}{a_i^*} - \frac{a_{lj}}{a_l^*}\right)^2 = \sum_{j=1}^{n} \frac{1}{a_j^*}\left(p_{ij} - p_{lj}\right)^2,\ i,l = 1,...,m\,,\ i \neq l\,, \qquad \text{(P2.5.2)}$$

kur p_{ij}, p_{lj} – atribūtu a_{ij}, a_{lj} vērtību profili.

P2.5.1. piemērs ilustrē χ_1^2 attālumu aprēķināšanas metodi. Par pamatu ņem sākotnējos datus no P2.2.1. tabulas. Ir nepieciešams aprēķināt kvadrātisko un standarta attālumu χ_1^2, χ_1 vērtības starp objektiem (to atribūtu vērtību vektoriem), izmantojot pirmo variantu svara matricas veidošanai.

Risinājums. P2.5.1. tabulas kreisajā daļā ir parādīti sākotnējie dati no P2.2.1. tabulas.

P2.5.1. tabula

Sākotnējie dati un provizorisko aprēķinu rezultāti P2.5.1. piemēram

	Sākotnējās atribūtu vērtības					Atribūtu vērtību profili			
	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_i^*	p_{i1}	p_{i2}	p_{i3}	p_{i4}
o_1	7	3	2	10	22	0.318	0.136	0.091	0.455
o_2	9	2	5	6	22	0.409	0.091	0.228	0.273
o_3	12	5	8	8	33	0.364	0.152	0.242	0.242
o_4	5	4	6	9	24	0.208	0.167	0.250	0.375
a_j^*	33	14	21	33					
$w_1 = \frac{1}{a_j^*}$	0.030	0.071	0.048	0.030					

Tabulas labajā daļā ir redzamas atribūtu p_{ij} profila vērtības, kas tiek aprēķinātas, dalot atribūtu a_{ij} vērtības ar atbilstošajām a_i^* vērtībām. Priekšpēdējā rindā dotas a_j^* vērtības, kas ir atribūtu vērtību summas attiecīgajās tabulas kolonnās. Pēdējā rindā ir norādīts svars $w_j = \frac{1}{a_j^*}$.

Veic aprēķinus, izmantojot vienādojumu (P2.5.2).

$$\chi_1^2(o_1 - o_2) = 0.030 * (0.318 - 0.409)^2 + 0.071 * (0.136 - 0.091)^2 + 0.048 * (0.091 - 0.228)^2 +$$
$$+0.030 * (0.455 - 0.273) =$$
$$= 0.030 * 0.00828 + 0.071 * 0.00202 + 0.048 * 0.01877 + 0.030 * 0.03312 =$$
$$= 0.00025 + 0.00014 + 0.00090 + 0.00010 \approx 0.0014;$$
$$\chi_1^2(o_1 - o_3) = 0.030 * (0.318 - 0.364)^2 + 0.071 * (0.136 - 0.152)^2 + 0.048 * (0.091 - 0.242)^2 +$$
$$+0.030 * (0.455 - 0.242) =$$
$$= 0.030 * 0.00212 + 0.071 * 0.00026 + 0.048 * 0.02280 + 0.030 * 0.04537 =$$
$$= 0.00006 + 0.00002 + 0.00109 + 0.00136 \approx 0.0025.$$

Pārējie aprēķini tiek veikti tādā pašā veidā. Aprēķinu rezultāti ir parādīti P2.5.2. tabulā.

P2.5.2. tabula

Kvadrātisko attālumu χ_1^2 aprēķinu rezultāti P2.5.1. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.0014	0.0025	0.0018
o_2	0.0014	0	0.0004	0.0019
o_3	0.0025	0.0004	0	0.0012
o_4	0.0018	0.0019	0.0012	0

Lai aprēķinātu standarta attālumus χ_1, vienādojums (P2.5.2) jāpārvērš formā:

$$\chi_1(o_i - o_l) = \sqrt{\sum_{j=1}^{n} \frac{1}{a_j^*} \left(\frac{a_{ij}}{a_i^*} - \frac{a_{lj}}{a_k^*} \right)^2} = \sqrt{\sum_{j=1}^{n} \frac{1}{a_j^*} \left(p_{ij} - p_{lj} \right)^2} \ \ldots \ i, l = 1, \ldots, m, \ i \neq l. \qquad \text{(P2.5.3)}$$

Šajā piemērā nav nepieciešams veikt aprēķinus, izmantojot vienādojumu (P2.5.3). Lai iegūtu nepieciešamās vērtības, pietiek ar kvadrātsakni no katras vērtības no P2.5.2. tabulas. Gala rezultāti ir parādīti P2.5.3. tabulā.

P2.5.3. tabula

Standarta attālumu χ_1 aprēķinu rezultāti P2.5.1. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.0374	0.0500	0.0424
o_2	0.0374	0	0.0200	0.0436
o_3	0.0500	0.0200	0	0.0346
o_4	0.0424	0.0436	0.0346	0

Otrajā variantā kvadrātisko attālumu χ^2 aprēķināšanai, tāpat kā pirmajā variantā, tiek aprēķinātas atribūtu vērtību summas datu tabulas rindās un kolonnās: $a_i^* = \sum_{j=1}^{n} a_{ij}$, $i = 1, \ldots, m$, $a_j^* = \sum_{i=1}^{m} a_{ij}$, $j = 1, \ldots, n$. Pamatojoties uz aprēķinātajām a_i^* vērtībām, tiek veidoti atribūtu vērtību profili.

Atšķirība starp variantiem slēpjas svara matricas **W** veidošanas principā. Otrajā variantā tiek aprēķināta visu atribūtu vērtību summa tabulā: $a_{ij}^{*} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}$. Pēc tam atribūtu smaguma centru vērtības aprēķina kā attiecību:

$$c_j = \frac{a_j^{*}}{a_{ij}^{*}},\ j = 1,...n. \qquad \text{(P2.5.4)}$$

Diagonālās svara matricas **W** komponentu vērtības tiek aprēķinātas kā:

$$w_j = \frac{1}{c_j}. \qquad \text{(P2.5.5)}$$

Tādējādi svara matrica ir šādā veidā:

$$\mathbf{W} = \begin{bmatrix} \frac{1}{c_1}...0...0...0...0 \\ ..... \\ 0...0...\frac{1}{c_j}...0...0 \\ ..... \\ 0....0...0....0...\frac{1}{c_n} \end{bmatrix}.$$

Kvadrātisko attālumu χ_2^2 noteikšanai tiek izmantots vienādojums:

$$\chi_2^2(o_i - o_l) = \sum_{j=1}^{n} w_i \left(p_{ij} - p_{lj}\right)^2,\ i,l = 1,...,m,\ l \neq k, \qquad \text{(P2.5.6)}$$

kur p_{ij}, p_{lj} – atribūtu a_{ij}, a_{lj} vērtību profili;

w_i tiek aprēķināts, izmantojot vienādojumu (P2.5.5).

P2.5.2. piemērs. Ņemot P2.5.1. piemēra sākotnējos datus, ir jāaprēķina kvadrātisko un standarta attālumu χ_2^2, χ_2 vērtības starp objektiem (atribūtu vērtību vektoriem), izmantojot otro variantu atribūtu svara matricas formēšanai.

Risinājums. P2.5.4. tabulas kreisajā pusē ir sākotnējie dati, bet labajā – atribūtu vērtību profili.

P2.5.4. tabula

Sākotnējie dati un provizorisko aprēķinu rezultāti P2.5.2. piemēram

	Sākotnējās atribūtu vērtības					Atribūtu vērtību profili			
	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_i^{*}	p_{i1}	p_{i2}	p_{i3}	p_{i4}
o_1	7	3	2	10	22	0.318	0.136	0.091	0.455
o_2	9	2	5	6	22	0.409	0.091	0.228	0.273
o_3	12	5	8	8	33	0.364	0.152	0.242	0.242
o_4	5	4	6	9	2 4	0.208	0.167	0.250	0.375
a_j^{*}	33	14	21	33	101				
c_j	0.327	0.138	0.208	0.327					
$w_1 = \frac{1}{c_j}$	3.058	7.246	4.808	3.058					

Skaitlis $a_{ij}^{*}=101$ rindas a_{i}^{*} un kolonnas a_{j}^{*} krustpunktā ir visu tabulas atribūtu vērtību summa. Vērtības c_j, $j=1,...,4$ aprēķina, dalot atbilstošo vērtību a_{j}^{*} ar vērtību $a_{ij}^{*}=101$. Vērtības w_j ir vērtību c_j apgrieztās vērtības.

Veic aprēķinus, izmantojot vienādojumu (P2.5.6). Nav nepieciešams veikt aprēķinus pilnā mērā, var izmantot aprēķinu rezultātus no P2.5.1. piemēra.

$$\chi_2^2\left(o_1-o_2\right)=3.058*0.00828+7.246*0.00202+4.808*0.01877+3.058*0.03312=$$
$$=0.02532+0.01464+0.09025+0.10128\approx 0.2315;$$
$$\chi_2^2\left(o_1-o_3\right)=3.058*0.00212+7.246*0.00026+4.808*0.02280+3.058*0.04537=$$
$$=0.00648+0.00188+0.10962+0.13874\approx 0.2567.$$

Pārējie aprēķini tiek veikti līdzīgi. Aprēķinu rezultāti ir parādīti P2.5.5. tabulā.

Lai aprēķinātu standarta attālumus χ_2 saskaņā ar otro variantu, vienādojums (P2.5.6) ir jāpārvērš šādā formā:

$$\chi_2\left(o_i-o_l\right)=\sqrt{\sum_{j=1}^{n} w_j\left(p_{ij}-p_{lj}\right)^2}\ \ldots\ i,l=1,...,m,\ i\neq l. \qquad \text{(P2.5.7)}$$

Nav nepieciešamības veikt aprēķinus, izmantojot vienādojumu (P2.5.7). Lai iegūtu standarta attālumus χ_2 šajā piemērā, pietiek ņemt kvadrātsaknes no P2.5.5. tabulā uzrādītajām $\chi_2^2\left(o_i-o_l\right)$ vērtībām. Galīgie rezultāti ir parādīti P2.5.6. tabulā.

P2.5.5. tabula

Kvadrātisko attālumu χ_2^2 aprēķinu rezultāti P2.5.2. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.2315	0.2567	0.1877
o_2	0.2315	0	0.0371	0.1995
o_3	0.2567	0.0371	0	0.1167
o_4	0.1877	0.1995	0.1167	0

P2.5.6. tabula

Standarta attālumu χ_2 aprēķinu rezultāti P2.5.2. piemēram

	o_1	o_2	o_3	o_4
o_1	0	0.4811	0.5067	0.4532
o_2	0.4811	0	0.1926	0.4466
o_3	0.5067	0.1926	0	0.3416
o_4	0.4332	0.4466	0.3416	0

Kurš no diviem iepriekš parādītajiem variantiem ir labāks? Atbilde uz šo jautājumu ir nepārprotama: abi varianti ir līdzvērtīgi. Aprēķinu apjoms abos variantos ir vienāds. Otrā varianta neliela priekšrocība ir tā, ka iegūtajiem attāluma aprēķiniem ir ievērojami lielākas absolūtās vērtības, tāpēc aprēķinu rezultātu noapaļošana mazāk ietekmē šīs aplēses.

Noslēdzot šo sadaļu, jāatzīmē, ka vērtības c_j tiek sadalītas saskaņā ar χ^2 likumu, tāpēc arī ir tāds attāluma aprēķinu nosaukums.

P2.6. Mahalanobisa distance

Iepriekšējā sadaļā tika apskatītas divas Eiklīda attāluma versijas, kurās tika izmantoti atribūtu svari. Šo svaru vērtības tika aprēķinātas īpašā veidā, pamatojoties uz atribūtu vērtību kombinācijām.

Šajā sadaļā ir apskatīts cits paplašinātā Eiklīda attāluma variants. Šo versiju ierosināja indiešu zinātnieks P.Mahalanobis savā darbā [Mahalanobis P.S., 1936]. Autors pētīja līdzības un atšķirības starp dažādām iedzīvotāju grupām Indijā un nonāca pie secinājuma, lai aprēķinātu attālumus starp attiecīgajām grupām daudzdimensiju datos, ir jāņem vērā atribūtu vērtību izkliedes pakāpe (variācija) katrā no grupām un kovariācija starp atsevišķiem atribūtu vērtību vektoriem katram grupu pārim. Šajā sadaļā tiks noteikts Mahalanobisa attālums, pamatojoties tikai uz atribūtu vērtību izmaiņām.

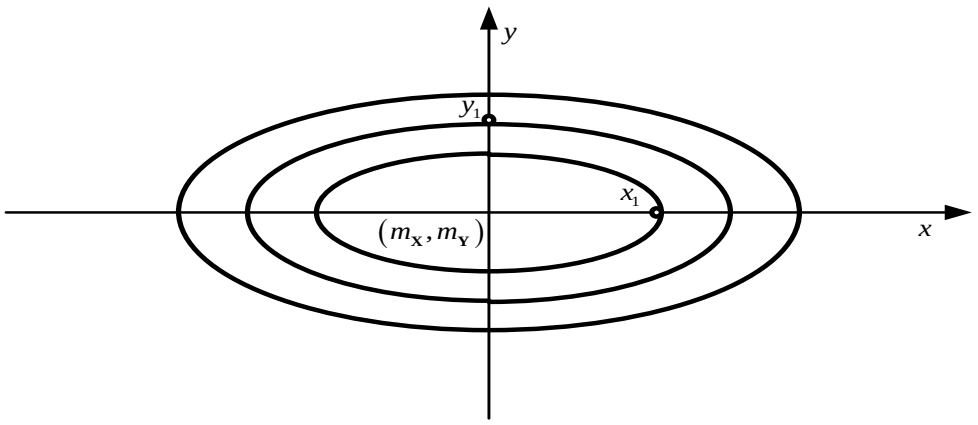
Lai iepazīstinātu ar Mahalanobisa attāluma definīcijas vispārējo ideju, der apskatīt piemēru. Pieņem, ka ir divdimensiju mainīgais $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$. Pamatojoties uz sākotnējo statistisko materiālu, tiek noteikta šī mainīgā lieluma vērtību sadalījuma funkcija $f(\mathbf{X}, \mathbf{Y})$. P2.6.1. attēlā parādīta šādas nosacītas sadalījuma funkcijas projekcijas mainīgo sākotnējo vērtību telpā $\mathbf{X}$ un $\mathbf{Y}$. Koordinātu sākuma punkts ir $(m_{\mathbf{X}}, m_{\mathbf{Y}})$, kur $m_{\mathbf{X}}$, $m_{\mathbf{Y}}$ ir mainīgo $\mathbf{X}$ un $\mathbf{Y}$ vidējās vērtības.

Katra ovāla līnija šajā attēlā apzīmē divdimensiju sadalījuma blīvuma funkcijas vērtību kopu $f(\mathbf{X}, \mathbf{Y})$.

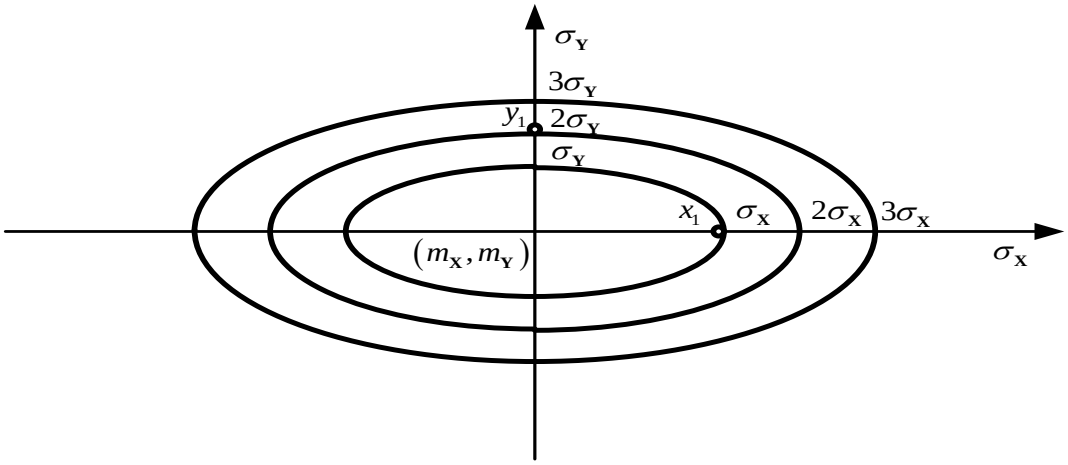
P2.6.2. attēlā parādītas tās pašas projekcijas, bet koordinātu sistēmā, kuras sastāvdaļas ir standartnovirzes $\sigma_{\mathbf{X}}$, $\sigma_{\mathbf{Y}}$.

Apskatot divus punktus x_1, y_1, kas parādīti P2.6.1. attēlā, acīmredzams, ka koordinātu sistēmā x, y punkts y_1 ir tuvāk koordinātu centram nekā punkts x_1. Pārnesot šos punktus uz koordinātu sistēmu $\sigma_{\mathbf{X}}$, $\sigma_{\mathbf{Y}}$ P2.6.2. attēlā, punktam x_1 ir koordinātes $(\sigma_{\mathbf{X}}, 0)$, punktam y_1 – koordinātes $(0, 2\sigma_{\mathbf{Y}})$.

Ja par attālumu mērīšanas mērvienībām ņem vērtības $\sigma_{\mathbf{X}}$, $\sigma_{\mathbf{Y}}$, redzams, ka punkts x_1 ir tuvāk koordinātu sākumam nekā punkts y_1. No tā izriet, ka standarta novirzes σ (un attiecīgi variācijas σ^2) var kalpot kā specifiski attāluma mēri, kuru pamatā ir gadījuma lielumu vērtību izkliedes pakāpe attiecībā pret noteiktiem centrālajiem punktiem. Šī ideja arī ir Mahalanobisa attāluma definīcijas pamatā.



P2.6.1. attēls. **Divdimensiju sadalījuma blīvuma funkcijas $f(\mathbf{X}, \mathbf{Y})$ projekciju grafiskais attēlojums vērtību $\mathbf{X}$ un $\mathbf{Y}$ telpā**



P2.6.2. attēls. **Funkciju $f(\mathbf{X},\mathbf{Y})$ projekciju grafiskais attēlojums koordinātu sistēmā $\sigma_{\mathbf{X}}$, $\sigma_{\mathbf{Y}}$**

Katra atribūta vērtības var korekti interpretēt kā gadījuma mainīgo realizāciju objektu kopā. Izmantojot atribūtu vērtību variācijas attiecībā pret to vidējām vērtībām kā atribūtu svaru, var formulēt šādu vispārīgu vienādojumu kvadrātiskajam attālumam starp objektiem:

$$d^2(o_i - o_l) = (\mathbf{a}_{ij} - \mathbf{a}_{lj})^T \mathbf{W} (\mathbf{a}_{ij} - \mathbf{a}_{lj}), \qquad \text{(P2.6.1)}$$

kur $\mathbf{a}_{ij}$, $\mathbf{a}_{lj}$ – atribūtu vērtību vektori objektiem o_i, o_k (šo vektoru vērtības parādītas sākotnējo datu tabulas rindās o_k, o_i;

T – matricas transponēšanas simbols;

$\mathbf{W}$ – svara diagonālā matrica, kuras elementi ir atbilstošo atribūtu variāciju apgrieztās vērtības. Datu tabulai, kas sastāv no n kolonnām, matricas $\mathbf{W}$ forma ir:

$$\mathbf{W} = \begin{pmatrix} \frac{1}{s_1^2} ....0....0....0....0 \\ ..... \\ 0....0....\frac{1}{s_j^2}....0....0 \\ ..... \\ 0....0....0.....0....\frac{1}{s_n^2} \end{pmatrix},,$$

Praktisku aprēķinu veikšanai vispārīgo vienādojumu (P2.6.1) var uzrādīt formā:

$$d^2(o_i - o_l) \sum_{j=1}^{n} \frac{1}{s_j^2} (a_{ij} - a_{lj})^2,\ i,l = 1,...m,\ i \neq l, \qquad \text{(P2.6.2)}$$

kur s_j^2 – atribūta j vērtību variācijas novērtējums;

a_{ij}, a_{lj} – j vērtības objektiem o_i, o_l.

(Jāņem vērā, ka vienādojums (P2.6.2) tiek izmantots, lai aprēķinātu kvadrātiskos Mahalanobisa attālumus, neņemot vērā kovariācijas starp atsevišķu atribūtu vērtībām).

P2.6.1. piemērs ilustrē Mahalanobisa attālumu starp objektiem aprēķinu. Par pamatu ņem sākotnējos datus no 4.2.1. tabulas. Šie dati ir reproducēti P2.6.1. tabulā. Nepieciešams aprēķināt kvadrātiskos Mahalanobisa attālumus starp visiem objektiem, izmantojot vienādojumu (P2.6.2).

Risinājums. Aprēķina atribūtu vidējās vērtības P2.6.1. tabulas kolonnās pēc vienādojuma:

$$m_j = \frac{\sum_{i=1}^{4} a_{ij}}{4},\ j = 1,...,4.$$

P2.6.1. tabula

Sākotnējie dati P2.6.1. piemēram

	a_{i1}	a_{i2}	a_{i3}	a_{i4}
o_1	7	3	2	10
o_2	9	2	5	6
o_3	12	5	8	8
o_4	5	4	6	9

$$m_1 = \frac{7+9+12+5}{4} = \frac{33}{4} = 8.25;\ m_2 = \frac{3+2+5+4}{4} = \frac{14}{4} = 3.50;$$
$$m_3 = \frac{2+5+8+6}{4} = \frac{21}{4} = 5.25;\ m_4 = \frac{10+6+8+9}{4} = \frac{33}{4} = 8.25.$$

Variāciju vērtības tabulas kolonnās tiek aprēķinātas, izmantojot vienādojumu:

$$s_j^2 = \frac{\sum_{i=1}^{4} \left(a_{ij} - m_j\right)^2}{3}.$$

$$s_1^2 = \frac{1}{3} * \left((7-8.25)^2 + (9-8.25)^2 + (12-8.25)^2 + (5-8.25)^2 \right) =$$
$$= \frac{1}{3}(1.5625 + 0.5625 + 14.0625 + 10.5650) = \frac{1}{3} * 26.2675 = 8.9167;$$
$$s_2^2 = \frac{1}{3} * \left((3-3.50)^2 + (2-3.50)^2 + (5-3.50)^2 + (4-3.50)^2 \right) =$$
$$= \frac{1}{3}(0.2500 + 2.2500 + 2.2500 + 0.2500) = \frac{1}{3} * 5.0000 = 1.6667;$$
$$s_3^2 = \frac{1}{3} * \left((2-5.25)^2 + (5-5.25)^2 + (8-5.25)^2 + (6-5.25)^2 \right) =$$
$$= \frac{1}{3} * (10.5625 + 0.0625 + 7.5625 + 0.5625) = \frac{1}{3} * 18.750 = 6.2500;$$
$$s_4^2 = \frac{1}{3} * \left((10-8.25)^2 + (6-8.25)^2 + (8-8.25)^2 + (9-8.25)^2 \right) =$$
$$= \frac{1}{3} * (3.0625 + 5.0625 + 0.0625 + 0.5625) = \frac{1}{3} * 8.7500 = 2.9167.$$

Aprēķina s_j^2 apgrieztās vērtības, $j = 1,2,3,4$.

$$\frac{1}{s_1^2} = \frac{1}{8.9167} = 0.1121;\ \frac{1}{s_2^2} = \frac{1}{1.6667} = 0.6000;$$
$$\frac{1}{s_3^2} = \frac{1}{6.2500} = 0.1600;\ \frac{1}{s_4^2} = \frac{1}{2.9167} = 0.3429.$$

Nepieciešamos aprēķinus veic, izmantojot vienādojumu (P2.6.2).

$$d^2(o_1 - o_2) = 0.1121 * (7-9)^2 + 0.6000 * (3-2)^2 + 0.1600 * (2-5)^2 + 0.3429 * (10-6)^2 =$$
$$= 0.1121 * 4 + 0.6000 * 1 + 0.1600 * 9 + 0.3429 * 16 =$$
$$= 0.4484 + 0.6000 + 1.4400 + 5.4864 \approx 7.975;$$

$$d^2(o_1-o_3)=0.1121*(7-12)^2+0.6000*(3-5)^2+0.1600*(2-8)^2+0.3429*(10-8)^2=$$
$$=0.1121*25+0.6000*4+0.1600*36+0.3429*4=$$
$$=2.8025+2.4000+5.7600+1.3716\approx 12.234;$$
$$d^2(o_1-o_4)=0.1121*(7-5)^2+0.6000*(3-4)^2+0.1600*(2-6)^2+0.3429*(10-9)^2=$$
$$=0.1121*4+0.6000*1+0.1600*16+0.3429*1=$$
$$=0.4484+0.6000+2.5600+0.3429\approx 3.951;$$
$$d^2(o_2-o_3)=0.1121*(9-12)^2+0.6000*(2-5)^2+0.1600*(5-8)^2+0.3429*(6-8)^2=$$
$$=0.1121*9+0.6000*9+0.1600*9+0.3429*4=$$
$$=1.0089+5.400+1.4400+1.3716\approx 9.220;$$
$$d^2(o_2-o_4)=0.1121*(9-5)^2+0.6000*(2-4)^2+0.1600*(5-6)^2+0.3429*(6-9)^2=$$
$$=0.1121*16+0.6000*4+0.1600*1+0.3429*9=$$
$$=1.7936+2.400+0.160+3.0861\approx 7.440;$$
$$d^2(o_3-o_4)=0.1121*(12-5)^2+0.6000*(5-4)^2+0.1600*(8-6)^2+0.3429*(8-9)^2=$$
$$=0.1121*49+0.6000*1+0.1600*4+0.3429*1=$$
$$=5.4929+0.6000+0.6400+0.3429\approx 7.076.$$

Iegūtie rezultāti ir apkopoti P2.6.2. tabulā.

Lai aprēķinātu standarta Mahalanobisa attālumu, vienādojums (P2.6.2) ir jāpārvērš formā:

$$d(o_i-o_l)=\sqrt{\sum_{j=1}^{n}\frac{1}{s_j^2}\left(a_{ij}-a_{lj}\right)^2}\ \ldots\ i,l=1,\ldots m,\ i\neq l. \qquad \text{(P2.6.3)}$$

P2.6.2. tabula

Kvadrātisko Mahalanobisa attālumu aprēķinu rezultāti P2.6.1. piemēram

	o_1	o_2	o_3	o_4
o_1	0	7.975	12.334	3.951
o_2	7.975	0	9.220	7.440
o_3	12.334	9.220	0	7.076
o_4	3.951	7.440	7.076	0

P2.6.2. piemērs. Izmantojot P2.6.1. piemērā iegūtos rezultātus, ir jāaprēķina standarta Mahalanobisa attālumi starp visiem objektu pāriem.

Risinājums. Šajā piemērā nav jāveic pilni aprēķini. No vienādojuma (P2.6.3) analīzes izriet, ka, lai aprēķinātu standarta Mahalanobisa attālumu starp objektu pāriem, pietiek aprēķināt kvadrātsaknes vērtību no atbilstošās kvadrātiskās Mahalanobisa attāluma vērtības P2.6.2. tabulā. Aprēķinu rezultāti ir parādīti P2.6.3. tabulā.

P2.6.3. tabula

Standarta Mahalanobisa attālumu aprēķinu rezultāti P2.6.2. piemēram

	o_1	o_2	o_3	o_4
o_1	0	2.824	3.512	1.988
o_2	2.824	0	3.036	2.728
o_3	3.512	3.036	0	2.660
o_4	1.988	2.728	2.660	0

P3. NORMĀLĀ SADALĪJUMA LIKUMA PĀRBAUDE

6. nodaļā tika apskatītas datu transformācijas. Viens no šādu transformāciju mērķiem ir pārveidot uz labo vai kreiso pusi vērstus sadalījumus normālos vai gandrīz normālos sadalījumos.

Pirms atribūtu vērtību transformācijas ir jāpārliecinās, kā šīs vērtības tiek sadalītas. Šeit ir divas iespējas: (1) attiecīgais sadalījums ir normāls vai tuvu normālam un nav nepieciešama atribūtu vērtību transformācija; (2) attiecīgais sadalījums nav normāls un ir nepieciešama tā transformācija.

Var atzīmēt, ka daudzas cilvēku populācijas raksturojošas īpašības pēc savas būtības ir sadalītas pēc normālā sadalījuma likuma (svars, augums, vecums, IQ koeficients).

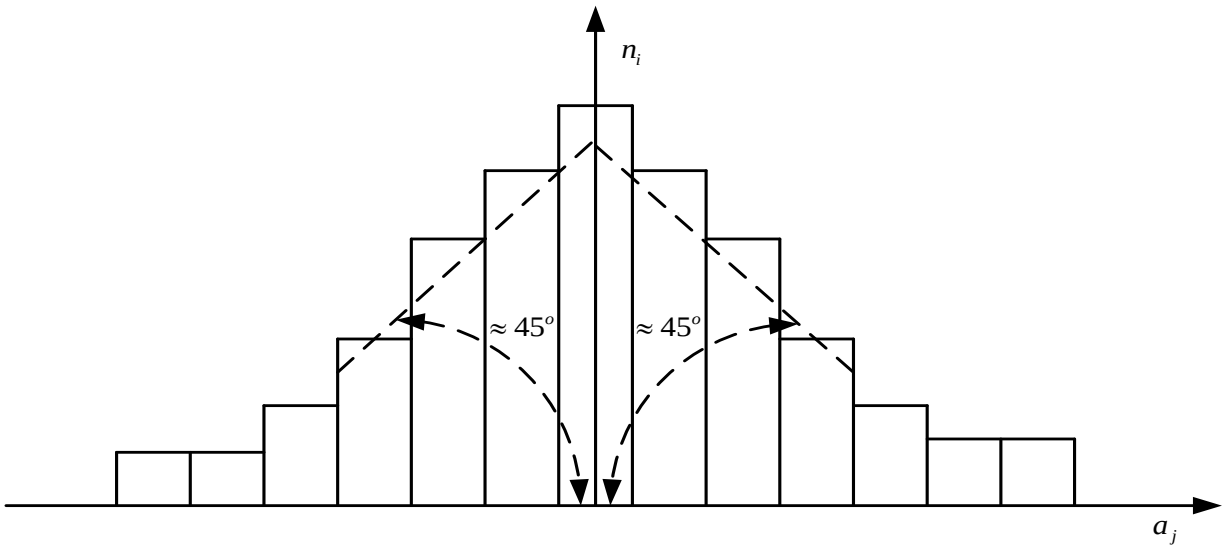
Šeit rodas vēl viens jautājums. Pieņem, ka reģiona iedzīvotāju vidū šo atribūtu vērtības patiešām ir sadalītas saskaņā ar normālo likumu. Bet socioloģiskajos un citos pētījumos tiek izmantoti attiecīgās populācijas izlases dati. Šādos gadījumos attiecīgo atribūtu sadalījuma likumi var lielākā vai mazākā mērā atšķirties no normālā sadalījuma. Turklāt sadalījuma parametri dažādiem paraugiem var atšķirties. Šī problēma ir saistīta ar izlases datu reprezentativitāti un ir jāatrisina pētījuma organizēšanas procesā.

Šajā pielikumā ir sniegtas dažādas pieejas normālā sadalījuma pārbaudei un dažādas aplēses, kas raksturo konkrētā sadalījumu novirzes pakāpi no normālā sadalījuma prasībām. Ērtības labad turpmāk bieži tiks lietots termins *normalitāte*, kas arī raksturo sadalījuma atbilstību normālajam sadalījuma likumam.

P3.1. Vienkāršākās metodes sadalījumu pārbaudei uz normalitāti

Vienkāršākais veids, kā novērtēt sadalījuma iespējamo normalitāti, ir grafiski attēlot atribūtu vērtību kopu histogrammas veidā. Viens šāds nosacīts attēlojums ir parādīts P3.1.1. attēlā.

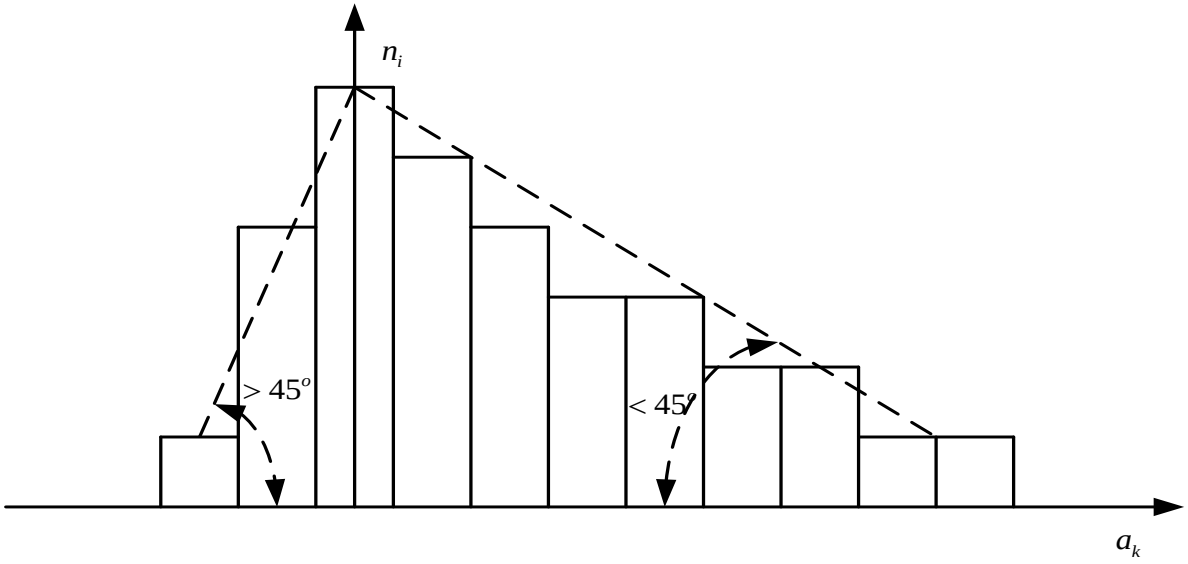
Sadalījuma histogramma tiek veidota šādi: atribūtu a_j vērtību diapazons ir sadalīts noteiktā intervālu skaitā. Šie intervāli ir parādīti attēlā uz horizontālās ass. Tālāk tiek uzskaitīts atribūtu vērtību skaits, kas ietilpst katrā intervālā. Šis skaitlis tiek uzdots kā kolonas augstums, kas attēlots katrā intervālā.



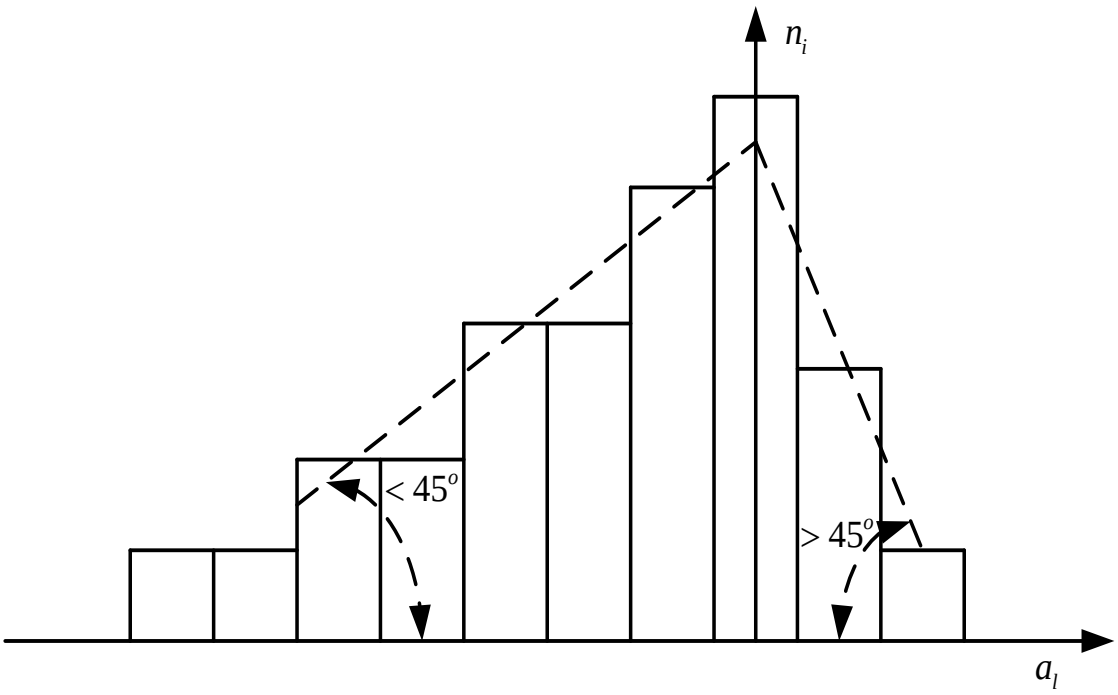
P3.1.1. attēls. **Atribūtu a_j vērtību nosacītā sadalījuma histogramma**

Kādus secinājumus var izdarīt no histogrammas analīzes P3.1.1. attēlā? Pirmkārt, atribūtu a_j vērtības ir simetriski sadalītas abās pusēs no centrālās vērtības. Otrkārt, taisnajām līnijām, kas atspoguļo vispārējās tendences atribūtu vērtību izmaiņās to vidējo vērtību zonās, ir aptuveni 45^0 slīpums attiecībā pret horizontālo asi. Šīs abas pazīmes norāda uz to, ka atribūtu vērtības a_j var tikt sadalītas pēc normālā sadalījuma likuma.

P3.1.2. un P3.1.3. attēlā parādītas atribūtu a_k , a_l vērtību sadalījuma histogrammas.



P3.1.2. attēls. **Atribūtu a_k vērtību nosacītā sadalījuma histogramma**



P3.1.3. attēls. **Atribūtu a_l vērtību nosacītā sadalījuma histogramma**

P3.1.2. attēlā redzams, ka atribūtu a_k vērtību sadalījums nav simetrisks. Datu punktu skaits sadalījuma labajā pusē ievērojami pārsniedz datu punktu skaitu tā kreisajā pusē. Līniju slīpumi, kas raksturo vispārējās atribūtu vērtību izmaiņu tendences, ir ļoti atšķirīgi. Līnijas slīpums sadalījuma kreisajā pusē ir ievērojami lielāks par 45^0, bet līnijas slīpums sadalījuma labajā pusē ir ievērojami mazāks par 45^0.

P3.1.3. attēlā redzams, ka atribūtu a_l vērtību sadalījums arī nav simetrisks. Datu punktu skaits sadalījuma kreisajā pusē ievērojami pārsniedz datu punktu skaitu labajā pusē. Līnijas slīpums sadalījuma kreisajā pusē ir ievērojami mazāks par 45^0, savukārt līnijas slīpums sadalījuma labajā pusē ir ievērojami lielāks par 45^0.

Histogramma P3.1.2. attēlā parāda, ka atribūtu a_k vērtību sadalījums ir tipisks uz labo pusi vērsts sadalījums, histogramma P3.1.3. attēlā parāda, ka atribūtu a_l vērtību sadalījums ir tipisks uz kreiso pusi vērsts sadalījums.

Atribūtu a_j vērtību sadalījums P3.1.1. attēlā var būt normāls tādā nozīmē, ka tas potenciāli var būt normāls, taču ir jāveic turpmāki novērtējumi, lai apstiprinātu šo faktu. Atribūtu vērtību a_k, a_l sadalījumi P3.1.2., P3.1.3. attēlā nevar būt normāli un nav nepieciešams cits šī fakta apstiprinājums.

Tādējādi vienkāršāko metožu mērķi sadalījuma pārbaudei uz normalitāti ir šādi: (1) noteikt, ka attiecīgais sadalījums var būt normāls un turpināt to pārbaudīt; (2) pārliecināties, ka attiecīgais sadalījums nav normāls un izstrādāt stratēģiju tā transformācijai.

P3.2. Sadalījuma normalitātes pārbaude, pamatojoties uz sadalījuma parametriem

Pieņem, ka visiem objektiem var norādīt sākotnējo atribūtu a_j vērtību kopu. Tad vienmēr var aprēķināt tā vērtību sadalījuma parametrus:

- *sadalījuma izvietojuma parametrs* – sagaidāmā vērtība:

$$\bar{a}_j = \sum_{i=1}^{n} a_{ij} p_i , \qquad \text{(P3.2.1)}$$

kur n – intervālu skaits, kas veidojas atribūtu a_j vērtību izmaiņu diapazonā;

a_{ij} – atribūta a_j vidējā vērtība i intervālā $i = 1,...,n$;

p_i – varbūtība, ka atribūta a_j vērtība nonāks i intervālā;

$$p_i = \frac{n_i}{n}, \qquad \text{(P3.2.2)}$$

kur n_i – atribūtu a_j vērtību skaits i intervālā;

n – kopējais atribūtu a_j vērtību skaits;

- s*adalījuma skalas parametrs* – atribūtu a_j vērtību standartnovirze no sagaidāmās vērtības:

$$s_j = \sqrt{\sum_{i=1}^{n} \left(a_{ij} - \bar{a}_j \right)^2 p_i} . \qquad \text{(P3.2.3)}$$

P3.2.1. piemērs. P3.2.1. tabulā ir parādīta sākotnējā atribūtu a_j vērtību kopa un dati par to vērtību sadalījumu. Ir nepieciešams novērtēt šī sadalījuma izvietojuma parametrus un novērtēt iespēju, ka sadalījums var būt normāls. Tabulas pirmajā rindā ir norādīti atribūtu a_j vērtību intervālu numuri. Šajā piemērā tiek pieņemts, ka katra intervāla garums ir vienāds ar vienu atribūtu a_j vērtību vienību. Otrajā tabulas rindā ir norādītas katra intervāla robežas.

Katrā intervālā tiek pieņemts, ka visas atribūtu a_j vērtības, kas ietilpst šajā intervālā, tiek aizstātas ar atribūta vidējo vērtību šajā intervālā. Tabulas trešajā rindā ir parādītas vidējās atribūtu a_j vērtības katrā intervālā.

Tabulas ceturtajā rindā ir parādīts atribūtu vērtību skaits katrā no intervāliem.

Tā kā šajā piemērā tiek runāts par atribūtu vērtību sadalījumu pa intervāliem, ir nepieciešamas atribūtu vērtību a_j varbūtības (biežumi) katrā intervālā. Aprēķina tabulas ceturtās rindas skaitļu summu. Tas ir vienāds ar 42. Lai aprēķinātu interesējošās varbūtības, katram intervālam ir jādala tabulas 4. rindas atbilstošajā šūnā esošais skaitlis ar 42. Iegūtie varbūtības aprēķini ir parādīti pēdējā tabulas rindā.

P3.2.1. tabula

Sākotnējo atribūtu a_j vērtību kopa un vērtību sadalījums

Intervāla numurs	1	2	3	4	5	6	7	8	9	10	11	12	13
Intervālu robežas	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5	10.5-11.5	11.5-12.5	12.5-13.5
Vidējā vērtība intervālā	1	2	3	4	5	6	7	8	9	10	11	12	13
Vērtību skaits intervālā	1	1	2	3	4	6	8	6	4	3	2	1	1
Varbūtība, ka atribūta vērtība nonāks intervālā	0.024	0.024	0.048	0.071	0.095	0.143	0.190	0.143	0.095	0.071	0.048	0.024	0.024

Atribūtu vērtību a_j sadalījuma histogramma ir parādīta P3.2.1. attēlā.

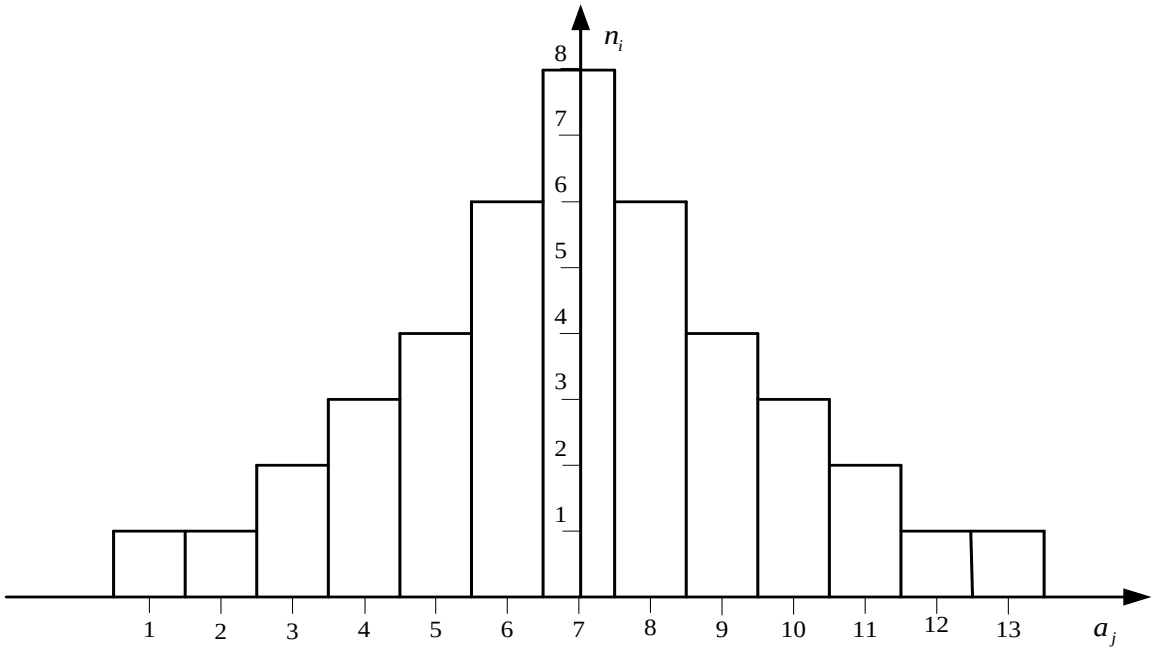
Aprēķina sadalījuma parametru vērtības:

$$\bar{a}_j = 1*0.024+2*0.024+3*0.048+4*0.071+5*0.095+6*0.143+7*0.190+$$
$$+8*0.143+9*0.095+10*0.071+11*0.048+12*0.024+13*0.024=$$
$$=0.024+0.048+0.144+0.284+0.476+0.858+1.330+1.144+$$
$$+0.855+0.710+0.528+0.288+0.312=6.926\approx 7\,.$$

Lai aprēķinātu atribūtu a_j vērtību standartnovirzi, vispirms aprēķina šī sadalījuma izkliedi (variāciju), izmantojot vienādojumu:

$$D_j=\sum_{i=1}^{n}\left(a_{ij}-\bar{a}_j\right)^2 p_i\,. \qquad \text{(P3.2.4)}$$

$$D_j=(1-7)_2*0.024+(2-7)^2*0.024+(3-7)^2*0.048+(4-7)^2*0.071+(5-7)^2*0.095+$$
$$+(6-7)^2*0.143+(7-7)^2*0.190+(8-7)^2*0.143+(9-7)^2*0.095+(10-7)^2*0.071+$$
$$+(11-7)^2*0.048+(12-7)^2*0.024+(13-7)^2*0.024=$$
$$=0.864+0.600+0.768+0.639+0.380+0.143+0+0.143+0.380+$$
$$+0.639+0.768+0.600+0.864=6.788\,.$$
$$s_j=\sqrt{D_j}=\sqrt{6.788}=2.60\,.$$



P3.2.1. attēls. **Atribūtu a_j vērtību sadalījuma histogramma datiem no P3.2.1. tabulas**

Kādā veidā, pamatojoties uz sadalījuma parametriem, var izdarīt secinājumus par sadalījuma normalitāti?

Jebkura simetriska sadalījuma raksturīga iezīme ir tā, ka tā sagaidāmā vērtība, moda un mediāna sakrīt. Sagaidāmā vērtība tika aprēķināta iepriekš: $\bar{a}_j = 7$.

Aprēķina šī sadalījuma mediānu. Lai to izdarītu, visas atribūtu a_j vērtības sakārto nedilstošā secībā:

1, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11, 12, 13

Kopumā mums ir 42 atribūtu a_j vērtības. Tā kā tas ir pārskaitlis, sadalījuma mediānu aprēķina kā vidējo vērtību no skaitļiem 21. un 22. pozīcijā iepriekš norādītajā vērtību secībā. Šajās pozīcijās ir atribūtu a_j vērtības, kas vienādas ar 7. Tāpēc:

$$med\left(a_j\right) = \frac{7+7}{2} = 7.$$

Moda ir atribūta a_j vērtība, kurai datu punktu skaits pa kreisi no šīs vērtības ir vienāds ar datu punktu skaitu pa labi no šīs vērtības. Acīmredzot šajā piemērā $\mathrm{mod}\left(a_j\right) = 7$.

Tā kā dažādo sadalījuma parametru vērtības sakrīt, var pamatoti secināt, ka atribūtu a_j vērtību sadalījums ir simetrisks sadalījums. Tas nozīmē, ka šis sadalījums var būt (bet ne vienmēr ir) normāls sadalījums.

Ja novērtējamais sadalījums patiešām ir normāls sadalījums, ir jāievēro šādas prasības:

- 68 % atribūtu vērtību atrodas standarta novirzes robežās no sagaidāmās vērtības;
- 95 % atribūtu vērtību atrodas divu standartnoviržu robežās no sagaidāmās vērtības;
- 99,7 % atribūtu vērtību atrodas trīs standartnoviržu robežās no sagaidāmās vērtības.

Šajā piemērā $\bar{a}_j = 7$, $s_j = 2.60$. Tas nozīmē, ka intervālā $\pm s_j$ no sagaidāmās vērtības ir intervāli 5, 6, 7, 8, 9. Šajā intervālā ir 4 + 6 + 8 + 6 + 4 = 28 atribūtu a_j vērtības jeb 67 % no visām tā vērtībām.

Intervālā $\pm 2s_j$ ir 38 atribūtu a_j vērtības jeb 91 % no visām tā vērtībām. Intervāls $\pm 3s_j$ satur 100 % atribūtu a_j vērtību.

Pamatojoties uz šiem rezultātiem, var apgalvot, ka atribūtu a_j vērtību sadalījums no P3.2.1. tabulas tuvināti ir normāls sadalījums.

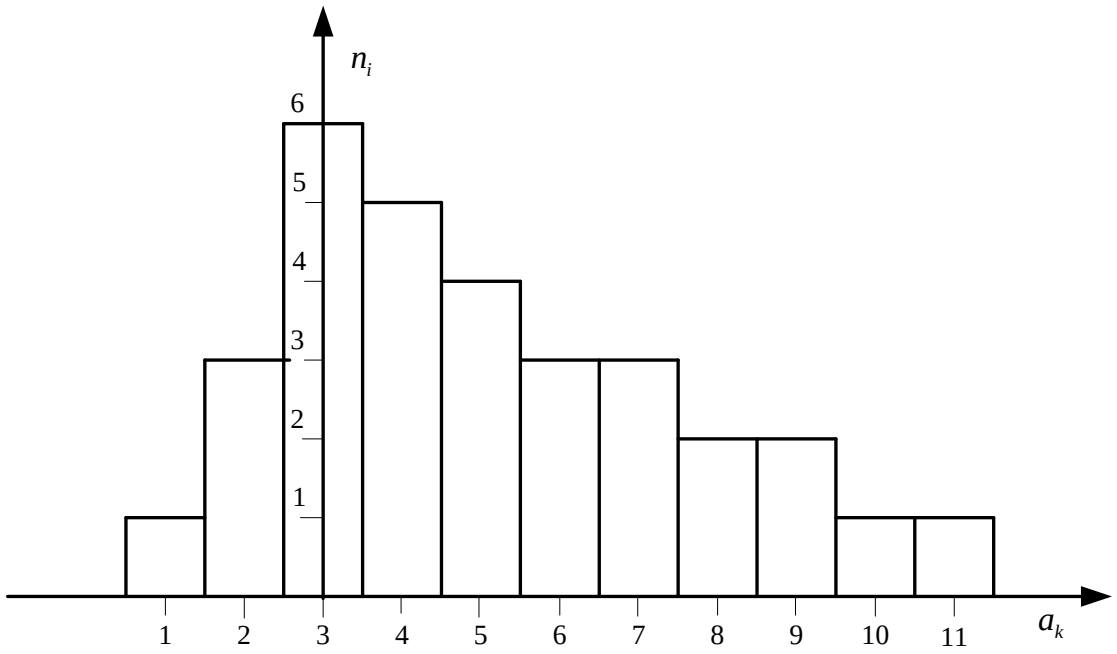
P3.2.2. piemērs. P3.2.2. tabulā ir parādītas dažādas sākotnējās atribūta a_k vērtības un citi dati par šī atribūta vērtību sadalījumu. Šīs tabulas datu interpretācija ir līdzīga P3.2.1. tabulas datu interpretācijai.

P3.2.2. tabula

Sākotnējā atribūta a_k vērtību kopa un vērtību sadalījums

Intervāla numurs	1	2	3	4	5	6	7	8	9	10	11
Intervālu robežas	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5	10.5-11.5
Vidējā vērtība intervālā	1	2	3	4	5	6	7	8	9	10	11
Vērtību skaits intervālā	1	3	6	5	4	3	3	2	2	1	1
Varbūtība, ka atribūta vērtība nonāks intervālā	0.032	0.098	0.194	0.161	0.128	0.098	0.098	0.064	0.064	0.32	0.032

Atribūtu a_k vērtību sadalījuma histogramma ir parādīta P3.2.2. attēlā.



P3.2.2. attēls. **Atribūtu a_k vērtību sadalījuma histogramma datiem no P3.2.2. tabulas**

Aprēķina atribūtu a_k vērtību sadalījuma parametrus.

$$\bar{a}_k = 1*0.032 + 2*0.098 + 3*0.194 + 4*0.161 + 5*0.128 + 6*0.098 +$$
$$+7*0.098 + 8*0.064 + 9*0.064 + 10*0.032 + 11*0.032 =$$
$$= 0.032 + 0.196 + 0.582 + 0.644 + 0.640 + 0.588 + 0.686 + 0.512 +$$
$$+0.576 + 0.320 + 0.352 = 5.127 \approx 5.$$

$$D_k = (1-5)^2 * 0.032 + (2-5)^2 * 0.098 + (3-5)^2 * 0.194 + (4-5)^2 * 0.161 + (5-5) * 0.129 +$$
$$+ (6-5)^2 * 0.098 + (7-5)^2 * 0.098 + (8-5)^2 * 0.064 + (9-5)^2 * 0.064 +$$
$$+ (10-5)^2 * 0.032 + (11-5)^2 * 0.032 = 0.512 + 0.576 + 0.776 + 0.111 + 0 +$$
$$+ 0.098 + 0.392 + 0.576 + 0.800 + 1.152 = 5.349.$$
$$s_k = \sqrt{5.349} = 2.31.$$

Nosaka sadalījuma mediānu. Atribūtu a_k vērtības sakārto nedilstošā secībā:

12,2,2,3,3,3,3,3,3,4,4,4,4,4,5,5,5,5,6,6,6,7,7,7,8,8,9,9,10,11.

Tā kā ir 31 atribūtu vērtība, mediāna būs vērtība, kas atrodas virknes 16. pozīcijā:

$$med(a_k) = 5.$$

Šajā piemērā $\mathrm{mod}(a_k) = 4$. Tā kā visi parametru aprēķini nav vienādi, atribūtu a_k vērtību sadalījums nav simetrisks un nevar būt normāls sadalījums.

Intervālā $\pm s_j$ no sagaidāmās vērtības atrodas intervāli 3, 4, 5, 6, 7. Šajā intervālā atrodas 21 atribūta a_k vērtība jeb 67 % no visām tā vērtībām.

Intervālā $\pm 2s_j$ ir 29 atribūtu a_j vērtības jeb 93 % no visām tā vērtībām.

Intervāls $\pm 3s_j$ satur 100 % atribūtu a_j vērtību.

Šie rezultāti arī norāda, ka atribūtu a_k vērtību sadalījums nav normāls sadalījums.

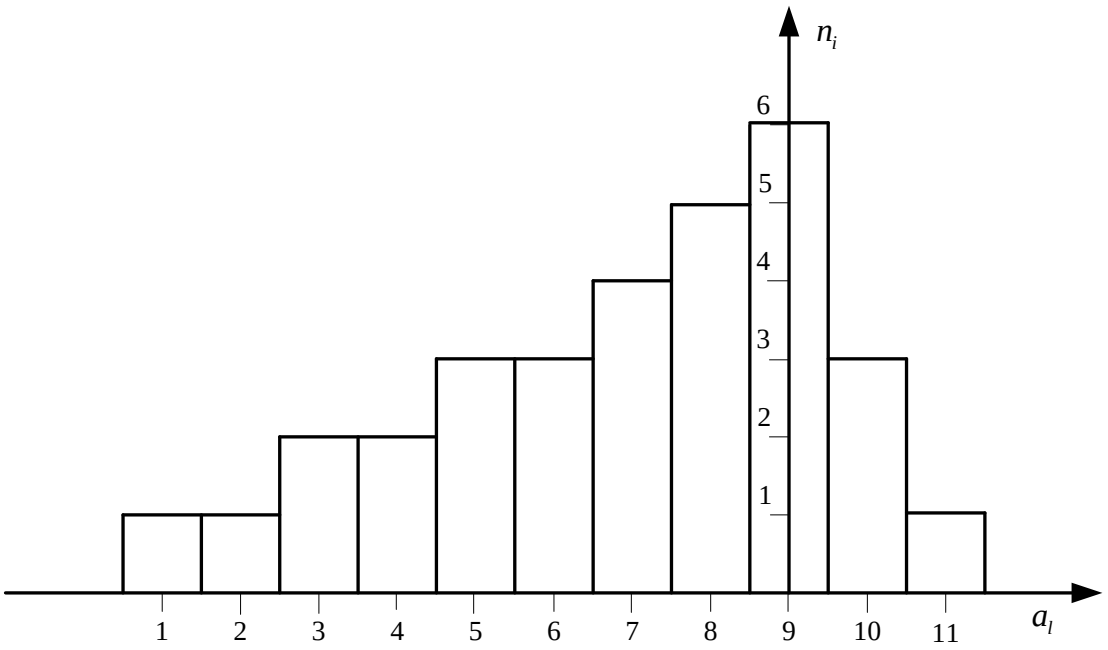
P3.2.3. piemērs. P3.2.3. tabulā ir norādītas atribūta a_l sākotnējās vērtības un citi dati par šī atribūta vērtību sadalījumu. Aprēķina šī atribūta parametru vērtības.

P3.2.3. tabula

Sākotnējā atribūta a_l vērtību kopa un vērtību sadalījums

Intervāla numurs	1	2	3	4	5	6	7	8	9	10	11
Intervālu robežas	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5	10.5-11.5
Vidējā vērtība intervālā	1	2	3	4	5	6	7	8	9	10	11
Vērtību skaits intervālā	1	1	2	2	3	3	4	5	9	3	1
Varbūtība, ka atribūta vērtība nonāks intervālā	0.032	0.032	0.064	0.064	0.097	0.097	0.129	0.161	0.194	0.098	0.032

Atribūtu a_l vērtību sadalījuma histogramma ir parādīta P3.2.3. attēlā.



P3.2.3. attēls. **Atribūtu a_l vērtību sadalījuma histogramma datiem no P3.2.3. tabulas**

Aprēķina atribūtu a_l vērtību sadalījuma parametrus.

$$\bar{a}_k = 1*0.032+2*0.032+3*0.064+4*0.064+5*0.097+6*0.097+$$
$$+7*0.129+8*0.161+9*0.194+10*0.098+11*0.032=$$
$$=0.032+0.064+0.192+0.256+0.485+0.582+0.903+0.1.288+$$
$$+1.746+0.980+0.352=6.880\approx 7.$$

$$D_l=(1-7)^2*0.032+(2-7)^2*0.032+(3-7)^2*0.064+(4-7)^2*0.064+(5-7)*0.097+$$
$$+(6-7)^2*0.097+(7-7)^2*0.129+(8-7)^2*0.161+(9-7)^2*0.194+$$
$$+(10-7)^2*0.098+(11-7)^2*0.032=1.152+0.800+1.024+0.570+0.388+$$
$$+0.097+0+0.161+0.776+0.882+0.512=6.368.$$

$$s_l=\sqrt{6.368}=2.52.$$

$$med(a_l)=7.$$

$$\mathrm{mod}(a_l)=8.$$

No tā izriet, ka atribūtu a_l vērtību sadalījums nevar būt normāls sadalījums.

67 % atribūtu vērtību ietilpst vienas standartnovirzes intervālā. 93 % atribūtu vērtību – divu standartnoviržu intervālā. 100 % atribūtu vērtību ietilpst trīs standarta noviržu intervālā. Tas arī norāda, ka atribūtu a_l vērtību sadalījums nav normāls sadalījums.

P3.3. Sadalījuma normalitātes pārbaude, pamatojoties uz augstākas kārtas parametriem

Atribūtu a_j vērtību sadalījuma asimetrijas rādītāju (*skewness*) S aprēķina pēc vienādojuma:

$$S(a_j)=\frac{n\sum_{i=1}^{n}\left(a_{ij}-\bar{a}_j\right)^3}{s_j^3(n-1)(n-2)}, \qquad \text{(P3.3.1)}$$

kur a_{ij} – atribūta a_j vērtība i intervālā;

$\bar{a}_j$ – atribūta a_j sagaidāmā vērtība;

s_j – atribūtu a_j vērtību standarta novirze no sagaidāmās vērtības;

n – kopējais atribūtu a_j vērtību skaits.

Stingri simetriskam sadalījumam S vērtība būs vienāda ar nulli. Uz kreiso pusi vērstam sadalījumam S vērtība ir mazāka par nulli, uz labo pusi vērstam sadalījumam S vērtība ir lielāka par nulli.

Simetrijas prasība ir nepieciešama, bet nav pietiekama, lai sadalījumu novērtētu kā normālu. Otrs faktors, no kura ir atkarīgi sadalījuma pārbaudes rezultāti attiecībā uz normalitāti, ir tā forma. Normālajam sadalījumam ir īpaša zvanveida forma (Gausa sadalījums). Bet ir iespējami simetriski sadalījumi, kuriem ir "plakanāka" vai "asāka" forma nekā parastajam sadalījumam.

Lai novērtētu sadalījuma formas atbilstību normalitātes prasībām, tiek izmantots īpašs parametrs – ekscesa rādītājs (*kurtosis*). Ekscesa vērtība tiek aprēķināta, izmantojot vienādojumu:

$$K\left(a_j\right)=\frac{\sum_{i=1}^{n}\left(a_{ij}-\bar{a}_j\right)^4/n}{s_j^4}-3, \qquad \text{(P3.3.2)}$$

kur a_{ij} – atribūtu a_j vērtības i objektam;

$\bar{a}_j$ – atribūtu a_j sagaidāmā vērtība;

s_j – atribūtu a_j vērtību standarta novirze no sagaidāmās vērtības;

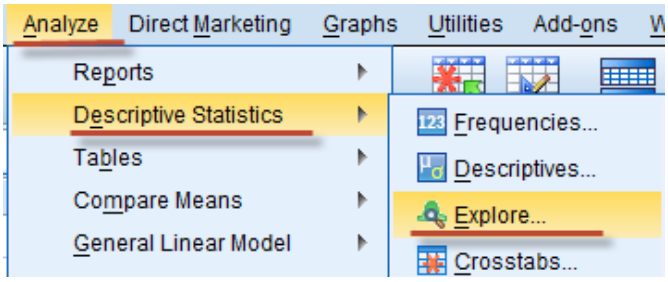
n – kopējais atribūtu a_j vērtību skaits.

Ja sadalījums ir "plakanāks" par parasto sadalījumu, ekscesa vērtība ir mazāka par nulli. Stingri normālam sadalījumam ekscesa vērtība ir nulle. Ja sadalījums ir "asāks" par parasto sadalījumu, ekscesa vērtība ir lielāka par nulli.

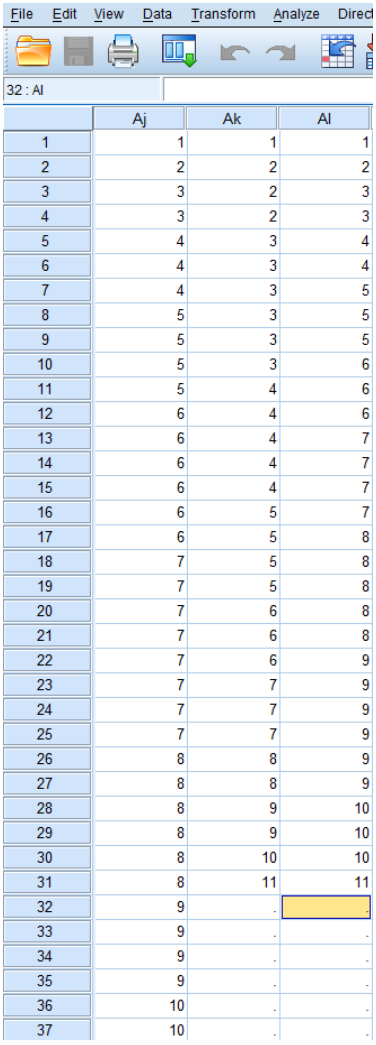
Ilustratīvi piemēri asimetrijas pakāpes novērtēšanai, pamatojoties uz parametru S.

P3.3.1. piemērs. Pēc P3.2.1. tabulas datiem nepieciešams novērtēt atribūtu a_j vērtību sadalījuma asimetrijas un ekscesa rādītājus.

Izmantojot SPSS programmatūras pakotni, asimetrijas un ekscesa aprēķināšanas procedūras tiek veiktas šādi: pēc SPSS palaišanas secīgi ir jāizvēlas opcijas "Analyze", "Descriptive Statistics" un "Explore".



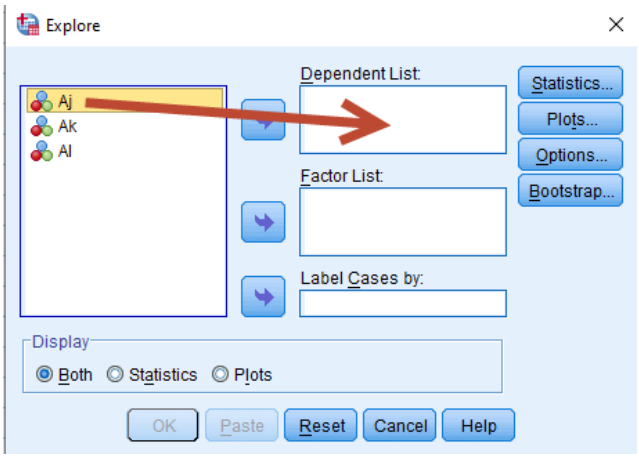
Logā, kas parādās, jāievada sākotnējie dati (*atribūts vai atribūtu vērtības, bet ne biežuma dati*).



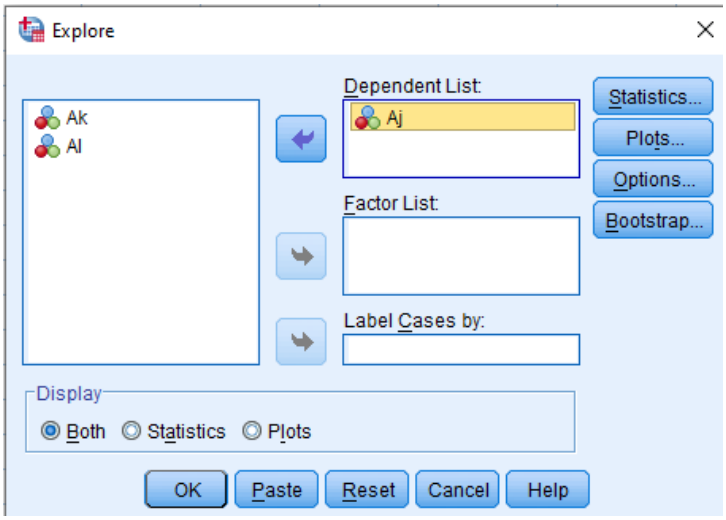
	Aj	Ak	Al
1	1	1	1
2	2	2	2
3	3	2	3
4	3	2	3
5	4	3	4
6	4	3	4
7	4	3	5
8	5	3	5
9	5	3	5
10	5	3	6
11	5	4	6
12	6	4	6
13	6	4	7
14	6	4	7
15	6	4	7
16	6	5	7
17	6	5	8
18	7	5	8
19	7	5	8
20	7	6	8
21	7	6	8
22	7	6	9
23	7	7	9
24	7	7	9
25	7	7	9
26	8	8	9
27	8	8	9
28	8	9	10
29	8	9	10
30	8	10	10
31	8	11	11
32	9	.	.
33	9	.	.
34	9	.	.
35	9	.	.
36	10	.	.
37	10	.	.

(Tehnisku iemeslu dēļ visas atribūtu a_j vērtības nav parādītas).

Pēc noklikšķināšanas uz pogas "OK" parādās logs.



Kreisajā panelī redzami testējamo atribūtu nosaukumi (šajā gadījumā – A_j, A_k, A_l). Jāizvēlas interesējošais atribūts A_j un ar bultiņas palīdzību tas jāpārsūta uz labo paneli.



Pēc pogas "OK" noklikšķināšanas tiek iegūti šādi rezultāti:

Descriptives

		Statistic	Std. Error
Mean		7,00	,406
95% ConfidenceIntervalforMean	LowerBound	6,18	
	UpperBound	7,82	
5% TrimmedMean		7,00	
Median		7,00	
Variance		6,927	
Std. Deviation		2,632	
Minimum		1	
Maximum		13	
Range		12	
InterquartileRange		4	
Skewness		,000	,365
Kurtosis		-,001	,717

Tabulas apakšējās divās rindās ir norādītas asimetrijas un ekscesa vērtības: $S(a_j) = 0.000$, $K = -0.001$. Tas norāda uz to, ka atribūtu a_j vērtības ir sadalītas pēc normālā sadalījuma likuma.

P3.3.2. piemērs. Atribūtu a_k vērtību sadalījuma asimetrijas vērtību nepieciešams novērtēt datiem no P3.2.2. tabulas. No šīs tabulas $S(a_k) = 0.581$, $K(a_k) = -0.427$. Atribūtu a_k vērtību sadalījums ir uz labo pusi vērsts sadalījums.

Izmantojot SPSS programmatūras pakotni, iegūst šādus rezultātus:

Descriptives

		Statistic	Std. Error
Mean		5,13	,461
95% ConfidenceIntervalforMean	LowerBound	4,19	
	UpperBound	6,07	
5% TrimmedMean		5,03	
Median		5,00	
Variance		6,583	
Std. Deviation		2,566	
Minimum		1	
Maximum		11	
Range		10	
InterquartileRange		4	
Skewness		,581	,421
Kurtosis		-,427	,821

P3.3.3. piemērs. Atribūtu a_l vērtību sadalījuma asimetrijas vērtību nepieciešams novērtēt datiem no P3.2.3. tabulas.

Izmantojot SPSS programmatūras pakotni, iegūst šādus rezultātus:

Descriptives

		Statistic	Std. Error
Mean		6,87	,461
95% ConfidenceIntervalforMean	LowerBound	5,93	
	UpperBound	7,81	
5% TrimmedMean		6,97	
Median		7,00	
Variance		6,583	
Std. Deviation		2,566	
Minimum		1	
Maximum		11	
Range		10	
InterquartileRange		4	
Skewness		-,581	,421
Kurtosis		-,427	,821

No šīs tabulas $S(a_l) = -0.581$, $K(a_l) = -0.427$. Atribūtu a_l vērtību sadalījums ir uz kreiso pusi vērsts sadalījums.

P3.4. Sadalījuma normalitātes neparametriskā pārbaude

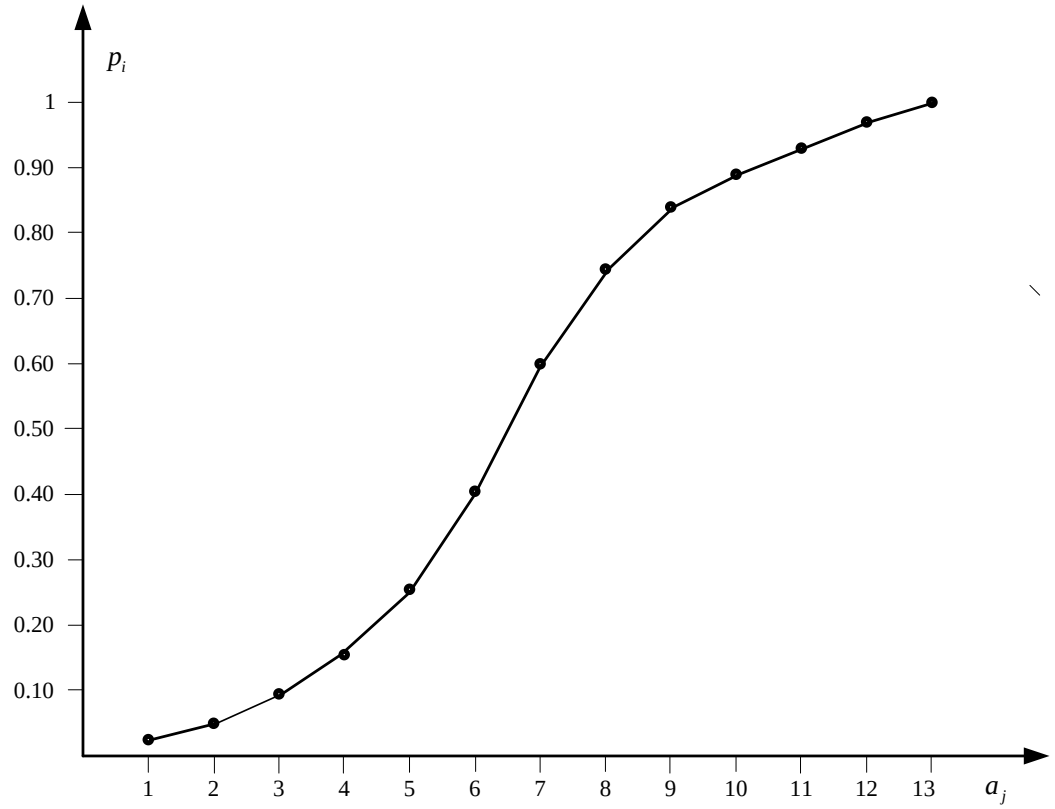
P3.2., P3.3. sadaļā tika izmantoti parametriskie testi sadalījumu normalitātes pārbaudei. P3.2. sadaļā šim nolūkam tika izmantoti pirmās kārtas sadalījumu parametri, P3.3. sadaļā – augstākās kārtas sadalījumu parametri.

Šajā sadaļā ir parādīti divi plaši izmantoti neparametriskie normalitātes testi: Kolmogorova-Smirnova tests (*Kolmogorov-Smirnov test*) un Šapiro-Vilksa tests (*Shapiro-Wilk test*).

1. *Kolmogorova-Smirnova tests*

Ja ir norādīts atribūtu a_j vērtību sadalījums, tad šim sadalījumam vienmēr var izveidot sadalījuma funkciju. P3.4.1. attēlā parādīts atribūtu a_j vērtību sadalījuma funkcijas grafiks datiem no P3.2.1. tabulas (P3.2.1. piemērs).

Grafiks tiek veidots šādi: uz horizontālās ass attēlotas atribūtu a_j vērtības (šajā piemērā tās ir vidējās atribūtu vērtības katrā intervālā). Vertikālā ass parāda to atribūtu vērtību varbūtības (biežumi), kas ietilpst atbilstošajos intervālos. Šeit ir runa par varbūtību uzkrātajām vērtībām.



P3.4.1. attēls. **Atribūtu a_j vērtību sadalījuma funkcijas grafiks datiem no P3.2.1. tabulas**

Pirmajam intervālam grafikā tiek attēlota tā varbūtības vērtība $p_1 = 0.024$. Šī vērtība ir atzīmēta ar punktu. Otrajam intervālam punkts parāda uzkrāto varbūtību $p_1 + p_2 = 0.024 + 0.024 = 0.048$ vērtības 1. un 2. intervālā. Trešais punkts parāda uzkrāto varbūtību vērtību $p_1 + p_2 + p_3 = 0.096$.

Pārējie punkti tiek noteikti līdzīgā veidā. Acīmredzams, ka pēdējā intervāla kumulatīvā varbūtības vērtība ir 1. Tas nozīmē, ka jebkura atribūta vērtība noteikti ietilpst tās vērtību diapazonā no 1 līdz 13.

Kolmogorova-Smirnova testa ideja ir šāda: atribūtu a_j vērtību diapazonā tiek konstruēta "ideālā" normālā sadalījuma funkcija. Šim normālajam sadalījumam ir tāda pati sagaidāmā vērtība un standarta novirze kā pārbaudāmajam empīriskajam sadalījumam. "Ideālā" sadalījuma funkcijas vērtības tiek salīdzinātas ar empīriskās sadalījuma funkcijas vērtībām dotajos punktos. Pēc tam testa statistiku aprēķina, izmantojot vienādojumu:

$$D = \sup_{a_j} \left| \left(F_0\left(a_{ij}\right) - F_e\left(a_{ij}\right) \right) \right|, \; i = 1, ..., n, \qquad \text{(P3.4.1)}$$

kur $F_0\left(a_{ij}\right)$ – "ideālā" sadalījuma funkcijas vērtība i punktā;

$F_e\left(a_{ij}\right)$ – empīriskā sadalījuma funkcijas vērtība i punktā;

n – datu punktu skaits (mūsu gadījumā vērtību intervālu skaits).

Citiem vārdiem sakot, vērtība D ir maksimālā atšķirība starp "ideālo" un empīrisko sadalījuma funkciju vērtību kādā i punktā.

Aprēķināto vērtību D salīdzina ar kritisko vērtību c. Ja $D < c$, nulles hipotēze tiek pieņemta: atribūtu a_j vērtības tiek sadalītas saskaņā ar normālo sadalījuma likumu. Ja $D > c$, nulles hipotēze tiek noraidīta, kas nozīmē, ka atribūtu a_j vērtības nav normāli sadalītas.

Kritiskās vērtības c ir tabulētas dažādiem nozīmīguma līmeņiem α. Šīs kritiskās vērtības ir norādītas P3.4.1. tabulā.

P3.4.1. tabula

Kolmogorova-Smirnova normalitātes testa kritiskās vērtības c

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.20$
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.382	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
>50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

Kolmogorova-Smirnova testa priekšrocības ir:

1) neatkarība no esošās empīriskās funkcijas sadalījuma;
2) rezultātu precizitāte;
3) izlases apjomam nav ierobežojumu.

Testa trūkumi ir:

1) tas attiecas tikai uz nepārtrauktiem sadalījumiem;
2) tests ir jūtīgāks pret vērtībām empīriskā sadalījuma centrā nekā pret vērtībām tā galos.

2. Šapiro-Vilksa tests

Šis tests ir balstīts uz empīriskās blīvuma funkcijas $f_e\left(a_{ij}\right)$ vērtību salīdzināšanu ar "ideālas" normālā blīvuma funkcijas vērtībām $f_0\left(a_{ij}\right)$, kas konstruētas noteiktā atribūtu a_j vērtību diapazonā. "Ideālajam" sadalījumam ir tāda pati sagaidāmā vērtība un standarta novirze kā empīriskajam sadalījumam.

Šapiro-Vilksa testa nulles hipotēze ir tāda, ka atribūta a_j sākotnējās vērtības ir izlase no populācijas, kurā šī atribūta vērtības ir normāli sadalītas.

Pārbaudes rezultātā tiek aprēķināta varbūtības p vērtība. Kritiskā vērtība ir $p = 0.5$. Ja aprēķinātā varbūtības vērtība ir $p < 0.05$, nulles hipotēze tiek noraidīta. Ja $p > 0.05$, tiek pieņemta nulles hipotēze, t. i., sākotnējās atribūtu a_j vērtības nāk no normālas populācijas.

Nosacījumu, ka atribūtu a_j vērtības nāk no normālas populācijas, var viegli apiet. Nelielas izmaiņas aprēķinu izteiksmēs ļauj tieši pārbaudīt sākotnējo sadalījumu bez atsauces uz pamatpopulāciju.

Tā kā abu testu praktiskai izpildei nepieciešams liels skaits aprēķinu, praksē šie testi tiek veikti, izmantojot piemērotus skaitļošanas rīkus, piemēram, statistikas pakotni SPSS. Lietojot P3.4. sadaļā aprakstītās procedūras, programma kopā ar izvades statistiku, balstoties uz Kolmogorova-Smirnova un Šapiro-Vilksa testu, rāda arī normalitātes sadalījuma pārbaudes rezultātus. Tomēr šie testi ir paredzēti sadalījumiem, kas satur lielu skaitu atbilstošu atribūtu vērtību. Ja sadalījumam ir neliels skaits atribūtu vērtību, testa rezultāti būs ļoti neprecīzi un neticami, tāpēc nav jēgas izmantot šos testus iepriekš sniegtajiem ilustratīvajiem piemēriem.