Statistiniai metodai medicinoje

JONĖ VENCLOVIENĖ

Statistiniai metodai medicinoje

Bendrasis vadovėlis aukštosioms mokykloms

Vytauto Didžiojo universitetas
Kaunas • 2010

UDK 311:61(075.8)
Ve-118

Recenzentai:
Prof. dr. A. Aksomaitis (KTU)
Prof. dr. J. Sapagovas (KMU)

Svarstyta ir rekomeduota spaudai Vytauto Didžiojo universiteto Gamtos mokslų fakulteto Aplinkotyros katedros posėdyje 2008-11-18; Gamtos mokslų fakulteto tarybos posėdyje 2008-11-19 (protokolo Nr. 2008/07, išrašas Nr. 07-17).

Lietuvos Respublikos švietimo ir mokslo ministerijos Bendrųjų vadovėlių leidybos komisijos rekomenduota 2008-11-21, Nr. 08-429.

Vadovėlio parengimą rėmė LVMSF (autorinės sutarties Nr. A-768).

ISBN 978-9955-12-558-7

Turinys

Pratarmė

Analizuodami surinktus duomenis, medikai-tyrėjai, biologai ir kitų biomedicinos mokslų atstovai tiria gausią informaciją. Surinktiems duomenims apdoroti taikomi ne tik standartiniai, bet ir šiuolaikiniai statistinės analizės metodai – logistinė regresija, išgyvenamumo analizė, daugiamačiai metodai. Vadovėlio tikslas – supažindinti medikus bei matematikus su šiuolaikiniais statistinės analizės metodais ir jų taikymu medicinos duomenims apdoroti. Vadovėlyje pateikti medicinos duomenų statistinės analizės pavyzdžiai padės specialistams geriau įsisavinti statistikos metodus ir jų taikymo subtilumus. Pateikti statistikos metodai taip pat gali būti taikomi ir kitų biomedicinos sričių – biologijos, biochemijos, ekologijos – duomenims apdoroti.

Vadovėlis skirtas medicinos mokslų krypties magistrantams, doktorantams ir moksliniams darbuotojams. Jis bus naudingas ir matematikams, taikantiems statistikos metodus biomedicinos duomenims apdoroti, bei kitų biomedicinos mokslų magistrantams, doktorantams ir mokslo darbuotojams.

Nuoširdžiai dėkoju VDU Aplinkotyros katedros ir Kardiologijos instituto Klinikinės kardiologijos laboratorijos mokslininkams už bendradarbiavimą aptariant statistinių metodų taikymą realių duomenų analizei. Ypač noriu padėkoti prof. A. Bikeliui, prof. A. Aksomaičiui, prof. J. Sapagovui, doc. R. Eidukevičiui už vertingas pastabas rankraščiui tobulinti.

Įvadas

Medikų, biologų bei epidemiologų tyrimo objektas – tam tikros populiacijos individai ir įvairių jų tyrimų duomenys. Ištirti visą populiaciją (arba matematinės statistikos terminu – generalinę visumą) dažnai neįmanoma, nes tam reikia daug laiko ir lėšų. Todėl dažniausiai tiriama tik populiacijos dalis, vadinama imtimi (*sample*). Ligonių imties sinonimas – ligonių kontingentas. Kad galėtume daryti išvadas apie populiaciją remdamiesi imties tyrimais, imtis turi reprezentuoti visą tiriamą populiaciją. Todėl individai, dalyvaujantys populiacijų tyrimuose, turi būti parinkti atsitiktinai. Imties reprezentatyvumas glaudžiai susijęs su dydžiu bei priklauso nuo sudarymo metodo (atrankos).

Parinkti atrankos schemą ir vertinti populiacijos būklę remiantis atrankos duomenimis – vienas statistikos uždavinių. Kai populiacija nėra labai didelė palyginti su imtimi, išvadoms apie populiaciją daryti naudojami atrankos iš baigtinių visumų metodai (Krapavickaitė, Plikusas, 2005). Šiame vadovėlyje daroma prielaida, kad tiriamos populiacijos individų skaičius yra labai didelis palyginti su imties dydžiu, todėl išvados apie populiaciją daromos statistiniais metodais, skirtais atrankoms iš begalinės populiacijos (visumos).

Praktikoje naudojami keli atsitiktinės atrankos būdai – paprasta atsitiktinė, stratifikuota, sisteminė bei klasterinė imtys. Šiame vadovėlyje apsiribosime paprastąja atsitiktine imtimi – atveju, kai iš visos populiacijos atsitiktinai atrenkamas tam tikras skaičius individų, nes duomenis, surinktus daugelio medikų tyrimų, galime laikyti gautais šia atranka. Pavyzdžiui, sergančiųjų infekciniu endokarditu (IE) komplikacijų priežastims tirti naudoti Kauno apskrities ligoninėse 1999–2001 m. nuo šios ligos gydytų 138 ligonių duomenys (Aržanauskienė, Zabiela, Jonkaitienė, 2002). Šiuos ligonius galima laikyti atsitiktine atranka iš ligonių, sergančių IE, populiacijos, nes į minėtas ligonines pateko dauguma sunkių IE sergančių ligonių.

Pasirinktos populiacijos medicininiams, biologiniams ar epidemiologiniams tyrimams atlikti pagal tyrimo tikslą formuluojami konkretūs uždaviniai. Jie sprendžiami keliais etapais: konkretizuojamas tiriamų ligonių kontingentas, tyrimo metodika, surenkama informacija apie tiriamų ligonių sveikatos rodiklius. Visa ši informacija perkeliama į duomenų bazę ir, panaudojus tam tikrus statistinius metodus, gaunamos išvados. Šią veiksmų seką galima pateikti tokia schema:

1. **Formuluoti tyrimo tikslą ir uždavinius.** Šiuo etapu apibrėžiamas tyrimo tikslas ir juo remiantis formuluojami konkretūs uždaviniai. Pavyzdžiui, mokslininkų grupės tyrimo tikslas – persirgusiųjų pirmuoju miokardo infarktu išgyvenamumo analizė. Minėtam tikslui pasiekti sprendžiami šie uždaviniai: (1) nustatyti rizikos veiksnius, didinančius mirties tikimybę; (2) stratifikuoti ligonius į mažos, vidutinės bei didelės rizikos sritis.

2. **Sudaryti duomenų rinkimo protokolą**. Šiuo etapu nustatoma, kokie medicininiai tyrimai (anamnezė, klinikiniai, kardiografiniai ar kiti) ir kokios ligonio charakteristikos bus naudojami tolesnei analizei. Duomenys parenkami remiantis anksčiau atliktų panašaus pobūdžio tyrimų rezultatais. Pavyzdžiui, žinoma, kad nepalankiai išeminės širdies ligos prognozei turi įtakos ligonio amžius, lytis, arterinė hipertenzija, cukrinis diabetas, cholesterolio kiekis, antsvoris, sumažėjęs fizinis aktyvumas. Todėl planuojant studiją, skirtą sergančiųjų išemine širdies liga išgyvenamumo analizei, be kitų ligonio rodiklių būtini duomenys apie šiuos rizikos veiksnius. Šiuo etapu sudaromos duomenų surinkimo anketos, protokolai ir t. t.

3. **Nustatyti ligonių atrankos kriterijus**. Nustatomas atsitiktinės imties sudarymo mechanizmas (atrankos schema) bei kriterijai, pagal kuriuos ligoniai įtraukiami ar neįtraukiami į studiją.

4. **Surinkti duomenis**.

5. **Sukurti duomenų bazę ir paruošti ją statistinei analizei bei duomenų grafiniam vaizdavimui**. Šiuo etapu surinkti anketų, ligos istorijų ar protokolų duomenys perkeliami į kompiuterį – sudaroma duomenų bazė. Duomenys vedami ar transformuojami taip, kad būtų galima atlikti statistinę analizę. Esant reikalui, apskaičiuojami ir išvestiniai rodikliai.

6. **Atlikti pirminę duomenų apžvalgą (žvalgomoji statistika).** Jos metu sudaromos dažnių ir tarpusavio dažnių lentelės, įvertinama kiekybinių rodiklių vidutinė reikšmė, išsibarstymas, skirstinio simetriškumas ir išskirtys. Dėl vaizdumo duomenų kitimas pateikiamas grafiškai.

7. **Tvarkyti duomenis ir sudaryti statistinį modelį.** Atsižvelgiant į pirminės apžvalgos rezultatus, duomenys atitinkamai pertvarkomi. Pavyzdžiui, jei pagal ligos sunkumą (pvz., CD laipsnį) ar kitą rodiklį ligoniai suskirstyti į kelias grupes ir vienoje grupėje jų yra tik keli, jie pergrupuojami; tarkime, vietoj trijų ligonių grupių (neserga CD, serga I^{o} CD, serga II^{o} CD) naudojamos dvi (neserga CD, serga CD). Matematinėje statistikoje individo rodiklio modelis – atsitiktinis dydis su tam tikru skirstiniu. Remiantis pirmine duomenų apžvalga, daromos išvados ir apie nustatyto rodiklio skirstinį ar reikiamą transformavimą. Pavyzdžiui, kai kurių kiekybinių rodiklių reikš-

mės dėl asimetriškumo logaritmuojamos. Taip pat nustatomos statistinei analizei nenaudotinos rodiklių reikšmės – neteisingai nustatytos ar klaidingai įvestos į duomenų bazę (pvz., cholesterolio reikšmė 0 ar 100 (mmol/l) tolesnei analizei nenaudojama).

8. **Taikyti statistinius metodus.** Statistiniai metodai parenkami pagal iškeltus uždavinius ir duomenų statistinį modelį.

9. **Pateikti išvadas.** Remiantis statistinių metodų taikymo rezultatais, daromi atitinkami sprendimai ir pateikiamos išvados.

Statistiniai metodai padeda klinicistams ar tyrėjams suprasti ir paaiškinti medicininių duomenų kitimą bei gauti informaciją, reikalingą ligoniams sėkmingiau gydyti. Trumpai apžvelgsime knygoje pateiktus statistinius metodus, skirtus medikų bei gimininguų sričių specialistų duomenų analizei. Pagal tyrėjų sprendžiamus uždavinius parenkami ir statistiniai metodai. Medikai, biologai bei epidemiologai, spręsdami savo srities problemas, dažniausiai susiduria su tokiais uždaviniais:

- populiacijos sveikatos būklės (sergamumo, mirtingumo) analizė, rizikos veiksnių populiacijoje paplitimo tyrimai;
- veiksnio (susirgimo, gydymo metodo) įtakos individo sveikatos būklei vertinimas;
- ligonio būklę charakterizuojančių rodiklių tarpusavio ryšio nustatymas bei vertinimas;
- rodiklių, nustatytų įvairiais tyrimais, modeliavimas;
- nepalankios išeities rizikos vertinimas; ligonių stratifikavimas į skirtingos rizikos grupes;
- ligonių išgyvenamumo analizė.

Analizuojant populiacijos sveikatos būklę, įvertinama nepalankaus įvykio (susirgimo, mirties) tikimybė su pasikliautinaisiais intervalais. Analizuojant kiekybinio rodiklio kitimą, skaičiuojamas vidurkis, jo standartinė paklaida ar pasikliautinieji intervalai bei kitos skaitinės charakteristikos (1, 4 skyrius); tikrinama hipotezė apie populiacijos vidurkio ar tikimybės lygybę normai (5.5, 6.1–6.2 skyriai).

Vertindami gydymo ar susirgimo įtaką individo sveikatos būklei, lyginame dviejų populiacijų skirstinio vidurkius (6.4 skyrius) ar skirstinius (6.8, 7.4–7.5 skyriai), dviejų ar daugiau kartotinių matavimų vidurkius (6.5, 6.10, 7.6 skyriai), kelių populiacijų skirstinio vidurkius (6.8, 12 skyriai).

Individo būklę charakterizuojančių kiekybinių rodiklių tarpusavio ryšiui vertinti skirta koreliacinė analizė (9 skyrius), o kokybinių rodiklių – kokybinių kintamųjų ryšio matai (7.7–7.8 skyriai).

Individo kiekybiniams rodikliams modeliuoti naudojama regresinė analizė (10 skyrius), dvinariams – logistinė regresinė analizė (11 skyrius).

Rizikai vertinti skirtas 8 skyrius. Ligoniams stratifikuoti į skirtingos rizikos grupes naudojama logistinė regresija (11.6 skyrius), Kokso proporcinga regresija (13.9–13.10 skyriai) bei diskriminantinė analizė (15.4–15.7 skyriai).

Ligonių išgyvenamumo analizė aptariama 13 skyriuje.

Kauno medicinos universiteto Fizikos ir matematikos katedros dėstytojų išleistoje mokomojoje knygoje „Statistikos ir informatikos pagrindai“[1] pateikta daug uždavinių, kuriais iliustruotas 1, 4–6, 9–12 skyriuose pateiktų statistikos metodų taikymas medicinos duomenų analizėje.

Ženklu * pažymėtiems skyriams skaityti reikia gilesnio matematinio pasiruošimo. Šie skyriai skirti matematikams, taikantiems statistikos metodus tiek medicinoje, tiek kituose biomedicinos moksluose.

[1] Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. Statistikos ir informatikos pagrindai, 2008, KMU leidykla, Kaunas, p. 98.

Terminų trumpinimai

Bendrų medicinos terminų trumpinimai:

AH – arterinė hipertenzija;
BC – bendrasis cholesterolis;
CD – cukrinis diabetas;
CV – kardiovaskulinis įvykis;
HPT – hiperparatiroidizmas;
KMI – kūno masės indeksas;
SAS – sistolinis arterinis kraujospūdis;
DAS – diastolinis arterinis kraujospūdis;
ŠSD – širdies skilvelių susitraukumo dažnis;
IŠL – išeminė širdies liga;
KA – krūtinės angina;
NKA – nestabili krūtinės angina;
UŠN – ūmus širdies nepakankamumas;
MI – miokardo infarktas;
ŪKS – ūmūs koronariniai sindromai;
PTCA – miokardo revaskuliarizacija.

Echoskopijos rodiklių trumpinimai:

KSGDD – kairiojo skilvelio galinis diastolinis dydis;
KSMI – kairiojo prieširdžio masės indeksas;
KPR – kairiojo prieširdžio dydis;
DPR – dešiniojo prieširdžio dydis;
USS – užpakalinės sienelės storis;
IF – išstūmimo frakcija;
DF – diastolinė funkcija.

Statistikos ir epidemiologijos terminų trumpinimai:

RV – rizikos veiksnys;
RR – santykinė rizika;
OR – rizikos santykis;
PR – proporcinga rizika;
Ats. d. – atsitiktinis dydis;
Ats. v. – atsitiktinis vektorius;
SE – standartinė paklaida;
PI – pasikliautinasis intervalas;
AIC – Akaike informacijos kriterijus;
BIC – Bajeso informacijos kriterijus;
SST – visa kvadratų suma;
ATM – apibendrinti tiesiniai modeliai;
ANOVA – dispersinė analizė;
ANCOVA – kovariancinė analizė;
MANOVA – daugiamatė dispersinė analizė.

1 SKYRIUS

Pirminė duomenų statistinė analizė

Atlikdami klinikinius ar epidemiologinius tyrimus, medikai susiduria su gausybe ligonio būklę charakterizuojančių duomenų. Visi šie duomenys kintantys – jie priklauso ne tik nuo to, ar ligoniui nustatytas susirgimas, ar ne, bet ir nuo individo biologinės būklės, jo aplinkos, gyvensenos kokybės, gretutinių susirgimų bei nuo įvairių kitų priežasčių, kurias galima apibūdinti atsitiktiniu faktoriumi.

Pirminė statistinė analizė skirta medicininių stebėjimų kitimui aprašyti. Remiantis jos rezultatais, daromos išvados apie tolesnę tyrimų eigą bei sudėtingesnių statistinių metodų taikymą.

1.1. Medicinos duomenys. Kintamojo sąvoka

Tiek pirminės statistinės analizės, tiek sudėtingesnių statistinių metodų taikymas duomenims apdoroti pirmiausiai priklauso nuo to, kokios charakteristikos aprašomos – kiekybinės ar kokybinės. Pavyzdžiui, sistolinis ir diastolinis arteriniai kraujospūdžiai (SAS ir DAS), širdies skilvelių susitraukimo dažnis (ŠSD) yra kiekybinės ligonio būklės charakteristikos. Kokybines individo charakteristikas apibūdina lytis, rasė, kraujo grupė, susirgimas (nėra, yra), skausmo lokalizacija, AH laipsnis ir kiti.

Kiekybiniai žymenys dažnai atspindi kokybinius organizmo pokyčius. Pavyzdžiui, esant pastoviam SAS>140 ar DAS>90, pasireiškia kokybiniai organizmo pokyčiai, todėl ligoniui diagnozuojama arterinė hipertenzija. Jei kūno masės indeksas (KMI) viršija 30, ligoniui nustatomas nutukimas. Jei bendrojo cholesterolio koncentracija nuolat viršija 5,2 mmol/l, pasireiškia vainikinių arterijų stenozės požymiai ir diagnozuojama hipercholesterinemija.

Atskirų individo charakteristikų (pvz., kraujospūdžio) rodikliai (sistolinis kraujospūdis mmHg st., diastolinis kraujospūdis, mmHg st.) statistinėje literatūroje vadinami kintamaisiais (*variable*). Literatūroje dažnai pateikiami kintamojo sinonimai – žymuo, rodmuo.

Atlikus tyrimus bei apklausus ligonį, nustatomos konkrečios tirtų kintamųjų reikšmės. Objektyvaus tyrimo metu registruojami kintamieji: svoris, ūgis, širdies ūžesiai (nėra, yra), ŠSD, SAS, DAS, karkalai plaučiuose (nėra, sausi, drėgni, mišrūs), kepenų būklė, kojų tinimas (nėra, yra) ir kt. Registruojami ir kiti ligonio nusiskundimų kintamieji: skausmai širdies plote (nėra, yra), skausmo lokalizacija (už krūtinkaulio, kairėje krūtinės pusėje, dešinėje krūtinės pusėje, visoje krūtinėje, kitur), skausmo intensyvumas (silpnas, stiprus, labai stiprus). Nustatomi anamnezės duomenų kintamieji: miokardo infarktas (MI) (nebuvo, buvo), MI rūšis (Q bangos, be Q bangos, reinfarktas, įtariamas), cukrinis diabetas (CD) (nėra, yra), arterinės hipertenzijos laipsnis ir t. t.

1.2. Kintamųjų tipai

Kintamuosius sąlygiškai galima priskirti vienam šių tipų: nominaliojo, dvinario (dichotominio), tvarkos, kiekybinio (tolydžiojo ar diskrečiojo), eiliškumo ir laiko eilučių.

Nominalieji kintamieji – nematuojami kintamieji, arba kategorijos, kiekybiškai tarpusavyje nepalyginami. Tai lytis, kraujo grupės (O, A, B ir AB), akių ar odos spalva, susirgimai, maisto grupės. Nominaliojo kintamojo reikšmių negalima sutvarkyti didėjimo ar mažėjimo tvarka ar kiekybiškai palyginti tarpusavyje.

Jau minėta, kad kintamasis „odos spalva“ yra nominalusis. Tačiau jis neatskiria normalios ir nenormalios odos spalvos, nors tai svarbu vertinant ligonio sveikatos būklę. Todėl tyrėjui pravartu sudaryti naują kintamąjį, kuris įgytų tik dvi reikšmes – normali odos spalva (pvz., koduojama 1) ir nenormali odos spalva (koduojama 2). **Dvinariai kintamieji** – tai kintamieji, įgyjantys tik dvi reikšmes (yra susirgimas arba jo nėra, 1 arba 2 ir t. t.). Juos patogu vartoti kalbant apie požymio buvimą ir nebuvimą, taip pat apie patekimą ir nepatekimą į tam tikrą aibę. Dažnai konstatuojama, kad individas serga (koduojama 1) arba neserga (koduojama 0) širdies liga, jo amžius daugiau kaip 65 metai arba ne, kepenys nepadidėjusios (koduojama 0) arba padidėjusios (koduojama 1).

Daugelis medicinos duomenų apibūdinami daugiau nei dviem reikšmėmis ir yra žinoma, kuri reikšmė geresnė, kuri blogesnė. Šie duomenys apibūdi-

nami tvarkos (ordinariaisiais) kintamaisiais. **Tvarkos (ordinarieji) kintamieji** įgyja diskrečias kiekybiškai palyginamas reikšmes. Tvarkos kintamųjų pavyzdžiai: AH laipsnis (1, 2, 3), UŠN klasės (1–4), skausmo intensyvumas (silpnas, vidutinis, stiprus, labai stiprus).

Nominalieji, dvinariai ir tvarkos kintamieji yra kokybiniai (kategorizuoti).

Kiekybiniai kintamieji – tai kintamieji, kurie kinta tolydžiojoje ar diskrečiojoje matavimo skalėje. Tolydieji kintamieji įgyja reikšmes neskaičioje skalėje (įgyjamų reikšmių sunumeruoti negalima), diskretieji – skaičioje (įgyjamas reikšmes galima sunumeruoti). Daugelis medicinos duomenų yra tolydieji: svoris, ūgis, SAS, DAS, BC koncentracija ir daugelis kitų. Diskrečiųjų kintamųjų pavyzdžiai: širdies skilvelių susitraukimo dažnis per minutę, per parą hospitalizuotų ligonių skaičius. Faktiškai matuodami gauname tik diskrečiuosius kintamuosius, nes tolydžiųjų kintamųjų reikšmių aibė yra skaičioji dėl matavimo prietaisų (svarstyklių, kraujospūdžio matuoklių) tikslumo – ūgis matuojamas centimetrais, arterinis kraujospūdis – mmHg. Kiekybiniai kintamieji gali įgyti tiek teigiamas, tiek neigiamas reikšmes. Dauguma medikų nustatomų kintamųjų reikšmių yra neneigiamos. Nemažai kintamųjų įgyja tik teigiamas reikšmes – ūgis, svoris, organų matmenys.

Analizuojant letalios baigties ar komplikacijos rizikos veiksnius, kartais vietoj kiekybinio kintamojo, tarkime, amžiaus, patogiau naudoti tvarkos – amžiaus grupę, arba dvinarį (amžius > 65 m. ir amžius ≤ 65 m.) kintamąjį. Vietoj susirgimo trukmės taip pat vartojamas tvarkos kintamasis: serga < 5 m., serga 5–10 metų, serga > 10 m. Kiekybinis kintamasis suteikia daugiau informacijos negu tvarkos, nes pagal kiekybinio kintamojo reikšmes galima palyginti bet kuriuos du individus, o pagal tvarkos – ne. Tačiau tvarkos kintamojo reikšmes patogiau interpretuoti.

Eiliškumo kintamieji vartojami dėl patogumo. Tai tirtų asmenų kodai, numeriai, pavardės ar kiti identifikatoriai. Matematiškai juos apdoroti nėra prasmės.

Medicinoje ar biologijoje taip pat analizuojamas rodiklio kitimas erdve ar laiku. Tai **laiko eilutės** – rodiklio, kintančio laiku ar erdve, reikšmių, išmatuotų vienodais laiko (erdvės) tarpais, sekos. Elektrokardiogramos signalo reikšmės, kasmetinės sergamumo ar mirtingumo rodiklių sekos, ligonio kasdienės medicininių ar biocheminių rodiklių matavimo sekos yra laiko eilutės.

Daugiamačiai kintamieji. Dažnai ligonio būklę charakterizuoja ne vienas, o keli kintamieji. Pavyzdžiui, naujagimio būklę nusako 5 tvarkos kintamieji, įgyjantys reikšmes 0, 1 ir 2: ŠSD, kvėpavimas, odos spalva, raumenų tonai,

refleksai. Vektorius (ŠSD, kvėpavimas, odos spalva, raumenų tonai, refleksai) bus vadinamas daugiamačiu (konkrečiau, penkiamačiu) kintamuoju. Pagal TNN sistemą vėžys charakterizuojamas trimačiu vektoriumi (auglys, mazgai, metastazės); vienmačiai kintamieji – auglys, mazgai, metastazės – vadinami šio vektoriaus komponentėmis (koordinatėmis).

Daugiamačio kintamojo komponenčių suminiam poveikiui ligonio būklei nustatyti įvedamas naujas kintamasis – indeksas (klasė, balas) (*indexes, factors, stages, systems, clases, scales, ratings, criteria*). Indekso reikšmės nustatomos pagal tam tikrą formulę, naudojant daugiamačio vektoriaus komponenčių reikšmes. Indeksas gali būti tiek tvarkos, tiek kiekybinis. Pavyzdžiui, naujagimio būklei apibūdinti naudojamas *Apgar* indeksas. Jis lygus visų penkių komponenčių sumai. *Apgar* indeksas kinta nuo 0 iki 10; jį geriau analizuoti kaip kiekybinį kintamąjį.

Remiantis trimačiu vektoriumi, charakterizuojančiu vėžio būklę, įvedamas naujas kintamasis – vėžio stadija, įgyjantis reikšmes I, II, III, IV. Tai tvarkos kintamasis.

Daugiamatį kintamąjį ir pagal jį sudarytą indeksą formaliai galima apibrėžti taip: turime k kintamųjų $X_1, X_2 \ldots X_k$. Vektorius $\boldsymbol{X} = (X_1, X_2 \ldots X_k)$ vadinamas daugiamačiu kintamuoju; $X_1, X_2 \ldots X_k$ – vektoriaus $\boldsymbol{X}$ komponentės (koordinatės). Viena formulių, nusakančių indekso Y apibrėžimą, yra:

$$Y = X_1 + X_2 + \ldots + X_k.$$

Vektoriaus $\boldsymbol{X}$ koordinatės gali būti ne tik atskiri kintamieji, bet ir to paties kintamojo, nustatyto tam tikrais laiko tarpais, reikšmės.

1.3. Duomenų paruošimas statistinei analizei bei grafiniam vaizdavimui

Daugelio ligonių tyrimų duomenų analizuoti be kompiuterio negalima net elementariai. Duomenų analizei ir grafiniam pateikimui būtina naudoti skaičiuoklę EXCEL arba statistinius programų paketus. Tarp medikų populiariausi STATISTICA ir SPSS statistikos paketai. Epidemiologinių duomenų analizei skirtas EPIINFO programų paketas.

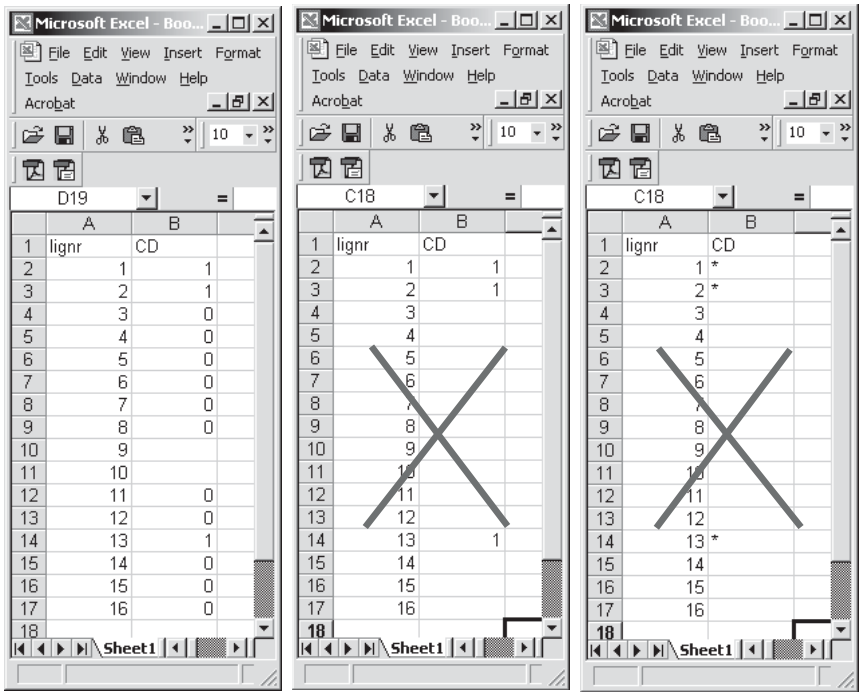
Kiekviename programų pakete naudojamos atitinkamo tipo duomenų bylos. Skaičiuoklėje EXCEL sudarytos bylos yra *.xls tipo. Pakete STATISTICA naudojamos *.STA, pakete SPSS – *.sav tipo duomenų bylos. STATISTICA ir SPSS paketams reikalingus duomenis galima perkelti (importuoti) iš *.xls bylų ir atvirkščiai. Todėl medikams paprasčiausia ir priimtiniausia duomenis kaupti EXCEL tipo bylose.

Statistinei analizei skirtus duomenis reikia tinkamai perkelti į kompiuterį, t. y. turi būti sudaryta duomenų bazė. Suvedant duomenis į bet kokio paketo bylą, kiekvienam kintamajam skiriamas stulpelis, kiekvienam individui – eilutė. Kintamojo pavadinimas (vardas) EXCEL skaičiuoklėje pateikiamas pirmoje eilutėje. SPSS ir STATISTICA paketuose kintamojo vardas (bei kintamojo tipas, formatas ir t. t.) fiksuojamas informacijai apie kintamąjį skirtame langelyje. Kiekybinio kintamojo reikšmės suvedamos tokios, kokios yra nustatytos – sveiki ar trupmeniniai skaičiai. Pavyzdžiui, išmatavus bendrojo cholesterolio koncentraciją, gauti dydžiai 5,2; 6,1; 4,9. Į kompiuterio bylą galima vesti šiuos skaičius arba be kablelio jiems proporcingus: 52; 61; 49. Pastaruoju atveju visuomet galima perskaičiuoti ir gauti tikrąsias reikšmes.

Nominaliojo kintamojo reikšmė gali būti koduojama skaičiais arba pateikta pavadinimu. Pavyzdžiui, įrašydami individo lytį, vyrus galime koduoti 1, moteris 2 arba V ir M, *male* ir *female*. Būtina stebėti, kad vienodos kintamojo reikšmės būtų koduojamos tais pačiais pavadinimais. Užrašus *Benigma* ir *Benigma.* kompiuteris laikys skirtingais ir dėl to apdorojant duomenis kils sunkumų.

Tvarkos kintamojo reikšmės koduojamos skaičiais, geriausia – didėjimo tvarka. Pavyzdžiui, kintamojo reikšmes *skausmas silpnas*, *vidutinis*, *stiprus* patogiausia koduoti 1, 2, 3 arba panašiai.

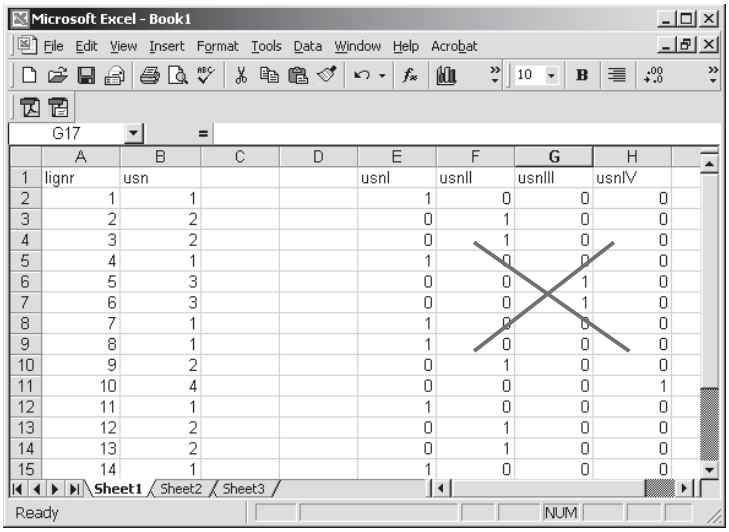
Statistiniuose paketuose susiduriame su terminu **praleista reikšmė** (*missing value*) – tai nenustatyta reikšmė. Vedant duomenis, nenustatytos reikšmės vietoje dažniausiai paliekamas tuščias langelis. Kartais (tai sukelia tam tikrų nepatogumų apdorojant duomenis) vietoj nežinomos reikšmės įrašomas tam tikras kodas (praleistos reikšmės kodas), ypač besiskiriantis nuo kitų kintamojo reikšmių, pavyzdžiui 0 ar –9999. Praleistos reikšmės kodas nurodomas statistiniame pakete aprašant kintamąjį. Todėl suvedant duomenis tuščias langelis paliekamas tik tada, jei nežinoma, kokia šio individo kintamojo reikšmė. Pavyzdžiui, reikalinga informacija apie sergamumą CD. Tam naudojamas kintamasis, lygus 1 (ligonis serga CD) ir 0 (jei CD neserga). Į duomenų bylą įrašoma 1, jei ligonis serga CD, 0 rašomas, jei neserga CD; jei nežinoma, ar ligonis serga CD, paliekamas tuščias langelis (1.1 pav.). Jei ne visiems duomenų bazės ligoniams atliktas koronarografinis tyrimas, įrašydami koronarų pažeidimus, rašome nulį, jei kraujagyslė nėra pažeista, bet nepaliekame tuščio langelio.



1.1 pav. Sergamumo CD surašymas į EXCEL tipo bylą

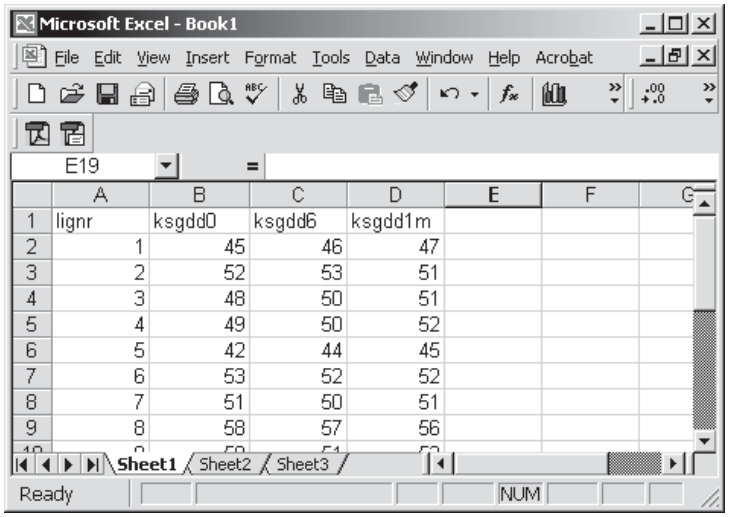
Surašant tvarkos kintamojo reikšmes, kartais vietoj vieno stulpelio rašoma į kelis. Pavyzdžiui, vedant UŠN klasės reikšmes, į atitinkamą stupelį rašoma klasė (1.2 pav., usn), bet nenaudojamai keli stulpeliai (1 klasė, 2 klasė ... 2 pav., usnI, usnII ...). Tvarkos kintamojo reikšmes rašyti keliais stulpeliais nėra klaidinga, tačiau taip suvestus duomenis reikia pertvarkyti, kad būtų galima analizuoti toliau, naudojantis statistiniu paketu.

Surašant kelių imčių duomenis, pavyzdžiui, tiriamos ir kontrolinės ligonių grupės, papildomas stulpelis skiriamas grupei identifikuoti. Kitaip tariant, įvedamas grupės kintamasis. Pavyzdžiui, kodas 1 – tiriama grupė, 0 – kontrolinė grupė. Kitaip suvedus duomenis, atsiranda sunkumų statistinei analizei atlikti. Pakartotiniai tyrimai rašomi į atskirus stulpelius. Pavyzdžiui, KSGDD – patekus į stacionarą, po 6 mėn., po metų – rašomi į atskirus stulpelius (1.3 pav., ksgdd0, ksgdd6, ksgdd1m kintamieji).



	A	B	C	D	E	F	G	H
1	lignr	usn			usnI	usnII	usnIII	usnIV
2	1	1			1	0	0	0
3	2	2			0	1	0	0
4	3	2			0	1	0	0
5	4	1			1	0	0	0
6	5	3			0	0	1	0
7	6	3			0	0	1	0
8	7	1			1	0	0	0
9	8	1			1	0	0	0
10	9	2			0	1	0	0
11	10	4			0	0	0	1
12	11	1			1	0	0	0
13	12	2			0	1	0	0
14	13	2			0	1	0	0
15	14	1			1	0	0	0

1.2 pav. UŠN klasė, suvesta į EXCEL tipo bylą



	A	B	C	D	E	F	G
1	lignr	ksgdd0	ksgdd6	ksgdd1m			
2	1	45	46	47			
3	2	52	53	51			
4	3	48	50	51			
5	4	49	50	52			
6	5	42	44	45			
7	6	53	52	52			
8	7	51	50	51			
9	8	58	57	56			

1.3 pav. Pakartotinis KSGDD tyrimas, suvestas į EXCEL tipo bylą

1.4. Pirminė kokybinio kintamojo analizė

Kokybinis kintamasis įgyja nedaug reikšmių palyginti su kiekybiniu. Šį kintamąjį visiškai apibūdinsime išvardiję jo įgyjamas reikšmes ir nurodę, kaip dažna imtyje yra atitinkama reikšmė. Todėl nominaliojo, dvinario ar tvarkos kintamojo įgyjamų reikšmių pasiskirstymas imtyje (duomenyse)

pateikiamas dažnių lentele. Joje nurodomi atvejų, kai kintamasis įgyja i-tąją reikšmę, skaičius imtyje n_i ir (arba) santykiniai i-tosios reikšmės dažniai $f_i = n_i/n$ (dalimis) ar procentais ($(n_i/n)\times 100$) (1.1 lentelė). Čia $n = n_1 + n_2 + ... + n_k$ – tirtų individų skaičius (imties dydis), k – kintamojo įgyjamų reikšmių (kategorijų) skaičius. Grafiškai ši informacija pateikiama stulpelio, skritulio ar kitokia diagrama (1.4–1.5 pav.).

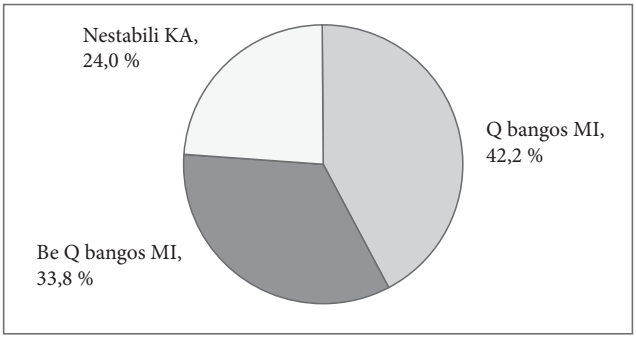
1.1 lentelė. Kokybinio kintamojo, įgyjančio k reikšmių, absoliučių ir santykinių (dalimis bei procentais) dažnių pasiskirstymas

Kintamojo reikšmės numeris	Absoliutūs dažniai (N)	Santykiniai dažniai	Procentai
1	n_1	$f_1 = n_1/n$	$(n_1/n)\times 100$
2	n_2	$f_2 = n_2/n$	$(n_2/n)\times 100$
...	...	...	...
k	n_k	$f_k = n_k/n$	$(n_k/n)\times 100$

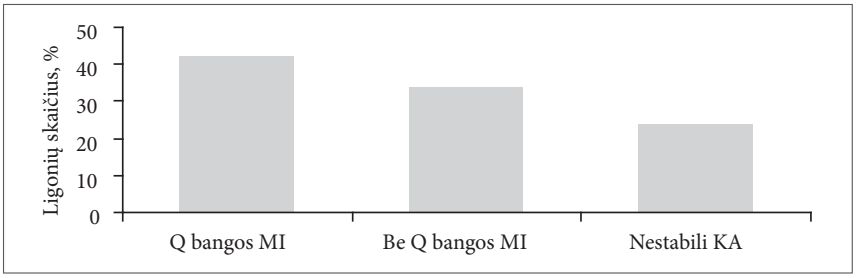
1.2 lentelėje pateikiame ligonių, sirgusių ūmiais koronariniais sindromais (ŪKS), kontingento pasiskirstymą pagal koronarinių sindromų pobūdį (duomenys paimti iš [2]). Ši informacija taip pat pateikta stulpelio bei skritulio diagrama (1.4, 1.5 pav.).

1.2 lentelė. Ligonių, sirgusių ŪKS, kontingento pasiskirstymas pagal koronarinių sindromų pobūdį

Koronariniai sindromai	N	Procentai
Q bangos MI	272	42,17
Be Q bangos MI	218	33,80
Nestabili KA	155	24,03



1.4 pav. Ligonių, sirgusių ŪKS, kontingento pasiskirstymo pagal koronarinių sindromų pobūdį skritulinė diagrama



1.5 pav. Ligonių, sirgusių ŪKS, kontingento pasiskirstymo pagal koronarinių sindromų pobūdį stulpelinė diagrama

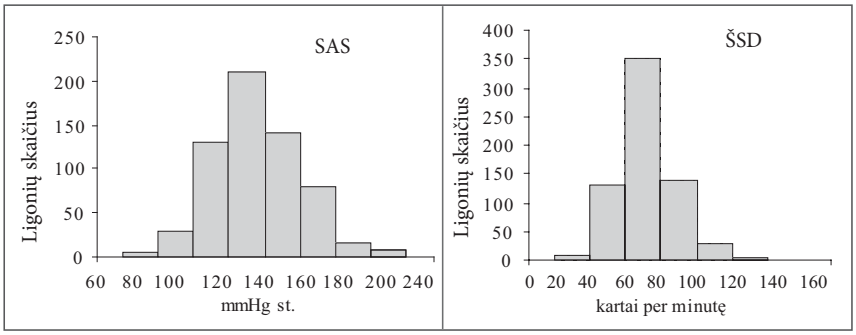
1.5. Pirminė kiekybinio kintamojo analizė

Kiekybinis kintamasis (rodiklis), pavyzdžiui, amžius, cholesterolio koncentracija kraujyje, įgyja daug skirtingų reikšmių, be to, ir gretimos reikšmės tarpusavyje gali labai mažai skirtis. Taigi dažnių lentelė nėra informatyvi. Todėl kiekybinio kintamojo reikšmės sugrupuojamos, po to pateikiama sugrupuotų duomenų dažnių (dažnių skirstinio) lentelė. Grupavimo intervalai paprastai parenkami vienodo pločio, nebent išskyrus pirmą ir paskutinį. Intervalų pločiai bei skaičius nustatomi atsižvelgiant į imties dydį, didžiausią ir mažiausią reikšmes bei kintamojo medicininę ar biologinę prasmę. Pavyzdžiui, ligonių kontingento amžiaus pasiskirstymui pateikti, amžius grupuojamas kas 5 ar 10 metų (1.3 lentelė). Grupuodami duomenis į 4 ar mažiau intervalų, gauname neinformatyvų dažnių skirstinį. Taip pat rekomenduotina naudoti ne daugiau kaip 15–20 grupavimo intervalų bei grupavimo intervalus parinkti taip, kad beveik visuose būtų bent po 5 reikšmes.

1.3 lentelė. Ligonių, sirgusių ūmiais koronariniais sindromais, kontingento pasiskirstymas pagal amžių

Amžiaus grupės	N	Procentai
< 30 m.	5	0,78
30–39	20	3,12
40–49	92	14,33
50–59	207	32,24
60–69	229	35,67
70–79	86	13,40
≥ 80	3	0,47
Iš viso	642	100

Kiekybinio kintamojo dažnių skirstinio grafinis vaizdas yra histograma. Ji braižoma taip: kiekvienas grupavimo intervalas atidedamas Ox ašyje, po to brėžiamas stačiakampis taip, kad pagrindas sutaptų su grupavimo intervalu, o aukštis būtų lygus arba proporcingas į jį patekusių kintamojo reikšmių skaičiui. Grupavimo intervalai turi būti vienodo pločio, nebent išskyrus pirmą ir paskutinį. 1.6 pav. pateiktos ligonių, sergančių ŪKS, SAS ir ŠSD histogramos ([2]).



1.6 pav. Ligonių, sirgusių ūmiais koronariniais sindromais, SAS ir ŠSD histogramos

Braižant histogramą, grupavimo intervalų skaičių k rekomenduojama imti: $k = 1+3{,}222\times\log_{10} n$. Tuomet grupavimo intervalo plotis apskaičiuojamas taip: $h = (x_{max} - x_{min})/(1+3{,}222\times\log_{10} n)$ (jei visi intervalai nustatomi vienodo pločio). Kai kurie autoriai intervalo plotį rekomenduoja imti lygų $(x_{max} - x_{min})\ (24\sqrt{\pi}\,/\,n)^{1/3}$ [5]. Tačiau grupavimo intervalų dydis bei skaičius parinktini atsižvelgiant ir į rodiklio medicininę ar biologinę prasmę. Pavyzdžiui, amžius grupuojamas kas 5 ar 10 metų, SAS, SŠD – kas 10 ar 20 vienetų.

Tačiau remiantis dažnių skirstinio lentele, sunku apibūdinti kintamąjį ar tarpusavyje palyginti keletą kintamųjų. Pavyzdžiui, analizuojama jaunų sveikų individų arterinio kraujospūdžio rodikliai – SAS ir DAS. Pagrindiniai jų skirtumai yra:

1) skirtingas dydis – SAS beveik dvigubai didesnis už DAS;
2) skirtingas kitimas: SAS reikšmės kinta nuo 90 iki 140, DAS – nuo 60 iki 80.

Bet kurį kiekybinį kintamąjį, kaip SAS ir DAS, glaustai apibūdina „centras“ – taškas, apie kurį išsidėsčiusios reikšmės, bei sklaidos apie šį „centrą“ įvertis. Tai patvirtina ir 1.6 pav. pateiktos histogramos. Iš jų matome, kad tirto ligonių kontingento dažniausiai pasitaikančios ŠSD reikšmės yra tarp

60–80 kartai/min. Analogiška situacija matyti ir SAS histogramoje: kintamojo reikšmės daugiau ar mažiau išsidėsčiusios apie imties „centrą“. Todėl reikalingos šio centro, išsidėstymo apie jį bei histogramos formos (dažnių skirstinio) skaitinės charakteristikos.

1.6. Kiekybinio kintamojo skaitinės charakteristikos

Kiekybinio kintamojo imties skaitinės charakteristikos skirstomos į šias grupes:

1) duomenų padėties (*location*), arba centro, charakteristikos (vidurkis, mediana, moda, geometrinis vidurkis, harmoninis vidurkis);

2) duomenų kitimo, arba sklaidos apie centrą, charakteristikos (dispersija, standartinis nuokrypis, kvartiliai, procentiliai, imties plotis, interkvartilinis plotis);

3) duomenų skirstinio (histogramos) formos charakteristikos (asimetrijos, ekscesо koeficientas);

4) išskirčių (ypač besiskiriančių ar blogai įvestų reikšmių, „šiukšlių“) charakteristikos (maksimali ir minimali imties reikšmė, stačiakampės diagramos, interkvartilinis plotis). Išskirtys – tai reikšmės, nesuderintos su visa duomenų mase, pavyzdžiui, kairiojo prieširdžio dydis negali būti 0 ar 1. Išskirtis apibrėžiama taip: tai imties reikšmė, nesiderinanti su statistiniu duomenų modeliu (žr. 3 skyrių).

Pateiksime visų minėtų charakteristikų apibrėžimus. Pažymėkime: x_1, x_2 ... x_n – tiriamo kintamojo imties skaitinės reikšmės; čia n – reikšmių skaičius imtyje. Skaičius n vadinamas imties dydžiu (*sample size*). Nemažėjimo tvarka išdėstyta dydžių x_1, x_2 ... x_n eilutė vadinama variacine seka ir žymima $x_{(1)}$, $x_{(2)}$ ... $x_{(n)}$. Pavyzdžiui, turime imtį: 3,1; 2,9; 3; 3,2; 2,8. Taigi $n = 5$, $x_1 = 3{,}1$; $x_2 = 2{,}9$; $x_3 = 3{,}0$; $x_4 = 3{,}2$; $x_5 = 2{,}8$; šios imties variacinė seka yra $x_{(1)} = 2{,}8$; $x_{(2)} = 2{,}9$; $x_{(3)} = 3{,}0$; $x_{(4)} = 3{,}1$; $x_{(5)} = 3{,}2$.

Duomenų centro įverčiai. Vidurkis (*mean*, *average*) $\overline{x}$:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Tai imties „masės centras“ – išmatuotų reikšmių aritmetinis vidurkis. Sumos ženklas Σ reiškia reikšmių x_i sumavimą. $\overline{x}$ yra nežinomo populiacijos vidurkio įvertis. Vidurkis labai jautrus išskirtims: pavyzdžiui, klaidingai įvesta labai didelė arba labai maža reikšmė žymiai pakeičia vidurkį. Tarkime, 5 bendrojo cholesterolio koncentracijos reikšmių (mmol/l) 4,1; 6,2; 5,8; 3,0; 5,4

vidurkis lygus: (4,1 + 6,2 + 5,8 + 3 + 5,4)/5 = 24,5/5 = 4,9. Jei į kompiuterį vietoj 3 įvesime reikšmę 13, vidurkis bus lygus (4,1 + 6,2 + 5,8 + 13 + 5,4)/5 = 34,5/5 = 6,9.

Mediana (*median*) x_{med} yra reikšmė, „dalijanti" variacinę seką pusiau: 50 % reikšmių yra ne didesnės už medianą, kitos 50 % – ne mažesnės. Jei n nelyginis, mediana lygi viduriniajam variacinės sekos nariui: $x_{med} = x_{((n+1)/2)}$. Jei n yra lyginis, mediana lygi viduriniųjų variacinės sekos narių vidurkiui: $x_{med} = (x_{(n/2)} + x_{(n/2+1)})/2$. Jei imties dažnių skirstinys yra simetriškas vidurkio atžvilgiu, mediana artima vidurkiui. Vertinant imties centrą, mediana ne tokia jautri išskirtims kaip vidurkis. Pavyzdžiui, 5 BC koncentracijos reikšmių (mmol/l) 4,1; 6,2; 5,8; 3,0; 5,4 variacinė seka yra 3,0; 4,1; 5,4; 5,8; 6,2, mediana lygi 5,4. Jei į kompiuterį vietoj 3 įrašysime reikšmę 13, mediana bus lygi 5,8. Ji pakito 0,4, nors vidurkis pakito 2. Todėl įtariant, kad imtyje yra „šiukšlių", imties centrą geriau įvertinti mediana. Jei mediana labai skiriasi nuo vidurkio, kintamojo pasiskirstymas nėra simetriškas.

Moda (*mode*) vadinama imtyje dažniausiai pasitaikanti reikšmė. Imtyje gali būti viena arba dvi ir daugiau modų. Tai vienintelis nominaliojo kintamojo imties centro įvertis, nes skaičiuoti vidurkį ar medianą nėra prasmės.

Geometrinis vidurkis. Dalis medicinos bei biologijos kintamųjų įgyja tik teigiamas reikšmes, be to, tarp matavimų pasitaiko labai didelių skaitinių reikšmių. Pavyzdžiui, tiriant vandens taršą bakterijomis, nustatomas (bakterijų skaičius)/ml. Dalies mėginių bakterijų skaičius kinta tarp šimto ir tūkstančio, nors keliuose mėginiuose pasitaiko ir keliasdešimt tūkstančių – bakterijų skaičius auga eksponentiškai. Kadangi aritmetinį vidurkį labai iškreipia ekstremalios reikšmės, tokių duomenų „sankaupos centrui" vertinti be medianos naudojamas ir geometrinis vidurkis:

$$\bar{x}_g = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

– n-tojo laipsnio šaknis iš imties reikšmių sandaugos. Pagal apibrėžimą geometrinio vidurkio logaritmas lygus logaritmuotų imties reikšmių vidurkiui:

$$\ln(\bar{x}_g) = (\ln(x_1) + \ln(x_2) + \ldots + \ln(x_n))/n; \quad \bar{x}_g = \exp(\ln(\bar{x}_g)).$$

Logaritmuojant imties reikšmes, mažinamas labai didelių reikšmių efektas – logaritmuotų reikšmių skirstinys simetriškesnis vidurkio atžvilgiu.

1.1 pavyzdys. Tiriant vandens taršą bakterijomis, šešiuose mėginiuose nustatytos tokios reikšmės: 100; 100; 1000; 1000; 10 000; 1 000 000. Logaritmuojant

šias reikšmes pagrindu 10, gaunama 2; 2; 3; 3; 4; 6; šių reikšmių vidurkis lygus 3,33. Antilogaritmuojant skaičių 3,33 pagrindu 10, gaunamas geometrinis vidurkis: $\bar{x}_g = 10^{3,33} = 2154,43$. Bakterijų skaičiaus aritmetinis vidurkis lygus 168 700 ir yra 78 kartus didesnis už geometrinį vidurkį.

Harmoninis vidurkis yra atvirkštinių imties reikšmių ($1/x_i$) vidurkio atvirkštinis dydis:

$$\bar{x}_h = n/(1/x_1 + 1/x_2 + ... + 1/x_n).$$

Harmoninis vidurkis naudotinas tik tada, kai kintamojo reikšmės yra teigiamos, jis tiksliau vertina vidutinį greitį nei aritmetinis. Harmoninio vidurkio taikymą iliustruosime pavyzdžiu. Tarkime, pirmą kartą apsilankęs donoras atidavė 250 ml kraujo 70 ml/min. greičiu, antrą kartą – 90 ml/min. greičiu. Vidutinį kraujo atidavimo greitį galima vertinti aritmetiniu vidurkiu: $\bar{x}$ = (70 + 90)/2 = 80 (min.). Tačiau pirmą kartą donoras kraują atidavė per 250 (ml)/70 (ml/min.) = 3,571 (min.), antrą kartą – per 250 (ml)/90 (ml/min.) = 2,778 (min.). 500 ml kraujo donoras atidavė per 3,571 + 2,778 = 6,349 (min.). Taigi vidutinis kraujo atidavimo greitis lygus 500 (ml)/6,349 (min.) = 78,750 (ml/min.). Šiuo atveju harmoninis vidurkis $\bar{x}_h = 2/(1/70 + 1/90) = 78,750$ vidutinį kraujo atidavimo greitį apibūdina tiksliau nei aritmetinis vidurkis (80 ml/min.).

Duomenų kitimo įverčiai. Duomenų sklaida apie vidurkį apibūdintina skirtumais $|x_i - \bar{x}|$, i = 1, 2 ... n. Išmatuotų reikšmių sklaidą apie vidurkį vertina imties dispersija (*variance*) s^2. Ji apskaičiuojama pagal formulę:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (1.1)$$

Iš formulės matyti, kad dispersija matuojama kvadratiniais x_i vienetais, taigi dispersijos negalima lyginti su vidurkiu. Todėl duomenų sklaidai vertinti naudojamas **standartinis nuokrypis** (*standard deviation*, arba *std. dev.*) s yra kvadratinė šaknis iš dispersijos. Jo matavimo vienetai tie patys, kaip ir kintamojo, todėl s galima lyginti su $\bar{x}$.

Bedimensė sklaidos charakteristika yra **variacijos koeficientas**:

$V = s / \bar{x}$.

Jis dažnai pateikiamas procentais ($s \times 100/\bar{x}$). Variacijos koeficientas parodo tyrimo tikslumą.

Imties plotis (*range*) yra absoliutus kitimo matas. Jis lygus maksimalios ir minimalios imties reikšmių skirtumui: $x_{(n)} - x_{(1)}$.

Kvartiliais (*quartile*) vadinami trys taškai, dalijantys kintamojo reikšmių seką, išsidėsčiusią nuo mažiausios iki didžiausios reikšmės, į keturias dalis, kurių kiekvienoje yra po 25 % imties reikšmių. Jie taip pat naudojami imties reikšmių kitimui įvertinti. Apatiniu kvartiliu (*lower quartile*) Q_1 laikoma reikšmė, už kurią 25 % imties reikšmių yra ne didesnės, viduriniu kvartiliu Q_2 laikoma mediana, o viršutiniu kvartiliu (*upper quartile*) Q_3 – reikšmė, už kurią 25 % imties reikšmių yra ne mažesnės. Kvartiliai nepriklauso nuo imties variacinės sekos kraštinių reikšmių, taigi jie nejautrūs išskirtims.

Procentiliai. j-osios eilės ($j < 100$) procentilis $P_{(j)}$ – tai skaitinė reikšmė, už kurią j % imties reikšmių yra ne didesnės. Pagal apibrėžimą apatinis ir viršutinis kvartiliai yra atitinkamai 25-osios ir 75-osios eilės procentiliai, o mediana – 50-osios eilės procentilis. Duomenų analizėje dažnai naudojami 33-osios ir 67-osios eilių procentiliai – terciliai, dalijantys variacinę seką į 3 lygias dalis. Terciliai ir kvartiliai naudojami kiekybiniam kintamajam transformuoti į tvarkos. Pavyzdžiui, vertinant pacientų gyvenimo kokybę, jiems pateikiamas klausimynas ir pagal atitinkamą metodiką nustatomas kiekybinis rodiklis – gyvenimo kokybės balas (kuo jis didesnis, tuo individas geriau vertina gyvenimo kokybę). Šio balo kitimo ribos būna gana plačios, pavyzdžiui, tarp 20 ir 60, todėl tolesnei analizei ir rezultatams interpretuoti patogiau individus suskirstyti į tris ar keturias grupes pagal gyvenimo kokybės laipsnį. Individai skirstomi į grupes, kurios sudaromos pagal tercilių ar kvartilių reikšmes. Pavyzdžiui, skirstant individus į 3 grupes, laikoma, kad gyvenimo kokybė prasta, jei balo reikšmė neviršija pirmo tercilio; gyvenimo kokybė vidutiniška, jei balo reikšmė yra tarp pirmo ir antro tercilio; gyvenimo kokybė gera, jei balo reikšmė viršija antrą tercilį. Analogiškai individai gali būti skirstomi į 4 grupes pagal kvartilius. Tokio skirstymo privalumas – visose grupėse individų skaičius beveik vienodas.

Interkvartilinis plotis (*quartile range*) $H = Q_3 - Q_1$, t. y. viršutinio ir apatinio kvartilio skirtumas, kuris dėl nejautrumo išskirtims naudojamas imties sklaidai įvertinti.

Vietoj interkvartilinio pločio imties sklaidai vertinti galima naudoti procentilių skirtumą.

Mažiausia imties reikšmė, kvartiliai bei didžiausia reikšmė vadinami penkių skaičių santrauka (*five-number summary*).

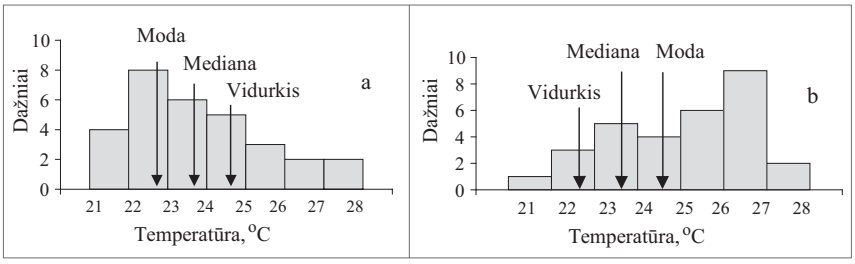
Histogramos formos įverčiai. Imties reikšmės ne visada būna išsidėsčiusios simetriškai „centro“ atžvilgiu. Jei intervalas nuo didžiausios imties reikšmės iki medianos yra ilgesnis už intervalą nuo medianos iki mažiausios reikšmės ($x_{max} - x_{med} > x_{med} - x_{min}$), turime teigiamą asimetriją. Jei intervalas nuo

mažiausios imties reikšmės iki medianos yra ilgesnis už intervalą nuo medianos iki didžiausios reikšmės ($x_{med} - x_{min} > x_{max} - x_{med}$), turime neigiamą asimetriją. Duomenų simetriškumas ar asimetriškumas atsispindi histogramoje (1.7 pav.). Jei asimetrija teigiama, vidurkis yra didesnis už medianą, jei neigiama – vidurkis mažesnis už medianą. Duomenų asimetrijai įvertinti naudojamas asimetrijos koeficientas (*skewness*) g_1:

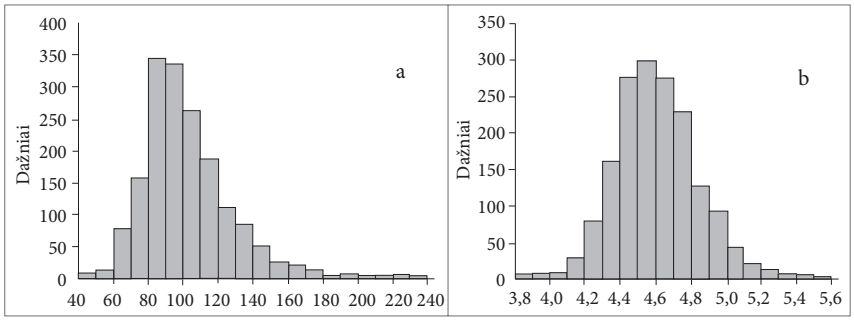
$$g_1 = n\sum_{i=1}^{n}(x_i - \bar{x})^3 /((n-1)(n-2)s^3).$$

Kai $g_1 > 0$ – asimetrija yra teigiama, kai $g_1 < 0$ – neigiama (1.7 pav.).

Jei rodiklis labai kinta ir įgyja tik teigiamas reikšmes, jo skirstinys dažniausiai turi teigiamą asimetriją. Tokie kintamieji nustatomi mikrobiologiniais bei serologiniais tyrimais: bakterijų skaičius mėginyje, troponono, kreatinino koncentracija ir pan. Teigiamą asimetriją sumažina kintamojo logaritmavimas (1.8 pav.). Pavyzdžiui, logaritmuodami pagrindu 10 skaičius 2; 20; 200; 2000, gauname atitinkamai 0,3; 1,3; 2,3; 3,3. Pirminių duomenų asimetrijos koeficientas lygus 1,94, logaritmuotų reikšmių – 0.



1.7 pav. Asimetriški duomenys: (a) teigiama asimetrija, (b) neigiama asimetrija



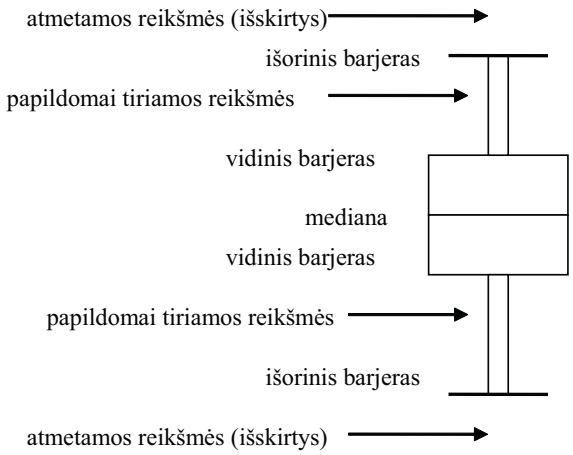
1.8 pav. Ligonių, sergančių ūmiais koronariniais sindromais (n = 1998), kreatinimo koncentracijos histograma: (a) pirminiai matavimai; (b) logaritmuotos reikšmės

Histogramos lėkštumo matas yra **eksceso koeficientas** (*kurtosis*):

$$g_2 = [n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4 /((n-1)s^4) - 3(n-1)^2]/[(n-2)(n-3)].$$

Kuo ekscesas didesnis, tuo histograma smailesnė.

Išskirtys. Išskirtis (*outliers*), arba „šiukšles“, rekomenduojama vertinti naudojant vidinius $[Q_1 - 1{,}5H; Q_3 + 1{,}5H]$ ir išorinius barjerus $[Q_1 - 3H; Q_3 + 3H]$ (1.9 pav.). Reikšmės, esančios už išorinio barjero ribų, yra išskirtys (*extreme outliers, extremes*) – jos atmetamos; reikšmės, esančios tarp vidinio ir išorinio barjero, yra sąlyginės išskirtys *(„mild“ outliers*). Jos tiriamos papildomai.



1.9 pav. Imties vidiniai ir išoriniai barjerai, atmetamos ir papildomai tiriamos reikšmės

Robastiniai įverčiai. Kaip minėta, imtyje esanti viena ar kelios išskirtys gerokai iškreipia vidurkį. Išskirtys taip pat labai keičia dispersiją, standartinį nuokrypį, asimetrijos ir eksceso koeficientus. Išskirčių poveikiui sumažinti naudojami robastiniai įverčiai (*robust estimations*) – įverčiai, nejautrūs išskirtims. Tai nupjautasis vidurkis (*trimmed mean*), vinzorizuota dispersija ir kt. Pavyzdžiui, imties „centrui“ vertinti be medianos dar naudojamas nupjautasis vidurkis (*trimmed mean*) bei vinzorizuotas vidurkis (*winsorized mean*). α % nupjautasis vidurkis skaičiuojamas atmetus α % mažiausių ir α % didžiausių imties reikšmių (čia $0 \le \alpha \le 100$). α % vinzorizuotas vidurkis skaičiuojamas α % mažiausių imties reikšmių pakeitus α eilės procentiliu ir α % didžiausių imties reikšmių pakeitus (100 – α) eilės procentiliu.

1.7. Skaitinių charakteristikų skaičiavimas ir grafinis pateikimas

1.4 lentelėje pateikti 30 moterų, vyresnių nei 50 m., sergančių išemine širdies liga, bendrojo cholesterolio koncentracijos reikšmės bei variacinė seka ([2]). 1.5 lentelėje pateiktos šio kontingento BC koncentracijos skaitinės charakteristikos. Iš 1.5 lentelės matome, kad vyresnių kaip 50 m. moterų, sergančių IŠL, BC koncentracija svyruoja nuo 3,83 iki 12,6. Cholesterolio vidurkis – 7,095, mediana lygi 7,12, t. y. nedaug skiriasi nuo vidurkio. Ketvirtadalio tirtų moterų BC vidurkis mažesnis nei 5,88, ketvirtadalio svyruoja tarp 5,88 ir 7,12 bei atitinkamai tarp 7,12 ir 7,76. Ketvirtadalio moterų BC koncentracija viršija 7,76 (mmol/l).

1.4 lentelė. Vyresnių nei 50 m. moterų, sergančių išemine širdies liga, BC koncentracijos (mmol/l) reikšmės, variacinė seka bei reikšmių rangai (i)

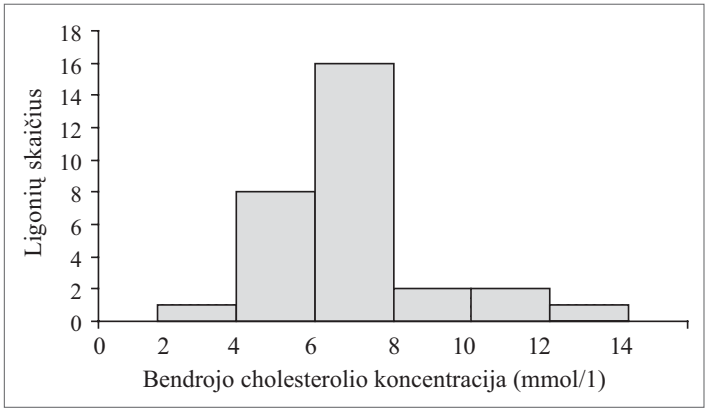
i	Imties reikšmės x_i	Variacinė seka $x_{(i)}$	i	Imties reikšmės x_i	Variacinė seka $x_{(i)}$
1	5,89	3,83	16	8,19	7,13
2	7,76	5,04	17	8,3	7,13
3	5,38	5,12	18	5,12	7,14
4	7,07	5,2	19	7,13	7,24
5	6,85	5,38	20	7,35	7,34
6	10,9	5,67	21	7,14	7,35
7	5,04	5,71	22	5,2	7,50
8	7,24	5,88	23	5,67	7,76
9	6,2	5,89	24	10,2	7,84
10	12,6	6,2	25	7,34	7,97
11	7,97	6,39	26	7,11	8,19
12	3,83	6,85	27	7,13	8,3
13	7,84	6,92	28	5,88	10,2
14	5,71	7,07	29	7,5	10,9
15	6,92	7,11	30	6,39	12,6

BC dispersija lygi 3,12, standartinis nuokrypis – 1,788, t. y. labai artimas interkvartiliniam pločiui (1,88). Imties skirstinys yra asimetriškas į dešinę (1.10 pav.), asimetrijos koeficientas – teigiamas, lygus 1,18. Apatinis vidinis barjeras $Q_1 - 1{,}5H$ lygus 3,06, už jį mažesnių reikšmių nėra. Viršutinis vidinis barjeras $Q_3 + 1{,}5H$ lygus 10,58, už jį didesnės dvi imties reikšmės –

10,9 ir 12,6. Tai sąlyginės išskirtys. Išorinis barjeras $Q_3 + 3H$ lygus 13,4, taigi atmetamų reikšmių imtyje nėra.

1.5 lentelė. 30 moterų, vyresnių nei 50 m., sergančių IŠL, BC koncentracijos (1.3 lentelė) skaitinės charakteristikos

Charakteristika	Žymėjimas	Reikšmė
Didžiausia reikšmė	x_{max}, $x_{(n)}$	12,6
Mažiausia reikšmė	x_{min}, $x_{(1)}$	3,83
Vidurkis	$\bar{x}$	7,095
Mediana	x_{med}	(7,11 + 7,13)/2 = 7,12
Imties plotis	$x_{max} - x_{min}$	12,6 – 3,83 = 8,77
Kvartiliai	Q_1, Q_2, Q_3	5,88; 7,12; 7,76
Interkvartilinis plotis	$H = Q_3 - Q_1$	7,76 – 5,88 = 1,88
Dispersija	s^2	3,1198
Standartinis nuokrypis	s	1,788
Variacijos koeficientas	$V = s / \bar{x}$	1,788/7,095 = 0,252
Asimetrijos koeficientas	g_1	1,18



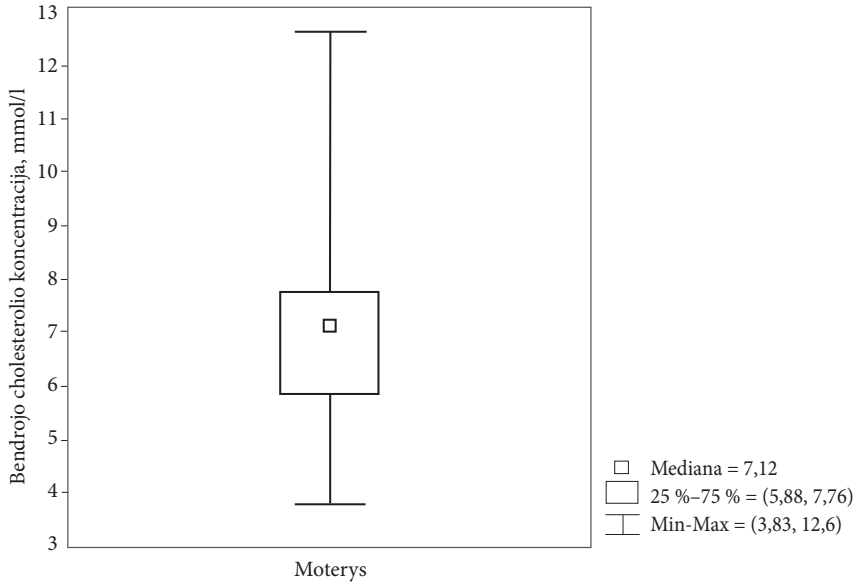
1.10 pav. Moterų, vyresnių nei 50 m. ir sergančių IŠL, BC koncentracijos (mmol/l) histograma

Apie bendrą kiekybinio kintamojo imties centro, sklaidos, simetriškumo bei ekstremalių (maksimalios ir minimalios) reikšmių vaizdą galime spręsti pagal stačiakampę diagramą (*Box-whisker plot*). Stačiakampės diagramos „dėžė“ – stačiakampis, braižomas nuo pirmojo Q_1 iki trečiojo kvartilio Q_3. Stačiakampio viduryje kvadratėliu ar brūkšniu pažymima mediana. (Kartais

diagramoje pliusu pažymimas ir vidurkis). Nuo stačiakampio šono brėžiami „ūsai“ – į viršų iki maksimalios ir į apačią iki minimalios reikšmės (1.11 pav.).

Stačiakampės diagramos „dėžės“ dydis charakterizuoja imties reikšmių išsibarstymą. Atstumai nuo „ūsų“ galo iki medianos charakterizuoja asimeriją. Jei viršutinis „ūsas“ daug ilgesnis už apatinį, kintamojo skirstinio asimetrija yra dešinioji, jei trumpesnis – kairioji. Jei „ūsai“ daug ilgesni už „dėžę“, galima įtarti imtyje esančias išskirtis. Stačiakampės diagramos leidžia palyginti kelių kintamųjų, matuotų tais pačiais vienetais, ar to paties kintamojo kelių imčių duomenis.

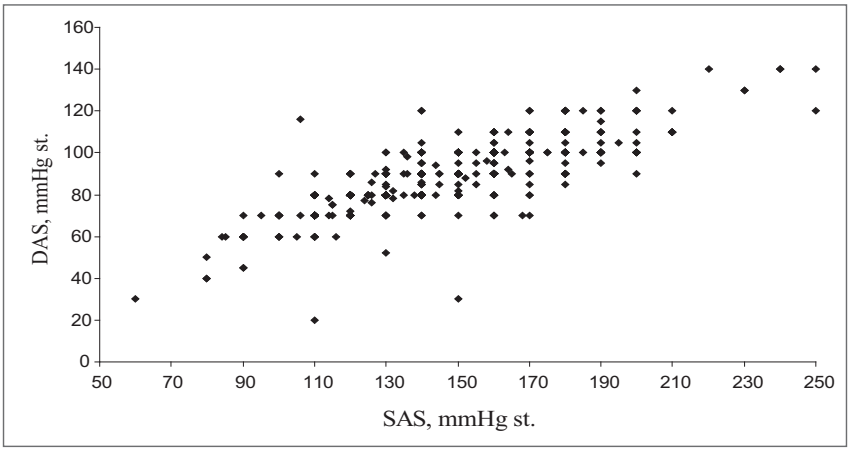
1.11 pav. pateikta vyresnių nei 50 m. ir sergančių išemine širdies liga moterų bendrojo cholesterolio koncentracijos stačiakampė diagrama.



1.11 pav. Vyresnių nei 50 m. ir sergančių išemine širdies liga moterų bendrojo cholesterolio koncentracijos stačiakampė diagrama

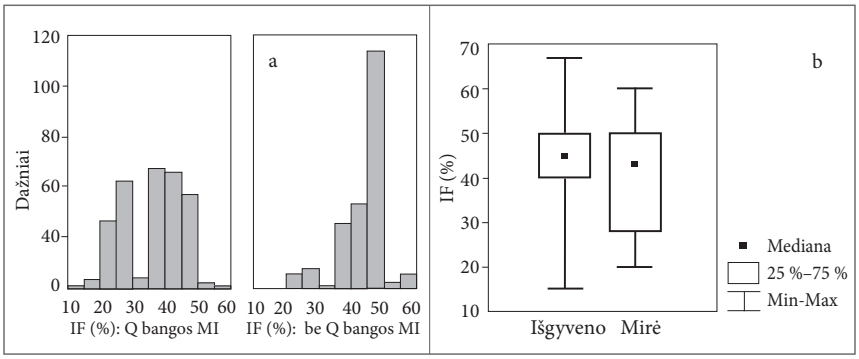
1.8. Grafinis dviejų kintamųjų pateikimas

Kintamųjų tarpusavio ryšio analizė pradedama nuo pirminės apžvalgos – grafinio vaizdo. Dviejų kiekybinių kintamųjų tarpusavio išsidėstymas pateikiamas skaidos diagrama (1.12 pav.). Ox ašyje atidedamos vieno kintamojo reikšmės, Oy – kito.



1.12 pav. Ligonių, sergančių ŪKS, sistolinio ir diastolinio kraujospūdžio skaidos diagrama

Kokybinio ir kiekybinio kintamojo tarpusavio pasiskirstymas gali būti pateiktas kategorizuota histograma (1.13 a pav.) ar kategorizuota stačiakampe diagrama (1.13 b pav.). 1.13 a pav. pateiktos ligonių su Q ir be Q bangos MI išstūmimo frakcijos (IF) histogramos. 1.13 b pav. pateiktos ligonių, sirgusių ŪKS, mirusių per 1 metus bei išgyvenusių, IF stačiakampės diagramos (naudoti tyrimo, aprašyto [2] duomenys). Iš šių diagramų matome, kad ketvirtadalis išgyvenusių ligonių turėjo išstūmimo frakciją, mažesnę nei 40, o ketvirtadalio mirusių per 1 metus ligonių IF buvo mažesnė nei 30. Be to, mirusių ligonių išstūmimo frakcijos interkvartilinis plotis (išsibarstymo matas) didesnis nei išgyvenusių.

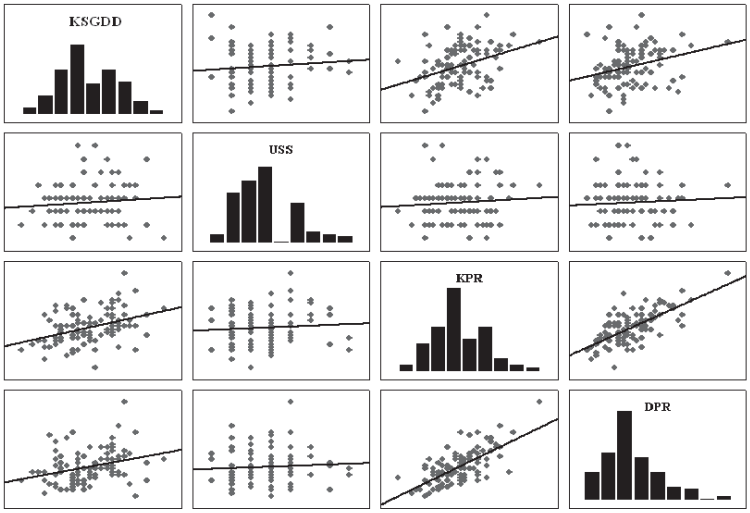


1.13 pav. Kategorizuota (Q bangos MI ir be Q bangos MI) išstūmimo frakcijos histograma ir kategorizuota (išgyvenusių ir mirusių per 1 metus) IF stačiakampė diagrama

1.9. Grafinis daugiamačio kintamojo pateikimas

Minėjome, kad dvimatis kintamasis, tarkime, (SAS, DAS), vaizduojamas plokštumoje, trimatis – erdvėje. Kai daugiamačio kintamojo koordinačių skaičius viršija 3, sunku grafiškai pavaizduoti jo reikšmes. Pateikiame keletą daugiamačių duomenų vaizdavimo būdų.

1. **Skaidos diagramų matrica (*scatter plot matrices*)**. Tai kvadratinė matrica, kurios gardelėse pateiktos atitinkamų kintamųjų skaidos diagramos (1.14 pav.).



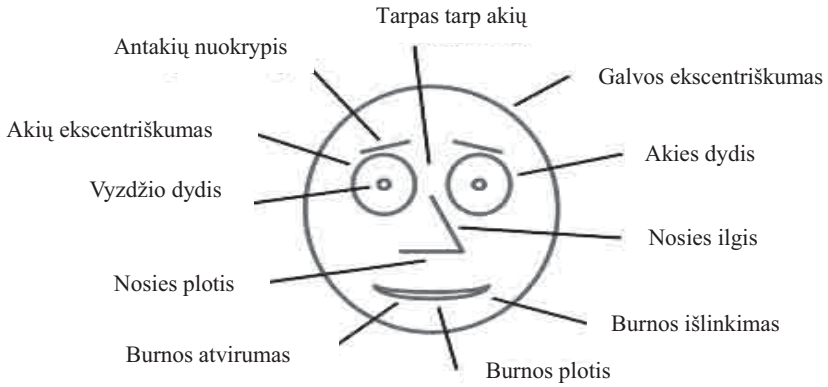
1.14 pav. Echoskopijos rodiklių skaidos diagramų matrica; čia KSGDD – širdies galinis diastolinis dydis, USS – užpakalinės sienelės dydis, KPR, DPR – kairiojo ir dešiniojo prieširdžio dydis

2. **Simboliai ženklai (*glyph techniques*)**. Individo daugiamačiam rodikliui pavaizduoti naudojami įvairūs simboliai – veidai, žvagždės ir t. t. Trumpai pateiksime, kaip tai daroma.

Veidai (*Chernoff faces*). Tai daugiamačių duomenų vaizdavimo metodas simboliniu veidu. Individo ar objekto daugiamačio kintamojo koordinatės pateikiamos kaip veido charakteristikos: galvos lėkštumas, akių, vyzdžių dydis, antakių, nosies ilgis ir plotis, burnos forma ir t. t. Visos koordinatės (vienmačiai kintamieji) transformuojamos į skaičių tarp 0 ir 1. Kiekviena koordinatė vaizduojama skirtingu veido bruožu, pavyzdžiui, veido pločiu, akių dydžiu, burnos kreivumu, nosies ilgiu ir t. t. 1.15 pav. pateiktas simbolinis veidas, apibūdintas 10 veido bruožų: kiekvienas veido parametras vaiz-

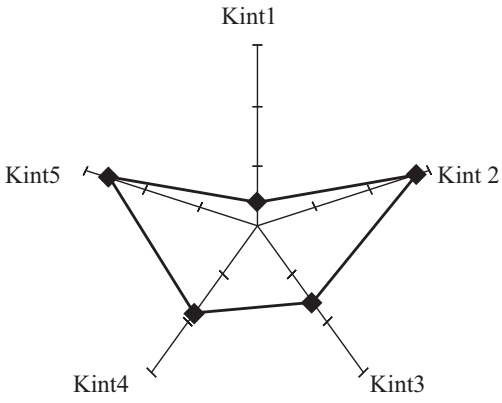
duoja skaičių, esantį tarp 0 ir 1. Veido bruožų yra tiek, kiek yra daugiamačio kintamojo koordinačių.

Simboliniu veidu galime pavaizduoti sergamumą įvairiomis vėžio rūšimis atskiruose rajonuose, skirtingų rasių individų morfometrinius rodiklius ir t. t.

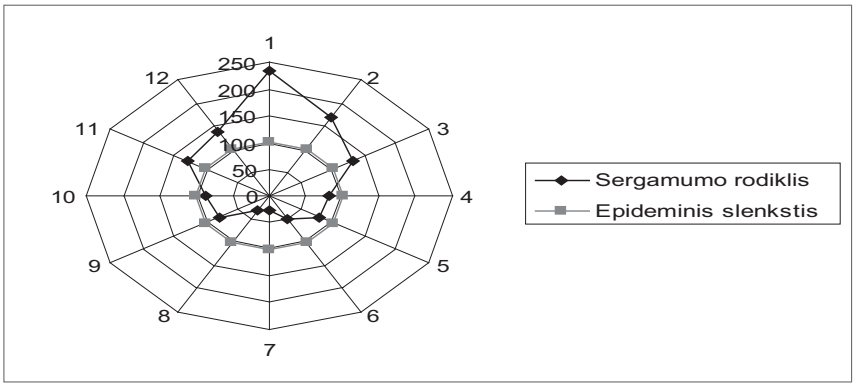


1.15 pav. Simbolinis veidas, skirtas daugiamačiam kintamajam grafiškai pateikti

Žvaigždės (*star plot*). Individas ar kitoks objektas vaizduojamas žvaigžde: kiekvieną koordinatę atitinka žvaigždės smailė, o smailės ilgis yra proporcingas koordinatės dydžiui (1.16 pav.). 1.17 pav. žvaigžde pateiktas 2004 m. sergamumo ūminėmis viršutinių kvėpavimo takų infekcijomis mėnesinis rodiklis (10 000 gyventojų) Vilniaus mieste (12 matavimų kintamasis) bei epideminis slenkstis (100 susirgimų, tenkančių 10 000 gyventojų).



1.16 pav. Penkiamačio kintamojo (Kint1, Kint2, Kint3, Kint4, Kint5) pateikimas žvaigžde



1.17 pav. Sergamumo ūmiomis viršutinių kvėpavimo takų infekcijomis rodiklio (10 000 gyventojų) Vilniaus mieste 2004 m. sezoniškumas

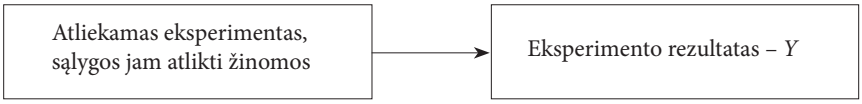
1 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Babarskienė M., Venclovienė J., Lukšienė D., Šlapikienė B., Milvydaitė I. Susirgusių miokardo infarktu klinikinės rizikos vertinimas 30 parų laikotarpiu. *Medicina.* 2001. 37 tomas, Nr. 12, p. 1418–1424.
3. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
4. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.
5. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
6. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
7. Prieiga per internetą: http://www.comp.leeds.ac.uk/kwb/ENV/lec2.ppt.
   Skaidrės apie daugiamačių duomenų vaizdavimą: skaidos diagramų matrica, lygiagrečių koordinačių diagrama, simbolinių ženklų diagramos.
8. Prieiga per internetą: http://www.epcc.ed.ac.uk/computing/training/document_archive/SciVis-course/SciVis.book_47.html#1.
   Daugiamačių duomenų vaizdavimas veidais.
9. Explanatory Data Analysis. *Engineering Statistics Handbook*. Prieiga per internetą: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3653.htm.

2 SKYRIUS

Pagrindinės tikimybių teorijos sąvokos ir formulės

Tiek medikų, tiek kitų sričių specialistų atliekamus tyrimus galime apibūdinti tokia schema:



Pavyzdžiui, elektros srovės priklausomybę nuo įtampos ir varžos nusako Omo dėsnis. Žinant įtampą U ir varžą R (eksperimento sąlygos), pagal formulę $I = U/R$ nustatomas srovės dydis I (eksperimento rezultatas Y). Tai yra determinuotas modelis – žinant eksperimento sąlygas, nusakomas eksperimento rezultatas. Tačiau yra nemažai eksperimentų, kurių išeities nusakyti neįmanoma, nors sąlygos ir žinomos. Pavyzdžiui, metame monetą (eksperimentas). Žinome, kad iškris arba herbas, arba skaičius, tačiau kas – tiksliai pasakyti neįmanoma. Šiuo atveju rezultatas Y yra atsitiktinis. Eksperimentai, kurių rezultatai atsitiktiniai, vadinami atsitiktiniais, tikimybiniais arba stochastiniais.

Tikimybių teorija yra atsitiktinių reiškinių matematinio modelio sudarymo ir analizės teorija. Žmogaus, kaip sudėtingos biologinės sistemos, rodikliai (SAS, DAS, cholesterolio kiekis ir t. t.) priklauso nuo daugybės faktorių. Šių faktorių tarpusavio sąveikos mechanizmas per daug sudėtingas, kad naudojantis determinuotu modeliu būtų galima nusakyti SAS, DAS, cholesterolio koncentraciją ir kitas panašaus pobūdžio charakteristikas. Todėl minėtus individo rodiklius tikslinga laikyti atsitiktiniais ir jų analizei naudoti tikimybinius modelius.

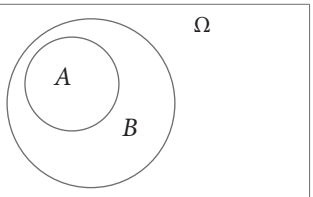
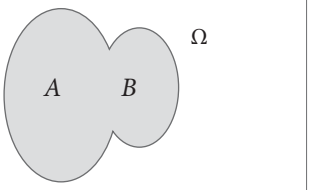
2.1. Atsitiktiniai įvykiai. Tikimybės

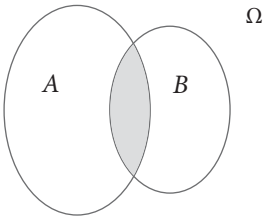
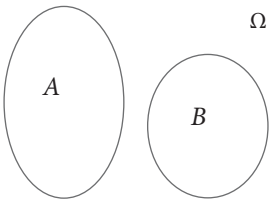
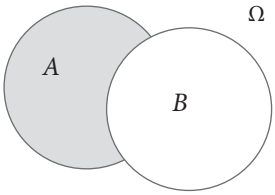
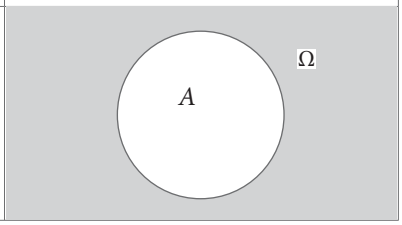
Viena pagrindinių tikimybių teorijos sąvokų – atsitiktinis įvykis.

Kaip minėta, eksperimento rezultatai yra atsitiktiniai. Atsitiktinės eksperimento baigtys vadinamos atsitiktiniais įvykiais. Eksperimento rezultatus, kurių negalime smulkiau išskaidyti, vadiname **elementariais atsitiktiniais įvykiais**. Pavyzdžiui, 2 kartus metame monetą. Galimi išeičių variantai: „iškrito abu herbai“, „iškrito herbas ir skaičius“, „iškrito abu skaičiai“. Tai atsitiktiniai įvykiai. Įvykis „iškrito herbas ir skaičius“ įvyks tuo atveju, kai pirmu metimu iškris herbas, antru – skaičius (H_1S_2) arba pirmu metimu iškris skaičius, antru – herbas (S_1H_2); t. y. įvykį „iškrito herbas ir skaičius“ galima išskaidyti į įvykius H_1S_2 ir S_1H_2. Toliau įvykių H_1S_2 ir S_1H_2 skaidyti neįmanoma, tad jie ir yra elementarūs. Taigi, metant monetą 2 kartus, gali įvykti šie elementarūs įvykiai: H_1H_2, H_1S_2, S_1H_2, S_1S_2. Visų elementarių įvykių aibė žymima Ω. Mūsų nagrinėtame pavyzdyje $\Omega = \{H_1H_2, H_1S_2, S_1H_2, S_1S_2\}$. Ω gali būti baigtinė: $\Omega = \{\omega_1, \omega_2 \ldots \omega_n\}$; čia ω_i – elementarus įvykis, arba begalinė. Pavyzdžiui, nustatomas kraujo krešėjimo laikas. Tai gali būti bet kuris intervalo $(0, T)$ taškas. Šiuo atveju Ω – visi intervalo $(0, T)$ taškai – yra begalinė.

Pateiksime supaprastintą atsitiktinio įvykio apibrėžimą (tikslus matematinis atsitiktinio įvykio apibrėžimas pateiktas vadovėliuose [2, 4, 8, 9]). **Atsitiktiniu įvykiu vadinamas bet kuris elementarių įvykių aibės Ω poaibis. Įvykis *A*, sutampantis su Ω ($A = \Omega$), vadinamas būtinu įvykiu. Įvykis, neturintis nė vieno elementaraus įvykio, vadinamas negalimu ir žymimas Ø.**

Veiksmai su įvykiais. Kadangi įvykiai yra aibės *Ω* poaibiai, su jais atliekami tokie patys veiksmai, kaip ir su aibėmis.

1. Įvykis A vadinamas įvykio B poaibiu ir žymimas $A \subset B$, jei įvykus A, įvyks ir B. Pavyzdžiui, tiriamas sergamumas CD. Kiekvienam individui nustatomas vienas iš atvejų: {„neserga“, „serga I^o CD“, „serga II^o CD“}. Tuomet „serga I^o CD“ $\subset$ serga CD“.	
2. Įvykių A ir B suma $A + B$ vadinamas įvykis, įvyksiantis tuomet, kai įvyks arba A, arba B, arba A ir B kartu. Pavyzdžiui, A = „šeimoje yra bent viena mergaitė“, B = „šeimoje yra bent vienas berniukas“; $A + B$ = „šeimoje yra bent vienas vaikas“.	

3. Įvykių A ir B sandauga AB vadinamas įvykis, įvyksiantis tuomet, kai įvyks ir A, ir B. Pavyzdžiui, 1891 m. surašant Anglijos ir Velso gyventojus, buvo renkami duomenys apie tėvų ir vaikų akių spalvą. Pažymėkime A = „tamsiaakis tėvas“, B = „tamsiaakis sūnus“. Tuomet AB = „tamsiaakis tėvas ir tamsiaakis sūnus“.	
4. Įvykiai A ir B nesutaikomi, jei jie negali įvykti vienu metu: $AB = Ø$. Pavyzdžiui, A = „individas serga I^o CD“, B = „individas serga II^o CD“. Tuomet $AB = Ø$.	
5. Įvykių A ir B skirtumu $A\backslash B$ vadinamas įvykis, įvykstantis tuomet, kai A įvyks, o B neįvyks. Pavyzdžiui, A = „individas serga CD“, B = „individas serga I^o CD“. Tuomet $A\backslash B$ = „individas serga II^o CD“.	
6. Įvykis $\overline{A} = \Omega\backslash A$ vadinamas priešingu įvykiui A, jei jis įvyks tuomet, kai neįvyks įvykis A. Pavyzdžiui, A = „individas serga CD“. Tuomet $\overline{A}$ = „individas neserga CD“.	

Kadangi eksperimento išeitimis gali būti keli atsitiktiniai įvykiai, pageidautina žinoti kiekvieno įvykio galimybę įvykti, t. y. norima įvertinti kiekvieno įvykio tikėtinumą. Dėl to įvedama *tikimybės* sąvoka. **Tikimybė yra atsitiktinio įvykio galimybės įvykti („dydžio“) skaitinis matas – skaičius tarp 0 ir 1.** Įvykio A tikimybė žymima $P(A)$.

Statistinis tikimybės apibrėžimas. Tarkime, tą patį eksperimentą kartojame n kartų. Šių n eksperimentų metu įvykis A įvyko m kartų. Pažymėkime:

$p_n = m/n$.

Sakykime, didėjant n, santykis p_n artėja prie skaičiaus p. Tuomet p vadinamas įvykio A tikimybe ir žymima $P(A) = p$.

Statistinė tikimybė nustatoma įvykiams „serga liga X“, „lankosi privačioje klinikoje“, „gimė berniukas“ ir t. t. Pavyzdžiui, ilgamečio naujagimių registravimo duomenys rodo, jog kasmet gimsta maždaug 51 % berniukų. Taigi

berniuko gimimo tikimybė laikoma lygia 0,51. Tikimybė procentais suprantama kaip įvykio A pasirodymų skaičius atlikus 100 eksperimentų.

Klasikinis tikimybės apibrėžimas. Šis apibrėžimas naudojamas tik tada, kai elementarių įvykių skaičius yra baigtinis ir visi elementarūs įvykiai vienodai galimi. Įvykio A tikimybė $P(A)$ apibrėžiama santykiu:

$P(A) = k/n$;

čia n – visų elementarių įvykių skaičius (galimų įvykių skaičius), k – į įvykį A įeinančių elementarių įvykių skaičius (palankių įvykiui A elementarių įvykių skaičius).

2.1 pavyzdys. Metame kauliuką. Kokia tikimybė, kad iškris iš 3 dalus akių skaičius? Pažymėkime: Ω – elementarių įvykių aibė, A = „iškrito iš 3 dalus akių skaičius“. Ω susideda iš 6 elementarių įvykių $\Omega = \{\omega_1, \omega_2 \ldots \omega_6\}$; čia ω_i = „iškrito i akių“, $A = \{\omega_3, \omega_6\}$. Galimų įvykių skaičius n lygus 6, $k = 2$, $P(A) = k/n = 2/6 = 1/3$.

Aksiominis tikimybės apibrėžimas (matematinis apibrėžimas pateiktas vadovėliuose [2, 4, 8, 9]). Įvykio A tikimybe vadinama skaitinė funkcija P, tenkinanti šias sąlygas:

1) $P(A) \geq 0$ – tikimybė neneigiama;
2) $P(\Omega) = 1$ – būtino įvykio tikimybė lygi 1;
3) $P(A + B) = P(A) + P(B)$, jei $AB = \emptyset$ – kai A ir B nesutaikomi, jų sumos tikimybė lygi A ir B tikimybių sumai.

Šis tikimybės apibrėžimas bendresnis už klasikinį ir statistinį: eksperimentų kartoti nereikia, elementarių įvykių aibė gali būti begalinė, o elementarūs įvykiai nebūtinai vienodai galimi.

2.2. Pagrindinės tikimybių skaičiavimo taisyklės. Sąlyginės tikimybės. Įvykių nepriklausomumas

Iš aksiominio tikimybės apibrėžimo 1–3 sąlygų išplaukia tikimybių skaičiavimo formulės:

1) $P(\emptyset) = 0$;

2) $P(A \setminus B) = P(A) - P(AB)$;

3) $P(A + B) = P(A) + P(B) - P(AB)$; jei A ir B nesutaikomi ($AB = \emptyset$), tuomet $P(A + B) = P(A) + P(B)$;

4) $P(\overline{A}) = 1 - P(A)$.

Šių savybių taikymą tikimybėms skaičiuoti iliustruosime pavyzdžiu. Draudimo agentas analizavo sutuoktinių lankymosi privačiose klinikose dėsnin-

gumus. Tyrimo metu atlikta abiejų sutuoktinių apklausa: fiksuota sutuoktinio lytis (reikšmės V, M) ir ar jis lankosi privačioje klinikoje (reikšmės +, –). Atsitiktinai parinkto asmens galimi šie atsakymų variantai (elementarūs įvykiai): $V+$, $V–$, $M+$, $M–$. Atsitiktinai parinktos šeimos galimi šie atsakymų variantai: ($V+$, $M+$), ($V–$, $M+$), ($V+$, $M–$), ($V–$, $M–$). Tyrimo metu nustatyta, kad 25 % šeimų abu sutuoktiniai lankosi privačiose klinikose, 25 % šeimų konstatuota ($V+$, $M–$), 25 % šeimų – ($V–$, $M+$) ir 25 % šeimų – ($V–$, $M–$). Taigi turime atitinkamų įvykių statistines tikimybes: $P((V+, M+)) = P((V+, M–)) = P((V–, M+)) = P((V–, M–)) = 0{,}25$.

Tikimybė, kad vyras lankosi privačioje klinikoje, skaičiuojama:

$P(V+) = P((V+, M+)$ *arba* $(V+, M–)) = (P((V+, M+)) + P((V+, M–)) = 0{,}5$

(įvykiai ($V+$, $M+$) ir ($V+$, $M–$) nesutaikomi – kartu įvykti negali). Analogiškai $P(M+) = 0{,}5$; $P(V–) = 1 – P(V+) = 0{,}5$.

Sąlyginė tikimybė. Įvykio A tikimybę galima skaičiuoti ir tuo atveju, kai įvykis B yra įvykęs. Tokia tikimybė vadinama įvykio A sąlygine tikimybe ir žymima $P(A|B)$ (įvykio A tikimybė su sąlyga, jei įvyko B). Sąlyginė tikimybė skaičiuojama naudojant besąlygines tikimybes:

$$P(A \mid B) = \frac{P(AB)}{P(B)}.$$

Pavyzdžiui, draudimo agentas norėjo nustatyti, ar vieno sutuoktinio lankymąsi privačioje klinikoje veikia faktas, kad joje lankosi kitas sutuoktinis. Kitaip tariant, draudimo agentas norėjo patikrinti, ar sąlyginės tikimybės $P(V+|M+)$ ir $P(V+|M+)$ yra lygios; čia $P(V+|M+)$ – tikimybė, jog vyras lankosi privačioje klinikoje su sąlyga, kad joje lankosi jo žmona. Tyrimo metu nustatyta: 50 % moterų lankosi privačiose klinikose. Taigi turime:

$$P(V+ \mid M+) = \frac{P(V+, M+)}{P(M+)} = 0{,}5 = P(V+).$$

Analogiškai $P(V+|M+) = 0{,}5 = P(V+)$.

Gavome, kad sąlyginės tikimybės $P(V+|M+)$ ir $P(V+|M+)$ yra lygios besąlyginėms, taigi žmonos lankymasis privačioje klinikoje neturi įtakos vyro lankymuisi. Analogiškai gauname, kad $P(M+|V+) = P(M+|V–) = 0{,}5$.

Įvykių nepriklausomumas. Įvykiai A ir B yra nepriklausomi, jei sąlyginė tikimybė $P(A|B)$ lygi besąlyginei: $P(A|B) = P(A)$. Atsižvelgę į $P(A|B)$ bei $P(AB)$, $P(B)$ tarpusavio priklausomybę, galime pateikti įvykių A ir B nepriklausomumo apibrėžimą: **įvykiai *A* ir *B* yra nepriklausomi, jei**:

$P(AB) = P(A)\,P(B)$.

Remiantis tyrimų duomenimis, galima tvirtinti, kad žmonos lankymasis ar nesilankymas privačioje klinikoje neturi įtakos vyro lankymuisi joje: $P(V+|M+) = P(V+|M-) = P(V+)$. Pagal apibrėžimą įvykiai $(V+)$ ir $(M+)$ yra nepriklausomi, nes jų sandaugos tikimybė lygi atitinkamų tikimybių sandaugai:

$P(V+ \text{ ir } M+) = P(V+)\,P(M+) = 0{,}5 \times 0{,}5 = 0{,}25.$

2.2 pavyzdys. Nustatysime tikimybę, kad iš 3 šeimoje gimusių vaikų nebuvo berniuko, buvo 1, 2 ir 3 berniukai. Daroma prielaida, kad vaiko lytis nepriklauso nuo anksčiau gimusio vaiko lyties, o tikimybė berniukui gimti lygi 0,51. Pažymėkime įvykį B = „gimė berniukas“, M = „gimė mergaitė“. Tuomet $P(B) = 0{,}51$, $P(M) = 1 - 0{,}51 = 0{,}49$, P(„iš 3 vaikų nėra berniuko“) = $P(MMM) = 0{,}49 \times 0{,}49 \times 0{,}49 = 0{,}11765$. Šeimoje, kurioje yra vienas berniukas ir dvi mergaitės, galimos šios vaikų lyčių kombinacijos: *BMM*, *MBM*, *MMB*. Kiekvienos kombinacijos tikimybė lygi $0{,}51 \times 0{,}49 \times 0{,}49 = 0{,}12245$. Įvykiai *BMM*, *MBM* ir *MMB* nesutaikomi, todėl jų sumos tikimybė lygi įvykių tikimybių sumai:

P („iš 3 vaikų 1 berniukas“) = $P(BMM + MBM + MMB) = 3 \times 0{,}12245 = 0{,}36745$. Analogiškai P(„iš 3 vaikų 2 berniukai“) = $P(BBM + MBB + BMB) = 3 \times 0{,}51 \times 0{,}51 \times 0{,}49 = 0{,}38235$ ir P(„visi 3 berniukai“) = $P(BBB) = 0{,}51 \times 0{,}51 \times 0{,}51 = 0{,}13265$.

2.3. Pilnosios tikimybės formulė. Bajeso formulė

Kartais įvykio A tikimybė žinoma tik tam tikroje populiacijos dalyje, t. y. žinoma $P(A|H_1)$ ... $P(A|H_k)$, čia H_1, H_2 ... H_k – įvykiai, nurodantys populiacijos dalį. Pavyzdžiui, populiacijos individai pagal amžių suskirstyti į tris grupes: iki 40 m.; 40–65 m.; daugiau kaip 65 m. Žinomos įvykio A tikimybės amžiaus grupėse $P(A|H_1)$, $P(A|H_2)$, $P(A|H_3)$; čia H_1 = „individas iki 40 m. amžiaus“, H_2 = „individo amžius 40–65 m.“, H_3 = „individo amžius daugiau kaip 65 m.“ Tokiu atveju įvykio A tikimybė visoje populiacijoje skaičiuojama pagal pilnosios tikimybės formulę, kurios taikymo sąlygas nurodo pateikta teorema.

Teorema. Sakykime, H_1, H_2 ... H_k – atsitiktiniai įvykiai, tenkinantys sąlygas:

1) $H_1 + H_2 + ... + H_k = \Omega$ (bent vienas šių įvykių būtinai įvyks);
2) $P(H_i) > 0$ visiems $i = 1, 2 \ldots k$ (tarp šių įvykių nėra negalimų);
3) $H_iH_j = \emptyset$, visiems $i \neq j$ (šie įvykiai negali įvykti vienu metu).

Tuomet teisinga **pilnosios tikimybės formulė**:

$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \ldots + P(A|H_k)P(H_k).$

Šios teoremos taikymą iliustruosime pavyzdžiu.

2.3 pavyzdys. Mokslinėje konferencijoje dalyvavo 20 psichologų, 20 psichiatrų ir 10 terapeutų. Tikimybė, kad preparato X vartojimą žino psichologas, lygi 0,2, psichiatras – 0,5, terapeutas – 0,7. Kokia tikimybė, kad atsitiktinai sutiktas konferencijos dalyvis žinos, kaip vartoti preparatą X?

Pažymėkime įvykius:

A = „konferencijos dalyvis žino, kaip vartoti preparatą X“,
H_1 = „konferencijos dalyvis psichologas“,
H_2 = „konferencijos dalyvis psichiatras“,
H_3 = „konferencijos dalyvis terapeutas“.

Šių įvykių tikimybės bei įvykio A sąlyginės tikimybės lygios:

$P(H_1) = 20/50 = 0{,}4$; $P(H_2) = 20/50 = 0{,}4$; $P(H_3) = 10/50 = 0{,}2$;
$P(A|H_1) = 0{,}2$; $P(A|H_2) = 0{,}5$; $P(A|H_3) = 0{,}7$.

Taikydami pilnosios tikimybės formulę, skaičiuojame besąlyginę įvykio A tikimybę:

$P(A) = 0{,}2\times0{,}4 + 0{,}5\times0{,}4 + 0{,}7\times0{,}2 = 0{,}08 + 0{,}2 + 0{,}14 = 0{,}42$.

Taigi tikimybė, kad atsitiktinai sutiktas konferencijos dalyvis žinos, kaip vartoti preparatą X, lygi 0,42.

Bajeso (*Bayes*) formulė. Pilnosios tikimybės formulė naudojama įvykio, įvyksiančio ar neįvyksiančio eksperimento metu, tikimybei apskaičiuoti. Sakykime, įvykiai H_1, H_2 ... H_k tenkina pilnosios tikimybės teoremos sąlygas. Šių įvykių tikimybės $P(H_i)$ žinomos iki eksperimento ir vadinamos apriorinėmis (žinomomis iki eksperimento). Sakykime, eksperimento metu įvykis A įvyko. Kokia tikimybė, kad įvyko įvykis H_j? Tikimybei $P(H_j|A)$ skaičiuoti naudojama **Bajeso formulė**:

$$P(H_j \mid A) = \frac{P(H_j)P(A \mid H_j)}{P(H_1)P(A \mid H_1) + P(H_2)P(A \mid H_2) + \ldots + P(H_k)P(A \mid H_k)}.$$

Tikimybę $P(H_j|A)$ galima apskaičiuoti tik po eksperimento, kurio metu įvyko įvykis A. Todėl tikimybės $P(H_j|A)$ vadinamos aposteriorinėmis (žinomomis po eksperimento).

Bajeso formulė yra viena dažniausiai medicinoje naudojamų tikimybių teorijos formulių. Pavyzdžiui, Bajeso formulė įgalina įvertinti infekcinio susirgimo, diagnozuojamo remiantis testo rezultatu, tikimybę pagal testo išeitį. Pažymėkime atsitiktinius įvykius: $T+$ ir $T-$ – testo rezultatas teigiamas ir neigiamas, $D+$ – individas serga, $D-$ – individas neserga. Bajeso formulė susirgimo tikimybei skaičiuoti, kai testo rezultatas teigiamas, yra tokia:

$$P(D+\mid T+)=\frac{P(T+\mid D+)P(D+)}{P(T+\mid D+)P(D+)+P(T+\mid D-)P(D-)}; \qquad (2.1)$$

čia $P(D+)$ – tikimybė, kad individas serga, ir $P(D-)$ – kad neserga. Šios tikimybės įvertinamos remiantis epidemiologiniais duomenimis. Tikimybės $P(T+|D+)$ ir $P(T+|D-)$ įvertinamos eksperimentiškai, atliekant testo mėginius sergantiems ir nesergantiems. Tai statistinės tikimybės.

2.4 pavyzdys ([5, 100 p.]). TBC testas yra tuberkulino odos mėginys. Laikoma, kad šio testo rezultatas teigiamas, jei paraudimo spindulys viršija 5 mm. Remiantis sergančių TBC ($D+$) ir nesergančių TBC ($D-$) vaikų testo tyrimo duomenimis, norima įvertinti vaikų susirgimo TBC tikimybę, kai tuberkulino testo rezultatas yra teigiamas. Tyrime dalyvavo 10 000 vaikų; iš jų 100 sirgo TBC. Sergančių ir nesergančių TBC vaikų testo rezultatai pateikti 2.1 lentelėje.

2.1 lentelė. Sergančių ir nesergančių TBC vaikų testo rezultatai

		SERGAMUMAS				Iš viso	
		Serga		Neserga			
		N	%	N	%	N	%
TESTO REZULTATAS	Teigiamas	96	(96)	594	(6)	69	(7)
	Neigiamas	4	(4)	9 306	(94)	9 310	(93)
	Iš viso	100	(100)	9 900	(100)	10 000	(100)

Vaikų susirgimo TBC tikimybę, kai tuberkulino testo rezultatas teigiamas, $P(D+|T+)$ skaičiuosime naudodamiesi Bajeso formule (2.1). 2.1 lentelės duomenimis, epidemiologinių duomenų pagrindu įvertintos tikimybės (statistinės) yra tokios: $P(D+) = 100/10\,000 = 0{,}01$; $P(D-) = 0{,}99$; $P(T+|D+) = 0{,}96$; $P(T+|D-) = 0{,}06$. Todėl:

$P(D+|T+) = 0{,}96\times0{,}01 / (0{,}96\times0{,}01 + 0{,}06\times0{,}99) = 0{,}139$.

Nustatėme, kad tik 13,9 % vaikų, kurių tuberkulino testo rezultatas teigiamas, serga TBC.

2.4. Atsitiktiniai dydžiai

Atsitiktinai atrinkto populiacijos individo sveikatos būklės bei kitus rodiklius galima laikyti atsitiktiniais. Taigi, fiksuodami atsitiktinai į studiją įtraukto ligonio tyrimų rezultatus, susiduriame su atsitiktiniais įvykiais. Atsitiktinis įvykis yra glaudžiai susijęs su atliekamu eksperimentu. Pavyzdžiui, klinikinėje epidemiologijoje tirdami koronarų pažeidimo būklę, susiduriame su atsitiktiniais įvykiais: „nėra pažeistų koronarų“, „koronarų

pažeidimai neviršija 50 %“, „koronarų pažeidimai viršija 50 %“. Epidemiologai, analizuojantys CD paplitimą, susiduria su įvykiais: „individas neserga CD“, „individas serga I^o CD“, „individas serga II^o CD“. Sociologas, atsitiktinai sutiktam piliečiui pateikęs klausimą: „Kaip vertinate politiko *X* veiklą?“, susiduria su atsitiktiniais įvykiais: „vertinu teigiamai“, „vertinu neigiamai“ ir „neturiu nuomonės“. Minėtų tyrimų (eksperimentų) tikslai skiriasi, skirtingi elementarūs įvykiai, tačiau visų trijų eksperimentų metu gaunami 3 elementarūs įvykiai, t. y. visų šių eksperimentų elementarių įvykių aibė yra $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Todėl reikalingi abstraktesni tikimybiniai modeliai, apibendrinantys daug iš prigimties skirtingų eksperimentų. Tam reikalinga *atsitiktinio dydžio* sąvoka (matematiškai tikslus atsitiktinio dydžio apibrėžimas pateiktas vadovėliuose [2, 4, 8, 9]).

Atsitiktinis dydis (ats. d.) yra atsitiktinio įvykio vienareikšmė skaitinė funkcija, t. y. ats. d. nusako taisyklę, pagal kurią kiekvienam atsitiktiniam įvykiui priskiriama skaitinė reikšmė. Pavyzdžiui, užuot nagrinėję visus 3 anksčiau minėtų tyrimų metu gautus įvykius, galime nagrinėti atsitiktinį dydį *X*:

$$X = \begin{cases} 0, \text{ jei individas neserga CD} \\ 1, \text{ jei individas serga I}^o \text{ CD} \\ 2, \text{ jei individas serga II}^o \text{ CD} \end{cases} \quad \text{arba} \quad X = \begin{cases} 1, \text{ jei „vertinu teigiamai“} \\ 2, \text{ jei „vertinu neigiamai“} \\ 3, \text{ jei „neturiu nuomonės“} \end{cases}$$

Atsitiktinis dydis vadinamas **diskrečiuoju**, jei jis įgyja baigtinį arba suskaičiuojamą reikšmių skaičių. Diskretusis atsitiktinis dydis *X* visiškai apibrėžiamas išvardijus įgyjamas reikšmes x_i ir jų įgijimo tikimybes p_i:

x	x_1	x_2	x_3	...
$P\{X=x\}$	p_1	p_2	p_3	...

Pavyzdžiui, ats. d. *X* – berniukų skaičius 3 vaikų šeimoje. Ats. d. *X* įgyjamos reikšmės yra 0, 1, 2, 3, šių reikšmių tikimybės apskaičiuotos 2.3 skyriuje. Ats. d. *X* skirstinys yra:

x	0	1	2	3
$P\{X=x\}$	0,11765	0,36745	0,38235	0,13265

Kokybinio kintamojo tikimybinis (statistinis) modelis yra diskretusis ats. dydis. Pavyzdžiui, ats. d., įgyjantis 1, 2, 3 reikšmes (arba bet kokias 3 skirtingas reikšmes) su nelygiomis 0 tikimybėmis, yra visų 3 anksčiau minėtų eksperimentų tikimybinis modelis. Ats. d. *Y*, įgyjantis 2 reikšmes – 1 su tikimybe π ir 0 su tikimybe $(1 - \pi)$ – yra daugelio fizine prasme skirtingų eksperimentų, kurių metu galimos tik 2 išeitys, tikimybinis modelis.

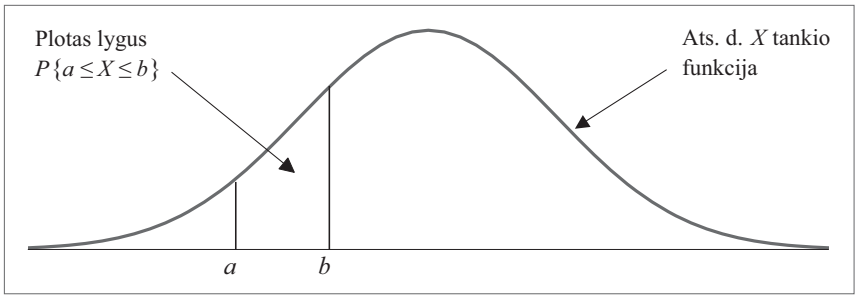
Atsitiktinis dydis X vadinamas **absoliučiai tolydžiuoju**, jei egzistuoja neneigiama funkcija $p(x)$, tokia, kad:

$$P\{X \le x\} = \int_{-\infty}^{x} p(y)dy.$$

Funkcija $F(x) = P\{X \le x\}$ vadinama ats. d. skirstinio funkcija. Ji visiškai charakterizuoja atsitiktinį dydį X. Šiame vadovėlyje absoliučiai tolydžius ats. dydžius vadinsime tiesiog tolydžiaisiais. Funkcija $p(x)$ vadinama atsitiktinio dydžio X **tankio funkcija,** arba **tankiu**. Tankio funkciją galima interpretuoti taip: tikimybė, kad tolydusis atsitiktinis dydis X pateks į intervalą $[a, b]$, lygi plotui, apribotam tankio kreive bei tiesėmis $y = 0$, $x = a$, $x = b$ (2.1 pav.). Jei atsitiktinis dydis X yra tolydusis, tuomet tikimybė, kad X lygus konkrečiai reikšmei, lygi 0:

$$P\{X = a\} = 0.$$

Atsitiktiniam dydžiui apibūdinti vartojamas terminas – **atsitiktinio dydžio skirstinys** – yra taisyklė, visiškai nusakanti ats. d. Tolydžiojo ats. d. atveju skirstinys – tankio nusakymo taisyklė (formulė). Diskrečiojo ats. d. atveju skirstinys – įgyjamų reikšmių ir tikimybių nustatymo taisyklė. Jei atsitiktinis dydis yra tolydusis, jo skirstinį vadinsime tolydžiuoju, o jei diskretusis – diskrečiuoju.



2.1 pav. Tankio funkcija ir ats. d. patekimo į intervalą [a, b] tikimybė

$$P\{a \le X \le b\} = \int_{a}^{b} p(y)dy$$

Jei tolydžiojo ats. d. X tankio funkcija yra simetriška (t. y. $p(x) = p(-x)$), ats. d. X skirstinys yra **simetriškas**.

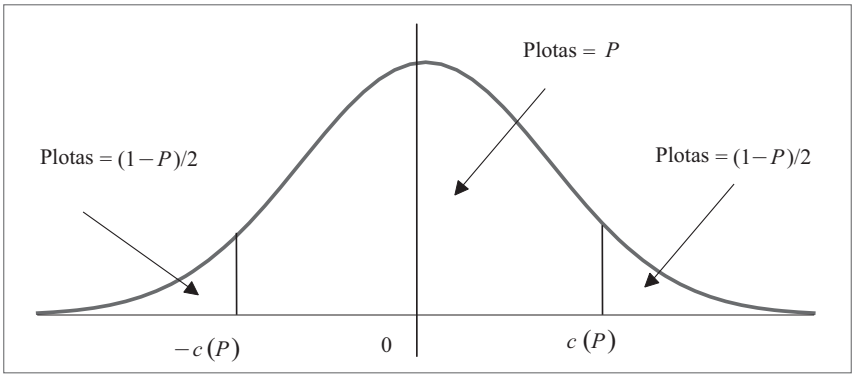
Kvantilio sąvoka. Žinodami ats. d. X tankio funkciją ir taškus a, b, galime apskaičiuoti tikimybę, kad X pateks į intervalą $[a, b]$, $P\{a \le X \le b\}$. Atvirkščiai, sakykime, turime skaičių P, esantį tarp 0 ir 1. Kaip rasti taškus $a(P)$ ir

$b(P)$, kad būtų teisinga lygybė $P\{a(P) \le X \le b(P)\} = P$? Atsakymas nėra vienintelis: tokių porų $a(P)$ ir $b(P)$ yra be galo daug. Uždavinį supaprastinsime: tegul ats. dydžio tankis yra simetriškas, $a(P) = -b(P) = c(P) > 0$ (2.2 pav.). Tuomet

$$P\{-c(P) \le X \le c(P)\} = P, \qquad (2.2)$$

o $P\{X > c(P)\} = (1 - P)/2$, $P\{X < -c(P)\} = (1 - P)/2$.

Šį uždavinį galime formuluoti ir taip: žinant plotą P po tankio kreive, reikia rasti jį ribojančius simetriškus taškus.



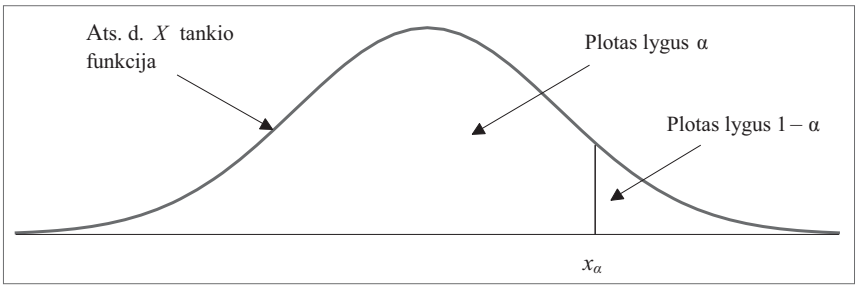
2.2 pav. Priklausomybės $P\{-c(P) \le X \le c(P)\} = P$ grafinis pateikimas, kai tankis yra simetriškas

Tokį uždavinį nesunku išspręsti įvedus kvantilio sąvoką. Sakykime, α yra skaičius tarp 0 ir 1, o X – atsitiktinis dydis. **Atsitiktinio dydžio X α lygio (eilės) kvantiliu (α kvantiliu) vadinamas skaičius x_α, tenkinantis nelygybę $P\{X < x_\alpha\} \le \alpha \le P\{X \le x_\alpha\}$.** Kitaip tariant, x_α yra mažiausia x reikšmė, tenkinanti nelygybę: $\alpha \le P\{X \le x\}$. Jei ats. d. X yra diskretusis, įgyjantis reikšmes $x_1, x_2 \ldots x_k \ldots$ su tikimybėmis $p_1, p_2 \ldots p_k \ldots$, tai x_k bus α lygmens kvantiliu, jei

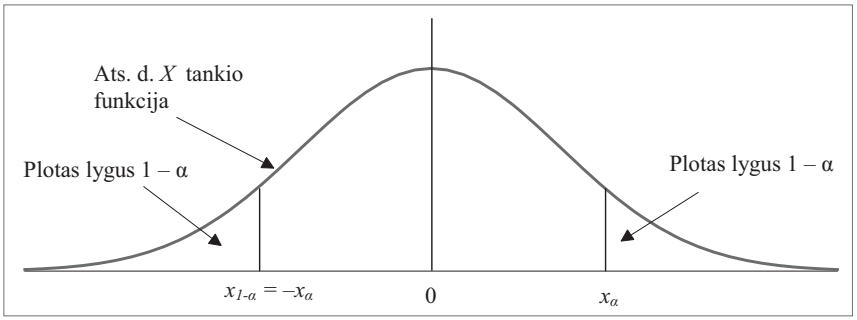
$$\sum_{i=1}^{k-1} p_i < \alpha \le \sum_{i=1}^{k} p_i .$$

Tolydžiojo atsitiktinio dydžio X kvantilis (2.3 pav.) apibrėžiamas lygybe: $P\{X \le x_\alpha\} = \alpha$ (jei skirstinio funkcija $P\{X \le x\}$ monotoniškai didėjanti; šis tvirtinimas teisingas visoms vadovėlyje pateiktoms tolydžiojo ats. d. funkcijoms), be to, ats. d. X teisinga lygybė $P\{X > x_\alpha\} = 1 - \alpha$. Kvantilių terminais, (2.2) formulėje $-c(P) = x_{(1-P)/2} = -x_{(1+P)/2}$; $c(P) = x_{(1+P)/2}$ arba $P\{-x_{(1+P)/2} \le X \le x_{(1+P)/2}\} = P$. Pažymėję $P = 1 - \alpha$, (2.1) formulę galime perrašyti taip: $P\{x_{\alpha/2} \le X \le x_{1-\alpha/2}\}$. Simetriško nulio atžvilgiu skirstinio kvantiniai tenkina sąlygą: $x_\alpha = -x_{1-\alpha}$ (2.4 pav.).

Kituose skyriuose susidursime dažniausiai su tolydžiųjų skirstinių kvantiliais. Kai kurių skirstinių kvantiliai pateikti lentelėse arba statistinėmis funkcijomis skaičiuoklėse. Matematinės statistikos lentelėse dažniausiai pateikiami ne skirstinio kvantiliai, o jiems ekvivalenčios charakteristikos – kritinės reikšmės. Pagal apibrėžimą, p-toji kritinė reikšmė sutampa su $(1 - p)$ lygio kvantiliu.



2.3 pav. Tolydžiojo ats. d. kvantilis



2.4 pav. Tolydžiojo simetriško skirstinio kvantilis

α lygio kvantilis, kai α pateikiamas procentais, vadinamas α procentiliu. Pavyzdžiui, $x_{0,25}$ yra 0,25 lygio kvantilis, arba 25 % procentilis. 25 %, 50 % ir 75 % eilės procentiliai (arba 0,25, 0,5 ir 0,75 lygio kvantiliai) vadinami ats. d. **kvartiliais**.

Atsitiktinių dydžių nepriklausomumas. Ats. dydžiai X ir Y vadinami nepriklausomais, jei $P\{X \le x,\ Y \le y\} = P\{X \le x\} \times P\{Y \le y\}$ visiems realiems x ir y.

Ats. d. skaitinės charakteristikos. Kaip jau minėjome, atsitiktinį dydį visiškai charakterizuoja jo skirstinio funkcija $F(x) = P\{X \le x\}$ arba tolydžiuoju atveju – tankis, o diskrečiuoju atveju – įgyjamos reikšmės su tikimybėmis. Tačiau praktikoje atsitiktiniam dydžiui apibūdinti dažnai pakanka ir ne tokių išsamių charakteristikų.

Vidurkis. Tai atsitiktinio dydžio įgyjamų reikšmių grupavimosi centras. Ats. d. X vidurkis žymimas EX. Jei ats. d. yra diskretusis, jo vidurkis skaičiuojamas:

$$EX = x_1p_1 + x_2p_2 + \ldots + x_kp_k + \ldots = \sum_i x_i p_i ;$$

čia (x_i, p_i) – ats. d. įgyjamos reikšmės ir jų įgijimo tikimybės. Jei ats. d. yra tolydusis, jo vidurkis skaičiuojamas:

$EX = \int_{-\infty}^{\infty} xp(x)dx$; čia $p(x)$ – ats. d. X tankis.

Pagrindinės vidurkio savybės:

1) pastovaus dydžio vidurkis lygus jam pačiam: $EC = C$.
2) pastovų dydį galima iškelti prieš vidurkio ženklą: $E(CX) = CEX$.
3) ats. d. sumos vidurkis lygus jų vidurkių sumai: $E(X + Y) = EX + EY$.

2.5 pavyzdys. Nustatysime, kiek vidutiniškai 3 vaikų šeimoje yra berniukų. Pažymėkime ats. d. X – berniukų skaičius 3 vaikų šeimoje. Tuomet X vidurkis lygus:

$EX = 0\times0{,}11765 + 1\times0{,}36745 + 2\times0{,}38235 + 3\times0{,}13265 = 1{,}5301.$

Taigi 3 vaikų šeimoje vidutiniškai yra 1,53 berniukų.

Moda. Tolydžiojo skirstinio atveju moda yra tankio maksimumo taškas. Jei tankis turi vieną maksimumą, toks skirstinys vadinamas unimodaliniu. Diskrečiojo skirstinio atveju moda yra „patikimiausia“ reikšmė, t. y. reikšmė X, su kuria tikimybė $P\{X = x_k\}$ yra didžiausia. Atsitiktinis dydis su unimodaliniu skirstiniu turi vieną ats. d. grupavimosi (sankaupos) tašką.

Dispersija apibūdina ats. d. reikšmių išsibarstymą aplink vidurkį. Ats. d. X dispersija lygi:

$DX = E(X - EX)^2$.

Ats. d. X dispersiją kartais patogiau skaičiuoti pagal tokią formulę: $DX = EX^2 - (EX)^2$. Dydis EX^2 diskrečiojo skirstinio atveju lygus:

$$EX^2 = \sum_i x_i^2 p_i$$

ir atitinkamai tolydžiojo skirstinio atveju:

$$EX^2 = \int_{-\infty}^{\infty} x^2 p(x)dx \cdot$$

Pagrindinės dispersijos savybės:

1) dispersija neneigiama: $DX \geq 0$;
2) pastovaus dydžio dispersija lygi 0: $DC = 0$;

3) pastovus dydis prieš dispersijos ženklą iškeliama pakėlus kvadratu: $D(CX) = C^2DX$;
4) jei ats d. X ir Y yra nepriklausomi, jų sumos dispersija lygi dispersijų sumai: $D(X + Y) = DX + DY$.

Standartinis nuokrypis σ yra kvadratinė šaknis iš dispersijos: $\sigma = \sqrt{DX}$. Standartinis nuokrypis taip pat charakterizuoja ats. d. išsibarstymą aplink vidurkį, tik σ dimensija sutampa su ats. d. X dimensija.

Asimetrijos koeficientas lygus $E(X - EX)^3/(DX)^{3/2}$. Jis apibūdina ats. d. skirstinio simetriškumą vidurkio atžvilgiu.

Ekscesas lygus $E(X - EX)^4/(DX)^2 - 3$. Tolydžiojo skirstinio atveju ekscesas apibūdina tankio kreivės smailiaviršūniškumą.

2.5. Normalusis skirstinys. Skirstiniai, susiję su normaliuoju

Statistikoje dažniausiai naudojamas skirstinys yra normalusis, arba Gauso. Sakoma, kad ats. d. X skirstinys yra normalusis, jei tankis apibrėžiamas pagal formulę:

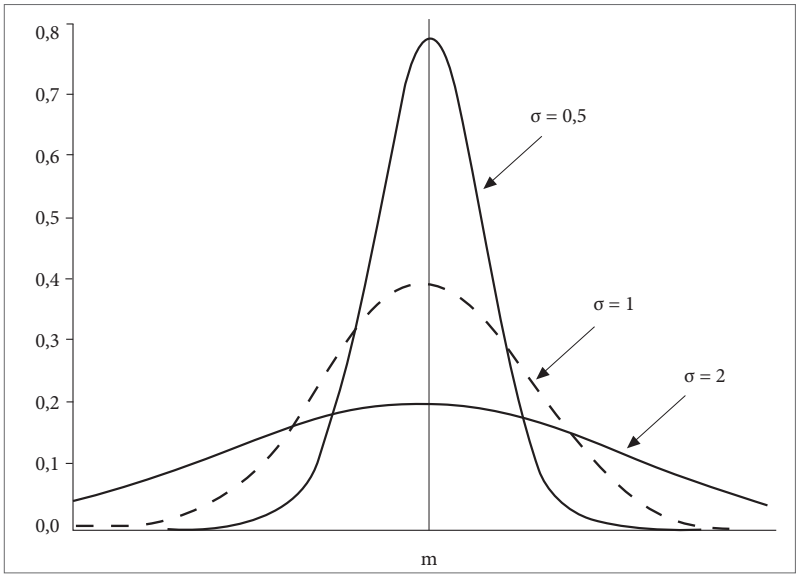
$$\varphi_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-(x-m)^2/(2\sigma^2)\}; \tag{2.3}$$

čia m ir σ – normaliojo skirstinio parametrai; m gali būti bet koks skaičius, σ – tik teigiamas. Normaliojo skirstinio tankio funkcija yra varpo formos, turinti maksimumą taške m bei simetriška taško m atžvilgiu. Parametras m reguliuoja tankio „centrą“ (varpo viršūnę), parametras σ – varpo „plotį“ (2.5 pav.). Kuo mažesnis σ, tuo labiau suspausta (2.3) kreivė, kuo didesnis σ, tuo labiau išsiplėtęs tankis (2.5 pav.).

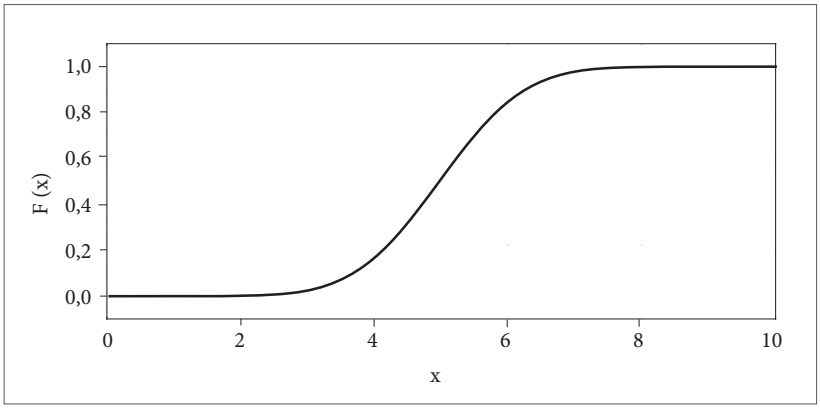
Jei atsitiktinio dydžio X skirstinys yra normalusis su parametrais m ir σ^2 (šis faktas simboliškai žymimas $X \sim N(m, \sigma^2)$), tai ats. d. X vidurkis lygus m, dispersija σ^2, standartinis nuokrypis – σ: $EX = m$, $DX = \sigma^2$. Parametras σ charakterizuoja atsitiktinio dydžio išsibarstymą aplink vidurkį. Normaliojo skirstinio asimetrijos koeficientas lygus 0 (tankis simetriškas vidurkio atžvilgiu), ekscesas lygus 0, o skirstinio funkcija lygi (2.6 pav.):

$$\Phi(x,m,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\{-(y-m)^2/(2\sigma^2)\}dy. \tag{2.4}$$

Jei ats. d. X skirstinys yra normalusis su parametrais (m, σ^2), tai ats. dydžio X tiesinės transformacijos $Y = aX + b$ (taip pat ats. dydžio) skirstinys bus normalusis su parametrais $(am + b, a^2\sigma^2)$. Tai įrodoma remiantis vidurkio ir dispersijos savybėmis: $EY = E(aX + b) = E(aX) + Eb = am + b$; $DY = D(aX + b) = D(aX) = a^2DX = a^2\sigma^2$.



2.5 pav. Normaliojo skirstinio tankis

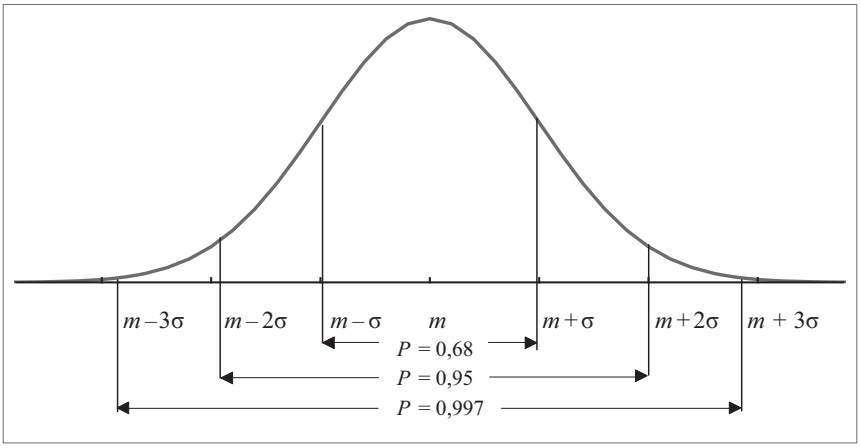


2.6 pav. Normaliojo ats. d. skirstinio funkcija (m = 5, σ = 1)

Normalusis skirstinys pasižymi stabilumu: nepriklausomų normaliųjų ats. d. suma taip pat yra normalusis atsitiktinis dydis. Jei nepriklausomo atsitiktinio dydžio X_i skirstinys yra normalusis su parametrais (m_i, σ_i^2), $i = 1 \ldots n$, tuomet sumos $S_n = (X_1 + X_2 + ... + X_n)$ skirstinys bus normalusis. Sumos S_n skirstinio parametrai ES_n ir DS_n nustatomi remiantis vidurkio ir dispersijos savybėmis: $ES_n = E(X_1 + X_2 + \ldots + X_n) = EX_1 + EX_2 + \ldots + EX_n = m_1 + m_2 + ... + m_n$, $DS_n = DX_1 + DX_2 + \ldots + DX_n = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$. Taigi sumos S_n skirstinys yra normalusis su parametrais $(m_1 + m_2 + ... + m_n,\ \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2)$.

Normaliajam skirstiniui teisinga vienos, dviejų ir trijų sigmų taisyklė:
- patekimo į intervalą $[m - \sigma, m + \sigma]$ tikimybė – maždaug 0,68;
- patekimo į intervalą $[m - 2\sigma, m + 2\sigma]$ tikimybė – maždaug 0,95;
- patekimo į intervalą $[m - 3\sigma, m + 3\sigma]$ tikimybė – maždaug 0,997.

Vienos, dviejų ir trijų sigmų taisyklė iliustruota 2.7 pav.



2.7 pav. Vienos, dviejų ir trijų sigmų taisyklė

Atskirą normaliojo skirstinio atvejį, kai $m = 0$, $\sigma = 1$, vadiname standartiniu normaliuoju skirstiniu. Jis žymimas $X \sim N(0, 1)$. Standartinio normaliojo skirstinio tankis lygus

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2 / 2\},$$

skirstinio funkcija –

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-y^2 / 2\} dy.$$

Tankio $\varphi(x)$ ir skirstinio funkcijos $\Phi(x)$ reikšmės naudojamos apytiksliams kai kurių tikimybių vertinimams, todėl pateikiamos lentelėmis ir skaičiuojamos duomenų apdorojimo paketais. Užtenka žinoti šių funkcijų reikšmes tik teigiamam argumentui, kadangi tankis $\varphi(x)$ yra simetriškas, o $\Phi(-x) = 1 - \Phi(x)$.

Standartinio normaliojo skirstinio α lygmens kvantilį žymėsime z_α. Kadangi skirstinys yra simetriškas, todėl teisinga lygybė $z_\alpha = -z_{1-\alpha}$. Normaliojo skirstinio kvantiliai z_α didesnėms nei 0,5 α reikšmėms statistikos vadovėliuose

pateikti lentele (1 lentelė) ir skaičiuoklėse statistine funkcija. Dažniausiai statistikos skaičiavimams naudojami standartinio normaliojo skirstinio kvantiliai: $z_{0,975} = 1,96$; $z_{0,95} = 1,645$; $z_{0,995} = 2,575$. Standartinio normaliojo skirstinio 25 %, 50 % ir 75 % eilės procentiliai (arba kvartiliai) atitinkamai lygūs 0,6745; 0 ir 0,6745.

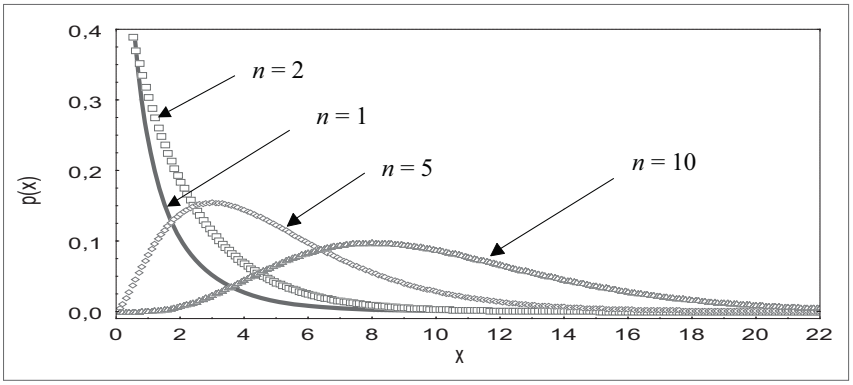
Jei ats. d. X skirstinys yra normalusis su parametrais (m, σ^2), tuomet standartizuoto ats. d. $Z = (X - m)/\sigma$ skirstinys bus standartinis normalusis. Jei $Z \sim N(0, 1)$, tai ats. d. $X = \sigma X + m \sim N(m, \sigma^2)$.

Skirstiniai, susiję su normaliuoju. Pateiksime atsitiktinių dydžių, sudarytų iš nepriklausomų standartinių normaliųjų dydžių funkcijų, skirstinius. Šie skirstiniai naudojami statistikoje hipotezėms tikrinti.

χ^2 skirstinys. Sakykime, $X_1, X_2 \ldots X_n$ – nepriklausomi standartinį normalųjį skirstinį turintys ats. d. Tada atsitiktinio dydžio

$$\chi_n^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

skirstinys vadinamas χ^2 skirstiniu su n laisvės laipsnių. Šio skirstinio tankio grafikas kelioms n reikšmėms pateiktas 2.8 pav. Ats. dydžio χ_n^2 vidurkis lygus laisvės laipsnių skaičiui n, dispersija lygi $2n$. χ^2 skirstinio su n laisvės laipsnių α lygmens kvantilį žymėsime $\chi_\alpha^2(n)$. Šio skirstinio kvantiliai statistikos vadovėliuose pateikiami lentele (2 lentelė) ir statistine funkcija skaičiuoklėse. Dažniausiai naudojami 0,95 lygio kvantiliai: $\chi_{0,95}^2(1) = 3,841$; $\chi_{0,95}^2(2) = 5,991$; $\chi_{0,95}^2(3) = 7,815$.

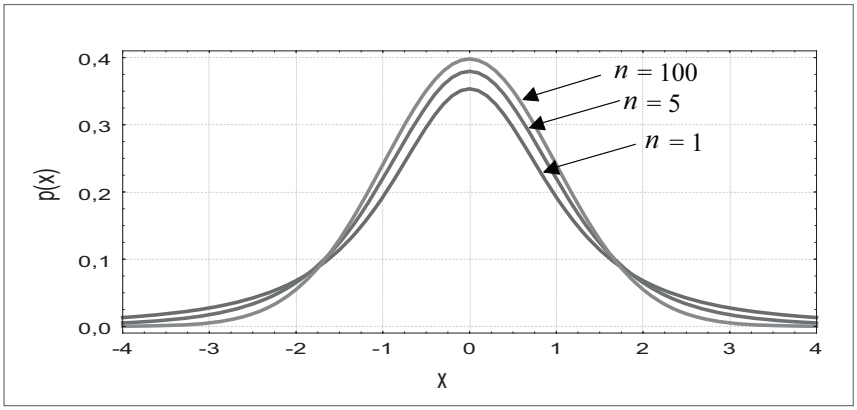


2.8 pav. χ^2 skirstinio tankis įvairiems n

Stjudento (*Student*) skirstinys. Sakykime, $X, X_1, X_2 \ldots X_n$ – nepriklausomi standartinį normalųjį skirstinį turintys ats. d. Tada atsitiktinio dydžio

$$t_n = X / \sqrt{(X_1^2 + X_2^2 + ... + X_n^2)/n} = X / \sqrt{\chi_n^2 / n}$$

skirstinys vadinamas Stjudento (arba *t*) skirstiniu su *n* laisvės laipsnių. Stjudento skirstinio tankio grafikas keletui *n* pateiktas 2.9 pav. Ats. dydžio t_n vidurkis lygus 0, dispersija lygi $\sqrt{n/(n-2)}$. t_n skirstinio α lygmens kvantilį žymėsime $t_\alpha(n)$. t_n skirstinio kvantiliai pateikti lentele (3 lentelė) ir statistine funkcija skaičiuoklėse. Kai kurios Stjudento skirstinio kvantilių reikšmės: $t_{0,95}(5) = 2{,}015$; $t_{0,95}(10) = 1{,}812$; $t_{0,975}(3) = 3{,}182$; $t_{0,975}(9) = 2{,}262$.

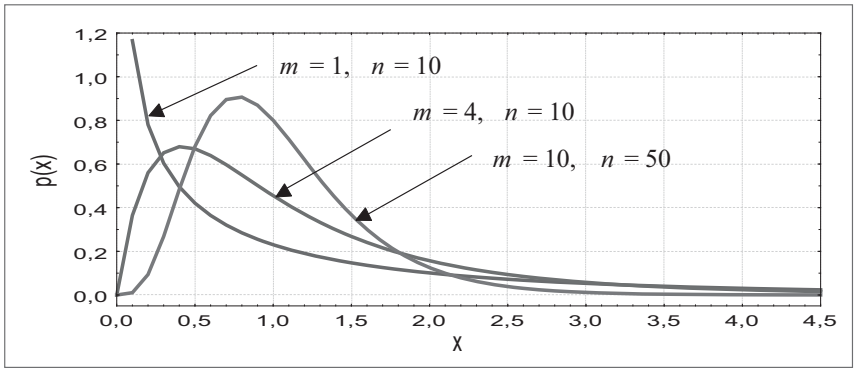


2.9 pav. Stjudento skirstinio tankis įvairiems n

Fišerio skirstinys. Sakykime, $X_1, X_2 ... X_m$, $Y_1, Y_2 ... Y_n$ – nepriklausomi standartinį normalųjį skirstinį turintys ats. d. Tada atsitiktinio dydžio

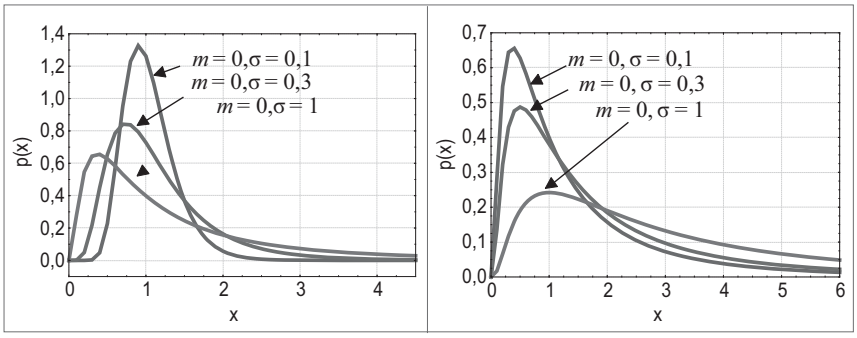
$$F_{m,n} = ((X_1^2 + X_2^2 + ... + X_m^2)/m)/((Y_1^2 + Y_2^2 + ... + Y_n^2)/n) =$$
$$(\chi_m^2 / m)/(\chi_n^2 / n)$$

skirstinys vadinamas Fišerio skirstiniu (*F* skirstiniu) su *m* ir *n* laisvės laipsnių. Fišerio skirstinio tankio grafikas įvairiems *m* ir *n* pateiktas 2.10 pav. Pastebėtina, kad ats. d. $F_{1,n}$ yra Stjudento ats. dydžio t_n kvadratas. $F_{m,n}$ skirstinio α lygmens kvantilį žymėsime $F_\alpha(m,n)$. $F_{m,n}$ skirstinio kvantiliai pateikti lentele (4 lentelė) ir statistine funkcija skaičiuoklėse. Fišerio skirstinio kvantiliams galioja lygybė: $F_\alpha(m,n) = F_{1-\alpha}(n,m)$, todėl užtenka žinoti tik mažesnio nei 0,5 lygio kvantilius. Pateikiame kai kurių Fišerio skirstinio kvantilių reikšmes: $F_{0,95}(1, 5) = 6{,}608$; $F_{0,95}(1, 10) = 4{,}965$; $F_{0,95}(4, 20) = 2{,}866$.



2.10 pav. Fišerio skirstinio tankis įvairiems m ir n

Lognormalusis skirstinys. Sakykime, ats. d. *X* įgyja tik teigiamas reikšmes. Ats. d. *X* skirstinys yra lognormalusis, jei ats. dydis ln*X* turi normalųjį skirstinį. Tai asimetriškas skirstinys. Šio skirstinio tankis įvairiems parametrams *m* ir σ pateiktas 2.11 pav.



2.11 pav. Lognormaliojo skirstinio tankis

Stjudentizuotas skirtumas*. Sakykime, X_1, X_2 ... X_n – nepriklausomi normalųjį skirstinį su parametrais (m, σ^2) turintys ats. d., s^2 – dispersijos σ^2 įvertis su k laisvės laipsnių:

$$s^2 = (\sum_{i=1}^{k+1}(X_i - \overline{X})^2)/k\,, \quad \overline{X} = (\sum_{i=1}^{k+1} X_i)/(k+1)\,.$$

Pagal apibrėžimą ks^2/σ^2 yra nuo X_1, X_2 ... X_n nepriklausantis atsitiktinis dydis, turintis χ^2 skirstinį su k laisvės laipsnių. Pažymėkime $R = max_i\, X_i - min_j\, X_j$. Atsitiktinis dydis $\xi = R/s$ vadinamas stjudentizuotu skirtumu su n ir k laisvės laipsnių. ξ skirstinys žymimas $Q(n, k)$. Jis naudojamas dispersinėje analizėje (žr. 12 skyrių). Nedideliems n ir k sudarytos šio skirstinio kvantilių lentelės.

2.6. Kiti dažnai naudojami skirstiniai

Tolydžiųjų skirstinių pavyzdžiai. Eksponentinis skirstinys. Ats. d. X skirstinys yra eksponentinis su parametru $\theta > 0$, jei X tankis lygus:

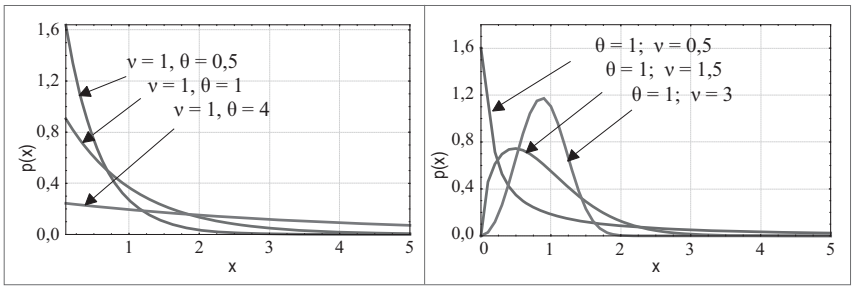
$p(x) = (1/\theta)\exp(-x/\theta)$, kai $x > 0$; ir $p(x) = 0$, kai $x \le 0$.

θ yra skirstinio mastelio parametras (*scale parameter*). Eksponentinis skirstinys taikomas išgyvenamumo duomenų analizei (žr. 13 skyrių). Ats. d. X skaitinės charakteristikos yra: $EX = \theta$, $DX = \theta^2$. Eksponentinio ats. d. skirstinio funkcija $F(x)$, kai $x > 0$, lygi $1 - \exp(-x/\theta)$.

Veibulo (*Weibull*) skirstinys. Ats. d. X skirstinys yra Veibulo su parametrais $\theta > 0$ ir $\nu > 0$, jei X tankis lygus:

$$p(x) = \frac{\nu}{\theta}\left(\frac{x}{\theta}\right)^{\nu-1} \exp\left(-\left(\frac{x}{\theta}\right)^{\nu}\right), \text{ kai } x > 0; \text{ ir } p(x) = 0, \text{ kai } x \le 0.$$

θ yra mastelio (*scale*), ν – formos (*shape*) parametras (2.13 pav.). Eksponentinis skirstinys yra Veibulo skirstinio atskiras atvejis ($\nu = 1$). Kaip ir eksponentinis skirstinys, Veibulo skirstinys taikomas išgyvenamumo duomenų analizei (žr. 13 skyrių). Veibulo ats. d. skirstinio funkcija $F(x)$, kai $x > 0$, lygi $1 - \exp(-(x/\theta)^{\nu})$. Veibulo skirstinio tankio grafikas įvairiems θ ir ν pateiktas 2.12 pav.



2.12 pav. Veibulo skirstinio tankis

Bernulio (*Bernoulli*) skirstinys. Sakoma, kad ats. d. X skirstinys yra Bernulio su parametru π, $0 < \pi < 1$, jei X įgyja dvi reikšmes – 1 ir 0 su tikimybėmis: $P\{X = 1\} = \pi$, $P\{X = 0\} = 1 - \pi$. Pagal apibrėžimą $P\{X = k\} = \pi^k(1 - \pi)^{1-k}$; čia $k = 0, 1$. Bernulio ats. d. skaitinės charakteristikos: $EX = \pi$; $DX = \pi(1 - \pi)$.

Binominis skirstinys. Tarkime, atlikta n nepriklausomų eksperimentų, kurių metu galimos tik dvi baigtys: „įvykis A įvyko“ ir „įvykis A neįvyko“. Kiekvieno eksperimento metu įvykio A tikimybė lygi π (įvykio A nepasi-

rodymo tikimybė lygi 1 – π); čia 0 < π < 1. Binominiu atsitiktiniu dydžiu *X* vadinamas įvykio *A* pasirodymų skaičius. Binominio ats. d. *X* skirstinys priklauso nuo dviejų parametrų – π ir *n* ir tai žymima $X \sim B(n, \pi)$. Tikimybė, kad binominis ats. d. įgis reikšmę, lygią *k*, yra

$P\{X = k\} = C_n^k \pi^k (1 - \pi)^{n-k}$, $k = 0, 1, 2 \ldots n$;

čia C_n^k – binominiai koeficientai, skaičiuojami pagal formulę:

$$C_n^k = \frac{n!}{k!(n-k)!} = \frac{n(n-1)...(n-k+1)}{1 \times 2 \times ... \times k},\ k!=1\times2\times\ldots\times k,\ 0!=1.$$

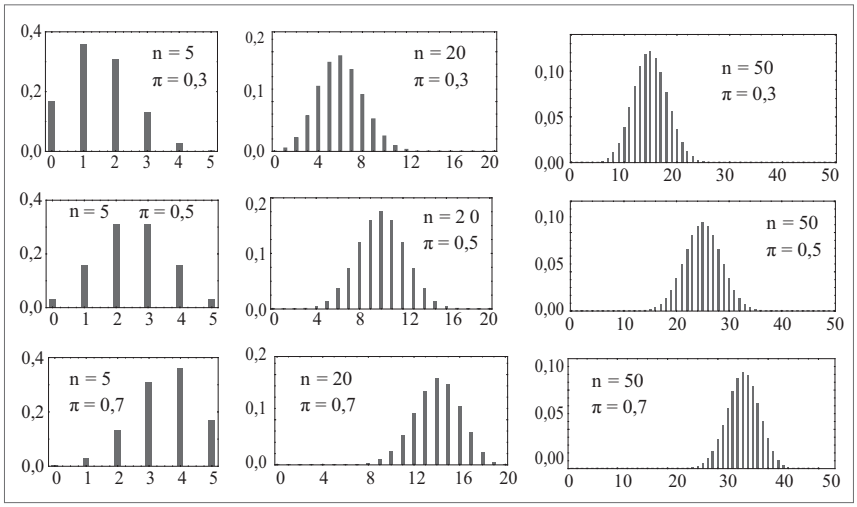
Pavyzdžiui, $C_5^2 = \frac{5!}{2!3!} = \frac{5 \times 4}{2} = 10$. Pagal apibrėžimą $C_n^1 = n$; $C_n^2 = n(n-1)/2$, $C_n^0 = 1$,

$P\{X = 0\} = (1 - \pi)^n$; $P\{X = 1\} = n\pi(1 - \pi)^{n-1}$; $P\{X = 2\} = (1/2)n(n - 1)\pi^2(1 - \pi)^{n-2}$.

Binominio ats. d. vidurkis lygus $n\pi$, dispersija – $n\pi(1 - \pi)$.

Binominis skirstinys naudojamas įvykiams per tam tikrą laiko tarpą modeliuoti. Pavyzdžiui, berniukų skaičius 3 vaikų šeimoje yra ats. dydis, turintis binominį skirstinį, kurio parametrai n = 3 ir π = P{gimė berniukas} = 0,51. 2.2 skyriaus pavyzdyje pateiktas šio ats. d. skirstinys.

Binominio skirstinio pavyzdžiai įvairiems π ir *n* pateikti 2.13 pav. Ats. dydžio $X \sim B(10, 0{,}3)$ skirstinys pateiktas 2.5 lentelėje. Joje matome, kad šio skirstinio 0,05; 0,95 ir 0,975 lygio kvantiliai yra: $x_{0,05} = 0$; $x_{0,95} = 5$; $x_{0,975} = 6$.



2.13 pav. Binominio skirstinio pavyzdžiai

2.5 lentelė. Binominis skirstinys; n = 10, π = 0,3

k	*P*{*X* = *k*}	*k*	*P*{*X* = *k*}
0	$(0,7)^{10} = 0,028248$	6	$210\ (0,3)^3(0,7)^7 = 0,03676$
1	$10\ (0,3)(0,7)^9 = 0,121061$	7	$120(0,3)^3(0,7)^7 = 0,009$
2	$45\ (0,3)^2(0,7)^8 = 0,233474$	8	$45\ (0,3)^3(0,7)^7 = 0,00135$
3	$120(0,3)^3(0,7)^7 = 0,266828$	9	$10\ (0,3)^3(0,7)^7 = 0,00014$
4	$210\ (0,3)^3(0,7)^7 = 0,20012$	10	$(0,3)^{10} = 0,0000059$
5	$252\ (0,3)^3(0,7)^7 = 0,10292$		

Puasono (*Poisson*) skirstinys. Šis skirstinys dar vadinamas retų įvykių skirstiniu. Puasono ats. d. *X* įgyja neneigiamas sveikas reikšmes su tikimybėmis:

$P\{X = k\} = (\lambda^k/k!)e^{-\lambda}$, $k = 0, 1, 2 \ldots$;

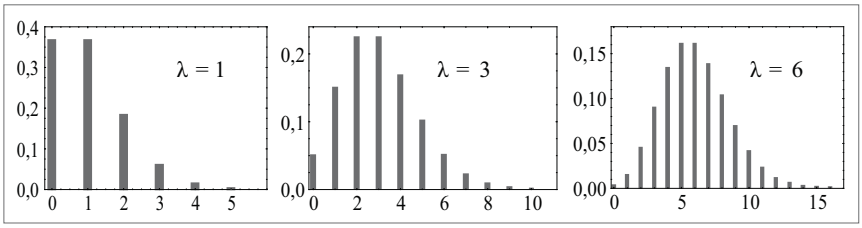
čia $\lambda > 0$ – skirstinio parametras.

Pagal apibrėžimą

$P\{X = 0\} = e^{-\lambda}$; $P\{X = 1\} = \lambda e^{-\lambda}$; $P\{X = 2\} = (1/2)\lambda^2 e^{-\lambda}$ …

Puasono skirstinio vidurkis ir dispersija lygūs λ. Jei *X* nusako įvykių, įvykusių per tam tikrą laiką, skaičių, tuomet λ – vidutinis įvykių per tam tikrą laiką skaičius. Puasono skirstinys įvairiems λ pateiktas 2.14 pav.

Puasono skirstinys naudojamas retiems įvykiams, pavyzdžiui, sergamumui leukemija, modeliuoti.



2.14 pav. Puasono skirstinys

Hipergeometrinis skirstinys. Sakykime, urnoje yra *n* elementų, iš kurių *m* yra pirmos rūšies, *n* – *m* – antros rūšies. Iš šios urnos atsitiktinai traukiame *k* elementų. Pažymėkime *X* – pirmos rūšies elementų skaičių iš *k* ištrauktų. Tuomet ats. d. *X* skirstinys yra hipergeometrinis, kurio parametrai (*n, m, k*) ($X \sim H(n, m, k)$):

$P\{X = i\} = C_m^i C_{n-m}^{k-i} / C_n^k$, $\max(0, m + k - n) \le i \le \min(m, k)$.

Hipergeometrinio skirstinio skaitinės charakteristikos yra:

$EX = km/n$, $DX = (km/n)(1 - k/n)(n - m)/(k - 1)$.

Hipergeometrinio skirstinio pavyzdys. Studijoje dalyvavo n ligonių, iš kurių m gydyti preparatu A, kiti $n - m$ – preparatu B. Iš studijos atsitiktinai atrinkta k ligonių. Tuomet preparatu A gydytų ligonių skaičiaus skirstinys yra hipergeometrinis, kurio parametrai yra (n, m, k).

2.7. Eksponentinių skirstinių šeima*

Sakoma, kad ats. d. X skirstinys priklauso eksponentinių skirstinių šeimai, jei jo tikimybinis tankis (tolydžiojo skirstinio atveju) ar tikimybė (diskrečiojo skirstinio atveju) $p(x)$ išreiškiama šita forma:

$$p(x) = \exp\left(\frac{x\theta - B(\theta)}{A(\varphi)} + C(x,\varphi)\right); \tag{2.5}$$

čia $A(\varphi)$, $B(\theta)$, ir $C(x, \varphi)$ – žinomo pavidalo funkcijos. Parametras θ vadinamas natūraliuoju skirstinio parametru, parametras $\varphi > 0$ yra papildomas dispersijos parametras (*scale parameter*). Funkcija $A(\varphi)$ priklauso tik nuo parametro φ, $B(\theta)$ – tik nuo parametro θ, o $C(x, \varphi)$ nuo parametro θ nepriklauso.

Eksponentinių skirstinių šeimai priklausančio ats. d. X vidurkis $m = E(X)$ ir dispersija $D(X)$ yra susiję su parametru θ ir φ tokia priklausomybe:

$$m = \partial B(\theta)/\partial\theta, \;\; DX = A(\varphi)\partial^2 B(\theta)/\partial\theta^2. \tag{2.6}$$

Daugelis iš 2.5–2.6 skyriuose pateiktų skirstinių priklauso eksponentinių skirstinių šeimai.

2.6 pavyzdys (normalusis skirstinys). Normaliojo skirstinio, kurio parametrai yra m ir σ^2, tankis $p(x)$ gali būti pertvarkomas taip:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) = \exp(-0{,}5\ln(2\pi\sigma^2))\exp\left(-\frac{x^2 - 2xm + m^2}{2\sigma^2}\right) =$$

$$= \exp\left(\frac{mx - m^2/2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right).$$

Pagal (2.5) apibrėžimą normalusis skirstinys priklauso eksponentinių skirstinių šeimai, kurios natūralus parametras lygus vidurkiui: $\theta = m$. Dispersijos parametras $\varphi = \sigma^2$, $A(\varphi) = \varphi$, $B(\theta) = \theta^2/2$, o $C(x, \varphi) = -0{,}5(x^2/\varphi + \ln(2\pi\sigma^2))$. Nesunku patikrinti, kad vidurkiui bei dispersijai teisinga (2.6) formulė:

$$\partial B(\theta)/\partial\theta = \partial/\partial\theta\,(0{,}5\theta^2) = \theta = m; \;\; A(\varphi)\partial^2 B(\theta)/\partial\theta^2 = \sigma^2.$$

2.7 pavyzdys (binominis skirstinys). Pagal apibrėžimą (2.6 skyrius) binominio skirstinio, kurio parametrai n ir π, tikimybė $p(x) = P\{X = x\}$, $x = 0, 1 \ldots n$ (čia n fiksuotas) lygi:

$$p(x) = C_n^x \pi^x (1-\pi)^{n-x} = \exp(x\ln(\pi) + (n-x)\ln(1-\pi) + \ln(C_n^x)) =$$
$$\exp(x\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln(C_n^x)).$$

Matome, kad binominio skirstinio atveju natūralus parametras θ lygus $\ln\left(\frac{\pi}{1-\pi}\right)$, $\varphi = 1$, $B(\theta) = -n\ln(1-\pi) = n\ln(1+\exp(\theta))$, $C(x, 1) = \ln(C_n^x)$.

2.8 pavyzdys (Bernulio skirstinys). Bernulio ats. d. X įgyja dvi reikšmes: 1 su tikimybe π ir 0 su tikimybe $1 - \pi$. Šio skirstinio tikimybė lygi:

$$p(x) = \pi^x (1-\pi)^{1-x} = \exp(x\ln(\pi) + (1-x)\ln(1-\pi)) =$$
$$\exp(x\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi));$$

čia $x = 0; 1$. Matome, kad ir Bernulio skirstinio atveju natūralus parametras θ lygus $\ln\left(\frac{\pi}{1-\pi}\right)$, $\varphi = 1$, o $B(\theta) = -\ln(1-\pi) = -\ln(1+\exp(\theta))$, $C(x, 1) = 0$.

2.9 pavyzdys (Puasono skirstinys). Puasono skirstinio tikimybė lygi:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} = \exp(x\ln(\lambda) - \lambda + \ln(x!)) = \exp(x\ln(\lambda) + \ln(x!));$$

čia $x = 0, 1 \ldots$, λ – skirstinio parametras. Šio skirstinio atveju natūralus parametras θ lygus $\ln(\lambda)$, $\varphi = 1$, $B(\theta) = \lambda$, $C(x, 1) = \ln(x!)$.

Eksponentinės skirstinių klasės išraiška naudojama apibendrintų tiesinių modelių (10.11 skyrius) ryšio funkcijai nustatyti.

2.8. Didžiųjų skaičių dėsnis, centrinė ribinė teorema

Nagrinėdami įvairius statistikos uždavinius, susiduriame su atsitiktinių dydžių sumomis. Kai atsitiktinių dydžių skaičius ganėtinai didelis, pasireiškia tam tikri jų sumos skirstinio dėsningumai, nors vieno atsitiktinio dydžio įgyjamos reikšmės numatyti negalima – žinoma tik įgyjamos reikšmės tikimybė. Šiuos dėsningumus nusako dvi ribinių teoremų rūšys: didžiųjų skaičių dėsnis ir centrinė ribinė teorema.

Didžiųjų skaičių dėsnis tvirtina: labai tikėtina, kad didelio nepriklausomų atsitiktinių dydžių skaičiaus aritmetinis vidurkis mažai skirsis nuo tikrojo vidurkio. Pateiksime formalų didžiųjų skaičių dėsnio apibrėžimą.

Sakykime, $X_1, X_2 \ldots X_n$ – nepriklausomų atsitiktinių dydžių, turinčių baigtinius vidurkius EX_i, $i = 1 \ldots n$, seka. Sekai $X_1, X_2 \ldots X_n$ galioja didžiųjų skaičių dėsnis, jei:

$$P\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{EX_1 + EX_2 + \ldots + EX_n}{n}\right| < \varepsilon\} \to 1,$$

kai n neaprėžtai auga. Nustatyta: jei ats. d. X_i dispersijos aprėžtos, t. y. $DX_i < C$ visiems $i = 1 \ldots n$, tuomet šiai sekai galioja didžiųjų skaičių dėsnis.

Didžiųjų skaičių dėsnio galiojimo prielaida nusako, kada teorinį sumos $(X_1 + X_2 + \ldots + X_n)/n$ vidurkį, dažnai interpretuojamą kaip $X_1 \ldots X_n$ skirstinio parametrą, galima vertinti ats. d. vidurkio reikšme. Pavyzdžiui, registruodami kai kurių tyrimų duomenis, specialistai susiduria su rodiklių reikšmėmis, kurias galima interpretuoti kaip nepriklausomų ats. d. $X_1 \ldots X_n$ reikšmes. Sakykime, X_i – prietaiso, registruojančio nervinio signalo stiprumą, reikšmės. Žinoma, kad organizmas siunčia α stiprumo signalą, bet prietaisas signalą registruoja su atsitiktine paklaida. Minėto tyrimo tikslas – išmatuotomis signalo reikšmėmis įvertinti dydį α. Pagal apibrėžimą:

$X_i = \alpha + \xi_i$;

čia ξ_i – matavimo paklaida, t. y. nepriklausomi ats. d. su $E\xi_i = 0$, $D\xi_i = \sigma_i^2$, $EX_i = \alpha$, $i = 1 \ldots n$. Signalo stiprumas α vertintinas parodymų vidutine reikšme: $(X_1 + X_2 + \ldots + X_n)/n$; šio ats. d. vidurkis lygus α. Kad šis vertinimas būtų teisėtas, tikimybė, jog $|(X_1 + X_2 + \ldots + X_n)/n - \alpha| < \varepsilon$ turi būti artima 1, t. y. sekai $X_1 \ldots X_n$ turi galioti didžiųjų skaičių dėsnis. Jei matavimo paklaidų dispersijos σ_i^2 nedidėja, kai matavimų daugėja, tuomet signalo stiprumą galima vertinti matavimų vidurkiu; jei augant i, ganėtinai greitai auga σ_i^2, minėtas įvertis gali ir netikti.

Nustatyta: jei atsitiktiniai dydžiai $X_1 \ldots X_n$ turi vienodą skirstinį ir baigtinį vidurkį, labai tikėtina, kad jų aritmetinis vidurkis mažai skirsis nuo skirstinio vidurkio m ($m = EX_i$):

$P\{|(X_1 + X_2 + \ldots + X_n)/n - m| < \varepsilon\} \to 1$, $n \to \infty$;

čia $\varepsilon > 0$ – labai mažas skaičius.

Centrinė ribinė teorema teigia: didinant sumuojamų atsitiktinių dydžių skaičių, jų sumų skirstiniai panašėja į normalųjį skirstinį. Centrinė ribinė teorema gali būti suformuluota taip: jei $X_1 \ldots X_n$ yra nepriklausomi, turintys

vienodą skirstinį atsitiktiniai dydžiai su vidurkiu m ir baigtine dispersija σ^2, tai jų centruotos ir normuotos sumos skirstinio funkcija artima standartinio normaliojo dydžio skirstinio funkcijai, t. y.

$P\{(X_1 + X_2 + ... + X_n - nm)/(\sigma\sqrt{n}) \le x\} \to \Phi(x),\ n \to \infty.$

2.9. Daugiamačiai atsitiktiniai dydžiai (atsitiktiniai vektoriai)

Daugiamačiu ats. dydžiu vadinamas vektorius $\boldsymbol{X} = (X^{(1)}, X^{(2)} \dots X^{(p)})$, kurio komponentės yra atsitiktiniai dydžiai; čia p – vektoriaus $\boldsymbol{X}$ matavimų skaičius. Atsitiktiniai dydžiai $X^{(1)}, X^{(2)} \dots X^{(p)}$ vadinami vektoriaus $\boldsymbol{X}$ koordinatėmis, arba komponentėmis. Atsitiktinis vektorius (ats. v.) $\boldsymbol{X}$ vadinamas diskrečiuoju, jei jo komponentės yra diskretieji ats. d. Atsitiktinis vektorius $\boldsymbol{X}$ vadinamas tolydžiuoju, jei jo komponentės – tolydieji ats. d. Jei $\boldsymbol{X} = (X^{(1)}, X^{(2)} \dots X^{(p)})$ – tolydusis ats. v., tuomet egzistuoja neneigiama funkcija $p(x_1, x_2 \dots x_n)$, tokia, kad tikimybę $P\{X^{(1)} \le x_1, X^{(2)} \le x_2 \dots X^{(p)} \le x_p\}$ galima išreikšti taip:

$$P\{X^{(1)} \le x_1, X^{(2)} \le x_2 \dots X^{(p)} \le x_p\} = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_p} p(y_1, y_2 \dots y_p)dy_1 dy_2 \dots dy_p .$$

Funkcija $F(x_1, x_2 \dots x_p) = P\{X^{(1)} \le x_1, X^{(2)} \le x_2 \dots X^{(p)} \le x_p\}$ vadinama ats. vektoriaus $\boldsymbol{X}$ skirstinio funkcija.

Ats. dydžiai $X_1, X_2 \dots X_n$ vadinami **nepriklausomais**, jei jų bendra skirstinio funkcija lygi atskirų skirstinio funkcijų sandaugai. Jei $X_1, X_2 \dots X_n$ – tolydieji atsitiktiniai dydžiai, jie yra nepriklausomi tuomet, kai jų bendras tankis $p(x_1, x_2 \dots x_n)$ lygus atskirų ats. d. tankių $p_i(x_i)$ sandaugai:

$$p(x_1, x_2 \dots x_n) = p_1(x_1)\, p_2(x_2) \dots p_n(x_n).$$

Plačiau apžvelgsime dvimačius atsitiktinius dydžius.

Dvimatis diskretusis atsitiktinis dydis. Sakykime, X ir Y yra du diskretieji ats. d.; X įgyja r, o Y – c skirtingų reikšmių. Dvimačio ats. v. (X, Y) tikimybinis skirstinys (jungtinis X ir Y skirstinys) pateikiamas lentele, turinčia r eilučių ir c stulpelių. Eilutės atitinka X, stulpeliai – Y įgyjamas reikšmes. Lentelės gardelės atitinka visas galimas $r \times c$ dvimačio dydžio (X, Y) įgyjamas kombinacijas. Šių kombinacijų įgijimo tikimybes pažymėkime $\{\pi_{ij}\}$; čia $\pi_{ij} = P\{X = x_i, Y = y_j\}$ – tikimybė, kad (X, Y) reikšmė pateks į i-tosios eilutės ir j-tojo stulpelio susikirtime esančią gardelę (2.6 lentelė). Skirstinys $\{\pi_{ij}\}$ vadinamas X ir Y jungtiniu (dvimačiu) skirstiniu.

X vienmatis skirstinys

x	x_1	x_2	x_3	...	x_r
$P\{X = x\}$	π_{1+}	π_{2+}	π_{3+}	...	π_{r+}

apibrėžiamas susumavus jungtinio skirstinio tikimybes eilutėse

$$\pi_{i+} = \sum_{j=1}^{c} \pi_{ij},\ i = 1, 2 \ldots r,$$

o Y skirstinys

y	y_1	y_2	y_3	...	y_c
$P\{Y = y\}$	π_{+1}	π_{+2}	π_{+3}	...	π_{+c}

analogiškai apibrėžiamas susumavus jungtinio skirstinio tikimybes stulpeliuose:

$$\pi_{+j} = \sum_{i=1}^{r} \pi_{ij},\ j = 1, 2 \ldots c.$$

2.6 lentelė. Dvimačio ats. d. (X, Y) skirstinys (jungtinis X ir Y skirstinys)

$X \setminus Y$	y_1	y_2	...	y_j	...	y_c
x_1	π_{11}	π_{12}	...	π_{1j}	...	π_{1c}
x_2	π_{21}	π_{22}	...	π_{2j}	...	π_{2c}
...	...	...	...	...	...	...
x_i	π_{i1}	π_{i2}	...	π_{ij}	...	π_{ic}
...	...	...	...	...	...	...
x_r	π_{r1}	π_{r2}	...	π_{rj}	...	π_{rc}

Sąlyginiai dvimačių diskrečiųjų atsitiktinių vektorių skirstiniai. Y sąlyginis skirstinys, kai X įgyja i-tąją reikšmę, visiškai apibūdinamas tikimybėmis ($\pi_{1|i}$, $\pi_{2|i}$ ... $\pi_{c|i}$); čia $\pi_{j|i}$ – tikimybė, kad Y įgis j-tąją reikšmę su sąlyga, jog X įgijo i-tąją reikšmę. Pagal apibrėžimą:

$\pi_{j|i} = \pi_{ij} / \pi_{i+},\ j = 1, 2 \ldots c.$

Jungtinis X ir Y, vienmatis X ir Y bei sąlyginis Y skirstiniai, kai X ir Y įgyja reikšmes 1 ir 2, pavaizduoti 2.7 lentelėje.

Diskretieji ats. dydžiai X ir Y yra nepriklausomi, jei visos jungtinio X ir Y skirstinio tikimybės lygios atitinkamų X ir Y tikimybių sandaugai:

$\pi_{ij} = \pi_{i+} \times \pi_{+j},\ i = 1, 2 \ldots r,\ j = 1, 2 \ldots c.$

Kai X ir Y nepriklausomi, tuomet Y sąlyginis skirstinys yra vienodas visoms X reikšmėms:

$\pi_{j|i} = \pi_{ij} / \pi_{i+} = \pi_{i+} \times \pi_{+j} / \pi_{i+} = \pi_{+j}$

ir X sąlyginis skirstinys bus vienodas visoms Y reikšmėms.

2.7 lentelė. Jungtinis (X, Y), vienmatis X ir Y bei sąlyginis Y skirstiniai

	Stulpeliai		
Eilutės	1	2	Iš viso
1	π_{11} $(\pi_{1\|1})$	π_{12} $(\pi_{2\|1})$	π_{1+} 1.0
2	π_{21} $(\pi_{1\|1})$	π_{22} $(\pi_{2\|1})$	π_{2+} 1.0
Iš viso	π_{+1}	π_{+2}	1.0

Dvimatis tolydusis atsitiktinis dydis. Sakykime, (X, Y) yra tolydusis ats. v. Tuomet jis turi dvimatį tankį $p(x, y)$, taip susietą su (X, Y) skirstinio funkcija $F(x, y)$:

$$F(x, y) = P\{X \le x,\ Y \le y\} = \int_{-\infty}^{x}\int_{-\infty}^{y} p(u,v)dudv\,.$$

Dviejų kintamųjų funkcijos, tarp jų ir tankio funkcijos geometrinė prasmė – paviršius trimatėje erdvėje (analogiškai vieno kintamojo funkcija – kreivė plokštumoje).

Tolydieji ats. d. X ir Y yra nepriklausomi, jei jų bendras tankis $p(u, v)$ lygus atskirų ats. d. tankių $p_x(u)$ ir $p_y(v)$ sandaugai: $p(u, v) = p_x(u)\, p_y(v)$.

Tikimybė, kad ats. v. (X, Y) gali patekti į plokštumos sritį S, skaičiuojama naudojant tankio funkciją pagal formulę:

$$P\{(X, Y) \text{ yra srityje } S\} = \iint_S p(x, y)dxdy\,.$$

Kai sritis S yra apskritimas ar stačiakampis, tuomet tikimybė (X, Y) patekti į sritį S lygi cilindro ar prizmės su pagrindu S, apribotu paviršiumi $p(x, y)$, tūriui.

2.10. Dvimačio atsitiktinio vektoriaus skaitinės charakteristikos

2.4 skyriuje minėta, kad atsitiktinio dydžio sankaupos tašką apibūdina vidurkis, o reikšmių išsibarstymą aplink vidurkį – dispersija. Analizuojant dvimatį atsitiktinį vektorių aktualu nustatyti ne tik jo koordinačių vidurkį bei dispersiją, bet ir įvertinti priklausomybę tarp šio vektoriaus koordinačių. Tam naudojama kovariacija ir koreliacijos koeficientas.

Kovariacija tarp ats. d. X ir Y apibrėžiama lygybe:

$cov(X,Y) = E(X - EX)(Y - EY)$.

Pertvarkę šią formulę ir pasinaudoję vidurkio savybėmis, gauname patogesnę formulę kovariacijai skaičiuoti:

$cov(X,Y) = E(XY - YEX - XEY + (EX)(EY)) = E(XY) - (EX)(EY)$.

Kovariacija turi šias savybes:

1) $cov(Y,X) = cov(X,Y)$; $cov(X,X) = DX$;

2) jei ats. d. X ir Y yra nepriklausomi, tai $E(XY) = (EX)(EY)$ ir $cov(X,Y) = 0$;

3) kovariacija tarp X ir Y absoliučiu dydžiu neviršija ats. d. X ir Y standartinių nuokrypių sandaugos:

$$|cov(X,Y)| \le \sigma_x \sigma_y = \sqrt{DX \times DY}.$$

Atsitiktiniai dydžiai, kurių kovariacija lygi 0, vadinami nekoreliuotais. Jei X ir Y yra nepriklausomi, tai jie yra nekoreliuoti. Atvirkščias teiginys nėra teisingas. Gali būti, kad kovariacija tarp X ir Y lygi 0, tačiau X ir Y nėra nepriklausomi: bendra jų skirstinio funkcija nelygi X ir Y skirstinio funkcijų sandaugai.

Kovariacijos reikšmė priklauso ne tik nuo X ir Y priklausomybės laipsnio, bet ir nuo X bei Y mastelio. Dėl to problemiška palyginti kelių atsitiktinių dydžių porų priklausomybės stiprumą. Todėl ats. d. priklausomybės matui vertinti įvedamas bedimensis dydis – koreliacijos koeficientas.

Koreliacijos koeficientas tarp ats. d. X ir Y lygus:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{DX \times DY}} = \frac{E(XY) - (EX)(EY)}{\sqrt{DX \times DY}}.$$

Korelicijos koeficientas kinta tarp –1 ir 1: $-1 \le \rho(X,Y) \le 1$. Jei X ir Y yra nepriklausomi, tai $\rho(X,Y) = 0$ (atvirkščias teiginys ne visada teisingas). Jei $\rho(X,Y) = \pm 1$, tai ats. d. X ir Y yra tiesiškai priklausomi: $X = aY + b$; $a > 0$, jei $\rho(X,Y) = 1$ ir $a < 0$, jei $\rho(X,Y) = -1$. Atvirkščiai, jei $X = aY + b$, tai $\rho(X,Y) = \pm 1$.

Koreliacijos koeficientas apibūdina monotoninį tiesinį tikimybinį ryšį: jei $\rho(X,Y) > 0$, tai didėjant vienam ats. d., antras turi tendenciją didėti. Jei $\rho(X,Y) < 0$, tuomet vienam dydžiui didėjant, kitas ats. d. turi tendenciją mažėti.

Kovariacijų matrica V charakterizuoja bendrą ats. v. (X, Y) kitimą. Ji apibrėžiama:

$$V = \begin{pmatrix} DX & cov(X,Y) \\ cov(X,Y) & DY \end{pmatrix}.$$

Koreliacijų matrica apibrėžiama:

$$R = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix}.$$

Pažymėkime $\rho = \rho(X,Y)$. Tuomet $cov(X, Y) = \rho\sqrt{DX \times DY}$, ir kovariacijų matricą V galime išreikšti taip:

$$V = \begin{pmatrix} DX & \rho\sqrt{DX \times DY} \\ \rho\sqrt{DX \times DY} & DY \end{pmatrix}.$$

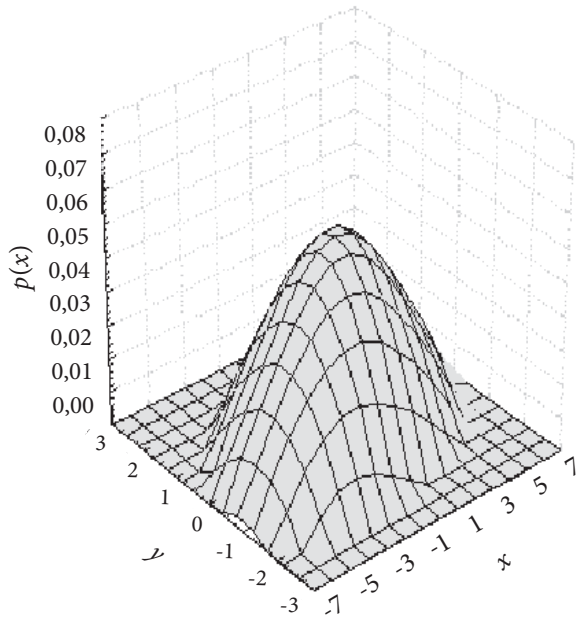
2.11. Dvimatis normalusis skirstinys

Sakoma, kad ats. vektoriaus (X, Y) skirstinys yra dvimatis normalusis, jei (X, Y) tankis lygus:

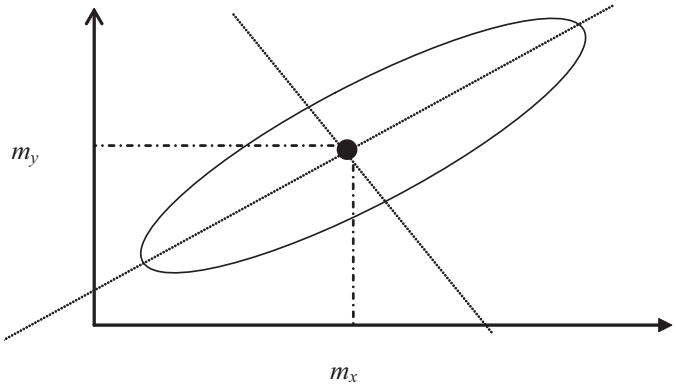
$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_x)^2}{\sigma_x^2} - \frac{2\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right)\right);$$

čia $m_x = EX$, $m_y = EY$ – ats. d. X ir Y vidurkiai, $\sigma_x^2 = DX$, $\sigma_y^2 = DY$ – ats. d. X ir Y dispersijos, $\rho = \rho(X, Y)$ – koreliacijos koeficientas tarp ats. d. X ir Y. Dvimačio normaliojo skirstinio tankis yra kupolo su centru (m_x, m_y) formos paviršius (2.15 pav.). Kupolo „ilgį“, „plotį“ ir „pasvirimą XOY plokštumoje“ nusako σ_x, σ_y bei ρ. Paviršiaus $z = p(x, y)$ pjūvis plokštuma $z = C$ yra elipsė (2.16 pav.):

$$\left(\frac{(x-m_x)^2}{\sigma_x^2} - \frac{2\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right) = C_1.$$



2.15 pav. Dvimačio normaliojo skirstinio tankis ($\sigma_x = 3$, $\sigma_y = 1$, $\rho = 0{,}7$)



2.16 pav. Dvimačio normaliojo skirstinio tankio pjūvis plokštuma

Elipsės didžioji ašis yra tiesėje $y = m_y + (\sigma_y/\sigma_x)\rho(x - m_x)$. Elipsės pasvirimo kampas bei suplotumas priklauso nuo ρ bei σ_x ir σ_y. Esant vienodoms ats. d. X ir Y dispersijoms, ilgesniosios ir trumpesniosios elipsės ašių ilgis atitinkamai proporcingas dydžiams $1/\sqrt{1-\rho}$ ir $1/\sqrt{1+\rho}$ – kuo ρ arčiau 1, tuo labiau suplota elipsė.

2 skyriaus literatūra

1. Agresti A. *Categorical Data Analysis*. 1996. New York: John Wiley & Sons, p. 558.
2. Aksomaitis A. *Tikimybių teorija ir statistika*. 2002. Kaunas: Technologija, 346 p.
3. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
4. Bagdonavičius V., Kruopis J. *Matematinė statistika*. I dalis. 2007. Vilnius, 359 p.
5. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
6. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
7. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.
8. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslo ir enciklopedijų leidykla, 416 p.
9. Kubilius J. *Tikimybių teorija ir matematinė statistika*. 1996. Vilnius: VU leidykla, 439 p.
10. Scheffe H. *The Analysis of Variance*. 1999. New York: John Wiley & Sons, p. 477.
11. Explanatory Data Analysis, *Engineering Statistics Handbook*. Prieiga per internetą: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3653.htm.

3 SKYRIUS

Statistiniai duomenų modeliai

Jau minėjome, kad medikų tiriamų rodiklių reikšmės skiriasi dėl įvairių priežasčių, kaip antai, dėl ligos, biologinių organizmo ypatybių, amžiaus, paveldimumo ir daugelio kitų priežasčių, kurias galima laikyti atsitiktinėmis. Kitimą populiacijoje galima modeliuoti nustatomus rodiklius (kintamuosius) laikant atsitiktiniais dydžiais, nusakomais tam tikru skirstiniu. Todėl analizuojant duomenis priimtina dalį kintamųjų laikyti atsitiktiniais su tam tikru skirstinio tipu (pvz., normaliuoju, Puasono).

Pateiksime kintamųjų bei priklausomybių tarp jų statistinius modelius, dažniausiai taikomus analizuojant medicinos ir biologijos duomenis.

3.1. Medicinoje naudojamų kintamųjų statistiniai modeliai

Atlikdami įvairius tyrimus, medikai nustato ligonių tiriamų rodiklių (SAS, DAS, KA klasės ir t. t.) skaitines reikšmes. Norint turimų duomenų pagrindu gauti argumentuotas išvadas apie tiriamų charakteristikų kintamumą, tarpusavio ryšį etc., būtina apibrėžti duomenų statistinį modelį.

Daugelį medikų bei epidemiologų rodiklių, nustatomų tam tikroje populiacijoje, galima laikyti atsitiktiniais dydžiais. Pavyzdžiui, IŠL sergančių 30–45 m. vyrų populiacijos arterinį sistolinį kraujospūdį dėl svyravimų populiacijoje galima laikyti atsitiktiniu dydžiu. Ligonio kokybinį kintamąjį, pavyzdžiui, KA klasę (I, II, III, IV), galima laikyti atsitiktiniu dydžiu, įgyjančiu reikšmes 1, 2, 3, 4. Kaip minėta 2 skyriuje, atsitiktinį dydį visiškai nusako jo skirstinys, todėl kintamojo statistinis modelis apibūdinamas tam tikru tikimybiniu skirstiniu. Kintamąjį modeliuojantis skirstinys pirmiausia parenkamas pagal kintamojo rūšį. Jei kintamasis yra kokybinis, statistinis jo modelis yra diskretusis skirstinys. Jei kintamasis yra kiekybinis, statistinis jo

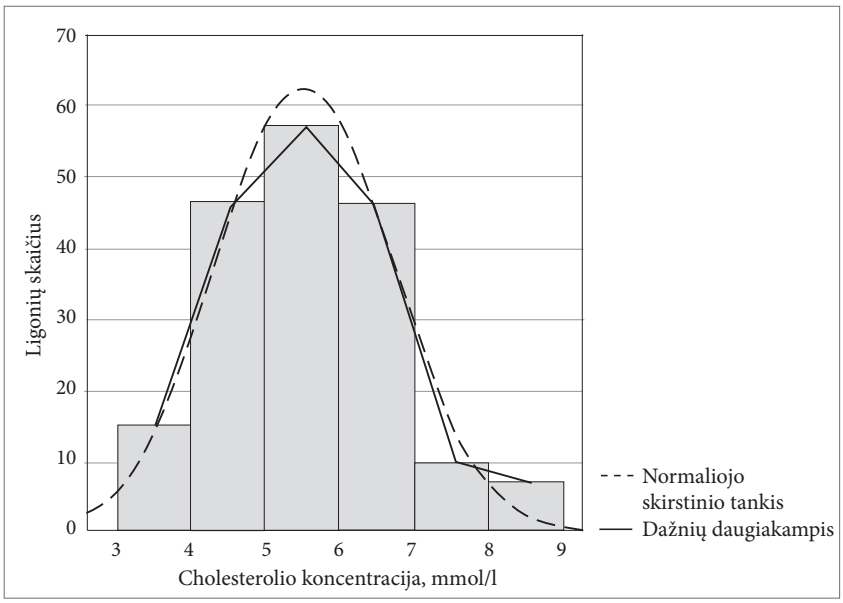
modelis dažniausiai yra tolydusis skirstinys. Tam tikrais atvejais diskrečiojo kiekybinio kintamojo modeliu naudojamas Puasono skirstinys.

Kokybinio kintamojo, įgyjančio k reikšmių, statistinis modelis – diskretusis atsitiktinis dydis, įgyjantis k reikšmių su nežinomomis tikimybėmis $\pi_1, \pi_2 \ldots \pi_k$. Šios tikimybės yra diskrečiojo skirstinio parametrai. Pavyzdžiui, KA klasės skirstinys – ats. d., įgyjantis 1, 2, 3, 4 reikšmes su atitinkamai π_1, π_2, π_3, π_4 tikimybėmis. Dvinarį kintamąjį galime laikyti atsitiktiniu dydžiu, įgyjančiu reikšmę 1 su tikimybe π ir 0 su tikimybe $1 - \pi$; čia π – skirstinio nežinomas parametras. Pavyzdžiui, tiriant rajono A sergamumą diabetu, individui nustatoma dvinario kintamojo reikšmė: 1, jei serga CD, ir 0 – jei neserga CD. Atsitiktinai parinktam individui sergamumo reikšmė – Bernulio ats. dydis su parametru π. Sergamumui kokia nors liga, sakykime, A, populiacijoje modeliuoti naudojamas binominis skirstinys. Jei daroma prielaida, kad sergamumo tikimybė vienoda visiems populiacijos individams ir liga A nėra reta, tai N dydžio populiacijoje sergančiųjų skaičiaus skirstinys yra binominis su parametrais (N, π); čia π – nežinoma tikimybė sirgti liga A.

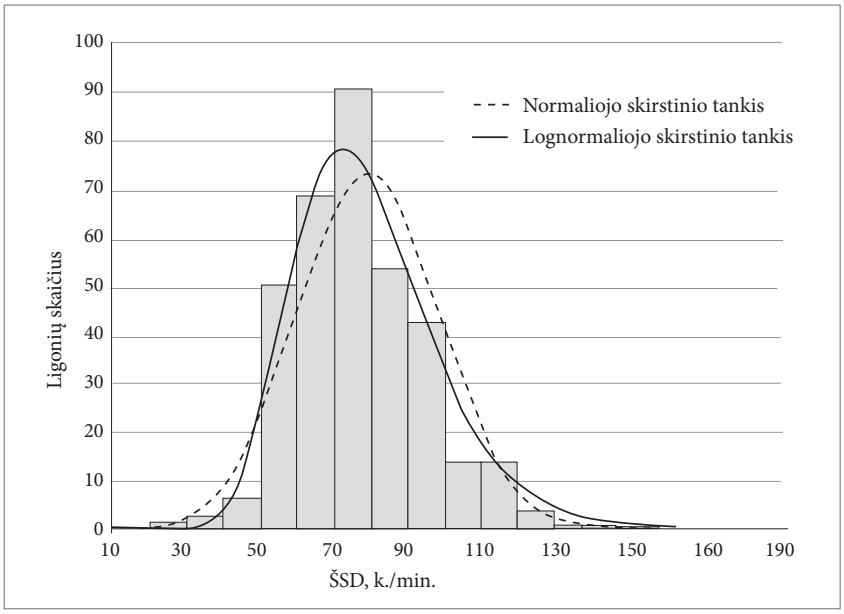
Puasono skirstiniu modeliuojami rodikliai, įgyjantys sveikas neneigiamas reikšmes, kaip antai, ištrauktų ar sugedusių vaiko dantų skaičius; vidutiniškai per dieną išgeriamų kavos ar arbatos puodelių skaičius etc., Puasono skirstiniu aprašomas retų įvykių pasiskirstymas populiacijoje. Pavyzdžiui, leukemija – retas vaikų susirgimas. Todėl fiksuoto dydžio vaikų populiacijoje sergamumui leukemija modeliuoti gerai tinka Puasono skirstinys.

Kai kurių kiekybinių kintamųjų, naudojamų medicinoje ar biologijoje, histogramos yra beveik simetriškos ir varpo formos. Sujungę histogramos stulpelių viršaus vidurio taškus, gausime kreivę (dažnių daugiakampį), panašią į normaliojo skirstinio tankį (3.1 pav.). Todėl kintamąjį, tiriamoje populiacijoje pasižymintį tokia savybe, su tam tikra išlyga galima laikyti atsitiktiniu dydžiu, turinčiu normalųjį skirstinį. Kitaip tariant, daroma prielaida, kad tiriamoje populiacijoje nagrinėjamas kintamasis yra atsitiktinis dydis, turintis normalųjį skirstinį su nežinomais parametrais.

Nemažai medikų ar biologų duomenų turi teigiamą asimetriją. Taip yra dėl to, kad daugelis kiekybinių kintamųjų, pavyzdžiui, arterinis kraujospūdis, cholesterolio koncentracija, mikrobiologinių tyrimų rodikliai įgyja tik teigiamas reikšmes, be to, pasitaiko gana didelių matavimo reikšmių. Tačiau teigiamą asimetriją turinčio kintamojo logaritmuotų reikšmių skirstinį dažnai galima laikyti normaliuoju. Tuomet daroma prielaida, kad šio kintamojo skirstinys yra lognormalusis (3.2 pav.).



3.1 pav. Ligonių, sergančių pirmine arterine hipertenzija, cholesterolio histograma, dažnių daugiakampis ir atitinkamo normaliojo skirstinio tankis



3.2 pav. Ligonių, sergančių Q bangos MI, ŠSD histograma ir populiacijos skirstinių tankiai

Didelėje populiacijoje atsitiktinai parinktas kintamojo X reikšmes – atsitiktinę imtį (x_1, x_2 ... x_n) galima laikyti nepriklausomais ats. dydžiais. Imtis (x_1, x_2 ... x_n) vadinama paprastąja, jei atsitiktinių dydžių x_1, x_2 ... x_n skirstiniai yra vienodi. Paprastoji imtis dažna daugelyje medikų tyrimų. Pavyzdžiui, daroma prielaida, kad IŠL sergančių 30–45 m. vyrų SAS yra atsitiktinis dydis, turintis normalųjį skirstinį su parametrais m ir σ^2. Šio kintamojo imties elementai – atsitiktinai atrinktų populiacijos vyrų SAS yra nepriklausomi atsitiktiniai dydžiai, turintys normalųjį skirstinį su vidurkiu m ir dispersija σ^2. Tačiau toks duomenų modelis – tiriamoje populiacijoje kintamasis yra atsitiktinis dydis su tam tikru skirstiniu – tinkamas ne visuose tyrimuose. Analizuodami ryšį tarp kelių individą apibūdinančių kintamųjų, susiduriame su modeliu: faktorius → atsakas (dozė → atsakas); t. y. vienus kintamuosius galime laikyti faktoriais (nepriklausomais kintamaisiais), kitus – atsaku į šių faktorių veikimą (priklausomais kintamaisiais). Pavyzdžiui, amžius laikomas faktoriumi, kūno masės indeksas – atsaku į amžiaus poveikį. Sakykime, populiacijoje analizuojamas sergamumas IŠL. Jis priklauso ir nuo ligonio amžiaus, lyties, gyvensenos veiksnių bei kitų faktorių. Todėl tvirtinti, kad „tikimybė sirgti IŠL yra pastovi“ galima tik kalbant apie labai siaurą populiaciją, pavyzdžiui, tam tikro amžiaus nerūkančius vyrus. Tiriamąją populiaciją išskaidyti į subpopuliacijas ir iš kiekvienos daryti atranką yra sudėtinga. Tokių duomenų analizei naudojami regresiniai modeliai (10–12 skyrius). Juose daroma prielaida, kad faktorius yra determinuotas (neatsitiktinis) dydis, o atsakas – atsitiktinis dydis, turintis skirstinį su parametrais, priklausančiais nuo faktoriaus reikšmių. Pavyzdžiui, vietoj modelio: „imties narių x_1, x_2 ... x_n skirstiniai yra normalieji ar lognormalieji su vienodais parametrais“ galime naudoti sudėtingesnį: „x_i skirstinys priklauso tam tikrai skirstinių šeimai (normaliojo, lognormaliojo), o skirstinio parametrai priklauso nuo tam tikro neatsitiktinio ligonį charakterizuojančio kintamojo (vienmačio ar daugiamačio) Z reikšmės, nustatytos i-tajam ligoniui z_i“. Šie modeliai vadinami regresiniais (jie nagrinėjami 10 skyriuje). Sudėtingesnis dvinario kintamojo modelis – tikimybė π priklauso ir nuo individą charakterizuojančio kintamojo z_i reikšmių: $\pi = \pi(z_i)$. Atskiras $\pi(z_i)$ modelio atvejis – logistinė regresija – nagrinėjamas 11 skyriuje.

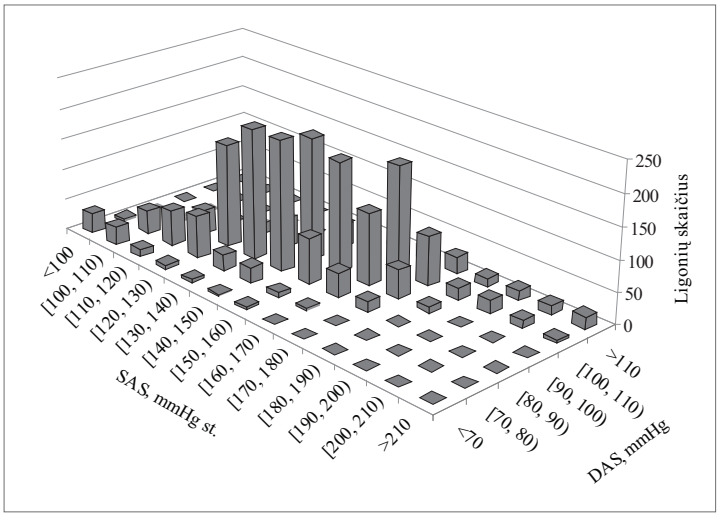
Medikams aktualu ne tik pavienio kintamojo dėsningumai, bet ir kelių kintamųjų tarpusavio priklausomybė. Todėl naudojami kelių kintamųjų jungtinio skirstinio modeliai. Pavyzdžiui, SAS ir DAS jungtinį skirstinį galima laikyti dvimačiu normaliuoju su nežinomais parametrais (3.3 pav.).

Kai kuriems statistinės analizės metodams reikalingos prielaidos apie k kintamųjų bendrą skirstinį. Atskirų kintamųjų normalumas leidžia visų k kin-

tamųjų daugiamatį skirstinį laikyti normaliuoju su nežinomais parametrais (plačiau apie daugiamatį normalųjį skirstinį – 15 skyriuje).

Apskritai duomenų skirstinių modeliai skirstomi į:

- parametrinius;
- pusiau parametrinius;
- neparametrinius.



3.3 pav. Ligonių, sergančių ūmiais koronariniais sindromais, (SAS, DAS) dvimatė histograma

Parametrinis modelis (dažniausiai naudojamas): kintamojo skirstinys priklauso tam tikrai parametrinių skirstinių šeimai $P = P(\theta_1, \theta_2 \ldots \theta_k)$, kuri apibrėžiama žinomos analizinės išraiškos skirstiniu, priklausančiu nuo k parametrų $(\theta_1, \theta_2 \ldots \theta_k)$. Žinodami šių parametrų reikšmes, visiškai žinome ir skirstinį. Parametrinių skirstinių šeimų pavyzdžiai: normaliųjų skirstinių šeima $P = P(m, \sigma)$ susideda iš visų galimų normaliųjų skirstinių; Puasono skirstinių šeima $P(\lambda)$ susideda iš visų Puasono skirstinių; Bernulio skirstinių šeima $P(\pi)$. Normaliųjų skirstinių šeimoje turime du nežinomus parametrus: θ_1 yra vidurkis m, o θ_2 – σ. Žinodami m ir σ, visiškai žinome normaliojo skirstinio tankį ar skirstinio funkciją. Puasono ir Bernulio skirstiniuose turime vieną nežinomą parametrą: θ_1 yra λ, arba π. Parametrinis modelis plačiau nagrinėjamas 3.2 skyriuje.

Pusiau parametrinis modelis: kintamojo skirstinys priklauso tam tikrai skirstinių klasei S, kuri yra platesnė už P. S klasės skirstiniai priklauso nuo nežinomų parametrų $(\theta_1, \theta_2 \ldots \theta_k)$, bet jų funkcinė išraiška (formulė) nėra

žinoma. *S* gali būti simetriškų, unimodaliųjų skirstinių, turinčių vidurkį *m* (parametras), klasė. Analizuojant gautus duomenis, ne visada reikia tiksliai apibūdinti kintamojo skirstinį. Kartais, ypač kai turima daug matavimų, užtenka prielaidos, kad šis skirstinys yra unimodalusis (ats. d. tankis ar tikimybė turi vieną maksimumo tašką).

Sakykime, norime įvertinti tam tikros populiacijos ligonių BC koncentracijos vidurkį arba palyginti dviejų didelių ligonių grupių ($n > 100$) BC koncentraciją. Turint dideles imtis ($n > 100$), nebūtina tiksliai apibūdinti BC koncentracijos skirstinio (pavyzdžiui, normalusis skirstinys), užtenka prielaidos, kad BC koncentracijos skirstiniai yra unimodalieji su vidurkiais m_1 ir m_2. Ši prielaida leidžia statistikos metodais (didžiųjų skaičių dėsniu, centrine ribine teorema) įvertinti ligonių grupių BC koncentracijos vidurkius ir juos palyginti (pusiau parametrinis modelis).

Turint nedidelę imtį, sunku patikrinti jos normalumą. Todėl dažnai apsiribojama prielaida, jog imties (ir populiacijos) skirstinys yra simetriškas. Lyginant dviejų nedidelių ligonių grupių kiekybinio kintamojo reikšmes, daroma prielaida, kad vienos ligonių grupės kintamojo reikšmės yra ats. dydžiai $X_1, X_2 \ldots X_n$, kitos ligonių grupės – ats. d. $\theta + X_{n+1}, \theta + X_{n+2} \ldots \theta + X_{n+m}$; čia $X_1, X_2 \ldots X_n, X_{n+1} \ldots X_{n+m}$ – nepriklausomi vienodai pasiskirstę ats. dydžiai, θ – poslinkio parametras. Kai yra toks kintamojo modelis, tvirtinti, kad abiejų ligonių grupių vidurkiai lygūs, yra tas pats, kaip tvirtinti, jog parametras θ lygus nuliui (pusiau parametrinis modelis).

Neparametrinis modelis: kintamojo skirstinys priklauso tam tikrai skirstinių klasei *F* (pavyzdžiui, tolydžiųjų, diskrečiųjų), kuri yra platesnė už *S*.

Norint nustatyti, ar klinikinio rodiklio A skirstiniai sergančių ir sveikų populiacijoje yra identiški, užtenka prielaidos, jog rodiklio A skirstinys yra tolydusis (neparametrinis modelis). Norint patikrinti hipotezę, kad tarp kiekybinių kintamųjų yra tiesinis ryšys, daroma prielaida, jog šių kintamųjų skirstiniai yra tolydieji.

Matematiniu požiūriu griežtesnis parametrinių, pusiau parametrinių ir neparametrinių modelių apibrėžimas pateiktas V. Bagdonavičiaus ir J. Kruopio vadovėlyje [2].

3.2. Parametrinis imties modelis, parametrų vertinimas*

Konkreti tiriamo rodiklio (ar kelių tiriamų rodiklių) imtis ($x_1, x_2 \ldots x_n$) yra atsitiktinės imties ($X_1, X_2 \ldots X_k$) realizacija; čia $X_1 \ldots X_k$ – nepriklausomi atsitiktiniai dydžiai. Sakoma, kad imties ($X_1, X_2 \ldots X_k$) modelis yra parametrinis, jei atsitiktinio vektoriaus $\boldsymbol{X} = (X_1, X_2 \ldots X_n)$ skirstinio analizinė išraiška žinoma

ir skirstinys priklauso nuo nežinomų parametrų $\theta_1, \theta_2 \ldots \theta_k$. Šiems skirstinio parametrams dažnai suteikiama tam tikra medicininė ar biologinė interpretacija; jie įvertinami remiantis turimomis imties reikšmėmis ($x_1, x_2 \ldots x_n$).

Nežinomi ($X_1, X_2 \ldots X_k$) skirstinio parametrai dažniausiai vertinami didžiausio tikėtinumo metodu. Šio metodo esmė – taip parinkti nežinomų parametrų įverčius $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_k$, kad tyrimo metu gauta imtis $x_1, x_2 \ldots x_n$ būtų labiausiai tikėtina.

Didžiausio tikėtinumo metodui pateikti apibrėšime imties tikėtinumo funkciją. Pažymėkime $p(\boldsymbol{x}, \boldsymbol{\theta})$ (arba $p_X(\boldsymbol{x}, \boldsymbol{\theta})$), $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_k)$ – vektoriaus $\boldsymbol{X}$ tankio funkcija, jei $\boldsymbol{X}$ koordinatės – tolydieji ats. d., arba tikimybė, kad $\boldsymbol{X}$ įgis reikšmę $\boldsymbol{x}$, jei $\boldsymbol{X}$ koordinatės – diskretieji ats. dydžiai. Funkcija $L(\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{X}, \boldsymbol{\theta}) = p(X_1, X_2 \ldots X_n, \boldsymbol{\theta})$ vadinama imties $(X_1, X_2 \ldots X_n)$ tikėtinumo funkcija. Kadangi imtis – ats. vektorius $\boldsymbol{X}$ su nepriklausomomis koordinatėmis, tai

$$L(\boldsymbol{X}, \boldsymbol{\theta}) = p_1(X_1, \boldsymbol{\theta}) p_2(X_2, \boldsymbol{\theta}) \ldots p_n(X_n, \boldsymbol{\theta});$$

čia $p_i(x, \boldsymbol{\theta})$ – ats. d. X_i tankis (jei X_i yra tolydusis) ar tikimybė (jei X_i yra diskretusis). Atvejis, kai imtį generuoja priklausomi ats. dydžiai, nagrinėjamas 10.11 skyriuje.

Sakoma, kad imties tikėtinumo funkcija priklauso eksponentinių skirstinių šeimai, jei ją galima išreikšti taip:

$$L(\boldsymbol{X}, \theta) = C(\theta)\exp\left(\sum_{j=1}^{k} Q_j(\theta_j) T_j(\boldsymbol{X})\right) h(\boldsymbol{X}); \qquad (3.1)$$

čia $C(\boldsymbol{\theta})$, $Q_j(\boldsymbol{\theta})$ – funkcijos, priklausančios tik nuo $\boldsymbol{\theta}$, funkcijos $T_1(\boldsymbol{x}) \ldots T_k(\boldsymbol{x})$, $h(\boldsymbol{x})$ nuo parametro $\boldsymbol{\theta}$ nepriklauso. Daugelis skirstinių, naudojamų medicinos duomenims modeliuoti, kaip antai, normalusis, binominis, Bernulio – priklauso eksponentinių skirstinių šeimai. Plačiau apie eksponentinių skirstinių šeimą pateikta [2].

3.1 pavyzdys. Imtis, generuota normalaus ats. d., t. y. vektoriaus $\boldsymbol{X} = (X_1, X_2 \ldots X_n)$ koordinatės, yra nepriklausomi, normalųjį skirstinį su vidurkiu m ir dispersija σ^2 turintys atsitiktiniai dydžiai. Šios imties tikėtinumo funkcija $L(\boldsymbol{X}, m, \sigma)$ (žr. 2.5 skyrių) lygi:

$$L(\boldsymbol{X}, m, \sigma) = (2\pi\sigma)^{-n/2} \exp\left\{-\frac{nm^2}{2\sigma^2}\right\} \exp\left\{\frac{m}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} X_i^2\right\};$$

t. y. (3.1) formulėje turime $\boldsymbol{\theta} = (m, \sigma)$, $k = 2$,

$C(\boldsymbol{\theta}) = C(m, \sigma) = (2\pi\sigma)^{-n/2} \exp\left\{-\frac{nm^2}{2\sigma^2}\right\}$, $h(\boldsymbol{x}) = 1$, $Q_1(\boldsymbol{\theta}) = m/\sigma^2$, $Q_2(\boldsymbol{\theta}) = -1/\sigma^2$,

$T_1(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$, $T_2(x) = \sum_{i=1}^{n} x_i^2$.

Analizuojant medikų sukauptus duomenis, dažniausiai susiduriama su paprastąja imtimi, kai (X_1, X_2 .., X_n) yra ats. vektorius su nepriklausomomis ir tą patį skirstinį turinčiomis koordinatėmis. Šiuo atveju imties tikėtinumo funkcija lygi:

$L(\boldsymbol{X}, \boldsymbol{\theta}) = p(X_1, \boldsymbol{\theta})\ p(X_2, \boldsymbol{\theta}) \ldots p(X_n, \boldsymbol{\theta})$;

čia $p(x, \boldsymbol{\theta})$ – ats. d. X_i tankis ar tikimybė. Skirstinių klasė P apibūdinama kaip vienmačio ats. d. skirstinių klasė.

Jei daroma prielaida, kad X_i skirstinys priklauso parametrinei skirstinių klasei $P(\boldsymbol{b})$, o X_i skirstinio parametras b_i priklauso nuo ligonio kito neatsitiktinio kintamojo reikšmės $\boldsymbol{z_i}$: $\boldsymbol{b_i} = \boldsymbol{b}(\boldsymbol{\theta}, \boldsymbol{z_i})$ (regresinis modelis), tai

$L(\boldsymbol{x}, \boldsymbol{\theta}) = p(X_1, \boldsymbol{b}(\boldsymbol{\theta}, \boldsymbol{z_1}))\ p(X_2, \boldsymbol{b}(\boldsymbol{\theta}, \boldsymbol{z_2})) \ldots p(X_n, \boldsymbol{b}(\boldsymbol{\theta}, \boldsymbol{z_n}))$.

Didžiausio tikėtinumo metodas. Labiausiai tikėtina ta diskretaus ats. dydžio reikšmė, kuriai įgyti tikimybė yra didžiausia. Kai skirstinys yra tolydusis, labiausiai „tikėtina“ ats. d. tankio maksimumo reikšmė. Todėl nežinomų parametrų įverčiais parenkamos tokios imties (X_1, X_2 ... X_n) funkcijos, su kuriomis tikėtinumo funkcija didžiausia. Ieškant tikėtinumo funkcijos maksimumo, ji logaritmuojama, surandamos logaritmuotos tikėtinumo funkcijos $l(\boldsymbol{X}, \boldsymbol{\theta}) = ln(L(\boldsymbol{X}, \boldsymbol{\theta}))$ išvestinės parametrų θ_1, θ_2 ... θ_k atžvilgiu ir prilyginamos nuliui. Gauta lygčių sistema

$$\frac{\partial l(...)}{\partial \theta_i} = 0,\ i = 1, 2 \ldots k$$

išsprendžiama θ_1, θ_2 ... θ_k atžvilgiu. Šios sistemos sprendiniai (tikslūs ar nustatyti artutiniais metodais) $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_k$ laikomi ieškomais nežinomų parametrų įverčiais. Pagal apibrėžimą $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_k$ yra atsitiktiniai dydžiai – imties (X_1, X_2 ... X_n) funkcijos (statistikos).

Didžiausio tikėtinumo metodu gauti parametrų įverčiai gana geri: kuo imtis didesnė, tuo labiau tikėtina, kad parametro įvertis labai mažai skirsis nuo tikrosios parametro reikšmės. Normaliojo, Puasono ar Bernulio skirstinių parametrų vertinimas didžiausio tikėtinumo metodu smulkiau išdėstytas [1, 2, 3, 7] vadovėliuose.

Dažniausiai naudojamų skirstinių parametrų įverčiai pateikti 3.1 lentelėje. Parametrinio modelio įverčiai gauti didžiausio tikėtinumo metodu.

Analizuojant kelių kintamųjų (faktoriaus ir atsako į jį) tarpusavio ryšį, naudojami sudėtingesni modeliai, kaip antai, dispersinė ar regresinė analizė (10–12 skyriai). Juose nežinomi parametrai vertinami artutiniais metodais, todėl ne visada galima pateikti parametrų įverčių formules, analogiškas pateiktoms 3.1 lentelėje.

3.1 lentelė. Parametrų įverčiai įvairiems kintamojo modeliams

Kintamasis	Skirstinys	Parametrai	Parametrų įverčiai
Kiekybinis	Normalusis	(m, σ^2)	$\hat{m} = \bar{x}, \sigma^2 = s^2$
Kiekybinis	Puasono	λ	$\hat{\lambda} = \bar{x}$
Kokybinis, įgyjantis k reikšmių	Diskretusis	$\pi_1, \pi_2 \ldots \pi_{k-1}$	$\hat{\pi}_i = p_i = n_i / n$, n_i – dažniai
Dvinaris	$P\{X = 1\} = \pi$, $P\{X = 0\} = 1 - \pi$	π	$\hat{\pi} = p = k / n$, k – vienetų skaičius imtyje; p – vienetų proporcija imtyje
Kiekybinis	–	Vidurkis m, dispersija σ^2	$\hat{m} = \bar{x}, \sigma^2 = s^2$

3.3. Parametrų vertinimas Bajeso ir pakartotinės atrankos metodu

Nežinomiems imties skirstinio parametrams vertinti taip pat naudojami Bajeso bei pakartotinės atrankos metodai.

Bajeso (*Bayes*) metodas*. Naudojant šį metodą, daroma prielaida, kad parametro $\boldsymbol{\theta}$ reikšmės nėra fiksuotos. Jos yra atsitiktinės su tam tikru žinomu skirstiniu. Taigi taikant didžiausio tikėtinumo metodą, nežinomų parametrų reikšmės nustatomos remiantis tik imties reikšmėmis, o Bajeso metodas papildomai naudoja ir informaciją apie $\boldsymbol{\theta}$.

Norint įvertinti $\boldsymbol{\theta}$, reikia nustatyti jo aposteriorinį skirstinį – skirstinį $p(\boldsymbol{\theta}|\boldsymbol{x})$ konkrečiai imčiai $\boldsymbol{x} = (x_1, x_2 \ldots x_n)$. Pagal Bajeso formulę

$p(\boldsymbol{\theta}|\boldsymbol{x}) = p(\boldsymbol{x}|\,\boldsymbol{\theta})\, p(\boldsymbol{\theta})/f(\boldsymbol{x})$;

čia $p(\boldsymbol{\theta})$ – žinomas skirstinys, $p(\boldsymbol{x}|\boldsymbol{\theta})$ – imties tikėtinumo funkcija (konkrečiai imčiai) esant fiksuotai $\boldsymbol{\theta}$ reikšmei, o $f(\boldsymbol{x}) = \int p\,(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ – imties $\boldsymbol{x}$ besąlyginis skirstinys (tankis arba tikimybė). Jei įmanoma, $p(\boldsymbol{\theta}|\boldsymbol{x})$ išreiškiamas analiziškai, naudojant $p(\boldsymbol{x}|\boldsymbol{\theta})$ ir $p(\boldsymbol{\theta})$ išraiškas. Kai $p(\boldsymbol{\theta}|\boldsymbol{x})$ analiziškai išreikšti neįmanoma, šis skirstinys modeliuojamas naudojant kompiuteriu generuotus reikiamus atsitiktinius dydžius.

Pakartotinės atrankos (*resampling*) metodai. Šių metodų esmė – naudojant turimas imties reikšmes, generuojamos naujos imtys, po to didžiausio tikėtinumo metodu įvertinami nežinomi parametrai kiekvienai sugeneruotai imčiai. Šių parametrų reikšmių vidurkis ir yra nežinomo parametro įvertis. Priklausomai nuo naujų imčių generavimo taisyklių naudojamas plėtros (*bootstrap*) bei atmestos reikšmės (*jackknife*) metodai.

Plėtros metodas. Iš n imties reikšmių sudaromos imtys, kurių kiekvienoje yra m ($m \le n$) narių; iš viso – m^n imčių. Imčių generavimo procedūra vadinama visa pakartotinė atranka (*complete resampling*), kai iš n dydžio imties reikšmių generuojamos visos galimos n dydžio imtys ($m = n$). Šį metodą iliustruosime pavyzdžiu.

3.2 pavyzdys ([4]). Turime 3 narių imtį, generuotą normaliojo ats. dydžio: 1; 6; 9. Šios imties vidurkis lygus $\overline{x} = 5{,}33$, standartinis nuokrypis $s = 4{,}04$ (s vertintas (1.1) formule). Plėtros metodu įvertinsime vidurkį ir standartinį nuokrypį. Iš 1; 6; 9 reikšmių galima sudaryti $3^3 = 27$ skirtingas imtis. 3.2 lentelėje pateiktos visos šios imtys, kiekvienos imties vidurkis ir standartinis nuokrypis. Šių 27 dydžių vidutinės reikšmės – 5,33 ir 2,95 – yra populiacijos skirstinio vidurkio ir standartinio nuokrypio įverčiai, gauti plėtros metodu. Matome, kad plėtros metodu gautas standartinio nuokrypio įvertis mažesnis, nei apskaičiuotas pagal (1.1) formulę.

Atmestos reikšmės metodas. Iš n dydžio imties vienas narys pašalinamas ir likusiais (n – 1) nariais vertinami nežinomi parametrai. Paeiliui šalinant vis skirtingą narį, ši procedūra kartojama n kartų. Iš gautų n įverčių reikšmių skaičiuojami vidurkiai. Jie ir laikomi nežinomų parametrų įverčiais.

3.2 lentelė. Imties 1; 6; 9 visa pakartotinė atranka bei vidurkio ir standartinio nuokrypio įverčiai, gauti plėtros metodu

Imtis	$\overline{x}$	s	Imtis	$\overline{x}$	s	Imtis	$\overline{x}$	s
1; 1; 1	1	0	6; 1; 1	2,67	2,89	9; 1; 1	3,67	4,62
1; 1; 6	2,67	2,89	6; 1; 6	4,33	2,89	9; 1; 6	5,33	4,04
1; 1; 9	3,67	4,62	6; 1; 9	5,33	4,04	9; 1; 9	6,33	6,62
1; 6; 1	2,67	2,89	6; 6; 1	4,33	2,89	9; 6; 1	5,33	4,04
1; 6; 6	4,33	2,89	6; 6; 6	6	0	9; 6; 6	7	1,73
1; 6; 9	5,33	4,04	6; 6; 9	7	1,73	9; 6; 9	8	1,73
1; 9; 1	3,67	4,62	6; 9; 1	5,33	4,04	9; 9; 1	6,33	6,62
1; 9; 6	5,33	4,04	6; 9; 6	7	1,73	9; 9; 6	8	1,73
1; 9; 9	6,33	6,62	6; 9; 9	8	1,73	9; 9; 9	9	0
Vidurkio ir standarinio nuokrypio įverčiai: Didžiausio tikėtinumo metodu: $\overline{x} = 5{,}33$, $s = 4{,}04$ Plėtros metodu: $\overline{x} = 5{,}33$, $s = 2{,}95$								

3.4. Parametrų įverčių kitimo charakteristika

Nežinomų parametrų, kaip antai, vidurkio, dispersijos, tikimybės įverčiai, gauti didžiausio tikėtinumo metodu, yra imties – atsitiktinių dydžių funkcijos, t. y. jie patys yra atsitiktiniai dydžiai. Todėl įvertinus nežinomus parametrus, būtina įvertinti jų patikimumą – sklaidą aplink vidurkį. Pažy-

mėsime, kad visi 3.1 lentelėje pateikti parametrų įverčiai yra nepriklausomų ats. dydžių x_1, x_2 ... x_n sumų funkcijų reikšmės. Kai imtis ganėtinai didelė, minėtų įverčių skirstinius galima laikyti normaliaisiais (remiantis centrine ribine teorema). Normaliojo ats. d. kitimą (sklaidą) charakterizuoja dispersija, arba standartinis nuokrypis. Todėl pateikiant įvertintus kintamojo (populiacijos) skirstinio parametrus, pateikiami ir jų standartinių nuokrypių įverčiai, dar vadinami standartinėmis paklaidomis (*standard error*).

Sakykime, kiekybinis kintamasis X turi unimodalų skirstinį su vidurkiu m ir dispersija σ^2. Vidurkio m įvertis yra imties vidurkis $\bar{x}$. Ats. d. $\bar{x}$ dispersija yra σ^2/n, standartinis nuokrypis – $\sigma/\sqrt{n}$, o standartinis nuokrypio įvertis – $s/\sqrt{n}$ žymimas SE($\bar{x}$), arba s_x. s_x vadinamas standartine vidurkio paklaida (*standard error of mean*, arba *SE*). 3.3 lentelėje pateiktos parametrų įverčių dispersijos, standartiniai nuokrypiai ir standartinės paklaidos. Tyrimų metu nustatytas kiekybinio kintamojo vidurkis pateikiamas kartu su standartine paklaida, pavyzdžiui, $\bar{x} \pm s_x$. 3.4 lentelėje pateikti kintamųjų vidurkių įverčiai su standartinėmis paklaidomis.

3.3 lentelė. Parametrų įverčių dispersija, standartinis nuokrypis ir standartinė paklaida

Skirstinys	Parametrai	Įvertis	Dispersija	St. nuokrypis	St. paklaida
Normalusis	Vidurkis m	$\bar{x}$	σ^2/n	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
	Dispersija σ^2	s^2	$2\sigma^4/n$	$\sqrt{2}\sigma^2/\sqrt{n}$	$\sqrt{2}s^2/\sqrt{n}$
Puasono	λ	$\bar{x}$	λ/n	$\sqrt{\lambda/n}$	$\sqrt{\bar{x}/n}$
Bernulio	Tikimybė π	$p = k/n$	$\pi(1-\pi)/n$	$\sqrt{\pi(1-\pi)/n}$	$\sqrt{p(1-p)/n}$

3.4 lentelė. Ligonių, sergančių ūmiais koronariniais sindromais, echoskopinių rodiklių vidurkiai ir standartinės paklaidos

Rodiklis	Vidurkis	St. paklaida
KSGDD	47,73	0,2
KSMI	128,95	1,15
KPR	54,46	0,23
DPR	47,69	0,25
IF	43,0	0,36

3.3 pavyzdys. Vidurkio, standartinio nuokrypio ir standartinės paklaidos skaičiavimą iliustruosime pavyzdžiu. 3.5 lentelėje pateikti 26 jaunų sveikų suaugusių asmenų arterinio kraujo spaudimo matavimo duomenys (SAS ir

DAS, mmHg). Kadangi šie asmenys rinkti atsitiktinai, tai, naudodami šios imties reikšmes, įvertinsime jaunų sveikų suaugusių asmenų populiacijos SAS vidurkį. Populiacijos SAS vidurkio įverčio galimą paklaidą įvertinsime vidurkio standartine paklaida SE. Naudodami 3.5 lentelės duomenis, turime:

- tyrimų skaičių, arba $n = 26$;
- vidurkį, arba $\bar{x} = 113{,}1$ mmHg;
- standartinis nuokrypis, arba $s = 10{,}3$ mmHg;
- standartinę paklaidą, arba $s_x = s/\sqrt{n} = 10{,}3/\sqrt{26} = 10{,}3/5{,}1 = 2{,}02$ (mmHg).

3.5 lentelė. 26 jaunų sveikų suaugusių asmenų SAS ir DAS reikšmės

Nr.	SAS (mmHg st.)	DAS (mmHg st.)	Lytis
1	108	62	M
2	134	74	V
3	100	64	M
4	108	68	M
5	112	72	V
6	112	64	M
7	112	68	M
8	122	70	V
9	116	70	V
10	116	70	V
11	120	72	V
12	108	70	M
13	108	70	M
14	96	64	M
15	114	74	V
16	108	68	V
17	128	86	V
18	114	68	V
19	112	64	V
20	124	70	M
21	90	60	M
22	102	64	M
23	106	70	V
24	124	74	V
25	130	72	V
26	116	70	M

Analizuojant sudėtingesnius modelius, pavyzdžiui, regresinius, vertinamas ne tik vidurkio, bet ir viso daugiamačio parametro $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_k)$ įverčio $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_k)$ – atsitiktinio vektoriaus – kitimas. Jis vertinamas kovariacijų matrica $cov(\hat{\theta})$. Kai $\hat{\theta}$ yra didžiausio tikėtinumo įvertis, ši kovariacijų matrica išreiškiama tokia formule: $cov(\hat{\theta}) = (-I(\boldsymbol{\theta}))^{-1}$, čia $I(\boldsymbol{\theta}) = (\nu_{ij})$ – Fišerio informacijos matrica, sudaryta iš elementų

$$\nu_{ij} = \frac{\partial^2 l(\ldots)}{\partial \theta_i \, \partial \theta_j}.$$

Kovariacijų matrica $cov(\hat{\theta})$ yra parametrų $\theta_1, \theta_2 \ldots \theta_k$ funkcija. Kovariacijų matricos $cov(\hat{\theta})$ įvertis $\boldsymbol{cov}(\hat{\theta})$ skaičiuojamas į $cov(\hat{\theta})$ išraišką vietoj $\boldsymbol{\theta}$ įrašius didžiausio tikėtinumo įvertį $\hat{\theta}$: $\boldsymbol{cov}(\hat{\theta}) = (-I(\boldsymbol{\theta}) | \, \boldsymbol{\theta} = \hat{\theta})^{-1}$.

Kai $\boldsymbol{\theta}$ vertinamas Bajeso metodu, $\boldsymbol{\theta}$ kintamumas nustatomas modeliuojant $p(\boldsymbol{\theta}|\boldsymbol{x})$ skirstinį arba naudojantis tikslia $p(\boldsymbol{\theta}|\boldsymbol{x})$ išraiška; čia $\boldsymbol{x} = (x_1, x_2 \ldots x_k)$.

3 skyriaus literatūra

1. Aksomaitis A. *Tikimybių teorija ir statistika*. 2002. Kaunas: Technologija, 346 p.
2. Bagdonavičius V., Kruopis J. *Matematinė statistika*. I dalis. 2007. Vilnius, 359 p.
3. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
4. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.
5. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
6. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, biostatistics and preventive Medicine*. 1996. London: Saunders, p. 297.
7. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslo ir enciklopedijų leidykla, 416 p.
8. Kubilius J. *Tikimybių teorija ir matematinė statistika*. 1996. Vilnius: VU leidykla, 439 p.
9. Shiryajev A. N. *Probability*. 1995. Second edition. Springer, 612 p.
10. Explanatory Data Analysis. *Engineering Statistics Handbook*. Prieiga per internetą: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3653.htm.
11. *Didžiausio tikėtinumo ir mažiausių kvadratų metodų, skirtų nežinomų parametrų įverčiams rasti, analizė ir palyginimas. Pateikti pavyzdžiai.* Prieiga per internetą: http://quantrm2.psy.ohio-state.edu/injae/mle-pub.pdf.

4 SKYRIUS

Pasikliautinieji intervalai ir jų naudojimas išvadoms gauti

4.1. Parametrų pasikliautinieji intervalai

Sakykime, X – tolydusis ats. dydis, turintis vidurkį m ir dispersiją σ^2, o $x_1, x_2 \ldots x_n$ – šio ats. d. paprastoji imtis (x_i – normalusis ats. d.). Vidurkio m įvertis yra imties vidurkis $\overline{x}$. Jis taip pat yra atsitiktinis: skirtingų konkrečių imčių $\overline{x}$ reikšmės, taip pat ir m įverčiai, skirsis. Pavyzdžiui, daroma prielaida, kad 30–45 m. vyrų SAS skirstinys yra normalusis. Trys rezidentai vertino SAS skirstinio vidurkį naudodami penkių atsitiktinai atrinktų 30–45 m. vyrų SAS reikšmes. Tyrimo metu gautos reikšmės: I rezidento – 130; 135; 120; 125; 120; II rezidento – 135; 115; 120; 125; 140; III rezidento – 120; 120; 140; 135; 125. Šių imčių vidurkiai, taip pat ir parametro m įverčio reikšmės, atitinkamai lygūs 126, 131 ir 128. Todėl aktualu įvertinti vidurkio įverčio patikimumą – nustatyti, kiek jis skiriasi nuo tikrosios parametro reikšmės.

Nežinomo parametro įverčio patikimumui vertinti pasitelkiama pasikliovimo lygmens sąvoka. Pasikliovimo lygmuo (*confidence level*) P – tikimybė, jog nežinomas populiacijos (skirstinio) parametras θ yra intervale $[\theta_{ap}(\boldsymbol{x}), \theta_{virš}(\boldsymbol{x})]$:

$$P = P\{\theta_{ap}(\boldsymbol{x}) \le \theta \le \theta_{virš}(\boldsymbol{x})\}; \tag{4.1}$$

čia $\boldsymbol{x} = (x_1, x_2 \ldots x_n)$ – atsitiktinė imtis – nepriklausomi ats. dydžiai.

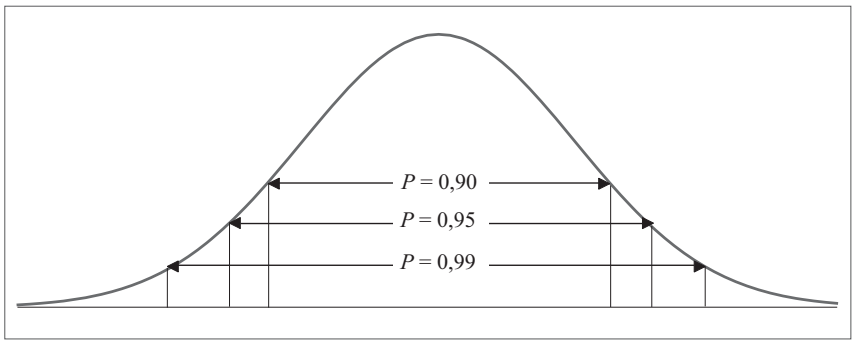
Intervalas $[\theta_{ap}(\boldsymbol{x}), \theta_{virš}(\boldsymbol{x})]$ vadinamas parametro θ pasikliautinuoju intervalu (*confidence interval*); atsitiktiniai dydžiai $\theta_{ap}(\boldsymbol{x})$ ir $\theta_{virš}(\boldsymbol{x})$ vadinami

pasikliautinojo intervalo apatine ir viršutine ribomis. Pasikliautinasis intervalas sutrumpintai žymimas PI. P reikšmės parenkamos artimos vienetui: $P = 0{,}9$; 0,95; 0,99; kartais P pateikiamas procentais. Nuo P parinkimo priklauso pasikliovimo intervalų ilgis – kuo P artimesnis vienetui, tuo pasikliautinasis intervalas yra platesnis (4.1 pav.). Pasikliautinasis intervalas $[\theta_{ap}(\boldsymbol{x}), \theta_{virš}(\boldsymbol{x})]$ nusako nežinomo parametro θ įvertinimo patikimumą – tai intervalas, kuriame su artima vienetui tikimybe yra tikroji parametro reikšmė. $\theta_{ap}(\boldsymbol{x})$ ir $\theta_{virš}(\boldsymbol{x})$ reikšmės priklauso nuo kintamojo modelio ($\boldsymbol{x}$ skirstinio), parametro θ įverčio $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ ir nuo pasikliovimo lygmens P.

Dažniausiai naudojamų parametrų pasikliautinieji intervalai sudaromi taip: nustatoma imties funkcija (statistika) $T(\theta, \boldsymbol{x})$, į kurią dažniausiai įeina ir parametro įvertis $\hat{\theta}(\boldsymbol{x})$. $T(\theta, \boldsymbol{x})$ skirstinys yra žinomas (pvz., standartinis normalusis, Stjudento, χ^2, ...). Parenkame skaičius u_1 ir u_2, nepriklausančius nuo imties reikšmių, kad

$$P = P\{u_1 \le T(\theta, \boldsymbol{x}) \le u_2\}. \tag{4.2}$$

Paprastai u_1 ir u_2 yra $T(\theta, \boldsymbol{x})$ skirstinio atitinkamo lygmens kvantiliai: $u_1 = t_{(1-P)/2}$, $u_2 = t_{(1+P)/2}$. Įrašę šias reikšmes į (4.1) formulę ir atitinkamai ją pertvarkę, gauname θ pasikliautinąjį intervalą.



4.1 pav. Pasikliautinieji intervalai su skirtingais pasikliovimo lygmenimis P

Konkrečiai imčiai nustatytos pasikliautinojo intervalo ribos θ_{ap} ir $\theta_{virš}$ yra skaičiai, o parametro pasikliautinasis intervalas ir (4.1) formulė interpretuojama taip: įvykio $\theta_{ap} \le \theta \le \theta_{virš}$ statistinė tikimybė – pasikliovimas – lygi P. Tai reiškia, kad, pakartotinai generuojant imtis, maždaug P procentų atvejų nežinomas parametras θ bus apskaičiuotame pasikliautinajame intervale.

4.2. Normaliojo skirstinio vidurkio pasikliautinasis intervalas

Pateiksime keletą pasikliautinųjų intervalų konstravimo pavyzdžių.

Normaliojo skirstinio vidurkio pasikliautinasis intervalas, kai dispersija žinoma

Sakykime, x_1, x_2 ... x_n – ats. dydžio X, turinčio normalųjį skirstinį su nežinomu vidurkiu m ir žinoma dispersija σ^2, imtis. Konstruodami vidurkio m (čia $\theta = m$) pasikliautinąjį intervalą, pirmiausia sudarykime funkciją $T(\theta, \boldsymbol{x})$. Kadangi x_i yra normalusis ats. d. su parametrais (m, σ^2), $i = 1 \ldots n$, tai $(x_1 + x_2 + \ldots + x_n)$ ir $(x_1 + x_2 + \ldots + x_n - nm)/(\sqrt{n}\sigma)$ yra atitinkamai normalieji ats. dydžiai su parametrais $(nm, n\sigma^2)$ ir (0, 1), ir statistika $T(\theta, \boldsymbol{x}) = \sqrt{n}(\bar{x} - m)/\sigma \sim N(0, 1)$. Todėl

$$P = P\{-z_{(1+P)/2} \le (x_1 + x_2 + \ldots + x_n - nm)/(\sqrt{n}\sigma) \le z_{(1+P)/2}\}$$

arba

$$P\{\bar{x} - z_{(1+P)/2}\,\sigma/\sqrt{n} \le m \le \bar{x} + z_{(1+P)/2}\,\sigma/\sqrt{n}\} = P.$$

Remdamiesi pastaraja formule, nustatome nežinomo vidurkio (dispersija žinoma) pasikliautinąjį intervalą $[m_{ap}; m_{virš}]$, kai patikimumas yra P; čia $m_{ap} = \bar{x} - z_{(1+P)/2}\,\sigma/\sqrt{n}$, $m_{virš} = \bar{x} + z_{(1+P)/2}\,\sigma/\sqrt{n}$.

Normaliojo skirstinio vidurkio PI, kai dispersija nežinoma

Sakykime, x_1, x_2 ... x_n – ats. dydžio X, turinčio normalųjį skirstinį su nežinomais parametrais (m, σ^2), imtis; čia $\theta = m$. Ats. d. $(x_1 + x_2 + \ldots + x_n - nm)$ yra normalusis su parametrais $(0, n\sigma^2)$, o $(x_1 + x_2 + \ldots + x_n - nm)/(\sqrt{n}s)$ pagal apibrėžimą yra ats. d., turintis Stjudento skirstinį su $n - 1$ laisvės laipsnių; taigi statistika $T(\theta, \boldsymbol{x})$ yra tokia: $T(\theta, \boldsymbol{x}) = \sqrt{n}(\bar{x} - m)/s \sim t(n - 1)$. Todėl

$$P = P\{-t_{(1+P)/2}(n-1) \le (x_1 + x_2 + \ldots + x_n - nm)/(s\sqrt{n}) \le t_{(1+P)/2}(n-1)\}$$

arba

$$P\{\bar{x} - t_{(1+P)/2}(n-1)\,s/\sqrt{n} \le m \le \bar{x} + t_{(1+P)/2}(n-1)\,s/\sqrt{n}\} = P.$$

Remdamiesi pastaraja formule, nusakome nežinomo vidurkio pasikliautinojo intervalo apatinę ir viršutinę ribas esant patikimumui P:

$$m_{ap} = \bar{x} - t_{(1+P)/2}(n-1)\,s/\sqrt{n},\ m_{virš} = \bar{x} + t_{(1+P)/2}(n-1)\,s/\sqrt{n}; \qquad (4.2)$$

čia $t_\alpha(n-1)$ – Stjudento skirstinio su $n - 1$ laisvės laipsniu α lygmens kvantilis, s – standartinis nuokrypis. Kai imtis didelė ($n > 100$), skaičiuojant vidurkio pasikliautinąjį intervalą, (4.2) formulėje vietoje Stjudento skirstinio

kvantilio galima naudoti $z_{(1+P)/2}$, nes $\sqrt{n}(\overline{x}\text{-}m)$/s skirstinys, augant *n,* artėja į standartinį normalųjį (didžiųjų skaičių dėsnis, centrinė ribinė teorema).

Kaip minėjome, dydis $s/\sqrt{n}$, naudojamas vidurkio pasikliautinojo intervalo formulėje (4.2), yra imties vidurkio standartinė paklaida; taigi žinant imties vidurkį ir standartinę paklaidą, galima apskaičiuoti vidurkio pasikliautinąjį intervalą bet kuriuo patikimumu.

4.1 pavyzdys. 3.5 lentelėje pateiktos 26 jaunų sveikų suaugusių asmenų SAS ir DAS reikšmės (mmHg). Naudodamiesi 3.3 skyriuje nustatytu SAS vidurkiu bei jo standartine paklaida, lygia 2,02, apskaičiuosime SAS vidurkio 95 % pasikliautinąjį intervalą. Remiantis (4.2) formule, vidurkio 95 % pasikliautinasis intervalas yra $\overline{x} \pm t_{0,975}(15)s_x$. 2 lentelėje randame $t_{0,975}(15) = 2,13$. Todėl SAS vidurkio 95 % PI yra:

$113,1 \pm 2,13 \times 2,02 = 113,1 \pm 4,303$ arba $m_{ap} = 108,8$ ir $m_{virš} = 117,4$.

SAS vidurkio 99 % PI skaičiuojamas analogiškai:

$113,1 \pm 2,95 \times 2,02 = 113,1 \pm 5,959$ arba tarp $m_{ap} = 107,14$ ir $m_{virš} = 119,06$, nes $t_{0,995}(15) = 2,95$.

Vidurkio PI skaičiavimas plėtros metodu. Skaičiuojant normaliojo skirstinio vidurkio PI pagal (4.2) formulę, vietoj standartinio nuokrypio σ įverčio *s*, apskaičiuoto pagal (1.1) formulę, naudojamas standartinio nuokrypio įvertis, gautas plėtros metodu (3.3 skyrius).

Vidurkio PI galima įvertinti ir tiesiogiai plėtros metodu. Generuojamas reikiamas imčių skaičius m^n; čia n – imties dydis, $m \leq n$. Apskaičiuojamas kiekvienos imties vidurkis ir visi gauti vidurkiai išdėstomi didėjimo tvarka. Atmetus $n(1 - P)/2$ didžiausių bei tiek pat mažiausių vidurkio reikšmių, likusi didžiausia ir mažiausia vidurkio reikšmės ir bus pasikliautinojo intervalo su patikimumu P ribos.

4.2 pavyzdys. Turime imtį 1; 6; 9, generuotą normaliojo ats. d. Apskaičiuosime vidurkio pasikliautinąjį intervalą su pasikliovimo lygmeniu $P = 0,93$.

- Vidurkio pasikliautinasis intervalas, apskaičiuotas pagal (4.2) formulę.

Turime $n = 3$; $\overline{x} = 5,33$, $s = 4,04$; čia s skaičiuotas naudojant (1.1) formulę; $SE = s/\sqrt{3} = 2,33$; $(1 + P)/2 = 0,965$; $t_{0,965}(2) = 3,76$. PI ribos lygios:

$m_{ap} = 5,33 - 3,76 \times 2,33 = -3,43$; $m_{virš} = 5,33 + 3,76 \times 2,33 = 14,09$.

- Vidurkio PI, apskaičiuotas pagal (4.2) formulę, kai s skaičiuotas plėtros metodu.

3.3 pavyzdyje pateiktas šios imties vidurkio ir standartinio nuokrypio skaičiavimas plėtros metodu naudojant visą pakartotinę atranką ($3^3 = 27$ imtys).

Minėtų parametrų įverčiai, gauti plėtros metodu, yra šie: $\overline{x} = 5{,}33$, $s = 2{,}95$. Tuomet standartinė paklaida lygi $2{,}95/\sqrt{3} = 1{,}7$, o pasikliautinojo intervalo ribos: $m_{ap} = 5{,}33 - 3{,}76 \times 1{,}7 = -1{,}07$; $m_{virš} = 5{,}33 + 3{,}76 \times 1{,}7 = 11{,}74$.

- Vidurkio PI, apskaičiuotas naudojant visą pakartotinę atranką.

Visų 27 imčių vidurkio reikšmės pateiktos 3.2 lentelėje. Atmetus 2 kraštines reikšmes iš 27 (1 ir 9), gautas toks vidurkio pasikliautinasis intervalas: $m_{ap} = 2{,}67$; $m_{virš} = 8$.

Iš pateiktų pavyzdžių matome, kad plačiausias vidurkio pasikliautinasis intervalas gautas naudojant (4.2) formulę su *s*, apskaičiuotu pagal (1.1) formulę, o siauriausias – visos pakartotinės atrankos metodu.

4.3. Dvinario kintamojo tikimybės pasikliautinasis intervalas

Sakykime, $x_1, x_2 \ldots x_n$ – dvinario kintamojo imtis; x_i – ats. dydžiai, įgyjantys dvi reikšmes – 1 ir 0 su tikimybėmis $P\{X = 1\} = \pi$, $P\{X = 0\} = 1 - \pi$; čia π – nežinomas parametras. Tikimybės π įvertis yra proporcija $p = k/n$; čia k – vienetų skaičius imtyje. Kadangi k yra binominis ats. d. su parametrais (n, π), tai tikimybės π tikslaus pasikliautinojo intervalo su patikimumu P apatinė ir viršutinė ribos p_{ap} ir $p_{virš}$ apibrėžiamos lygybėmis:

$$(1 - P)/2 = \sum_{j=0}^{k} C_n^j p_{ap}^j (1 - p_{ap})^{n-j}, \quad (1 - P)/2 = \sum_{j=k}^{n} C_n^j p_{virs}^j (1 - p_{virs})^{n-j}.$$

Naudojantis šiomis formulėmis, tikimybės pasikliautinąjį intervalą nustatyti yra sudėtinga. Tikslios tikimybės PI ribos skaičiuojamos statistiniais paketais, pavyzdžiui, EPIINFO, bei pateikiamos lentelėmis (žr. [6]). Kai n ganėtinai didelis, o p nėra labai maža, t. y. $np > 5$ ir $n(1-p) > 5$, tikimybės π pasikliautinojo intervalo ribos nustatomos remiantis centrine ribine teorema: statistikos

$$\frac{\sqrt{n}(p - \pi)}{\sqrt{\pi(1 - \pi)}} \quad \text{arba} \quad \frac{(k - n\pi)}{\sqrt{n\pi(1 - \pi)}} \tag{4.3}$$

asimptotinis skirstinys yra standartinis normalusis. Kadangi k yra diskretusis, o normalusis ats. d. – tolydusis, į (4.3) išraišką įrašoma tolydumo pataisa; nustatyta, kad statistikos

$$Z = \frac{|k - n\pi| - 1/2}{\sqrt{n\pi(1 - \pi)}} \tag{4.4}$$

skirstinys artimesnis standartiniam normaliajam skirstiniui nei (4.3). Kai n labai didelis, tolydumo pataisa nebūtina. Remdamiesi (4.3–4.4) atsitiktinių dydžių normalumu, pateiksime tris variantus tikimybės apytikslėms

PI riboms skaičiuoti. Pažymėkime z – standartinio normaliojo skirstinio $(1 + P)/2$ lygio kvantilis, P – nustatytas patikimumas.

I. PI ribos nustatomos remiantis ats. d. Z (4.4) normalumu:

$$\frac{\sqrt{n}(p - p_{ap} - 1/2)}{\sqrt{p_{ap}(1 - p_{ap})}} = z; \quad \frac{\sqrt{n}(p - p_{virš} + 1/2)}{\sqrt{p_{virš}(1 - p_{virš})}} = -z$$

arba $p_{ap} = ((2k + z^2 - 1) - z\sqrt{z^2 - (2 + 1/n) + 4k((n-k)+1)/n})/(2(n + z^2))$,

$p_{virš} = ((2k + z^2 + 1) + z\sqrt{z^2 + (2 + 1/n) + 4k((n-k)-1)/n})/(2(n + z^2))$.

II. Kai n didelis, pataisa diskretumui nebūtina, todėl π pasikliautinąjį intervalą galima apibrėžti nelygybe:

$$| p - \pi | \le z\sqrt{\pi(1-\pi)/n}. \qquad (4.5)$$

Pakėlę abi nelygybės (4.5) puses kvadratu ir išsprendę lygtį $(p - \pi)^2 = z^2\pi(1 - \pi)/n$ π atžvilgiu, gauname apytikslį tikimybės π pasikliautinąjį intervalą $[p_{ap}, p_{virš}]$:

$p_{ap} = (2np + z^2 - z\sqrt{z^2 + 4np(1-p)})/(2(n + z^2))$,

$p_{virš} = (2np + z^2 + z\sqrt{z^2 + 4np(1-p)})/(2(n + z^2))$.

III. π pasikliautinasis intervalas gaunamas (4.5) formulės dešinėje pusėje pakeitus π jos įverčiu p: $p_{ap} = p - z\sqrt{p(1-p)/n}$, $p_{ap} = p + z\sqrt{p(1-p)/n}$.

Pateiksime tikimybės įverčio ir pasikliautinojo intervalo skaičiavimo pavyzdį.

4.2 pavyzdys. Ištyrus 100 atsitiktinai atrinktų individų, 65 nustatėme požymį A, taigi π įvertis lygus: $p = 65/100 = 0{,}65$. Apskaičiuosime tikimybės 95 % pasikliautinąjį intervalą. Tikslus šios tikimybės PI yra [0,548; 0,743]; įvertintas I metodu – [0,548; 0,741], II metodu – [0,552; 0,736], III metodu – $0{,}65 \pm 1{,}96(0{,}65\times0{,}35/100)^{1/2} = 0{,}65 \pm 1{,}96\times0{,}0477$, arba [0,557; 0,743]. Matome, visais metodais apskaičiuoti tikimybės PI beveik nesiskiria.

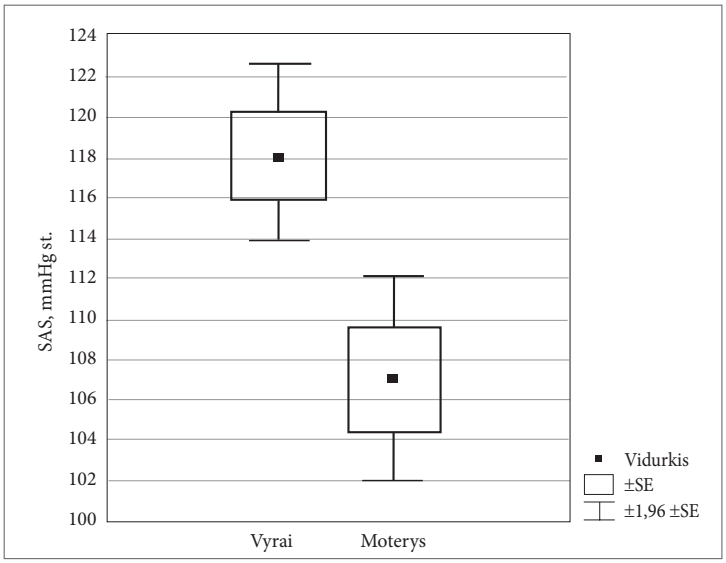
Palyginsime nedidelės imties tikimybės pasikliautinuosius intervalus. Sakykime, n =20, p = 0,25. Tikslus 95 % PI: [0,087; 0,491]; I metodu – [0,096; 0,494], II metodu – [0,112; 0,469], III metodu – [0,06; 0,440]. Matome, kad III metodu nustatytos atitinkamos PI ribos yra aiškiai mažesnės, nei nustatytos I ar II metodu. Šiaip III metodą PI nustatyti rekomenduotina, kai np ir $n(1 - p)$ viršija 10. II metodas naudotinas, kai dydžiai np_{ap} ir $n(1 - p_{virš})$ didesni už 5.

4.4. Pasikliautinųjų intervalų grafinis pateikimas

Pasikliautinuosius intervalus, apskaičiuotus naudojant konkrečios imties reikšmes, galima grafiškai pateikti stačiakampe diagrama. Standartinė stačiakampė diagrama brėžiama taip: stačiakampio viduryje kvadratėliu pažymimas vidurkis; stačiakampis braižomas nuo *vidurkis – standartinė paklaida* ($\bar{x} - s/\sqrt{n}$) iki *vidurkis + standartinė paklaida* ($\bar{x} + s/\sqrt{n}$). Nuo stačiakampio apačios brėžiamas apatinis „ūsas“ tęsiasi iki ($\bar{x} - 1{,}96s/\sqrt{n}$), o viršutinis „ūsas“ prasideda nuo stačiakampio viršaus ir brėžiamas iki reikšmės ($\bar{x} + 1{,}96s/\sqrt{n}$). 1,96 yra normaliojo skirstinio 0,975 lygio kvantilis, taigi stačiakampės diagramos „ūsai“ nurodo vidurkio pasikliautinąjį intervalą, skaičiuotą standartiniam patikimumui $P = 0{,}95$ ir gana didelei imčiai (4.2 pav.). Statistiniai paketai STATISTICA ir SPSS suteikia galimybę stačiakampe diagrama pavaizduoti ir vidurkio pasikliautinąjį intervalą, nustatomą pagal formulę (4.2).

Pateikus kelių imčių (ar kelių parametrų) vidurkio pasikliautinąjį intervalą stačiakampe diagrama, nesunku palyginti populiacijos skirstinio vidurkius. Jei pasikliautinieji intervalai nesusikerta, galima daryti išvadą, kad populiacijų skirstinio vidurkiai yra skirtingi (plačiau apie vidurkių palyginimą – 5 skyriuje).

4.2 pav. stačiakampe diagrama pateikti vyrų ir moterų SAS vidurkio pasikliautinieji intervalai (3.5 lentelės duomenys).



4.2 pav. Vyrų ir moterų SAS vidurkio pasikliautinieji intervalai

Iš 4.2 pav. daroma išvada: „jaunų sveikų suaugusių vyrų SAS vidurkis (118,3 mmHg st.) reikšmingai didesnis nei tokio pat amžiaus moterų (107,0 mmHg st.)“. Tai suprantama, kad jaunų sveikų vyrų populiacijos SAS skirstinio vidurkis yra didesnis už jaunų sveikų moterų populiacijos SAS skirstinio vidurkį.

Dažniausiai naudojami tolydžiojo skirstinio parametrų pasikliautinieji intervalai pateikti 4.1 lentelėje. Tikimybės pasikliautinojo intervalo formulės pateiktos 4.3 skyriuje.

4.1 lentelė. Tolydžiojo skirstinio parametrų pasikliautinieji intervalai

Ats. d. X, generuojančio imtį, skirstinys	Parametras	Pasikliautinasis intervalas
$X \sim N(m, \sigma^2)$, m ir σ – nežinomi	m	$m_{ap} = \bar{x} - t_{(1+P)/2}(n-1)\, s/\sqrt{n}$, $m_{virš} = \bar{x} - t_{(1+P)/2}(n-1)\, s/\sqrt{n}$,
$X \sim N(m, \sigma^2)$, m ir σ – nežinomi	σ	$\sigma_{ap} = (n-1)s^2/\chi^2_{(1-P)/2}(n-1)$, $\sigma_{virš} = (n-1)s^2/\chi^2_{(1+P)/2}(n-1)$,
Skirstinys unimodalus, imtis didelė ($n > 30$)	m	$m_{ap} = \bar{x} - z_{(1+P)/2}\, s/\sqrt{n}$, $m_{virš} = \bar{x} + z_{(1+P)/2}\, s/\sqrt{n}$

4 skyriaus literatūra

1. Aksomaitis A. *Tikimybių teorija ir statistika*. 2002. Kaunas: Technologija, 346 p.
2. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed. Blackwell Science, p. 817.
3. Bagdonavičius V., Kruopis J. *Matematinė statistika*. I dalis. 2007. Vilnius, 359 p.
4. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
5. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.
6. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslas, 416 p.
7. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.

5 SKYRIUS

Hipotezių tikrinimas

5.1. Statistinės hipotezės

Hipoteze vadinamas bet kuris nepatvirtintas teiginys. Teiginį apie stebimo ats. dydžio ar kelių ats. dydžių skirstinį vadinsime **statistine hipoteze**.

Kasdieniuose tyrimuose medikai susiduria su aibe hipotezių, kurias reikia patvirtinti arba atmesti. Pateiksime keletą medikų klinicistų hipotezių:

(I) Sergančių I tipo CD moterų iki menopauzės kaulinio audinio tankis mažesnis nei sveikų moterų.

(II) Ramiprilio vartojimas mažina kardiovaskulinio įvykio 5 metų laikotarpiu tikimybę.

(III) Ligonių sergamumas X liga neturi įtakos širdies kairiojo skilvelio galiniam diastoliniam dydžiui (KSGDD).

(IV) Ligonių, sergančių X liga, bendrojo cholesterolio koncentracija kraujyje yra padidėjusi.

(V) Sergančiuosius arterine hipertenzija gydant X preparatu, ramybės ir fizinio krūvio metu sumažėjo SAS.

(VI) Ligonių, turinčių širdies nepakankamumą, tikimybė išgyventi vienerius metus po MI mažesnė nei neturinčių širdies nepakankamumo.

Laikant ligonį charakterizuojančius kintamuosius atsitiktiniais dydžiais, visas šias hipotezes galima paversti statistinėmis hipotezėmis.

Jei duomenų skirstinio modelis yra parametrinis, t. y. kintamojo skirstinys yra žinomo pavidalo (pavyzdžiui, normalusis), bet su nežinomais parametrais, tai (I–VI) hipotezės performuluojamos į hipotezes apie skirstinio parametrus. Pavyzdžiui, individo BC koncentracija priklauso nuo įvairių indi-

vido savybių – amžiaus, mitybos, biologinių savybių ir t. t. Todėl priimtina prielaida, kad atsitiktinai atrinktų populiacijos individų BC koncentracija kraujyje – atsitiktinis dydis, turintis tam tikro pavidalo skirstinį su nežinomu vidurkiu m; čia vidurkis m yra BC skirstinio parametras. Šiuo modeliu aiškinant cholesterolio svyravimus, (IV) hipotezė–tvirtinimas „susirgimas X padidina BC koncentraciją“ reiškia, kad „ligonių, sergančių X liga, BC skirstinio vidurkis yra didesnis nei sveikų ligonių BC skirstinio vidurkis“. Matematiškai tai išreiškiama taip: $m > m_0$; čia m_0 – BC koncentracijos norma. Jei (IV) pasitvirtino, sakoma: „sergančių X liga BC koncentracijos vidurkis reikšmingai didesnis už normą“. Aprašant statistines išvadas, vietoj „populiacijos vidurkis yra didesnis...“, teigiama „vidurkis reikšmingai didesnis...“. Frazė „vidurkis didesnis“ suprantama, kad didesnis yra apskaičiuotas konkrečios imties vidurkis, o išvada apie populiacijos vidurkį nedaroma.

Performuluosime (I) hipotezę į statistinę. Darome prielaidą, kad sergančiųjų I tipo CD moterų kaulinio audinio tankis turi normalųjį skirstinį su nežinomu vidurkiu m_1, sveikų moterų kaulinio audinio tankis taip pat turi normalųjį skirstinį su nežinomu vidurkiu m_2. Tuomet (I) hipotezę galima suformuluoti taip: $m_1 < m_2$, arba „sergančiųjų I tipo CD moterų kaulinio audinio tankio vidurkis yra reikšmingai mažesnis už sveikų moterų kaulinio audinio tankio vidurkį“.

Nagrinėdami (III) hipotezę, darykime prielaidą: ligonių, sergančių ligomis X ir Y, KSGDD turi to paties tipo skirstinį su vidurkiais m_x ir m_y. Tuomet (III) hipotezė formuluojama taip: $m_x = m_y$, arba „ligonių, sergančių ligomis X ir Y, KSGDD vidurkiai reikšmingai nesiskyrė“. (V) hipotezė performuluojama analogiškai.

(II) hipotezė: kardiovaskulinį įvykį (CV) nurodantis kintamasis Y ($Y = 1$, jei CV įvyko penkerių metų laikotarpiu, $Y = 0$, jei per penkerius metus CV neįvyko) yra ats. dydis su Bernulio skirstiniu. Sakykime, kad vartojusių ramiprilio CV atsiradimo tikimybė $P\{Y = 1\}$ yra π_1, vartojusių placebo – π_2. Tuomet vietoj (II) hipotezės galime tikrinti hipotezę $\pi_1 < \pi_2$, arba „vartojusių ramiprilio CV penkerių metų laikotarpiu tikimybė yra mažesnė nei ramiprilio nevartojusiųjų“. Atitinkamai ir (VI) hipotezė performuluojama į hipotezę apie skirstinio parametrus (6 skyrius).

Taip pat nagrinėjamos hipotezės, nesusijusios su skirstinio parametrais. Tai hipotezės apie imties skirstinį (dažniausiai skirstinio normalumą), dviejų ar daugiau imčių skirstinių vienodumą bei dviejų ar daugiau kintamųjų nepriklausomumą. Pateiksime tokio tipo hipotezių pavyzdžių.

(VII) Tiriamoje ligonių populiacijoje visos KA klasės vienodai dažnos.
(VIII) Tiriamos populiacijos SAS skirstinys yra normalusis.

(IX) A ir B populiacijų ligonių BC koncentracijos skirstiniai yra vienodi.

(VII) Hipotezę galime sukonkretinti: „KA klasės skirstinys yra diskretusis, įgyjantis 4 reikšmes su vienodomis tikimybėmis“:

x	1	2	3	4
$P\{X=x\}$	0,25	0,25	0,25	0,25

(5.1)

(VIII) hipotezėje nenurodyti normaliojo skirstinio parametrai, jie nežinomi. Todėl ši hipotezė – „SAS skirstinys normalusis“ – suprantama, kad SAS skirstinys yra normalusis su nežinomais parametrais m ir σ^2.

5.2. Nulinė hipotezė ir alternatyva

Tikrinant statistines hipotezes, apibrėžiama nulinė hipotezė bei jai alternatyvi hipotezė – alternatyva. **H_0 – nulinė hipotezė** – hipotezė apie tiriamo kintamojo skirstinį, kurią galima patvirtinti arba atmesti; **H_i – alternatyva.** Ji nurodo H_0 atmetimo kryptį.

Jei kintamojo skirstinys priklauso parametrinių skirstinių šeimai, tuomet H_0 virsta hipoteze apie nežinomo parametro (ar kelių nežinomų parametrų) reikšmę. Analogiškai formuluojama ir alternatyva. Pateiksime keletą nulinės ir alternatyvios hipotezių variantų apie parametro m (pvz., SAS populiacijos vidurkio) reikšmę:

$$H_0: m \le m_0;\ H_1: m > m_0, \tag{5.2}$$

$$H_0: m = m_0;\ H_1: m > m_0, \tag{5.3}$$

$$H_0: m = m_0;\ H_2: m < m_0, \tag{5.4}$$

$$H_0: m = m_0;\ H_3: m \ne m_0. \tag{5.5}$$

(5.2) hipotezę ir alternatyvą formuluotume taip: H_0 – SAS vidurkis (suprantama, populiacijos skirstinio) neviršija nustatytos reikšmės m_0; H_1 – SAS vidurkis viršija reikšmę m_0. (5.3) hipotezę ir alternatyvą formuluotume taip: H_0 – SAS vidurkis lygus m_0; H_1 – SAS vidurkis viršija m_0. (5.4) ir (5.5) nulinę hipotezę formuluotume analogiškai: H_0 – SAS vidurkis lygus m_0; H_2 – tirtų ligonių SAS vidurkis mažesnis nei m_0; H_3 – SAS vidurkis nelygus m_0.

Alternatyva H_1 vadinama dešiniapuse, H_2 – kairiapuse. Abi šios alternatyvos vadinamos vienpusėmis. Alternatyva H_3 vadinama dvipuse.

Nulinė hipotezė, arba alternatyva, vadinama **paprasta**, jeigu nežinomo parametro įgyjamų reikšmių aibė yra vienas taškas. Priešingu atveju nulinė hipotezė, ar alternatyva, vadinama **sudėtinga**. 5.3–5.5 pavyzdžiai pateikia

paprastą nulinę hipotezę ir sudėtingą alternatyvą. 5.2 pavyzdžio ir nulinė hipotezė, ir alternatyva yra sudėtingos.

Tikrinant (I–VI) hipotezes, nulinė hipotezė ir alternatyva formuluojamos taip:

(I): H_0: tiek sergančiųjų I tipo CD, tiek sveikų moterų kaulinio audinio tankio vidurkis nesiskiria ($m_1 = m_2$);
H_1: sergančiųjų I tipo CD moterų kaulinio audinio tankio vidurkis reikšmingai mažesnis už sveikų moterų kaulinio audinio tankio vidurkį ($m_1 < m_2$);

(II): H_0: ramiprilio vartojimas neturi įtakos CV penkerių metų laikotarpiu tikimybei ($\pi_1 = \pi_2$);
H_1: ramiprilio vartojimas sumažina CV penkerių metų laikotarpiu tikimybę ($\pi_1 < \pi_2$);

(IV): H_0: ligonių, sergančių X liga, vidutinė BC koncentracija atitinka normą ($m = 5{,}2$);
(H_1): ligonių, sergančių X liga, BC koncentracijos vidurkis yra padidėjęs ($m > 5{,}2$).

Tikrinant (VII) hipotezę, nulinę ir alternatyvią hipotezes galime formuluoti taip:

H_0: „KA skirstinys sutampa su (5.1) skirstiniu“;
H_4: „KA skirstinys nesutampa su (5.1) skirstiniu“. (5.6)

Tikrindami (IX) hipotezę, nulinę hipotezę formuluojame taip: H_0: „A ir B ligonių populiacijų BC koncentracijos skirstiniai yra indentiški“; o alternatyvią – H_5: „A ir B ligonių populiacijų BC koncentracijos skirstiniai nėra indentiški“, arba H_6: „B ligonių populiacijos BC skirstinio tankis pasislinkęs į dešinę nuo A populiacijos skirstinio“.

Alternatyvos H_4, H_5 yra dvipusės, H_6 – dešiniapusė. 5.5 pavyzdyje nulinė hipotezė yra paprasta, nes įgyjamų reikšmių aibė – vienas taškas (vienas skirstinys (5.1)). (IX) pavyzdyje tiek nulinė hipotezė, tiek alternatyva yra sudėtingos.

Iš pateiktų nulinės ir alternatyvios hipotezės pavyzdžių matome, kad nulinė hipotezė formuluojama: „nėra skirtumo...“ (tarp populiacijų vidurkių, tarp skirstinių etc.). Alternatyvi hipotezė formuluojama „yra skirtumas...“, t. y. vienos populiacijos vidurkis (tikimybė) yra didesnis ar mažesnis už kitą, parametrai ir skirstiniai nėra lygūs etc. Alternatyvi hipotezė formuluojama pagal gautus tyrimo rezultatus. Pavyzdžiui, sergančiųjų X liga atsitiktinai parinktų ligonių BC vidurkis $\bar{x}$ lygus 6,3 (mmol/l). Tyrimo rezultatai rodo,

kad $\bar{x} > m_0$; čia $m_0 = 5{,}2$ (mmol/l). Reikia patikrinti, ar ir $m > m_0$; čia m – sergančiųjų X liga populiacijos BC vidurkis. Jei nustatėme, kad vartojusiųjų ramiprilio CV proporcija $p_1 = 0{,}14$, o vartojusiųjų placebą – CV proporcija $p_2 = 0{,}18$, išvadai apie teigiamą ramiprilio poveikį reikia patvirtinti, kad ir $\pi_1 < \pi_2$, todėl formuluojama ir tokia alternatyvi hipotezė.

Taikant statistikos metodus biomedicinoje (taip pat ir kituose tyrimuose) ir tikrinant hipotezes apie skirstinio parametrus, dažniausiai parenkama vienpusė alternatyva. Tyrėjams svarbu, kurios populiacijos parametras yra didesnis, kurios – mažesnis: pagal tai ir daromos medicininės išvados.

5.3. Hipotezių tikrinimas

Tikrinant iškeltą hipotezę H_0, galima priimti teisingą sprendinį arba padaryti dviejų rūšių klaidas (5.1 lentelė). Galima atmesti hipotezę, nors ji yra teisinga. Tokia klaida vadinama I rūšies klaida, arba α rūšies klaida. Taip pat galima padaryti išvadą, kad hipotezė yra teisinga, nors iš tikrųjų ji yra klaidinga. Tai II rūšies klaida, arba β rūšies klaida. Šių klaidų pasėkmės gali būti labai skirtingos. Pavyzdžiui, tikriname hipotezę, kad ligonis serga X liga. Jei padarysime I rūšies klaidą ir sergančio ligonio negydysime, jis gali mirti. O padarę II rūšies klaidą ir gydydami nesergantį ligonį, tik sukeliame jam nepatogumą. I rūšies klaidos tikimybė žymima α, o II rūšies klaidos tikimybė – β.

5.1 lentelė. Hipotezės tikrinimo rezultatai

	H_0 teisinga	H_0 klaidinga
atmetame H_0	I rūšies klaida	teisingas sprendimas
neatmetame H_0	teisingas sprendimas	II rūšies klaida

Taisyklė, pagal kurią, remdamiesi kintamojo imtimi $\boldsymbol{x} = (x_1, x_2 \ldots x_n)$ (x_i – ats. dydis), nulinę hipotezę atmetame arba jai neprieštaraujame, vadiname **statistiniu kriterijumi**. Statistiniam kriterijui apibrėžti naudojama kriterijaus statistika – imties $\boldsymbol{x}$ funkcija. Sakoma, kad imties funkcija $t = t(\boldsymbol{x})$ yra kriterijaus hipotezei H_0 tikrinti statistika, jei:

a) esant teisingai H_0, yra žinomas t skirstinys arba bent jau t asimptotinis skirstinys (asimptotinis skirstinys – skirstinys, į kurį, augant n, artėja t skirstinys);

b) kuo didesnė ar (ir) kuo mažesnė t reikšmė, tuo labiau tikėtina alternatyva.

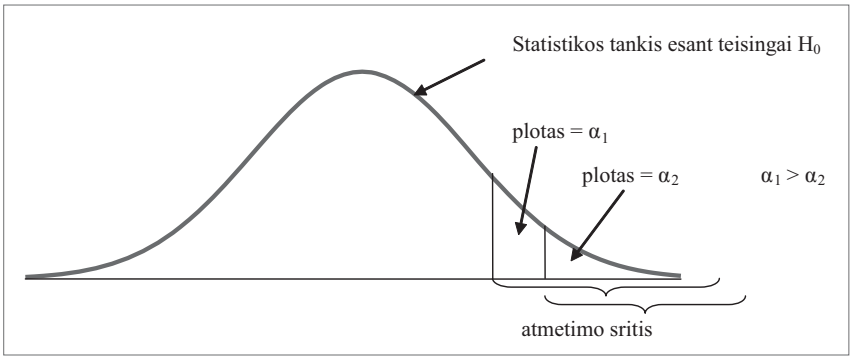
Statistikos $t = t(\boldsymbol{x})$ reikšmių, kurioms esant nulinę hipotezę atmetame, aibė vadinama **kritine sritimi**, arba **atmetimo sritimi**. Remiantis statistikos

b) savybe, atmetimo sritis yra viena iš šių aibių: $\{t: t > a\}$, $\{t: t < a\}$ arba $\{t: |t| > a\}$. Taškas a, ribojantis atmetimo sritį, vadinamas **kritine reikšme**.

Statistinis kriterijus yra tuo geresnis, kuo mažesnės abiejų rūšių klaidų tikimybės. Praktiškai neįmanoma sudaryti kriterijaus, kad abiejų rūšių klaidų tikimybės būtų lygios nuliui. Nenorėdami beprasmiškai atmesti nulinės hipotezės, nustatome mažą leistiną I rūšies klaidos tikimybę α. Ši pasirinkta tikimybė α vadinama **reikšmingumo lygmeniu**. α skaitinės reikšmės gali būti bet koks mažas skaičius. Paprastai naudojamos α reikšmės yra 0,1; 0,05; 0,01; 0,001 ar pan. Standartinė α reikšmė – 0,05. Tai reiškia, kad, naudodami parinktą statistinį kriterijų, vidutiniškai penkiais atvejais iš šimto klaidingai atmesime nulinę hipotezę.

Nustačius reikšmingumo lygmenį α, atmetimo sritis ir kartu kritinė reikšmė priklauso nuo α – kuo α mažesnė, tuo mažesnė ir atmetimo sritis (5.1 pav.). Pažymėkime nulinės hipotezės atmetimo sritį ω_α. ω_α priklauso nuo α ir nuo kriterijaus statistikos tankio esant teisingai H_0. Kadangi tai pačiai H_0 tikrinti gali būti naudojamas ne vienas statistinis kriterijus, būtina palyginti šių kriterijų kokybę. Tolygiai galingiausiu (tam tikra prasme optimaliu) statistiniu kriterijumi laikomas toks, kuriam II rūšies klaidos tikimybė β yra mažiausia esant fiksuotai I rūšies klaidos tikimybei. II rūšies klaidos tikimybė bus mažiausia tada, kai atmetimo sritis ω_α bus didžiausia.

Kriterijai, sudaryti remiantis duomenų skirstinio parametriniu modeliu (šiuo atveju kintamojo skirstinio funkcinė išraiška žinoma, ir statistinės hipotezės yra hipotezės apie skirstinio parametrus), vadinami **parametriniais**. Parametrinio kriterijaus statistikos skirstinio funkcinė išraiška žinoma esant tiek nulinei hipotezei, tiek alternatyvai.



5.1 pav. Atmetimo srities priklausomybė nuo reikšmingumo lygmens

Analizuojant duomenis, ne visuomet galima apsiriboti parametriniu kintamojo modeliu. Kartais daroma prielaida, kad kintamojo skirstinys priklauso

skirstinių šeimai F, kuri yra bendresnė už $P(\theta_1, \theta_2 \ldots \theta_k)$. Šeima F gali būti simetrinių, tolydžiųjų ar diskrečiųjų skirstinių šeima. Šiuo atveju nežinome konkretaus kintamojo skirstinio funkcinės išraiškos; statistinė hipotezė gali būti apie kai kurias kintamojo skaitines charakteristikas (vidurkį, medianą) arba apie patį skirstinį. Statistiniai kriterijai, sudaryti neatsižvelgiant į kintamojo skirstinio funkcinę išraišką, vadinami **neparametriniais**. Tokių kriterijaus statistikų skirstinys dažnai yra diskretusis, priklausantis nuo n. Nagrinėjant hipotezes, nesusijusias su skirstinio parametrais, ne visada įmanoma nustatyti kriterijaus statistikos skirstinį esant alternatyvai.

5.4. Parametrinio kriterijaus, turinčio didžiausią atmetimo sritį, sudarymas*

Parametrinių hipotezių tikrinimo teoriniai aspektai išdėstyti [3] vadovėlyje. Šiame skyriuje pateiksime ne tokį matematiškai griežtą dėstymo variantą.

Ne visoms parametrinėms hipotezėms galima sudaryti kriterijų, turintį didžiausią atmetimo sritį. Čia pateiksime atvejus, kai egzistuoja kriterijus su didžiausia atmetimo sritimi, ir nurodysime, kaip šis kriterijus sudaromas.

Pažymėkime $p(\boldsymbol{x}, \boldsymbol{\theta})$ – imties $\boldsymbol{x} = (x_1, x_2 \ldots x_n)$ (x_i – atsitiktiniai dydžiai) tikėtinumo funkcija. Tegul $p(\boldsymbol{x}, \boldsymbol{\theta})$ priklauso parametrinių skirstinių šeimai $P(\boldsymbol{\theta})$; čia $\boldsymbol{\theta} = (\theta_1 \ldots \theta_k)$ – k-matis parametras. Pasirenkame reikšmingumo lygmenį, lygų α.

1) Tikrinama paprasta nulinė hipotezė H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ su paprasta alternatyva H_A: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. Tuomet didžiausia atmetimo sritis ω_α apibrėžiama taip:

$$\omega_\alpha = \{\boldsymbol{x}: p(\boldsymbol{x}, \boldsymbol{\theta}_1) / p(\boldsymbol{x}, \boldsymbol{\theta}_0) > c_\alpha\}. \tag{5.7}$$

Praktikoje retai sutinkama paprasta nulinė hipotezė ir paprasta alternatyva. Dažniausiai susiduriame su paprasta nuline hipoteze ir sudėtinga alternatyva arba su sudėtinga tiek nuline hipoteze, tiek alternatyva. Pateiksime $p(\boldsymbol{x}, \boldsymbol{\theta})$ atvejus, kai galima nustatyti didžiausią atmetimo sritį.

2) $k = 1$, t. y. parametras θ vienmatis bei $p(\boldsymbol{x}, \boldsymbol{\theta})$ priklauso vienparametrei eksponentinių skirstinių šeimai, t. y. egzistuoja vienmatės funkcijos $C(\theta)$, $Q(\theta)$ – monotoninė, $T(\boldsymbol{x})$ ir $h(\boldsymbol{x})$, tokios, kad imties tikėtinumo funkciją $p(\boldsymbol{x}, \theta)$ galime išreikšti:

$$p(\boldsymbol{x}, \theta) = C(\theta)\exp(Q(\theta)T(\boldsymbol{x}))h(\boldsymbol{x}).$$

Tikrinant paprastą nulinę hipotezę H_0: $\theta = \theta_0$ su dešiniapuse alternatyva H_1: $\theta > \theta_0$, didžiausia atmetimo sritis ω_α apibrėžiama:

$$\omega_\alpha = \{\boldsymbol{x}: T(\boldsymbol{x}) > c_{\alpha 1}\}; \tag{5.8}$$

tikrinant H_0 su kairiapusia alternatyva H_2: $\theta < \theta_0$, didžiausia atmetimo sritis ω_α apibrėžiama:

$$\omega_\alpha = \{\boldsymbol{x}: T(\boldsymbol{x}) < c_{\alpha 2}\}; \tag{5.9}$$

jei $T(\boldsymbol{x})$ – simetrinė funkcija, tuomet tikrinant minėtą nulinę hipotezę su dvipuse alternatyva H_3: $\theta \neq \theta_0$, didžiausia atmetimo sritis apibrėžiama:

$$\omega_\alpha = \{\boldsymbol{x}: |T(\boldsymbol{x})| > c_{\alpha 3}\}. \tag{5.10}$$

3) $k > 1$, $p(\boldsymbol{x}, \boldsymbol{\theta})$ priklauso eksponentinių skirstinių šeimai. Daroma prielaida, kad daugiamatį parametrą $\boldsymbol{\theta}$ galima išreikšti pavidalu $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$ taip, kad statistinę hipotezę galima performuluoti į nulinę hipotezę H_0: $\psi = \psi_0$ ir alternatyvą, konkretizuojančią tik vienmatį parametrą ψ, o $(k - 1)$ matavimo parametro $\boldsymbol{\lambda}$ atžvilgiu jokios hipotezės nekeliamos. Parametras $\boldsymbol{\lambda}$ šiuo atveju vadinamas trukdančiu.

Atitinkamai pakeitus, (3.1) formulę galima perrašyti taip:

$$p(\boldsymbol{x}, \psi, \boldsymbol{\lambda}) = C(\psi, \boldsymbol{\lambda})\exp(\psi U(\boldsymbol{x}) + \sum\nolimits_{j=1}^{k-1} \lambda_j T_j(\boldsymbol{x}))h(\boldsymbol{x}); \tag{5.11}$$

čia $U(\boldsymbol{x})$ ir $\boldsymbol{T}(\boldsymbol{x}) = (T_1(\boldsymbol{x}), ..., T_{k-1}(\boldsymbol{x}))$ – imties funkcijos, nepriklausančios nuo parametrų $(\psi, \boldsymbol{\lambda})$. Sakykime, $h(U, \boldsymbol{T}) = a(\boldsymbol{T})U + b(\boldsymbol{T})$; čia $a(\boldsymbol{T}) > 0$ – imties funkcija, sudaryta taip, kad $h(U, \boldsymbol{T})$ nepriklauso nuo $\boldsymbol{T}$, kai $\psi = \psi_0$. Tuomet didžiausia H_0 atmetimo sritis ω_α yra tokia:

$\omega_\alpha = \{\boldsymbol{x}: h(U, \boldsymbol{T}) > c_{\alpha 1}\}$, kai alternatyva H_1: $\psi > \psi_0$;

$\omega_\alpha = \{\boldsymbol{x}: h(U, \boldsymbol{T}) < c_{\alpha 2}\}$, kai alternatyva H_2: $\psi < \psi_0$;

$\omega_\alpha = \{\boldsymbol{x}: h(U, \boldsymbol{T}) > c_{\alpha 3}$ arba $h(U, \boldsymbol{T}) < c_{\alpha 4}\}$, kai alternatyva H_3: $\psi \neq \psi_0$.

Paprastos H_0 ir paprastos alternatyvos atveju kriterijaus, turinčio didžiausią atmetimo sritį, statistika T_n (turinti žinomą skirstinį, kai H_0 yra teisinga) sudaroma remiantis tankių santykio $p(\boldsymbol{x}, \boldsymbol{\theta}_1)/p(\boldsymbol{x}, \boldsymbol{\theta}_0)$ skirstiniu. Jei tenkinamos 2 punkto sąlygos, kriterijaus statistika T_n sudaroma kaip statistikos $T(\boldsymbol{x})$ funkcija: $T_n = f(T(\boldsymbol{x}))$; čia funkcija $f(x)$ parenkama tokia, kad T_n tenkintų kriterijaus statistikos sąlygas. Pavyzdžiui, jei $T(\boldsymbol{x})$ yra n nepriklausomų ar silpnai priklausomų ats. d. suma, statistika gali būti standartizuotas santykis:

$$T_n = \frac{T(x) - T_0}{SE_0(T)}; \tag{5.12}$$

čia T_0 – funkcijos $T(\boldsymbol{x})$ vidurkis, esant teisingai nulinei hipotezei, $SE_0(T)$ – $T(\boldsymbol{x})$ standartinio nuokrypio įvertis, esant teisingai H_0. Remiantis centrine ribine teorema, T_n skirstinys yra asimptotiškai normalusis. Analogiškai kri-

terijaus statistika sudaroma ir hipotezėms, tenkinančioms 3 punkto sąlygas, tik vietoj $T(\boldsymbol{x})$ rašoma $h(U, T)$.

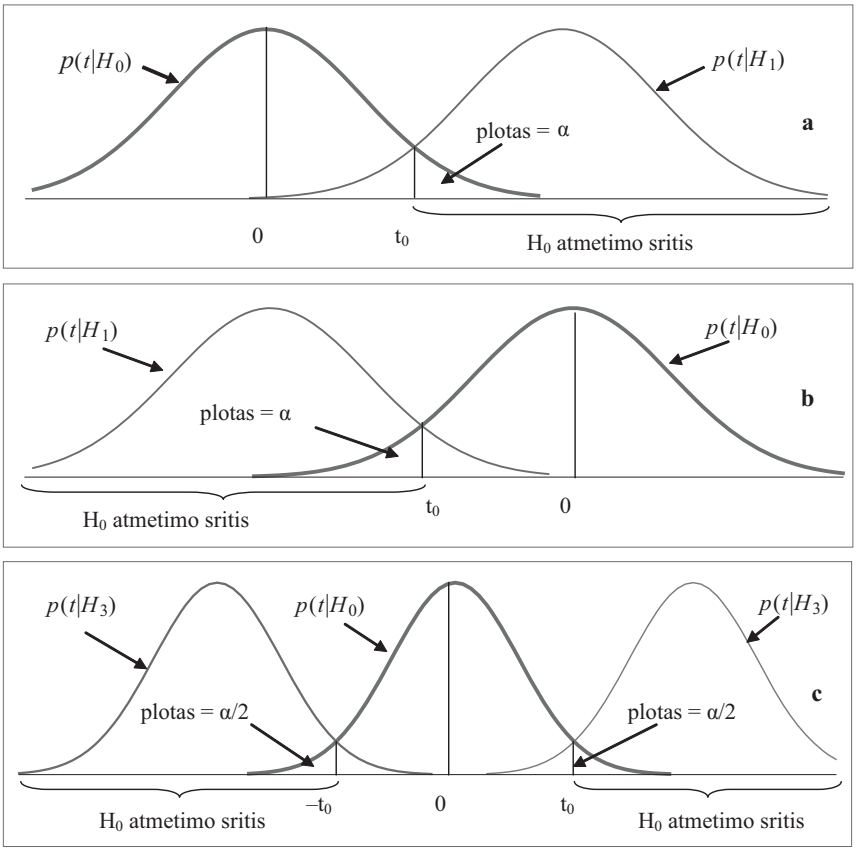
5.5. Atmetimo srities nustatymas parametrinės hipotezės atveju

Sakykime, θ yra vienmatis, $p(\boldsymbol{x}, \theta)$ priklauso eksponentinių skirstinių šeimai (tenkinamos 2 punkto sąlygos). Pažymėkime T_n – kriterijaus statistika, sudaryta remiantis $T(\boldsymbol{x})$ skirstiniu; T_{n0} – ats. d., turintis T_n skirstinį esant teisingai H_0; $p(t|H_0)$ ir $p(t|H_A)$ – T_n tankis (arba tikimybė), esant nulinei hipotezei ir alternatyvai.

1. Tikrinama paprasta nulinė hipotezė H_0: $\theta = \theta_0$ su dešiniapuse alternatyva H_1: $\theta > \theta_0$. Kadangi tikėtinumo funkcija priklauso eksponentinių skirstinių šeimai, statistikos T_n tankio funkcija, esant teisingai alternatyvai, bus T_n tankio funkcijos, nustatytos esant teisingai nulinei hipotezei, dešinėje (5.2 a pav.). Atmetimo sritis ω_α bus į dešinę nuo kritinės reikšmės t_0, hipotezės priėmimo sritis – į kairę nuo t_0. Pirmos rūšies klaidos tikimybė priklauso nuo t_0 ir lygi $\alpha(t_0) = P\{\boldsymbol{x}: T_n \geq t_0 \,|H_0\}$, antros rūšies klaidos tikimybė $\beta(t_0) = P\{x: T_n < t_0 \,|H_1\}$. Kadangi su tikimybe α leidžiama atmesti teisingą nulinę hipotezę, todėl t_0 reikšmė apibrėžiama lygybe: $\alpha = P\{T_n \geq t_0 \,|H_0\} = P\{\boldsymbol{x}: T_{n0} \geq t_0\}$. Pagal apibrėžimą, t_0 yra ats. dydžio T_{n0} skirstinio 1 – α lygio kvantilis; taigi atmetimo sritis bus visos statistikos T_n reikšmės, didesnės už ats. dydžio T_{n0} 1 – α lygio kvantilį $t_{1-\alpha;n}$ (5.2 a pav.).

2. Tikriname H_0: $\theta = \theta_0$ su kairiapuse alternatyva H_2: $\theta < \theta_0$. Tuomet T_n tankis, esant teisingai alternatyvai, bus pasislinkęs į kairę nuo T_n tankio funkcijos, nustatytos esant teisingai H_0 (5.2 b pav.). Atmetimo sritis ω_α bus kritinės reikšmės t_0, nusakomos lygybe: $\alpha = P\{\boldsymbol{x}: T_n \leq t_0 \,|H_0\} = P\{T_{n0} \leq t_0\}$, kairėje; taigi atmetimo sritis bus T_n reikšmės, mažesnės už ats. dydžio T_{n0} α lygio kvantilį $t_{\alpha;n}$. (5.2 b pav.).

3. Tikriname H_0: $\theta = \theta_0$ su su dvipuse alternatyva H_3: $\theta \neq \theta_0$. Tuomet statistikos T_n tankio funkcija, esant teisingai alternatyvai, $p(t|H_3)$ gali būti tiek kairėje, tiek dešinėje statistikos T_n tankio funkcijos $p(t|H_0)$ pusėje (5.2 c pav.). Atmetimo sritis yra: $\omega_\alpha = \{\boldsymbol{x}: T_n < t_{01}$ arba $T_n > t_{02} \,|H_0\}$. Kritinės reikšmės t_{01} ir t_{02} nusakomos lygybėmis $\alpha/2 = P\{\boldsymbol{x}: T_n < t_{01} \,|H_0\}$ ir $\alpha/2 = P\{T_n > t_{02} \,|H_0\}$; taigi t_{01} ir t_{02} yra T_n skirstinio, esant teisingai H_0, atitinkamai α/2 ir 1 – α/2 lygio kvantiliai. Jei $p(t|H_0)$ simetriška, tai $t_{01} = |t_{02}| = t_0$ ir atmetimo sritis yra statistikos T_n reikšmės, absoliučiu dydžiu didesnės už T_{n0} skirstinio 1 – α/2 lygio kvantilį $t_{1-\alpha/2;n}$: $\omega_\alpha = \{\boldsymbol{x}: |T_{n0}| > t_{1-\alpha/2;n}\}$ (5.2 c pav.).



5.2 pav. Kritinių reikšmių ir atmetimo sričių pavyzdžiai; t_0 – kritinė reikšmė

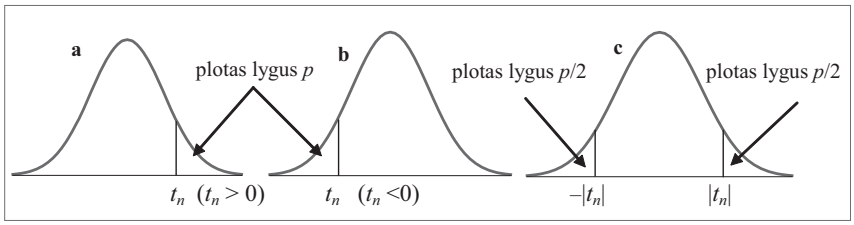
Ats. dydžio T_{n0} skirstinys, augant n, gali artėti prie tam tikro ribinio skirstinio T_0. Tuomet sakoma, kad T_{n0} asimptotinis skirstinys yra T_0. Jei T_{n0} skirstinys priklauso nuo n, tuomet nedideliam n jo kvantiliai pateikiami lentelėse. Dideliems n vietoj T_{n0} skirstinio kvantilių naudojami asimptotinio skirstinio T_0 atitinkamos eilės kvantiliai.

Praktiškai tikrinant hipotezę, konkretaus tyrimo metu gautai imčiai skaičiuojama statistikos T_n reikšmė t_n ir lyginama su T_{n0} skirstinio atitinkamo lygio kvantiliu $t_{1-\alpha;n}$; $t_{\alpha;n}$; $t_{1-\alpha/2;n}$. Pavyzdžiui, tikrinant H_0 su alternatyva H_1, jei $t_n > t_{1-\alpha;n}$, nulinę hipotezę atmetame, pasirenkame alternatyvą – $\theta > \theta_0$. Jei $t_n \leq t_{1-\alpha;n}$, nulinei hipotezei – „$\theta = \theta_0$“ – neprieštaraujame.

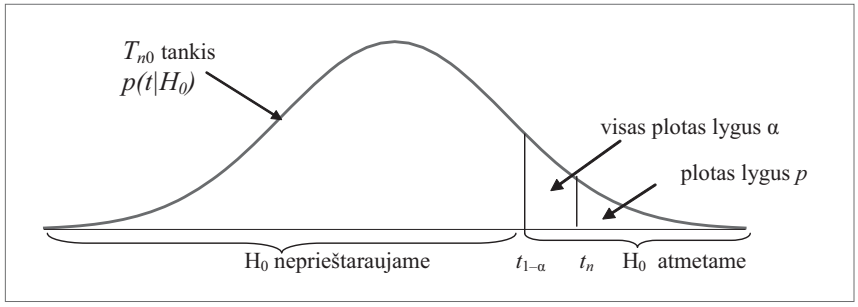
Nustatyti, ar tyrimo metu gauta imtis patenka į kritinę sritį, galima ne tik kriterijaus statistikos T_n konkrečią reikšmę t_n lyginant su T_{n0} skirstinio atitinkamo lygio kvantiliu, bet ir skaičiuojant kriterijaus p reikšmę. Vienpusė

(*one-sided*) kriterijaus *p* reikšmė apibrėžiama $p = P\{T_n > t_n \,|H_0\} = P\{T_{n0} > t_n\}$, jei naudojama dešiniapusė alternatyva ($t_n > 0$), bei $p = P\{T_n < t_n \,|H_0\}$, jei naudojama kairiapusė alternatyva (5.3 a, b pav.). Dvipusė *p* reikšmė (*two-sided*, 2-*tailed*) apibrėžiama: $p = P\{|T_n|>|t_n| \,|H_0\} = P\{|T_{n0}|>|t_n|\}$ (5.3 c pav.). Jei kriterijaus statistikos, esant nulinei hipotezei, skirstinys yra χ^2 arba Fišerio, skaičiuojamos tik vienpusė *p* reikšmė $p = P\{T_{n0} > t_n\}$. Standartinio normaliojo ar Stjudento skirstinio atveju skaičiuojamos tiek vienpusė, tiek dvipusė *p* reikšmės.

Jei statistikos *p* reikšmė mažesnė nei reikšmingumo lygmuo α, imties reikšmė patenka į atmetimo sritį (5.4 pav.). Tuomet nulinę hipotezę atmetame ir priimame alternatyvą; priešingu atveju nulinei hipotezei neprieštaraujame. Kriterijaus statistikos dvipusė ar vienpusė *p* reikšmės pateikiamos statistiniuose paketuose.



5.3 pav. Vienpusė ir dvipusė p reikšmė



5.4 pav. Atmetimo sritis ir vienpusė p reikšmė

5.6. Hipotezės apie normaliojo skirstinio vidurkį tikrinimas

Šiame skyriuje pateiksime pavyzdį, kaip sudaromas kriterijus, skirtas konkrečiai hipotezei tikrinti. Sakykime, nagrinėjamas kintamasis yra kiekybinis, turintis normalųjį skirstinį su nežinomais parametrais – vidurkiu *m* ir dispersija σ^2. Tikrinsime hipotezę apie vidurkio lygybę skaičiui. Šiuo atveju turime dvimatį parametrą $\boldsymbol{\theta} = (m, \sigma^2)$, imties tikėtinumo funkcija $p(x, m, \sigma)$

priklauso eksponentinių skirstinių šeimai. Nulinė hipotezė H_0: $m = m_0$ konkretizuoja tik vidurkį m; σ yra trukdantis parametras. Pažymėkime $\psi = m/\sigma^2$, $\lambda = -1/(2\sigma^2)$; matome, kad nulinė hipotezė konkretizuoja tik parametrą ψ. Mūsų nagrinėjamas atvejis tenkina 3 punkto sąlygas su funkcijomis $U(x)$, $T_1(x)$ ir $h(U, T)$ (5.12) formulėje:

$$U(\boldsymbol{x}) = \sum_{i=1}^{n} x_i,\ T_1(x) = \sum_{i=1}^{n} x_i^2,\ (\bar{x} = U(\boldsymbol{x})/n,\ s^2 = (T_1(\boldsymbol{x}) - (U(\boldsymbol{x}))^2/n)/(n-1)),$$

$$h(U, T) = \frac{\sqrt{n}(\bar{x} - m_0)}{s} = \frac{\sum_{i=1}^{n} x_i - nm_0}{n\sqrt{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 / n}} = \frac{U - nm_0}{n\sqrt{T - U^2/n}}. \quad (5.13)$$

Kadangi imties skirstinys yra normalusis, tai $h(U, T)$ skirstinys, kai $m = m_0$, nuo imties funkcijos T nepriklauso (žr. [6, 8]) ir yra Stjudento skirstinys su $(n - 1)$ laisvės laipsnių. Todėl kriterijus hipotezei apie vidurkio lygybę skaičiui tikrinti vadinamas ***t* kriterijumi vienai imčiai** (*one sample t-test*). Pateiksime šio kriterijaus aprašymą.

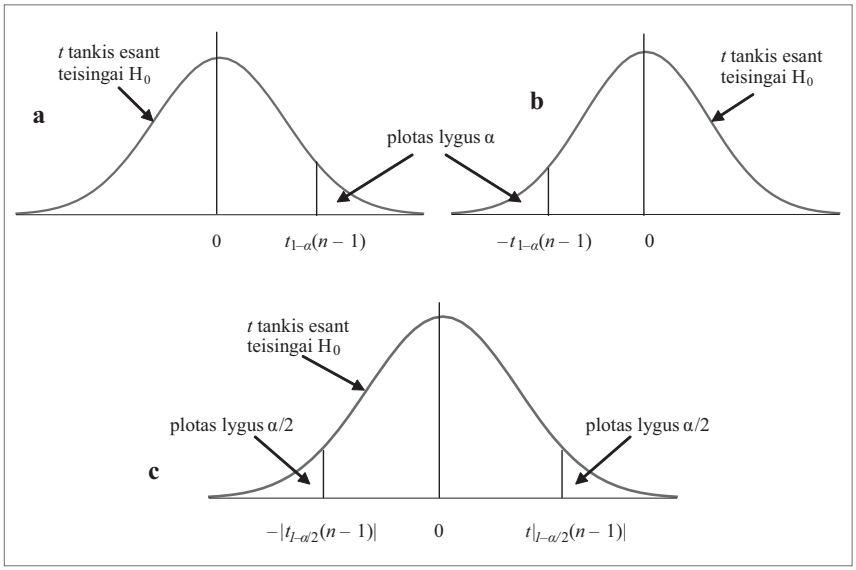
t kriterijus vienai imčiai. Tikriname nulinę hipotezę H_0: $m = m_0$ (tiriamo kintamojo skirstinio vidurkis m lygus konkrečiam skaičiui m_0; pvz., „BC koncentracijos vidurkis lygus 5,2“). Alternatyvos šiai nulinei hipotezei gali būti H_1: $m > m_0$; H_2: $m < m_0$; H_3: $m \neq m_0$, α – parinktas reikšmingumo lygmuo.

H_0 tikrinti naudosime kriterijaus statistiką t, sutampančią su funkcija $h(U, T)$:

$$t = \frac{\sqrt{n}(\bar{x} - m_0)}{s}. \quad (5.14)$$

Esant teisingai nulinei hipotezei, statistika t turi Stjudento skirstinį su $(n - 1)$ laisvės laipsniu. Pasirinkus dešiniapusę alternatyvą H_1, H_0 atmetimo sritis yra $t > t_{1-\alpha}(n-1)$; čia $t_{1-\alpha}(n-1)$ – Stjudento skirstinio su $(n - 1)$ laisvės laipsnių $(1 - \alpha)$ lygio kvantilis (5.5 a pav.). Tikrinant H_0 su kairiapuse alternatyva H_2, hipotezės atmetimo sritis yra $t < -t_{1-\alpha}(n-1)$, o tikrinant H_0 su dvipuse alternatyva H_3: $m \neq m_0$ – atmetimo sritis $|t| > t_{1-\alpha/2}(n-1)$ (5.5 c pav.).

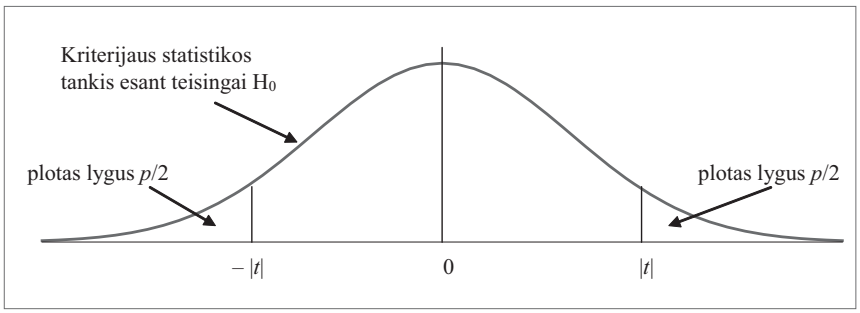
Statistiniuose paketuose pateikiama t kriterijaus vienai imčiai statistikos reikšmė, laisvės laipsniai ir dvipusė p reikšmė (5.6 pav.). Kaip jau minėjome, jei p mažesnė už pasirinktą reikšmingumo lygmenį, H_0 atmetame ir laikome, kad teisinga yra dvipusė alternatyva. Tikrinant H_0 su vienpuse alternatyva, H_0 atmetame, jei $p/2$ mažesnė už parinktą reikšmingumo lygmenį α. Atmetimo sritis, sprendinio priėmimo taisyklė pagal kriterijaus dvipusę reikšmę pateikta 5.2 lentelėje.



5.5 pav. t kriterijaus vienai imčiai nulinės hipotezės atmetimo sritis

5.2 lentelė. H_0: $m = m_0$ atmetimo sritis, sprendinio priėmimo taisyklė pagal kriterijaus dvipusę reikšmę (t – statistikos reikšmė)

Alternatyva	Atmetimo sritis	H_0 atmetimo taisyklė pagal kriterijaus dvipusę *p* reikšmę
H_1: $m > m_0$	$t > t_{1-\alpha}(n-1)$	$\bar{x} > m_0$, $p/2 < \alpha$
H_2: $m < m_0$	$t < -t_{1-\alpha}(n-1)$	$\bar{x} < m_0$, $p/2 < \alpha$
H_3: $m \neq m_0$	$\lvert t \rvert > t_{1-\alpha/2}(n-1)$	$p < \alpha$



5.6 pav. t kriterijaus dvipusė p reikšmė; t – statistikos reikšmė

5.7. Hipotezių tikrinimas ranginiais kriterijais

Kaip minėta 3 skyriuje, daugeliu atvejų tiriamo rodiklio skirstinį galima laikyti normaliuoju. Tačiau analizuojant medikų surinktus duomenis, pasitaiko:

- išskirčių imtyje;
- turime nedaug matavimų ir negalime daryti išvados apie duomenų normalumą;
- sunku nustatyti tikslias rodiklio reikšmes, tačiau pagal jį nesunku palyginti individus tarpusavyje – suranguoti.

Minėtais atvejais hipotezėms tikrinti netinka kriterijai, sudaryti remiantis konkrečiu imties skirstiniu (parametriniu modeliu). Tačiau hipotezėms tikrinti yra daug kriterijų, nesusijusių su konkrečiu tiriamo kintamojo skirstiniu. Tai – neparametriniai kriterijai. Jie nereikalauja tokių griežtų prielaidų kintamojo skirstiniui kaip normalumas, tačiau jų atmetimo sritis nėra didžiausia. Neparametrinių kriterijų taikymo prielaidos kuklesnės: tai skirstinio simetriškumas ar kelių skirstinių formos vienodumas. Dauguma neparametrinių kriterijų naudoja ne kintamojo reikšmes, o jų rangus – reikšmės vietą variacinėje sekoje.

Apibrėšime imties reikšmių rangus. Jei imtyje nėra pasikartojančių reikšmių, jų rangai sutampa su eilės numeriu variacinėje sekoje. Jei imtyje yra pasikartojančių reikšmių, vienodoms reikšmėms priskiriamas vidutinis rangas. Rangų skaičiavimą iliustruosime pavyzdžiais.

5.1 pavyzdys. Turime imties reikšmes $x_1 = 3{,}1$; $x_2 = 2{,}9$; $x_3 = 3{,}0$; $x_4 = 3{,}2$; $x_5 = 2{,}8$. Šios imties variacinė seka yra $x_{(1)} = 2{,}8$; $x_{(2)} = 2{,}9$; $x_{(3)} = 3{,}0$; $x_{(4)} = 3{,}1$; $x_{(5)} = 3{,}2$. Reikšmės 3,1 rangas – 4, reikšmės 2,9 rangas – 2, 3,0 – 3, 3,2 – 5, 2,8 – 1.

5.2 pavyzdys. Turime imties reikšmes $x_1 = 6$; $x_2 = 5$; $x_3 = 7$; $x_4 = 3$; $x_5 = 6$; $x_6 = 6$. Šios imties variacinė seka yra $x_{(1)} = 3$; $x_{(2)} = 5$; $x_{(3)} = 6$; $x_{(4)} = 6$; $x_{(5)} = 6$; $x_{(6)} = 7$. Reikšmės 3 rangas yra 1, 5 rangas – 2. Imtyje yra trys šešetai, jų numeriai variacinėje sekoje – 3, 4 ir 5. Šių numerių vidurkis lygus 4, taigi reikšmės 6 rangas yra 4, o 7 rangas – 6:

Variacinė seka:	3	5	6	6	6	7
Rangai:	1	2	4	4	4	6

Statistiniai kriterijai, naudojantys ne kintamojo reikšmes, o rangus, vadinami ranginiais kriterijais. Ranginiai kriterijai yra gana paprasti ir suprantami. Juos galima taikyti ir tuomet, kai pateiktos ne išmatuotos reikšmės, o jų rangai. Pateiksime ranginį kriterijų, skirtą dviejų kartotinių matavimų vidurkiams palyginti.

Vilkoksono kriterijus dviejų kartotinių matavimų vidurkiams palyginti (*Wilcoxon signed rank test)*. Tiriama atsitiktinai atrinktų n individų. i-tajam individui nustatoma kiekybinio kintamojo reikšmė prieš veiksnį x_i ir reikšmė po veiksnio y_i, $i = 1, 2 \ldots n$. Pažymėkime $z_i = y_i - x_i$. Darome tokią prielaidą z_i skirstiniui: $z_i = \theta + e_i$; čia e_i – nepriklausomi, tą patį tolydųjį simetrišką skirstinį turintys ats. d. Kitaip tariant, daroma prielaida, kad $y_i - x_i$ skirstinys yra tolydusis ir simetriškas vidurkio atžvilgiu.

Veiksnio įtakai (poveikiui) konstatuoti tikrinsime nulinę hipotezę H_0: $\theta = 0$ su viena alternatyvų: $\theta > 0$, $\theta < 0$ ir $\theta \neq 0$. Skaičiuojant Vilkoksono kriterijaus statistiką, atliekami šie veiksmai:

1. Skaičiuojami nenulinių skirtumų absoliutieji dydžiai $|z_1|$, $|z_2|$, ..., $|z_n|$ (nuliui lygūs z_i atmetami). Dydžio $|z_i|$ rangą pažymėkime R_i.

2. Skaičiuojame φ_i: $\varphi_i = 1$, jei $z_i > 0$, ir $\varphi_i = 0$, jei $z_i < 0$.

3. Sudarome kriterijaus statistiką

$$T^+ = \sum_{i=1}^{n} R_i \varphi_i \text{ arba } T^- = \sum_{i=1}^{n} R_i (1 - \varphi_i).$$

Į T^+ ir T^- sumas neįtraukiami tie $R_i\varphi_i$ ir $R_i(1 - \varphi_i)$, kurių $z_i = 0$. Kadangi daryta prielaida, kad z_i skirstinys yra tolydusis, reikšmių $z_i = 0$ imtyje neturėtų būti daug. Sandauga $R_i\varphi_i$ vadinama z_i teigiamo ženklo rangu. Jei z_i neigiamas, tuomet $R_i\varphi_i = 0$, jei z_i teigiamas, tuomet $R_i\varphi_i$ lygus $|z_i|$ rangui. Statistika T^+ yra teigiamo, T^- – neigiamo ženklo rangų suma, $T^+ + T^- = n(n + 1)/2$; čia n – nelygių 0 z_i skaičius. Jei $\theta > 0$, tuomet teigiamo ženklo rangų yra daugiau negu neigiamo ir labai tikėtina, kad T^+ yra didesnis už T^-.

Kriterijaus statistika T^+ – atsitiktinis dydis, įgyjantis sveikas reikšmes nuo 0 iki $n(n + 1)/2$. T^+ skirstinį galima gauti iš priklausomybės $T^+ = \sum_{i=1}^{b} r_i$; čia r_i – teigiamų z_i rangų suma, $r_1 < r_2 < \ldots < r_b$, b – teigiamų z_i skaičius. Esant teisingai nulinei hipotezei, kiekvieno iš 2^n skirtingų rangų rinkinių ($r_1 \ldots r_b$) pasirodymo tikimybė lygi $1/2^n$. Todėl T^+ turi vidurkio atžvilgiu simetrišką diskretųjį skirstinį, priklausantį tik nuo n. 5.3 lentelėje pateiktas T^+ skirstinys, kai $n = 3$. Lentelėse pateikiamos T^+ skirstinio (nedideliam n) p reikšmės arba atitinkamo lygio kvantiliai $t(1 - \alpha, n)$, arba $t(\alpha, n)$; čia α – parinktas reikšmingumo lygmuo. Kadangi T^+ skirstinys yra simetriškas vidurkio atžvilgiu, todėl teisinga priklausomybė: $t(\alpha, n) + t(1 - \alpha, n) = n(n + 1)/2$. 5 lentelėje pateikti T^+ skirstinio 0,025; 0,05; 0,95 ir 0,975 lygio kvantiliai.

Nulinės hipotezės H_0: $\theta = 0$ tikrinimas nedideliam n ($n \leq 30$):

- alternatyva H_1: $\theta > 0$: H_0 atmetame, jei $T^+ > t(1 - \alpha, n)$ (arba $T^+ > n(n + 1)/2 - t(\alpha, n)$); H_0 neprieštaraujame, jei $T^+ \leq t(1 - \alpha, n)$;

- alternatyva H_2: $\theta < 0$: H_0 atmetame, jei $T^+ < n(n+1)/2 - t(1-\alpha, n)$ (arba $T^+ < t(\alpha, n)$); H_0 neprieštaraujame, jei $T^+ \geq t(\alpha, n)$;
- alternatyva H_3: $\theta \neq 0$: H_0 atmetame, jei arba $T^+ > t(1-\alpha/2, n)$ arba $T^+ < n(n+1)/2 - t(1-\alpha/2, n)$, H_0 neprieštaraujame, jei $n(n+1)/2 - t(1-\alpha/2, n) \leq T^+ \leq t(1-\alpha/2, n)$ (arba $t(\alpha/2, n) \leq T^+ \leq t(1-\alpha/2, n)$).

5.3 lentelė. T^+ skirstinys, $n = 3$

b	$(r_1 \ldots r_b)$	$P_0\{(r_1 \ldots r_b)\}$	$T^+ = \sum_{i=1}^{b} r_i$
0		1/8	0
1	$r_1 = 1$	1/8	1
1	$r_1 = 2$	1/8	2
1	$r_1 = 3$	1/8	3
2	$r_1 = 1, r_2 = 2$	1/8	3
2	$r_1 = 1, r_2 = 3$	1/8	4
2	$r_1 = 2, r_2 = 3$	1/8	4
3	$r_1 = 1, r_2 = 2, r_3 = 3$	1/8	5

Dideliam n ($n > 30$) H_0 tikrinti naudojamas asimptotinis kriterijus. Esant teisingai nulinei hipotezei, (5.13) tipo statistika

$$T^* = (T^+ - ET^+)/se(T^+) = (T^+ - n(n+1)/4)/\sqrt{n(n+1)(2n+1)/24} \quad (5.15)$$

turi asimptotinį standartinį normalųjį skirstinį. Dideliems n ($n > 30$) H_0 atmetimo ar priėmimo taisyklė pagal alternatyvą yra:

- alternatyva $\theta > 0$: H_0 atmetame, jei $T^+ > z_{1-\alpha}$, H_0 neprieštaraujame, jei $T^+ \leq z_{1-\alpha}$;
- alternatyva $\theta < 0$: H_0 atmetame, jei $T^+ < -z_{1-\alpha}$, H_0 neprieštaraujame, jei $T^+ \geq -z_{1-\alpha}$;
- alternatyva $\theta \neq 0$: H_0 atmetame, jei $|T^+| > z_{1-\alpha/2}$; H_0 neprieštaraujame, jei $|T^+| \leq z_{1-\alpha/2}$.

Statistiniu paketu taikant Vilkoksono kriterijų, pateikiamos kriterijaus statistikos T^+ ir T^- reikšmės bei tikslaus ir asimptotinio kriterijaus p reikšmės. Nulinės hipotezės ar alternatyvos priėmimo taisyklė pagal p reikšmę analogiška taisyklei, pateiktai 5.2 lentelėje, tik vietoj sąlygos $\bar{x} > m_0$ ar $\bar{x} < m_0$ naudojama $T^+ > T^-$ ar $T^+ < T^-$.

5.8. Tikėtinumų santykio kriterijus*

Kaip minėta 5.4 skyriuje, tikrinant paprastą hipotezę su paprasta alternatyva, didžiausia atmetimo sritis nusakoma tikėtinumų santykiu $p(\boldsymbol{x}, \theta_1)/p(\boldsymbol{x}, \theta_0)$ (žr. (5.7)). Jei tikėtinumo funkcija priklauso eksponentinių skirstinių šeimai (tikėtinumo funkcija apibrėžiama (5.9) formule), tikrinant paprastą nulinę hipotezę H_0: $\theta = \theta_0$ (θ vienmatis) su vienpuse ar dvipuse alternatyva, didžiausia atmetimo sritis (5.10–5.12) nusakoma $T(\boldsymbol{x})$ skirstiniu arba tikėtinumų santykiu

$$p(x, \hat{\theta})/p(x, \theta_0); \tag{5.16}$$

čia $\hat{\theta}$ – parametro θ didžiausio tikėtinumo įvertis.

Panagrinėsime bendresnį parametrinės hipotezės atvejį. Sakykime, kintamojo imties $\boldsymbol{x} = (x_1, x_2 \ldots x_n)$ skirstinys priklauso parametrinių skirstinių šeimai $P(\theta_1 \ldots \theta_k)$. Tikrinama nulinė hipotezė, konkretizuojanti tik dalį šių parametrų. Jei $\boldsymbol{\theta} = (\psi, \lambda)$ (čia ψ – vienmatis), H_0 konkretizuoja tik ψ parametrą ir tikėtinumo funkcija yra (5.12) pavidalo, tuomet statistinis kriterijus, turintis didžiausią atmetimo sritį, apibrėžiamas funkcija $h(U, T)$ (5.4 skyrius). Iš tikėtinumo funkcijos nereikalaujant (5.12) pavidalo, didžiausią atmetimo sritį turintį kriterijų galima pateikti tik tada, jei egzistuoja statistika $t(\boldsymbol{x})$, kurios skirstinys priklauso tik nuo parametro ψ. Jei tikėtinumo santykis

$$p_T(t, \psi_1)/p_T(t, \psi_0) \tag{5.17}$$

(čia $p_T(t, \psi)$ – $t(\boldsymbol{x})$ tankis) monotoninis t atžvilgiu, tuomet didžiausia H_0: $\psi = \psi_0$ su alternatyva $\psi > \psi_0$ atmetimo sritis ω_α apibrėžiama:

$$\omega_\alpha = \{t = t(\boldsymbol{x}): p_T(t, \psi_1)/p_T(t, \psi_0) > c_\alpha\}.$$

Tačiau tikrinant sudėtingas hipotezes, pasitaiko atvejų, kai:

1) ne visada galima rasti funkciją $h(U, T)$ arba $t(x)$ tokią, kad tikėtinumo santykis (5.17) būtų monotoninis;

2) galimos kelios statistikos $t(x)$ ir būna sunku jas sujungti į vieną kriterijų.

Todėl tikrinant daug sudėtingų hipotezių, didelės imties atveju naudojamas tikėtinumų santykio kriterijus. Pateiksime šį kriterijų.

Sakykime, statistinė hipotezė konkretizuoja tik pirmuosius s parametrus. Parametrų vektorių $\boldsymbol{\theta}$ išskaidykime į dvi dalis: $\boldsymbol{\theta} = (\theta_1 \ldots \theta_k) = (\theta_1 \ldots \theta_s, \theta_{s+1} \ldots \theta_k) = (\boldsymbol{\theta}_s, \boldsymbol{\theta}_{k-s})$, $\boldsymbol{\theta}_s$ – s-matis, o $\boldsymbol{\theta}_{k-s}$ – $(k - s)$ matavimų vektorius. Tegul nulinė hipotezė konkretizuoja pirmuosius s parametrus: H_0: $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{s0}$ su alternatyva

$\boldsymbol{\theta}_s \neq \boldsymbol{\theta}_{s0}$. Tikėtinumų santykio kriterijaus, naudojamo šiai nulinei hipotezei tikrinti, statistika lygi:

$$\gamma^{(n)} = -2(l_0 - l);$$

čia $l = \ln L(\boldsymbol{x}, \hat{\theta}_s, \hat{\theta}_{k-s})$; $l_0 = \ln L(\boldsymbol{x}, \theta_{s0}, \hat{\theta}^*_{k-s})$; čia $L(\boldsymbol{x}, \hat{\theta}_s, \hat{\theta}_{k-s})$ – tikėtinumo funkcija su parametrais, įvertintais didžiausio tikėtinumo metodu, o $L(\boldsymbol{x}, \theta_{s0}, \hat{\theta}^*_{k-s})$ – tikėtinumo funkcija, esant teisingai nulinei hipotezei, kai likę nežinomi $(k - s)$ parametrai įvertinti didžiausio tikėtinumo metodu. Tikėtinumų santykio kriterijaus atmetimo sritis $\{\boldsymbol{x}: \gamma^{(n)} > c_\alpha\}$ nėra didžiausia. Tačiau statistikos $\gamma^{(n)} = -2(l_0 - l)$ asimptotinis skirstinys yra χ^2 su $(k - s)$ laisvės laipsnių ir, esant ganėtinai dideliam n, H_0 atmetimo sritis yra:

$$\{\boldsymbol{x}: \gamma^{(n)} > \chi^2_{1-\alpha}(k-s)\};$$

čia $\chi^2_{1-\alpha}(k-s)$ – χ^2 skirstinio su $(k - s)$ laisvės laipsnių $1 - \alpha$ lygio kvantilis, α – parinktas reikšmingumo lygmuo.

Nors ir neduodantis didžiausios atmetimo srities, tikėtinumų santykio kriterijus plačiai taikomas hipotezėms apie nežinomų parametrų reikšmes tikrinti logistinėje regresijoje (11 skyrius), išgyvenamumo analizėje (13 skyrius) ir daugiamatėje analizėje (15–16 skyrius).

5 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Bagdonavičius V., Kruopis J. *Matematinė statistika*. I dalis. 2007. Vilnius, 359 p.
3. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
4. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
5. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.
6. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
7. Кокс Д., Хинкли Д. *Теоретическая статистика*. 1978. Москва: Мир, 560 с.
8. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslas, 416 p.
9. Леман Э. *Проверка статистических гипотез*. 1979. Москва: Наука, 408 с.
10. Miller J. C., Miller J. N. *Statistics for Analycal Chemistry*. Second ed. 1988. New York: John Wiley & Sons, p. 227.
10. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
12. Холлендер М., Вулф Д. А. *Непараметрические методы статистики*. 1983. Москва: Наука, 516 с.
13. Watts S., Halliwell L. *Essential Environmental Science. Methods and Techniques*. 1996, p. 512.

6 SKYRIUS

Statistiniai kriterijai hipotezėms tikrinti ir jų taikymas

Šiame skyriuje pateiksime kriterijus, skirtus tikrinti hipotezes apie imties vidurkio ar medianos lygybę skaičiui (normai), dviejų populiacijų vidurkių lygybę, dviejų kartotinių matavimų vidurkių lygybę, imties skirstinį, dviejų skirstinių tapatumą; kelių imčių centro charakteristikų tapatumą.

6.1. Ženklų kriterijus

Tai kriterijus, skirtas tikrinti hipotezę apie Bernulio skirstinio (2.6 skyrius) parametro π lygybę 0,5. Šis kriterijus naudojamas kelioms neparametrinėms hipotezėms – apie simetriško skirstinio medianą ir apie dvimačio skirstinio simetriškumą – tikrinti.

Turime dvinario kintamojo, įgyjančio 1 arba 0 reikšmes, n dydžio imtį. Šios imties nariai x_i yra Bernulio atsitiktiniai dydžiai, įgyjantys reikšmes 1 ir 0 su tikimybėmis π ir 1 – π. Tikrinsime nulinę hipotezę H_0: „1 ir 0 imtyje yra vienodai tikėtini“ ($\pi = 0{,}5$).

Sakykime, imtyje yra m vienetukų. Tikrinant H_0 su alternatyva H_1: $\pi < 0{,}5$, nulinė hipotezė atmetama, kai m patenka į atmetimo sritį $m \le m1$; čia $m1$ – didžiausias neneigiamas skaičius, tenkinantis nelygybę:

$$\sum_{k=0}^{m1} C_n^k (0{,}5)^n \le \alpha; \qquad (6.1)$$

α – parinktas reikšmingumo lygmuo. Iš (6.1) priklausomybės matome, kad $m1$ priklauso tik nuo n ir α, pagal apibrėžimą, $(m1 + 1)$ yra binominio skirstinio su parametrais $(n, 0{,}5)$ α lygio kvantilis, žymimas $b(n, \alpha)$. H_0 atmetame ir priimame alternatyvą H_1, jei $m < b(n, \alpha)$; jei $m \ge b(n, \alpha)$, H_0 neprieštaraujame.

Tikrindami H_0 su alternatyva H_2: $\pi > 0{,}5$, H_0 atmesime, kai $m \geq m2$; čia $m2$ – didžiausias teigiamas skaičius, teikinantis sąlygą

$$\sum_{k=m2}^{n} C_n^k (0{,}5)^n \geq \alpha.$$

Pagal apibrėžimą, $m2 - 1 = b(n, 1 - \alpha)$, todėl H_0 atmesime, kai $m > b(n, 1 - \alpha)$. Binominis skirstinys yra simetriškas vidurkio $n\pi$, arba taško $n/2$, atžvilgiu, taigi binominio skirstinio kvantiliams teisinga $b(n, \alpha) + b(n, 1 - \alpha) = n$.

Tikrindami H_0 su dvipuse alternatyva H_3: $\pi \neq 0{,}5$, H_0 atmetame, jei

$m > b(n, 1 - \alpha/2)$ arba $m < b(n, \alpha/2)$.

Binominio skirstinio su $\pi = 0{,}5$ kvantiliai pateikti 6 lentelėje.

6.2. Kriterijai apie populiacijos vidurkio lygybę skaičiui (normai)

t kriterijus vienai imčiai. Prielaida: nagrinėjamas kintamasis yra kiekybinis ir turi normalųjį skirstinį su nežinomais parametrais – vidurkiu ir dispersija. Šis kriterijus aprašytas 5.4 skyriuje. Pateiksime šio kriterijaus taikymo pavyzdį.

6.1 pavyzdys. 3.5 lentelėje pateiktos 26 jaunų sveikų suaugusių asmenų SAS ir DAS reikšmės (mmHg). Šios SAS imties vidurkis lygus 113,1; jis mažesnis nei 120. Nustatysime, ar ir populiacijos vidurkis m yra mažesnis už 120.

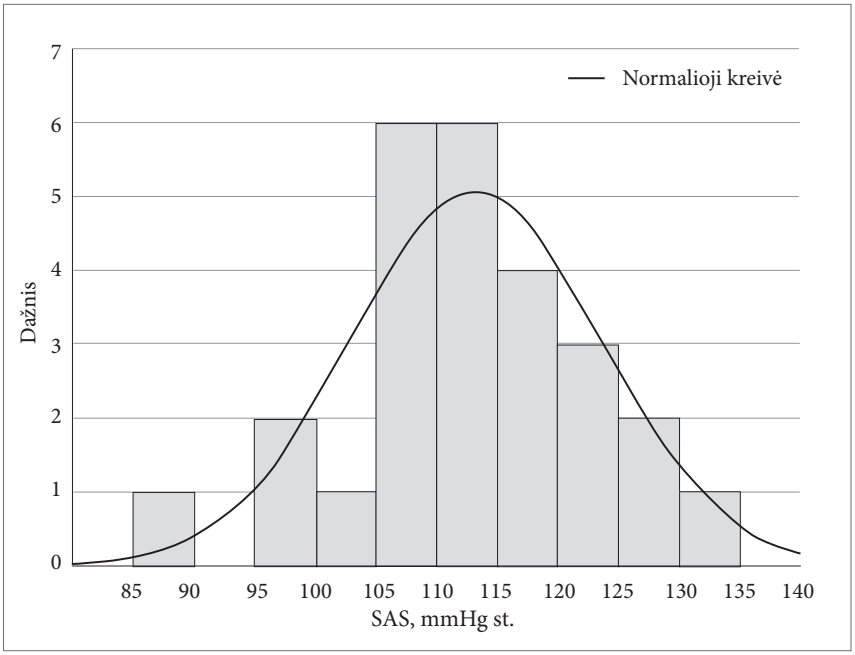
SAS histograma (6.1 pav.) leidžia daryti prielaidą, kad SAS skirstinys yra normalusis. Todėl hipotezei apie SAS populiacijos vidurkio m lygybę skaičiui tikrinti naudosime t kriterijų.

Tikrinsime nulinę hipotezę H_0: $m = 120$ (mmHg) su alternatyva H_2: $m < 120$ bei reikšmingumo lygmeniu $\alpha = 0{,}05$. Remiantis SAS imties duomenimis (3.4 skyrius):

imties dydis	$n = 26$;
vidurkis	$\bar{x} = 113{,}1$ (mmHg);
standartinis nuokrypis	$s = 10{,}3$ (mmHg);
norma	$m_0 = 120$ (mmHg);
reikšmingumo lygmuo	$\alpha = 0{,}05$.

Skaičiuojame kriterijaus statistikos reikšmę: $t = \sqrt{26}\,(113{,}1 - 120)/10{,}3 = 5{,}1 \times (-6{,}9)/10{,}3 = -3{,}84$. t kriterijaus dvipusė p reikšmė lygi 0,000746, o vienpusė – 0,000373, aišku, mažesnė už parinktą α. t skirstinio su 25 laisvės laipsniais $1 - 0{,}05 = 0{,}95$ lygio kvantilis $t_{0{,}95}(25)$ lygus 1,708. Taigi $-3{,}84 < -1{,}708$. Todėl nulinę hipotezę atmetame ir priimame alternatyvą:

jaunų sveikų suaugusių asmenų SAS populiacijos vidurkis mažesnis nei 120 mmHg (sakoma: SAS vidurkis reikšmingai mažesnis už 120 mmHg).



6.1 pav. 26 jaunų sveikų suaugusių asmenų SAS histograma ir normaliojo skirstinio kreivė

Tikrinsime kitą nulinę hipotezę H_0: $m = 110$ (mmHg) su alternatyva H_3: $m \neq 110$ bei reikšmingumo lygmeniu $\alpha = 0{,}01$. Kriterijaus statistikos reikšmė yra $t = \sqrt{26}\,(113{,}1 - 110)/10{,}3 = 5{,}1 \times 3{,}1)/10{,}3 = 1{,}535$. t kriterijaus dvipusė p reikšmė lygi 0,137 (didesnė už parinktą α), t skirstinio su 25 laisvės laipsniais $1 - 0{,}01/2 = 0{,}995$ lygio kvantilis $t_{0{,}995}(25)$ lygus 3,078. Taigi $1{,}535 < 3{,}078$, ir remiantis šio tyrimo duomenimis, nulinei hipotezei – „jaunų sveikų suaugusių asmenų SAS vidurkis lygus 110“ – prieštarauti nėra pagrindo.

Normalumo prielaidą galima sušvelninti, reikalaujant iš generuojančio imtį atsitiktinio dydžio skirstinio tik simetriškumo vidurkio atžvilgiu. Tuomet populiacijos vidurkis ir mediana bus vienodi; hipotezę apie vidurkio lygybę skaičiui galima pakeisti hipoteze apie medianos lygybę skaičiui. Šiai hipotezei (apie medianos lygybę skaičiui) tikrinti gali būti naudojamas ženklų kriterijus (6.1 skyrius) arba Vilkoksono kriterijus vienai imčiai. Kriterijų taikymo prielaida – kintamojo skirstinys yra tolydusis ir simetriškas vidurkio atžvilgiu.

Tarkime, $x_1, x_2 \ldots x_n$ – tolydųjį simetrišką vidurkio atžvilgiu skirstinį turinčio kintamojo imtis. Tikrinsime nulinę hipotezę H_0: $x_{med} = m_0$ su viena iš alternatyvų H_1: $x_{med} > m_0$; H_2: $x_{med} < m_0$; H_3: $x_{med} \neq m_0$.

Ženklų kriterijus. Skaičiuojame skirtumus $x_i - m_0$, $i = 1, 2 \ldots n$. Kadangi x_i skirstinys yra tolydusis ir simetriškas vidurkio atžvilgiu, todėl esant teisingai nulinei hipotezei dydžiai $x_i - m_0$ yra teigiami arba neigiami su vienodomis tikimybėmis, lygiomis 0,5: $P\{x_i - m_0 > 0\} = P\{x_i - m_0 < 0\} = 0{,}5$; ($P\{x_i = m_0\} = 0$, nes x_i skirstinys yra tolydusis). Sudarome naujus atsitiktinius dydžius y_i: $y_i = 1$, jei $x_i - m_0 > 0$, ir $y_i = 0$, jei $x_i - m_0 < 0$. Reikšmių $x_i = m_0$ į skaičiavimus netraukiame. Bernulio atsitiktinių dydžių $y_1, y_2 \ldots y_n$ sekai taikome ženklų kriterijų, aprašytą 6.1 skyriuje. Jei ženklų kriterijumi atmesime hipotezę $P\{y_i = 1\} = 0{,}5$ ir priimsime vieną iš alternatyvų (H_1: $P\{y_i = 1\} > 0$, H_2: $P\{y_i = 1\} < 0$, H_3: $P\{y_i = 1\} \neq 0$), tuomet atmesime nulinę hipotezę $x_{med} = m_0$ ir priimsime atitinkamą alternatyvą.

6.2 pavyzdys ([5, 139 p.]. Farmakologiniame preparate esančios komponentės A koncentracijos skirstinys yra simetriškas vidurkio atžvilgiu. Ar esant 5 % reikšmingumo lygmeniui galima tvirtinti, jog komponentės A koncentracija farmakologiniame preparate yra 8 %, jei atskirose partijose ji buvo atitinkamai 7,3; 7,1; 7,9; 9,1; 8,0; 7,1; 6,8 ir 7,3 %.

Turime $n = 8$, $m_0 = 8$. Dydžių $x_i - m_0$ reikšmės yra: –0,7; –0,9; –0,1; 1,1; 0; –0,9; –1,2; – 0,7. Atitinkamai nustatome ir 7 y_i reikšmes (išmetę reikšmę $x_i = 8$): 0 0 0 1 0 0 0. Tikrinsime H_0 su H_1 alternatyva. Mūsų atveju $m = 1$, $n = 7$, $\alpha = 0{,}05$. 6 lentelėje randame $b(7; 0{,}05) = 1$. Kadangi m nėra mažesnis už $b(7; 0{,}05)$ ($m = b(7; 0{,}05)$, todėl nulinei hipotezei prieštarauti negalime.

Vilkoksono kriterijus vienai imčiai. Pažymėkime $z_i = x_i - m_0$. Toliau Vilkoksono kriterijaus statistika skaičiuojama taip pat, kaip ir kartotinių imčių atveju (5.6 skyrius). Išvados gaunamos analogiškai.

6.3 pavyzdys ([5, 144 p.]). 7 vaikų kraujyje švino koncentracija nustatyta atitinkamai 104; 79; 98; 150; 87 ir 101 (pg/ml). Švino koncentracijos skirstinys laikomas simetrišku vidurkio atžvilgiu. Ar galime tvirtinti, kad tai duomenys iš populiacijos, kurios mediana (vidurkis) lygi 95 pg/ml?

Tikrinsime nulinę hipotezę H_0: $x_{med} = 95$ su dvipuse alternatyva H_3: $x_{med} \neq m_0$. Reikšmingumo lygmuo $\alpha = 0{,}05$, $n = 7$. Skaičiavimai, kurių reikia Vilkoksono statistikai T^+ nustatyti, pateikti 6.1 lentelėje.

Iš 6.1 lentelės gauname $T^+ = \sum_{i=1}^{n} R_i \varphi_i = 20$. T^+ gali svyruoti nuo 0 iki $7 \times 8/2 = 28$. Nulinę hipotezę atmesime, jei statistikai T^+ galios nelygybė $T^+ > t(1 - \alpha, n)$. T^+ skirstinio, esant teisingai nulinei hipotezei, 0,95 lygio kvan-

tilis $t(0{,}95, 7)$ lygus 25 (5 lentelė). Kadangi $T^+ < 25$, todėl nulinei hipotezei neprieštaraujame: galima tvirtinti, kad tai duomenys iš populiacijos, kurios mediana lygi 95 pg/ml.

6.1 lentelė. Vilkoksono kriterijaus vienai imčiai statistikos skaičiavimas

i	x_i	$z_i = x_i - m_0$	$\lvert z_i \rvert$	R_i	φ_i	$R_i \varphi_i$
1	104	9	9	4	1	4
2	79	–16	16	5	0	0
3	98	3	3	1	1	1
4	150	55	55	7	1	7
5	87	–8	8	3	0	0
6	126	41	41	6	1	6
7	101	6	6	2	1	2

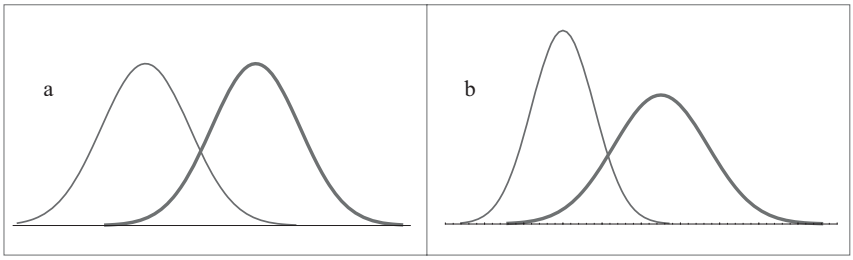
6.3. Dviejų populiacijų vidurkių palyginimas

Analizuojant susirgimo ar patologijos ypatumus, tenka lyginti ligonių, turinčių patologiją, tyrimų duomenis su atitinkamais kontrolinės grupės duomenimis. Laikant, kad tiriamo kintamojo skirstiniai abiejose populiacijose (su patologija ar be patologijos) yra normalieji, populiacijų lyginimas tampa atitinkamo kintamojo vidurkių lyginimu (I hipotezė 5.1 skyriuje).

Nepriklausomų imčių t kriterijus (*t-test for independent samples*). Prielaida: kintamojo iš X populiacijos (pvz., sergančių ligonių) ir iš Y populiacijos (pvz., sveikų asmenų) skirstiniai yra normalieji.

Sakykime, kintamojo iš X populiacijos skirstinys yra normalusis su vidurkiu m_x ir dispersija σ_x^2; iš Y populiacijos – normalusis su vidurkiu m_y ir dispersija σ_y^2. Uždavinys – palyginti dviejų populiacijų X ir Y kintamojo vidurkius m_x ir m_y. Todėl formuluojama nulinė hipotezė H_0: $m_x = m_y$ ir pagal imčių vidurkių reikšmes pasirenkama dešiniapusė, kairiapusė ar dvipusė alternatyva: H_1: $m_x > m_y$; H_2: $m_x < m_y$; H_3: $m_x \neq m_y$. Pavyzdžiui, I hipotezei (5.1 skyrius) tikrinti formuluojama nulinė hipotezė: $m_{CD} = m_{SV}$ su dešiniapuse alternatyva: $m_{CD} > m_{SV}$.

Nulinei hipotezei tikrinti naudojamos imtys iš X ir Y populiacijų. Jau minėjome, kad kintamojo skirstinys X ir Y populiacijose yra normalusis su vidurkiu m_x ir m_y. t kriterijaus statistika priklauso nuo to, ar populiacijų normalieji skirstiniai skiriasi tik poslinkiu (6.2 a pav.), ar dar ir forma (6.2 b pav.). Kitaip tariant, kriterijaus statistika priklauso nuo to, ar kintamojo dispersijos abiejose populiacijose vienodos, ar ne.



6.2 pav. Skirstinių dispersijos: a – vienodos; b – skirtingos

a) Kintamojo skirstinio dispersijos populiacijose lygios: $x_i \sim N(m_x, \sigma^2)$, $y_j \sim N(m_y, \sigma^2)$, $i = 1, 2 \ldots n$; $j = 1, 2 \ldots m$; čia $x_1, x_2 \ldots x_n$ – kintamojo iš X populiacijos imtis, o $y_1, y_2 \ldots y_m$ – kintamojo iš Y populiacijos imtis, n ir m – imčių dydžiai. Tuomet hipotezei apie vidurkių lygybę tikrinti naudojamas t kriterijus su (5.12) tipo statistika:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_p^2 (1/n + 1/m)}}; \tag{6.2}$$

čia $\overline{x}$ ir $\overline{y}$ – imčių vidurkiai, s_p^2 – sujungta dispersija: $s_p^2 = [(n-1)s_x^2 + (m-1)s_y^2]/(n+m-2)$, s_x^2 ir s_y^2 – imčių dispersijos. Jei nulinė hipotezė yra teisinga, statistika t turi Stjudento skirstinį su $(n + m - 2)$ laisvės laipsnių. Nulinės hipotezės atmetimo sritis bei sprendinio priėmimas, remiantis t kriterijaus dvipuse p reikšme, priklausomai nuo alternatyvos, pateikta 6.2 lentelėje.

b) Populiacijų dispersijos nelygios: $x_i \sim N(m_x, \sigma_x^2)$, $y_i \sim N(m_y, \sigma_y^2)$. Tuomet hipotezei apie vidurkių lygybę tikrinti taip pat naudojama (5.12) tipo kriterijaus statistika

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2 / n + s_y^2 / m)}}; \tag{6.3}$$

čia $\overline{x}$ ir $\overline{y}$ – imčių vidurkiai, s_x^2 ir s_y^2 – imčių dispersijos, n ir m – imčių dydžiai. Jei nulinė hipotezė teisinga, t skirstinys yra Stjudento su k laisvės laispnių, k – mažiausias sveikas skaičius, tenkinantis nelygybę: $k \le (s_x^2 / n + s_y^2 / m)^2 / (s_x^4 / n^3 + s_y^4 / m^3)$. Nulinės hipotezės atmetimo sritis bei sprendinio priėmimas, remiantis šio t kriterijaus dvipuse p reikšme, yra analogiškas vienodų dispersijų atveju ir pateiktas 6.2 lentelėje (tik joje vietoj $(n + m - 2)$ laisvės laipsnių skaičiaus reikia imti k).

6.2 lentelė. H_0: $m_x = m_y$ atmetimo sritis, sprendinio priėmimo taisyklė pagal kriterijaus dvipusę reikšmę

Alternatyva	Atmetimo sritis	H_0 atmetimo taisyklė pagal kriterijaus dvipusę p reikšmę
H_1: $m_x > m_y$	$t > {}_{1-\alpha}(n + m - 2)$	$\bar{x} > \bar{y}$, $p/2 < \alpha$
H_2: $m_x < m_y$	$t < -t_{1-\alpha}(n + m - 2)$	$\bar{x} < \bar{y}$, $p/2 < \alpha$
H_3: $m_x \neq m_y$	$\lvert t \rvert > t_{1-\alpha/2}(n + m - 2)$	$p < \alpha$

Kaip minėta, kriterijaus statistika, naudojama dviejų populiacijų vidurkiams palyginti, priklauso nuo populiacijų dispersijų: jei dispersijos lygios, naudojame (6.2), jei nelygios – (6.3) statistiką. Todėl, prieš lyginant dviejų populiacijų vidurkius, būtina patikrinti hipotezę apie populiacijų dispersijų σ_x^2 ir σ_y^2 lygybę.

Dviejų populiacijų dispersijų palyginimas. Prielaida: kintamojo skirstinys populiacijose yra normalusis.

Tikriname nulinę hipotezę H_0: $\sigma_x^2 = \sigma_y^2$. Apsiribosime dvipuse alternatyva: $\sigma_x^2 \neq \sigma_y^2$. Sakykime, $x_1, x_2 \ldots x_n$ ir $y_1, y_2 \ldots y_n$ – imtys iš tiriamų populiacijų, s_x^2 ir s_y^2 – imčių dispersijos, n ir m – imčių dydžiai. H_0 tikrinti naudojamas F kriterijus. Šio kriterijaus statistika lygi:

$$F = \frac{\text{didesnioji imties dispersija}}{\text{mažesnioji imties dispersija}}. \qquad (6.4)$$

Sakykime, $s_x^2 > s_y^2$. Tuomet $F = s_x^2 / s_y^2$. Esant teisingai nulinei hipotezei, statistika F turi Fišerio skirstinį su $(n - 1)$ ir $(m - 1)$ laisvės laipsnių (2.6 skyrius). Hipotezę apie dispersijų lygybę atmetame, jei $F > F_{1-\alpha}(n - 1, m - 1)$; čia $F_{1-\alpha}(n - 1, m - 1)$ – F skirstinio su $(n - 1)$ ir $(m - 1)$ laisvės laipsnių $1 - \alpha$ lygio kvantilis, α – reikšmingumo lygmuo. Priešingu atveju H_0 neprieštaraujame. Statistiniuose paketuose pateikiama šio kriterijaus vienpusė p reikšmė. Jei $p < \alpha$, teigiama, kad skirstinių dispersijos nelygios, o jei $p \geq \alpha$ – dispersijų lygybei neprieštaraujame.

6.4 pavyzdys. Dviejų populiacijų vidurkių palyginimas naudojant nepriklausomų imčių t kriterijų ([4, 132 p.]). Šį kriterijų taikysime 14 vyrų ir 12 moterų SAS vidurkiams palyginti. Remdamiesi 3.4 lentelėje pateiktais SAS duomenimis, tikrinsime nulinę hipotezę, kad jaunų sveikų suaugusių vyrų ir moterų SAS vidurkiai m_v ir m_m vienodi su vienpuse alternatyva: vyrų SAS vidurkis didesnis negu moterų ($m_v > m_m$). Pradiniai duomenys ir t statistikos skaičiavimai pateikti 6.3 lentelėje.

6.3 lentelė. t statistikos, naudotos 14 vyrų ir 12 moterų SAS vidurkiams palyginti, skaičiavimas

1. Pradiniai duomenys (3.4 lentelė)

imties dydis	$n = 14$ (vyrai)
	$m = 12$ (moterys)
vidurkis	$\bar{x} = 118{,}29$ mmHg (vyrų)
	$\bar{y} = 107{,}0$ mmHg (moterų)
dispersija	$s_x^2 = 70{,}07$ (vyrų)
	$s_y^2 = 82{,}55$ (moterų)
reikšmingumo lygmuo	$\alpha = 0{,}05$

2. t statistikos skaičiavimas, kai skirstinio dispersijos populiacijose yra lygios

$$s_p^2 = [(n-1)s_x^2 + (m-1)s_y^2]/(n+m-2) = (13 \times 70{,}07 + 11 \times 82{,}55)/24 = 75{,}79$$

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2(1/n + 1/m)}} = (118{,}29 - 107)/\sqrt{75{,}79(1/14 + 1/12)} = 11{,}29/3{,}425 = 3{,}3$$

laisvės laipsnių skaičius: $df = n + m - 2 = 14 + 12 - 2 = 24$; $t_{0{,}975}(24) = 2{,}064$
t kriterijaus dvipusė p reikšmė: $p = 0{,}003$

3. t kriterijaus skaičiavimas, kai skirstinio dispersijos populiacijose nelygios

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m)}} = (118{,}29 - 107)/\sqrt{75{,}79/14 + 82{,}5/12)} = 11{,}29\sqrt{5{,}01 + 6{,}88}/$$

$$= 11{,}29/\sqrt{11{,}89} = 11{,}29/3{,}45 = 3{,}28$$

laisvės laipsnių skaičius: $df = 22{,}67$; $t_{0{,}975}(22) = 2{,}069$
t kriterijaus dvipusė p reikšmė: $p = 0{,}003$

4. F kriterijaus skaičiavimas dispersijoms palyginti

$$F = s_y^2 / s_x^2 = 82{,}55/70{,}07 = 1{,}178$$

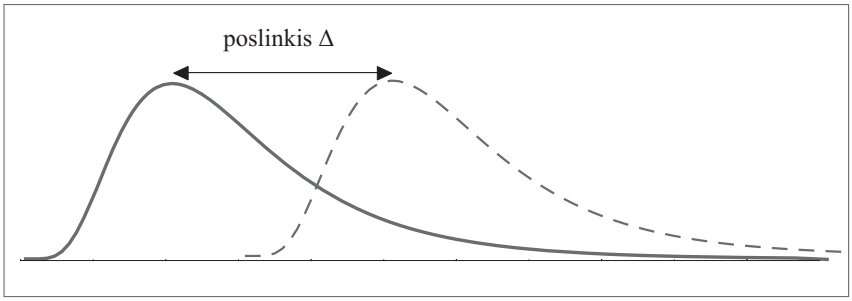
laisvės laipsnių skaičius: $n - 1 = 14 - 1 = 13$; $m - 1 = 12 - 1 = 11$
F kriterijaus p reikšmė: $p > 0{,}7$

Statistikos *t* formulė nulinei hipotezei tikrinti priklauso nuo to, ar SAS vyrų ir moterų dispersijos yra lygios, ar ne. Dispersijų lygybei tikrinti skaičiuojame F kriterijaus statistikos reikšmę. Ji lygi 1,178; kriterijaus *p* reikšmė didesnė nei 0,7 (*F* skirstinio su 11 ($m - 1$) ir 13 ($n - 1$) laisvės laipsnių $1 - 0{,}05/2 = 0{,}975$ lygio kvantilis $F_{0{,}975}(11, 13)$ lygus 3,2; $1{,}178 < 3{,}2$). Todėl galime tvirtinti, kad vyrų ir moterų SAS populiacijos dispersijos yra lygios. 6.3 lentelėje matome, kad statistikos *t* reikšmė lygi 3,3, o dvipusė p reikšmė mažesnė nei 0,003. Todėl nulinę hipotezę atmetame ir priimame alternatyvą: vyrų SAS vidurkis reikšmingai didesnis nei moterų.

Kaip minėta, t kriterijus dviejų populiacijų vidurkiams palyginti taikomas tuomet, kai kintamojo skirstiniai populiacijose yra normalieji. Tačiau normalumo prielaidą sunku patikrinti nedidelių imčių atveju; be to, kartais

kintamojo skirstiniai populiacijoje būna asimetriški (6.3 pav.), imtyse pasitaiko išskirčių. Pateiksime kriterijų, reikalaujantį silpnesnių prielaidų nei normalumas.

U kriterijus dviejų populiacijų vidurkiams palyginti (*Mann-Whitney U test*). Prielaidos: kintamojo reikšmės (imtys) iš X populiacijos (pvz., sergančiųjų) $x_1, x_2 \ldots x_n$ ir iš Y populiacijos (pvz., sveikų asmenų) $y_1, y_2 \ldots y_m$ aprašomos pagal modelį: $x_i = e_i$, $i = 1 \ldots n$; $y_j = e_{n+j} + \Delta$, $j = 1 \ldots m$. Čia $e_1, e_2 \ldots e_n$, $e_{n+1} \ldots e_{n+m}$ – nepriklausomi nestebimi ats. d., turintys tą patį tolydųjį skirstinį, Δ – nežinomas poslinkis, sukeltas veiksnio (pvz., susirgimo), skiriančio populiacijas. Kitaip tariant, skirstiniai populiacijose yra tolydieji ir skiriasi tik poslinkiu (6.3 pav.). Dėl perėjimo iš vienos populiacijos į kitą atsiradusiam poslinkiui konstatuoti naudojamas U kriterijus.



6.3 pav. Kintamojo skirstiniai populiacijose

Tikrinsime nulinę hipotezę H_0: $\Delta = 0$ su viena iš alternatyvų: $\Delta > 0$, $\Delta < 0$ ir $\Delta \neq 0$. H_0 tikrinti naudojamo U kriterijaus statistika skaičiuojama naudojant ne imčių reikšmes, o rangus. Abiejų imčių kartu sudėtos reikšmės suranguojamos. R_i pažymėkime y_i rangą bendroje sekoje. Kadangi kintamojo skirstinys yra tolydusis, abiejose imtyse mažai tikėtinos tos pačios reikšmės, arba vienodi rangai. U kriterijaus statistika W lygi imties iš Y populiacijos rangų sumai:

$$W = \sum_{j=1}^{m} R_j,\ U = m \times n + m(m+1)/2 - W. \qquad (6.4)$$

W įgyja reikšmes nuo 0 iki $m(n + m + 1)$. Esant teisingai nulinei hipotezei, rangų sumos W skirstinys yra diskretusis, simetriškas vidurkio atžvilgiu ir priklauso tik nuo n ir m. Nedideliems n ir m lentelėse pateikiami statistikos U skirstinio atitinkamo lygio kvantiliai $w(1 - \alpha, n, m)$ arba $w(\alpha, n, m)$; čia α – parinktas reikšmingumo lygmuo (7 lentelė). Pažymėsime, kad $w(1 - \alpha, n, m) + w(\alpha, n, m) = m \times n$, nes U įgyja reikšmes nuo 0 iki $m \times n$.

H_0 tikrinimas nedideliam n ir m ($n, m \leq 30$):

- alternatyva $\Delta > 0$: H_0 atmetame, jei $U > w(1 - \alpha, n, m)$, H_0 neprieštaraujame, jei $U \leq w(1 - \alpha, n, m)$;
- alternatyva $\Delta < 0$: H_0 atmetame, jei $U < w(\alpha, n, m)$, H_0 neprieštaraujame, jei $U \geq w(\alpha, n, m)$;
- alternatyva $\Delta \neq 0$: H_0 atmetame, jei arba $U > w(1 - \alpha/2, n, m)$, arba $U < w, (\alpha/2, n, m)$, H_0 neprieštaraujame, jei $w(\alpha/2, n, m) \leq U \leq w(1 - \alpha/2, n, m)$.

Vietoj statistikos U nulinei hipotezei tikrinti galima naudoti analogišką statistiką, skaičiuotą naudojant imties $x_1, x_2 \ldots x_n$ rangus. H_0 tikrinti dideliems n ir m ($n > 30$, $m > 30$) naudojama (5.12) tipo statistika:

$$W^* = (W - EW) / se(W) = (W - m(n + m + 1) / 2) / \sqrt{mn(n + m + 1) / 12}, \qquad (6.5)$$

turinti asimptotinį standartinį normalųjį skirstinį. Taikant U kriterijų, statistiniu paketu pateikiamos kriterijaus statistikos bei tikslaus ir asimptotinio kriterijaus dvipusė p reikšmė. H_0 atmetimo ar priėmimo taisyklė, priklausomai nuo alternatyvos, yra analogiška Vilkoksono kriterijaus atveju.

Pateiksime U kriterijaus taikymo pavyzdį.

6.5 pavyzdys ([7, 88 p.]). Lyginta tričio vandens difuzija per placentą normaliai pagimdžiusių moterų ir moterų, nutraukusių nėštumą tarp 12 ir 24 savaitės. Medikus domino, ar pagimdžiusiųjų laiku placentos pralaidumas yra didesnis. Tyrimui naudoti 10 normaliai pagimdžiusių ir 5 nutraukusių nėštumą moterų placentos audinio mėginiai – konkrečios imtys $x_1, x_2 \ldots x_{10}$ ir $y_1, y_2 \ldots y_5$. Šie duomenys pateikti 6.4 lentelėje. Tikrinama nulinė hipotezė: laiku ir nelaiku pagimdžiusiųjų placentos pralaidumas vienodas ($\Delta = 0$) su vienpuse alternatyva: $\Delta < 0$. Nulinei hipotezei tikrinti naudosime U kriterijų su statistika (6.4). y_j rangai pateikti 6.4 lentelėje.

6.4 lentelė. Tričio vandens difuzija per placentą (10^{-4} cm/s)

Pagimdžiusių laiku (x_j)	j	Nutraukusių nėštumą (y_j)	R_j
0,8	1	1,15	8
0,83	2	0,88	5
1,89	3	0,9	6
1,04	4	0,74	2
1,45	5	1,21	9
1,38	6	–	–
1,91	7	–	–
1,64	8	–	–
0,73	9	–	–
1,46	10	–	–

Skaičiuojame: $W = R_1 + R_2 + R_3 + R_4 + R_5 = 30$, $U = 5\times10 + 5\times6/2 - 30 = 35$. Iš 7 lentelės matyti, kad $w(0{,}05, 10, 5) = 10$. Apskaičiuojame $w(0{,}95, 10, 5) = 50 - 11 = 39$. Kadangi $W = 35 < w(0{,}95, 10, 5) = 39$, todėl nulinei hipotezei su reikšmingumo lygmeniu 0,05 neprieštaraujame.

Didelės imties atveju naudotina statistika (6.14) lygi:

$W^* = (30 - 5\times16/2)/(5\times10\times16/12)1/2 = -1{,}225$.

Šio kriterijaus vienpusė *p* reikšmė lygi 0,11. Taigi tiek tikslus *U* kriterijus, tiek normalioji jo aproksimacija nurodo, kad remiantis pateiktais duomenimis nėra prielaidos tvirtinti, jog pagimdžiusiųjų laiku placentos pralaidumas yra didesnis, palyginti su nutraukusiųjų nėštumą tarp 12 ir 24 savaitės.

6.4. Dviejų kartotinių matavimų vidurkių palyginimas

Medikams ypač aktualu nustatyti veiksnio – vaistų ar terapijos – poveikį ligonio būklei. Dėl to ligonio būklę charakterizuojantys rodikliai nustatomi prieš veiksnį (vaistų naudojimą, terapijos seansą, operaciją) ir po jo: sudaromos kartotinės imtys – vienam individui atliekami keli matavimai. Medikus dažniausiai domina kiekybinio rodiklio, apibūdinančio ligonio būklę, vidurkio ar medianos pokytis. Nustačius, kad populiacijos vidurkis sumažėjo ar padidėjo, daroma atitinkama išvada apie veiksnio įtaką ligonio būklei. Kartotinių matavimų vidurkiams palyginti naudojamas kartotinių imčių t kriterijus ir Vilkoksono kriterijus.

Kartotinių imčių t kriterijus (*t-test for dependent samples*). Prielaida: tiriamo kintamojo, nustatyto prieš veiksnį ir po jo, skirstiniai yra normalieji. Todėl poveikio nustatymas tampa atitinkamo kintamojo vidurkių palyginimu (5.1 skyriaus V hipotezė).

Populiacijos vidurkių pokyčiui nustatyti tikrinsime nulinę hipotezę H_0: $m_x = m_y$, čia m_x ir m_y – skirstinių prieš veiksnį ir po jo vidurkiai su dešiniapuse ($m_x > m_y$), kairiapuse ($m_x < m_y$) ar dvipuse alternatyvomis. Sakykime, $x_1, x_2 \ldots x_n$ kintamojo reikšmės (imtis) prieš veiksnį, $y_1, y_2 \ldots y_n$ – po veiksnio išmatuotos tų pačių individų reikšmės. Nulinei hipotezei tikrinti skaičiuojama kriterijaus statistika:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_d^2 / n}}; \qquad (6.6)$$

čia $\overline{x}$ ir $\overline{y}$ – imčių prieš ir po veiksnio vidurkiai, s_d^2 – skirtumų $x_i - y_i$ dispersija. Esant teisingai nulinei hipotezei, statistika *t* turi Stjudento skirstinį su $(n - 1)$ laisvės laipsniu. Nulinės hipotezės atmetimo sritis bei sprendinio priėmimas, remiantis šio t kriterijaus dvipuse *p* reikšme, pateiktas 5.2 lentelėje.

Vilkoksono kriterijus. Prielaidos: kintamojo, nustatyto prieš veiksnį ir po jo, reikšmių skirtumo skirstinys yra simetriškas vidurkio θ atžvilgiu. Tuomet hipotezė apie veiksnio poveikį tapati hipotezei apie poslinkio θ lygybę 0 H_0: θ = 0 (5.5 skyrius). Vilkoksono kriterijus šiai hipotezei tikrinti pateiktas 5.7 skyriuje.

Pateikiame pavyzdį, kaip taikomas Vilkoksono kriterijus.

6.6 pavyzdys ([7, 49 p.]). Tirtas trankvilizatoriaus T poveikis neuroze sergantiems ligoniams. Paciento būklė vertinta IV Hamiltono depresijos skale (kuo mažesnis nustatytas koeficientas, tuo geresnė ligonio būklė). Pacientų būklė tirta gydymo pradžioje (x_i reikšmės) ir pabaigoje (y_i reikšmės). Šios IV Hamiltono skalės reikšmės bei teigiamo ženklo rangų skaičiavimai pateikti 6.5 lentelėje. Reikšmingumo lygmuo parenkamas standartinis – α = 0,05.

Iš 6.5 lentelės matome, kad Vilkoksono kriterijaus statistika $T^+ = \sum_{i=1}^{n} R_i \varphi_i =$ 3 + 2 = 5. 5 lentelėje randame kvantilį t(0,05, 9) = 8. Kadangi $T^+ = 5 < 8$, H_0 atmetame ir priimame alternatyvą: θ < 0.

6.5 lentelė. IV Hamiltono skalės reikšmės gydymo pradžioje bei pabaigoje ir rangų skaičiavimai

i	x_i	y_i	z_i	$\|z_i\|$	R_i	φ_i	$R_i\varphi_i$
1	1,83	0,878	-0,952	0,952	8	0	0
2	0,50	0,647	0,147	0,147	3	1	3
3	1,62	0,598	-1,022	1,022	9	0	0
4	2,48	2,05	-0,430	0,430	4	0	0
5	1,68	1,06	-0,620	0,620	7	0	0
6	1,88	1,29	-0,590	0,590	6	0	0
7	1,55	1,06	-0,490	0,490	5	0	0
8	3,06	3,14	0,080	0,080	2	1	2
9	1,3	1,29	-0,010	0,010	1	0	0

Statistika T^* (5.15), naudojama didelės imties atveju, yra $T^* = (5 - 10 \times 9/4)/(9 \times 10 \times 19/24)^{1/2} = -17{,}5/8{,}44 = -2{,}07$. Šio kriterijaus atitinkama p reikšmė lygi 0,0192 ir yra mažesnė už parinktą reikšmingumo lygmenį. Taigi ir tikslus Vilkoksono kriterijus bei jo aproksimacija didelės imties atveju leidžia daryti išvadą, kad trankvilizatoriaus T poveikis pagerino pacientų būklę, vertintą IV Hamiltono depresijos skalės koeficientu.

6.5. Hipotezės apie populiacijos vidurkį tikrinimas didelių imčių atveju

Šiame skyriuje daroma prielaida, kad nagrinėjamas rodiklis yra kiekybinis su unimodaliu skirstiniu. Tai reiškia, kad kintamojo reikšmės išsibarsčiusios apie tam tikrą sankaupos tašką – skirstinio „centrą“, dažniausiai vidurkį. Tokiu atveju yra prasmė tikrinti hipotezes apie šio „centro“ konkrečią reikšmę, poslinkį po tam tikro poveikio ar palyginti kelių imčių „centrus“. Daugiau prielaidų apie skirstinį nedaroma – nereikalaujama normalumo kaip 5.4 skyriuje. Tačiau nagrinėjama imtis (ar dvi imtys) turi būti gana didelė: būtinai $n > 30$, o apsidraudžiant nuo atsitiktinių svyravimų – $n > 100$, kad hipotezėms tikrinti galėtume taikyti asimptotinius kriterijus.

6.2–6.4 skyriuose hipotezėms apie vidurkio lygybę skaičiui, dviejų populiacijų ar dviejų kartotinių matavimų vidurkių lygybę tikrinti naudojome standartizuoto santykio statistiką (5.12) su funkcija $T(\boldsymbol{x})$, lygia $\overline{x}$ ar $\overline{x} - \overline{y}$. Jei kintamojo skirstinys yra normalusis, standartizuoto santykio statistikos atskirų atvejų (5.14), (6.2), (6.6) skirstiniai yra Stjudento su atitinkamu laisvės laipsnių skaičiumi. $T(\boldsymbol{x})$ šiuo atveju yra nepriklausomų ats. d. suma, taigi dėl centrinės ribinės teoremos standartizuoto santykio (5.14), (6.2), (6.6) asimptotinis skirstinys yra normalusis. Todėl esant didelėms imtims ($n > 100$, $m > 100$ arba bent jau $n > 30$, $m > 30$), galima laikyti, kad statistikų (5.14), (6.2), (6.6) skirstiniai yra normalieji su vidurkiu, lygiu nuliui, ir vienetine dispersija. Dėl to didelių imčių atveju nulinės hipotezės atmetimo sritys 5.2–5.3 lentelėse gali būti apibrėžiamos ne Stjudento, o standartinio normaliojo skirstinio atitinkamos eilės kvantiliais z_α, kadangi Stjudento skirstinio kvantilis $t_\alpha(n)$ dideliam n yra artimas z_α.

6.7 pavyzdys. Ligoniams, persirgusiems Q ir be Q miokardo infarktu, atliktas EKG tyrimas; jo metu nustatytos ir ŠSD reikšmės. Tyrimas atliktas 270 sergančių Q bangos MI ir 216 sergančių be Q bangos MI ligonių; taigi imtys gana didelės. Nustatyta, kad sergančių Q bangos MI ŠSD vidurkis – 78,5 (k./min.), be Q bangos MI – 73,7 (k./min.). Norėdami patvirtinti, kad „sergančiųjų Q bangos infarktu populiacijos ŠŠD vidurkis didesnis nei sergančiųjų be Q bangos MI populiacijos“, naudosime t kriterijų populiacijų vidurkiams palyginti. Sergančiųjų Q ir be Q bangos MI ligonių ŠSD standartiniai nuokrypiai yra atitinkamai 19,6 ir 16,3. Populiacijų dispersijos reikšmingai skiriasi ($\alpha = 0{,}05$), nes F kriterijaus statistika lygi 1,46, jos p reikšmė – 0,004. t kriterijaus statistikos, skaičiuotos pagal 5.10 formulę, reikšmė t lygi 2,93, normaliojo skirstinio kvantilis $z_{1-0{,}05} = z_{0{,}95} = 1{,}645$. Turime $t > z_{0{,}95}$, taigi nulinę hipotezę atmetame ir tvirtiname, kad, turimų duomenų pagrindu, Q bangos MI susirgusių ligonių SŠD reikšmingai didesnis nei susirgusių be Q bangos MI.

6.6. Hipotezės apie kintamojo skirstinį (suderinamumo hipotezės)

Kaip minėta, 6.2–6.4 skyriuose pateikto t kriterijaus (t kriterijaus vienai imčiai, nepriklausomų imčių t kriterijaus, kartotinių matavimų t kriterijaus) taikymo prielaida – skirstinių normalumas. Tai reiškia, kad prieš taikant t kriterijų, būtina patikrinti, ar normalumo prielaida yra teisinga. Šiame skyriuje pateiksime keletą kriterijų, skirtų hipotezėms apie kintamojo skirstinį. Tokios hipotezės vadinamos suderinamumo hipotezėmis.

Sakykime, tiriamo kintamojo imtis yra nepriklausomi atsitiktiniai dydžiai su tuo pačiu skirstiniu. Pažymėkime: F_0 – žinomas skirstinys, pavyzdžiui, standartinis normalusis, ar (5.1) formule apibrėžtas skirstinys. Formuluokime nulinę hipotezę H_0: „kintamojo skirstinys yra F_0“. Ši hipotezė vadinama paprasta suderinamumo hipoteze.

Sudėtinga suderinamumo hipoteze vadiname tokią hipotezę: „kintamojo skirstinys priklauso parametrinių skirstinių šeimai $P(\theta_1, \theta_2 \ldots \theta_m)$“; čia $P(\theta_1, \theta_2 \ldots \theta_m)$ – skirstinių, priklausančių nuo parametrų $\theta_1, \theta_2 \ldots \theta_m$, šeima. P skirstinio (tankio ar tikimybės) analizinė išraiška yra žinoma; tai gali būti normaliųjų skirstinių šeima – ji susideda iš visų galimų normaliųjų skirstinių, parametras θ_1 yra vidurkis m, o θ_2 – σ. P gali būti Puasono skirstinių šeima (susideda iš visų Puasono skirstinių), Bernulio skirstinių šeima.

Suderinamumo hipotezėms tikrinti naudojamas χ^2 kriterijus. Paprastai suderinamumo hipotezei tikrinti kiekybinio kintamojo atveju taikomas dar ir Kolmogorovo–Smirnovo kriterijus.

χ^2 kriterijus paprastai suderinamumo hipotezei tikrinti. Tikrinama H_0: „kintamojo skirstinys yra F_0“ su alternatyva H_a: „kintamojo skirstinys nėra lygus F_0“. Sakykime, $x_1, x_2 \ldots x_n$ – tiriamo kintamojo imtis. Imties reikšmių intervalą (nuo mažiausios reikšmės iki didžiausios) padalijame į k nesusikertančių intervalų $[c_{i-1}, c_i)$, $i = 1, 2 \ldots k$ taip, kad į kiekvieną intervalą patektų bent po 5 imties reikšmes. Pažymėkime O_i – imties reikšmių, patekusių į i-tąjį intervalą, skaičius; p_i – tikimybė, kad atsitiktinis dydis, sakykime, ξ, turintis skirstinį F_0, pateks į intervalą $[c_{i-1}, c_i)$: t. y. $p_i = P\{\xi < c_i\} - P\{\xi < c_{i-1}\}$, $i = 1, 2 \ldots k$. Dydis O_i vadinamas stebimu (*observed*) dažniu, E_i – tikėtinu (*expected*) dažniu, nes E_i rodo, kiek imties narių turėtų patekti į intervalą $[c_{i-1}, c_i)$, jei imtis būtų iš populiacijos su skirstiniu F_0.

Nulinės hipotezės ar alternatyvos pasirinkimas remiasi χ^2 kriterijaus statistika:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i . \qquad (6.7)$$

Jei teisinga nulinė hipotezė, t. y. jei analizuojamo kintamojo skirstinys yra F_0, tuomet ši statistika turi asimptotinį χ^2 skirstinį su $(k-1)$ laisvės laipsnių. Jei n ganėtinai didelis ($n > 30$) ir stebėti dažniai ne mažesni už 5, galime tvirtinti, kad statistikos (6.7) skirstinys artimas χ^2 skirstiniui su $(k-1)$ laisvės laipsnių. Todėl nulinę hipotezę atmetame, jei $\chi^2 > \chi^2_{1-\alpha}(k-1)$; čia $\chi^2_{1-\alpha}(k-1)$ – χ^2 skirstinio su $(k-1)$ laisvės laipsniu $1-\alpha$ lygio kvantilis arba jei χ^2 kriterijaus vienpusė p reiksmė neviršys parinkto reikšmingumo lygmens α. Priešingu atveju nulinei hipotezei neprieštaraujame.

χ^2 kriterijaus nerekomenduotina taikyti mažoms imtims. Jei tikėtini dažniai O_i mažesni už 5, χ^2 kriterijumi gautos išvados gali būti neteisingos, todėl grupavimo intervalų nereikia imti daug.

Pateiksime χ^2 kriterijaus taikymo paprastai suderinamumo hipotezei tikrinti pavyzdį.

6.8 pavyzdys ([7, 104 p.]). Pastebėta, kad žmogus dažniausiai stato dešinę koją, pakreipdamas pėdą į dešinę pusę. Šiam teiginiui pagrįsti buvo tirta 300 žmonių su tokia pėdos orientacija: pėda krypsta į kairę, į dešinę arba dedama tiesiai. Atlikus eksperimentą, paaiškėjo, kad 80 žmonių pėdą kreipia į kairę, 96 deda tiesiai, 124 – pasuka į dešinę. Tikrinant teiginį, ar pėdos orientacija į dešinę yra dažnesnė, tikrinama nulinė hipotezė H_0: „pėdos orientacija į kairę, į dešinę ir tiesiai vienodai dažna“ (tikimybė, kad žmogus kreips pėdą į dešinę, į kairę, statys tiesiai, lygi 1/3). Jeigu šios trys padėtys būtų vienodai dažnos, turėtų būti, jog apie 100 žmonių (iš 300) pėdą kreipia į kairę, 100 – į dešinę, 100 stato tiesiai. Tikrindami nulinę hipotezę, χ^2 kriterijaus statistika lyginame stebėtus dažnius O_i (80, 96, 124) su tikėtinais E_i (100, 100, 100). Apskaičiavę gauname: $\chi^2 = (80-100)^2/100 + (96-100)^2/100 + (124-100)^2/100 = 4 + 0{,}16 + 5{,}76 = 9{,}92$, laisvės laipsniai = 2, $p = 0{,}007$. Net su 99 % pasikliovimo lygmeniu konstatuojame, kad pėdos orientacija į dešinę yra dažnesnė.

***χ^2* kriterijus sudėtingai suderinamumo hipotezei tikrinti.** Tikrinama nulinė hipotezė H_0: „kintamojo skirstinys priklauso skirstinių šeimai $P(\theta_1, \theta_2 \ldots \theta_m)$“ su alternatyva H_a: „kintamojo skirstinys nepriklauso parametrinių skirstinių šeimai $P(\theta_1, \theta_2 \ldots \theta_m)$“; čia $\theta_1, \theta_2 \ldots \theta_m$ – nežinomi skirstinio parametrai. Analogiškai kintamojo reikšmių sritis padalijama į k nesusikertančių intervalų, $[c_{i-1}, c_i)$, $i = 1, 2 \ldots k$. Nežinomi skirstinio parametrai $\theta_1, \theta_2 \ldots \theta_m$ įvertinami didžiausio tikėtinumo metodu, o gauti įverčiai naudojami patekimo į intervalą $[c_{i-1}, c_i)$ tikimybei p_i skaičiuoti, čia $p_i = P\{\xi < c_i\} - P\{\xi < c_{i-1}\}$, $i = 1, 2 \ldots k$, ξ – ats. d. su konkrečiu skirstiniu $P(\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_m)$. Pažymėkime O_i – imties reikšmių patekusių į i-tąjį intervalą skaičius, $E_i = np_i$, $i = 1, 2 \ldots k$ – tikėtini dažniai. Nulinės hipotezės atmetimo sritis sudaroma remiantis kriterijaus statistika (6.7). Esant teisingai nulinei hipotezei,

statistika (6.7) turi asimptotinį χ^2 skirstinį su $(k - m - 1)$ laisvės laipsniu. Nulinę hipotezę atmetame, jei $\chi^2 > \chi^2_{1-\alpha}(k-m-1)$; čia $\chi^2_{1-\alpha}(k-m-1)$ – χ^2 skirstinio su $(k - m - 1)$ laisvės laipsniu $1 - \alpha$ lygio kvantilis, arba χ^2 kriterijaus p reikšmė neviršys parinkto reikšmingumo lygmens α. Priešingu atveju nulinei hipotezei neprieštaraujame.

χ^2 kriterijaus nerekomenduotina taikyti mažoms imtims, nes gautos išvados gali būti neteisingos. Taikant χ^2 kriterijų sudėtingai suderinamumo hipotezei tikrinti, grupavimo intervalų skaičius k turi viršyti $(m + 1)$ (m – nežinomų skirstinio parametrų skaičius), tikėtini dažniai O_i – viršyti 4.

Pateiksime χ^2 kriterijaus taikymo sudėtingai suderinamumo hipotezei tikrinti pavyzdį.

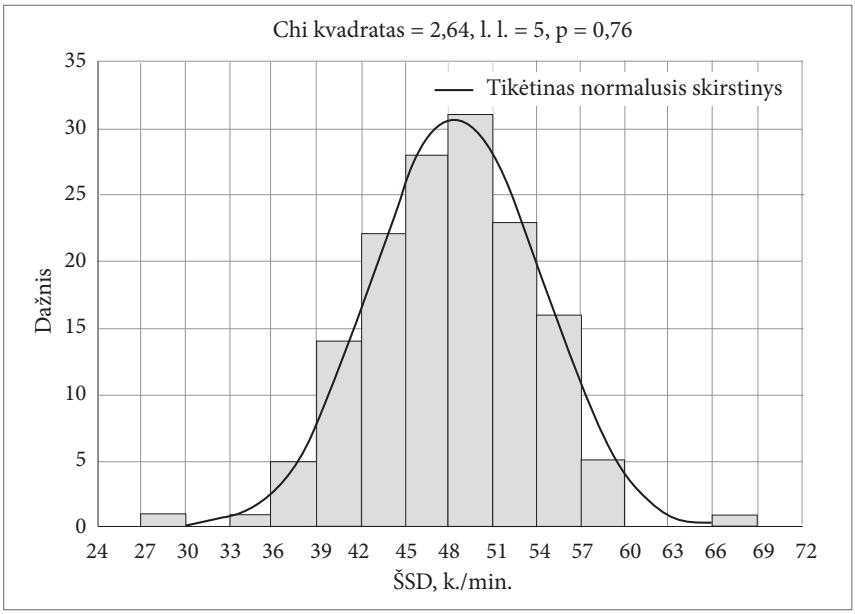
6.9 pavyzdys. Nagrinėsime ligonių, sergančių nestabilia KA, dešiniojo prieširdžio (DPR) dydžio skirstinį. Tikrinsime nulinę hipotezę H_0: „DPR skirstinys yra normalusis“, alternatyva – „DPR skirstinys nėra normalusis“. Hipotezei tikrinti naudosime 147 ligonių echoskopijos tyrimų duomenis.

DPR histograma pateikta 6.4 pav. DPR reikšmės svyravo nuo 30 iki 69, vidurkis ir dispersija lygūs: $\overline{x} = 48{,}56$, $s^2 = 32{,}86$; $s = 5{,}73$. DPR reikšmes sugrupavome į 8 intervalus ($k = 8$). Patekimo į intervalą tikimybes ir tikėtinus dažnius skaičiavome naudodami normaliojo skirstinio su $m = 48{,}56$ ir $\sigma = 5{,}73$ (tokie didžiausio tikėtinumo įverčiai) skirstinio funkcijos $\Phi(x, 48{,}56, 5{,}73)$ (2.3) reikšmes. Grupavimo intervalai, stebėti ir tikėtini dažniai pateikti 6.6 lentelėje.

6.6 lentelės duomenimis, gauname, kad χ^2 kriterijaus statistika (6.7) lygi 2,611, laisvės laipsnių skaičius yra 8 – 1 – 2 = 5, vienpusė p reikšmė lygi 0,76. Todėl nulinei hipotezei prieštarauti nėra pagrindo – DPR skirstinys yra normalusis.

6.6 lentelė. Ligonių, sergančių nestabilia KA, stebėti ir tikėtini dešiniojo prieširdžio dažniai

Grupavimo intervalai	O_i	p_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
[30, 42)	7	0,048	7.056	–0.056	0.000444
[42, 45)	14	0,079	11.466	2.534	0.560017
[45, 48)	22	0,141	20.727	1.273	0.078184
[48, 51)	28	0,194	28.518	–0.518	0.009409
[51, 54)	31	0,204	29.988	1.012	0.034152
[54, 57)	23	0,164	24.108	–1.108	0.050924
[57, 60)	16	0,101	14.847	1.153	0.089541
[60, 70)	6	0,070	10.29	–4.29	1.788542



6.4 pav. DPR dydžio histograma ir tikėtinas normalusis skirstinys

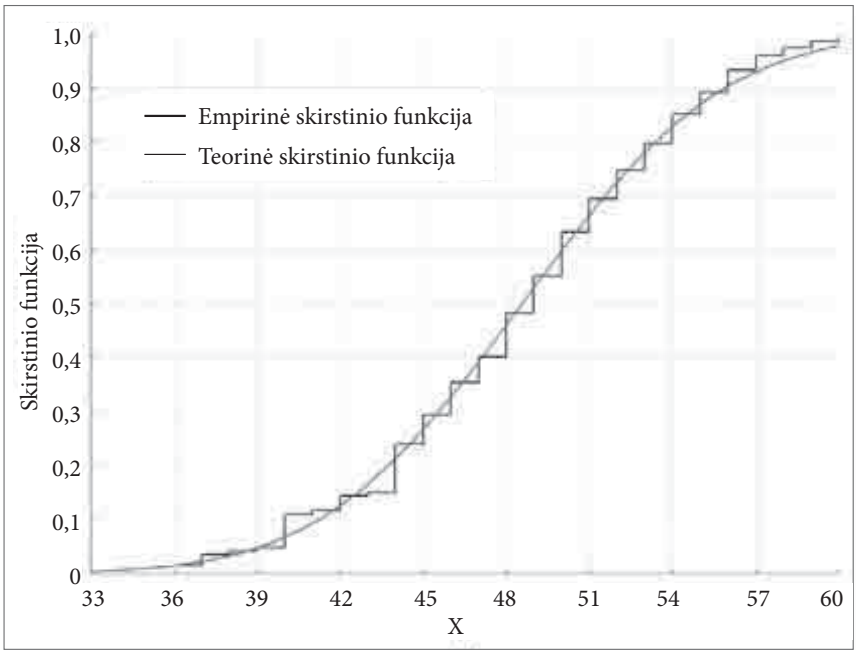
Kolmogorovo–Smirnovo kriterijus. Prielaida: kintamojo skirstinys yra tolydusis. Kriterijus naudojamas tikrinti paprastą suderinamumo hipotezę H_0: „kintamojo skirstinys sutampa su F_0“ su alternatyva H_a: „kintamojo skirstinys nesutampa su F_0“. Šis kriterijus yra jautresnis nukrypimams nuo duoto skirstinio variacinės sekos centre negu galuose.

Kolmogorovo–Smirnovo kriterijus grindžiamas empirinės ir populiacijos (teorinio) skirstinio funkcijų skirtumu. Imties x_1, x_2 ... x_n empirinė skirstinio funkcija $E_n(x)$ apibrėžiama $E_n(x) = n(x)/n$; čia $n(x)$ – imties reikšmių, ne didesnių už x, skaičius (6.5 pav.). Teorinė (populiacijos) skirstinio funkcija $F_0(x) = P\{\xi \le x\}$; čia ξ – ats. d. su skirstiniu F_0. Kolmogorovo–Smirnovo kriterijaus statistika lygi $D_n = \max|E_n(x) - F_0(x)|$ (6.4 pav.) arba

$$D_n = \max |F_0(x_{(i)}) - i/n|,\ i = 1, 2 \ldots n; \qquad (6.7)$$

čia $x_{(1)}$, $x_{(2)}$ ... $x_{(n)}$ – tiriamo kintamojo variacinė seka.

Statistikos D_n skirstinys, esant teisingai H_0, priklauso ne nuo skirstinio F_0, o tik nuo n. Augant n, šis skirstinys artėja prie ribinio skirstinio D. H_0 atmetimo sritis yra D_n reikšmės, viršijančios skirstinio D $(1 - \alpha)$ lygio kvantilio reikšmę. Statistiniuose paketuose pateikiama Kolmogorovo–Smirnovo kriterijaus p reikšmė. H_0 atmetama, jei p reikšmė yra mažesnė už pasirinktą reikšmingumo lygmenį.



6.5 pav. Empirinė ir teorinė skirstinio funkcijos

6.7. Normalumo tikrinimas

Kaip minėta 5.5, 6.2–6.4 skyriuose, t ir F kriterijų taikymo prielaida – duomenų normalumas. Taigi prieš taikant šiuos kriterijus, rekomenduotina patikrinti, ar kintamojo skirstinys yra normalusis. Dažniausiai normaliojo skirstinio parametrų – vidurkio ir dispersijos – nežinome, todėl normalumo tikrinimas – atskiras sudėtingos suderinamumo hipotezės atvejis.

Nulinei hipotezei H_0: „kintamojo skirstinys yra normalusis“ su alternatyva H_a: „Kintamojo skirstinys nėra normalusis“ tikrinti naudotinas χ^2 kriterijus, jei imtis nėra maža ($n > 25$), nes grupių skaičius k ir stebėtini dažniai (6.7) formulėje turi viršyti 4. Kadangi t ir F kriterijai yra jautresni išskirtims negu nukrypimams nuo normaliojo skirstinio formos duomenų centre, todėl tikrinant H_0 tikslinga pasirinkti alternatyvą, susijusią su išskirtimis. Galima naudoti alternatyvą H_1: „kintamojo skirstinys yra normaliųjų skirstinių su parametrais (m, σ^2) ir $(m, t\sigma^2)$, $t > 1$, mišinys“ arba H_2: „Kintamojo skirstinys yra normaliųjų skirstinių su parametrais (m, σ^2) ir $(m + d, \sigma^2)$, $d \neq 0$, mišinys“. H_1 alternatyvos atveju pagrindinė imtis (normalusis ats. d. su parametrais (m, σ^2)) būna „apšiukšlinta“ reikšmėmis su

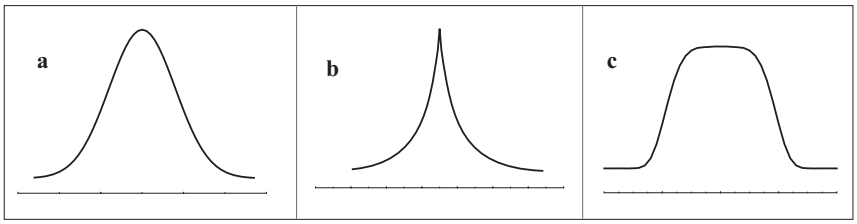
didesne dispersija – išskirtys gali būti tiek labai didelės, tiek labai mažos. H_2 alternatyvos atveju „šiukšlės“ (ats. d. su vidurkiu $m + d$) yra vienoje pusėje.

Nulinei hipotezei su alternatyva H_1 ar H_2 tikrinti naudojami kriterijai, besiremiantys ekstremalių (didžiausių ar mažiausių) imties reikšmių skirstiniu. Be jų, naudojami ir kiti kriterijai, susiję su normaliojo skirstinio savybėmis. 2.5 skyriuje pateikta vienos, dviejų ir trijų sigmų taisyklė: jei ats. d. $X \sim N(m, \sigma^2)$, tikimybė X patekti į intervalą $[m - \sigma, m + \sigma]$ lygi 0,68; į intervalą $[m - 1{,}96\sigma, m + 1{,}96\sigma]$ – 0,95, į intervalą $[m - 2{,}58\sigma, m + 2{,}58\sigma]$ – 0,99. Taigi, jeigu kintamojo skirstinys yra normalusis, intervale $[\overline{x} - s;\ \overline{x} + s]$, $[\overline{x} - 1{,}96s; \overline{x} + 1{,}96s]$, $[\overline{x} - 2{,}58s;\ \overline{x} + 2{,}58s]$ turėtų būti atitinkamai 68 %, 95 % ir 99 % imties reikšmių. Praktiškai visos imties reikšmės (99,7 %) turėtų būti intervale $[\overline{x} - 3s; \overline{x} + 3s]$. Taip pat maždaug 50 % imties reikšmių turėtų būti mažesnės už vidurkį $\overline{x}$, o 50 % – didesnės už $\overline{x}$, apie 16 % imties reikšmių mažesnės už $\overline{x} - s$, apie 16 % reikšmių turėtų viršyti $\overline{x} + s$. Jei imties reikšmių skaičius atitinkamuose intervaluose ryškiai skirsis nuo minėtų procentinių reikšmių, tuomet galima abejoti duomenų normalumu.

6.10 pavyzdys. Turime imtį: 1,26; 0,34; 0,7; 1,75; 50,57; 1,55; 0,08; 0,42; 0,5; 3,2; 0,15; 0,49; 0,95; 0,24; 1,37; 0,17; 6,98; 0,1; 0,94; 0,38. Šios imties skaitinės charakteristikos: $\overline{x} = 3{,}607$, $n = 20$, $s = 11{,}1646$, kvartiliai: $Q_1 = 0{,}29$; $Q_2 = 0{,}6$; $Q_3 = 1{,}46$. Jei imtis būtų generuota normaliojo ats. d., maždaug pusė reikšmių turėtų būti mažesnės nei vidurkis (3,607). Tačiau daugiau nei ¾ imties reikšmių mažesnės už vidurkį, nes $Q_3 < 3{,}607$; ir nėra reikšmių, mažesnių už $\overline{x} - s$ (teoriškai turėtų būti ape 16 %). Todėl tvirtiname, kad ši imtis nėra atrinkta iš normalųjį skirstinį turinčios populiacijos.

Jei priimame alternatyvą H_1 ar H_2 (imtis „apšiukšlinta“), išskirtis galima pašalinti. Paprasčiausias metodas – atmesti reikšmes, besiskiriančias nuo vidurkio trimis ar net dviem s.

Normalumo kriterijai, susiję su histogramos forma. Normalumo prielaidą atmetame, jei histogramoje išsiskiria keletas maksimumų (įtariama, kad tai kelių populiacijų duomenys). 2.4 skyriuje nurodyti du rodikliai, apibūdinantys skirstinio tankio (arba tikimybės) formą: asimetrijos koeficientas γ_1, charakterizuojantis skirstinio simetriškumą vidurkio atžvilgiu, ir ekscesas γ_2, apibūdinantis tankio kreivės lėkštumą. Normaliojo skirstinio asimetrijos ir ekscesо koeficientai lygūs 0. Jei skirstinio tankis asimetriškas į dešinę, asimetrijos koeficientas $\gamma_1 > 0$, jei asimetriškas į kairę – $\gamma_1 < 0$. Jei skirstinio tankis bukesnis už normalųjį skirstinį, jo ekscesas yra teigiamas, jei smailesnis – ekscesas neigiamas (6.6 pav.)



6.6 pav. Skirstinių tankiai su skirtingais ekscesais: a) ekscesas lygus 0; b) teigiamas ekscesas; c) neigiamas ekscesas

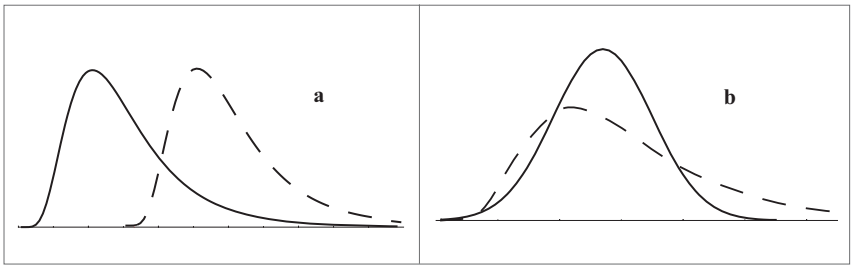
Populiacijos skirstinio asimetrijos ir eksceso koeficientai vertinami imties asimetrijos (g_1) ir eksceso (g_2) koeficientais; g_1 formulė pateikta 1.8 skyriuje. Nustatyta, kad augant *n,* rodikliai g_1 ir g_2 artėja prie atitinkamų skirstinio formos rodiklių γ_1 ir γ_2 (čia artėjimas suprantamas tikimybine prasme: tikimybė, kad g_i skirsis nuo γ_i labai mažu dydžiu, augant n, artėja į 1). Žinoma, kad normaliojo skirstinio atveju g_1 ir g_2 yra asimptotiškai normalieji su nuliniu vidurkiu, o standartizuotų dydžių $g_1/\mathrm{se}(g_1)$ ir $g_2/\mathrm{se}(g_2)$ asimptotinis skirstinys yra standartinis normalusis; čia se(...) – rodiklio standartinė paklaida. Statistiniuose paketuose (SPSS, STATISTICA) pateikiamos g_1, g_2 ir jų standartinių paklaidų reikšmės. Jei nustatoma, kad $|g_1/\mathrm{se}(g_1)| > z_{1-\alpha/2}$ ($|g_2/\mathrm{se}(g_2)| > z_{1-\alpha/2}$), tuomet tvirtinama, kad γ_1 (γ_2) nėra lygus 0 – taigi populiacijos skirstinys nėra normalusis (α – reikšmingumo lygmuo). Jei $|g_i/\mathrm{se}(g_i)| \le z_{1-\alpha/2}$ – teiginiui $\gamma_i = 0$ (kartu ir normalumui) neprieštaraujame, čia i = 1, 2.

6.11 pavyzdys. 6.10 pavyzdyje pateiktos imties asimetrijos ir eksceso koeficientai lygūs: $g_1 = 4{,}33$, $g_2 = 19{,}08$; jų standartiniai nuokrypiai atitinkamai lygūs 0,512 ir 0,992, $\alpha = 0{,}05$, $z_{0{,}975} = 1{,}96$. Statistikų, naudojamų tikrinti hipotezę „populiacijos asimetrijos ir eksceso koeficientai lygūs nuliui“, reikšmės yra $4{,}33/0{,}512 = 8{,}457 > 1{,}96$ ir $19{,}08/0{,}992 = 19{,}23 > 1{,}96$; atitinkamos p reikšmės mažesnės nei 0,001. Todėl tvirtiname, jog asimetrijos ir eksceso koeficientai reikšmingai skiriasi nuo nulio. Tai normaliajam skirstiniui nebūdinga, todėl imtis nėra iš normalųjį skirstinį turinčios populiacijos.

6.8. Hipotezės apie dviejų populiacijų skirstinių tapatumą

Kaip minėta, vienas medikams kylančių aktualių uždavinių – palyginti ligonių su patologija (imtis iš X populiacijos) tyrimų duomenis su kontrolinės grupės (imtis iš Y populiacijos) atitinkamais duomenimis. Dažniausiai

lyginami abiejų ligonių grupių tiriamo kintamojo vidurkiai, t. y. tikrinama hipotezė H_0: $m_x = m_y$ su dešiniapuse, kairiapuse ar dvipuse alternatyva. Tačiau patologijos poveikis gali pasireikšti ne tik skirstinio vidurkio poslinkiu (6.7 a pav.), bet ir skirstinio iškraipymu (6.7 b pav.). Todėl patologijos poveikiui konstatuoti kartais tikslinga ir tikrinti hipotezę apie kintamojo skirstinių abiejose ligonių grupėse tapatumą (arba populiacijų identiškumą). Šiam tikslui naudojamas Kolmogorovo–Smirnovo kriterijus.



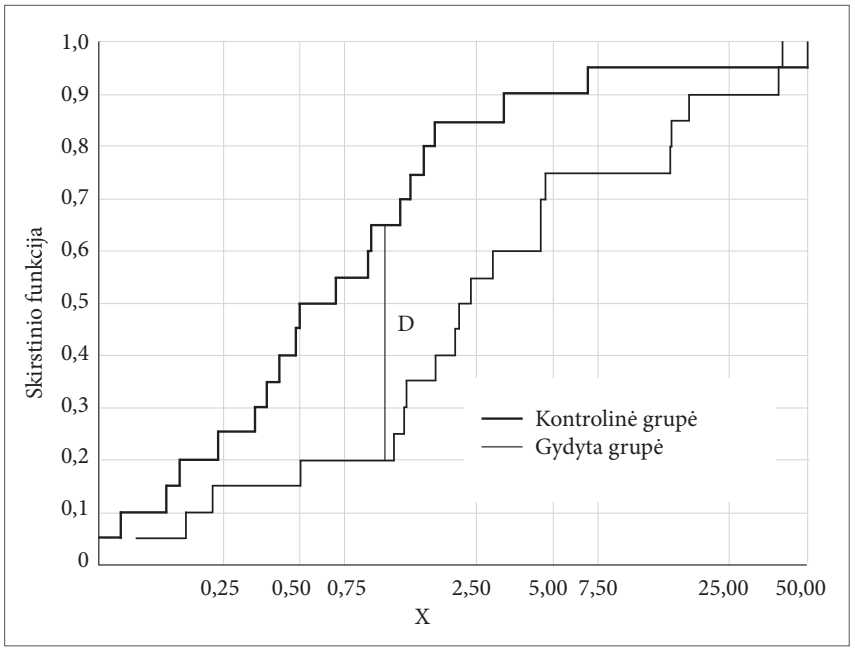
6.7 pav. Poveikis pasireiškia skirstinio poslinkiu (a), skirstinio iškraipymu (b)

Kolmogorovo–Smirnovo kriterijus. Taikymo prielaida – kintamojo skirstiniai abiejose populiacijose yra tolydieji.

Sakykime, turime dvi imtis: imtį iš X populiacijos (pvz., sergančių ligonių) $x_1, x_2 \ldots x_n$ ir imtį iš Y populiacijos (pvz., sveikų asmenų) $y_1, y_2 \ldots y_m$. Kintamojo iš populiacijos X skirstinio funkciją pažymėkime $F(x)$, iš populiacijos Y – $G(x)$. Tikrinsime nulinę hipotezę H_0: „populiacijų skirstiniai identiški“ ($F(x) \neq G(x)$) su alternatyva „populiacijų skirstiniai nėra identiški“ ($F(x) \neq G(x)$). Kolmogorovo–Smirnovo kriterijaus statistika grindžiama dviejų empirinių skirstinio funkcijų $E_{1n}(x)$ ir $E_{2m}(x)$ skirtumu (6.8 pav.). Ji lygi:

$$D_{n,m} = \max| E_{1n}(x) - E_{2m}(x)|;$$

čia $E_{1n}(x)$ ir $E_{2m}(x)$ – imčių $x_1, x_2 \ldots x_n$ ir $y_1, y_2 \ldots y_m$ empirinės skirstinio funkcijos. Statistikos $D_{n,m}$ skirstinys, esant teisingai nulinei hipotezei, nepriklauso nuo skirstinio funkcijų $F(x)$ ir $G(x)$, o priklauso tik nuo n ir m. Išvados apie H_0 atmetimą ar jai neprieštaravimą paprastai grindžiamos kriterijaus p reikšme, pateikiama statistiniuose paketuose: jei p viršija ar lygi 0,05 – teigiama, kad populiacijų skirstiniai indentiški; jei $p < 0{,}05$ – skirstiniai nėra identiški (čia $\alpha = 0{,}05$).



6.8 pav. Dviejų imčių empirinės skirstinio funkcijos, $D_{n,m}$ statistika. X pateiktas logaritminiu masteliu

6.12 pavyzdys [8]. Sakykime, turime gydytą ir kontrolinę ligonių grupes. Tyrimo metu nustatėme kintamojo reikšmes:

Kontrolinė grupė: 1,26; 0,34; 0,7; 1,75; 50,57; 1,55; 0,08; 0,42; 0,5; 3,2; 0,15; 0,49; 0,95; 0,24; 1,37; 0,17; 6,98; 0,1; 0,94; 0,38.

Gydyta grupė: 2,37; 2,16; 14,82; 1,73; 41,04; 0,23; 1,32; 2,91; 39,41; 0,11; 17,44; 4,51; 0,51; 4,5; 0,18; 14,68; 4,66; 1,3; 2,06; 1,19.

Gydytos ir kontrolinės grupių imties skaitinės charakteristikos pateiktos 6.7 lentelėje.

6.7 lentelė. Gydytos ir kontrolinės imties skaitinės charakteristikos

	Vidurkis	St. nuokrypis	Mediana	Min.	Maks.	Q_1	Q_3
Kontrolinė gr.	3,607	11,16	0,6	0,08	50,57	0,29	1,46
Gydyta gr.	8,357	12,82	2,27	0,11	41,04	1,25	9,67

Iš 6.7 lentelės matyti, kad imčių negalima laikyti normaliosiomis – nėra reikšmių, mažesnių nei $\overline{x} - s$, o jų turėtų būti apie 16 %. Todėl imčių vidur-

kiams palyginti negalime taikyti *t* kriterijaus. Imtims, o kartu ir populiacijoms palyginti naudosime Kolmogorovo–Smirnovo statistiką $D_{n,m}$.

Kontrolinės ir gydytos grupių empirinių skirstinio funkcijų grafikai logaritmine argumento skale pateikti 6.8 pav. Didžiausias empirinių skirstinio funkcijų skirtumas yra arti taško $x = 1$; jis lygus 0,45. Šio skirtumo atitinkama *p* reikšmė mažesnė už 0,05, todėl hipotezę apie populiacijų tapatumą atmetame su 0,05 reikšmingumo lygmeniu.

6.9. Kelių populiacijų medianų palyginimas

Atliekant duomenų analizę, medikams tenka lyginti ir daugiau nei dviejų populiacijų rodiklius. Pavyzdžiui, vienai ligonių grupei duota 20 mg vaisto dozė, kitai – 10 mg, trečiai – placebo. Po to visų trijų grupių ligoniams išmatuojamas SAS (ar kitas rodiklis). Vaisto dozės poveikiui nustatyti būtina palyginti visų 3 grupių SAS vidurkius ar medianas. Kelių vidurkių lyginimas, kai kintamojo skirstinys yra normalusis, nagrinėjamas 12 skyriuje. Šiame skyriuje pateiksime kriterijų, skirtą *k* ($k > 2$) medianoms palyginti.

Kruskalio–Voliso (***Kruskal–Walis***) **kriterijus**. Taikymo prielaida: kintamojo skirstinys visose *k*, $k > 2$ populiacijose yra simetriškas vidurkio atžvilgiu. Tuomet hipotezė apie kintamojo vidurkio populiacijose tapatumą identiška hipotezei apie kintamojo medianų populiacijose lygybę.

Tikrinama H_0: „visų skirstinių medianos lygios“ su alternatyva: „bent dvi medianos nėra lygios“. Nulinei hipotezei tikrinti skaičiuojama kriterijaus statistika naudoja imties reikšmių rangus. Visos imtys sujungiamos, jų reikšmės suranguojamos; po to skaičiuojama Kruskalio–Voliso kriterijaus statistika pagal formulę:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1);$$

čia $n_1, n_2 \ldots n_k$ – imčių dydžiai, $N = n_1 + \ldots + n_k$ – bendras matavimų skaičius, R_i – *i*-tos imties rangų suma.

Esant teisingai nulinei hipotezei, statistikos *H* skirstinys artimas χ^2 skirstiniui su $(k - 1)$ laisvės laipsnių. Nulinė hipotezė atmetama, jei atitinkama kriterijaus *p* reiksmė neviršys parinkto reikšmingumo lygmens α (0,05) arba kai $H > \chi^2_{1-\alpha}(k-1)$. Priešingu atveju nulinei hipotezei neprieštaraujame.

6.10. Kelių kartotinių matavimų palyginimas

Sakykime, atliekama to paties individo kiekybinio rodiklio *k* kartotinių matavimų. Pavyzdžiui, echoskopija atliekama ligoniams prieš PTCA, išrašant iš stacionaro, po 6 mėn., po metų; t. y. atliekami 4 pakartotiniai tyrimai ir jų metu nustatomi kiekybiniai echoskopijos rodikliai – kintamieji KSGDD, IF ... . Analizuojant operacijos rezultatus, aktualu nustatyti, ar tiriamos ligonių populiacijos KSGDD vidurkiai pakito, ar ne. Keliems kartotiniems matavimams palyginti naudojamas Frydmano kriterijus.

Frydmano (***Friedman***) **kriterijus**. Prielaida: kartotinių matavimų skirtumų skirstiniai yra simetriški nulio atžvilgiu. Tuomet hipotezė apie poveikio nebuvimą tapati hipotezei apie visų kartotinių matavimų medianų (ar vidurkių) lygybę.

Tikrinsime nulinę hipotezę H_0: „visi kartotiniai matavimai iš populiacijų su ta pačia mediana“ su alternatyva: „bent dviejų kartotinių matavimų medianos skiriasi“.

Kriterijaus statistika skaičiuojama taip. Sakykime, *n* individams atlikta po *k* kartotinių matavimų. Pirmiausia suranguojamos kiekvieno individo reikšmės (ats. d.); gautų rangų suma lygi $k(k + 1)/2$. Po to sudedami rangai, tekę kiekvienam kartotiniam matavimui. Gaunamos rangų sumos R_1, R_2 ... R_k.

Frydmano kriterijaus statistika lygi:

$$S = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \cdot$$

Esant teisingai nulinei hipotezei, *S* skirstinys artimas χ^2 skirstiniui su $(k - 1)$ laisvės laipsnių. Nulinė hipotezė atmetama, jei kriterijaus *p* reikšmė neviršys parinkto reikšmingumo lygmens α (0,05) arba kai $S > \chi^2_{1-\alpha}(k-1)$. Priešingu atveju nulinei hipotezei neprieštaraujame.

6.13 pavyzdys ([5, 151 p.]). Pesticidų kiekis 4 augalo mėginiuose nustatytas trimis metodais A, B ir C (6.8 lentelė). Ar galima tvirtinti, jog visais metodais gauti tie patys rezultatai?

Prielaida, kad bet kuriais dviem metodais gauto pesticidų kiekio skirtumai turi skirstinį, simetrišką vidurkio atžvilgiu, nėra didelė. Skaičiuojame Frydmano kriterijaus statistiką. Turime $k = 3$, $n = 4$, $\alpha = 0{,}05$. Kiekvieną bandinį atitinkantys rangai pateikti 6.8 lentelėje. Metodų A, B ir C rangų sumos R_A, R_B, R_C atitinkamai lygios 5,5, 8,5 ir 10. $S = 12(5{,}5 + 8{,}5 + 10)/(4 \times 3 \times 4) - 3 \times 4 \times 4 = 2{,}65$.

6.8 lentelė. Pesticidų kiekio 4 augalo mėginiuose absoliučios reikšmės ir rangai

Mėginys	Nustatytos reikšmės			Rangai		
	A	B	C	A	B	C
1	4,7	5,8	5,7	1	3	2
2	7,7	7,7	8,5	1,5	1,5	3
3	9,0	9,9	9,5	1	3	2
4	2,3	2,0	2,9	2	1	3

Turime $\chi^2_{0,95}(2) = 5{,}99 > S$. Todėl daroma išvada: nėra pagrindo teigti, kad šiais metodais gaunami skirtingi rezultatai.

6 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. 2000. Vilnius: TEV, 238 p.
3. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
4. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.
5. Miller J. C., Miller J. N. *Statistics for Analycal Chemistry*. Second ed. 1988. New York: John Wiley & Sons, p. 227.
6. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
7. Холлендер М., Вулф Д. А. *Непараметрические методы статистики*. 1983. Москва: Наука, 516 с.
8. Watts S., Halliwell L. *Essential Environmental Science. Methods and Techniques*. 1996, p. 512.
9. *Kolmogorovo–Smirnovo kriterijus dviejų populiacijų tolydiesiems skirstiniams palyginti.* Prieiga per internetą: http://www.physics.csbsju.edu/stats/KS-test.html.

7 SKYRIUS

Porinės dažnių lentelės analizė

Analizuojant susirgimo ar patologijos priežastis, aktualu nustatyti ligonio rodiklius, susijusius su susirgimu ar patologijos laipsniu. Dalis individo charakteristikų apibūdinamos kokybiniais kintamaisiais – nominaliu, dvinariu ar tvarkos. Susirgimo ar patologijos laipsnis išreiškiamas ir kokybiniu kintamuoju. Todėl sprendžiant minėtus uždavinius, reikalingi dvimačių kokybinių kintamųjų statistiniai modeliai bei kokybinių kintamųjų tarpusavio ryšio (*association*) matai. Porine dažnių lentele pateikiami ir kokybinio kintamojo tyrimo rezultatai, gauti iš įvairių epidemiologinių studijų.

7.1. Porinė dažnių lentelė

Sakykime, X ir Y – du kokybiniai kintamieji: kintamasis X įgyja r reikšmių (kategorijų), kintamasis Y – c reikšmių; abiejų kintamųjų reikšmės nustatytos n individų. Šių dviejų kintamųjų reikšmių tarpusavio išsidėstymas imtyje pateikiamas porine $r \times c$ dažnių lentele, turinčia r eilučių ir c stulpelių. Lentelės eilutės atitinka kintamojo X, stulpeliai – kintamojo Y reikšmes. Individų, turinčių kintamojo X i-tąją reikšmę ir kintamojo Y j-tąją reikšmę, skaičius n_{ij} (absoliutus dažnis) pateikiamas porinės dažnių lentelės i-tosios eilutės ir j-tojo stulpelio susikirtimo gardelėje (7.1 lentelė). Individų, turinčių kintamojo X i-tąją reikšmę, skaičius žymimas n_{i+}, turinčių kintamojo Y j-tąją reikšmę skaičius – n_{+j}; n – tirtų individų skaičius. Šie skaičiai susieti priklausomybe:

$$
\begin{aligned}
n_{i+} &= n_{i1} + n_{i2} + \ldots + n_{ic} = \sum_{j=1}^{c} n_{ij}, \\
n_{+j} &= n_{1j} + n_{2j} + \ldots + n_{rj} = \sum_{i=1}^{r} n_{ij}, \\
n &= n_{1+} + n_{2+} + \ldots + n_{r+} = n_{+1} + n_{+2} + \ldots + n_{+c}.
\end{aligned}
\tag{7.1}
$$

Taip pat $r \times c$ lentelėje gali būti pateikti:

- santykiniai dažniai $p_{ij} = n_{ij}/n$;
- santykiniai dažniai eilutėse $p_{j|i} = n_{ij}/n_{i+}$, atspindintys kintamojo Y skirstinį, kai kintamasis X įgyja i-tąją reikšmę;
- santykiniai dažniai stulpeliuose $p_{j|j} = n_{ij}/n_{+j}$, atspindintys kintamojo X reikšmių skirstinį, kai kintamasis Y įgyja j-tąją reikšmę.

Santykiniai dažniai gali taip pat būti pateikti ir procentais (eilutėse ir stulpeliuose).

7.1 lentelė. Porinė r×c dažnių lentelė

$X \setminus Y$	1	2	...	j	...	c	Iš viso
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_{1+}
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_{2+}
...	...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_{i+}
...	...	...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	n_{r+}
Iš viso	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+c}	n

7.1 pavyzdys. 7.2 lentelėje pateikti duomenys apie ligonių, sergančių ŪKS, išgyvenimo 30 parų laikotarpiu priklausomybę nuo koronarinio sindromo ([3]). Čia kintamasis X – koronarinis sindromas (I reikšmė – Q bangos MI, II reikšmė – be Q bangos MI, III reikšmė – nestabili KA), kintamasis Y – išgyvenimas 30 parų laikotarpiu (I reikšmė – mirė, II reikšmė – išgyveno). Šių kintamųjų tarpusavio priklausomybė pateikiama 3×2 porine dažnių lentele. 7.2 lentelėje pateikti atvejų skaičiai – mirusių 30 parų laikotarpiu ligonių: su Q bangos MI $n_{11} = 20$, be Q bangos MI – $n_{21} = 4$, nestabilios KA – $n_{31} = 1$; išgyvenusių: su Q bangos MI $n_{12} = 252$, be Q bangos MI – $n_{22} = 214$, nestabilios KA – $n_{32} = 155$. Q bangos MI ligonių iš viso buvo $n_{1+} = 272$, be Q bangos $n_{2+} = 218$, nestabilios KA – $n_{3+} = 156$. Mirusių ligonių buvo $n_{+1} = 25$, išgyvenusių – $n_{+2} = 621$. Iš viso tirta $n = 646$ ligonių. 7.2 lentelėje taip pat pateikti santykiniai dažniai eilutėse procentais.

7.2 lentelė. Ligonių, sergančių ŪKS, išgyvenimo 30 parų laikotarpiu priklausomybė nuo koronarinio sindromo (atvejų skaičius ir santykiniai dažniai eilutėse procentais)

	Mirė		Išgyveno		Iš viso
	N	% eilutėse	N	% eilutėse	
Q bangos MI	20	7,35	252	92,65	272
Be Q bangos MI	4	1,83	214	98,17	218
Nestabili KA	1	0,64	155	99,36	156
Iš viso	25	3,87	621	96,13	646

7.2. Dviejų kokybinių kintamųjų statistinis modelis

Imties dydis *n* yra fiksuotas. Sakykime, populiacijoje atsitiktinai atrinkta n individų; kiekvienam individui nustatytos kokybinių kintamųjų X ir Y reikšmės; X įgyja r, o Y – c skirtingų reikšmių. Tokio tyrimo statistinis modelis – X ir Y yra atsitiktiniai dydžiai; šiuo modeliu aprašomi kohortinės bei momentinės epidemiologinės studijos duomenys. Individo įgyjama (X, Y) reikšmių kombinacija yra atsitiktinė, apibrėžiama jungtiniu dvimačiu tikimybiniu skirstiniu su $r \times c - 1$ nežinomų parametrų $\pi_{ij} = P\{X = x_i; Y = y_j\}$ (2.6 lentelė); čia π_{ij} – tikimybė, kad X įgis i-tąją, Y – j-tąją reikšmę. X ir Y atskirų (vienmačių) skirstinių tikimybes žymėsime: $\pi_{i+} = P\{X = x_i\} = \pi_{i1} + ... + \pi_{ic}$, $\pi_{+j} = P\{Y = y_j\} = \pi_{1j} + ... + \pi_{rj}$.

Nežinomų (X, Y) skirstinio parametrų – tikimybių π_{ij} įverčiai, gauti didžiausio tikėtinumo metodu, yra santykiniai dažniai $p_{ij} = n_{ij} / n$. X ir Y skirstinių tikimybės π_{i+} ir π_{+j} įverčiai yra atitinkamai:

$p_{i+} = n_{i+} / n$, $p_{+j} = n_{+j} / n$, $i = 1, 2 ... r$, $j = 1, 2 ... c$;

čia n_{i+} ir n_{+j} – ats. d., apibrėžti pagal (7.1) formulę. Santykiniai dažniai p_{ij} yra atsitiktiniai dydžiai, p_{ij} vidurkis lygus π_{ij}, dispersija lygi $\pi_{ij}\ (1 - \pi_{ij})/n$, standartinio nuokrypio įvertis, arba standartinė paklaida – $\sqrt{n_{ij}(1 - n_{ij} / n)}/n$. p_{i+} ir p_{+j} standartinės paklaidos skaičiuojamos analogiškai.

Vienas iš kintamųjų X, Y yra neatsitiktinis. Porine dažnių lentele (7.1 lentelė) pateikiami ir klinikinio eksperimento (*clinic trials*) tyrimų duomenys. Klinikinio eksperimento metu iš r populiacijų atsitiktinai atrenkama n_{1+}, $n_{2+} ... n_{r+}$ individų ir kiekvienam jų nustatoma kokybinio kintamojo Y reikšmė. Tokio tyrimo rezultatus aprašančioje porinėje dažnių lentelėje kintamojo X reikšmės yra neatsitiktinės – jos skirtos populiacijai (vyrų, moterų; vartojo vaistą, placebo) identifikuoti. Dažniai n_{ij} yra kintamojo Y j-tosios reikšmės dažnis, nustatytas i-toje populiacijoje; dažniai n_{1+}, $n_{2+} ... n_{r+}$ yra imčių dydžiai (neatsitiktiniai). n_{ij} yra atsitiktiniai dydžiai; skirtingų eilučių dažniai – nepriklausomi ats. dydžiai.

Kokybinio kintamojo Y skirstinys i-toje populiacijoje nusakomas tikimybėmis $\pi_{1|i}$, $\pi_{2|i}$, ..., $\pi_{c|i}$, čia $\pi_{j|i} = P\{Y = y_j | X = x_i\}$. Sąlyginių tikimybių $\pi_{j|i}$ (2.8 skyrius) įverčiai $p_{j|i}$ yra:

$p_{j|i} = p_{ij} / p_{i+} = n_{ij} / n_{i+}$.

Sąlyginės tikimybės $p_{j|i}$ standartinė paklaida lygi $\sqrt{p_{j|i}(1 - p_{j|i}) / n_{j+}}$.

7.2 pavyzdys [7]. Siekiant palyginti hipertenzine (H), išemine (IŠ) ir idiopatine dilatacine (ID) kardiomiopatija sergančiųjų klinikinių, elektrokardiografinių, echoskopijos ir perfuzijos tyrimų duomenis, atsitiktinai atrinkta

po 30 kiekvienos patologijos ligonių (atsitiktinė atranka iš 3 populiacijų). Be kitų rodiklių, kiekvienam ligoniui nustatyta ir širdies ritmo sutrikimų – kintamasis *Y*: *Y* = 1, jei ritmo sutrikimų buvo, ir *Y* = 0, jei ritmo sutrikimų nebuvo. Tuomet ritmo sutrikimų pasiskirstymas tarp sergančiųjų hipertenzine (*X* = 1), išemine (*X* = 2) ir idiopatine dilatacine (*X* = 3) kardiomiopatija pateikiamas 3×2 porine dažnių lentele (7.3 lentelė). 7.4 lentelėje pateikti ritmo sutrikimo (Y) skirstinio tikimybių įverčiai kiekvienoje populiacijoje.

7.3 lentelė. Ritmo sutrikimų pasiskirstymas pagal kardiomiopatijos rūšį

Kardiomiopatijos rūšis	Ritmo sutrikimų nėra	Ritmo sutrikimų yra	Iš viso
Hipertenzinė	18	12	30
Išeminė	6	24	30
Idiopatinė dilatacinė	13	17	30

7.4 lentelė. Ritmo sutrikimų procentinis pasiskirstymas sergančiųjų hipertenzine, išemine ir idiopatine dilatacine kardiomiopatija

Kardiomiopatijos rūšis	Ritmo sutrikimų nėra		Ritmo sutrikimų yra	
	$p_{1\|i}$	$p_{1\|i}$ (%)	$p_{2\|i}$	$p_{2\|i}$ (%)
Hipertenzinė	0,6	60	0,4	40
Išeminė	0,2	20	0,8	80
Idiopatinė dilatacinė	0,433	43,3	0,567	56,7

Iš 7.4 lentelės matyti, kad ritmo sutrikimų pasitaikė 40 % ligonių, sergančių hipertenzine kardiomiopatija, 80 % ligonių, sergančių išemine, ir 56,7 % ligonių, sergančių ID kardiomiopatija.

Atvejo–kontrolės studijose atsitiktinai atrenkamas fiksuotas skaičius atvejų (*Y* = 1) ir atsitiktinai atrenkama kontrolė (*Y* = 0), t. y. *Y* neatsitiktinis, reikšmės n_{+1} ir n_{+2} – fiksuotos. Tuomet galima analizuoti *X* skirstinį atvejams ir kontrolei.

7.3. Kokybinių kintamųjų nepriklausomumas ir jo tikrinimas

Medicinoje ir epidemiologijoje dažnai susiduriama su ryšio tarp kokybinių kintamųjų problema: yra tarp kintamųjų ryšys ar jo nėra. Medikams svarbu žinoti, ar:

1) vaisto vartojimas turėjo įtakos kraujospūdžio sumažėjimui;

2) ŪŠN klasė ar ritmo sutrikimai, nustatyti susirgus MI, turėjo įtakos mirties 1 metų laikotarpiu dažniui;

3) moterims, sergančioms I laipsnio CD, retinopatija turi įtakos osteopenijai atsirasti;

4) odos tipas turi įtakos melanomai atsirasti.

Minėtoms hipotezėms patvirtinti ar atmesti sudaromos ligonių studijos. Pagal galimybes ir keliamą uždavinį tyrėjas parenka ligonių studijos tipą, pavyzdžiui, kohortinę, momentinę ar atvejo–kontrolės. Pateiksime porą kriterijų, skirtų nepriklausomumo hipotezėms tikrinti.

Pirsono χ^2 kriterijus kokybinių kintamųjų nepriklausomumui tikrinti. Prielaida: (X, Y) – dvimatis diskretus atsitiktinis vektorius su tikimybiniu skirstiniu $\{\pi_{ij}\}$. Todėl hipotezės apie kintamųjų X ir Y ryšio buvimą ar nebuvimą tikrinimas yra hipotezės apie dvimačio skirstinio dedamųjų nepriklausomumą (žr. 2.8 skyrius) tikrinimas.

Nulinė hipotezė H_0 apibrėžiama: „kintamieji X ir Y nepriklausomi“. Pagal apibrėžimą, ats. d. X ir Y yra nepriklausomi, jei tikimybė, kad X įgis i-tąją ir Y j-tąją reikšmę π_{ij}, lygi tikimybių, kad X įgys i-tąją, o Y – j-tąją reikšmę, sandaugai:

$\pi_{ij} = \pi_{i+} \times \pi_{+j}$ visiems i ir j.

Alternatyvą galima išreikšti taip: „bent vienai i ir j porai $\pi_{ij} \neq \pi_{i+} \times \pi_{+j}$“. Nulinei hipotezei tikrinti naudojamas Pirsono χ^2 kriterijus su statistika, pagrįsta stebėtų ir tikėtinų dažnių įverčių skirtumu. Stebėti dažniai yra n_{ij} – dažniai, pateikti porinėje dažnių lentelėje. Esant teisingai nulinei hipotezei, kintamojo X i-tąją ir kintamojo Y j-tąją reikšmę turėtų maždaug $n(\pi_{i+} \times \pi_{+j})$ individų. Skaičiai $n(\pi_{i+} \times \pi_{+j})$ vadinami tikėtinais dažniais. Kadangi tikimybių π_{i+} ir π_{+j} nežinome, χ^2 statistikoje tikėtinus dažnius pakeičiame jų įverčiais $E_{ij} = np_{i+}p_{+j} = n_{i+}n_{+j}/n$, nes parametrų π_{i+} ir π_{+j} didžiausio tikėtinumo įverčiai yra p_{i+} ir p_{+j}. Nulinei hipotezei tikrinti naudojamo Pirsono χ^2 kriterijaus statistika lygi:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}. \qquad (7.2)$$

Jei ats. d. X ir Y yra nepriklausomi (tarp kintamųjų X ir Y ryšio nėra), tuomet statistikos χ^2 asimptotinis skirstinys yra χ^2 skirstinys su $(r - 1)(c - 1)$ laisvės laipsnių. Esant gana dideliam n ir nenuliniams stebėtiems dažniams, galime tvirtinti, jog statistika (7.2) turi χ^2 skirstinį su $(r - 1)(c - 1)$ laisvės laipsnių. Statistiniuose paketuose pateikiama χ^2 ir atitinkama p reikšmė. Jei $p < \alpha$ (čia α – parinktas reikšmingumo lygmuo (I rūšies klaidos tikimybė; dažniausiai naudojama $\alpha = 0{,}05$)), H_0 atmetama ir priimama alternatyva – tarp kintamųjų X ir Y yra ryšys. Jei $p \geq \alpha$ – kintamųjų nepriklausomumui neprieštaraujame.

Didžiausio tikėtinumo santykio kriterijus (***ML*** χ^2). Prielaida ta pati, kaip ir Pirsono χ^2 kriterijui. *ML* χ^2 yra kriterijus hipotezei apie kintamųjų *X* ir *Y* nepriklausomumą (ryšio nebuvimą) tikrinti. Jo statistika G^2 lygi:

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln(n_{ij} / E_{ij}).$$

Jei *X* ir *Y* yra nepriklausomi, G^2 taip pat turi asimptotinį χ^2 skirstinį su $(r - 1)(c - 1)$ laisvės laipsnių. Taigi gana dideliam *n* statistiką G^2 galima laikyti ats. dydžiu, turinčiu χ^2 skirstinį. Jei *ML* χ^2 kriterijaus $p < \alpha$, H_0 atmetama ir priimama alternatyva. Priešingu atveju H_0 neprieštaraujame.

χ^2 ir *ML* χ^2 kriterijaus negalima taikyti esant nuliniams dažniams porinėje dažnių lentelėje. Kokybinių požymių nepriklausomumui tirti rekomenduojama χ^2 ir *ML* χ^2 kriterijus taikyti tik tuomet, kai visi dažniai n_{ij} viršija 4 arba, geriausiu atveju, kai $n/r \times c \geq 5$. Priešingu atveju išvados, gautos remiantis šiais kriterijais, gali būti neteisingos. Esant mažai imčiai ar mažiems dažniams, kintamųjų nepriklausomumui tikrinti rekomenduojama taikyti tikslius kriterijus.

7.3 pavyzdys ([3]). Naudodami χ^2 ir *ML* χ^2 kriterijus, tikrinsime hipotezę, ar ligonių, persirgusių ŪKS, mirčių skaičius 1 metų laikotarpiu priklauso nuo koronarinio sindromo. 7.5 lentelėje pateikti stebėti dažniai (n_{ij}), tikėtinų dažnių įverčiai bei kiti duomenys χ^2 kriterijaus statistikai skaičiuoti. Naudodamiesi statistiniu paketu nustatėme, jog $\chi^2 = 10{,}65$, $G^2 = 12{,}13$. Laisvės laipsnių skaičius lygus 2, abiejų kriterijų *p* reikšmės mažesnės nei 0,0001. Todėl galime tvirtinti, kad mirčių skaičius 1 metų laikotarpiu reikšmingai priklauso nuo koronarinio sindromo.

7.5 lentelė. Letalios baigties 1 metų laikotarpiu pasiskirstymas priklausomai nuo koronarinio sindromo: stebėti dažniai ir tikėtinų dažnių įverčiai

Koronarinis sindromas	Mirė		Išgyveno		Iš viso
	n_{i1}	E_{i1}	n_{i2}	E_{i2}	
Q MI	34	269×58/642=24,3	235	269×584/642=244,7	269
Be Q MI	19	218×58/642=19,7	199	218×584/642=198,3	218
NKA	5	155×58/642=14,0	150	155×584/642=141,0	155
Iš viso	58		584		642
$\chi^2 = (34 - 24{,}3)^2/24{,}3 + (19 - 19{,}7)^2/19{,}7 + (5 - 14)^2/14 + (235 - 244{,}7)^2/244{,}7 + (199 - 198{,}3)^2/198{,}3 + (150 - 141)^2/141 = 10{,}65$					

Jei porinėje dažnių lentelėje pateikiami kelių nepriklausomų imčių kokybinio rodiklio matavimų rezultatai, t. y. *X* – neatsitiktinis ir dažniai n_{i+} fiksuoti, nepriklausomumo hipotezė virsta homogeniškumo hipoteze. Sakykime,

X – kintamasis, nurodantis populiaciją, pavyzdžiui, arba vartotas vaistas ($X = 1$), arba placebas ($X = 0$). Tuomet Y priklausomybė nuo X nustatoma tikrinant hipotezę, ar Y skirstinys yra identiškas visoms X reikšmėms, t. y. tikrinama nulinė hipotezė H_0: „$\pi_{j|1} = \pi_{j|2} = ... = \pi_{j|r} = \pi_j$, $j = 1 ... c$.

H_0 tikrinti naudojami χ^2 ir *ML* χ^2 kriterijai. Šių kriterijų statistikos tokios pat kaip ir statistikos, naudojamos nepriklausomumo hipotezei tikrinti.

7.4. 2×2 porinė dažnių lentelė

Daugelio priežastinių reiškinių tyrimo rezultatai pateikiami 2×2 porine dažnių lentele (7.5 lentelė). Pavyzdžiui, UŠN sunkumo laipsnį galime vertinti dvinariu kintamuoju X: $X = 1$, jei UŠN klasė yra I arba II, ir $X = 2$, jei UŠN klasė yra III arba IV. Tuomet letalios baigties 30 dienų laikotarpiu priklausomybė nuo UŠN pateikiama 2×2 porine dažnių lentele (7.6 lentelė).

7.5 lentelė. Standartinė 2×2 porine dažnių lentelė, pateikianti dvinarių kintamųjų tarpusavio išsidėstymą

$X \setminus Y$	I kategorija	II kategorija	Iš viso
I kategorija	a	b	a + b
II kategorija	c	d	c + d
Iš viso	$a + c$	$b + d$	$a + b + c + d$

7.6 lentelė. UŠN sunkumo laipsnio ir letalios baigties 30 dienų laikotarpiu pasiskirstymas

	Mirė	Išgyveno
I–II UŠN klasė	12	416
III–IV UŠN klasė	12	19

Jei visi a, b, c, d ne mažesni už 5, hipotezei apie dvinarių kintamųjų X ir Y nepriklausomumą tikrinti galima naudoti χ^2 kriterijų; šio kriterijaus statistika (7.2) po pertvarkymų yra tokia:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \qquad (7.3)$$

Statistikos (7.3) ribinis skirstinys χ^2 yra tolydusis, bet naudojamas diskretaus skirstinio aproksimacijai. Todėl hipotezei apie dvinarių kintamųjų nepriklausomumą tikrinti vietoj χ^2 kriterijaus su statistika (7.2) naudojamas pataisytas χ^2, arba **Jeitso** (*Yates*) **χ^2 kriterijus**, kurio statistika yra χ^2_Y:

$$\chi_Y^2 = \frac{n(|\, ad - bc\,| - 0{,}5n)^2}{(a+b)(a+c)(c+d)(b+d)}.$$

Tiek χ^2, tiek χ_Y^2, esant kintamųjų nepriklausomumui, turi asimptotinį χ^2 skirstinį su 1 laisvės laipsniu. Statistiniuose paketuose pateikiamos šių kriterijų *p* reikšmės. Jei $p < 0{,}05$ ($\alpha = 0{,}05$), hipotezė apie požymių nepriklausomumą atmetama ir teigiama, kad tarp kintamųjų yra ryšys. Priešingu atveju kintamųjų nepriklausomumui neprieštaraujame.

7.4 pavyzdys. Tikrinant hipotezę, ar letalios baigties tikimybė priklauso nuo UŠN laipsnio (7.6 lentelė), statistikų χ^2, χ_Y^2 ir *ML* χ^2 reikšmės atitinkamai lygios 75,2; 37,5 ir 68,1; visos *p* reikšmės yra mažesnės nei 0,00001. Todėl daroma išvada, kad UŠN sunkumo laipsnis turi įtakos letalios baigties tikimybei.

Kriterijai χ^2, χ_Y^2 ir *ML* χ^2 naudotini tik tais atvejais, kai dydžiai *a*, *b*, *c*, ir *d* nėra mažesni už 5. Jei atvejų skaičiai 2×2 lentelėje mažesni nei 5, hipotezė apie kintamųjų nepriklausomumą tikrinama pagal **tikslų Fišerio kriterijų** (***Fisher's exact test***). Jo statistika naudoja tikslias stebimų dažnių įgijimo tikimybes.

Jei *X* ir *Y* dažnius $(a + b)$, $(a + c)$, $(b + d)$ ir $(c + d)$ laikysime fiksuotais, tuomet, kai *X* ir *Y* yra nepriklausomi, tikimybė, kad porinės dažnių gardelių reikšmės yra *a*, *b*, *c*, *d*, lygi

$$P(a,b,c,d) = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{a!b!c!d!n!}.$$

Tikslios tikimybės $P(a,b,c,d)$ skaičiavimą iliustruosime pavyzdžiu. Sakykime, turime 13 individų; 9 iš jų nustatyta kintamojo *X* I kategorija ($a + b = 9$), 5 – II kategorija ($c + d = 5$). Analogiškai, 6 individams konstatuota kintamojo Y I kategorija (($a + c = 6$), 8 – kintamo *Y* II kategorija ($b + d = 8$). Visos galimos *a*, *b*, *c*, *d* reikšmių kombinacijos ir šių kombinacijų tikimybės pateiktos 7.7 lentelėje.

7.7 lentelė. Visos galimos a, b, c, d reikšmės ir jų įgijimo tikimybės ($a + b = 9$, $c + d = 5$, $a + c = 6$, $b + d = 8$)

1	8	2	7	3	6	4	5	5	4	6	3
5	0	4	1	3	2	2	3	1	4	0	5
$p = 0{,}003$		$p = 0{,}06$		$p = 0{,}28$		$p = 0{,}42$		$p = 0{,}21$		$p = 0{,}028$	

Tikslaus Fišerio kriterijaus statistika skaičiuojama taip: skaičiuojama tikimybė, kad pirmą stulpelį ir pirmą eilutę atitinkančių gardelių reikšmė nėra didesnė už stebėtą:

$$P_F^- = \sum_{z \le a} P(z, b + a - z, c + a - z, d + a - z).$$

Jei $P_F^- < \alpha$ (čia α – parinktas reikšmingumo lygmuo), hipotezę apie kintamųjų nepriklausomumą atmetame. Priešingu atveju skaičiuojame tikimybę:

$$P_F^+ = \sum_{z \ge a} P(z, b + a - z, c + a - z, d + a - z).$$

Jei $P_F^+ < \alpha$, teigiame, kad tarp kintamųjų yra ryšys. Priešingu atveju (jei $P_F^- \ge \alpha$ ir $P_F^+ \ge \alpha$) tvirtinimui, kad kintamieji yra nepriklausomi, prieštarauti nėra pagrindo.

7.5 pavyzdys [4, 15 p.]. Tirtas psichiatrinių – psichinių ir neurotinių – ligonių polinkis į savižudybę. Tirta 20 psichinių ligonių, iš jų polinkis į savižudybę nustatytas dviem (10 %). Iš 20 tirtų neurotinių ligonių polinkis į savižudybę nustatytas 6 (30 %) (7.8 lentelė). Kyla klausimas, ar galima tvirtinti, jog neurotiniai ligoniai į savižudybę linkę labiau nei psichiniai. Hipotezei, kad polinkis į savižudybę nuo ligos pobūdžio nepriklauso, tikrinti negalima naudoti χ^2 kriterijaus, nes dažniai 7.8 lentelėje mažesni nei 5. Minėtai hipotezei tikrinti skaičiuojamos tikimybės, reikalingos tikslaus Fišerio kriterijaus statistikai $P_F^- = P(2, 18, 6, 14) + P(1, 19, 7, 13) + P(0, 20, 8, 12)$:

$P(2, 18, 6, 14) = \frac{8!3\,2!2\,0!2\,0!}{2!6!1\,8!1\,4!2\,0!} = 0{,}09576$; $P(1, 19, 7, 13) = 0{,}02016$; $P(0, 20, 8, 12) = 0{,}00164$,

$P_F^- = 0{,}11756 > 0{,}05$.

Kadangi $P(2, 18, 6, 14) > 0{,}05$, todėl P_F^+ skaičiuoti nėra pagrindo. Remdamiesi tiksliu Fišerio kriterijumi, negalime tvirtinti, kad neurotiniai ligoniai į savižudybę linkę labiau nei psichiniai.

7.8 lentelė. Psichinių ir neurotinių ligonių polinkio į savižudybę tyrimo rezultatai

	Psichiniai ligoniai	Neurotiniai ligoniai	Iš viso
Į savižudybę linkę	2	6	8
Į savižudybę nelinkę	18	14	32
Iš viso	20	20	40

7.5. Kartotinių testų analizė

2×2 lentele pateikiami ir pakartotinio testo, atlikto tiems patiems asmenims, rezultatai. Sakykime, grupei individų atliekamas testas. Rezultatas gali būti teigiamas arba neigiamas. Po tam tikro veiksmo (mokymo, poveikio, praėjus tam tikram laikui) tiems patiems asmenims vėl atliekamas tas pats testas.

Gauti rezultatai pateikiami 7.9 tipo lentele.

7.9 lentelė. Kartotinio testo rezultatai

		Po veiksnio +	Po veiksnio –	Iš viso
Prieš veiksnį	+	*a*	*b*	*a* + *b*
	–	*c*	*d*	*c* + *d*
Iš viso		*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Analizuojant pakartotinio testo rezultatus, aktualu nustatyti, ar poveikis – laikas, gydymas, paskaita, diskusijos, papildomas mokymas – turėjo įtakos nuomonės pasikeitimui ar ligonių sveikatos būklei, apibūdinamai dvinariu kintamuoju. Kitaip tariant, norima nustatyti, ar teigiamo (neigiamo) testo rezultato tikimybė prieš ir po veiksnio buvo vienoda, t. y. ar $\pi_{1+} = \pi_{+1}$ (arba $\pi_{2+} = \pi_{+2}$). Ši lygybė ekvivalenti tikimybių simetriškumui: $\pi_{12} = \pi_{21}$. Taigi, tikrinant, ar testo rezultato tikimybė prieš ir po veiksnio buvo vienoda, reikia patikrinti nulinę hipotezę H_0: „skirtingų rezultatų „+ –“ ir „–+“ tikimybės vienodos“ ($\pi_{12} = \pi_{21}$) su alternatyva „skirtingų rezultatų „+–“ ir „ –+“ tikimybės skiriasi“ ($\pi_{12} \neq \pi_{21}$). Šiai hipotezei tikrinti naudojamas **Maknamaro (*McNemar*)** χ^2 kriterijus su statistika

$$Mc\chi^2 = \frac{(|b-c|-1)^2}{b+c}.$$

Jei nulinė hipotezė teisinga, tai statistikos $Mc\chi^2$ asimptotinis skirstinys yra χ^2 su 1 laisvės laipsniu. Nulinę hipotezę atmetame su reikšmingumu 0,05; jei $Mc\chi^2 > 3{,}84$ arba jei kriterijaus *p* reikšmė mažesnė nei 0,05, čia 3,84 yra χ^2 skirstinio su vienu laisvės laipsniu 0,95 lygio kvantilis.

7.6 pavyzdys [6, 150 p.]. Nagrinėta, kaip pasikeitė medicinos mokyklos pirmo kurso moksleivių nuomonė apie aukštesniųjų kursų moksleivių humoro jausmą po moksleivių renginio „Kitų metų šou“. Renginyje pirmakursiai artimiau susipažino su aukštesniųjų kursų moksleiviais. Apklausta 200 pirmo kurso moksleivių prieš ir po renginio. Apklausos metu pirmakursiai aukštesniųjų kursų moksleivių humoro jausmą vertino teigiamai arba neigiamai (7.10 lentelė).

7.10 lentelė. Pirmakursių apklausos rezultatai

		Po renginio +	Po renginio –	Iš viso
Prieš renginį	+	150	22	172
	–	8	20	28
Iš viso		158	42	200

Iš 7.10 lentelės matyti, kad prieš renginį aukštesniųjų kursų moksleivių humoro jausmą teigiamai vertino 86 % (172 iš 200) pirmakursių, po renginio – tik 79 % (158 iš 200) pirmakursių. Tikrinsime hipotezę, ar renginys turėjo įtakos nuomonės pasikeitimui, t. y. ar pasikeitimų iš teigiamos į neigiamą ir iš neigiamos į teigiamą tikimybės vienodos ($\pi_{12} = \pi_{21}$) su alternatyva $\pi_{12} \neq \pi_{21}$. Maknamaro χ^2 kriterijaus statistika lygi:

$$Mc\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|22-8|-1)^2}{22+8} = \frac{(13)^2}{30} = \frac{169}{30} = 5{,}63;\ p < 0{,}025.$$

Kadangi $p < 0{,}025 < 0{,}05$, nulinę hipotezę atmetame ir tvirtiname, kad renginys turėjo įtakos nuomonei pasikeisti: reikšmingai dažnesnė permaina iš teigiamos nuomonės į neigiamą negu atvirkščiai.

7.6. Nominaliųjų kintamųjų ryšio matai

Naudodami χ^2 kriterijų, 7.3–7.4 skyriuose tikrinome hipotezes apie dviejų kokybinių kintamųjų ryšio (*association*) buvimą ar nebuvimą. Be to, tyrėjui nepakanka vien konstatuoti ryšį tarp kintamųjų, jį domina ir sąryšio tarp kintamųjų stiprumas (*relationship*). Todėl reikalingi kokybiniai kintamųjų ryšio stiprumo rodikliai. Šiame skyriuje pateiksime keletą ryšio tarp nominaliųjų kintamųjų matų. Visiems pateiktiems ryšio matams galioja taisyklė: kuo jie didesni absoliučiu dydžiu, tuo stipresnis ryšys tarp kintamųjų.

Ryšio tarp kokybinių kintamųjų buvimą ar nebuvimą konstatuojame remdamiesi χ^2 statistikos reikšme. Tačiau χ^2 reikšmė (7.1) labai priklauso nuo imties didumo: χ^2 kriterijus, jei *n* labai didelis, fiksuos net labai nežymų dažnių procentinį skirtumą; jei *n* nedidelis, tuomet tik esant gana ryškiems procentiniams dažnių skirtumams, χ^2 kriterijumi konstatuosime ryšį tarp kintamųjų (H_0 atmetimą).

7.7 pavyzdys [6, 153 p.]. Tirta kraujo testo rezultato (įgyjamos reikšmės: teigiamas, neigiamas) priklausomybė nuo lyties. Studijoje dalyvavo 20 000 asmenų: 10 000 vyrų ir 10 000 moterų; rezultatai pateikti 7.11 lentelėje. Analogišką tyrimą atliko ir magistrantūros studijų studentai. Jų organizuotoje studijoje dalyvavo 200 asmenų: 100 vyrų ir 100 moterų; tyrimo rezultatai taip pat pateikti 7.11 lentelėje.

Skaičiuodami χ^2 statistiką, didelės studijos atveju gavome $\chi^2 = 32$, atitinkama *p* reikšmė yra mažesnė nei 0,0001. Taigi galime tvirtinti, kad kraujo tyrimo rezultatai reikšmingai priklauso nuo lyties. Skaičiuodami χ^2 statistiką nedidelės studijos atveju, gavome $\chi^2 = 0{,}32$, atitinkama *p* reikšmė lygi 0,572. Todėl remdamiesi šios studijos rezultatais, negalime tvirtinti, jog kraujo testo rezultatai priklauso nuo lyties.

7.11 lentelė. Kraujo testo rezultatų pasiskirstymas pagal lytį

	Lytis			
Testo rezultatas	Vyrai		Moterys	
	N	%	N	%
Nedidelės apimties studija (n = 200)				
Teigiamas	52	52	48	48
Neigiamas	48	48	52	52
Didelės apimties studija (n = 20 000)				
Teigiamas	5 200	52	4 800	48
Neigiamas	4 800	48	5 200	52

Ryšio tarp nominaliųjų kintamųjų stiprumui vertinti naudosime ψ, Kramerio V, kontingencijos, Julo, Gudmano–Kruskalio λ bei entropijos koeficientus. Šiais koeficientais galima vertinti ir sąryšį tarp tvarkos bei dvinarių kintamųjų, tačiau minėti koeficientai dažniausiai vartojami nominaliųjų kintamųjų atveju.

Koeficientas ψ (*phi*). Jis apibrėžiamas: $\psi = \sqrt{\chi^2 / n}$. ψ dažniausiai naudojamas 2×2 lentelėje, nes didelės apimties lentelės atveju ψ gali viršyti 1. 2×2 lentelės atveju ψ lygus:

$$\psi = \sqrt{\chi^2 / n} = \frac{|ad - bc|}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}. \quad (7.4)$$

Koeficientas ψ (7.4) kinta nuo 0 iki 1. Jei kintamieji visiškai priklausomi (b = 0, c = 0), tada ψ = 1. Jei χ^2 artimas 0, tuomet ir ψ artimas 0.

7.7 pavyzdyje kraujo testo priklausomybė nuo lyties vertinta naudojant didelės (n = 20 000) ir mažos (n = 200) apimties studijos duomenis. χ^2 reikšmės gautos 32 ir 0,32, tačiau abiem atvejais ψ lygus 0,04:

$$\psi = \sqrt{32 / 20000} = \sqrt{0{,}32 / 200} = 0{,}04.$$

Julo asociacijos koeficientas Q (*Yule's Q*) naudojamas 2×2 lentelės atveju. Jis lygus:

$$Q = \frac{ad - bc}{ad + bc};$$

Q kinta nuo –1 iki 1. Visada $|Q| \le \psi$. Q tinka vertinti tiek nominaliųjų, tiek tvarkos kintamųjų ryšio stiprumą.

Kontingencijos koeficientas (*contingency coefficient Pearson's C*):

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

C kinta nuo 0 iki 1. Maksimalią reikšmę 1 pasiekia tik didelėms lentelėms. Kai kurie tyrėjai jį rekomenduoja taikyti tik 5×5 bei didesnės apimties lentelėms.

Kramerio *V* koeficientas (*Cramer's V*):

$$V = \frac{\chi^2}{n \min(r-1, c-1)}.$$

V kinta nuo nulio iki 1. 2×2 lentelėje *V* lygus ψ^2.

Visų minėtų koeficientų, susijusių su χ^2 statistika, trūkumas – sudėtinga jų tikimybinė interpretacija. Gudmano–Kruskalio λ bei entropijos koeficientų tikimybinė interpretacija pateikiama nesunkiai.

Gudmano–Kruskalio (*Goodman–Kruskal*) *λ*. Jis naudojamas visiems kokybiniams kintamiesiems. λ parodo, kiek sumažėja vieno kintamojo klasifikacijos santykinė klaida, žinant kito kintamojo kategoriją. Šio ryšio mato skaičiavimą paaiškinsime pavyzdžiu. 7.12 lentelėje pateikti duomenys apie psichiatrinių ligonių diagnozės pasiskirstymą pagal socialinę klasę [4, 56 p.]. Reikia įvertinti, kiek pacientų socialinės klasės nustatymas padeda prognozuoti diagnozę.

Dažniausiai pasitaikanti tirto ligonių kontingento diagnozė – depresija. Šios diagnozės klaida vertinama dydžiu:

$P_1 = 1 - P$(pacientas serga depresija) = 1 – 91/284 = 0,68.

7.12 lentelė. Psichiatrinių ligonių diagnozės pasiskirstymas pagal socialinę klasę

		Diagnozė (kintamasis *Y*)				
		Neurotiniai	Sergantys depresija	Turintys asmenybės problemų	Šizofrenikai	Iš viso
Socialinė klasė (kintamasis *X*)	1	45	25	21	18	109
	2	10	45	24	22	101
	3	17	21	18	18	74
Iš viso		72	91	63	58	284

Sakykime, ligonio socialinė klasė yra žinoma. Iš 1 socialinės klasės ligonių daugiausia buvo neurotinių, todėl šios diagnozės klaidos tikimybę galima vertinti dydžiu:

$p_1 = 1 - P$(ligonis iš 1 socialinės klasės neurotinis) = 1 – 45/109 = 0,59.

Analogiškai, 2 ir 3 socialinės klasės ligonių dažniausios diagnozės klaidos tikimybė, žinant paciento socialinę klasę, įvertinama taip:

$p_2 = 1 - P$(ligonis iš 2 socialinės klasės serga depresija) = 1 – 45/101 = 0,55;

$p_3 = 1 - P$(ligonis iš 3 socialinės klasės serga depresija) = 1 – 21/74 = 0,72.

Klaidingos dažniausios diagnozės tikimybės įvertis, žinant paciento socialinę klasę, yra:

$P_2 = p_1 \times P$(ligonis iš 1 socialinės klasės) + $p_2 \times P$(ligonis iš 2 socialinės klasės) + p_3 P(ligonis iš 3 socialinės klasės) = 0,59×109/284 + 0,55×101/284 + 0,72×74/284 = 0,61.

Matyti, kad, žinant paciento socialinę klasę, sumažėja dažniausiai pasitaikančios diagnozės klaidos tikimybės įvertis nuo 0,68 iki 0,61. Santykinis dažniausios diagnozės (Y) klaidos sumažėjimas, žinant socialinę klasę (X), įvertinamas dydžiu:

$\lambda_Y = (P_1 - P_2)/P_1 = (0{,}68 - 0{,}64)/0{,}68 = 0{,}103$.

λ_Y yra Gudmano–Kruskalio koeficientas, vertinantis kintamojo Y dažniausios kategorijos klaidingos prognozės tikimybės santykinį sumažėjimą žinant kintamojo X kategoriją. Bendru atveju λ_Y skaičiuojamas:

$$\lambda_Y = \frac{\sum_{i=1}^{r} \max_j (n_{ij}) - \max_j (n_{+j})}{n - \max_j (n_{+j})}.$$

Santykinis dažniausios socialinės klasės prognozės klaidos sumažėjimas, žinant diagnozę, skaičiuojamas analogiškai:

$$\lambda_X = \frac{\sum_{j=1}^{c} \max_i (n_{ij}) - \max_i (n_{i+})}{n - \max_i (n_{i+})}.$$

Koeficientai λ_Y ir λ_X naudojami tuomet, kai vieną kintamąjį galime laikyti nepriklausomu (faktoriumi), kitą – priklausomu (atsaku). Jie kinta tarp 0 ir 1. Jei vienas kintamasis neturi įtakos kitam, tuomet λ_Y ir λ_X yra artimi 0. Bendrą klasifikacijos klaidos sumažėjimą vertiname koeficientu λ:

$$\lambda = \frac{\sum_{i=1}^{r} \max_j (n_{ij}) + \sum_{j=1}^{c} \max_i (n_{ij}) - \max_j (n_{+j}) - \max_i (n_{i+})}{2n - \max_j (n_{+j}) - \max_i (n_{i+})}.$$

7.12 lentelės duomenimis, koeficientas λ lygus:

$\lambda = [(45 + 45 + 21) + (45 + 45 + 24 + 22) - 91 - 109]/(2\times284 - 91 - 109) = 0{,}128$.

Entropijos koeficientas (*uncertainty coefficient, entropy coefficient*) *UC*.

UC pagrįstas entropijos sąvoka. Jis parodo santykinį priklausomo kintamojo entropijos sumažėjimą, žinant nepriklausomo kintamojo kategoriją. Analogiškai koeficientui λ, *UC* skaičiuojamas:

$UC(Y) = [H(X) + H(Y) - H(XY)]/H(Y)$ – Y – priklausomas kintamasis;

$UC(X) = [H(X) + H(Y) - H(XY)]/H(X)$ – X – priklausomas kintamasis;

čia $H(X) = -\sum_{i=1}^{r}(n_{i+}/n)\ln(n_{i+}/n)$ – kintamojo X entropija;

$H(Y) = -\sum_{j=1}^{c}(n_{+j}/n)\ln(n_{+j}/n)$ – kintamojo Y entropija;

$H(XY) = -\sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij}/n)\ln(n_{ij}/n)$ – kintamojo (X, Y) entropija.

Vertinant vieno kintamojo entropijos sumažėjimą ir žinant kito kintamojo kategoriją bei neatsižvelgiant į priklausomybę, naudojamas entropijos koeficientas UC:

$UC = 2[H(X) + H(Y) - H(XY)]/[H(X) + H(Y)]$.

7.12 lentelės duomenimis, $UC(X) = 0{,}0515$, $UC(Y) = 0{,}0408$, $UC = 0{,}0455$.

7.7. Tvarkos kintamųjų ryšio matai

Tarkime, X ir Y yra tvarkos kintamieji, t. y. tiek X, tiek Y įgyjamas reikšmes galima kiekybiškai palyginti tarpusavyje. Tvarkos kintamųjų ryšio stiprumo laipsnis vertinamas suderintų ir nesuderintų porų skirtumu. Sakoma, kad dvi kintamųjų reikšmių poros (x_i, y_i) ir (x_j, y_j) yra suderintos (*concordant*), jei skirtumai $x_i - x_j$ ir $y_i - y_j$ turi vienodą ženklą: arba $x_i < x_j$ ir $y_i < y_j$ arba $x_i > x_j$ ir $y_i > y_j$. Analogiškai, reikšmių poros (x_i, y_i) ir (x_j, y_j) yra nesuderintos (*discordant*), jei $x_i < x_j$ ir $y_i > y_j$ arba $x_i > x_j$ ir $y_i < y_j$. Pažymėkime P – suderintų porų skaičių porinėje dažnių lentelėje, Q – nesuderintų porų skaičių porinėje dažnių lentelėje.

P ir Q skaičiavimą iliustruosime pavyzdžiu. 7.13 lentelėje pateikti duomenys apie polinkio į savižudybę laipsnio pasiskirstymą pagal depresijos klasę [4, 132 p.]. Šio tyrimo duomenims apskaičiuosime suderintų porų skaičių P ir nesuderintų porų skaičių Q.

7.13 lentelė. Polinkio į savižudybę laipsnio pasiskirstymas pagal depresijos klases

		Depresijos skalė (kintamasis Y)				
		<20	20–29	30–39	>39	Iš viso
Polinkio į savižudybę laipsnis (kintamasis X)	1	10	14	8	2	34
	2	2	4	7	2	15
	3	5	9	11	17	42
Iš viso		17	27	26	21	91

Suderintų porų skaičius P skaičiuojamas taip: paeiliui imamos visos porinės dažnių lentelės gardelės ir jose esantys dažniai dauginami iš dažnių, esančių

šios gardelės dešinėje ir apačioje. Po to visos sandaugos sudedamos. Gardelės, esančios tame pačiame stulpelyje ar eilutėje, ignoruojamos. 7.13 lentelės duomenimis, suderintų porų skaičius lygus:

$P = 10\times(4 + 7 + 2 + 9 + 11 + 17) + 14\times(7 + 2 + 11 + 17) + 8\times(2 + 17) + 2\times(9 + 11 + 17) + 4\times(11 + 17) + 7\times17 = 1475.$

Nesuderintų porų skaičius Q nustatomas analogiškai: visose lentelės gardelėse esantys dažniai dauginami iš dažnių, esančių kairėje ir viršuje. Po to visos sandaugos sudedamos. Gardelės, esančios tame pačiame stulpelyje ar eilutėje, ignoruojamos. 7.13 lentelės duomenimis:

$Q = 14\times(2 + 5) + 8\times(2 + 4 + 5 + 9) + 2\times(2 + 4 + 7 + 5 + 9 + 11) + 4\times5 + 7\times(5 + 9) + 2\times(5 + 9 + 11) = 502.$

Pažymėkime X_0 ir Y_0 – kintamojo X ir Y didėjimų skaičių. X_0 (Y_0) skaičiuojamas taip: paeiliui imami dažniai ir dauginami iš tos pačios eilutės (stulpelio) dažnių, esančių dešinėje (apačioje). Po to visos sandaugos sudedamos. 7.13 lentelės duomenimis:

$X_0 = 10\times(14 + 8 + 2) + 14\times(8 + 2) + 8\times2 + 2\times(4 + 7 + 2) + 4\times(7 + 2) + 7\times2 + 5\times(9 + 11 + 17) + 9\times(11 + 17) + 11\times17 = 1096$

ir

$Y_0 = 10\times(2 + 5) + 2\times5 + 14\times(4 + 9) + 4\times9 + 8\times(7 + 11) + 7\times11 + 2\times(2 + 17) + 2\times17 = 591.$

Tvarkos kintamųjų ryšio stiprumui vertinti naudojami Kendalo τ_b ir τ_c, Gudmano–Kruskalio γ, Somero d koeficientai.

Kendalo τ_b ir τ_c (*Kendall's tau statistics, tau b, tau c*). Šios statistikos, vertinančios ryšio tarp dviejų tvarkos kintamųjų stiprumą, remiasi suderintų ir nesuderintų porų skirtumu $K = P - Q$. τ_b ir τ_c yra lygūs:

$$\tau_b = \frac{2K}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}}, \qquad (7.5)$$

$$\tau_c = \frac{2mK}{n^2(m-1)}, \quad m = \min(r, c).$$

τ_b ir τ_c kinta nuo –1 iki 1. Jų trūkumas – neaiški tikimybinė interpretacija.

Pagal apibrėžimą, (7.5) formulėje esantys P, Q, X_0 ir Y_0 yra atsitiktiniai dydžiai, taigi τ_b ir τ_c taip pat yra atsitiktiniai. Jų kitimui vertinti statistiniuose paketuose (SPSS, SAS) pateikiamos ir (7.5) įverčių standartinės paklaidos. Jei tarp kintamųjų X ir Y ryšio nėra, dideliam n τ_b ir τ_c skirstinys yra normalusis su lygiu 0 vidurkiu ir dispersija, priklausančia nuo n.

Pagal 7.13 lentelę:

$$\tau_b = 2(1475 - 502)\Big/\sqrt{(1475 + 502 + 591)(1475 + 502 + 1096)} = 0{,}69,$$

$$\tau_c = 2 \times 3(1475 - 502)\Big/(2(91)^2) = 0{,}35.$$

Gudmano–Kruskalio *γ* (*Goodman and Kruskal's gamma*). Jis lygus:

$\gamma = K/(P + Q)$.

γ = 1, jei nenuliniai dažniai yra tik diagonalėje, einančioje iš viršutinio kairio kampo į apatinį dešinį (7.1 a pav.), t. y. jei didėjant *X*, didėja ir *Y*. γ = –1, jei nenuliniai dažniai yra tik diagonalėje, einančioje iš apatinio kairio kampo į viršutinį dešinį (7.1 b pav.), t. y jei didėjant *X*, *Y* mažėja. Jei *X* ir *Y* yra nepriklausomi, tuomet *γ* artimas 0, nes kai *X* ir *Y* yra nepriklausomi, *P* nedaug skiriasi nuo *Q*.

a)

	Y		
X	15	0	0
	0	15	0
	0	0	15

b)

	Y		
X	0	0	15
	0	15	0
	15	0	0

7.1 pav. Kintamųjų X ir Y ryšys: γ = 1 (a); γ = –1 (b)

γ interpretuojamas kaip tikimybės, kad *X* panašus į *Y*, ir tikimybės, kad *X* nepanašus į *Y*, skirtumas. Pagal 7.13 lentelę γ = (1475 – 502)/(1475 + 502) = 0,49.

2×2 lentelės atveju (7.3 lentelė) turime $P = ad$, $Q = bc$, o $\gamma = (ad - bc)/(ad + bc)$; t. y. koeficientas γ sutampa su Julo asociacijos koeficientu.

Somero koeficientas *d* (*Somer's d*). Sakykime, *X* – nepriklausomas (faktorius), *Y* – priklausomas kintamasis (atsakas). Ryšys tarp *X* ir *Y* vertinamas koeficientu

$d_{YX} = K/(P + Q + Y_0)$.

Analogiškai, $d_{XY} = K/(P + Q + X_0)$ (*X* – atsakas, *Y* – faktorius).

d_{YX} tikimybinė interpretacija analogiška γ. Tarp τ_b ir *d* koeficiento yra teisinga priklausomybė: $(\tau_b)^2 = d_{YX} \times d_{XY}$. Pagal 7.13 lentelę: d_{XY} = (1475 – 502)/(1475 + 502 + 1096) = 0,32.

Matyti, kad polinkio į savižudybę laipsnio priklausomybę nuo depresijos laipsnio vertinant koeficientais τ_b, τ_c, γ ir d_{XY}, konstatuojama teigiama priklausomybė, nors koeficientų reikšmės gana skirtingos. Todėl galima daryti išvadą, jog, didėjant depresijos laipsniui, didėja polinkis į savižudybę.

7.8. Daugiamatės dažnių lentelės

Analizuojant susirgimo ar kitos patologijos priežastis, susiduriama ne tik su vieno, bet ir su kelių kokybinių rodiklių įtaka susirgimui. Sakykime, tirtos dviejų amžiaus grupių – 40–64 m. ir 65–74 m. – ligonių išeitys (mirė, išgyveno) po miokardo infarkto per 90 dienų, priklausomai nuo metoprololio vartojimo. Šio tyrimo rezultatai pateikti 3 matavimais – 2×2×2 dažnių lentele (7.14 lentelė), gauta sujungus dvi 2×2 dažnių lenteles.

7.14 lentelė. 40–64 m. ir 65–74 m. ligonių išeitys po miokardo infarkto per 90 dienų pagal metoprololio vartojimą

Amžius	Metoprololio vartojimas	Mirė	Išgyveno
40–64 m.	metoprololis placebas	21 26	443 427
65–74 m.	metoprololis placebas	19 36	235 208

Analogiškai sudaromos *r*×*c*×*m* ir didesnių matavimų lentelės. Analizuojant 3 ir daugiau matavimų lentelių duomenis, tikrinama hipotezė apie visų kintamųjų bendrą nepriklausomumą, sąlyginį nepriklausomumą (fiksuojant vieno kintamojo kategoriją) bei dažnių ir atskirų kintamųjų reikšmių ryšį (naudojami logtiesiniai modeliai).

7 skyriaus literatūra

1. Agresti A. *Categorical Data Analysis*. 1996. New York: John Wiley & Sons, p. 558.
2. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
3. Babarskienė M., Venclovienė J., Lukšienė D., Šlapikienė B., Milvydaitė I. Susirgusių miokardo infarktu klinikinės rizikos vertinimas 30 parų laikotarpiu. *Medicina*. 2001, 37 tomas, Nr. 12, p. 1418–1424.
4. Everitt B. S. *The Analysis of Contingency Tables*. Second edition. 1992, p. 163.
5. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.
6. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.
7. Ragaišytė N. *Sergančiųjų idiopatine dilatacine, išemine ir hipertenzine kardiomiopatija palyginamieji duomenys*. Daktaro disertacija. 2003. Kaunas.
8. *Kokybinių kintamųjų ryšio matai.* Prieiga per internetą: http://www2.chass.ncu.edu/garson/pa765/association.htm.
9. *Binarinių kintamųjų ryšio matai.* Prieiga per internetą: http://www2.chass.ncu.edu/garson/pa765/assoc2x2.htm.
10. *Tvarkos kintamųjų ryšio matai.* Prieiga per internetą: http://www2.chass.ncu.edu/garson/pa765/assocordinal.htm.
11. *Medicinos statistika.* Prieiga per internetą: http://www.shef.ac.uk/nickfieller/tampere/clinic.pdf.

8 SKYRIUS

Rizikos vertinimas epidemiologinėse studijose

8.1. Kintamieji epidemiologinėse studijose

Nagrinėjant priežastinį ryšį epidemiologijoje, susiduriama su dvejopo pobūdžio kintamaisiais:

- kintamuoju, apibūdinančiu veiksnio, sukeliančio neigiamą poveikį individui, buvimą arba veiksnio stiprumo laipsnį. Jis vadinamas nepriklausomu kintamuoju (*independent variable*), arba faktoriumi (*explanatory variable, factor*);
- kintamuoju, apibūdinančiu kokybinį atsaką į veiksnio sukeltą nepalankų poveikį. Jis vadinamas priklausomu kintamuoju (*dependent variable*), arba atsaku (*outcome*).

Priklausomas kintamasis dažniausiai konstatuoja susirgimo, mirties ar kito nepalankaus įvykio buvimą ar nebuvimą. Nepriklausomas kintamasis, nusakantis poveikį, pavyzdžiui, sergamumui, gali būti:

- mitybos faktorius (dieta, tam tikro produkto vartojimas);
- aplinkos taršos faktorius (radiacija, oro, vandens tarša);
- elgsenos faktorius (rūkymas, alkoholio vartojimas);
- fiziologinės charakteristikos (cholesterolio kiekis ar aukštas arterinis kraujospūdis);
- amžius;
- medicininė intervencija (vaistų vartojimas, skiepai, operacija) ir kt.

Nepriklausomas kintamasis apskritai gali būti tiek kokybinis, tiek kiekybinis.

Paprasčiausias epidemiologinio tyrimo atvejis – nagrinėjamas rizikos veiksnio (RV) poveikis sergamumui ar kitam nepalankiam įvykiui. Čia nepriklau-

somas kintamasis – (RV yra, RV nėra) ir priklausomas kintamasis (serga, neserga) yra dvinariai. Tokio epidemiologinio tyrimo rezultatai pateikiami 2×2 porine dažnių lentele (8.1 lentelė).

8.1 lentelė. Standartinė 2×2 porinė dažnių lentelė, sudaroma vertinant ryšį tarp susirgimo ir RV

		SUSIRGIMAS		
		Yra	Nėra	Iš viso
RIZIKOS	Yra	*a*	*b*	*a* + *b*
VEIKSNYS	Nėra	*c*	*d*	*c* + *d*
Iš viso		*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Lentelės paaiškinimai:
a = sergantys ir turintys RV individai;
b = nesergantys ir turintys RV individai;
c = sergantys be RV individai;
d = nesergantys be RV individai;

a + *b* = visi individai, turintys RV;
c + *d* = visi individai be RV;
a + *c* = visi sergantys individai;
b + *d* = visi nesergantys individai;

a + *b* + *c* + *d* = visi studijoje dalyvavę individai.

Šių duomenų statistinis modelis – dvimatis diskretusis atsitiktinis dydis, kurio skirstinio parametrai – nežinomos tikimybės π_{ij} (8.2 lentelė): π_{11} – tikimybė, kad individas turi RV ir serga; π_{12} – tikimybė, kad individas turi RV ir neserga; π_{21} ir π_{22} – tikimybės, kad be RV individas serga ir neserga. Kohortinėje studijoje, kurioje individai atsitiktinai atrenkami iš populiacijos su RV ir be RV, o po to nustatomas susirgimas, sąlyginė tikimybė $\pi_{1|1} = \pi_{11}/(\pi_1 + \pi_{12})$ yra susirgimo tikimybė populiacijoje su RV, o $\pi_{1|2} = \pi_{21}/(\pi_{21} + \pi_{22})$ – susirgimo tikimybė populiacijoje be RV. Atvejo–kontrolės studijoje, kai daroma atranka iš sergančių ir nesergančių populiacijos, sąlyginės tikimybės $\pi_{1|1}$ ir $\pi_{1|2}$ nėra susirgimo tikimybės populiacijose atitinkamai su RV ir be RV. Čia sąlygines tikimybes $\pi_{11}/(\pi_{11} + \pi_{21})$ bei $\pi_{12}/(\pi_{12} + \pi_{22})$ galima interpretuoti kaip RV tikimybę sergančių ir nesergančių populiacijoje.

8.2 lentelė. Jungtinis (X, Y), vienmatis X ir Y skirstinys (2.8 skyrius)

	Serga	Neserga	Iš viso
Yra RV	π_{11}	π_{12}	π_{1+}
Nėra RV	π_{21}	π_{22}	π_{2+}
Iš viso	π_{+1}	π_{+2}	1,0

Nepriklausomas kintamasis gali turėti ir daugiau negu dvi reikšmes (pvz., nerūko, rūkė anksčiau, rūko ir t. t.). Tokiu atveju jo ir susirgimo priklausomybės tyrimo rezultatai pateikiami $r \times 2$ porine dažnių lentele.

Vertinant ryšį tarp atsako ir faktoriaus, pavyzdžiui, tarp susirgimo ir rizikos veiksnio buvimo ar nebuvimo, epidemiologams aktualu atsakyti į šiuos klausimus:

- nustatyti, ar RV buvimas padidina sergamumą;
- įvertinti susirgimo rizikos laipsnį, atsiradusį dėl RV buvimo (kad būtų galima palyginti atskirų faktorių įtaką).

Pirmo punkto uždavinys – kokybinių kintamųjų nepriklausomumo nustatymas – pateiktas 7.3–7.4 skyriuose. Šiam tikslui naudojamas χ^2 kriterijus bei tikslus Fišerio kriterijus. Antro punkto uždavinys – kokybinių kintamųjų ryšio stiprumo skaičiavimas (7.6–7.7 skyriai). Tačiau 7.6–7.7 skyriuose pateiktų ryšio stiprumo rodiklių reikšmės nenusako tam tikro požymio buvimo sukeltos rizikos susirgimui, todėl šiame skyriuje pateiksime specifinius epidemiologų bei klinicistų rizikos įverčio rodiklius. Epidemiologinėse studijose priežastiniam ryšiui tarp faktoriaus ir atsako nustatyti ir ryšio stiprumui vertinti naudojamos tikimybių π_{ij} funkcijos. Priežastinio ryšio epidemiologijoje analizę pirmiausia sunkina tai, kad vienu metu atsaką veikia keletas nepriklausomų faktorių. Be to, egzistuoja ir šių veiksnių tarpusavio sąveika. Pavyzdžiui, žinoma, jog hipertenziją sąlygoja amžius ir lytis. Šių faktorių tarpusavio sąveika pasireiškia tuo, kad 50 m. ir jaunesniems asmenims – hipertenzija dažnesnė vyrams, o vyresniems nei 50 m. asmenims – hipertenzija būdingesnė moterims. Antra, sunku nustatyti taršos lygį ar trukmę, kuriuos galėtume laikyti rizikos veiksniu susirgimui.

8.2. Rizikos įverčiai kohortinėje ar prospektyvinėje studijose

Kohortinėje ar prospektyvinėje studijose individams, turintiems faktoriaus X (užterštumo lygio, rizikos veiksnio) i-tąjį lygį, sąlyginė tikimybė $\pi_{1|i}$ (2.8 skyrius) yra tikimybė susirgti esant faktoriaus i-tajam lygiui, o $(\pi_{1|i}, 1 - \pi_{1|i})$ – sergamumo sąlyginis skirstinys. Matome, kad šis skirstinys visiškai nusakomas tik sąlygine tikimybe $\pi_{1|i}$. Todėl sąlyginė tikimybė $\pi_{1|i}$ vadinama susirgimo **rizika**. Susirgimo rizikos įvertis yra proporcija (santykinis dažnis) $p_{1/i} = n_{i1}/n_{i+}$; čia n_{i1}, n_{i2} – sergančių ir nesergančių individų su faktoriaus X i-tuoju lygiu, skaičius, $n_{i+} = n_{i1} + n_{i2}$. Toliau šį įvertį vadinsime tiesiog rizika. Sakykime, tarp užterštoje zonoje gyvenančių asmenų a serga, b neserga. Atitinkamai neužterštoje zonoje serga c ir neserga d asmenų. Tuomet susirgimo rizika užterštoje zonoje yra:

$R_{užt.} = a/(a + b) = p_{1/1}$,

rizika neužterštoje zonoje: $R_{neužt.} = c/(c + d) = p_{1/2}$.

Rizikų skirtumas. Lyginant pagal sergamumą dvi kintamojo *X* kategorijas, sakykime, *i* ir *h*, naudojamas proporcijų arba rizikų skirtumas $\pi_{1|i} - \pi_{1|h}$. Šio skirtumo įvertis – rizikų skirtumas, dar vadinamas absoliučiu rizikos sumažėjimu (*absolute risk reduction*, *ARR*), arba atributine rizika (*attributable risk*, *AR*). Atributinė rizika (rizikų skirtumas) lygi $p_{1|i} - p_{1|h}$, jos standartinė paklaida $\sqrt{p_{1|i}(1-p_{1|i})/n_{i+} + p_{1|h}(1-p_{1|h})/n_{h+}}$.

2×2 lentelės atveju turime $p_{1|1} = a/(a + b)$, $p_{1|2} = c/(c + d)$, todėl rizikų skirtumas, arba absoliutus rizikos sumažėjimas, yra:

$$ARR = a/(a + b) - c/(c + d). \tag{8.1}$$

Jei faktorius įgyja dvi reikšmes – užtersta, neužtersta – tuomet *ARR* lygi sergamumo užterštoje ir neužterštoje aplinkoje skirtumui: $ARR = R_{užt.} - R_{neužt.}$. Rizikų skirtumas padeda palyginti sergamumo (ar kito nepalankaus įvykio) dažnumą dviejose populiacijose, pavyzdžiui, gyvenančių užterštoje ir neužterštoje aplinkose, vartojusių vaisto ir jo nevartojusių. *ARR* standartinė paklaida *se*(*ARR*) 2×2 lentelės atveju yra $se(ARR) = \sqrt{ab/(a+b)^3 + cd/(c+d)^3}$. Kai studijos apimtis ganėtinai didelė, *ARR* skirstinys artimas normaliajam. Todėl *ARR* pasikliautinieji intervalai yra $ARR \pm z_{(1+P)/2}\, se(ARR)$.

8.1 pavyzdys. Randomizuotoje prospektyvinėje HOPE studijoje [7] tirtas ramiprilio efektas kardiovaskulinio įvykio (CV) dažniui. Į studiją įtraukti pacientai stebėti 5 metus. Ramiprilio efektas tirtas lyginant CV proporcijas ramiprilio ir placebo grupėse. Tyrimo rezultatai pateikti lentelėje:

	CV įvyko	CV neįvyko
Ramiprilio grupė (n = 4645)	651	3994
Placebo grupė (n = 4652)	826	3826

Santykinis CV dažnis, arba rizika, ramiprilio grupėje lygi 651/4645 = 0,1402 (14,02 %). Atitinkamai CV rizika placebo grupėje: 826/4652 = 0,1776 (17,76 %). Absoliutus rizikų skirtumas lygus:

ARR = 0,1776 – 0,1402 = 0,0374 (3,74 %).

Atlikto tyrimo duomenimis, ramiprilio vartojimas 3,74 % sumažina kardiovaskulinio įvykio per 5 metus riziką. ARR standartinė paklaida lygi:

$$se(ARR) = \sqrt{651\times 3994/(4645)^3 + 826\times 3826/(4652)^3} = 0{,}00756,$$

ARR pasikliautinasis intervalas: 0,0374 ± $(0{,}00756)^{1/2}$ = 0,0374 ± 0,0148, arba [0,02256; 0,05224].

Prospektyvinėse studijose, skirtose gydymo efektui tirti, naudojamas rodiklis ***NNT* = 1/*ARR*** – atvirkštinis *ARR* dydis (*number need to treatment*). *NNT* rodo, kiek pacientų turėtų priklausyti poveikio (*treatment*) (vartojusių vaisto) grupei, kad vaisto vartojimas užkirstų kelią 1 nepalankiam įvykiui (susirgimui). 8.1 pavyzdyje *NNT* lygus:

NNT = 1/0,0374 = 26,7.

Taigi, HOPE studijos duomenimis, iš ramiprilio vartojusių 27 pacientų 5 metų laikotarpiu turėtų būti 1 CV mažiau negu iš 27 pacientų, nevartojusių ramiprilio.

Santykinė rizika (***risk ratio, RR***). Rizikų skirtumas neatspindi, kiek rizika pakito, palyginti su pradine (sergamumu). Pavyzdžiui, skirtumas tarp 0,001 ir 0,01 bei 0,401 ir 0,41 vienodas, tačiau pirmu atveju pokytis kelis kartus didesnis už pradinę riziką, antru atveju pokytis sudaro tik 0,009 % pradinės rizikos (0,401). 2×2 lentelėje santykinė rizika yra sąlyginių tikimybių santykis:

$\pi_{1|1}/\ \pi_{1|2}$.

Santykinė rizika yra neneigiamas dydis. Jei santykinė rizika lygi 1 ($\pi_{1|1} = \pi_{1|2}$), sergamumas nuo faktoriaus nepriklauso.

Santykinės rizikos įvertis *RR* lygus rizikų santykiui:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{p_{1|1}}{p_{1|2}}. \qquad (8.2)$$

RR kinta nuo 0 iki ∞. Jei nepriklausomas kintamasis įgyja dvi reikšmes (RV yra, RV nėra) ir *RR* > 1, tuomet individų su RV sergamumas yra didesnis, palyginti su individais be RV.

Kai individų, dalyvaujančių epidemiologiniame tyrime, nėra daug, *RR* gali būti gerokai didesnis ar mažesnis už 1, nors susirgimas abiejose veiksnio kategorijose vienodai tikėtinas ($\pi_{1|1} = \pi_{1|2}$). Kadangi *RR* yra atsitiktinis dydis, būtina įvertinti jo patikimumą tikrinant nulinę hipotezę „*RR* = 1“ ar skaičiuojant *RR* pasikliautinąjį intervalą. Nustatyta, kad ln(*RR*) asimptotinis skirstinys yra normalusis. Todėl, kai *n* gana didelis, skaičiuojamas rizikos santykio logaritmo pasikliautinasis intervalas:

$$\ln[a(c+d)/c(a+b)] \pm z_{(1+P)/2}se(\ln(RR)); \qquad (8.3)$$

čia *P* – patikimumas, *se*(ln(*RR*)) yra ln(*RR*) standartinė paklaida:

$$\sqrt{1/a - 1/(a+b) + 1/c - 1/(c+d)}.$$

Todėl *RR* pasikliautinasis intervalas gana dideliam *n* yra toks:

$$RR\exp(\pm z_{(1+P)/2}\sqrt{1/a-1/(a+b)+1/c-1/(c+d)}).$$

Standartinam patikimumui $P = 0{,}95$, $z_{(1+P)/2} = z_{0{,}975} = 1{,}96$. Jei *a, b, c, d* mažesni nei 5, skaičiuojamas tikslus *RR* pasikliautinasis intervalas. PI skaičiuojamas SPSS ar EPIINFO statistinių programų paketais.

Pagal santykinės rizikos pasikliautinąjį intervalą daroma išvada, ar RV buvimas reikšmingai didina susirgimo riziką. Jei $RR > 1$ ir apatinė *RR* pasikliautinojo intervalo riba viršija 1, tuomet $\pi_{1|1} > \pi_{1|2}$ (arba $\pi_{1|1}/\pi_{1|2} > 1$), t. y. RV buvimas reikšmingai didina (su duotu patikimumu *P*) susirgimo riziką. Jei apatinis pasikliautinojo intervalo rėžis mažesnis nei 1, negalime prieštarauti teiginiui, kad $\pi_{1|1} = \pi_{1|2}$, t. y. RV buvimas susirgimo rizikos nedidina. Jei veiksnys (faktorius) *X* įgyja reikšmes „yra požymis“, „nėra požymio“ (vietoj „yra RV“, „nėra RV“ 8.1 lentelėje) ir $RR < 1$, tuomet, jei *RR* pasikliautinojo intervalo viršutinis rėžis mažesnis už 1, požymio buvimas reikšmingai mažina susirgimo riziką ($\pi_{1|1} < \pi_{1|2}$). Jei $RR < 1$ ir PI viršutinis rėžis didesnis už 1, prieštarauti rizikų lygybei nėra pagrindo.

8.2 pavyzdys. 8.1 pavyzdyje nustatyta, kad kardiovaskulinio įvykio (CV) rizika ramiprilio ir placebo grupėse yra atitinkamai 0,1402 ir 0,1776. Šių rizikų santykis lygus $RR = 0{,}1402/0{,}1776 = 0{,}789$, taigi, HOPE studijos duomenimis, ramiprilio vartojimas 21,1 % mažina CV tikimybę. Rizikos santykio logaritmo standartinė paklaida lygi:

$$se(\ln(RR)) = \sqrt{1/651-1/(4645)+1/826-1/4652} = \sqrt{0{,}0023} = 0{,}048,$$

RR 95 % pasikliautinasis intervalas lygus $0{,}789\times\exp(\pm 1{,}96\times 0{,}048)$ arba [0,718; 0,867]. Abu pasikliautinojo intervalo rėžiai mažesni už vienetą, $RR < 1$, todėl galima tvirtinti, kad ramiprilio vartojimas reikšmingai mažina kardiovaskulinio įvykio per 5 metus riziką.

Tiek $\ln(RR)$, tiek jo standartinė paklaida neapibrėžti, kai $a = 0$ ar $c = 0$. Todėl vietoj santykinės rizikos įverčio (8.2) naudojamas pataisytas santykinės rizikos įvertis RR^*:

$$RR^* = \frac{(a/(a+b)+0{,}5)}{(c/(c+d)+0{,}5)}.$$

Kai *n* ganėtinai didelis, $\ln(RR^*)$ skirstinį galima laikyti normaliuoju. Tuo remiantis nustatomas rizikos santykio logaritmo pasikliautinasis intervalas:

$$\ln(RR^*) \pm z_{(1+P)/2}((a+0{,}5)^{-1} - (a+b+0{,}5)^{-1} + (c+0{,}5)^{-1} - (c+d+0{,}5)^{-1})^{1/2}.$$

Santykinė atributinė rizika (*attributable risk percent, attributable fraction*). Atributinė rizika (rizikų skirtumas) neparodo, kiek pakito rizika, palyginti su pradine. Todėl rizikos pokyčiui vertinti skaičiuojama santykinė atributinė rizika *AR* %:

$$AR\ \% = \frac{R_{užt.} - R_{neužt.}}{R_{užt.}} \times 100 = \frac{RR - 1}{RR} \times 100;$$

čia $RR = R_{užt}/R_{neužt}$, $R_{užt}$, $R_{neužt}$ – nepalankaus įvykio (susirgimo, mažo gimimo svorio) rizika atitinkamai užterštoje (RV yra) ir neužterštoje (RV nėra) aplinkoje.

Populiacijos atributinė rizika (*population attributable risk*). Ji parodo, kiek visoje populiacijoje padidėjo susirgimo rizika, dalyje populiacijos atsiradus rizikos veiksniui (pvz., aplinkos taršai). Skaičiuojama ir populiacijos atributinė rizika *PAR*, ir populiacijos atributinė rizika procentais *PAR* %:

$$PAR = R_{bendr.} - R_{neužt.}; \quad PAR\ \% = \frac{R_{bendr.} - R_{neužt.}}{R_{bendr.}} \times 100\ .$$

Pateiksime rizikos įverčių skaičiavimo pavyzdį.

8.3 pavyzdys [8, 6 skyrius]. 1986 m. duomenimis, JAV pagal amžių standartizuotas vyrų mirtingumas nuo plaučių vėžio buvo 72,5 iš 100 000 per metus. Standartizuotas pagal amžių rūkančių vyrų mirtingumas nuo plaučių vėžio buvo 191 iš 100 000 per metus, nerūkančių – tik 8,7 iš 100 000 per metus. Remdamiesi pateiktais rodikliais, vertinsime rūkymo sukeltą riziką mirtingumui nuo plaučių vėžio.

Sąlyginės susirgimo plaučių vėžiu tikimybės:

$p_{1|1}$ = 191/100 000 (rūkančių populiacijoje);
$p_{1|2}$ = 8,7/100 000 (nerūkančių populiacijoje).
Atributinė rizika lygi: *AR* = 191/100 000 – 8,7/100 000 = 182,3/100 000.
Atributinė rizika procentais: *AR* % = (191 – 8,7)100/191 = 182 300/191 = 95,4 %.

Santykinė rizika: *RR* = (191/100 000)/(8,7/100 000) = 22.

Populiacijos atributinė rizika: *PAR* = 72,5/100 000 – 8,7/100 000 = 63,8/100 000;

PAR % = (72,5 – 8,7)100/72,5 = 88 %.

Remdamiesi skaičiavimais, galime daryti išvadą, kad rūkymas padidina mirtingumą nuo plaučių vėžio 182,3/100 000 per metus (*AR* = 182,3/100 000). Mirčių nuo plaučio vėžio sumažėtų 95,4 %, jei nebūtų

rūkoma (*AR* % = 95,4 %). Rūkančių vyrų mirtingumas nuo plaučių vėžio 22 kartus didesnis negu nerūkančių (*RR* = 22).

8.3. Rizikos vertinimas atvejo–kontrolės ir momentinėje studijose. Rizikos santykis

Atvejo–kontrolės, momentinėje ar kohortinėje studijose rizikai vertinti apibrėžiamas šansas – sąlyginių tikimybių santykis. Pažymėkime Ω_1 – šansas susirgti populiacijoje su RV ($X = 1$):

$\Omega_1 = \pi_{1|1}/ \pi_{2|1} = \pi_{11}/ \pi_{12}$.

Analogiškai populiacijoje be RV ($X = 0$) šansas susirgti yra:

$\Omega_1 = \pi_{1|2}/ \pi_{2|2}$.

Vietoj šanso susirgti galima skaičiuoti šansą kitam nepalankiam įvykiui (mažam gimimo svoriui, susirgimui atsinaujinti, sveikatai pablogėti).

Ω_1 ir Ω_2 santykis vadinamas **rizikos santykiu (*odds ratio*),** arba šansų santykiu, ir žymimas θ: $\theta = \Omega_1/ \Omega_2$. Rizikos santykis apibrėžiamas per (X, Y) jungtinio skirstinio tikimybes:

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}.$$

θ kinta nuo 0 iki ∞. Jei θ = 1, sergamumas nuo rizikos veiksnio X nepriklauso. Jei 0 < θ <1, susirgimas labiau tikėtinas antroje 8.1 lentelės eilutėje nei pirmoje. Jei θ > 1, susirgimas labiau tikėtinas pirmoje (8.1) lentelės eilutėje nei antroje. Rizikos santykis nuo santykinės rizikos skiriasi nežymiai, jei $\pi_{1|1}$ ir $\pi_{1|2}$ nėra didelės. Tarp rizikos santykio ir santykinės rizikos yra toks ryšys:

(Rizikos santykis) = (Santykinė rizika) $(1 - \pi_{1|2})/(1 - \pi_{1|1})$.

θ įvertis yra:

$$\hat{\theta} = OR = \frac{(a/n)(d/n)}{(b/n)(c/n)} = \frac{ad}{bc}. \qquad (8.4)$$

OR vadinamas rizikos santykiu. *OR* neapibrėžtas, kai $b = 0$ arba $c = 0$. Tuomet vietoj (8.4) įverčio naudojamas pataisytas rizikos santykis:

$OR^* = [(a + 0{,}5)(d + 0{,}5)]/[(c + 0{,}5)(d + 0{,}5)]$.

Apskaičiavus *OR*, būtina įvertinti jo reikšmingumą. Tiek *OR*, tiek OR^* logaritmai turi asimptotinį normalųjį skirstinį. Todėl gana dideliam n rizi-

kos santykio logaritmo ln(*OR*) pasikliautinasis intervalas skaičiuojamas analogiškai (8.3):

$\ln(OR) \pm z_{(1+P)/2} se(\ln(OR))$;

čia $se(\ln(OR)) = (1/a + 1/b + 1/c + 1/d)^{1/2}$. *OR* pasikliautinasis intervalas, kai $P = 0{,}95$, yra: $OR \times \exp(\pm 1{,}96\sqrt{1/a + 1/b + 1/c + 1/d})$. Jei *a, b, c, d* mažesni nei 5, patartina naudoti SPSS ir EPIINFO statistiniais paketais skaičiuojamą tikslų *OR* pasikliautinąjį intervalą.

Jei bent vienas iš dažnių *a, b, c, d* lygus nuliui, tuomet vietoj įverčio *OR* naudojamas OR^*, o jo standartinė paklaida ir pasiklautinasis intervalas skaičiuojamas kaip *OR* įverčio atveju, tik prie dažnių *a, b, c, d* pridedama po 0,5.

Išvados apie *OR* reikšmingumą, remiantis pasikliautinuoju intervalu, daromos analogiškai išvadoms apie santykinę riziką. Jei $OR > 1$ ir *OR* pasikliautinojo intervalo apatinis rėžis ne mažesnis už 1, daroma išvada, kad $\theta > 1$, arba populiacijoje su RV šansas susirgti reikšmingai didesnis nei populiacijoje be RV. Jei pasikliautinojo intervalo apatinis rėžis mažesnis už 1, nėra pagrindo prieštarauti, kad sergamumas nuo RV nepriklauso ($\theta = 1$). Jei $OR < 1$ ir viršutinis pasikliautinojo intervalo rėžis ne didesnis už 1, tuomet $\theta < 1$, arba reikšmė $X = 1$ reikšmingai mažina šansą susirgti, palyginti su $X = 0$.

Rizikos santykiu atvejo–kontrolės, kohortinėje ar momentinėje studijose vertinama veiksnio įtaka susirgimui, letaliai baigčiai ir pan. Klinikinio eksperimento studijoje veiksnio įtaka vertinama naudojant santykinę riziką.

8.4 pavyzdys. Tirta amžiaus įtaka letaliai baigčiai po miokardo infarkto 30 parų laikotarpiu. Nustatyta, kad per 30 parų mirė 9 (2,8 %), išgyveno 315 ligonių, turinčių 65 m. ir jaunesnių. Vyresnių nei 65 m. ligonių mirė 15 (11,2 %), išgyveno – 119. Amžiaus per 65 m. keliamą riziką mirčiai 30 parų laikotarpiu įvertinsime rizikos santykiu *OR*. Tyrimo duomenimis, $a = 15$; $b = 119$; $c = 9$; $d = 315$ (8.1 lentelės žymėjimai):

$$OR = \frac{ad}{bc} = \frac{15 \times 315}{9 \times 119} = 4{,}41.$$

Nustatėme, kad ligonių, vyresnių nei 65 m., letalios baigties 30 parų laikotarpiu šansas 4,41 karto didesnis negu 65 m. ir jaunesnių. *OR* patikimumui apskaičiuosime 95 % pasikliautinojo intervalo ribas:

$se(\ln(OR)) = (1/9 + 1/315 + 1/15 + 1/119)^{1/2} = (0{,}18936)^{1/2} = 0{,}435$;

$OR \times \exp(-1{,}96 \times 0{,}435) = 4{,}41 \times 0{,}426 = 1{,}88$; $OR \times \exp(1{,}96 \times 0{,}435) = 4{,}41 \times 2{,}346 = 10{,}35$.

Rizikos santykio 95 % PI yra [1,88; 10,35]; abi ribos viršija vienetą – taigi rizikos santykis už 1 reikšmingai didesnis.

8.5 pavyzdys. Tirta įvairių ligonio požymių įtaka letaliai baigčiai (LB) po miokardo infarkto 30 parų laikotarpiu. 8.3 lentelėje pateiktas šių požymių (reikšmės *yra*, *nėra*) rizikos santykis su 95 % pasikliautinaisiais intervalais. Iš lentelės matyti, kad didžiausią riziką LB sukėlė prieširdžių arba skilvelių virpėjimas – *OR* = 7,71. Reikšmingą įtaką LB turėjo ir amžius bei ST pakilimas aukščiau kaip 3 mm – visų šių rodiklių rizikos santykio PI ribos viršija 1. Moteriškoji lytis ir laidumo sutrikimai nedidino LB rizikos, nes šių rodiklių rizikos santykis nuo 1 reikšmingai nesiskyrė – vienetas pateko į PI ribas.

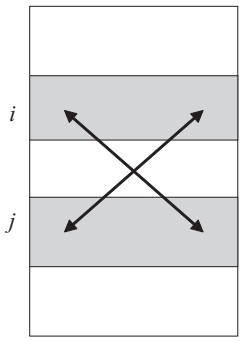
8.3 lentelė. Ligonių, persirgusių MI, letalios baigties 30 parų laikotarpiu rizikos santykis su 95 % pasikliautinaisiais intervalais

Požymis	OR	95 % PI
Moteriškoji lytis	1,33	0,54–3,29
Amžius > 65 m.	4,41	1,88–10,4
ST pakilimas > 3 mm	3,43	1,49–7,87
Prieširdžių ar skilvelių virpėjimas	7,71	3,26–18,3
Laidumo sutrikimai	1,24	0,85–1,81

Pagal sergamumą dviem kintamojo X kategorijoms, sakykime, i ir j, lyginti naudojamas rizikos santykis tarp i-tos ir j-tos eilutės (8.1 pav.):

$$OR_{ij} = \frac{n_{i1} n_{j2}}{n_{i2} n_{j1}}. \tag{8.5}$$

Šis rizikos santykis parodo, kiek kartų susirgimas labiau tikėtinas i-tame faktoriaus X lygyje, palyginti su faktoriaus j-tuoju lygiu. Analogiškai (8.5) formulei apibrėžiama ir santykinė rizika tarp i-tos ir j-tos eilučių RR_{ij}. Išvados apie OR_{ij} (ar RR_{ij}) taip pat daromos remiantis pasikliautinaisiais intervalais.



8.1. pav. Rizikos santykis tarp i-tos ir j-tos eilučių

8.4. Rizikos analizė *r*×2 lentelėje

Vertindami veiksnio įtaką susirgimui ar kitam įvykiui, susiduriame ir su atveju, kai veiksnį apibūdinantis kintamasis yra ne dvinaris, o kiekybinis, tvarkos ar nominalusis. Pavyzdžiui, tiriant rūkymo įtaką mirčiai nuo IŠL 5 metų laikotarpiu, pagal rūkymo įtaką sveikatai ligonį galima chrakterizuoti: nerūko, rūkė anksčiau, rūko. Rūkymą taip pat galima vertinti pagal surūkytų cigarečių kiekį. Pateiksime rizikos analizės pavyzdį.

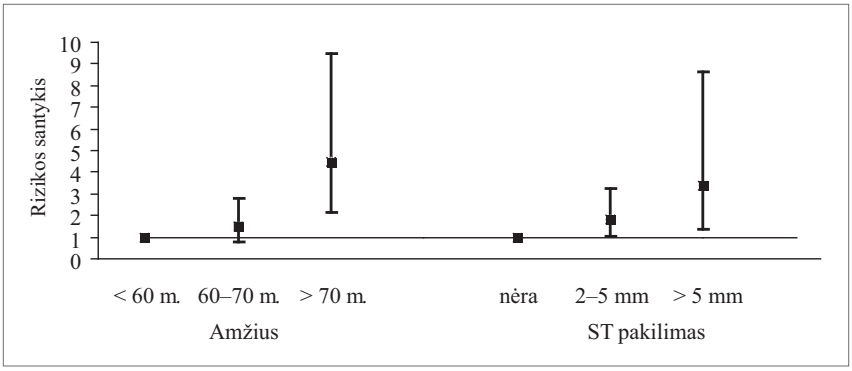
8.6 pavyzdys. Tiriant amžiaus įtaką letaliai baigčiai po stacionarizacijos dėl miokardo infarkto ar nestabiliosios KA 1 metų laikotarpiu, konstatuota, kad iš ligonių iki 60 m. per 1 metus mirė 6,7 %, iš 60–70 m. amžiaus ligonių – 9,7 %, vyresnių nei 70 m. ligonių – 24,6 % (8.4 lentelė). Analogiškai, iš ligonių, neturėjusių ST pakilimo stacionarizuojant, mirė 6,8 %, iš ligonių su 2–5 mm ST pakilimu – 11, 8%, o iš ligonių, kurių ST pakilimas viršijo 5 mm – 20 %. Abiem nagrinėtais atvejais faktorius (veiksnys) yra tvarkos kintamasis, gautas pertvarkius kiekybinį kintamąjį į kokybinį. Mirusių ir išgyvenusių dažnių išsidėstymas amžiaus grupėse ir ST pakilimo laipsniais pateiktas 3×2 lentele (8.4 lentelė). Analizuojant šių veiksnių įtaką mirštamumui 1 metų laikotarpiu, aktualu įvertinti mirties rizikos augimą didėjant amžiui ir ST pakilimo laipsniui.

8.4 lentelė. Ligonio amžiaus ir ST pakilimo stacionarizuojant rizika letaliai baigčiai vienerių metų po stacionarizacijos laikotarpiu

Kintamasis	Kintamojo reikšmė	Letali baigtis N	%	Išgyveno N	%	Rizikos santykis	95 % PI
	< 60 m.	22	6,7	304	93,3	1,0	
Amžius	60–70 m.	22	9,7	204	90,3	1,490	0,80–2,78
	> 70 m.	14	24,6	43	75,4	4,499	2,14–9,45
ST pakilimas	nėra	23	6,8	313	93,2	1,0	
	2–5 mm	28	11,8	209	88,2	1,823	1,02–3,25
	> 5 mm	7	20,0	28	80,0	3,402	1,34–8,63

Vertinant kelių lygių (kategorijų) veiksnio įtaką sergamumui ar mirtingumui, viena veiksnio kategorija, tarkime, i_0, laikoma referentine ir jos rizika ar šansas prilyginamas 1. Jei veiksnys yra tvarkos kintamasis, referentine dažniausiai laikoma mažiausia (pirmoji), pvz., „nerūko“, „amžius < 60 m.“ Kitų veiksnio kategorijų rizika ar susirgimo šansas lyginami su rizika ar šansu, nustatytu referentinėje kategorijoje – skaičiuojame RR_{ii0} ar OR_{ii0} su pasikliautinaisiais intervalais. Referentine kartais nustatoma ir ta veiksnio kategorija, kurioje daugiausia individų.

8.4 lentelėje ir 8.2 pav. pateiktas mirties 1 metų po stacionarizacijos dėl MI ar NKA laikotarpiu rizikos kitimas pagal amžių ir ST pakilimo laipsnį. Čia referentinės kategorijos – „amžius < 60 m." ir „nėra ST pakilimo": $i_0 = 1$ (8.5 formulėje $j = 1$). Vertindami amžiaus sukeltą riziką, nustatėme OR_{21} =1,49; OR_{31} = 4,5. Šiuos rezultatus galime interpretuoti taip: 60–70 m. ligonių galimybė (rizika) mirčiai 1 metų laikotarpiu po susirgimo MI ar NKA yra 1,49 karto didesnė nei jaunesnių negu 60 m. ligonių. Tačiau šis rizikos padidėjimas nėra reikšmingas, nes rizikos santykio pasikliautinojo intervalo apatinė riba lygi 0,8 ir yra mažesnė už 1. Vyresnių nei 70 m. ligonių galimybė mirti 4,5 karto (ir reikšmingai, nes PI reikšmės viršija 1) didesnė negu jaunesnių nei 60 m. ligonių. Vertinant ST pakilimo stacionarizuojant riziką, nustatyta, kad ST pakilimas 2–5 mm letalios baigties 1 metų laikotarpiu galimybę didina 1,82 karto, ST pakilimas daugiau kaip 5 mm – 3,4 karto, palyginti su ligoniais, neturėjusiais ST pakilimo.



8.2 pav. Letalios baigties rizikos santykis su pasikliautinaisiais intervalais, priklausomai nuo ligonio amžiaus ir ST pakilimo laipsnio

8.5. Koreguotas rizikos santykis

Tiek rizikos santykis, tiek kiti rizikos rodikliai leidžia įvertinti vieno RV buvimo įtaką susirgimui ar kitam dvinariam atsakui. Tačiau vienu metu atsaką veikia ne vienas, o keli nepriklausomi faktoriai, todėl aktualu įvertinti atskiro RV sukeltą susirgimo riziką, izoliavus pašalinių veiksnių įtaką. Tam skaičiuojamas koreguotas (standartizuotas) rizikos santykis (*adjasted OR*, *Mantel–Haenszel OR*) OR_K. Aptariant kelių faktorių įtaką sergamumui, rizikos santykis *OR*, apibrėžtas 8.3 skyriuje, vadinamas izoliuotu rizikos santykiu (*crude OR*).

Sakykime, visi tirti individai suskirstyti į grupes pagal kokybinio faktoriaus X_1 reikšmes. Tegu faktoriaus X_1 i-tąją reikšmę turi a_i sergančių individų su RV, b_i nesergančių individų su RV bei c_i ir d_i sergančių ir nesergančių individų be RV. Koreguotas rizikos santykis, izoliuojantis faktoriaus X_1 įtaką RV sukeltai rizikai, skaičiuojamas taip:

$$OR_K = \left(\sum_i \frac{a_i d_i}{n_i} \right) / \left(\sum_i \frac{b_i c_i}{n_i} \right), \; n_i = a_i + b_i + c_i + d_i.$$

Paketu EPIINFO skaičiuojamas ir šio rodiklio pasikliautinasis intervalas.

8.4 lentelėje pateikti duomenys apie 15–24 m., 25–39 m. ir ≥ 40 m. individų, gyvenančių didelėje bei mažoje oro taršos zonoje, sergamumą bronchitu. Įvertinsime oro taršos sukeltą bronchito riziką atmetus amžiaus įtaką.

8.4 lentelė. Sergamumas bronchitu, priklausomai nuo oro taršos amžiaus grupėse

Amžius	Oro taršos lygis	Bronchitas	
		Serga	Neserga
15–24 m.	aukštas	20	382
	žemas	9	214
25–39 m.	aukštas	10	172
	žemas	7	120
≥ 40 m.	aukštas	12	327
	žemas	6	183

Koreguotas pagal amžiaus grupes rizikos santykis lygus:

$$OR_K = \frac{6{,}85 + 3{,}88 + 4{,}16}{5{,}50 + 3{,}90 + 3{,}72} = 1{,}13.$$

OR_K pasikliautinasis intervalas yra [0,84; 1,42]; taigi negalima tvirtinti, kad aukštas oro taršos lygis padidina sergamumą bronchitu.

8 skyriaus literatūra

1. Agresti A. *Categorical Data Analysis*. 1996. New York: John Wiley & Sons, p. 558.
2. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
3. Babarskienė M., Venclovienė J., Lukšienė D., Šlapikienė B., Milvydaitė I. Susirgusių miokardo infarktu klinikinės rizikos vertinimas 30 parų laikotarpiu. *Medicina*. 2001, 37 tomas, Nr. 12, p. 1418–1424.
4. Everitt B. S. *The Analysis of Contingency Tables*. Second edition. 1992, p. 163.
5. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.

6. Grabauskas V. J., Misevičienė I., Padaiga Ž. ir kt. *Fundamentinė epidemiologija*. 2003. Kaunas: KMU, 144 p.
7. The Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *The New England Journal of Medicine*. 2000, 342, p. 145–153.
8. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.

9 SKYRIUS

Koreliacinė analizė

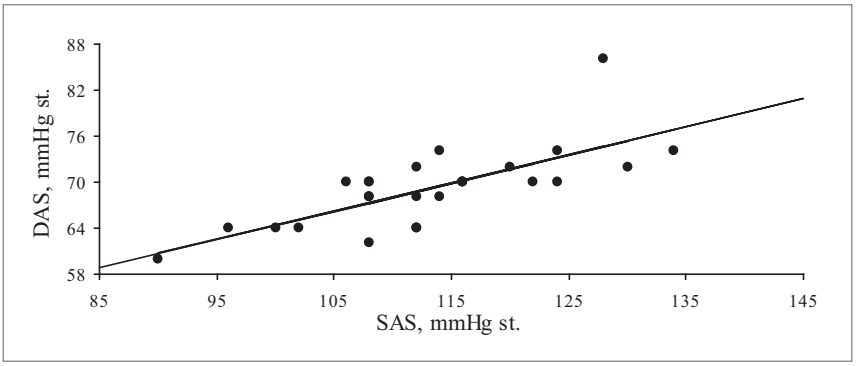
Vienas iš daugelio medikų-tyrėjų uždavinių – nustatyti, įvertinti ir paaiškinti ryšį tarp ligonį charakterizuojančių klinikinių rodiklių. Šiame skyriuje analizuosime gautų kiekybinių kintamųjų reikšmių tarpusavio sąryšį ir jo kiekybinius įverčius, kuriais remiantis daromos išvados ir apie pačių kintamųjų ryšį (*relationship*). Tačiau statistiniais metodais konstatuotas ryšys tarp kintamųjų nepaaiškina priežastinio ryšio tarp jų – šį faktą reikia pagrįsti ir mediko-tyrėjo išvadomis.

9.1. Dviejų kiekybinių kintamųjų ryšio aspektai

Fizikos moksluose tarp dviejų kintamųjų dažnai konstatuojamas funkcinis ryšys. Pavyzdžiui, tarp srovės ir įtampos esančią tiesinę priklausomybę nusako Omo dėsnis. Dėl sudėtingos žmogaus, kaip biologinės būtybės, sandaros neįmanoma aptikti funkcinio ryšio tarp dviejų individą charakterizuojančių rodiklių. Tarp panašaus pobūdžio rodiklių stebimas tik statistinis (tikimybinis) ryšys. Pavyzdžiui, 9.1 pav. pateikta 26 jaunų sveikų suaugusių asmenų SAS ir DAS skaidos diagrama (3.4 lentelės duomenys). Diagramoje pastebima: didėjant SAS, DAS reikšmės taip pat didėja, t. y. tikėtina, kad, esant didesniam SAS, ir DAS bus didesnis. Tai ir yra statistinis ryšys. Be to, SAS ir DAS reikšmės išsidėsčiusios arti tiesės – pastebima tiesinio didėjimo tendencija.

Apskritai analizuojant nustatytų kiekybinių rodiklių X ir Y reikšmių kitimą ir tarpusavio sąryšį, išskiriami trys aspektai:

- tendencija;
- forma;
- ryšio stiprumo laipsnis.



9.1 pav. 26 jaunų sveikų suaugusių asmenų SAS ir DAS skaidos diagrama

Ryšio tarp *X* ir *Y* tendencija – tai *X* ir *Y* kitimo pobūdis: kaip kinta *Y*, didėjant ar mažėjant *X*. Pavyzdžiui, didėjant *X*, 9.2 pav. (2) ir (3) pastebima *Y* didėjimo, o (4) ir (5) – mažėjimo tendencija. Jei visoms *X* reikšmėms ryšio tarp *X* ir *Y* tendencija vienoda, sakoma, kad tarp šių kintamųjų yra monotoninis ryšys (9.2 pav.). Jei tiriamoje *X* reikšmių aibėje ryšio tarp *X* ir *Y* reikšmių tendencija skirtinga, turime nemonotoninį ryšį. Nemonotoninio ryšio atvejis pateiktas 9.3 pav. – didėjant *X*, iki *a* taško *Y* turi tendenciją didėti (netiesiškai), o už *a* taško – mažėti.

Statistinio ryšio forma gali būti:

- tiesinė;
- netiesinė (kvadratinė, eksponentinė, logaritminė).

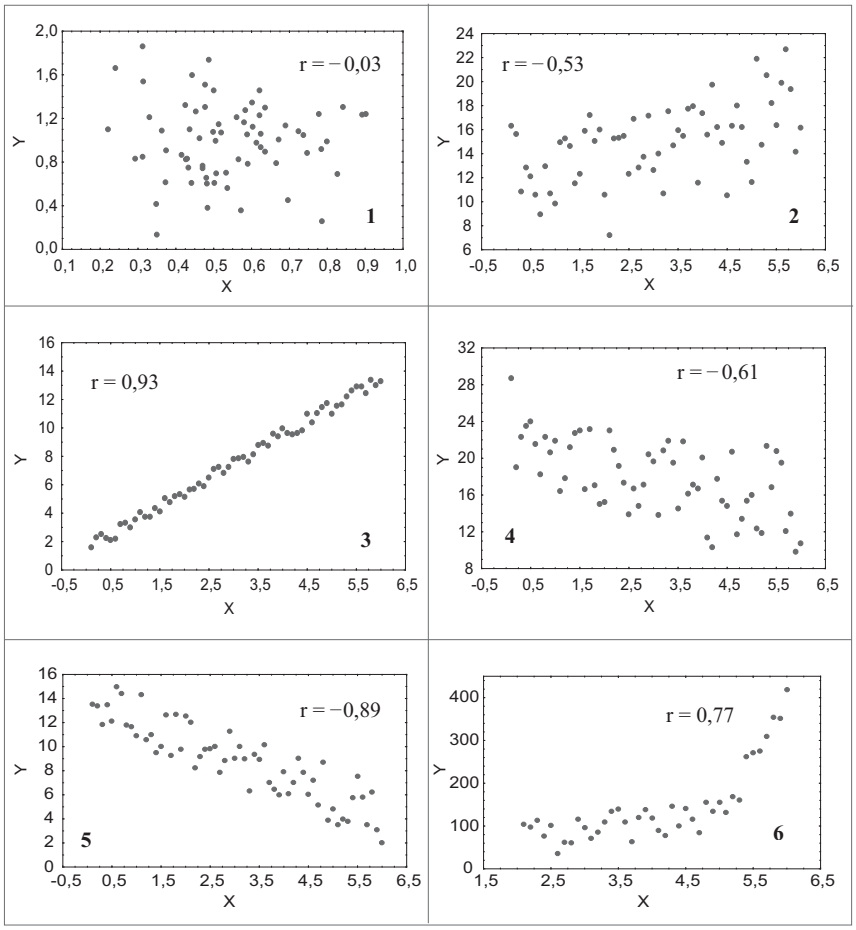
Tiesinė ir netiesinė funkcinė priklausomybė tarp kintamųjų *X* ir *Y* pateikta 9.4 pav.: (a) tiesinė; (b) netiesinė.

Skaidos diagrama analizuojant medicinos ar biologijos kiekybinių kintamųjų *X* ir *Y* reikšmių tarpusavio kitimą, labai tikėtinas vienas iš 9.2 pav. pateiktų atvejų. Remdamiesi 9.2 pav. pavaizduotomis skaidos diagramomis, galime daryti išvadą, kad:

1) ryšio tarp *X* ir *Y* reikšmių nėra;
2) stebima tendencija: didėjant *X*, didėja ir *Y*;
3) beveik tiesinis ryšys: didėjant *X*, didėja ir *Y*;
4) stebima tendencija: didėjant *X*, mažėja *Y*;
5) gana glaudus tiesinis ryšys: didėjant *X*, mažėja *Y*;
6) tarp *X* ir *Y* reikšmių yra netiesinis ryšys.

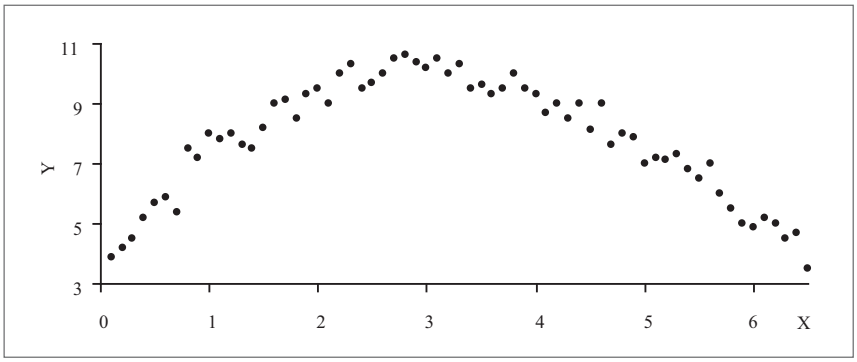
Nagrinėjant 9.2 pav. (2) ir (3) bei (4) ir (5) atvejus, pastebima vienoda (monotoninė) kitimo tendencija bei tiesinė ryšio forma; skiriasi tik *X*

ir *Y* ryšio stiprumo laipsnis. 9.2 pav. (2) stebima tik kitimo tendencija, o 9.2 pav. (3) – jau beveik tiesinis ryšys. Todėl reikalingas ryšio tarp kintamųjų reikšmių glaudumą vertinantis rodiklis, pagal kurį būtų daroma išvada apie pačių kintamųjų ryšį bei nustatoma šio ryšio stiprumo kiekybinė išraiška.

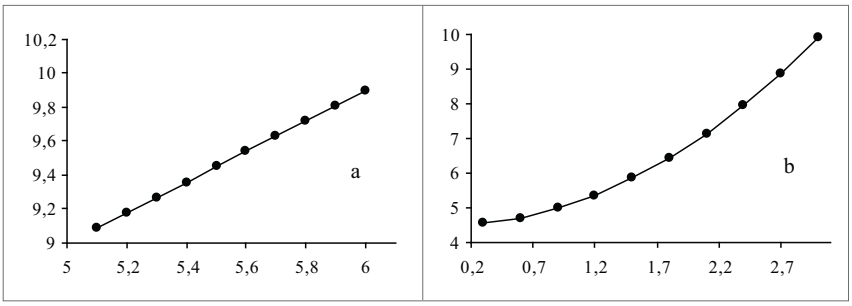


9.2 pav. Ryšio tarp kiekybinių kintamųjų atvejai:

1) tarp X ir Y ryšio nėra;
2) stebima tendencija: didėjant X, didėja ir Y;
3) beveik tiesinis ryšys: didėjant X, didėja ir Y;
4) stebima tendencija: didėjant X, mažėja Y;
5) gana glaudus tiesinis ryšys: didėjant X, mažėja Y;
6) tarp X ir Y yra eksponentinis ryšys: didėjant X, Y didėja eksponentiškai



9.3 pav. Nemonotoninis ryšys (antro laipsnio polinomas) – kai didėjant X, X<3, Y didėja (netiesiškai), o kai X>3 – Y mažėja



9.4 pav. Tiesinė (a) ir netiesinė funkcinė priklausomybė (b) tarp kintamųjų

9.2. Koreliacijos sąvoka, koreliacijos koeficientas

Koreliacija suprantama kaip tiesinio ryšio tarp kiekybinių kintamųjų analizė. Statistinių metodų, skirtų tiesinio monotoninio ryšio tarp kiekybinių kintamųjų analizei, visuma vadinama koreliacine analize. Pagrindinis koreliacinėje analizėje naudojamas rodiklis – koreliacijos koeficientas *r*, apskaičiuotas naudojant kintamųjų imties reikšmes. Aptarsime šio koreliacijos koeficiento skaičiavimo ir interpretavimo aspektus.

Išvados apie kiekybinių rodiklių *X* ir *Y* tarpusavio ryšį remiasi gautų duomenų – dvimačio kintamojo (*X*, *Y*) imties (x_1, y_1), (x_2, y_2) ... (x_n, y_n) analize. Šių reikšmių kitimą bei išsibarstymą plokštumoje sunku paaiškinti priežastiniu ryšiu, todėl daroma prielaida, kad (x_1, y_1) ... (x_n, y_n) yra nepriklausomi dvimačiai atsitiktiniai dydžiai, turintys tą patį tolydųjį populiacijos skirstinį. Taigi dvimačio kintamojo (*X*, *Y*) statistinis modelis – dvimatis ats. d., todėl

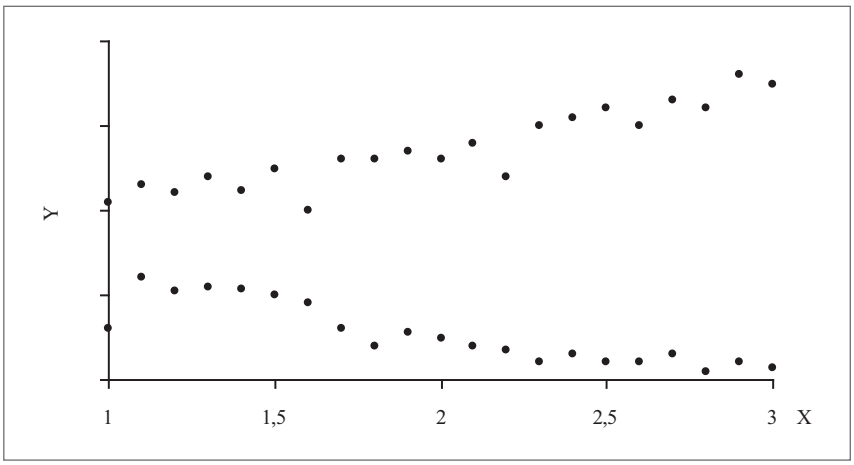
ir tiesinis ryšys tarp X ir Y turi būti vertinamas tiesinės priklausomybės tarp atsitiktinių dydžių rodikliu. Kaip minėta 2.9 skyriuje, tiesinės priklausomybės tarp dviejų ats. dydžių matas – koreliacijos koeficientas tarp jų ρ. Tai bedimensis dydis, kintantis nuo –1 iki +1. Pagal ρ reikšmes daroma išvada apie atsitiktinių dydžių tiesinę priklausomybę. Sakykime, ρ – koreliacijos koeficientas tarp ats. dydžių ξ ir η. Jei ρ yra lygus 1 ar –1, tarp kintamųjų yra tiesinė priklausomybė, kurią galima išreikšti lygtimi: $\eta = a\xi + b$. Jei $\rho = 1$, $a > 0$ – didėjant ξ, didėja η; 9.4 pav. a); jei $\rho = -1$, $– a < 0$ – didėjant ξ, mažėja η. Jei ξ ir η yra nepriklausomi ats. d., tuomet $\rho = 0$. Jei $\rho \neq 0$, sakoma, kad tarp ats. dydžių ξ ir η yra koreliacija, arba jie koreliuoja; jei $\rho = 0$, sakoma, kad ats. dydžiai ξ ir η nekoreliuoja. Analogiškai tiesinio ryšio arba koreliacijos tarp kiekybinių kintamųjų X ir Y stiprumo matas yra koreliacijos koeficientas ρ, nusakomas (X, Y) populiacijos skirstiniu.

Teorinis dydis – populiacijos koreliacijos koeficientas ρ nėra žinomas; ρ įverčiai yra imties koreliacijos koeficientai, skaičiuojami naudojant (X, Y) imties reikšmes. Imties koreliacijos koeficientas r, apskaičiuotas naudojant konkrečios imties reikšmes, pasižymi tomis pačiomis savybėmis, kaip ir populiacijos koreliacijos koeficientas ρ:

- $-1 \leq r \leq 1$;
- jei $r > 0$, didėjant X reikšmėms, Y reikšmės turi tendenciją didėti (9.2 pav. (2) ir (3)); be to, kuo glaudesnis ryšys tarp X ir Y reikšmių, tuo r didesnis; 9.2 pav. (3) r didesnis nei (2);
- jei $r < 0$, tai didėjant X reikšmėms, Y reikšmės turi tendenciją mažėti (9.2 pav. (4) ir (5)); be to, kuo glaudesnis ryšys tarp X ir Y reikšmių, tuo r artimesnis – 1, arba $|r|$ didesnis; 9.2 pav. (4) $|r|$ mažesnis nei (5);
- jei tarp X ir Y reikšmių tiesinio ryšio nėra (9.2 pav. (1)), r artimas 0.

Kad pagal r reikšmę, apskaičiuotą naudojant konkrečius X ir Y matavimus, galėtume daryti išvadą ir apie ryšį tarp kintamųjų, turi galioti koreliacijos koeficiento taikymo prielaidos. Atkreiptinas dėmesys, kad:

- koreliacijos koeficientas neatspindi nemonotoninio ryšio. Nors tarp kintamųjų reikšmių yra labai glaudus netiesinis ryšys (9.3 pav.), r artimas nuliui.
- tiesiniam ryšiui tarp kintamųjų vertinti koreliacijos koeficientas nenaudotinas ir tuo atveju, jei (X, Y) imties reikšmės nėra generuotos to paties atsitiktinio dydžio. Jei (X, Y) reikšmės yra imčių iš skirtingų populiacijų mišinys, r gali būti artimas 0, nors atskirose populiacijose stebimas tiesinis ryšys tarp X ir Y reikšmių (9.5 pav.).



9.5 pav. r ≈ 0; X ir Y reikšmės generuotos dviejų dvimačių atsitiktinių dydžių

Pagal prielaidas (*X*, *Y*) populiacijos skirstiniui, kintamųjų tiesiniam ryšiui (populiacijos koreliacijos koeficientui ρ) vertinti naudojami Pirsono, Spirmeno, Kendalo koreliacijos koeficientai (atitinkamai žymimi r, r_S, r_K). Pirsono koreliacijos koeficientui skaičiuoti naudojamos kintamųjų (*X*, *Y*) atsitiktinės imties reikšmės. Spirmeno koreliacijos koeficientas skaičiuojamas naudojant ne imties reikšmes, o jų rangus – eilės numerį variacinėje sekoje. Todėl Spirmeno koreliacijos koeficientas vadinamas ranginiu. Kurį koreliacijos koeficientą panaudoti, sprendžiama remiantis turimu duomenų kiekiu, duomenų statistiniu modeliu bei norima koreliacijos koeficiento interpretacija.

Pagal koreliacijos koeficiento, apskaičiuoto turimai imčiai, absoliutaus dydžio reikšmes apibrėžiamas kokybinis ryšio įvertis – ryšys silpnas, vidutinis, stiprus. Sakoma, kad tarp kintamųjų yra silpnas ryšys, jei $|r| \leq 0{,}3$, vidutinio stiprumo ryšys, jei $0{,}3 < |r| \leq 0{,}6$, ir stiprus ryšys, jei $|r| > 0{,}6$. Ryšio tarp kintamųjų glaudumui matuoti naudojamas ir koreliacijos koeficiento kvadratas r^2.

Apskaičiavus koreliacijos koeficientą r, aktualu įvertinti ir jo patikimumą – kaip r skiriasi nuo ρ (populiacijos (*X*, *Y*) skirstinio koreliacijos koeficiento). Koreliacijos koeficientas r, apskaičiuotas naudojant atsitiktinės imties reikšmes – ats. d. (x_i, y_i), yra atsitiktinis dydis. Ats. dydžio r vidurkis lygus ρ, o r dispersija – kitimo įvertis – priklauso nuo imties dydžio: kuo n mažesnis, tuo didesnė r dispersija. Pavyzdžiui, atsitiktinai plokštumoje dedant 3 ar 4 taškus, labai tikėtina, kad jie išsidėstys arti tiesės ir $|r|$ bus artimas 1, nors $\rho = 0$. Todėl prieš konstatuojant ryšio tarp *X* ir *Y* buvimą ($\rho \neq 0$), tikrinama nulinė hipotezė H_0: $\rho = 0$ (populiacijos koreliacijos koeficientas lygus 0 – nėra

tiesinio ryšio tarp kintamųjų) su viena iš alternatyvų: dešiniapuse H_1: $\rho > 0$; kairiapuse H_2: $\rho < 0$; arba dvipuse H_3: $\rho \neq 0$. Dešiniapusė alternatyva H_1 apibūdinama taip: „koreliacijos koeficientas reikšmingai viršija nulį“ arba „tarp kintamųjų yra reikšmingai teigiama koreliacija“. Kairiapusė alternatyva H_2 apibūdinama taip: „koreliacijos koeficientas reikšmingai mažesnis už nulį“ arba „tarp kintamųjų yra reikšmingai neigiama koreliacija“. Dvipusė alternatyva H_3 apibūdinama taip: „koreliacijos koeficientas reikšmingai skiriasi nuo nulio“ arba „tarp kintamųjų yra patikima (reikšminga) koreliacija“. H_0 tikrinti naudojami atitinkami statistiniai kriterijai su statistika, priklausančia nuo apskaičiuoto koreliacijos koeficiento r ir nuo tirtų individų skaičiaus n.

Koreliacijos koeficientai dažniausiai skaičiuojami statistiniu paketu. Kaip ir tikrinant visas statistines hipotezes, pakete pateikiama kriterijaus statistikos bei jo atitinkama dvipusė p reikšmė (5.4 pav.). Pagal šią p reikšmę arba atmetame hipotezę apie koreliacijos koeficiento lygybę nuliui, arba jai neprieštaraujame. Jei nulinę hipotezę atmetame, sakoma, kad „tarp kintamųjų X ir Y yra reikšminga koreliacija“.

9.3. Pirsono koreliacijos koeficientas

Turime dviejų kiekybinių rodiklių X ir Y atsitiktinių matavimų poras (x_1, y_1), (x_2, y_2) ... (x_n, y_n) – dvimačius atsitiktinius dydžius, turinčius tą patį skirstinį. Jei nėra pagrindo prieštarauti prielaidai, kad X ir Y jungtinis skirstinys yra dvimatis normalusis (praktiškai tai bus visuomet, jei X ir Y skirstiniai laikytini normaliaisiais), tuomet tiesiniam ryšiui tarp kintamųjų X ir Y vertinti naudojamas Pirsono (*Pearson*) koreliacijos koeficientas

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}; \qquad (9.1)$$

čia $\bar{x}$ ir $\bar{y}$ – imčių vidurkiai. Pagal apibrėžimą r yra atsitiktinis dydis. Į r formulę įstačius konkrečios imties duomenis, gaunama konkreti Pirsono koreliacijos koeficiento reikšmė. 9.2 pav. pateiktiems duomenims gauta: 2) $r = 0{,}53$; 3) $r = 0{,}93$; 4) $r = -0{,}61$; 5) $r = -0{,}89$.

Pirsono koreliacijos koeficientas jautrus išskirtims – imties „apšiukšlinimui“. Jei imtyje yra narių iš kitos populiacijos (ats. d. su kitu skirstiniu), r skirstinys gali žymiai keistis, ypač jei imtis nedidelė. Konkrečios imties atveju r reikšmę gali labai iškreipti viena ar kelios išskirtys. Pavyzdžiui, naudojant 3.5 lentelėje pateiktus 26 jaunų sveikų suaugusių asmenų SAS ir DAS duomenis, koreliacijos koeficiento tarp SAS ir DAS reikšmė lygi 0,73. Tačiau

prie šių matavimų pridėjus asmens iš kitos populiacijos SAS ir DAS reikšmę, pavyzdžiui, (100, 80), koreliacijos koeficiento reikšmė smarkiai pakinta – gauname $r = 0{,}57$.

Kaip minėta 9.2 skyriuje, koreliacijos koeficientas r, apibrėžtas (9.1) formule, yra populiacijos koreliacijos koeficiento ρ įvertis. Jis priklauso nuo imties; taigi r yra atsitiktinis dydis – jo reikšmė gali būti didelė, nors ρ lygus 0. Todėl remiantis konkrečiais X ir Y matavimais nustatyta r reikšme, dar negalima daryti išvados apie ryšio tarp X ir Y buvimą, t. y. teigti, kad $\rho \neq 0$. Todėl būtina tikrinti nulinę hipotezę H_0: $\rho = 0$ su viena iš 9.2 skyriuje pateiktų alternatyvų. Šiai hipotezei tikrinti naudojamas t kriterijus su statistika:

$$t = r\sqrt{n-2} \,/\, \sqrt{1-r^2}\,. \tag{9.2}$$

Jei kintamieji X ir Y yra nepriklausomi ($\rho = 0$), t turi asimptotinį Stjudento skirstinį su $(n - 2)$ laisvės laipsnių (tai reiškia, kad didėjant n, t tankis, esant $\rho = 0$, artės prie Stjudento skirstinio su $(n - 2)$ laisvės laipsniais tankio). Praktiškai, kai $n > 10$, t skirstinį, esant teisingai H_0, galima laikyti Stjudento skirstiniu. Hipotezė apie koreliacijos koeficiento reikšmingumą dažniausiai tikrinama statistiniais paketais. Juose pateikiama koreliacijos koeficiento reikšmė, t kriterijaus statistikos (9.2) ir kriterijaus dvipusė p reikšmė (5.4 pav.). H_0 priėmimo ar atmetimo taisyklė su pasirinktu reikšmingumo lygmeniu α, pagal alternatyvą, pateikta 9.1 lentelėje. Jei n yra nedidelis ($n \leq 10$), H_0 tikrinti naudojamas tikslus kriterijus: gauta r reikšmė lyginama su statistikos (9.1) tikslaus skirstinio, esant $\rho = 0$ (simetrišku, priklausančio tik nuo n), atitinkamo lygio kvantiliu (8 lentelė). H_0 priėmimo ar atmetimo taisyklė analogiška pateiktoje 9.1 lentelėje. Nedideliems n ($n \leq 30$) lentelėse nurodomi ir Pirsono koreliacijos koeficiento tikslūs pasikliautinieji intervalai.

9.1 lentelė. Koreliacijos koeficiento reikšmingumo tikrinimas (čia $t_{1-\alpha}(n-2)$ – Stjudento skirstinio su $(n-2)$ l. l. 1 – α lygio kvantilis)

Alternatyva	H_0 atmetimo sritis	H_0 atmetimo taisyklė pagal kriterijaus dvipusę p reikšmę
H_1: $\rho > 0$	$t > t_{1-\alpha}(n-2)$	$r > 0$, $p/2 < \alpha$
H_2: $\rho < 0$	$t < -t_{1-\alpha}(n-2)$	$r < 0$, $p/2 < \alpha$
H_3: $\rho \neq 0$	$\lvert t\rvert > t_{1-\alpha/2}(n-2)$	$p < \alpha$

9.1 pavyzdys. 3.5 lentelėje pateikta 26 jaunų sveikų individų SAS ir DAS reikšmės. Pirsono koreliacijos koeficientas tarp SAS ir DAS lygus 0,73. Jo reikšmingumui tikrinti skirtos t kriterijaus statistikos reikšmė: $t = 0{,}73\sqrt{24} \,/\, \sqrt{1-(0{,}73)^2} = 5{,}23$, $p = 0{,}000023$. Todėl daroma išvada, kad $\rho > 0$, ir didėjant jaunų sveikų individų SAS, didėja ir DAS.

9.4. Ranginiai koreliacijos koeficientai. Spirmeno (*Spearman*) koreliacijos koeficientas

Analizuojant medikų sukauptus duomenis, dažnai pasitaiko atvejų, kai negalime tvirtinti, kad kintamųjų X ir Y jungtinis skirstinys yra normalusis. Taip būna tuomet, kai:

- tarp (X, Y) reikšmių yra išskirčių;
- turime nedaug (X, Y) matavimų ir negalime daryti išvados dėl normalumo;
- kiekybinių kintamųjų X ir Y reikšmės pateiktos rangais.

Visais šiais atvejais tiesinio ryšio tarp kintamųjų stiprumui vertinti naudojami ranginiai koreliacijos koeficientai. Kaip minėta, ranginiai koreliacijos koeficientai nesusiję su jungtiniu (X, Y) skirstiniu. Jie skaičiuojami naudojant ne kintamųjų reikšmes, o jų rangus. Todėl ranginiams koreliacijos koeficientams išskirtys nedaro tokios didelės įtakos kaip Pirsono koreliacijos koeficientui. Ranginių koreliacijos koeficientų naudojimo prielaida: jungtinis (X, Y) populiacijos skirstinys yra tolydusis. Ši prielaida leidžia teigti, kad konkrečiose imtyse $x_1, x_2 \ldots x_n$ ir $y_1, y_2 \ldots y_n$ pasikartojančių reikšmių yra labai mažai, t. y. $x_1, x_2 \ldots x_n$ ir $y_1, y_2 \ldots y_n$ rangai yra jų eilės numeriai variacinėje sekoje.

Spirmeno (*Spearman*) koreliacijos koeficientas r_S skaičiuojamas taip: turime kiekybinio kintamojo (X, Y), kurio modelis – dvimatis tolydusis ats. d., atsitiktinę imtį (x_1, y_1), $(x_2, y_2) \ldots (x_n, y_n)$. Reikšmės $x_1, x_2 \ldots x_n$ ir $y_1, y_2 \ldots y_n$ suranguojamos atskirai ir imtyje $(x_1, y_1) \ldots (x_n, y_n)$ esančios reikšmės pakeičiamos rangais. Gaunamos rangų poros (R_{x1}, R_{y1}), $(R_{x2}, R_{y2}) \ldots (R_{xn}, R_{yn})$. Spirmeno koreliacijos koeficientas r_S apibrėžiamas:

$$r_S = \frac{\sum_{i=1}^{n}(R_{x1} - (n+1)/2)(R_{yi} - (n+1)/2)}{(\sum_{i=1}^{n}(R_{xi} - (n+1)/2)^2 \sum_{i=1}^{n}(R_{yi} - (n+1)/2)^2)^{1/2}}. \tag{9.3}$$

Iš šios formulės matyti, kad r_S yra Pirsono koreliacijos koeficientas, apskaičiuotas naudojant ne kintamojo reikšmes, o jų rangus (vidutinis n dydžio imties rangas lygus $(n+1)/2$). Pertvarkius (9.3) formulę gaunama, kad Spirmeno koreliacijos koeficientas lygus:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}; \tag{9.4}$$

čia $D_i - x_i$ ir y_i rangų skirtumas $D_i = R_{xi} - R_{yi}$. Spirmeno koreliacijos koeficiento skaičiavimo pavyzdys pagal (9.4) formulę pateiktas 9.2 lentelėje.

9.2 lentelė. Spirmeno koreliacijos koeficiento skaičiavimo pavyzdys

x_i	y_i	R_{xi}	R_{yi}	$(R_{xi} - R_{yi})^2$
1,1	5	1	5	16
2,1	3	2	3	1
3,1	4	3	4	1
4,1	2	4	2	4
5,1	1,5	5	1	16
Iš viso				38
$r_S = 1 - 6 \times 38/(5 \times (25 - 1)) = 1 - 228/120 = 1 - 1{,}9 = -0{,}9$				

(9.3) formulę galima pertvarkyti ir taip: rangų porų seka (R_{x1}, R_{y1}), (R_{x2}, R_{y2}) ... (R_{xn}, R_{yn}) išdėstoma taip, kad vietoj Y rangų būtų skaičiai 1, 2 ... n. Šioje rangų sekoje pažymimi atitinkami X rangai Z_1, Z_2 ... Z_n. Tuomet r_S skaičiuojamas analogiškai (9.4):

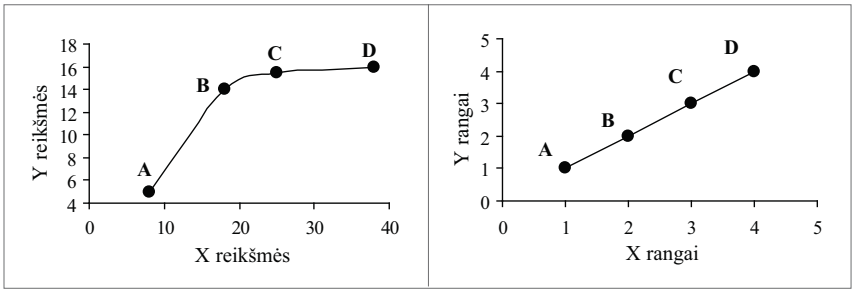
$$r_S = 1 - \frac{6\sum_{i=1}^{n}(Z_i - i)^2}{n(n^2 - 1)} = 1 - \frac{6}{n(n^2 - 1)} S \cdot \qquad (9.5)$$

Jei tarp X ir Y yra tiesinis ryšys ir didėjant X, didėja ir Y, tuomet $Z_i = i$, ir $S = 0$. Jei didėjant X, mažėja Y, tuomet $Z_i = n - i$, ir S lygus $n(n^2 - 1)/3$.

Hipotezei H_0: $\rho = 0$ su atitinkama vienpuse arba dvipuse alternatyva tikrinti mažiems n ($n \leq 10$) naudojamas tikslus kriterijus. Nustatyta, kad (9.5) formulėje pateiktos statistikos S skirstinys, kai X ir Y imties reikšmės (x_i, y_i) yra nepriklausomi atsitiktiniai dydžiai ($\rho = 0$), $\widetilde{S}$ priklauso tik nuo n. Lentelėse pateiktos šio skirstinio funkcijos arba kvantilių reikšmės. 9 lentelėje pateiktos $\widetilde{S}$ skirstinio kvantilių reikšmės. Išvada apie ρ reikšmę gaunama analogiškai (9.1) lentelėje nurodytai taisyklei, lyginant gautą S reikšmę su lentelėse pateiktu $\widetilde{S}$ skirstinio atitinkamo lygio kvantiliu.

Didesniems n ($n > 10$) hipotezei H_0: $\rho = 0$ tikrinti su atitinkama vienpuse arba dvipuse alternatyva naudojamas t kriterijus su statistika: $t = r_S\sqrt{n-2}\,/\sqrt{1-r_S^2}$. t asimptotinis skirstinys, kai $\rho = 0$, yra Stjudento su ($n - 2$) laisvės laipsnių. Todėl H_0 priėmimo ir atmetimo taisyklė tokia pati, kaip ir Pirsono koreliacijos koeficiento atveju (9.1 lentelė).

Kai tarp kintamųjų X ir Y yra netiesinis monotoninis ryšys, tuomet X ir Y ryšio stiprumui vertinti geriau naudoti Spirmeno koreliacijos koeficientą, nes tarp X ir Y reikšmių esant netiesiniam ryšiui, tarp X ir Y rangų gali būti tiesinis ryšys (9.6 pav.).



9.6 pav. Priklausomybė tarp reikšmių ir tarp rangų

9.5. Kendalo koreliacijos koeficientas

Spirmeno koreliacijos koeficientą, apibrėžiamą (9.3–9.5) formulėmis, sunku interpretuoti. Jei reikalinga koreliacijos koeficiento interpretacija, esant teisingai alternatyvai, naudotinas Kendalo koreliacijos koeficientas.

Sakykime, (x_1, y_1), (x_2, y_2) ... (x_n, y_n) – kiekybinio kintamojo (X, Y), turinčio dvimatį tolydųjį skirstinį, imtis. Kendalo koreliacijos koeficientas, kaip ir ryšio tarp tvarkos kintamųjų rodikliai, skaičiuojamas naudojant suderintų ir nesuderintų imties porų skaičių. Reikšmių (x_i, y_i) ir (x_j, y_j) pora yra suderinta, jei skirtumai $x_i - x_j$ ir $y_i - y_j$ turi vienodus ženklus, ir nesuderinta, jei skirtumų ženklai skirtingi. Jei duomenys pateikti rangais, suderinta ir nesuderinta poros apibrėžiamos analogiškai, tik vietoj skirtumų tarp imties reikšmių imami rangų skirtumai.

Pažymėkime P – suderintų porų skaičių imtyje, Q – nesuderintų porų skaičių bei K – suderintų ir nesuderintų porų skaičiaus skirtumą: $K = P - Q$. Visų galimų porų $P + Q$ skaičius lygus $n(n - 1)/2$. K gali kisti nuo $-n(n - 1)/2$ iki $n(n - 1)/2$. Jei X ir Y yra nepriklausomi atsitiktiniai dydžiai, suderintų ir nesuderintų porų yra beveik vienodai ir K yra artimas 0.

Kendalo koreliacijos koeficientas r_K apibrėžiamas:

$$r_K = 2K / [n(n - 1)]. \tag{9.6}$$

Pagal apibrėžimą Kendalo koreliacijos koeficientas kinta nuo –1 iki 1. Jei konkrečioje imtyje didėjant X, reikšmės Y turi tendenciją mažėti, tai imtyje nesuderintų porų yra daugiau nei suderintų, ir r_K reikšmė neigiama. Jei didėjant X, reikšmės Y turi tendenciją didėti, suderintų porų yra daugiau nei nesuderintų ir r_K reikšmė teigiama.

Tikrinant hipotezę H_0: $\rho = 0$ su viena iš alternatyvų ($\rho > 0$; $\rho < 0$; $\rho \neq 0$) nedideliam n ($n \leq 10$) naudojamas tikslus kriterijus. Pažymėkime $\widetilde{K}$ – atsi-

tiktinį dydį, lygų suderintų ir nesuderintų porų skaičių skirtumui dvimatėje nepriklausomų ats. d. n dydžio imtyje. Kai X ir Y yra nepriklausomi, K skirstinys sutampa su $\widetilde{K}$ skirstiniu. Tikrinant H_0, apskaičiuota K reikšmė lyginama su $\widetilde{K}$ skirstinio atitinkamo lygio kvantiliu (10 lentelė) arba, naudojantis $\widetilde{K}$ skirstinio funkcijos lentelėmis, nustatoma statistikos K p reikšmė. Išvada apie r_K reikšmingumą daroma analogiškai 9.1 lentelėje pateiktai taisyklei.

Esant gana dideliam n, ($n > 10$), H_0 tikrinti naudojamas asimptotinis kriterijus. Statistikos

$$K/[n(n-1)(2n+5)/18\}^{-0,5} \tag{9.6}$$

asimptotinis skirstinys, esant teisingai nulinei hipotezei, yra standartinis normalusis, todėl hipotezės H_0: $\rho = 0$ priėmimo ar atmetimo taisyklė analogiška taisyklei, pateiktai 9.1 lentelėje, tik vietoj Stjudento skirstinio kvantilio $t_{1-\alpha}(n-2)$ ir $t_{1-\alpha/2}(n-2)$ naudojami standartinio normaliojo skirstinio kvantiliai $z_{1-\alpha}$ ir $z_{1-\alpha/2}$.

9.6. Hipotezė apie koreliacijos koeficiento lygybę skaičiui. Dviejų koreliacijos koeficientų lyginimas

Analizuojant sąsajas tarp kiekybinių rodiklių, aktualu konstatuoti ne tik tiesinio ryšio buvimą ar nebuvimą, bet ir patikimai įvertinti ryšio stiprumą – patvirtinti, kad ne tik $r > 0,3$, bet ir $\rho > 0,3$ (ryšys yra reikšmingai vidutinio stiprumo) ar $\rho \geq 0,6$ (yra reikšmingai stiprus ryšys tarp rodiklių). Norint konstatuoti, kad ryšys tarp kintamųjų X ir Y yra tam tikro stiprumo arba populiacijos koreliacijos koeficientas ((X, Y) skirstinio parametras) ρ yra tam tikro dydžio, sakykime, ρ_0, būtina tikrinti nulinę hipotezę H_0: $\rho = \rho_0$ su viena iš alternatyvų: $\rho > \rho_0$, $\rho < \rho_0$ ar $\rho \neq \rho_0$.

Sakykime, ρ – ryšio tarp X ir Y stiprumo rodiklį vertiname Pirsono koreliacijos koeficientu r (prielaida – (X, Y) skirstinys yra dvimatis normalusis). Hipotezei H_0 tikrinti, kai $n > 30$, naudojamo Z kriterijaus statistika lygi:

$$z = 0,5(\ln\frac{1+r}{1-r} - \ln\frac{1+\rho_0}{1-\rho_0})\sqrt{n-3}\,.$$

Esant teisingai nulinei hipotezei, z asimptotinis skirstinys yra standartinis normalusis. Kai n gana didelis, $n > 30$, z skirstinį galima laikyti standartiniu normaliuoju. H_0 priėmimo ir atmetimo taisyklė, priklausomai nuo alternatyvos, pateikta 9.4 lentelėje.

9.4 lentelė. Sprendinio apie koreliacijos koeficiento lygybę skaičiui priėmimo taisyklė

Alternatyva	H_0 atmetimo sritis	H_0 atmetimo taisyklė pagal kriterijaus dvipusę p reikšmę
H_1: $\rho > \rho_0$	$t > z_{1-\alpha}$	$r > 0$, $p/2 < \alpha$
H_2: $\rho < \rho_0$	$t < -z_{1-\alpha}$	$r < 0$, $p/2 < \alpha$
H_3: $\rho \neq \rho_0$	$\lvert t\rvert > z_{1-\alpha/2}$	$p < \alpha$

Medikams dažnai tenka lyginti dvi ligonių grupes, pavyzdžiui, placebo ir vartojusių vaisto. Analizuojant vaisto poveikį, dažniausiai lyginami kintamųjų vidurkiai. Tačiau poveikis gali pasireikšti ne tik kintamojo vidurkio poslinkiu, bet ir ryšio tarp individo rodiklių pokyčiu. Tokiu atveju lyginami kiti populiacijų skirstinio parametrai – koreliacijos koeficientai.

Sakykime, ρ_1 ir ρ_2 – koreliacijos koeficientai tarp kintamųjų (ats. dydžių) X ir Y dviejose populiacijose. Tikrinsime nulinę hipotezę H_0: $\rho_1 = \rho_2$ (abiejose populiacijose ryšys tarp X ir Y vienodai stiprus) su viena iš alternatyvų H_1: $\rho_1 > \rho_2$ (jei $\rho_2 > 0$, tai I populiacijoje ryšys tarp X ir Y yra stipresnis nei II); H_2: $\rho_1 < \rho_2$ (jei $\rho_1 > 0$, tai I populiacijoje ryšys tarp X ir Y yra silpnesnis nei II); H_3: $\rho_1 \neq \rho_2$ (abiejose populiacijose ryšio tarp X ir Y stiprumas nevienodas).

Pažymėkime: r_1 ir r_2 – imčių iš I ir II populiacijos Pirsono koreliacijos koeficientai, $z_1 = 0{,}5\ln[(1 + r_1)/(1 - r_1)]$, $z_2 = 0{,}5\ln[(1 + r_2)/(1 - r_2)]$ – koreliacijos koeficiento Fišerio transformacijos, n_1 ir n_2 – imčių dydžiai. Kai $n_1 > 30$ ir $n_2 > 30$, H_0 tikrinti naudojamas Z kriterijus su statistika:

$$Z = (z_1 - z_2)/((n_1 - 3)^{-1} + (n_2 - 3)^{-1})^{-1/2}.$$

Esant teisingai nulinei hipotezei, Z asimptotinis skirstinys yra standartinis normalusis. Nulinės hipotezės priėmimo ir atmetimo taisyklė tokia pati, kaip 9.4 lentelėje, tik atitinkamai keičiasi alternatyvos.

9.7. Koreliacijų matrica

Atliekant sudėtingus medicininius tyrimus, pavyzdžiui, biocheminį, echoskopiją ir t. t., fiksuojama keletas kiekybinių rodiklių. Analizuodami tokių tyrimų informatyvumą, medikai-tyrėjai turi vertinti ryšį tarp rodiklių, gautų atlikus vieną tyrimą (pvz., echoskopiją) arba vertinti ryšį tarp rodiklių, gautų atlikus skirtingus tyrimus (pvz., echoskopiją ir biocheminį tyrimą). Tokiu atveju tenka skaičiuoti ir analizuoti ne vieną koreliacijos koeficientą, o kelis. Informaciją apie kelių kintamųjų tarpusavio (porinių) ryšių įverčius patogu pateikti koreliacijų matrica arba dalimi koreliacijų matricos.

Koreliacijų matrica yra kvadratinė; jos eilutės ir stulpeliai atitinka analizuojamus kintamuosius. Gardelėje, esančioje koreliacijų matricos eilutės ir stulpelio susikirtime, yra koreliacijos koeficientas tarp eilutę ir stulpelį atitinkančių kintamųjų. Be to, gardelėje pateikiamas ne tik atitinkamas koreliacijos koeficientas, bet ir jo reikšmingumo rodiklis – statistinio kriterijaus p reikšmė bei imties dydis. Koreliacijų matricos diagonalėje yra vienetai; be to, matrica yra simetrinė. Koreliacijų matricos dalis – tai koreliacijos koeficientų tarp kintamųjų $X_1 ... X_k$ ir $Y_1 ... Y_m$ matrica; čia $X_1 ... X_k$ – vieno, o $Y_1 ... Y_m$ – kito tyrimo metu gauti kintamieji. Šios matricos eilutės atitinka kintamuosius $X_1 ... X_k$, stulpeliai – kintamuosius $Y_1 ... Y_m$; i-tosios eilutės ir j-tojo stulpelio susikirtimo gardelėje rašomas koreliacijos koeficientas tarp X_i ir Y_j bei jo patikimumo rodiklis (p reikšmė). Statistiniuose paketuose pateikiama visa koreliacijų matrica arba jos dalis.

9.2 pavyzdys. 9.5 lentelėje pateikta 487 ligonių, sirgusių ūmiais koronariniais sindromais, echoskopijos rodiklių – širdies galinio diastolinio dydžio (KSGDD), užpakalinės sienelės storio (USS), kairiojo ir dešiniojo prieširdžio dydžio (KPR ir DPR) bei išstūmimo frakcijos (IF) – koreliacijų matrica.

9.5 lentelė. Echoskopijos rodiklių Pirsono koreliacijos koeficientai ir jų p reikšmės

	KSGDD	USS	KPR	DPR	IF
KSGDD	1	0,0589 p=0,21	0,227 p<0,001	0,230 p<0,001	–0,163 p=0,001
USS	0,0589 p=0,21	1	0,163 p=0,001	0,084 p=0,18	–0,081 p=0,12
KPR	0,227 p<0,001	0,163 p=0,001	1	0,529 p<0,001	–0,058 p=0,22
DPR	0,230 p<0,001	0,084 p=0,18	0,529 p<0,001	1	0,007 p=0,88
IF	–0,163 p=0,001	–0,081 p=0,12	–0,058 p=0,22	0,007 p=0,88	1

Iš 9.5 lentelės matyti, kad KSGDD reikšmingai koreliuoja su KPR, DPR ir IF; tarp KSGDD ir minėtų rodiklių ryšys yra silpnas. Tarp KSGDD ir USS reikšmingo ryšio ($\rho > 0$) konstatuoti negalima, nes $p/2 = 0,21/2 > 0,05$. Kadangi koreliacijos koeficientai tarp KSGDD ir KPR bei DPR atitinkamai lygūs 0,227 ir 0,23 ($p < 0,001$), todėl galima daryti išvadą, kad, didėjant širdies galiniam diastoliniam dydžiui, tiek kairysis, tiek dešinysis prieširdžiai turi tendenciją didėti. Kadangi r tarp KSGDD ir IF lygus $–0,136 < 0$ ir $p = 0,001 < 0,05$, galima daryti išvadą, kad, didėjant KSGDD, išstūmimo frakcija mažėja.

USS reikšmingai koreliuoja tik su KPR. Kadangi $r = 0{,}163 > 0$, galima daryti išvadą, kad, didėjant KPR, USS turi tendenciją didėti. Tarp KPR ir DPR stebimas vidutinio stiprumo ryšys: $r = 0{,}529$, $p < 0{,}001$ – koreliacija reikšminga. Taigi didėjant vienam prieširdžiui, kitas irgi didėja. Tarp IF ir KPR bei DPR reikšmingo ryšio konstatuoti negalima – r artimas 0 bei $p > 0{,}2$.

9.8. Dalinis koreliacijos koeficientas

Minėta, kad priežastiniam ryšiui tarp dviejų kiekybinių rodiklių pagrįsti, be teorinės reiškinio analizės, skaičiuojamas koreliacijos koeficientas tarp šių rodiklių ir vertinamas jo patikimumas. Tačiau priežastinio ryšio mechanizmas sudėtingas: reikšminga koreliacija tarp kintamųjų *X* ir *Y* gali būti ir tuo atveju, kai:

- *X* ir *Y* sąlygoti kito kintamojo *Z*;
- *X* gali būti *Z* priežastis, o *Z* – *Y* priežastis *Z*: $X \rightarrow Z \rightarrow Y$.

Priežastiniam ryšiui tarp 3–5 kiekybinių rodiklių analizuoti naudojami daliniai koreliacijos koeficientai. Dalinis koreliacijos koeficientas (*partial correlation*) $r_{XY.Z}$ yra koreliacijos koeficientas tarp kintamųjų *X* ir *Y*, apskaičiuotas izoliavus kintamojo *Z* įtaką (arba kontroliuojant kintamąjį *Z*). $r_{XY.Z}$ formulė tokia:

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}; \tag{9.7}$$

čia r_{XY}, r_{XZ} ir r_{YZ} – koreliacijos koeficientai tarp nurodytų kintamųjų, vadinamieji poriniai koreliacijos koeficientai. *Z* vadinamas kontroliuojančiu kintamuoju. Analogiškai apibrėžiamas $r_{XY.ZV}$ – koreliacijos koeficientas tarp *X* ir *Y*, apskaičiuotas izoliavus *Z* ir *V* įtaką, ir t. t. Lyginant dalinį koreliacijos koeficientą su poriniu, t. y. lyginant $r_{XY.Z}$ su r_{XY}, daroma išvada apie kintamųjų *X*, *Y* ir *Z* ryšį (9.7 pav.).

Iš (9.7) formulės matome: kai r_{XZ} ir r_{YZ} lygūs 0, $r_{XY.Z} = r_{XY}$. Taigi daroma išvada: jei $r_{XY.Z}$ ir r_{XY} nesiskiria, kintamasis *Z* nesusijęs (neturi ryšio) su *X* ir *Y* (9.7 a pav.). Jei $r_{XY.Z} = 0$, koreliacija tarp *X* ir *Y* atsiranda dėl paaiškinamo (*explanationary*) kontroliuojančio kintamojo *Z* efekto: arba *Z* veikia *X* ir *Y* (9.7 b pav.) arba kintamasis *X* veikia *Z*, o *Z* savo ruožtu veikia *Y* (9.7 c pav.). Tai dar vadinama kontroliuojančio kintamojo efektu.

Jei $r_{XY} > r_{XY.Z} > 0$ (taip bus, jei r_{XZ} ir r_{YZ} yra to paties ženklo, t. y. didėjant *Z*, *X* ir *Y* kitimo tendencija vienoda) arba $r_{XY} < r_{XY.Z} < 0$ pasireiškia dalinis paaiškinimo efektas (*partial explanation*) – stipresnė koreliacija tarp *X* ir *Y* atsiranda dėl *Z* įtakos abiem rodikliams (9.7 d pav.) arba dėl to, kad *X* yra

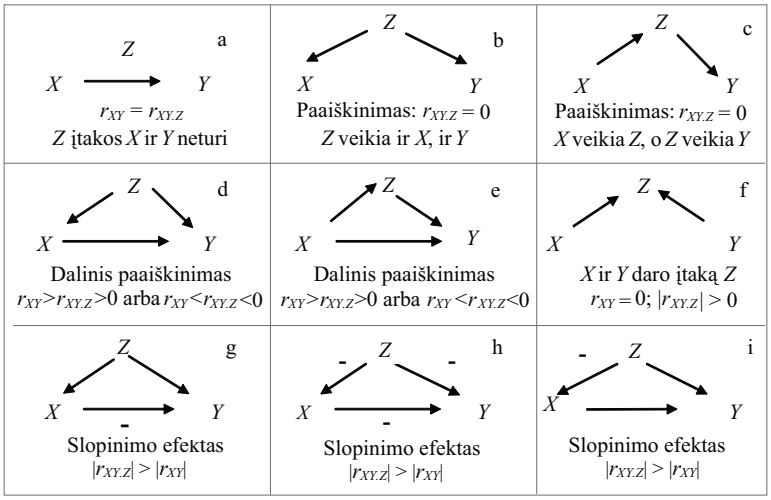
Z priežastis, o *Y* – *Z* priežastis, (9.7 e pav.). Jei $r_{XY} = 0$, o $|r_{XY.Z}| > 0$, daroma išvada, kad *X* ir *Y* daro įtaką kintamajam *Z* (9.7 f pav.).

Jei (9.7) formulėje $r_{XY} < 0$, o r_{XZ} ir r_{YZ} yra vienodo ženklo, gaunama, kad $|r_{XY.Z}| > |r_{XY}|$. Tokiu atveju pasireiškia kontroliuojančiojo kintamojo slopinimo efektas: dėl *Z* poveikio kintamieji *X* ir *Y* turi tendenciją kartu arba didėti, arba mažėti; toks *Z* poveikis mažina koreliaciją tarp *X* ir *Y* (9.7 g, h pav.). Jei (9.7) formulėje $r_{XY} > 0$, o r_{XZ} ir r_{YZ} yra skirtingų ženklų, gaunama, kad $r_{XY.Z} > r_{XY}$. Tuomet pasireiškia toks kontroliuojančio kintamojo slopinimo efektas: dėl *Z* poveikio vienas kintamasis didėja, kitas mažėja (9.7 i pav.). Izoliavus kintamojo *Z* poveikį, koreliacija tarp *X* ir *Y* sustiprėja.

9.3 pavyzdys. Analizuojant 669 MI sergančiųjų moterų išstūmimo frakcijos (IF) ryšį su biocheminiais žyminiais, nustatyta, kad Spirmeno koreliacijos koeficientas tarp IF ir kreatinino koncentracijos K lygus –0,195 ($p < 0,05$). Tačiau stebima šių rodiklių reikšminga koreliacija su kontroliuojančiu kintamuoju – amžiumi (A): $r_{(IF, A)} = -0,227$; $r_{(K, A)} = 0,227$ – didėjant amžiui, išstūmimo frakcija turi tendenciją mažėti, o kreatinino koncentracija – didėti. Dalinis koreliacijos koeficientas tarp IF ir K, izoliavus amžiaus įtaką, lygus:

$$r_{(IF, K).A} = \frac{-0,195 + 0,274 \times 0,227}{\sqrt{(1 - 0,227 \times 0,227)(1 - 0,274 \times 0,274)}} = -0,141.$$

Gavome, kad $r_{(IF, K).A}$ absoliučiu dydžiu mažesnis už $r_{(IF, K)}$ (d atvejis). Todėl galima tvirtinti, kad didesnė koreliacija tarp IF ir K atsiranda dėl amžiaus įtakos minėtiems rodikliams.



9.7 pav. Dalinė koreliacija ir išvados apie kintamųjų priežastinį ryšį

9 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. I dalis. Vilnius: TEV, 2000, 238 p.
3. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslo ir enciklopedijų leidykla, 416 p.
6. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
5. Холлендер М., Вулф Д. А. *Непараметрические методы статистики*. 1983. Москва: Наука, 516 с.
6. Watts S., Halliwell L. *Essential Environmental Science. Methods and Techniques*. 1996, p. 512.
7. *Modeliavimo pratimai, grafiškai iliustruojantys koreliacijos koeficiento ir duomenų skaidos priklausomybę.* Prieiga per internetą: http://noppa5.pc.helsinki.fi/koe/corr/index.html.
8. *Dalinės koreliacijos apibrėžimas ir interpretacija.* Prieiga per internetą: http://www2.chass.ncsu.edu/garson/pa765/partialr.htm.
9. *Koreliacinė analizė.* Prieiga per internetą: http://www.psy.ucsd.edu/~sky/Psyc%2060%20Correlation.ppt.

10 SKYRIUS

Regresinė analizė

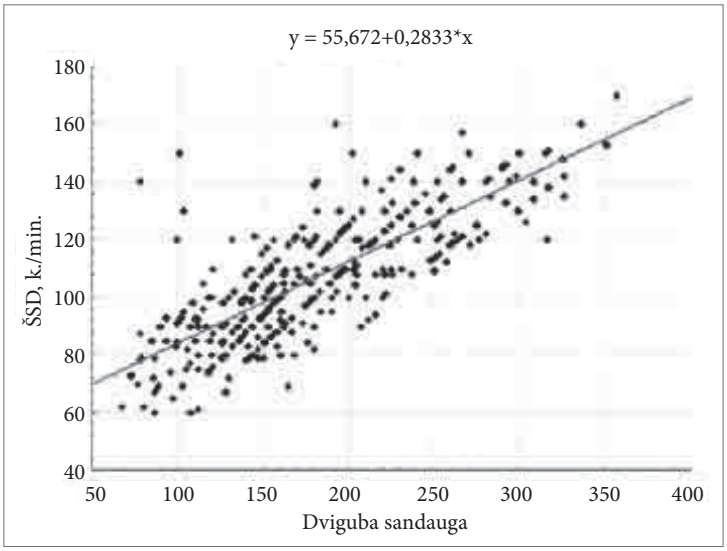
10.1. Regresijos sąvoka

Tiek medicinoje, tiek kituose moksluose kai kuriuos kintamuosius galima laikyti atsaku (*outcome, response variable*) į jį veikiantį vieną ar keletą faktorių (*factor, covariate, explanatory variable*):

Pavyzdžiui, ūmiais koronariniais sindromais sirgusiems ligoniams atliekant fizinio krūvio mėginį, ŠSD, SAS ir DAS reikšmės, nustatytos prieš nutraukiant krūvį, priklauso nuo rodiklio, nustatyto fizinio krūvio metu, vadinamu dviguba sandauga. Didėjant dvigubai sandaugai, ŠSD taip pat turi tendenciją didėti (10.1 pav.). Šiuo atveju faktorius, arba nepriklausomas kintamasis X, yra dviguba sandauga, o atsakas, arba priklausomas kintamasis Y, yra ŠSD prieš nutraukiant krūvį. Atsaką gali veikti keli faktoriai $X^{(1)}, X^{(2)} \ldots X^{(k)}$; taigi faktorius X gali būti ir daugiamatis: $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(k)})$.

Tarp kintamųjų konstatavus faktoriaus–atsako ryšį ir žinant faktoriaus (vienmačio ar daugiamačio) reikšmę, tikslinga prognozuoti ir Y reikšmę. Faktoriaus–atsako priklausomybei vertinti naudojami regresiniai modeliai. Regresinio modelio prielaidos:

- X reikšmės determinuotos (neatsitiktinės);
- priklausomo kintamojo reikšmės $y_1, y_2 \ldots y_n$ yra atsitiktinės ir nekoreliuotos;
- nepriklausomų kintamųjų $X^{(1)}, X^{(2)} \ldots X^{(k)}$ reikšmės yra tiesiškai nepriklausomos (kelių faktorių atveju).



10.1 pav. ŠSD, nustatyto prieš nutraukiant krūvį, priklausomybė nuo dvigubos sandaugos

Funkcija $f(\boldsymbol{x}) = E(Y|\boldsymbol{X}=\boldsymbol{x})$ vadinama Y regresijos funkcija $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(k)})$ atžvilgiu; $f(\boldsymbol{x})$ yra k-matis paviršius $(k+1)$ matavimo erdvėje. Kai $\boldsymbol{x} = (x^{(1)}, x^{(2)})$ $(k = 2)$, funkcija $f(\boldsymbol{x})$ yra paviršius trimatėje erdvėje. Vieno faktoriaus atveju $(k = 1)$ $f(x)$ yra kreivė plokštumoje; šiuo atveju $f(x)$ vadinama regresijos kreive.

Regresijos funkcijos $f(\boldsymbol{x})$ interpretacija priklauso nuo atsako Y statistinio modelio, t. y. nuo Y skirstinio. 10.1 lentelėje pateikti keli dažniausiai naudojami regresiniai modeliai, klasifikuojami pagal Y skirstinį.

10.1 lentelė. Regresinio modelio klasė pagal Y skirstinį

Y skirstinys	Regresinis modelis
Normalusis	Normalioji regresija, arba tiesiog regresija
Dvinaris	Dvinarė regresija, jos atskiras atvejis – logistinė regresija (11 skyrius)
Polinominis (Y įgyja keletą reikšmių)	Polinominė regresija
Puasono	Puasono regresija

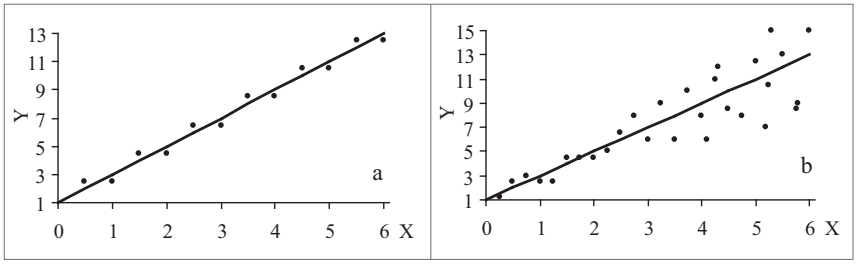
10.2–10.10 skyriuose nagrinėjamas regresinis modelis, kai Y skirstinys yra normalusis. Tuomet atsako reikšmę y_i galime išreikšti taip:

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, 2 \ldots n; \qquad (10.1)$$

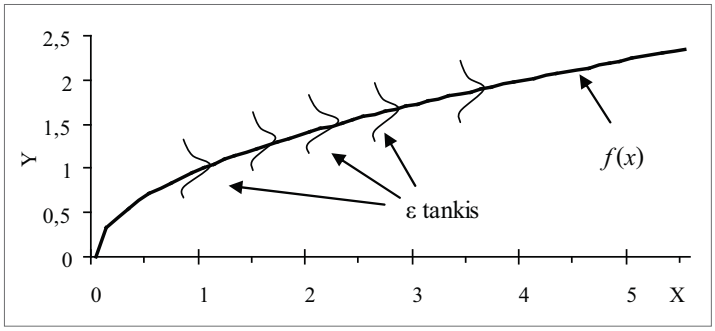
čia (x_i, y_i) – i-tojo individo faktoriaus ir atsako reikšmė, ε_i – normalusis atsitiktinis dydis su vidurkiu, lygiu nuliui, ir nežinoma dispersija, bendru atveju priklausančia nuo x_i: $D\varepsilon_i = D(Y|X = x_i) = \sigma^2(x_i)$, be to, dydžiai ε_1, ε_2 ... ε_n yra nepriklausomi. Jei Y sąlygoja keletas faktorių, (10.1) modelyje vietoj x_i naudojama daugiamačio faktoriaus reikšmė $\boldsymbol{x}_i$.

Jei Y dispersija vienoda visoms faktoriaus reikšmėms, t. y. $D(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma^2$, sakoma, kad duomenys yra homoskedastiški (10.2 a pav.). Jei ši sąlyga netenkinama, t. y. $D(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma^2(\boldsymbol{x})$; čia $\sigma^2(\boldsymbol{x})$ yra x funkcija; sakoma, kad duomenys yra heteroskedastiški (10.2 b pav.).

(10.1) regresiniame modelyje daroma prielaida, kad tarp atsako ir faktoriaus yra funkcinis ryšys $Y = f(X)$, kitaip tariant, sakoma, kad regresinio modelio forma (10.1) yra korektiška. Tačiau tikslios $f(x_i)$ reikšmės išmatuoti negalima – $f(x_i)$ reikšmė matuojama su paklaida ε_i, turinčia normalųjį skirstinį (10.3 pav.). Be to, nustatant atsako reikšmę, sisteminės paklaidos nedaroma ($E\varepsilon_i = 0$); tačiau matavimo tikslumas gali būti nevienodas ($D\varepsilon_i = \sigma^2(x_i)$). Taigi atsako reikšmės y_i yra išsibarsčiusios apie regresijos kreivę (arba dviejų faktorių atveju – apie paviršių) (10.3 pav.).



10.2 pav. (a) Duomenys su pastovia dispersija D(Y|x); (b) didėjant x, dispersija D(Y|x) didėja



10.3 pav. Atsako Y priklausomybė nuo jį veikiančio faktoriaus: paklaidų skirstinys yra normalusis

Pagal regresijos funkciją bei faktorių skaičių išskiriama keletas regresinių modelių tipų. Regresinių modelių klasifikacija ir regresinės analizės etapai pateikti 10.2 skyriuje.

10.2. Regresinio modelio tipai ir regresinės analizės etapai

Regresijos funkcija $f(x)$ gali būti:

- parametrinė (žinoma $f(x)$ formulė, bet nežinomi jos parametrai);
- neparametrinė.

Jei $f(x)$ yra neparametrinė, sakoma, kad turime neparametrinę regresiją; jei $f(x)$ parametrinė – turime parametrinę regresiją. Populiariausi medikų naudojami regresiniai modeliai – $f(x)$ parametrinė funkcija, turinti 2 ar 3 nežinomus parametrus. $f(x)$ parenkama taip, kad atspindėtų realų atsako kitimą didėjant ar mažėjant faktoriaus reikšmei. Jei tarp faktoriaus ir atsako reikšmių stebimas tiesinis ryšys, naudojama tiesinė regresijos funkcija $f(x) = \beta_0 + \beta_1 x$ (arba $f(x) = \alpha + \beta x$). Jei tarp atsako ir faktoriaus reikšmių stebimas netiesinis ryšys, naudojama netiesinė funkcija. Dažniausiai naudojamos: kvadratinė $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$; logaritminė $f(x) = \beta_0 + \beta_1 \log(x)$ ar eksponentinė

$$f(x) = \beta_0 + \beta_1 \exp(\beta_2 x) \tag{10.3}$$

funkcijos; čia β_0, β_1, β_2 – regresijos funkcijos koeficientai – regresinio modelio parametrai. Vertinant kelių faktorių įtaką atsakui, dažniausiai naudojama tiesinė kelių kintamųjų funkcija

$$f(x_1, x_2 \ldots x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{10.4}$$

arba

$$f(x_1, x_2 \ldots x_k) = \alpha + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k). \tag{10.5}$$

Pagal nepriklausomų kintamųjų skaičių ir regresijos funkcijos $f(\boldsymbol{x})$ pobūdį parametriniai regresiniai modeliai skirstomi į:

- vieno kintamojo tiesinę regresiją: $f(x) = \alpha + \beta x$;
- kelių kintamųjų (daugialypę) tiesinę regresiją:

  $f(x_1, x_2 \ldots x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$;
- netiesinę regresiją: $f(x)$ ar $f(\boldsymbol{x})$ netiesinė parametrinė funkcija ($f(x)$ gali būti kvadratinė, logaritminė ar eksponentinė; kelių kintamųjų netiesinės regresijos pavyzdys – (10.5) funkcija).

Sudarant bet kurio tipo regresinį modelį, skiriami 3 etapai: 1) modelio (regresijos funkcijos) parinkimas; 2) regresijos funkcijos vertinimas; 3) modelio diagnostika. Trumpai juos aptarsime.

1. Regresijos kreivė turi atspindėti biologiniais principais grindžiamą ryšį tarp atsako ir faktoriaus. Apie ryšio tarp faktoriaus ir atsako reikšmių pobūdį galima spręsti pagal X ir Y skaidos diagramą. Jei didėjant X, atsako reikšmės turi tendenciją didėti ar mažėti tiesiškai, naudojama tiesinė regresija. Jei Y kitimas yra netiesinis arba nemonotoninis, parenkama atitinkama netiesinė ar nemonotoninė parametrinė regresijos kreivė. Jei yra daug ($n > 100$) X ir Y matavimų bei ryšio tarp X ir Y pobūdis sudėtingas, galima naudoti neparametrinę regresiją. Sudarant daugiamatį regresinį modelį, būtina nustatyti, kokie faktoriai reikalingi regresiniam modeliui, t. y. būtina biologiškai ir statistiškai pagrįsti ryšį tarp į daugiamatį modelį įtraukiamo faktoriaus ir atsako.

2. Parinkus regresijos funkcijos tipą, būtina šią funkciją identifikuoti (įvertinti) remiantis duomenimis. Jei $f(x)$ – neparametrinė funkcija, ji vertinama neparametriniais metodais (10.10 skyrius). Jei $f(x)$ – parametrinė funkcija, pavyzdžiui, (10.3–10.5), būtina įvertinti jos nežinomus parametrus β_0, β_1 ... β_k. Šių parametrų įverčiai parenkami taip, kad regresijos funkcija būtų kuo „arčiau“ Y reikšmių. Tai atliekama mažiausių kvadratų metodu minimizuojant nežinomų parametrų atžvilgiu tikslo funkciją:

$$L = \sum_{i=1}^{n} w_i (y_i - f(x_i^{(1)},...x_i^{(k)}))^2; \qquad (10.7)$$

čia w_i, $i = 1, 2 \ldots n$ – svoriai, dažniausiai parenkami atvirkščiai proporcingi paklaidos dispersijai $D\varepsilon_i$. Funkcijos L minimizavimas suprantamas taip: $f(x)$ parametrai parenkami taip, kad L reikšmė būtų mažiausia. Netiesinės regresijos atveju naudojami šie svoriai $w_i = (D\varepsilon_i)^{-1/2}$, $w_i = (y_i)^{-1}$ arba $w_i = (f(\boldsymbol{x}_i))^{-1}$. Nežinomi regresijos funkcijos parametrai apskaičiuojami tikslo funkcijos L dalines išvestines parametrų atžvilgiu prilyginus 0 ir išsprendus lygčių sistemą.

3. Įvertinus nežinomus regresinio modelio parametrus, būtina patikrinti, ar turimi duomenys (x_i, y_i), $i = 1, 2 \ldots n$ atitinka pasirinktą regresinį modelį – turi būti atliekama modelio suderinamumo analizė (*significant test*) ir vertinamas duomenų atitiktis modeliui (*goodness of fit*).

Tiriant regresinio modelio suderinamumą, būtina nustatyti, ar faktoriaus X kitimu, kuris pateiktas regresijos kreive, galima paaiškinti tam tikrą dalį atsako kitimo. Jei taip nėra, parinktas regresinis modelis netinka; jį reikia pertvarkyti parinkus kitą regresijos kreivę arba konstatuoti ryšio tarp šio faktoriaus ir atsako nebuvimą. Jei $f(x)$ yra parametrinė funkcija ir konsta-

tuojama, kad dalį Y kitimo galima paaiškinti faktoriaus kitimu (tai nustatoma atmetus atitinkamą nulinę hipotezę), būtina patikrinti, ar būtent tokio pavidalo funkcija turi būti regresiniame modelyje. Pavyzdžiui, nustatyta, kad kvadratinės funkcijos $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ kitimu galima paaiškinti 20 % atsako kitimo. Tačiau gal tą pačią atsako kitimo dalį būtų galima paaiškinti paprastesne – tiesine funkcija $f(x) = \beta_0 + \beta_1 x$? Kitaip tariant, būtina patikrinti hipotezę, ar visi $f(x)$ parametrų įverčiai reikšmingai skiriasi nuo 0. Jei į regresinį modelį įtraukti keli faktoriai, būtina patikrinti hipotezę, ar visi faktoriai modelyje reikalingi, t. y. ar regresinio modelio koeficientų prie šių faktorių įverčiai reikšmingai skiriasi nuo 0 (yra reikšmingi). Jei kai kurie regresijos funkcijos koeficientų įverčiai nereikšmingi, regresijos funkciją reikia pertvarkyti: vieno faktoriaus atveju – $f(x)$ supaprastinti, kelių kintamųjų tiesinės regresijos atveju – į modelį netraukti faktorių, prie kurių esantys koeficientai nėra reikšmingi.

Konstatavus regresinio modelio tinkamumą, būtina patikrinti, ar jis gerai atspindi turimus duomenis, t. y. ar skirtumų $y_i - \hat{f}(x_i)$ skirstinys yra normalusis su vidurkiu, lygiu 0, bei nuo X nepriklausančia dispersija, ir ar skirtumai $y_i - \hat{f}(x_i)$ nėra koreliuoti; čia $\hat{f}(x_i)$ – funkcijos $f(x_i)$ įvertis.

10.3. Tiesinė vieno kintamojo regresija

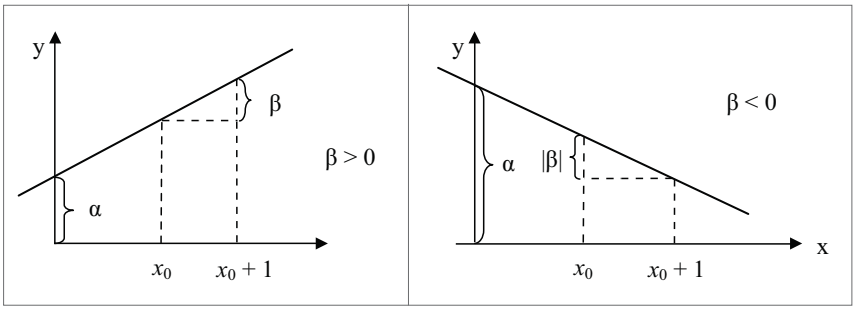
Tiesinės vieno kintamojo regresijos modelis apibrėžiamas taip:

$$y_i = \alpha + \beta x_i + \varepsilon_i; \tag{10.8}$$

čia (x_1, y_1), (x_2, y_2) ... (x_n, y_n) – individo faktoriaus ir atsako į jį reikšmės, ε_i – nepriklausomi normalieji ats. d., α ir β – skaitiniai koeficientai – modelio parametrai. Šiame modelyje apsiribosime homoskedastiškais duomenimis (10.2 a pav.), t. y. $D\varepsilon_i = \sigma^2$, $i = 1, 2 \ldots n$.

Tiesė $y = \alpha + \beta x$ vadinama regresijos tiese; α ir β – tiesės parametrai. α yra lygties laisvasis narys (*intercept*); α – tai atkarpa, kurią tiesė $\alpha + \beta x$ atkerta Y ašyje (10.4 pav.). β yra tiesės krypties koeficientas (*slope*), vadinamas regresijos koeficientu. β parodo, kiek pakinta y, kai x pakinta vienetu (10.4 pav.). Jei β yra teigiamas, didėjant x, didėja ir y; jei β neigiamas, didėjant x, mažėja y. Kuo β absoliučiu dydžiu didesnis, tuo staigiau tiesė kyla ar leidžiasi (10.4 pav.).

Kaip minėta, tiesinės regresijos prielaidos: ε_i – nepriklausomi, be to $\varepsilon_i \sim N(0, \sigma^2)$, x – neatsitiktinis bei tarp Y ir X yra tiesinė priklausomybė (modelio forma yra korektiška). Todėl esant fiksuotai faktoriaus reikšmei x, priklausomas kintamasis y turi normalųjį skirstinį su vidurkiu $\alpha + \beta x$ ir dispersija σ^2.



10.4 pav. Tiesės $y = \alpha + \beta x$ parametrų α ir β geometrinė prasmė

Regresijos modelyje (10.8) esantys parametrai α ir β – nežinomi, juos reikia įvertinti remiantis turimais duomenimis. Nežinomų parametrų α ir β įverčiai *a* ir *b* randami minimizavus sumą (*a* ir *b* funkciją)

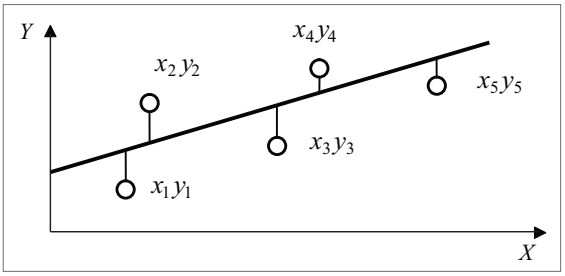
$$\sum_{i=1}^{n}(y_i - a - bx_i)^2 \qquad (10.9)$$

– atskirą tikslo funkcijos *L* atvejį su $w_i = 1$, $i = 1, 2 \ldots n$. Suma (10.9) yra y_i atstumų nuo tiesės $a + bx_i$ (10.5 pav.) kvadratų suma. Kitaip tariant, regresijos tiesės parametrų įverčiai *a* ir *b* parenkami taip, kad regresijos tiesė būtų kuo arčiau y_i reikšmių.

Nustačius funkcijos (10.9) išvestines nežinomų parametrų atžvilgiu, prilyginus jas nuliui ir išsprendus lygčių sistemą, gaunama, kad suma (10.9) yra mažiausia, kai:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}; \qquad (10.10)$$

čia $\bar{x} = \sum_{i=1}^{n} x_i / n$, $\bar{y} = \sum_{i=1}^{n} y_i / n$. Nežinomos atsitiktinės paklaidos dispersijos σ^2 įvertis s_0^2 lygus: $s_0^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-2), \hat{y}_i = a + bx_i$.



10.5 pav. y_i atstumas nuo regresijos tiesės

Parametrų įverčiai *a* ir *b* yra atsitiktiniai dydžiai (nes y_i yra atsitiktiniai), todėl būtina įvertinti ne tik jų reikšmes, bet ir kitimo rodiklį – standartinę paklaidą. *a* ir *b* standartinių nuokrypių įverčiai – standartinės paklaidos – lygūs:

$$se(a) = s_0\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2}, \quad se(b) = \frac{s_0}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}}.$$

Tiesinės regresijos lygties koeficientų α ir β pasikliautinieji intervalai konstruojami analogiškai normaliojo skirstinio vidurkio pasikliautiniesiems intervalams (4.2 skyrius):

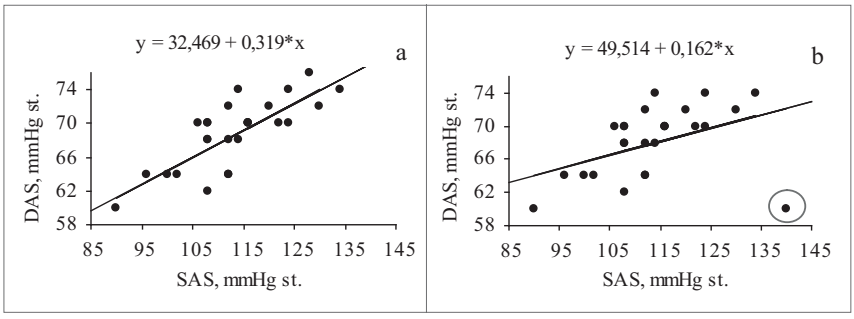
$$a \pm t_{(1+P)/2}(n-2)se(a); \quad b \pm t_{(1+P)/2}(n-2)se(b).$$

Parametrų α ir β įverčiai *a* ir *b* yra jautrūs išskirtims – ε_i nuokrypiams nuo normalumo. 10.6 pav. pateiktos regresijos kreivės parametrų įverčių reikšmės duomenims be išskirties (a) ir atsiradus išskirčiai (b). 10.6 pav. matyti, kad, atsiradus išskirčiai, regresijos koeficiento *b* reikšmė sumažėjo beveik dvigubai – nuo 0,319 iki 0,162.

Regresijos tiesės $\alpha + \beta x$ įvertis yra $a + bx$. Visai regresijos tiesei taip pat nustatoma pasikliautinoji sritis (10.7 pav.):

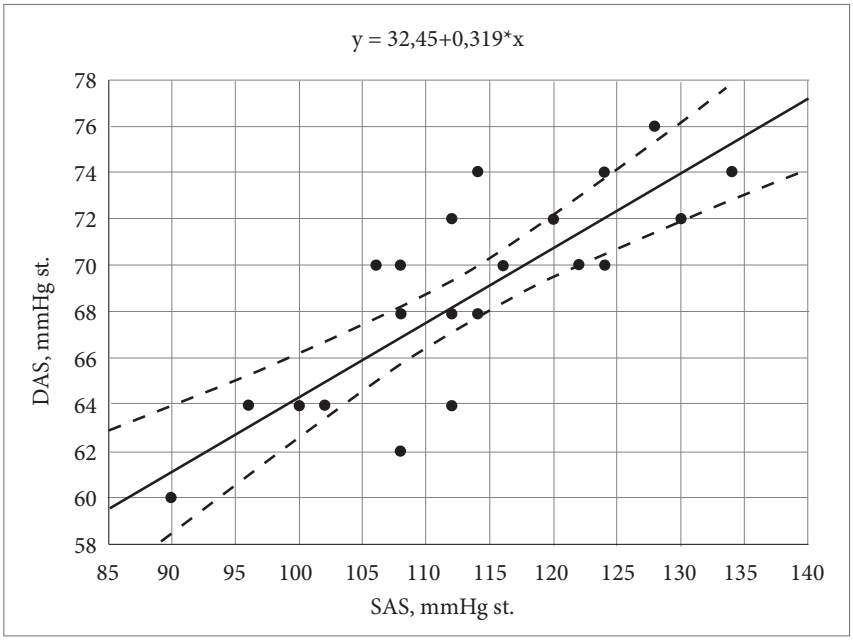
$$\bar{y} + \hat{b}(x - \bar{x}) \pm t_{(1+P)/2}(n-2)s_0\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2};$$

čia *P* – pasikliovimo lygmuo.



10.6 pav. Regresijos kreivės parametrų įverčiai:

a – be išskirties (a = 32,47; se(a) = 6,03; b = 0,319; se(b) = 0,053);
b – atsiradus išskirčiai (a = 49,51; se(a) = 7,93; b = 0,162; se(b) = 0,07)



10.7 pav. Regresijos tiesės pasikliovimo sritis

Dydžiai $\hat{y}_i = a + bx_i$ yra regresiniu modeliu įvertintos atsako reikšmės. Skirtumai tarp atsako faktinių ir modeliuotų reikšmių $e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$ vadinami likučiais (*residual*). Likučių analizė padeda įvertinti sudaryto modelio adekvatumą – atitiktį realiems duomenims.

10.4. Regresijos tiesės tinkamumo (adekvatumo) tyrimas

Regresijos tiesė reikalinga tam, kad, žinant faktoriaus reikšmę, ja remiantis būtų galima įvertinti atsako reikšmę. Tačiau regresijos modelis neturi prasmės, jei paklaidų (likučių) dispersija nėra mažesnė už atsako. Taigi, įvertinus regresijos kreivės parametrus, reikia patikrinti hipotezę, ar faktoriaus *X* tiesinis kitimas reikšmingai sumažina atsako *Y* dispersiją, t. y. ar modelyje (10.8) $\beta \neq 0$. Atsako dispersija vertinama y_i skirtumų nuo vidurkio kvadratų suma. Šią sumą galima išskaidyti į dvi dedamąsias taip:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (10.11)$$

Skirtumų nuo vidurkio kvadratų suma *SST* (*Total SS*)	Kvadratų suma, sąlygota regresijos *SSR* (*Regress SS*)	Skirtumų nuo regresijos tiesės (likučių) kvadratų suma *SSE* (*Residual SS*)

SST reikšmė vertina visą atsako kitimą, *SSR* reikšmė apibūdina, kaip $\hat{y}_i$ reikšmės skiriasi nuo atsako reikšmių vidurkio, *SSE* reikšmė vertina regresinio modelio paklaidas.

Jei konkrečioms *X* ir *Y* reikšmėms apskaičiuota atsitiktinių paklaidų nuo regresijos tiesės kvadratų sumos *SSE* reikšmė nedaug skiriasi nuo *SST*, t. y. jei *SSR* reikšmė yra palyginti nedidelė, galima įtarti, kad nebūtina faktoriaus įtraukti į regresinį modelį (10.8) – kintant *X*, atsako vidurkis tiesiškai nekinta. Todėl, analizuojant (10.8) modelio tinkamumą, tikrinama nulinė hipotezė H_0: $\beta = 0$ su alternatyva H_3: $\beta \neq 0$. Nulinei hipotezei tikrinti naudojamas F kriterijus su statistika:

$$F = \frac{SSR}{s_0^2} = \frac{(SST - SSE)}{SSE/(n-2)} = \frac{SSE_{H0} - SSE}{SSE/(n-2)};$$

čia $SSE_{H0} = SST$ – likučių kvadratų suma, kai $\beta = 0$ (teisinga H_0). Esant teisingai nulinei hipotezei, statistika *F* turi Fišerio skirstinį su (1, *n* – 2) laisvės laipsnių. Jeigu *F* kriterijaus *p* reikšmė yra mažesnė už 0,05 (arba $F > F_{0,95}(1; (n-2)$; čia $F_{0,95}(1; (n-2))$ yra Fišerio skirstinio su (1, *n* – 2) laisvės laipsnių 0,95 eilės kvantilis), galima teigti, kad regresijos tiesė reikšmingai sumažina atsako dispersiją, t. y. *Y* kinta ne tik dėl atsitiktinumo, bet ir dėl *X* kitimo. Jei $p \geq 0{,}05$, hipotezei H_0 neprieštaraujama (čia $\alpha = 0{,}05$).

Nulinei hipotezei H_0: $\beta = 0$ (kintant *X*, atsako vidurkis nekinta) tikrinti naudojamas ir t kriterijus su statistika $t = b/se(b)$. Esant teisingai nulinei hipotezei, *t* skirstinys yra Stjudento su (*n* – 2) laisvės laipsnių. Analogiškai tikrinama ir H_0: $\alpha = 0$.

Regresinių modelių koeficientai dažniausiai vertinami statistiniais paketais. Juose pateikiami koeficientų įverčiai *a* ir *b*, jų standartinės paklaidos, t kriterijaus, skirto nulinėms hipotezėms „$\beta = 0$“ ir „$\alpha = 0$“ tikrinti, statistikos ir jo dvipusės *p* reikšmės. Tikrindami H_0: $\beta = 0$ ($\alpha = 0$) su alternatyva H_3: $\beta \neq 0$ ($\alpha \neq 0$), H_0 atmetame, jei *p* mažesnė už pasirinktą reikšmingumo lygmenį. Teigiama, kad *b* reikšmingai skiriasi nuo 0. Jei $b > 0$ ir H_0 tikrinama su vienpuse alternatyva H_1: $\beta > 0$, tuomet H_0 atmetama, kai $p/2$ mažesnė už pasirinktą reikšmingumo lygmenį, ir teigiama, kad koeficientas *b* reikšmingai viršija 0. Tokiu atveju daroma išvada, kad, didėjant *X*, atsako vidurkis irgi didėja. H_0 su altenatyva ($\beta < 0$) tikrinama analogiškai.

Kiekybinis tiesinės regresijos „gerumo“ matas yra daugialypis koreliacijos koeficientas *R*, lygus koreliacijos koeficiento tarp *Y* ir *X* reikšmių moduliui, ir determinacijos koeficientas R^2 – koreliacijos koeficiento tarp *Y* ir *X* reikšmių kvadratas. R^2 išreiškiamas taip:

$R^2 = 1 - SSE/SST = SSR/SST$.

R^2 reikšmė rodo, kokią dalį atsako kitimo galima paaiškinti tiesinės regresijos modeliu – faktoriaus X kitimu. Kuo R^2 artimesnis 1, tuo regresinis modelis geresnis: tuo daugiau atsako kitimo paaiškinama faktoriaus kitimu. F kriterijus, pateiktas skyriaus pradžioje, ir yra skirtas tikrinti hipotezę H_0: $R^2 = 0$.

Nustačius, kad regresijos koeficientas β (10.8) modelyje nėra lygus 0, skirtumu $y_i - \hat{y}_i$ įvertinamos paklaidų ε_i reikšmės bei tikrinama modelio atitiktis duomenims: ar likučiai (*residuals*) $e_i = y_i - \hat{y}_i$ nekoreliuoti ir ar jų skirstinys yra normalusis su dispersija, nepriklausančia nuo X. Taip pat aktualu rasti likučių išskirtis – (x_i, y_i) reikšmes, neatitinkančias sudaryto modelio.

Ar regresinis modelis tinkamas duomenims modeliuoti, galima spręsti ir pagal likučių grafikus. Pavyzdžiui, norint nustatyti, ar e_i, $i = 1, 2 \ldots n$, dispersija vienoda visiems x_i (duomenys homoskedastiški), brėžiama (x_i, e_i) skaidos diagrama. Homoskedastiškumo prielaidą galima patikrinti ir taikant F kriterijų dviejų populiacijų dispersijų lygybei. Šiuo atveju lyginamos likučių, skaičiuotų reikšmėms $x_i \le x_0$ ir $x_i > x_0$, dispersijos; čia x_0 – reikšmė iš faktoriaus reikšmių intervalo (vidurkis, mediana, kvartilis).

Likučių $e_1, e_2 \ldots e_n$ normalumą su parametrais $(0, s_0^2)$ galima tikrinti naudojant Kolmogorovo–Smirnovo bei asimetrijos ir ekscesо kriterijus (6.7 skyrius). Nukrypimui nuo normalumo bei likučių išskirtims nustatyti kiekvienai (x_i, y_i) reikšmei statistiniuose paketuose skaičiuojama: stebėjimo įtakos indeksas (*leverage statistics*), standartizuota liekana (*studentized residuals*) bei Kuko matas (*Cook distance*).

Stebėjimo (x_i, y_i) įtakos indeksas h_i parodo, kaip regresijos koeficientų įvertčiai priklauso nuo i-tojo individo reikšmės (x_i, y_i). h_i lygus:

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

Kuo didesnę įtaką (x_i, y_i) turi a ir b dydžiui, tuo h_i didesnis. Reikšmė (x_i, y_i) laikoma išskirtimi, jei $h_i > 4/n$ arba net $h_i > 6/n$.

Standartizuota liekana SR_i yra standartizuota e_i reikšmė. Ji skaičiuojama taip: $SR_i = e_i / (s_0 \sqrt{(1 - h_i)})$. Standartizuotų liekanų vidurkis lygus 0, imties dispersija – 1; taigi SR_1, SR_2 ... SR_n skirstinys yra standartinis normalusis. Todėl SR_i reikšmės turi atitikti sigmų taisyklę (2.5 skyrius): 95 % SR_i reikšmių turi būti tarp –1,96 ir 1,96 bei praktiškai visos SR_i reikšmės absoliučiu dydžiu turi būti mažesnės už 3. Jei taip nėra, galima įtarti, kad duomenys regresinio modelio neatitinka. Reikšmė (x_i, y_i) laikoma išskirtimi, jei $|SR_i| > 3$.

Kuko matas D_i parodo, kaip pasikeičia regresijos koeficientų įverčiai, gauti naudojant visus duomenis ir duomenis be *i*-tojo individo reikšmės (x_i, y_i). Reikšmė (x_i, y_i) laikoma išskirtimi, jei $D_i > 1$. Kai kurie autoriai rekomenduoja (x_i, y_i) laikyti išskirtimi, jei $D_i > 4/(n - 2)$ ar $D_i > 4/n$.

e_i nekoreliuotumui tikrinti naudojamas **Darbino–Vatsono (*Darbin–Watson*)** kriterijus. Jis skirtas tikrinti hipotezę, ar koreliacijos koeficientas tarp ε_i ir ε_{i-1} ρ lygus 0: H_0: ρ = 0; H_A: ρ ≠ 0. Darbino–Vatsono kriterijaus statistika *d* lygi:

$$d = \left(\sum_{i=2}^{n} (e_i - e_{i-1})^2\right) / \left(\sum_{i=1}^{n} e_i^2\right).$$

Ji kinta nuo 0 iki 4. Jei *d* artima 2, labai tikėtina, kad reikšmės e_i nėra koreliuotos. Jei *d* artimas 0 arba 4, galima tvirtinti, kad koreliacijos koeficientas tarp ε_i ir ε_{i-1} ρ nelygus 0. H_0 priėmimo ar atmetimo taisyklė pateikta (10.8 pav.): pagal *n* ir pasirinktą reikšmingumo lygmenį α lentelėse parenkamos d_U ir d_L reikšmės. Jei $d < d_L$ arba $d > 4 - d_L$, H_0 atmetama ir tvirtinama, kad ρ ≠ 0 (ε_i – koreliuoti dydžiai). Jei $d_U < d < 4 - d_U$, prieštarauti H_0 nėra pagrindo (ρ = 0). Jei $d_L \le d \le d_U$ arba $4 - d_U \le d \le 4 - d_L$, statistinės išvados apie likučių koreliuotumą daryti negalima.

ρ ≠ 0	Neaišku	ρ = 0	Neaišku	ρ ≠ 0

0 d_L d_U 2 $4 - d_U$ $4 - d_L$ 4

10.8 pav. Darbino–Vatsono statistikos reikšmių interpretavimas

10.5. Tiesinės regresijos modelio sudarymo pavyzdys

10.1 pavyzdys. Žinoma: kuo didesnis naujagimio gimimo svoris (GS), tuo mažesnis procentinis jo svorio padidėjimas per 3 mėnesius. 10.2 lentelėje pateikti 32 naujagimių gimimo svorio ir jo padidėjimo tarp 70 ir 100 dienos procentais duomenys ([1]). Šie rodikliai tarpusavyje susiję: koreliacijos koeficientas tarp jų absoliučiu dydžiu yra gana didelis – 0,67. Be to, naujagimio svorio procentinis padidėjimas yra atsakas į faktorių – gimimo svorį. Todėl tikslinga sudaryti tiesinės regresijos modelį svorio padidėjimo per 3 mėn. procentui įvertinti, kai žinomas naujagimio gimimo svoris.

Tiesinės regresijos parametrų įverčiai *a* ir *b*, jų standartinės paklaidos, t kriterijaus, skirto hipotezei apie parametrų reikšmingumą (lygybę 0) tikrinti, statistikos ir atitinkama *p* reikšmė pateikti 10.3 lentelėje. Lentelės apačioje nurodyti regresinio modelio kokybės rodikliai: *R*, R^2, *F* ir *p* reikšmės bei paklaidų ε_i standartinio nuokrypio įvertis s_0.

10.2 lentelė. 32 naujagimių gimimo svorio (GS) ir jo padidėjimo tarp 70 ir 100 dienos procentais duomenys

Nr.	GS, g	Padidėjimas, %	Nr.	GS, g	Padidėjimas, %
1.	3150	68	17.	5469	27
2.	4900	63	18.	5513	60
3.	4856	66	19.	5338	71
4.	4681	72	20.	5513	88
5.	5206	52	21.	5556	63
6.	4025	75	22.	3763	88
7.	5513	76	23.	6213	53
8.	3500	118	24.	577	50
9.	3544	120	25.	3806	111
10.	3675	114	26.	5381	59
11.	5031	29	27.	5819	76
12.	5163	42	28.	4638	72
13.	5600	48	29.	4506	90
14.	5600	50	30.	5163	68
15.	5381	69	31.	4988	93
16.	5075	59	32.	4113	91

10.3 lentelė. Parametrų įverčiai, jų standartinės paklaidos, t kriterijaus statistikos ir p reikšmės

Parametrai	Parametrų įverčiai	Standartinė paklaida	*t*	*p*
a	167,87	19,883	8,44	<0,00001
b	–0,0198	0,004	–4,92	0,00003

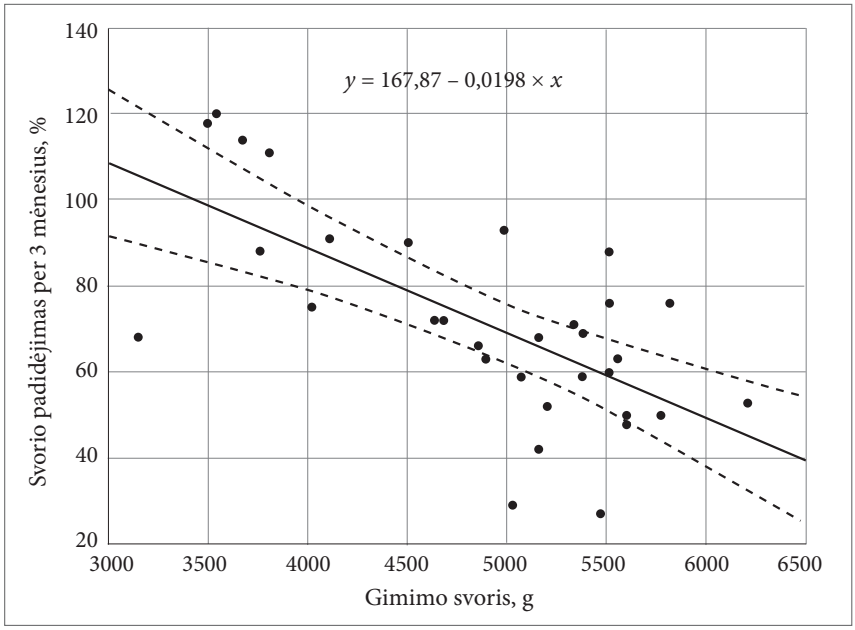
$R = 0{,}668$; $R^2 = 0{,}4465$; $F = 24{,}2$; $p < 0{,}00003$; $s_0 = 17{,}8$.

Taigi regresijos lygtis, skirta naujagimio svorio padidėjimo procentui, priklausomai nuo gimino svorio, apskaičiuoti, yra tokia: $y = 167{,}87 - 0{,}0198 \times x$, arba

(svorio padidėjimo %) = 167,87 – 0,0198 × (gimimo svoris).

Abu koeficientai yra reikšmingi, nes t kriterijaus *p* reikšmės mažesnės net už 0,001. Determinacijos koeficientas lygus 0,4465 – jis reikšmingai viršija 0, nes F kriterijaus *p* reikšmė mažesnė už 0,00003. Remdamiesi tokia determinacijos koeficiento reikšme, galime tvirtinti, kad 44,65 % svorio padidėjimo kitimo galima paaiškinti GS kitimu. Gimimo svorio ir jo procentinio padidėjimo skaidos diagrama bei regresijos tiesė su pasikliautinąja sritimi pateikta 10.9 pav. 10.4 lentelėje nurodomos faktoriaus x_i, atsako y_i, svorio

padidėjimo procentais reikšmės, įvertintos regresiniu modeliu $\hat{y}_i$, likučių $e_i = y_i - \hat{y}_i$, standartizuotų likučių SR_i, stebėjimo įtakos indekso h_i bei Kuko mato D_i reikšmės.



10.9 pav. Gimimo svorio ir jo padidėjimo procentais skaidos diagrama bei regresijos tiesė su pasikliautinąja sritimi

10.4 lentelė. Faktoriaus (x_i), faktinės (y_i) ir įvertintos ($\hat{y}_i$) atsako reikšmės, likučių e_i, standartizuotų likučių SR_i, stebėjimo įtakos indeksų h_i bei Kuko mato D_i reikšmės

Nr.	x_i	y_i	$\hat{y}_i$	e_i	SR_i	h_i	D_i
1.	3150	68	105,64	–37,64	–2,115	0,1852	0,6239
2.	4900	63	71,07	–8,07	–0,453	0,0313	0,0034
3.	4856	66	71,93	–5,93	–0,333	0,0313	0,0019
4.	4681	72	75,39	–3,39	–0,190	0,0334	0,0006
5.	5206	52	65,01	–13,01	–0,731	0,0364	0,0105
6.	4025	75	88,35	–13,35	–0,750	0,0693	0,0225
7.	5513	76	58,96	17,04	0,957	0,0510	0,0260
8.	3500	118	98,72	19,28	1,083	0,1295	0,1002
9.	3544	120	97,86	22,14	1,244	0,1234	0,1243

Nr.	x_i	y_i	$\hat{y}_i$	e_i	SR_i	h_i	D_i
10.	3675	114	95,27	18,73	1,053	0,1063	0,0737
11.	5031	29	68,47	–39,47	–2,218	0,0323	0,0848
12.	5163	42	65,88	–23,88	–1,342	0,0351	0,0339
13.	5600	48	57,24	–9,24	–0,519	0,0570	0,0086
14.	5600	50	57,24	–7,24	–0,407	0,0570	0,0053
15.	5381	69	61,56	7,44	0,418	0,0436	0,0042
16.	5075	59	67,61	–8,61	–0,484	0,0330	0,0041
17.	5469	27	59,83	–32,83	–1,845	0,0484	0,0908
18.	5513	60	58,96	1,04	0,058	0,0510	0,0001
19.	5338	71	62,42	8,58	0,482	0,0415	0,0052
20.	5513	88	58,96	29,04	1,631	0,0510	0,0754
21.	5556	63	58,10	4,90	0,275	0,0539	0,0023
22.	3763	88	93,54	–5,54	–0,311	0,0959	0,0057
23.	6213	53	45,14	7,86	0,442	0,1204	0,0152
24.	5775	50	53,78	–3,78	–0,212	0,0712	0,0019
25.	3806	111	92,67	18,33	1,030	0,0909	0,0583
26.	5381	59	61,56	–2,56	–0,144	0,0436	0,0005
27.	5819	76	52,91	23,09	1,297	0,0753	0,0740
28.	4638	72	76,25	–4,25	–0,239	0,0345	0,0011
29.	4506	90	78,84	11,16	0,627	0,0387	0,0082
30.	5163	68	65,88	2,12	0,119	0,0351	0,0003
31.	4988	93	69,34	23,66	1,330	0,0317	0,0299
32.	4113	91	86,62	4,38	0,246	0,0620	0,0021

Iš 10.4 lentelės matyti, kad nė viena SR_i reikšmė absoliučiu dydžiu neviršija 3 bei dvi (6,25 %) $|SR_i|$ reikšmės (1 ir 11) viršija 1,96. Taigi, remiantis standartizuotais likučiais, e_i normalumui prieštarauti nėra pagrindo. Nustatant išskirtis pagal stebėjimo įtakos indeksą, y_i laikytume išskirtimi, jei $h_i > 4/n = 0{,}125$ arba kai $h_i > 6/n = 0{,}187$. 10.4 lentelėje matyti dvi reikšmės: $h_1 = 0{,}1852$ ir $h_8 = 0{,}1295$ viršija 0,125; taigi 1 ir 8 individo svorio padidėjimo procento reikšmes galima įtarti esant išskirtimis. Tačiau šios abi reikšmės neviršija 0,187. Visos D_i reikšmės 10.4 lentelėje mažesnės už 1, tik $D_1 = 0{,}6239$ viršija $4/n = 0{,}125$. Taigi pagal Kuko mato, stebėjimo įtakos indekso ir standartizuotos liekanos reikšmes išskirtimi galima įtarti pirmo naujagimio svorio padidėjimo procentą.

Darbino–Vatsono statistikos reikšmė d lygi 1,39. Iš lentelės sužinome: $d_L = 1{,}26$; $d_U = 1{,}39$. Kadangi $d_L \leq d \leq d_U$, daryti išvadą apie paklaidų koreliuotumą nėra pagrindo.

10.6. Kai kurie tiesinės regresijos naudojimo aspektai

Svorinė regresija. Sakykime, kiekvienai faktoriaus reikšmei atlikta po keletą atsako matavimų ir kiekviename taške x_i apskaičiuotos šių duomenų skaitinės charakteristikos: atsako reikšmių vidurkis $\overline{y}_i$, standartinis nuokrypis s_i ir standartinė paklaida se_i (10.5 lentelė). Tokia tyrimo schema gali būti atliekama, pavyzdžiui, vertinant vaisto dozės X įtaką ligonio sistoliniam kraujospūdžiui: 10.5 lentelėje x_1, x_2 ... x_k – vaisto dozė, y_{i1}, y_{i2} ... y_{ini} – n_i ligonių, kurie vartojo x_i dozę, SAS reikšmės, $\overline{y}_i$, s_i, $se_i = s_i/\sqrt{n_i}$ – SAS vidurkis, standartinis nuokrypis ir standartinė paklaida.

10.5 lentelė. Atsako matavimai įvairioms faktoriaus reikšmėms

X reikšmė	Atsako reikšmės	Vidurkis	Standartinis nuokrypis	Standartinė paklaida
x_1	$y_{11}, y_{12} \dots y_{1n1}$	$\overline{y}_1$	s_1	se_1
x_2	$y_{21}, y_{22} \dots y_{2n2}$	$\overline{y}_2$	s_2	se_2
...	...	...	...	...
x_k	$y_{k1}, y_{k2} \dots y_{knk}$	$\overline{y}_k$	s_k	se_k

Nustačius, kad vaisto dozė daro įtaką SAS dydžiui, aktualu sudaryti regresinį modelį, skirtą įvertinti $\overline{y}_1$, $\overline{y}_2$ ... $\overline{y}_k$ pagal x_1, x_2 ... x_k reikšmes, t. y. sudaryti tiesinės regresijos modelį:

$$\overline{y}_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, 2 \dots k. \tag{10.12}$$

Jei $y_{i1}, y_{i2} \dots y_{ini}$ skirstinys yra normalusis su vidurkiu $\alpha + \beta x_i$ ir dispersija σ_i^2, tai $\overline{y}_i$ skirstinys yra normalusis su tuo pačiu vidurkiu ir dispersija σ_i^2/n_i. Taigi ε_i modelyje (10.12) yra normalusis ats. d. su nuliniu vidurkiu ir dispersija, lygia σ_i^2/n_i, ir jei σ_i^2 bei n_i yra nevienodi – 10.5 lentelės duomenys heteroskedastiški (10.3 pav.): vienai faktoriaus reikšmei Y vidurkis įvertintas patikimiau negu kitai ir t. t. Todėl tas pats regresijos tiesės atstumas iki taško $\overline{y}_i$, sakykime d, bus nevienodas standartinio nuokrypio ar standartinės paklaidos atžvilgiu – jis bus lygus d/s_i standartinių nuokrypių. Kuo s_i mažesnis, tuo d_i/s_i didesnis ir atvirkščiai – taigi regresijos tiesė turėtų būti atitinkamai arčiau taškų $\overline{y}_i$ su mažesne s_i ar se_i. Todėl regresijos tiesės parametrų įverčiai randami minimizavus (10.9) sumą su svoriais:

$$\sum_{i=1}^{k} w_i (\overline{y}_i - a - bx_i)^2;$$

čia $w_i = \frac{n(1/se_i^2)}{\sum_{i=1}^{k}(1/se_i^2)}$ arba $w_i = \frac{n(1/s_i^2)}{\sum_{i=1}^{k}(1/s_i^2)}$, $\sum_{i=1}^{k} w_i = n$.

Tuomet regresijos lygties koeficientų įverčiai lygūs:

$$b = \frac{\sum_{i=1}^{k} w_i x_i \bar{y}_i - n\bar{x}_w \bar{y}_w}{\sum_{i=1}^{k} w_i x_i^2 - n\bar{x}_w^2}, \quad a = \bar{y}_w - b\bar{x}_w, \quad \bar{x}_w = (1/n)\sum_{i=1}^{k} w_i x_i,$$

$$y_w = (1/n)\sum_{i=1}^{k} w_i \bar{y}_i.$$

Hipotezė apie regresijos tiesės parametrus. Kartais regresijos tiesė sudaroma tam, kad būtų galima palyginti kiekybinio rodiklio reikšmes, gautas skirtingais metodais. Pavyzdžiui, arterinis kraujospūdis matuojamas prietaisu A ir patobulintu prietaisu B. Žinoma, dėl atsitiktinių paklaidų abiem prietaisais tų pačių SAS reikšmių negausime; tačiau viso arterinio kraujospūdžio diapozono skirtumai tarp abiem prietaisais išmatuotų SAS reikšmių turi būti atsitiktiniai su vidurkiu, lygiu 0 (sisteminės paklaidos neturi būti), ir pastovia dispersija. Prietaisų kokybei palyginti atliekamas eksperimentas: tam pačiam individui prietaisu A nustatoma SAS reikšmė x_i, prietaisu B – reikšmė y_i. Jei patobulintas prietaisas fiksuoja SAS reikšmes be sisteminės paklaidos, turi būti teisingas toks y_i priklausomybės nuo x_i modelis: $y_i = x_i + \varepsilon_i$, čia $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$. Taigi regresijos tiesė tarp SAS matavimų B ir A prietaisais tokia: $y = x$ – modelyje (10.8) turi būti $\alpha = 0$, $\beta = 1$. Norint patvirtinti šią prielaidą, būtina tikrinti hipotezę H_0: $\alpha = 0$, $\beta = 1$.

Hipotezė apie regresijos tiesės parametrų α ir β lygybę konkrečiam skaičiui tikrinama taip. Sakykime, tikrinama H_0: $\beta = \beta_0$ su alternatyva H_3: $\beta \neq \beta_0$. H_0 tikrinti skaičiuojama t kriterijaus statistika $t = (b - \beta_0)/se(b)$. Kadangi, esant teisingai H_0, statistikos t skirstinys yra Stjudento su $(n - 2)$ laisvės laipsnių, H_0 atmetama, jei $|t| > t_{1-\alpha/2}(n - 2)$; čia α – reikšmingumo lygmuo, $t_{1-\alpha/2}(n - 2)$ – Stjudento skirstinio su $(n - 2)$ laisvės laipsnių $(1 - \alpha/2)$ lygio kvantilis. Kai $|t| \leq t_{1-\alpha/2}(n - 2)$, neprieštaraujama teiginiui, kad parametras β lygus duotam β_0. Jei H_0 tikrinama su alternatyva H_1: $\beta > \beta_0$, H_0 atmetama, kai $t > t_{1-\alpha}(n - 2)$, ir H_0 neprieštaraujama, kai $t \leq t_{1-\alpha}(n - 2)$. Analogiškai tikrinama ir hipotezė apie parametro α lygybę skaičiui: H_0: $\alpha = \alpha_0$ su alternatyva H_3: $\alpha \neq \alpha_0$. Šiuo atveju t kriterijaus statistika lygi: $t = (a - \alpha_0)/se(a)$. Minėtoms nulinėms hipotezėms tikrinti gali būti skaičiuojami regresijos tiesės parametrų α ir β pasikliautinieji intervalai. Jei į parametro α ir β pasikliautinuosius intervalus patenka atitinkamai 0 ir 1, galima tvirtinti, kad regresijos tiesė tarp rodiklio matavimų B ir A metodais yra $y = x$ ir abiem metodais gaunami vienodi rezultatai.

10.7. Kelių kintamųjų (daugialypė) tiesinė regresija

Sakykime, kintamasis Y yra atsakas į kelių tiesiškai nepriklausomų faktorių $X^{(1)}$, $X^{(2)}$ ... $X^{(k)}$ (arba daugiamačio faktoriaus $\boldsymbol{X}$) poveikį. Jei tarp kiekvieno iš šių faktorių ir Y konstatuojamas tiesinis ryšys bei suminis faktorių poveikis atsakui išreiškiamas tiesine faktorių kombinacija, tai Y priklausomybei nuo $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(k)})$ modeliuoti dažniausiai naudojamas daugialypės tiesinės regresijos modelis:

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i; \qquad (10.13)$$

čia $(x_1^{(1)}, x_1^{(2)} \ldots x_1^{(k)}, y_1) \ldots (x_n^{(1)}, x_n^{(2)} \ldots x_n^{(k)}, y_n)$ – individų faktoriaus ir atsako imtis, β_0, β_1 ... β_k – modelio koeficientai, ε_i – atsitiktinės paklaidos. Šiame modelyje daroma prielaida, kad:

- ε_i – nepriklausomi ats. dydžiai, turintys normalųjį skirstinį su vidurkiu, lygiu 0, ir dispersija σ^2;
- faktorių reikšmės yra neatsitiktinės ir tiesiškai nepriklausomos, t. y. faktoriaus $X^{(j)}$ reikšmių negalima išreikšti likusių faktorių tiesine kombinacija.

Dviejų faktorių atveju ($k = 2$) regresijos funkcija $y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)}$ yra plokštuma trimatėje erdvėje, o y_i reikšmės išsibarsčiusios aplink šią plokštumą. Kai $k > 2$, regresijos funkcija $E(Y|\boldsymbol{X}=\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_k x^{(k)}$ yra hiperplokštuma $k + 1$ matavimo erdvėje; ją įsivaizduoti sunku.

Nežinomų koeficientų β_0, β_1 ... β_k įverčiai b_0, b_1 ... b_k nustatomi mažiausių kvadratų metodu, minimizuojant funkciją L, apibrėžtą pagal (10.7) formulę su $w_i \equiv 1$ ir $f(...) = b_0 + b_1 x_i^{(1)} + \ldots + b_k x_i^{(k)}$. Kitaip tariant, koeficientai b_0, b_1 ... b_k parenkami taip, kad y_i reikšmės būtų kuo arčiau modeliuotos reikšmės $\hat{y}_i = b_0 + b_1 x_i^{(1)} + \ldots + b_k x_i^{(k)}$. Dviejų faktorių atveju koeficientai b_0, b_1, b_2 parenkami taip, kad y_i reikšmės būtų kuo arčiau plokštumos $y = b_0 + b_1 x^{(1)} + b_2 x^{(2)}$. Funkcija L minimizuojama taip: randamos L išvestinės parametrų b atžvilgiu ir prilyginamos 0. Išsprendus šią $(k + 1)$ tiesinių lygčių sistemą, nustatomi parametrų β įverčiai b. Statistiniuose paketuose pateikiami koeficientų β įverčiai b ir jų standartinės paklaidos; taip pat pateikiamas atsitiktinės paklaidos dispersijos įvertis s_0^2. Koeficiento b_i reikšmę galima interpretuoti taip: b_i parodo, kiek pakinta atsako reikšmė, $X^{(i)}$ padidėjus vienetu, kai likusių faktorių reikšmės yra fiksuotos.

Nustačius koeficientus b_0, b_1 ... b_k, pereinama prie kito regresinės analizės etapo – modelio suderinamumo analizės. Prieš interpretuojant ar naudojant (10.13) modelį, būtina nustatyti, ar reikalingi visi į modelį įtraukti faktoriai.

Gali būti, kad supaprastinus modelį, t. y. atsisakius kelių faktorių, paklaidų ε_i išsibarstymas (dispersija) liks toks pat, kaip ir sudėtingesnio modelio. Tokiu atveju geriau naudoti modelį su mažiau faktorių, nes tokį modelį paprasčiau interpretuoti, be to, mažesnis modeliuojamų reikšmių išsibarstymas.

Jei regresiniame modelyje faktorius $X^{(i)}$ nereikalingas, tai (10.13) formulėje $\beta_i = 0$ – koeficiento β_i įvertis b_i nėra reikšmingas. Todėl vertinant, ar $X^{(i)}$ reikalingas regresiniame modelyje, tikrinama nulinė hipotezė H_0: $\beta_i = 0$ su viena iš alternatyvų: H_1: $\beta_i > 0$; H_2: $\beta_i < 0$; H_3: $\beta_i \neq 0$. H_0 tikrinti naudojamas t kriterijus su statistika $t = b_i/se(b_i)$; čia $se(b_i)$ – b_i standartinė paklaida. Jei H_0 teisinga, t skirstinys yra Stjudento su $(n - k - 1)$ laisvės laipsnių. Statistiniuose paketuose pateikiama statistikos t bei kriterijaus p reikšmės. H_0 tikrinama remiantis p reikšme taip pat, kaip ir tiesinės regresijos atveju.

Pagal koeficientų reikšmingumą daroma išvada apie regresinio modelio tinkamumą. Jei nėra pagrindo prieštarauti, kad $b_i = 0$, daroma išvada, jog faktorius $X^{(i)}$ regresiniam modeliui nereikalingas. Rekomenduojama šį kintamąjį pašalinti iš (10.13) modelio, po to vėl vertinti naujo modelio koeficientus, tikrinti jų reikšmingumą ir t. t.

Ar kintamasis $X^{(i)}$ reikalingas regresiniame modelyje, galima nustatyti tikrinant tokią hipotezę H_0: „$X^{(i)}$ pašalinimas iš regresinio modelio nepadidina paklaidų dispersijos“. H_0 tikrinti naudojamas F kriterijus su statistika $F = (SSE_H - SSE) \times (n - k - 1)/SSE$; čia SSE (*Residual SS*) – likučių kvadratų suma (10.11) formulėje su $\hat{y}_i = b_0 + b_1 x_i^{(1)} + ... + b_k x_i^{(k)}$, o SSE_H – likučių kvadratų suma daugialypės regresijos modelyje, sudarytame be faktoriaus $X^{(i)}$, t. y. (10.13) formulėje $\beta_i = 0$. Jei F kriterijaus p reikšmė nėra mažesnė už pasirinktą reikšmingumo lygmenį, daroma išvada, kad $X^{(i)}$ pašalinimas iš regresinio modelio nepadidina ε_i dispersijos (arba reikšmingai nepadidina SSE), ir todėl $X^{(i)}$ regresiniame modelyje (10.15) nėra reikalingas.

Daugialypės regresijos modelio kokybei vertinti ir keliems regresiniams modeliams palyginti naudojamas determinacijos koeficientas R^2:

$R^2 = 1 - (SSE)/(SST) = (SSR)/(SST)$;

čia SSE, SSR ir SST – kvadratų sumos (10.11) formulėje su $\hat{y}_i = b_0 + b_1 x_i^{(1)} + ... + b_k x_i^{(k)}$. R^2 yra kiekybinis rodiklis, nusakantis modelio „gerumą“. Jis rodo, kurią atsako dispersijos dalį sąlygoja tiesinė k faktorių įtaka.

R^2 yra atsitiktinis dydis, todėl gali pasitaikyti ir nemaža R^2 reikšmė, nors iš tikrųjų ε_i ir Y dispersijos nesiskiria, t. y. $\beta_1 = \beta_2 = ... = \beta_k = 0$. Tokiu atveju regresinis modelis (10.13) nėra tinkamas. Determinacijos koeficiento reikšmingumui – nulinei hipotezei H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$ su alternatyva „bent

vienas $\beta_i \neq 0$“ tikrinti naudojamas F kriterijus su statistika $F = (SSR/k)/(SSE/(n - k - 1))$. Esant teisingai nulinei hipotezei, statistikos F skirstinys yra Fišerio su $(k, n - k - 1)$ laisvės laipsnių. Statistiniuose paketuose pateikiama F kriterijaus statistikos bei kriterijaus p reikšmės. Jei $p \geq 0{,}05$ (čia $\alpha = 0{,}05$), daroma išvada, kad regresijos modelis nėra tinkamas, nes, į modelį įtraukus tiesinę faktorių kombinaciją, atsako kitimas nesumažėja. Ypač aktualu įvertinti R^2 reikšmingumą tuo atveju, kai regresinis modelis sudaromas naudojant kelių ar keliolikos individų duomenis, o modelyje yra palyginti nemažai faktorių, pavyzdžiui, $n < 10$, $k = 2, 3, 4$. Tuomet gaunama, kad R^2 artimas 1, tačiau jis gali nebūti reikšmingas.

R^2 tuo didesnis, kuo mažesnė *SSE*, t. y. kuo daugiau kintamųjų įtraukta į modelį. Tačiau į modelį įtraukus papildomus kintamuosius, modeliuojamų reikšmių išsibarstymas didėja, o R^2 patikimumas mažėja (F kriterijaus p reikšmė didėja). Todėl daugialypės regresijos modelio „gerumui“ vertinti skaičiuojamas ir koreguotas determinacijos koeficientas:

AdjR^2 (*Adjasted* R^2) $= 1 - (n - 1) \times SSE/(SST \times (n - k - 1))$.

Jis parodo, kuri atsako dispersijos dalis paaiškinama tiesine k faktorių įtaka, atsižvelgiant į imties dydį ir lygtyje esančių kintamųjų skaičių. AdjR^2 tuo didesnis, kuo mažesnis dydis $SSE/(n - k - 1)$. Mažėjant *SSE*, AdjR^2 didėja, o didėjant k – mažėja. Todėl **rekomenduojama naudoti regresinį modelį su didžiausiu AdjR^2**. Regresinio modelio kokybei vertinti, be AdjR^2, naudojami ir kiti rodikliai, priklausantys nuo *SSE* ir n bei k. Dažniausiai naudojama:

- Akaike informacijos kriterijus:
  $AIC = n \ln(SSE/(n - k - 1)) + 2k$;
- Bajeso informacijos kriterijus:
  $BIC = n \ln(SSE/(n - k - 1)) + k \ln(n)$.

SSE mažėjant, *AIC* ir *BIC* mažėja, o didėjant faktorių skaičiui k – didėja. Vertinant regresinio modelio kokybę *AIC* ar *BIC* kriterijumi, rekomenduojama naudoti modelį su mažiausiomis *AIC* ar *BIC* reikšmėmis.

Kvadratinė šaknis iš determinacijos koeficiento R vadinama daugialypiu koreliacijos koeficientu. Konkrečios imties atveju R reikšmė lygi Pirsono koreliacijos koeficiento tarp atsako reikšmių ir modeliuotų reikšmių $b_1X^{(1)} + b_2X^{(2)} + \ldots + b_kX^{(k)}$ moduliui. R parodo tiesinio ryšio tarp atsako ir regresijos funkcijos reikšmių stiprumą.

Sudarius optimalų daugialypės regresijos modelį, skirtumu $e_i = y_i - \hat{y}_i$ (čia $\hat{y}_i = b_0 + b_1 x_i^{(1)} + \ldots + b_k x_i^{(k)}$) vertinamos paklaidų ε_i reikšmės. Modelio

atitiktis duomenims tikrinama naudojant e_i – kaip duomenų atitiktis tikrinama tiesinės regresijos atveju.

10.8. Optimalaus daugialypės regresijos modelio sudarymas

Optimaliu laikytinas toks daugialypės regresijos modelis, kurio AdjR^2 yra didžiausias. Sudaryti optimalų modelį sudėtinga, jei faktorių skaičius k nėra mažas. Iš k faktorių galima sudaryti $2^k - k - 1$ skirtingų daugialypės regresijos modelių (į kiekvieną modelį bus įtraukti bent 2 faktoriai). Pavyzdžiui, naudojant 6 faktorius, galima sudaryti $2^6 - 7 = 57$ skirtingus modelius. Todėl konstruojant daugialypės regresijos modelį, rekomenduotina apsiriboti tik faktoriais (kintamaisiais), medicinine ar biologine prasme susijusiais su atsaku bei reikšmingai koreliuojančiais su atsaku. Jei su atsaku reikšmingai koreliuojančių faktorių nėra daug, modeliui sudaryti galima naudoti ir tuos faktorius, kurių kriterijaus, skirto tikrinti koreliacijos tarp Y ir $X^{(i)}$ reikšmingumą, p reikšmė neviršija 0,2 (kartais net ir kai $p < 0,5$). Regresinio modelio kintamieji $X^{(i)}$ turi būti tiesiškai nepriklausomi, nes kai $X^{(i)}$ labai koreliuoti, modelio koeficientų įverčiai būna nestabilūs. Jei koreliacijos koeficientas tarp 2 faktorių absoliučiu dydžiu viršija 0,8, į modelį verta traukti tik vieną iš jų arba modelyje naudoti tam tikrą šių kintamųjų funkciją.

Optimaliam modeliui sudaryti dažniausiai naudojami: kintamųjų įtraukimo (*Forward Entry*), kintamųjų eliminavimo (*Backward Removal*), kintamųjų įtraukimo žingsninis (*Forward Stepwise*) ir kintamųjų eliminavimo žingsninis (*Backward Stepwise*) metodai. Visi jie pagrįsti žingsnine procedūra – algoritmu, leidžiančiu į modelį laipsniškai įtraukti „geriausią“ kintamąjį arba (ir) pašalinti „blogiausią“. Plačiau aprašysime kintamųjų įtraukimo žingsninį metodą.

0 žingsnis. Į modelį įtraukiame kintamąjį, kurio koreliacija su atsaku yra didžiausia. Savaime suprantama, kad ši koreliacija yra reikšminga ($p < 0,05$), nes kitaip sudaryti daugialypės tiesinės regresijos modelį nėra tikslinga. Šio modelio likučių kvadratų sumą pažymėkime SSE_1.

I žingsnis. Pažymėkime SSE_{2j} – likučių kvadratų suma, apskaičiuota papildomai į sudarytą modelį įtraukus j-tąjį kintamąjį. Kiekvienam į modelį neįtrauktam kintamajam skaičiuojame F – įtraukimo kriterijaus statistikos reikšmę $F = (SSE_1 - SSE_{2j})(n - 3)/SSE_{2j}$ bei F – įtraukimo p reikšmę. F skirta tikrinti hipotezę, ar reikšmingai sumažėjo likučių dispersija, įtraukus kintamąjį $X^{(i)}$. Iš visų p reikšmių išrenkame mažiausią p_{min} ir lyginame ją su p – įtraukimo (p – *entry*) reikšme p_E. Dažniausiai naudojama $p_E = 0,05$. Jei $p_{min} \geq p_E$, modelį sudaryti baigiame. Jei $p_{min} < p_E$, į modelį įtraukiame tą kintamąjį, kurio

F – įtraukimo p reikšmė yra mažiausia. Naujai sudaryto modelio (iš 2 faktorių) likučių kvadratų sumą pažymime SSE_2 ir vykdome II žingsnį.

L žingsnis. Sakykime, regresiniame modelyje yra L kintamųjų; šio modelio likučių kvadratų suma lygi SSE_L. Pažymėkime $SSE_{L+1,j}$ – modelio su įtrauktu j-tuoju kintamuoju likučių kvadratų suma. Kiekvienam į modelį neįtrauktam kintamajam skaičiuojame F – įtraukimo statistiką: $F = (SSE_L - SSE_{L+1,j})(n - L - 2)/SSE_{L+1,j}$ ir F – įtraukimo p reikšmę. Jei visos p viršija p_E, modelį sudaryti baigiame; jei ne, į modelį įtraukiame kintamąjį, kurio p reikšmė yra mažiausia. Šio modelio likučių kvadratų sumą žymime SSE_{L+1}.

Įtraukus į modelį naują kintamąjį, tikrinama, ar modelis nepablogės, iš jo pašalinus kurį nors kintamąjį. Dėl to kiekvienam modelyje esančiam kintamajam skaičiuojama F – eliminavimo kriterijaus statistika $F = (SSE_{L,j} - SSE_{L+1})(n - L - 2)/SSE_{L+1}$ bei šio kriterijaus p reikšmė; čia $SSE_{L,j}$ – likučių kvadratų suma, kai iš sudaryto modelio pašalintas j-tasis kintamasis. Iš apskaičiuotų p reikšmių išrenkame didžiausią, sakykime, p_{max}, ir lyginame ją su p – eliminavimo (p – *remove*) reikšme p_R. p_R parenkama didesnė už p_E, dažniausiai naudojama $p_R = 0{,}1$. Jei $p_{max} < p_R$, modelio sudarymą tęsiame ir pereiname prie $(L + 1)$ žingsnio. Jei $p_{max} \geq p_R$, iš modelio pašaliname kintamąjį, kurio F – eliminavimo p yra didžiausia, ir tęsiame modelio sudarymą.

Kintamųjų įtraukimo metodas nuo kintamųjų įtraukimo žingsninio metodo skiriasi tuo, kad į modelį įtraukti kintamieji nešalinami.

Kintamųjų eliminavimo žingsninė procedūra vykdoma taip: iš pradžių į modelį įtraukiami visi kintamieji. Po to kiekvienam jų skaičiuojama F – eliminavimo statistika, iš modelio po vieną šalinami nereikšmingi kintamieji. Jei modelį reikšmingai pagerina įtrauktas kintamasis, kuris prieš tai buvo pašalintas, į modelį tą kintamąjį ir įtraukiame. Kintamųjų eliminavimo metodas nuo atitinkamo žingsninio skiriasi tik tuo, kad pašalinti kintamieji į modelį neįtraukiami.

10.9. Netiesinė regresija

Kaip minėta 10.2 skyriuje, netiesinės regresijos modelis apibrėžiamas taip:

$y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, 2 \ldots n$;

čia $\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)} \ldots x_i^{(k)})$, y_i – i-tojo individo daugiamačio faktoriaus ir atsako reikšmės, $f(\boldsymbol{x})$ – netiesinė bendru atveju daugiamačio argumento funkcija, ε_i – nepriklausomi ats. dydžiai, $\varepsilon_i \sim N(0, \sigma^2)$. Netiesine regresija modeliuojami daugelis individe vykstančių biocheminių procesų. Pavyzdžiui, sergančių ŽIV antigeno P24 lygis E_t, skaičiuotas praėjus t laikui nuo gydymo pradžios, aprašomas tokiu regresijos modeliu [10, 12]:

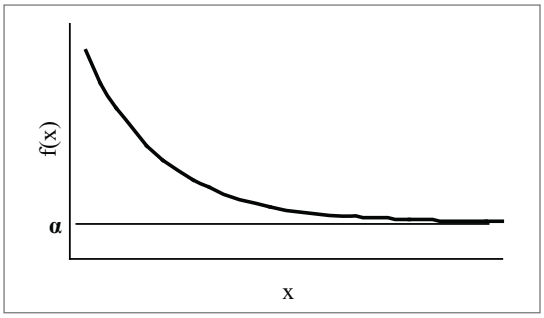
$$E_i = E_0(1-A)e^{-k_{out}t} + E_0 A + \varepsilon_i; \quad (10.14)$$

čia $\varepsilon_i \sim N(0, \sigma^2)$ – nepriklausomi ats. dydžiai; E_0, A, k_{out} yra regresinio modelio koeficientai. Juos galima interpretuoti taip: E_0 – pradinis antigeno lygis, A – antigeno P24 redukcijos koeficientas, k_{out} – antigeno mažėjimo greitis. Šiame modelyje regresijos funkcija $f(x)$ yra eksponentinė: $f(x) = \alpha + \beta\exp(\gamma x)$. Eksponentinės funkcijos parametrų α, β, γ, kai $\gamma < 0$, prasmė tokia: α – reikšmė, prie kurios, augant x, artėja funkcija, β – kreivės formos parametras, γ – mastelio parametras (10.10 pav.).

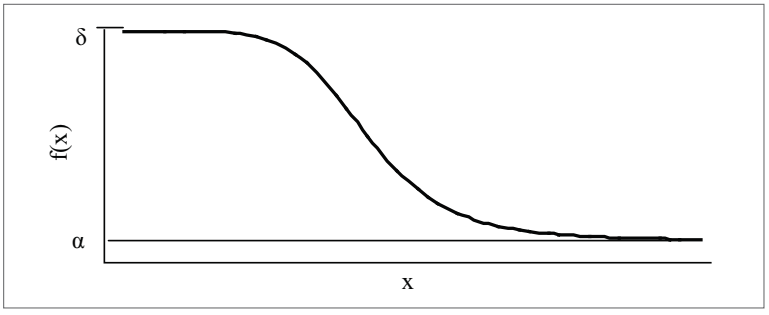
Leukocitų chemokinezės tyrimų rezultatams modeliuoti ([11, 13]) naudojama 4 parametrų logistinė regresijos funkcija (10.11 pav.):

$$f(x) = \alpha + \frac{\delta - \alpha}{1 + (x/\gamma)^{\beta}}, \; x \geq 0.$$

Šios funkcijos parametrų α, β, γ ir δ, kai $\beta > 0$, prasmė tokia: δ – didžiausia $f(x)$ reikšmė, $\delta = f(0)$, α – reikšmė, prie kurios, augant x, artėja $f(x)$. β ir γ – atitinkamai formos ir mastelio parametrai, kai $x = \gamma$, $f(x) = (\delta + \alpha)/2$.



10.10 pav. Eksponentinės funkcijos α + βexp(γx), kai γ < 0 grafikas



10.11 pav. Logistinės funkcijos grafikas

Iš pavyzdžių matyti, jog regresijos funkcija $f(\boldsymbol{x})$ parenkama taip, kad atspindėtų biologinį ryšį, esantį tarp faktorių ir atsako. Pasirinkus netiesinę regresijos funkciją $f(x)$, kitas modelio sudarymo etapas – įvertinti nežinomus $f(x)$ parametrus. Netiesinės regresijos modelį galima suvesti į tiesinę ar tiesinę daugialypę regresiją, kai:

- $f(\mathrm{x})$ yra tiesinė argumento $\boldsymbol{x}$ funkcijų atžvilgiu;
- $f(\boldsymbol{x})$ netiesinė transformacija (pvz., logaritmas, atvirkštinė funkcija ir t. t.) yra tiesinė argumentų ar jų funkcijų atžvilgiu.

Pavyzdžiui, funkcija $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ yra tiesinė kintamojo $Z_1 = x$ ir $Z_2 = x^2$ funkcija, $f(x_1, x_2) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$ – tiesinė argumentų $Z_1 = \ln(x_1)$ ir $Z_2 = \ln(x_2)$ funkcija. Tuomet nežinomi regresinio modelio parametrai ir modelio adekvatumas vertinamas taip pat, kaip ir daugialypės tiesinės regresijos atveju, tik kintamųjų $X_1, X_2, \ldots$ funkcijos keičiamos naujais kintamaisiais $Z_1, Z_2, \ldots$.

Jei $f(\boldsymbol{x})$ – multiplikatyvinė funkcija, pavyzdžiui, $f(\boldsymbol{x}) = \alpha(x_1)^{\beta}(x_2)^{\gamma}(x_3)^{\lambda}$, tada šios funkcijos logaritmas yra tiesinė kintamųjų $\ln(x_1)$, $\ln(x_2)$ ir $\ln(x_3)$ funkcija: $\ln(f(\boldsymbol{x})) = \ln\alpha + \beta \ln(x_1) + \gamma \ln(x_2) + \lambda \ln(x_3)$. Šiuo atveju atsaką Y pakeitę $\ln(Y)$, kintamąjį X_i – $\ln(X_i)$, gauname daugialypės tiesinės regresijos modelį.

Kai funkcijos $f(\boldsymbol{x})$ neįmanoma transformuoti į tiesinį pavidalą, pavyzdžiui, $f(x)$ yra eksponentinė funkcija, nežinomi parametrai vertinami iteraciniais metodais: pirmiausia parenkamos pradinės parametrų reikšmės $b_{00}, b_{10} \ldots b_{k0}$, po to tam tikromis formulėmis, į kurias įeina prieš tai nustatytos parametrų reikšmės b_{00}, b_{10} ... b_{k0}, skaičiuojama kita reikšmė b_{01}, b_{11} ... b_{k1} ir t. t. Iteracinė procedūra baigiama, jei skirtumas tarp b_j, $j = 0, 1 \ldots k$, reikšmių gautų $(i - 1)$-osios ir i-tosios iteracijos metu, absoliučiu dydžiu mažesnis už artimą 0 skaičių, sakykime, 0,0001. Tokiu atveju sakoma, kad iteracinis metodas konverguoja. Jei funkcijoje $f(\boldsymbol{x})$ yra palyginti daug nežinomų parametrų, sakykime, $k = 5, 6, \ldots$, iteracinis metodas gali ir nekonverguoti. Tuomet rekomenduojama keisti pradines parametrų reikšmes ar patį iteracinį metodą.

10.10. Neparametrinė regresija

Analizuojant ryšį tarp faktoriaus X ir atsako Y, galima susidurti su situacija, kai regresijos funkciją $E(Y|X = x) = f(x)$ sudėtinga išreikšti parametriškai. 10.12 pav. pateikti mirtingumo priklausomybės nuo amžiaus duomenys: x ašyje atidėtas amžius metais, y ašyje – mirtingumas 100 000 gyventojų. Savaime suprantama – mirtingumas su amžiumi didėja. 10.12 pav. pateikta

ir regresijos tiesė, skirta mirtingumo priklausomybei nuo amžiaus vertinti ([5]). Tačiau pastebima, kad (x_i, y_i) reikšmės aplink tiesę nėra išsibarsčiusios atsitiktinai: kai amžius 25–40 m., mirtingumo reikšmės yra regresijos tiesės apačioje; 20–22 m. ir 42–45 m. amžiuje mirtingumo reikšmės yra gerokai virš regresijos tiesės. Tokio pobūdžio realius duomenis atitinkantį kitimą parametrine funkcija išreikšti sunku. Todėl naudojamas $f(x)$ neparametrinis įvertis – realūs Y duomenys tam tikru būdu suglodinami.

Paprasčiausias $f(x)$ neparametrinis įvertis – slenkančio vidurkio kreivė. Ji gaunama taip: sakykime, faktoriaus ir atsako reikšmių poros (x_i, y_i), $i = 1, 2 \ldots n$, išdėstytos x reikšmių didėjimo tvarka ir skirtumai tarp gretimų X reikšmių vienodi, t. y. $x_{i+1} - x_i = x_{j+1} - x_j$ visiems i ir j. Paprasčiausias $f(x)$ neparametrinis įvertis taške x_i yra $(2m + 1)$ ilgio slenkančioji:

$$\hat{f}(x_i) = \frac{y_{i-m} + y_{i-m+1} + \ldots + y_i + y_{i+1} + \ldots + y_{i+m}}{2m+1},\ i = m+1 \ldots n-m.$$

Taip pat naudojamos slenkančiosios su svoriais:

$$\hat{f}(x_i) = c_1 y_{i-m} + c_2 y_{i-m+1} + \ldots + c_{m+1} y_i + c_{m+2} y_{i+1} + \ldots + c_{2m+1} y_{i+m};$$

čia $c_1 + \ldots + c_{2m+1} = 1$, $c_i \geq 0$, i = 1, ..., $2m + 1$.

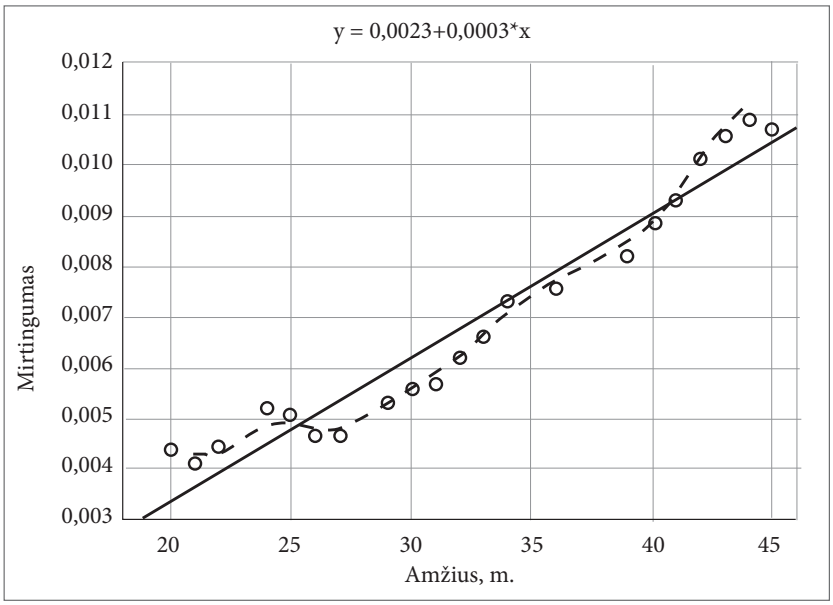
10.12 pav. pateiktas paprasčiausias neparametrinis regresijos kreivės įvertis – suglodinta kreivė (3 ilgio ($m = 1$) slenkančioji) $\hat{f}(\mathrm{x_i}) = (y_{i-1} + y_i + y_{i+1})/3$; čia x_i reikšmės pateiktos didėjimo tvarka.

Vienas neparametrinių regresijos kreivės vertinimo metodų – lokali regresija (*local regression*). Lokalios regresijos metodu $f(x)$ įvertis $\hat{f}(x)$ konstruojamas taip: kiekvienai reikšmei x skaičiuojami koeficientai a ir b, minimizuojantys tokią sumą:

$$\sum_{i=1}^{n} w((x_i - x)/h)(y_i - (a + b(x_i - x)))^2; \qquad (10.15)$$

čia $w(z) = (1 - |z|^3)^3$, kai $|z| < 1$, ir $w(z) = 0$, kai $|z| \geq 1$; x_i, y_i – i-tojo individo faktoriaus ir atsako reikšmė, $h = h(x)$ – suglodinimo lango ilgis. $h(x)$ parenkamas pagal suglodinimo parametrą (*smoothing parameter*) α; čia α – duomenų dalis, naudojama $f(x)$ vertinti. Taigi $h(x)$ parenkamas taip, kad intervale – $h \leq x_i - x \leq h$ būtų 100α % x_i reikšmių.

Sakykime, suma (10.15) yra mažiausia, kai $a = a_0$ ir $b = b_0$. Tuomet $f(x)$ įvertis yra $\hat{f}(x) = a_0$. Lokaliosios regresijos metodas naudojamas ir daugiamatės regresijos funkcijos neparametriniam vertinimui. Lokalios regresijos metodui realizuoti skirtas programų paketas *Lockfit* bei SAS programų paketo procedūra LOESS.



10.12 pav. Mirtingumo duomenys, regresijos tiesė, suglodinta kreivė

***Splain* funkcijų naudojimas.** Neparametriniam regresijos kreivės vertinimui naudojamos ir *splain* (*spline*) funkcijos. Pateiksime *splain* funkcijos apibrėžimą: intervale $[a, b]$ duota k taškų: x_0, x_1 ... x_k, vadinamų mazgais, tokių, kad $a = x_0 < x_1 < ... < x_k = b$. *Splain* funkcija $S(x)$, kai $a \le x \le b$, konstruojama taip: intervale $[x_i, x_{i+1}]$ $S(x)$ lygi n-tojo laipsnio polinomui. Polinomo koeficientai parenkami taip, kad $S(x)$ būtų vientisa – be trūkio taškų (tolydi) funkcija. *Splain* funkcijos trūkumas – už intervalo $[a, b]$ ją pratęsti sudėtinga.

Regresijos funkcijos neparametriniam vertinimui dažniausiai naudojama kubinė *splain* funkcija:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \theta_j (x - x_j)_+, \; a \le x \le b;$$

čia β_0, β_1, β_2, β_3, θ_1 ... θ_k – nežinomi *splain* funkcijos parametrai, $u_+ = 1$, kai $u > 0$, ir $u_+ = 0$, kai $u \le 0$. Nežinomi *splain* funkcijos parametrai dažniausiai vertinami mažiausių kvadratų metodu.

10.11. Apibendrinti tiesiniai modeliai

10.1–10.7 skyriuose pateiktas tiesinės regresijos (vieno kintamojo ir daugialypės) bei tolesniuose skyriuose nagrinėjami logistinės regresijos (11 skyrius),

dispersinės analizės (12 skyrius) modeliai yra platesnės regresinių modelių klasės, vadinamos **apibendrintais tiesiniais modeliais (*generalized linear model*)** (ATM), atskiri atvejai.

Apibendrinti tiesiniai modeliai charakterizuojami trimis komponentėmis:

- atsitiktine dedamąja, apibrėžiama atsako skirstiniu;
- sistemine dedamąja, išreiškiama faktorių reikšmių tiesine kombinacija;
- ryšio funkcija (*link*), apibrėžiančia funkcinį ryšį tarp sisteminės dedamosios ir atsako skirstinio vidurkio.

ATM atsitiktinė dedamoji apibrėžiama atsako Y skirstiniu, daugeliu atvejų priklausančiu eksponentinių skirstinių šeimai. Sakykime, i-tajam individui nustatyta atsako Y ir faktorių $X^{(1)}$ ... $X^{(k)}$ reikšmė yra $(y_i, x_i^{(1)} \ldots x_i^{(k)})$; y_i – nepriklausomi, faktorių reikšmės – neatsitiktinės. y_i skirstinys priklauso eksponentinių skirstinių šeimai su parametru θ_i, priklausančiu nuo faktoriaus reikšmių, t. y. $\theta_i = f(x_i^{(1)} \ldots x_i^{(k)})$. Todėl ats. d. y_i tikėtinumo funkcija – tolydžiojo skirstinio atveju tankis, diskrečiojo – tikimybė – išreiškiama pertvarkyta (2.5) formule (be dispersijos parametro):

$p(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i))$, θ_i – skirstinio parametras, $i = 1, 2 \ldots n$.

Natūralus skirstinio parametras – funkcija $Q(\theta_i)$, vadinama kanonine funkcija. Kaip minėta 2.7 skyriuje, eksponentinių skirstinių šeimai priklauso dažniausiai biomedicinos duomenims modeliuoti naudojami skirstiniai: normalusis, Bernulio, Puasono.

Sisteminė dedamoji nusakoma neatsitiktiniu dydžiu

$$\eta_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)}.$$

Trečioji ATM dedamoji – ryšio funkcija tarp atsitiktinės ir sisteminės dedamosios apibrėžiama taip:

$\eta_i = g(\mu_i) = Q(\theta_i)$, $\mu_i = Ey_i$.

Ryšys $\eta_i = Q(\theta_i)$ vadinamas kanoniniu ryšiu.

AT modelių parametrai $\beta_0, \beta_1 \ldots \beta_k$ vertinami didžiausio tikėtinumo metodu: tikėtinumo funkcijos logaritmo išvestinės parametrų atžvilgiu prilyginamos 0 ir sprendžiamos artutiniais metodais.

Pateiksime kelis ATM atvejus.

***Logit* modeliai.** Sakykime, y_i – Bernulio ats. d. su parametru $\theta_i = \pi_i = P\{y_i = 1\}$, $i = 1, 2 \ldots n$. Tuomet y_i tikėtinumo funkcija lygi:

$$p(y_i; \theta_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i} = (1-\theta_i)\exp\left(y_i \ln\left(\frac{\theta_i}{1-\theta_i}\right)\right).$$

Funkcija $Q(\theta_i) = \ln(\theta_i/(1 - \theta_i))$ yra parametro $\theta_i = \pi_i$ *logit* funkcija. Kadangi $Ey_i = \pi_i$, kanoninė ryšio funkcija lygi:

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \log it(\pi_i) \text{ arba } \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)};$$

čia $\eta_i = \beta_0 + \beta_1 x_i^{(1)} + ... + \beta_k x_i^{(k)}$.

AT modeliai su *logit* funkcija vadinami *logit* modeliais. Šių modelių atskiras atvejis – logistinė regresija (11 skyrius).

Logtiesiniai modeliai (Puasono regresija). Sakykime, y_i – Puasono ats. d. su parametru $\theta_i = \lambda_i = Ey_i$, $i = 1, 2 ... n$. Tuomet y_i tikėtinumo funkcija lygi

$$p(y_i;\lambda_i) = \frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!} = \frac{\exp(-\lambda_i)}{y_i!}\exp(y_i \ln(\lambda_i)),$$

kanoninė ryšio funkcija lygi $Q(\lambda_i) = \ln(\lambda_i)$. Taigi ryšys nusakomas logaritmine funkcija:

$\eta_i = \ln(\lambda_i)$ arba $\lambda_i = \exp(\eta_i)$.

Kai y_i skirstinys yra normalusis su visiems i pastovia dispersija, kanoninė ryšio funkcija lygi $Q(\theta_i) = \theta_i = Ey_i = \eta_i$ – turime tiesinę regresiją. AT modelių tipai pagal atsako skirstinį ir faktorių tipą pateikti 10.6 lentelėje.

10.6 lentelė. AT modelių tipai pagal atsako skirstinį ir faktorių tipą

Atsako skirstinys (ats. dedamoji)	Ryšys	Sisteminė dedamoji	Modelis
Normalusis	$\eta = \mu$	Kiekybiniai kintamieji	Tiesinė regresija (10 sk.)
Normalusis	$\eta = \mu$	Kokybiniai kintamieji	Dispersinė analizė (12 sk.)
Normalusis	$\eta = \mu$	Kiekybiniai ir kokybiniai kintamieji	Kovariancinė analizė (12.9 sk.)
Bernulio	*Logit* funkcija	Kiekybiniai ir kokybiniai kintamieji	Logistinė regresija (11 sk.)
Puasono	Logaritminė funkcija	Kiekybiniai ir kokybiniai kintamieji	Logtiesiniai modeliai (10.12 sk.)

Regresiniai modeliai, naudojantys kanoninį ryšį, yra natūralūs. Pavyzdžiui, dvinario atsako atveju galima naudoti ir tiesinės regresijos modelį:

$Ey_i = \pi_i = \alpha + \beta x_i + \varepsilon_i$.

Tačiau šio modelio trūkumas – funkcija $\alpha + \beta x_i$ gali įgyti ir neigiamas bei didesnes už 1 reikšmes, o π_i kinta tik tarp 0 ir 1.

10.12. Puasono regresija ir jos taikymas sveikatos duomenims modeliuoti

Kai kuriais mikrobiologijos ar epidemiologijos tyrimais nustatomi kintamieji, įgyjantys sveikas neneigiamas reikšmes, pavyzdžiui, sugedusių dantų skaičius, susirgimų, mirčių ar apskritai atvejų skaičius tam tikroje populiacijoje. Jei įgyjamos reikšmės nėra didelės, tokį kintamąjį galima laikyti atsitiktiniu dydžiu, turinčiu Puasono skirstinį. Jei populiacija charakterizuojama rodikliais, kuriuos galima laikyti faktoriais, pavyzdžiui, vidutiniu amžiumi, rūkymo paplitimu, socialine struktūra, tuomet susirgimų skaičių šioje populiacijoje galima laikyti atsaku į faktorių rinkinį. Šio atsako reikšmę y_i galima laikyti atsitiktiniu dydžiu, turinčiu Puasono skirstinį su parametru

$$\lambda_i(\boldsymbol{x}_i) = \exp(\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)}); \tag{10.16}$$

čia $(x_i^{(1)} \ldots x_i^{(k)})$ – i-tos populiacijos faktorių reikšmės. Jei y_i – susirgimų skaičius rajone, turinčiame N_i gyventojų, tuomet y_i laikomas Puasono ats. dydžiu su parametru

$$\lambda_i^{(1)}(\boldsymbol{x}_i) = N_i \lambda_i(\boldsymbol{x}_i) = \exp(\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)} + \ln(N_i)). \tag{10.17}$$

Puasono skirstinio parametras yra šio skirstinio vidurkis, taigi i-tosios populiacijos susirgimų skaičiaus skirstinio parametrą $\lambda_i(\boldsymbol{x})$ galima interpretuoti kaip vidutinį susirgimų skaičių šioje populiacijoje, jei $\lambda_i(\boldsymbol{x})$ apibrėžiamas pagal (10.16) formulę. Jei $\lambda_i(\boldsymbol{x})$ apibrėžiamas pagal (10.17) formulę, jis interpretuojamas kaip susirgimų skaičius tam tikram gyventojų skaičiui.

Puasono regresijos parametrai vertinami didžiausio tikėtinumo metodu. Sakykime, (y_1, y_2 ... y_n) – atsako imtis yra nepriklausomi Puasono ats. dydžiai. Šios imties tikėtinumo funkcija lygi:

$$L = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i};$$

čia $\lambda_i = \lambda(\boldsymbol{x}_i)$ apibrėžti pagal (10.16) ar (10.17) formules. Regresinio modelio parametrų β_0, β_1 ... β_k įverčiai b_0, b_1 ... b_k parenkami taip, kad L bei jos logaritmas

$$l = \ln(L) = \sum_{i=1}^{n} (y_i (b_0 + b_1 x_i^{(1)} + \ldots + b_k x_i^{(k)})) - \exp(\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)})$$

įgytų didžiausią reikšmę.

Puasono regresijos modelis su padidėjusia dispersija (*Poisson regression with overdispersion*). Realiems duomenims modeliuoti naudojant Puasono skirstinį, kartais stebimas neatitikimas tarp faktinio ir modeliuoto duomenų kitimo. Šis faktas vadinamas **dispersijos padidėjimu (*overdispersion arba extra Poisson variation*)**. Jis atsiranda dėl to, kad Puasono skirstinio vidurkis ir dispersija yra vienodi (lygūs skirstinio parametrui λ). Dispersijos padidėjimas aiškinamas tuo, kad tiriama populiacija susideda iš atskirų subpopuliacijų. Kadangi imties reikšmių priklausomybė subpopuliacijoms nežinoma, nėra galimybės į regresijos modelį įtraukti papildomą parametrą.

Modeliuojant sveikaskaitinius duomenis, padidėjusios dispersijos efektui atmesti naudojamas **nuliu praplėstas Puasono skirstinys (*zero-inflated Poisson distribution*)** su parametrais λ ir π, $0 < \pi < 1$:

$P\{Y = 0\} = \pi \times e^{-\lambda} + 1 - \pi$,

$P\{Y = k\} = \pi \times \lambda^k e^{-\lambda}/k!$, $k = 1, 2 \ldots$.

Nežinomi šio skirstinio parametrai λ ir p vertinami didžiausio tikėtinumo metodu.

10.13. Daugiapakopiai (*multilevel*) modeliai

Iki šiol analizuotuose regresiniuose modeliuose nežinomų parametrų įverčiai bei išvados buvo gautos darant prielaidą, kad atsako reikšmės y_1, y_2 ... y_n – nepriklausomi atsitiktiniai dydžiai. Tačiau daugelyje medikų ar epidemiologų fiksuojamų duomenų stebimas ryšys tarp atsako reikšmių. Pavyzdžiui, matuojamas tam tikro rajono suaugusių gyventojų arterinis kraujospūdis. Visiškai priimtinas teiginys, kad skirtingų individų arterinio kraujospūdžio reikšmės yra nepriklausomos, tačiau to negalima teigti apie giminingų asmenų tyrimus, nes stebimas paveldimas polinkis į hipertenziją. Kitas pavyzdys: ligoniai, sergantys cukriniu diabetu, gydomi skirtinguose gydymo centruose. Analizuojant šių ligonių glikemijos lygį, stebima ne tik ligonio būklės rodiklių, bet ir centro – tam tikros gydymo metodikos – įtaka. Analogiška situacija įmanoma analizuojant šalies gyventojų sergamumą, sakykime, liga X. Tam turi įtakos ne tik individo rizikos veiksniai, bet ir prevencinės priemonės, vykdomos rajone (rajono įtaka), bei atskiro gydytojo gydymo subtilybės (gydytojo įtaka).

Priklausomybė tarp minėtų rodiklių reikšmių kyla dėl individo priklausymo tam tikrai grupei (rajonui, šeimai). Šio pobūdžio duomenis galima traktuoti kaip herarchinius; tokių duomenų ryšiui faktorius → atsakas modeliuoti naudojami **daugiapakopiai (*multilevel*) modeliai**.

Daugiapakopių modelių sudarymo principus iliustruosime pavyzdžiu [1, 266 psl.]. Kvėpavimo takų infekcijos paplitimui tirti organizuota studija, kurioje dalyvavo 18 šeimų. Kiekvienoje šeimoje – po 5 asmenis: tėvas, motina ir 3 vaikai; iš viso 90 asmenų. Kiekvienam asmeniui tam tikrą laiką imti *Pneumococcus* mėginiai bei nustatytas teigiamų mėginių skaičius. Pažymėkime y_{ij} – teigiamų mėginių skaičių, nustatytą i-tosios šeimos j-tajam individui. Paprasčiausias šių duomenų statistinis modelis – y_{ij} reikšmės atsitiktinai išsibarsčiusios apie vidurkį m:

$$y_{ij} = m + \varepsilon_{ij}; \tag{10.18}$$

čia ε_{ij} – nepriklausomi ats. d. su vidurkiu, lygiu 0, ir dispersija, lygia σ^2. Tačiau šis modelis neatspindi šeimos aplinkos įtakos kvėpavimo takų infekcijos paplitimui; be to, y_{ij} kitimas tarp šeimų didesnis negu šeimų viduje. Todėl į (10.18) modelį įtraukiamas atsitiktinis dydis ξ_i, atspindintis kitimą tarp šeimų:

$$y_{ij} = m + \xi_i + \varepsilon_{ij}; \tag{10.19}$$

čia ξ_i – ats. d. su vidurkiu, lygiu 0, ir dispersija σ_F^2; ξ_i ir ε_{ij} – nepriklausomi ats. dydžiai. Pagal šį modelį ats. d. y_{ij} dispersija lygi $\sigma^2 + \sigma_{F_2}^2$, o kovariacija tarp y_{ij} ir y_{il} (reikšmių, nustatytų vienoje šeimoje) lygi σ_F^2. Koreliacijos koeficientas tarp y_{ij} ir y_{il} lygus $\sigma_F^2/(\sigma^2 + \sigma_F^2)$; jis vadinamas tarpklasinės koreliacijos koeficientu.

Minėtoje studijoje nustatytas ir gyvenamasis plotas, tenkantis vienam asmeniui. Todėl vietoj (10.19) modelio tikslinga naudoti modelį, papildytą faktoriumi:

$$y_{ij} = m + \beta x_{ij} + \xi_i + \varepsilon_{ij};$$

čia x_{ij} – i-tosios šeimos j-tajam individui tenkantis gyvenamas plotas, β – regresijos koeficientas.

Pateiksime kitą duomenų aprašymo daugiapakopiu modeliu pavyzdį. Tam tikros aštuonmečių moksleivių grupės matematikos žinios įvertintos balais. Po 3 metų kiekvieno šių moksleivių matematikos žinios vėl įvertintos atitinkamu testu. Antro testo metu gauto balo Y priklausomybė nuo prieš 3 metus nustatyto balo X įvertinta regresiniu modeliu:

$$y_i = \alpha + \beta x_i + \varepsilon_i; \tag{10.20}$$

čia (x_i, y_i) – i-tojo moksleivio balų reikšmės, ε_i – nepriklausomi ats. dydžiai, $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$, α ir β – tiesinės regresijos koeficientai. Tačiau šie koeficientai, nustatyti atskirų mokyklų moksleiviams, ganėtinai skyrėsi, nes ne visose mokyklose matematikos žinios vienodai ugdytos. Todėl daroma prielaida, kad

koeficientai α ir β – atsitiktiniai dydžiai, atspindintys mokyklos įtaką. Šiam reiškiniui įvertinti (10.20) modelis pakeičiamas dviejų pakopų modeliu:

$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij} = \alpha_0 + \xi_{0i} + \beta_0 x_{ij} + \xi_{1i} x_{ij} + \varepsilon_{ij}$;

čia α_0, β_0 – modelio koeficientai, ξ_{0i} ir ξ_{1i} – nepriklausomi atsitiktiniai dydžiai su vidurkiu, lygiu 0, (x_{ij}, y_{ij}) – i-tosios mokyklos j-tojo moksleivio balų reikšmės.

Daugiapakopio modelio parametrų vertinimas*. Nagrinėsime dviejų pakopų regresinį modelį (10.18), kuriame $i = 1 \ldots L$ – lygių skaičius, $j = 1 \ldots n$ – atitinkamo lygio individų skaičius, ξ_i – nepriklausomi ats. d., $i = 1 \ldots L$, su $E\xi_i = 0$, $D\xi_i = \sigma_F^2$; ξ_i ir ε_{ij} – nepriklausomi, ε_{ij} – nepriklausomi ats. d. su $E\varepsilon_{ij} = 0$, $D\varepsilon_{ij} = \sigma^2$. Šį modelį galima perrašyti matricos pavidalu:

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\delta}$;

čia $\boldsymbol{Y}$ – $(L \times n) \times 1$ ilgio vektorius-stulpelis (15.1 skyrius), $\boldsymbol{X}$ – $(L \times n) \times 2$ matrica, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ – dvimatis vektorius, $\boldsymbol{\delta}$ – $(L \times n) \times 1$ ilgio vektorius-stulpelis, kurio reikšmės išdėstytos „blokais“ pagal lygio reikšmes: $(\xi_1 + \varepsilon_{1j}, j = 1 \ldots n)$, $(\xi_2 + \varepsilon_{2j}, j = 1 \ldots n)$, ..., $(\xi_L + \varepsilon_{Lj}, j = 1 \ldots n)$. Remiantis ξ_i ir ε_{ij} apibrėžimu, vektoriaus $\boldsymbol{\delta}$ koordinačių vidurkiai lygūs 0, o kovariacijų matrica V, atspindinti ξ_i efektą, yra tokia:

$$V = \begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ldots & \\ & & & V_L \end{pmatrix};$$

čia V_i – $n \times n$ – matrica, kurios diagonalėje yra $\sigma^2 + \sigma_F^2$, o likusieji elementai – σ_F^2. Jei σ_F^2 ir σ^2 žinomi, parametrų β_0 ir β_1 įverčiai b_0 ir b_1 yra tokie:

$$(b_0, b_1) = (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} \boldsymbol{Y}. \qquad (10.21)$$

Kadangi σ_F^2 ir σ^2 nežinomi, β_0 ir β_1 vertinami iteraciniu metodu. Pirmiausia (10.21) formulėje daroma prielaida, kad V – vienetinė matrica. Pagal (10.21) formulę nustačius pradinius įverčius b_{00} ir b_{10}, apskaičiuojamas vektorius $\boldsymbol{\delta_0}$. Kadangi vektoriaus $\boldsymbol{\delta}$ koordinačių dispersijos lygios $\sigma^2 + \sigma_F^2$, kovariacijos tarp $\boldsymbol{\delta}$ blokų elementų lygios σ_F^2, tai σ_{F0}^2 ir σ_0^2 įvertinami atitinkamų $\boldsymbol{\delta_0}$ elementų vidurkiu. Po to skaičiuojami nauji įverčiai V_1, b_{01} ir b_{11}, $\boldsymbol{\delta_1}$ ... Iteracinę procedūrą baigiame, kai β_0 ir β_1 įverčiai beveik nesikeičia arba baigėsi nustatytas iteracijų skaičius.

10 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
3. Дрейпер Н., Смит Г. *Прикладной регрессионный анализ*. 1986. Москва: Финансы и статистика, с. 365.
4. Kruopis J. *Matematinė statistika*. 1993. Vilnius: Mokslas, 416 p.
5. Loader C. *Local Regression and Likelihood*. 1999. New York: Springer, p. 308.
6. Miller J. C., Miller J. N. *Statistics for Analycal Chemistry*. Second ed. 1988. New York: John Wiley & Sons, p. 227.
7. Olmsted S., Khann K. V., Ming E., Whitten S. T., Johnson O. N., Markham R. B., Cone R. A., Moench T. R. Low pH immobilizes and kills human leukocytes and prevents transmission of cell – associated HIV in a mouse model. 2005. *BMC Infectious Diseases*, 5, p. 79.
8. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
9. Sasomsin P., Mentre F., Diquet B., Simon F., Brun-Vezinet F. Relationship between exposure to zidovudine and decrease of P24 antigemia in HIV – infected patients in monotherapy. 2002. *Fundamental & Clinical Pharmacology*. Vol. 16, p. 347–352.
10. Себер Д. *Линейный регрессионный анализ*. 1980. Москва: Мир, с. 456.
11. *Lokalioji regresija ir jos taikymas (SAS programų paketo LOESS aprašymas)*. Prieiga per internetą: http://www.math.wpi.edu/saspdf/stat/chap38.pdf.
12. Winner L. *Introduction to Biostatistics*. 2004, p. 204. Prieiga per internetą: http://www.stat.ufl.edu/~winner/sta6934/st4170_int.pdf.
13 Nonlinear Regression: Numeric response and explanatory variables, with non-straight line **...** Source: Sasomsin et al. (2002). *Example – P24 Antigens and AZT*. Prieiga per internetą: www.stat.ufl.edu/~winner/sta6934/lognlreg.ppt.

11 SKYRIUS

Logistinė regresija

11.1. Logistinės regresijos sąvoka

Medikams aktualu nustatyti ligonio rodiklius, keliančius riziką ligai atsirasti, atsinaujinti ar mirti. Tačiau, žinant šių rodiklių reikšmes, jiems svarbu įvertinti ligos atsiradimo, atsinaujinimo ar mirties tikimybę. Taigi vėl susiduriame su modeliu faktorius → atsakas (10.1 pav.), kuriame faktorius X – ligonio būklės rodiklis, atsakas Y – dvinaris kintamasis – (serga (1), neserga (0)), (mirė (1), išgyveno (0))... . Pateiksime konkretų ryšio faktorius → dvinarinis atsakas pavyzdį.

11.1 pavyzdys ([3], I skyrius). Išeminė širdies liga (IŠL) dažnesnė vyresniems žmonėms. Todėl vertinti IŠL priklausomybę nuo amžiaus ir kitų rizikos veiksnių parengta studija. 11.1 lentelėje pateikti šimto asmenų, dalyvavusių šioje studijoje, amžius (amz) metais ir sergamumas išemine širdies liga (ISL). Kintamasis ISL koduojamas 1, jei asmuo serga, ir 0, jei neserga. 11.1 lentelėje taip pat pateiktas asmens eilės numeris (nr) ir amžiaus grupės kodas (amzgr).

Vertinant sergamumo IŠL priklausomybę nuo amžiaus, kiekvienoje amžiaus grupėje skaičiuota IŠL proporcija procentais (11.2 lentelė). 11.1 pav. pateiktas IŠL proporcijos pasiskirstymas amžiaus grupėse.

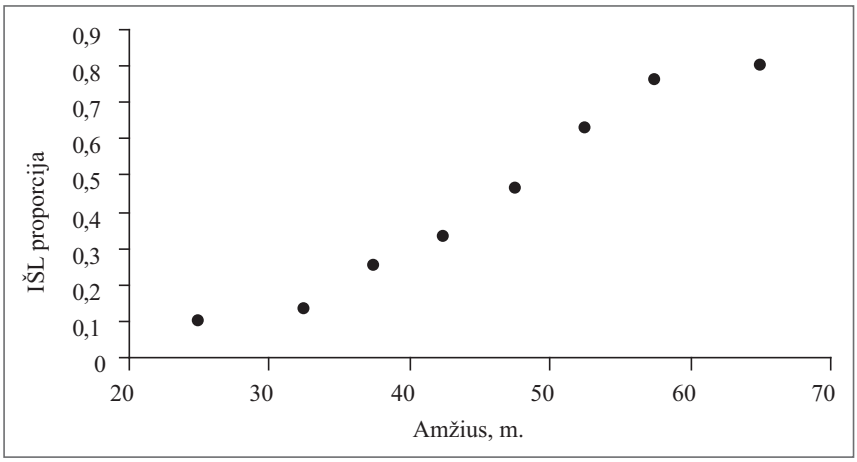
Pažymėsime, kad IŠL proporcija, skaičiuota amžiaus grupėms, yra susirgimo IŠL tikimybės atitinkamoje amžiaus grupėje įvertis. Todėl, remdamiesi 11.2 lentelėje ir 11.1 pav. pateiktais rezultatais, galime daryti išvadą, kad tikimybė susirgti IŠL jaunesniems negu 40 m. individams auga nežymiai, 45–60 m. laikotarpiu tikimybė ima augti greičiau, o po 60 m. – auga nežymiai.

11.1 lentelė. Šimto asmenų amžius ir sergamumas išemine širdies liga ([3], 3 p.)

nr	amzgr	amz	ISL	nr	amzgr	amz	ISL	nr	amzgr	amz	ISL	nr	amzgr	amz	ISL
1	1	20	0	26	3	35	0	51	4	44	1	76	7	55	1
2	1	23	0	27	3	35	0	52	4	44	1	77	7	56	1
3	1	24	0	28	3	36	0	53	5	45	0	78	7	56	1
4	1	25	0	29	3	36	1	54	5	45	1	79	7	56	1
5	1	25	1	30	3	36	0	55	5	46	0	80	7	57	0
6	1	26	0	31	3	37	0	56	5	46	1	81	7	57	0
7	1	26	0	32	3	37	1	57	5	47	0	82	7	57	1
8	1	28	0	33	3	37	0	58	5	47	0	83	7	57	1
9	1	28	0	34	3	38	0	59	5	47	1	84	7	57	1
10	1	29	0	35	3	38	0	60	5	48	0	85	7	57	1
11	2	30	0	36	3	39	0	61	5	48	1	86	7	58	0
12	2	30	0	37	3	39	1	62	5	48	1	87	7	58	1
13	2	30	0	38	4	40	0	63	5	49	0	88	7	58	1
14	2	30	0	39	4	40	1	64	5	49	0	89	7	59	1
15	2	30	0	40	4	41	0	65	5	49	1	90	7	59	1
16	2	30	1	41	4	41	0	66	6	50	0	91	8	60	0
17	2	32	0	42	4	42	0	67	6	50	1	92	8	60	1
18	2	32	0	43	4	42	0	68	6	51	0	93	8	61	1
19	2	33	0	44	4	42	0	69	6	52	0	94	8	62	1
20	2	33	0	45	4	42	1	70	6	52	1	95	8	62	1
21	2	34	0	46	4	43	0	71	6	53	1	96	8	63	1
22	2	34	0	47	4	43	0	72	6	53	1	97	8	64	0
23	2	34	1	48	4	43	1	73	6	54	1	98	8	64	1
24	2	34	0	49	4	44	0	74	7	55	0	99	8	65	1
25	2	34	0	50	4	44	0	75	7	55	1	100	8	69	1

11.2 lentelė. IŠL dažnis amžiaus grupėse

	IŠL				
Amžiaus grupė	*n*	neserga	serga	Proporcija	Proporcija procentais
20–29	10	9	1	0,10	10
30–34	15	13	2	0,13	13
35–39	12	9	3	0,25	25
40–44	15	10	5	0,33	33
45–49	13	7	6	0,46	46
50–54	8	3	5	0,63	63
55–59	17	4	13	0,76	76
60–69	10	2	8	0,80	80
Iš viso	100	57	43	0,43	43

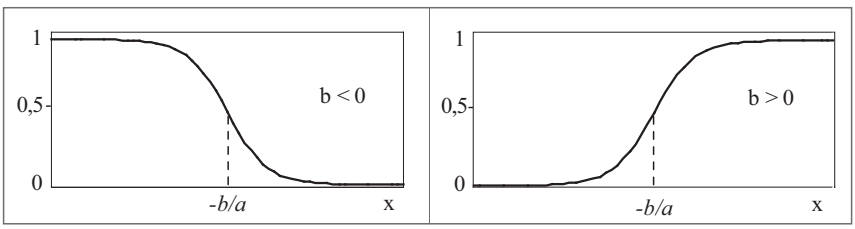


11.1 pav. IŠL proporcijos pasiskirstymas amžiaus grupėse

Analogišką 11.1 pav. pateiktą dinamiką stebėsime ir tirdami organizmo žuvimo tikimybės kitimą priklausomai nuo gautos toksinės medžiagos kiekio. Pažymėkime x – toksinės medžiagos, veikiančios gyvą organizmą, kiekį, $P\{Y = 1|x\}$ – tikimybę organizmui žūti, per tam tikrą laiko tarpą gavus toksinės medžiagos kiekį x. Kol dozė nedidelė, žuvimo tikimybė taip pat nėra didelė. Didėjant dozei, iš pradžių $P\{Y = 1|x\}$ didėja lėtai, nes gyvas organizmas mobilizuoja jėgas kovai su toksine medžiaga – stresoriumi. Tačiau nuo tam tikros kritinės dozės žuvimo tikimybė ima sparčiai didėti – organizmas nepajėgia priešintis toksinams. Kai toksinės medžiagos kiekis x yra didelis, žuvimo tikimybės $P\{Y = 1|x\}$ didėjimas su x vėl sulėtėja, nes $P\{Y = 1|x\}$ jau yra artima vienetui.

Tokį atsaką į stresoriaus poveikį gyvam organizmui gerai atspindi logistinė funkcja (11.2 pav.):

$$\frac{e^{a+bx}}{1+e^{a+bx}}.$$



11.2 pav. Logistinė funkcija

Todėl susirgimo ar žuvimo tikimybės $P\{Y = 1\}$ kitimas priklauso nuo faktoriaus X kitimo ir yra modeliuojamas logistine funkcija:

$$P\{Y = 1|X = x\} = \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}; \quad (11.1)$$

čia α ir β – skaitiniai koeficientai – modelio parametrai. Jei $\beta > 0$, tai didėjant x, didėja $\pi(x)$; jei $\beta < 0$, tai didėjant x, mažėja $\pi(x)$. (11.1) modelis vadinamas logistinės regresijos modeliu. Kaip ir tiesinėje regresijoje, (11.1) modelyje daroma prielaida, kad X yra determinuotas (neatsitiktinis) dydis.

Paaiškinsime logistinės regresijos ir tiesinės regresijos ryšį. Kaip minėta 10.11 skyriuje, logistinė regresija yra atskiras apibendrinto tiesinio modelio atvejis (kaip ir tiesinė regresija). Tiesinėje regresijoje $E(Y|X = x) = \alpha + \beta x$, o logistinėje regresijoje

$$E(Y|X = x) = P\{Y = 1|X = x\} = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Remiantis (11.1) formule, nustatoma tikimybė nesirgti: $P\{Y = 0\} = 1 - P\{Y = 1\} = (1 + \exp(\alpha + \beta x))^{-1}$, susirgimo šansas $\pi(x) / (1 - \pi(x)) = \exp(\alpha + \beta x)$ bei apibrėžiama *logit* funkcija:

$$\text{logit}\, \pi(x) = \ln\{\pi(x) / (1 - \pi(x))\} = z = \alpha + \beta x. \quad (11.2)$$

Funkcijai $\pi(x) = P\{Y = 1|X = x\}$ modeliuoti naudojama ne tik logistinė, bet ir kitos funkcijos, kintančios intervale [0, 1], pavyzdžiui:

$\pi(x) = 1 - \exp(-\exp(\alpha + \beta x))$ (log–log modelis);

$\pi(x) = (2\pi)^{-1/2} \int_{-\infty}^{\alpha+\beta x} \exp(-y^2 / 2) dy$ (*probit* modelis) ir kitos.

11.2. Logistinės regresijos modelio parametrų vertinimas

Sakykime, (x_i, y_i), $i = 1, 2 \ldots n$ – nepriklausomų tyrimų imtis. Čia y_i – i-tojo individo dvinario atsako, x_i – neatsitiktinio faktoriaus reikšmė. y_i reikšmė koduojama 1 arba 0. Daroma prielaida, kad $P\{Y = 1\}$ priklausomybė nuo faktoriaus reikšmės vertinama logistiniu modeliu (11.1) su nežinomais parametrais α ir β, kuriuos būtina įvertinti remiantis turimais duomenimis (x_i, y_i).

Tiesinėje regresijoje nežinomi parametrai vertinami mažiausių kvadratų metodu, minimizavus sumą (10.9). Bendresnis nežinomų parametrų vertinimo metodas regresiniuose modeliuose yra didžiausio tikėtinumo metodas (3.2 skyrius). Jo esmė – nežinomų parametrų įverčiai parenkami taip, kad atsako imties $(y_1, y_2 \ldots y_n)$ tikėtinumo funkcija būtų didžiausia. Kai atsako

skirstinys yra normalusis, didžiausio tikėtinumo metodas tapatus mažiausių kvadratų metodui, nes tikėtinumo funkcijos maksimumo ieškojimas yra tas pat, kaip sumos (10.9) minimizavimas.

Pateiksime logistinio modelio parametrų vertinimą didžiausio tikėtinumo metodu. Pažymėkime

$$P\{y_i = 1 | X = x_i\} = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)} = \hat{\pi}(x_i).$$

Pagal apibrėžimą (3.2 skyrius), y_i tikėtinumo funkcija lygi $[\hat{\pi}(x_i)]^{y_i}[1 - \hat{\pi}(x_i)]^{1-y_i}$, imties ($y_1$, y_2 ... y_n) tikėtinumo funkcija $L(a,b)$ bei jos logaritmas $l(a, b)$ (log–tikėtinumo funkcija) lygūs:

$$L(a, b) = \prod_{i=1}^{n} [\hat{\pi}(x_i)]^{y_i}[1 - \hat{\pi}(x_i)]^{1-y_i}, \tag{11.3}$$

$$l(a, b) = \ln(L(a, b)) = \sum_{i=1}^{n} \{y_i \ln[\hat{\pi}(x_i)] + (1 - y_i) \ln[1 - \hat{\pi}(x_i)]\}. \tag{11.4}$$

Radę (11.4) funkcijos išvestines parametrų a ir b atžvilgiu bei prilyginę jas 0, gauname lygčių sistemą:

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\pi}(x_i)) = 0, \\ \sum_{i=1}^{n} (y_i - \hat{\pi}(x_i)) x_i = 0. \end{cases}$$

Išsprendus šią sistemą artutiniais metodais, randami parametrų α ir β įverčiai a ir b. Ši vertinimo procedūra atliekama statistiniais paketais. Statistiniuose paketuose (SPSS ar STATISTICA) taip pat pateikiamos a ir b kitimo charakteristikos – standartinių nuokrypių įverčiai $se(a)$ ir $se(b)$.

Naudojant 11.1 pavyzdžio duomenis (11.1 lentelė), logistinės regresijos metodu vertinta IŠL tikimybė pagal amžių. Skaičiavimo rezultatai pateikti 11.3 lentelėje.

11.3 lentelė. Logistinės regresijos modelio, skirto IŠL tikimybei vertinti pagal amžių, parametrų įverčiai

Kintamasis	Koeficientas	Koeficiento įvertis	Standartinė paklaida	Koef./SE
Amžius	β	0,111	0,024	4,61
konstanta	α	–5,310	1,134	–4,68

Kaip matome, parametrų α ir β didžiausio tikėtinumo įverčiai yra a = –5,321 ir b = 0,111. Tikimybės susirgti IŠL įvertis lygus:

$$\hat{\pi}(amzius) = \frac{\exp(-5{,}31 + 0{,}111 \times amzius)}{1 + \exp(-5{,}31 + 0{,}111 \times amzius)}$$

11.3. Logistinio modelio tinkamumo tyrimas (*testing significance*)

Įvertinus logistinio modelio parametrus, natūraliai kyla klausimas apie sudaryto modelio tinkamumą. Gali būti, kad $P\{Y = 1\}$ nuo faktoriaus reikšmių nepriklauso ir vietoj (11.1) modelio galima naudoti paprastesnį – redukuotą modelį: $P\{Y = 1\} \equiv \pi$. Todėl būtina nustatyti, ar modelyje (11.1) reikalingas parametras β, t. y. reikia patikrinti nulinę hipotezę: $\beta = 0$.

Norint nustatyti, ar β lygus 0, būtina palyginti modeliu (11.1) įvertintas (prognozuojamas) atsako reikšmes su realiomis. Jei modeliu (11.1) įvertintos reikšmės labiau atitinka realius duomenis nei įvertintos modeliu $P\{Y = 1\} \equiv \pi$, galime tvirtinti, kad, faktorių X įtraukus į $P\{Y = 1\}$ modelį, reikšmingai pagerėja prognozės kokybė arba modelis (11.1) tinkamesnis $P\{Y = 1\}$ vertinti nei redukuotas modelis, t. y. $\beta \neq 0$. Šiuo atveju klausimas, kaip tiksliai prognozės atitinka realius duomenis, nekeliamas.

Modelio tinkamumas gerai iliustruojamas analizuojant tiesinį regresinį modelį (10.3 skyrius). Tiesinėje regresijoje atsako kitimą apibūdina visa kvadratų suma *SST*, o likučių kvadratų suma *SSE* ((10.11) formulė) apibūdina skirtumo tarp atsako ir regresinio modelio reikšmių (paklaidų) kitimą. Kitaip tariant, *SST* atspindi skirtumą tarp realių reikšmių ir vertintų (10.8) modeliu su $\beta = 0$, o *SSE* – skirtumą tarp realių reikšmių ir vertintų (10.8) modeliu. *SST* ir *SSE* skirtumas *SSR* parodo, kokią atsako išsibarstymo dalį sąlygoja regresijos funkcija. Tiesinėje regresijoje kriterijaus, skirto H_0: $\beta = 0$ tikrinti, statistika išreiškiama per *SST* ir *SSE*.

Logistinėje regresijoje elgiamasi analogiškai: lyginamos charakteristikos, atspindinčios duomenų atitiktį atsako tikimybiniam modeliui. Logistinėje regresijoje duomenų ir modelio (11.1) atitikimą charakterizuoja tikėtinumo funkcijos L arba jos logaritmo l maksimumas. Todėl vietoj F kriterijaus statistikos tiesinėje regresijoje (10.4 skyrius) skaičiuojama G statistika (tikėtinumų santykio statistika, arba χ^2 statistika):

$$G = -2\ln\frac{L_0}{L_1} = -2(l_0 - l_1) = -2\ln\frac{\text{tikėtinumo funkcija į modelį neįtraukiant faktoriaus}}{\text{tikėtinumo funkcija į modelį įtraukiant faktorių}}; \quad (11.5)$$

čia L_0 – tikėtinumo funkcijos maksimumas, kai $P\{Y = 1\} \equiv \pi$ ($\beta = 0$), L_1 – tikėtinumo funkcijos (11.3) maksimumas, $l_0 = \ln(L_0)$, $l_1 = \ln(L_1)$. Statistika G naudojama nulinei hipotezei H_0: $\beta = 0$ su alternatyva $\beta \neq 0$ tikrinti. Jei $\beta = 0$, tuomet G statistika turi asimptotinį χ^2 skirstinį su 1 laisvės laipsniu. Statistiniuose paketuose pateikiamos G statistikos ir kriterijaus p reikšmės. Jei $p < \alpha$, čia α – reikšmingumo lygmuo, tvirtiname, kad $\beta \neq 0$; o jei $p \geq \alpha$, neprieštaraujama H_0.

Analizuodami 11.1 lentelės duomenis, gavome: $l_0 = -53{,}677$, $l_1 = -68{,}322$, $G = 29{,}31$, $p < 0{,}0001$. Aišku, kad $G > \chi^2_{1-\alpha}(1)$, $\alpha = 0{,}001$, čia $\chi^2_{1-\alpha}(1) - \chi^2$ skirstinio su 1 laisvės laipsniu $1 - \alpha$ lygio kvantilis. Galime tvirtinti, jog $\beta \neq 0$ (*b* reikšmingas).

Kintamasis *X* vadinamas informatyviu rodikliu įvykiui $\{Y = 1\}$ atsirasti, jei jį įtraukus į $P\{Y = 1\}$ modelį reikšmingai sumažėja tikėtinumo funkcija. Logistinio modelio (11.1) atveju *X* yra informatyvus, jei patvirtinama alternatyva $\beta \neq 0$.

Nulinei hipotezei $\beta = 0$ tikrinti taip pat naudojami Valdo (*Wald*) ir pokyčio (*score*) kriterijai. Valdo kriterijaus statistika lygi: $W = (b/se(b))^2$. Esant teisingai nulinei hipotezei ($\beta = 0$), *W* turi asimptotinį χ^2 skirstinį su 1 laisvės laipsniu. H_0 atmetama, jei *W* reikšmė viršija 3,84 – χ^2 skirstinio su 1 laisvės laipsniu 0,95 lygio kvantilį. Nagrinėjamame pavyzdyje $W = 0{,}111/0{,}024 = 4{,}61 > 3{,}84$, todėl galima tvirtinti, kad $\beta \neq 0$.

Pokyčio kriterijaus statistika skaičiuojama:

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n} (x_i - \overline{x})^2}};$$

čia $\overline{x}$ ir $\overline{y}$ – atitinkamų kintamųjų vidurkiai. $\overline{y}$ yra parametro π įvertis redukuotame modelyje. Jei $\beta = 0$, dydžio $\sum_{i=1}^{n} x_i y_i$ vidurkis lygus $\sum_{i=1}^{n} x_i \overline{y}$, dispersija – $\overline{y}(1-\overline{y})\sum_{i=1}^{n} (x_i - \overline{x})^2$. Todėl esant teisingai H_0, *ST* turi asimptotinį normalųjį skirstinį.

11.1 lentelės duomenimis, $ST = 296{,}66/\sqrt{3333.742} = 5{,}14$. Turime $P\{|z| > 5.14\} < 0{,}001$ (čia *z* – ats. d., turintis standartinį normalųjį skirstinį), todėl H_0 su 0,001 reikšmingumo lygmeniu atmetame.

11.4. Daugialypė logistinė regresija

Kadangi daugeliu atvejų dvinarį atsaką veikia ne vienas, o keletas faktorių, būtina įvertinti šių faktorių kompleksinį poveikį atsakui. Šiam tikslui naudojama daugialypė logistinė regresija.

Sakykime, $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2) \ldots (y_n, \boldsymbol{x}_n)$, $i = 1, 2 \ldots n$ – *n* individų dvinario atsako ir *k* tiesiškai nepriklausomų neatsitiktinių faktorių (*explanatory variables, covariate*) duomenys; čia $\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)} \ldots x_i^{(k)})$. Daugialypės logistinės regresijos modelis apibrėžiamas taip:

$$P\{Y = 1 | \boldsymbol{X} = \boldsymbol{x}\} = \pi(\boldsymbol{x}) = \frac{\exp(g(\boldsymbol{x}))}{1 + \exp(g(\boldsymbol{x}))}, \; g(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_k x^{(k)}. \quad (11.6)$$

Jeigu tarp faktorių yra nominaliųjų kintamųjų, kaip antai rasė, lytis, gydymo grupė, tuomet šiuos kintamuosius į modelį galėsime įtraukti tik pavertę skaitiniais. Šiam tikslui sudaromi vadinamieji fiktyvūs kintamieji (*design variables, dummy variables*). Pavyzdžiui, turime nominalųjį kintamąjį „akių spalva“, įgyjantį 3 reikšmes: ruda, mėlyna, kita spalva. Perkodavimui į skaitines reikšmes reikalingi du fiktyvūs kintamieji D_1 ir D_2. Vienas galimų perkodavimo variantų gali būti toks (11.4 lentelė): asmeniui, kurio akys mėlynos, priskiriamos nulinės D_1 ir D_2 reikšmės. Asmeniui, turinčiam rudas akis, priskiriame $D_1 = 1$, o $D_2 = 0$. Jei akių spalva kitokia, D_1 reikšmė nustatoma nulinė, o D_2 – vienetinė. Galimi ir kiti perkodavimo variantai.

11.4 lentelė. Nominaliojo kintamojo perkodavimas

	Fiktyvūs kintamieji	
Akių spalva	D_1	D_2
mėlyna	0	0
ruda	1	0
kita	0	1

Sakykime, j-tasis faktorius $X^{(j)}$ yra nominalusis, įgyjantis p_j reikšmių. Perkodavimui reikės $p_j - 1$ fiktyvių kintamųjų D_{ju}, $u = 1, 2 \ldots p_j - 1$. Tuomet daugialypės regresijos modelyje (11.6) $g(\boldsymbol{x})$ yra tokia:

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \sum_{u=1}^{p_j - 1} \beta_{ju} D_{ju} + \ldots + \beta_k x^{(k)}.$$

Nežinomi (11.6) modelio parametrai β_0, $\beta_1 \ldots \beta_k$ vertinami didžiausio tikėtinumo metodu (žr. 3.3, 11.2 skyriai). Jų įverčiai b_0, b_1, $b_2 \ldots$ yra $k + 1$ lygčių sistemos

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\pi}(\boldsymbol{x}_i)) = 0 \\ \sum_{i=1}^{n} (y_i - \hat{\pi}(\boldsymbol{x}_i)) x_i^{(j)} = 0, \quad j = 1, 2 \ldots k \end{cases}$$

sprendiniai, gaunami artutiniais metodais; čia $\hat{\pi}(\mathbf{x})$ yra tikimybės $\pi(\boldsymbol{x})$ įvertis:

$$\hat{\pi}(\boldsymbol{x}) = \exp(\hat{g}(\boldsymbol{x}))/(1 + \exp(\hat{g}(\boldsymbol{x}))), \ \hat{g}(\boldsymbol{x}) = b_0 + b_1 x^{(1)} + b_2 x^{(2)} + \ldots + b_k x^{(k)}. \quad (11.7)$$

Kadangi įverčiai b_0, b_1, $b_2 \ldots b_k$ yra atsitiktiniai, būtina įvertinti jų kitimą, t. y. standartines paklaidas. Didžiausio tikėtinumo įverčių b_0, b_1, $b_2 \ldots b_k$ asimptotinė kovariacijų matrica vertinama matrica I^{-1}, atvirkštine informacijos matricai I:

$$I = -\frac{\partial^2 l}{\partial b_l \partial b_j} = \sum_{i=1}^{n} x_i^{(l)} x_i^{(j)} \pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i)) = X^T \hat{V} X;$$

čia X – daugiamačio faktoriaus reikšmių matrica, turinti $(k + 1)$ stulpelį ir n eilučių. Pirmame stulpelyje yra vienetai, antrame, trečiame ... ir $(k + 1)$ stulpeliuose – 1, 2 ... k-tojo faktoriaus reikšmės. $\hat{V}$ matrica yra $n \times n$ diagonalinė matrica; jos i-tasis diagonalinis elementas lygus $\hat{\pi}(\boldsymbol{x}_i)(1 - \hat{\pi}(\boldsymbol{x}_i))$. Matricos I^{-1} diagonalinis elementas yra $se^2(b_j)$; čia $se(b_j)$ – įverčio b_j standartinė paklaida.

Pateiksime daugiamačio logistinio modelio taikymo pavyzdį.

11.2 pavyzdys ([3, 4 skyrius]). Tirta, kaip motinos būklė nėštumo metu susijusi su naujagimio mažu gimimo svoriu. Laikoma, kad naujagimio gimimo svoris yra mažas, jei gimęs jis svėrė mažiau nei 2500 gramų. 1986 m. Beisteito (*Baystate*) medicinos centre (Spingfildas, Masačiusetso valstija, JAV) surinkti duomenys apie 189 moterų būklę ir nėštumo eigą. 59 motinos pagimdė mažo gimimo svorio kūdikius, o 130 – normalaus svorio. Buvo fiksuotas visų moterų amžius, svoris nėštumo pradžioje, rasė, lankymosi dažnis poliklinikoje pirmą nėštumo pusę bei kiti rodikliai.

Logistiniu modeliu vertinta mažo gimimo svorio tikimybė pagal motinos būklę. Logistiniame modelyje naudoti šie faktoriai: amžius (amz), rasė, svoris (svor), rūkymas (ruk = 1 – motina rūko, ruk = 0 – nerūko), arterinė hipertenzija (AH = 1 – serga arterine hipertenzija, AH = 0 – neserga) bei lankymosi konsultacinėje poliklinikose pirmą nėštumo pusę dažnis (LD). Rasė yra nominalusis kintamasis, įgyjantis reikšmes „baltasis“, „juodasis“, „kita“, todėl naudojome du fiktyvius kintamuosius: R1=0 ir R2=0, jei asmuo baltasis; R1 = 1, R2 = 0, jei asmuo juodasis; R1 = 0, R2 = 1, jei kita. Logistinio modelio koeficientų įverčiai ir jų standartinė paklaida pateikti 11.5 lentelėje.

11.5 lentelė. Mažo gimimo svorio tikimybės daugiamačio logistinio modelio parametrų įverčiai

Kintamasis	β įverčiai	β standartinė paklaida	Valdo statistika	Valdo statistikos *p* reikšmė
amžius(amz)	–0,022	0,036	0,402	0,526
svoris (svor)	–0,039	0,015	6,429	0,011
R1	1,243	0,527	5,567	0,018
R2	0,929	0,429	4,687	0,030
ruk	1,073	0,391	7,539	0,006
AH	1,749	0,693	6,375	0,012
LD	0,037	0,167	0,046	0,830
Konstanta	0,789	1,157	0,465	0,495

Logtikėtinumas = –103,9.

Funkcijos $g(\boldsymbol{x})$ (11.6) įvertis $\hat{g}(\boldsymbol{x})$ yra:

$\hat{g}(\boldsymbol{x}) = 0{,}789 - 0{,}022 \times \text{amz} - 0{,}039 \times \text{svor} + 1{,}243 \times \text{R1} + 0{,}929 \times \text{R2} + 1{,}073 \times \text{ruk} + 1{,}748 \times \text{AH} + 0{,}037 \times \text{LD}.$

Nustatę $\hat{g}(\boldsymbol{x})$, randame ir mažo gimimo svorio tikimybės įvertį $\hat{\pi}(\boldsymbol{x}) = \exp(\hat{g}(\boldsymbol{x}))/(1 + \exp(\hat{g}(\boldsymbol{x})))$.

11.5. Daugiamačio logistinio modelio tinkamumo analizė (*testing significance*)

Sudarius daugiamatį logistinį modelį, t. y. įvertinus modelio (11.6) nežinomus parametrus, būtina analizuoti modelio tinkamumą. Gali būti, kad, atsisakius kelių faktorių, modelis liks toks pat informatyvus, kaip ir modelis su daugiau faktorių. Todėl, įvertinus modelio parametrus, būtina patikrinti jų reikšmingumą – tikrinti H_0: $\beta_j = 0$ (t. y. ar $X^{(j)}$ (11.6) modeliui reikalingas).

Jei j-tasis kintamasis $X^{(j)}$ įgyja skaitines reikšmes (nėra nominalusis), tuomet hipotezei H_0: $\beta_j = 0$ su alternatyva $\beta_j \neq 0$ tikrinti naudojamas Valdo kriterijus su statistika $W = (b_j/se(b_j))^2$. Jeigu $\beta_j = 0$, tuomet W turi asimptotinį χ^2 skirstinį su 1 laisvės laipsniu.

Nominaliojo kintamojo atveju Valdo statistika skaičiuojama kitaip. Sakykime, j-tasis faktorius yra nominalusis, $\boldsymbol{b}_j$ – koeficientų prie m–1 fiktyvių kintamųjų maksimalaus tikėtinumo įverčio vektorius, C – vektoriaus $\boldsymbol{b}_j$ kovariacijų matricos įvertis. Tuomet Valdo statistika apibrėžiama:

$$W_j = \boldsymbol{b}_j^T C^{-1} \boldsymbol{b}_j.$$

Jei $\boldsymbol{\beta}_j = 0$, tuomet W_j asimptotinis skirstinys yra χ^2 su $(m - 1)$ laisvės laipsnių.

Kadangi logistinio modelio gerumas vertinamas logtikėtinumo funkcija, tai H_0: $\beta_j = 0$ tikrinti naudojamas ir tikėtinumų santykio kriterijus su statistika:

$$G = -2\ln(L_0/L_1) = -2\ln \frac{\text{tikėtinumo funkcija be } X(j)}{\text{tikėtinumo funkcija su } X(j)}.$$

Esant teisingai nulinei hipotezei, statistikos G asimptotinis skirstinys yra χ^2 su 1 laisvės laipsniu. Remiantis šiuo faktu, galima tvirtinti: kai $G < \chi^2_{1-\alpha}(1)$, modelio su $X^{(j)}$ ir be $X^{(j)}$ tikėtinumo funkcijos reikšmingai nesiskiria ir (11.6) modelis be $X^{(j)}$ teikia tiek pat informacijos, kiek ir su $X^{(j)}$. Jei $G > \chi^2_{1-\alpha}(1)$ arba G kriterijaus $p < \alpha$, tvirtinama, kad, įtraukus $X^{(j)}$, modelis pagerėja.

Analogiškai norima nustatyti, ar į daugiamatį logistinį modelį papildomai įtraukus kurį nors faktorių rinkinį (bloką) ($X^{(i1)}$, $X^{(i2)}$, ..., $X^{(ip)}$), $p = 1, 2 \ldots k$,

pagerėja $P\{Y = 1\}$ įvertis. Norint atsakyti į šį klausimą, reikia tikrinti tokią nulinę hipotezę:

$H_{0p}: \beta_{i_1} = \beta_{i_2} = ... = \beta_{i_p} = 0$

su alternatyva: bent vienas $\beta_{i_j} \neq 0$.

H_{0p} tikrinti naudojama G statistika (*block Chi-square*):

$$G = -2\ln \frac{\text{tikėtinumo funkcija be papildomo } p \text{ faktorių rinkinio}}{\text{tikėtinumo funkcija su papildomu } p \text{ faktorių rinkiniu}}.$$

Esant teisingai nulinei hipotezei, statistikos G asimptotinis skirstinys yra χ^2 su p laisvės laipsnių.

Ar daugiamatis logistinis modelis apskritai tinkamas naudoti, t. y. ar įtraukus faktorius $X^{(1)}$, $X^{(2)}$ ... $X^{(k)}$ į logistinį modelį, pagerėja $P\{Y = 1\}$ įvertis, nustatoma patikrinus nulinę hipotezę:

$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$ su alternatyva:

H_A: bent vienas $\beta_j \neq 0$.

H_0 tikrinti naudojama tokia tikėtinumų santykio statistika (*model Chi-square*):

$$G = -2\ln\frac{L_0}{L_1} = -2(l_0 - l_1) = -2\ln \frac{\text{tikėtinumo funkcija be faktorių}}{\text{tikėtinumo funkcija į modelį įtraukiant } k \text{ faktorių}}.$$

Jei H_0 teisinga, statistikos G asimptotinis skirstinys yra χ^2 su k laisvės laipsnių.

Nulinei hipotezei $H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$ tikrinti taip pat naudojami Waldo ir pokyčio kriterijai. Valdo kriterijaus statistika šiuo atveju lygi:

$$W = \hat{b}^T (X^T \hat{V} X)\hat{b}.$$

Esant teisingai nulinei hipotezei, W asimptotinis skirstinys yra χ^2 su k laisvės laipsnių.

Koeficientų reikšmingumo tikrinimo pavyzdys. Nagrinėtame 11.2 pavyzdyje (11.5 lentelė) į logistinį modelį įtraukus kintamuosius amz, svor, R1, R2, ruk, AH, LD, tikėtinumo funkcijos logaritmo reikšmė l_1 lygi –103,9. Darant prielaidą, kad mažo gimimo svorio tikimybė pastovi, tikėtinumo funkcijos logaritmo reikšmė l_0 lygi –117,3. Todėl G = –2((–117,3) – (–103,9)) = –2(–13,4) = 26,8. Esant teisingai H_0 (11.5 lentelėje pateikti kintamieji neturi įtakos mažo gimimo svorio tikimybei), statistikos G asimptotinis skirstinys yra χ^2_7. Tačiau $P\{\chi^2_7 > 26{,}8\} < 0{,}0001$, todėl nulinę hipotezę su reikšmingumu $\alpha = 0{,}05$ atmetame ir priimame alternatyvą – bent vienas iš koeficientų b_j, j = 1, 2 ... 5 nelygus 0.

11.5 lentelėje, be logistinio modelio koeficientų įverčių, pateikta ir Valdo statistikos bei atitinkama *p* reikšmės. Lentelėje matome, kad koeficiento prie kintamojo LD Valdo statistikos *p* reikšmė yra didžiausia ir lygi 0,83; taigi šis koeficientas nuo 0 reikšmingai nesiskiria. Logistinio modelio, sudaryto be kintamojo LD, logtikėtinumas lygus –103,94. Jis beveik nesiskiria nuo pateikto 11.5 lentelėje – taigi LD modelyje nereikalingas. Logistiniame modelyje, sudarytame be LD, koeficiento prie kintamojo amz Valdo kriterijaus *p* reikšmė lygi 0,545 > 0,05; taigi kintamasis amz modelyje irgi nereikalingas. Sudarę logistinį modelį be šio kintamojo, gavome, kad visi modelio koeficientai reikšmingai skiriasi nuo 0 (11.6 lentelė). Šio modelio logtikėtinumas lygus –104,1.

11.6 lentelė. Optimalaus logistinio modelio, skirto mažo gimimo svorio tikimybei vertinti, koeficientų įverčiai, Valdo statistikos p reikšmė bei rizikos įverčiai

Kintamasis	*b*	Valdo statistikos *p* reikšmė	e^b	e^{β} pasikliautinieji intervalai
svoris (svor)	–0,040	0,008	0,961	0,933–0,990
R1	1,288	0,014	3,624	1,304–10,075
R2	0,944	0,026	2,569	1,121–5,891
ruk	1,072	0,006	2,920	1,366–6,241
AH	1,749	0,011	5,750	1,485–22,268
Konstanta	0,352	0,7	–	–

Logtikėtinumas = –104,1.

Skirtumas tarp 11.6 ir 11.5 lentelėse pateiktų modelių logtikėtinumo reikšmių lygus 0,2, $G = 0{,}4$. Remiantis 11.6 lentelėje pateiktais koeficientų įverčiais, mažo gimimo svorio tikimybė lygi: $\hat{\pi}(\boldsymbol{x}) = \exp(\hat{g}(\boldsymbol{x}))/(1 + \exp(\hat{g}(\boldsymbol{x})))$; čia

$$\hat{g}(\boldsymbol{x}) = 0{,}352 - 0{,}04 \times \text{svor} + 1{,}288 \times \text{R1} + 0{,}944 \times \text{R2} + 1{,}072 \times \text{ruk} + 1{,}749 \times \text{AH}. \quad (11.8)$$

Logistinis modelis (11.8) tam tikra prasme yra optimalus – visi koeficientai reikšmingai skiriasi nuo 0, papildomai įtraukus kintamąjį, modelis nepagerėja, modelio logtikėtinumo funkcija yra didžiausia.

Atrenkant kintamuosius daugiamačiam modeliui, būtina atsižvelgti į jų medicininę prasmę bei kuo labiau pagrįsti jų įtaką atsakui. Faktas, kad modelio (11.1) *G* kriterijaus *p* reikšmė mažesnė už 0,05, dar nerodo priežasties ryšio tarp faktoriaus ir dvinario atsako. Kintamieji $X^{(j)}$ logistiniame modelyje turi būti tiesiškai nepriklausomi; jei koreliacijos koeficientas tarp dviejų

faktorių absoliučiu dydžiu viršija 0,8, rekomenduojama į modelį traukti tik vieną iš jų. Į daugiamatį modelį rekomenduojama traukti tik tuos kintamuosius, kurių G kriterijaus p reikšmė vienmačiame modelyje mažesnė už 0,2. Jei n nėra didelis arba įvykių $\{Y = 1\}$ nėra daug, į daugiamatį modelį traukiame tik informatyvius kintamuosius (G kriterijaus $p < 0{,}05$). Kintamųjų įtraukimo (*forward stepwise*) ar kintamųjų eliminavimo (*backward stepwise*) žingsninis metodas vykdomas taip pat, kaip ir daugialypės regresijos atveju (10.8 skyrius), tik vietoj F – įtraukimo ir F – eliminavimo statistikos naudojama G statistika (11.7).

11.6. Logistinio modelio parametrų interpretacija. Rizikos vertinimas. Rizikos balo skaičiavimas

Sakykime, faktorius X įgyja dvi reikšmes: 0 (pavyzdžiui, rizikos veiksnio (RV) nėra) ir 1 (RV yra). Tuomet į (11.1) formulę įrašę X reikšmę 0 ir 1, gauname:

$P\{Y = 1|X = 0\} = \exp(\alpha)/(1 + \exp(\alpha)) = \pi(0)$,

$P\{Y = 1|X = 1\} = \exp(\alpha + \beta)/(1 + \exp(\alpha + \beta)) = \pi(1)$.

Naudodami *logit* funkcijos apibrėžimą (11.2), išreiškiame α ir β:

$$\begin{aligned}&\alpha = \text{logit}(\pi(0)) = \ln \pi(0)/(1 - \pi(0)),\\ &\alpha + \beta = \text{logit}(\pi(1)) = \ln \pi(1)/(1 - \pi(1)),\end{aligned} \qquad (11.9)$$

$$\beta = \text{logit}(\pi(1)) - \text{logit}(\pi(0)) = \ln \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}.$$

Analogiškai parametro β įvertis $b = \text{logit}(\hat{\pi}(1)) - \text{logit}(\hat{\pi}(0)) = \ln \frac{\hat{\pi}(1)/(1-\hat{\pi}(1))}{\hat{\pi}(0)/(1-\hat{\pi}(0))}$.

Pagal apibrėžimą (8.3 skyrius), tikimybių santykis $\pi(1)/(1 - \pi(1))$ ir $\pi(0)/(1 - \pi(0))$ yra įvykio $\{Y = 1\}$ šansas, kai faktoriaus reikšmės atitinkamai lygios 1 ir 0. Epidemiologinėse studijose rizikos veiksnio ($X = 1$) įtaka susirgimui $\{Y = 1\}$ vertinama rizikos santykiu (šansų santykiu) θ. θ parodo, kiek kartų susirgimo šansas, esant RV, yra didesnis nei susirgimo šansas, nesant RV. Pagal apibrėžimą,

$$\theta = \frac{P\{Y = 1 \mid x = 1\}/(1 - P\{Y = 1 \mid x = 1\})}{P\{Y = 1 \mid x = 0\}/(1 - P\{Y = 1 \mid x = 0\})} = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = e^{\beta}. \qquad (11.10)$$

Populiacijos rizikos santykio θ įvertis yra OR. Pagal (11.9–11.10) formules, $OR = \exp(b)$. Taigi, jei $P\{Y = 1\}$ apibrėžiama (11.1) formule su X reikšmėmis 0 (RV nėra) ir 1 (RV yra), tuomet $b = \ln(OR)$.

Jei X yra kiekybinis kintamasis, pagal apibrėžimą:

$$b = \text{logit}(\hat{\pi}(x+1)) - \text{logit}(\hat{\pi}(x)) = \ln\frac{\hat{\pi}(x+1)/(1-\hat{\pi}(x+1))}{\hat{\pi}(x)/(1-\hat{\pi}(x))}.$$

Dydis $\frac{\hat{\pi}(x+1)/(1-\hat{\pi}(x+1))}{\hat{\pi}(x)/(1-\hat{\pi}(x))} = e^b$ parodo, kiek kartų padidėja susirgimo ($Y = 1$) rizikos santykis, faktoriaus reikšmei padidėjus 1.

Parametrų interpretacijos pavyzdys. 11.1 pavyzdyje pateikti tyrimo, skirto IŠL priklausomybei nuo amžiaus vertinti, duomenys, o IŠL tikimybės įverčio pagal amžių logistinės funkcijos parametrai – 11.3 lentelėje. Koeficientas b lygus 0,111, e^b = 1,117, todėl galima daryti išvadą, kad, padidėjus amžiui 1 metais, rizika susirgti IŠL vidutiniškai padidėja 1,117 karto.

Sakykime, į logistinį modelį (11.6) įtraukiami du faktoriai: $X^{(1)}$, įgyjantis reikšmę 1 arba 0, ir $X^{(2)}$, įgyjantis m reikšmių. Tuomet koeficientas b_1 lygus koreguoto rizikos santykio, skaičiuoto izoliavus $X^{(2)}$ įtaką, logaritmui: $b_1 = \ln(OR_K)$.

Daugiamačio logistinio modelio koeficientas b_j parodo, kiek, padidėjus j-tojo faktoriaus reikšmei 1, padidėja susirgimo ($Y = 1$) šansų santykio, koreguoto pagal kitus faktorius, logaritmas. Taigi $\exp(b_j)$ yra kintamojo $X^{(j)}$ rizikos santykis, koreguotas pagal likusius (11.6) modelio faktorius; $\exp(b_j)$ vadinamas j-tojo faktoriaus **standartizuotu (koreguotu) rizikos santykiu**. Dydis e^b, nustatytas vieno kintamojo logistiniame modelyje, vadinamas **izoliuotu rizikos santykiu**. Statistiniuose paketuose, be b_j ir $se(b_j)$, pateikiamas ir dydis $\exp(b_j)$ su pasikliautinaisiais intervalais. 11.6 lentelėje pateikti kintamųjų, naudotų logistiniame modelyje mažo gimimo svorio tikimybei vertinti, koreguoti rizikos santykiai su pasikliautinaisiais intervalais. Matome, kad rūkančiai motinai mažo gimimo svorio rizika padidėja 2,9 karto, sergančiai arterine hipertenzija – 5,75 karto; juodaodės moters rizika pagimdyti mažo gimimo svorio kūdikį 3,62 karto didesnė, palyginti su baltaode.

Remiantis daugialypės logistinės regresijos modeliu, nepalankaus įvykio (mažo gimimo svorio, susirgimo, mirties) rizikai vertinti skaičiuojamas **rizikos balas** (sudaromas *risk score model*). Rizikos balas sudaromas taip: į daugialypės regresijos modelį įtrauktų kintamųjų reikšmės dauginamos iš svorių, proporcingų dydžiams e^{b_j}, o gautos reikšmės sudedamos. 11.7 lentelėje pateiktos kintamųjų – rasės, mažo motinos svorio (low = 1, jei motinos svoris nėštumo pradžioje neviršija 50 kg), ruk ir AH – izoliuota rizika (koeficientai e^b vienmačiame modelyje) su 95 % PI, standartizuota rizika (e^{b_j} daugiamačiame modelyje) su 95 % PI bei rizikos svoriai.

11.7 lentelė. Kintamųjų izoliuota rizika, standartizuota rizika su pasikliautinaisiais intervalais bei rizikos svoriai

Kintamasis	Izoliuota rizika		Standartizuota rizika		Svoris
	e^{β}	e^{β} 95 % PI	$e^{\beta j}$	$e^{\beta j}$ 95 % PI	
low	2,679	1,378–5,206	2,630	1,281–5,401	1
R1	2,328	0,939–5,772	3,238	1,191–8,804	2
R2	1,889	0,955–3,736	2,582	1,136–5,869	1
ruk	2,022	1,081–3,783	2,764	1,304–5,859	1
AH	3,365	1,021–11,088	3,774	1,085–13,134	3

Naudojant šiuos svorius, pagal formulę

BAL = low + 2 × R1 + R2 + ruk + 3 × AH

skaičiuojamas mažo gimimo svorio rizikos balas.

11.7. Logistinio modelio adekvatumo analizė (*assessing fit*)

Sudarius vienmatį ar daugiamatį logistinį modelį, vertinamas ir jo adekvatumas – modeliuoti atsako duomenys palyginami su faktiniais.

Sakykime, į logistinį modelį (11.6) įeina k nepriklausomų kintamųjų $X^{(1)}$, $X^{(2)}$ ... $X^{(k)}$. Pažymėkime J – skirtingų šio faktorių rinkinio reikšmių (*covariate pattern*) imtyje skaičių. Jei kelių individų faktorių rinkinio (kovariatės) $\boldsymbol{x} = (x^{(1)}, x^{(2)} \ldots x^{(k)})$ reikšmės vienodos, tuomet $J < n$. Pažymėkime individų su faktorių vektoriumi $\boldsymbol{x} = \boldsymbol{x}_j$, $j = 1, 2 \ldots J$, skaičių m_j; turime $m_1 + m_2 + \ldots + m_J = n$. Tegu $\overline{y}_j$ – atvejų $\{Y = 1\}$ skaičius, kai $\boldsymbol{x} = \boldsymbol{x}_j$. Logistiniu modeliu prognozuojamų $\{Y = 1\}$ atvejų skaičius, kai $\boldsymbol{x} = \boldsymbol{x}_j$, yra:

$$m_j \hat{\pi}_j = m_j \exp(\hat{g}(\boldsymbol{x}_j))/(1 + \exp(\hat{g}(\boldsymbol{x}_j))).$$

Pateiksime keletą skirtumo tarp modeliuotų $m_j \hat{\pi}_j$ ir faktinių atsako reikšmių $\overline{y}_j$ skaičiaus matų.

Adekvatumo χ^2 kriterijus (*goodness of fit*) su statistika:

$$\chi^2_A = \sum_{j=1}^{J} \frac{(y_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}.$$

Jei logistinis modelis (11.6) teisingas, t. y. $\hat{\pi}_j = y_j / m_j$, tuomet fiksuotam J statistikos χ^2_A skirstinys yra asimptotinis χ^2 su $J - (k + 1)$ laisvės laipsnių. Šį kriterijų modelio adekvatumui realiems duomenims vertinti galima taikyti tik tuo atveju, jei $J < n$ bei y_j ne mažesni už 5.

Jei n ganėtinai didelis ($n > 400$) ir J gana didelis, tada modelio adekvatumui vertinti naudojamas Hosmerio–Lemešou (*Hosmer–Lemeshow*) kriterijus. Skaičiuojant Hosmerio–Lemešou kriterijaus statistikos reikšmę, kiekvienam individui pagal modelį apskaičiuojamos tikimybės (11.7). Jos surikiuojamos nuo mažiausios iki didžiausios ir dalijamos į g grupių; dažniausiai imama $g = 10$. Po to visų individų apskaičiuotos tikimybės dalijamos į deciles (į 10 grupių pagal procentilius). Pirmoje grupėje yra apie $n/10$ individų, turinčių mažiausias $\{Y = 1\}$ įvertintas tikimybes. Kitoje grupėje yra apie $n/10$ individų, turinčių didesnes įvertintas tikimybes. Paskutinėje, dešimtoje grupėje yra apie $n/10$ individų, turinčių didžiausias įvertintas tikimybes.

Hosmerio–Lemešou kriterijaus statistika C yra χ^2 tipo statistika, naudojanti stebėtus ir tikėtinus $\{Y = 1\}$ ir $\{Y = 0\}$ dažnius $2 \times g$ lentelėje:

$$C = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \vec{\pi}_k (1 - \bar{\pi}_k)};$$

čia n_k – individų skaičius k-toje grupėje, O_k – įvykių $\{Y = 1\}$ dažnis k-toje grupėje,

$$\bar{\pi}_k = (\sum \hat{\pi}_j) / n_k$$

– vidutinė k-tosios grupės įvykio $\{Y = 1\}$ tikimybė.

Modeliavimo rezultatais parodyta: jei tarp faktorių yra kiekybinių ir logistinis modelis yra adekvatus realiems duomenims, tai C skirstinys gerai aproksimuojamas χ^2 skirstiniu su $(g - 2)$ laisvės laipsnių.

Kaip minėta, modelio adekvatumui vertinti Hosmerio–Lemešou kriterijų rekomenduojama naudoti tik turint didelę imtį: $n > 400$.

(11.8) modelio Hosmerio–Lemešou statistikos reikšmė lygi 6,528, laisvės laipsniai $g - 2 = 8$, kriterijaus p reikšmė lygi 0,59. Taigi sudarytas modelis gerai atitinka duomenis.

Pseudodeterminacijos koeficientai (pseudo – R^2), arba R^2 tipo adekvatumo matai. Kaip minėta 10 skyriuje, tiesinės bei daugialypės regresijos modelio atitikties realiems duomenims laipsnis vertintas determinacijos koeficientu R^2, lygiu kvadratų sumos, sąlygotos regresijos, ir visos kvadratų sumos santykiui. Tiesinėje regresijoje R^2 parodo, kokią dalį atsako kitimo galima paaiškinti regresijos funkcijos kitimu. Kadangi tiesinėje regresijoje paklaidų skirstinys yra normalusis, todėl skirtumas tarp visos kvadratų sumos ir kvadratų sumos, sąlygotos regresijos, *SSE* proporcingas modelio (10.8) tikėtinumo funkcijai.

Pažymėkime l_0 ir l_k – logtikėtinumo funkcijos, nustatytos modelyje (10.15) atitinkamai naudojant tik konstantą ir konstantą plius k faktorių. Tiesinėje regresijoje R^2 galima išreikšti:

$R^2 = (l_0 - l_k)/l_0 = 1 - l_k/l_0$.

Analogiškos statistikos, vadinamos pseudodeterminacijos koeficientais, naudojamos ir logistinėje regresijoje. Dažniausiai naudojami šie pseudo – R^2:

Kokso–Šnelio (*Cox–Snell*) pseudo – R^2:

$$R_C^2 = 1 - (L(0)/L(k))^{2/n};$$

čia $L(0)$ – tikėtinumo funkcija, (11.3) modelyje naudojant tik konstantą; $L(k)$ – tikėtinumo funkcija, (11.3) modelyje naudojant konstantą plius k faktorių. R_C^2 maksimali reikšmė – $1 - (L(0))^{2/n}$. Ji mažesnė už 1, todėl vietoj R_C^2 naudojamas **Nagelkerke R^2**, kuris kinta tarp 0 ir 1:

$$R_N^2 = R_C^2/(1 - (L(0))^{2/n}).$$

Makfadeno (*McFadden*) pseudo – R^2:

$$R_M^2 = 1 - G/(-2\ln(L(0))) = 1 - \ln(L(k))/\ln(L(0)).$$

Teoriškai R_M^2 gali įgyti reikšmes tarp 0 ir 1. Tačiau, analizuojant realius duomenis, statistika $G = -2(\ln(L(0)) - \ln(L(k)))$ nėra didelė, palyginti su $-2\ln(L(0))$. Jei $0{,}2 \le R_M^2 \le 0{,}4$, tuomet sakoma, kad logistinis modelis puikiai atspindi realius duomenis.

11.8. Polinominė regresija

Sakykime, Y – nominalusis ar tvarkos kintamasis; jo reikšmės užkoduotos skaičiais 0, 1, 2 ... k. Faktorių $X^{(1)}$ ... $X^{(p)}$ įtaka tokiam atsakui vertinama polinominės regresijos modeliu. Polinominės regresijos modelio prielaida: sąlyginės tikimybės $P\{Y = 0 | \boldsymbol{X} = \boldsymbol{x}\}$, $P\{Y = 1 | \boldsymbol{X} = \boldsymbol{x}\}$ ... priklausomai nuo faktoriaus reikšmių, kinta taip:

$$P\{Y = 0 \mid \boldsymbol{X} = \boldsymbol{x}\} = 1/(1 + e^{g_1(\boldsymbol{x})} + e^{g_2(\boldsymbol{x})} + ... + e^{g_k(\boldsymbol{x})}),$$

$$P\{Y = i \mid \boldsymbol{X} = \boldsymbol{x}\} = e^{g_i(\boldsymbol{x})}/(1 + e^{g_1(\boldsymbol{x})} + e^{g_2(\boldsymbol{x})} + ... + e^{g_k(\boldsymbol{x})}), \qquad (11.11)$$

$i = 1, 2 ... k$; čia $g_i(x) = \beta_{i0} + \beta_{i1}x^{(1)} + ... + \beta_{ip}x^{(p)}$.

Koeficientas prie $x^{(j)}$ β_{ij} parodo, kiek padidėja $\ln(P\{Y = i | \boldsymbol{X} = \boldsymbol{x}\} / P\{Y = 0 | \boldsymbol{X} = \boldsymbol{x}\})$, kai $x^{(j)}$ padidėja 1 ir kai kitų faktorių reikšmės yra fiksuotos. Nežinomi modelio koeficientai β vertinami didžiausio tikėtinumo metodu; jų

įverčiai skaičiuojami statistinių programų paketais SPSS ir SAS. Ar tikimybės $P\{Y = 0\}$, $P\{Y = 1\}$ ... $P\{Y = k\}$ priklauso nuo faktorių reikšmių, nustatoma lyginant modelio su $g_i(x) = g_i$ ir (11.11) tikėtinumo funkcijos logaritmus: skaičiuojama statistika G, kaip ir logistinėje regresijoje. Jei (11.11) modelyje visi koeficientai β prie $x^{(i)}$ lygūs nuliui, tuomet statistikos G skirstinys χ^2 yra su $p \times k$ laisvės laipsnių. Kiekvieno koeficiento β_{ij} reikšmingumui tikrinti skaičiuojama Valdo statistika ir jos p reikšmė.

11 skyriaus literatūra

1. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
2. Everitt B. S. *The Analysis of Contingency Tables*. Second edition. 1992, p. 163.
3. Hosmer D. W., Lemeshow S. *Applied Logistic Regression*. 1989. New York: John Wiley & Sons.
4. Fortescue E. B., Kahn K., Bates D. W. Development and Validation of a Clinical Prediction Rule for Major Adverse Outcomes in Coronary Bypass Grafting. 2001. *The American Journal of Cardiology*. Vol. 88, 1251–1257 p.
5. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.
6. *Logistinės regresijos charakteristikų skaičiavimo pavyzdys SPSS paketu. Pateikti paaiškinimai.* Prieiga per internetą: http://www2.chass.ncsu.edu/garson/pa765/logispss.htm.
7. *Logistinė regresija: modelis, parametrų vertinimas, žingsninis parametrų parinkimo algoritmas, modelio suderinamumo tyrimas, modelio adekvatiškumo bei likučių charakteristikos*. Prieiga per internetą: http://www.rrz.uni-hamburg.de/RRZ/Software/SPSS/Algorith.120/logistic_regression.pdf.
8. *Logistinė regresija su pavyzdžiu, skaičiuotu SPSS.* Prieiga per internetą: http://www.education.umd.edu/EDMS/LRA/LRA.pdf.
9. *Logistinio modelio adekvatiškumo charakteristikos.* Prieiga per internetą: http://www.upa.pdx.edu/IQA/newsom/da2/ho_logistc3.goc.
10. *Apie R^2 tipo statistikas logistinėje regresijoje*. Prieiga per internetą: http://www.soziologie.uni-halle.de/langer/pdf/papers/rc33langer.pdf.

12 SKYRIUS

Dispersinė analizė

12.1. Dispersinės analizės sąvoka

Apibūdinti dispersinę analizę pradėsime keliomis problemomis, kylančiomis medikams apdorojant tyrimų duomenis, kaip antai:

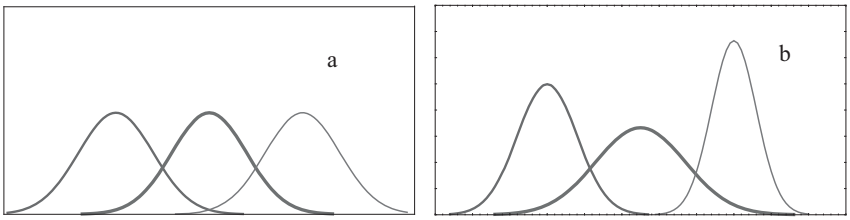
- Ligoniams atliekamas skirtingo galingumo fizinio krūvio mėginys: 50 W, 75 W, 150 W, 200 W. Mėginys kiekvienu galingumu atliktas 5 ligoniams. Ar fizinio krūvio metu ligonio būklę nusakantys parametrai (SAS, DAS, ŠSD, …) priklauso nuo galingumo?
- Tiriamas vaisto poveikis kraujo spaudimui. Tyrime dalyvauja 20 ligonių, po penkis kiekvienoje grupėje: placebo, vartojusio vaisto dozę 5 mg × 3, vartojusio dozę 10 mg × 2 ir 15 mg × 2. Ar galima tvirtinti, kad vaistas darė įtaką SAS?
- Nustatyti ligonių, sergančių idiopatine dilatacine, išemine ir hipertenzine kardiomiopatija, echoskopijos parametrai (išstūmimo frakcija, KSGDD, …). Ar šių parametrų vidutinės reikšmės susijusios su kardiomiopatijos rūšimi?
- Tiriamas vaistų A ir B poveikis kraujo spaudimui. Atsitiktinai parinkti pacientai suskirstomi į 4 grupes: I – vartojo tik vaisto A, II – vartojo tik vaisto B, III – vartojo abiejų vaistų, IV – placebo. Ar vaistai turėjo poveikį SAS? Koks šių vaistų poveikis SAS?
- Tirta lyties ir amžiaus (iki 50 m.; 50 m. ir daugiau) įtaka SAS dydžiui. Ar SAS priklauso nuo lyties ir nuo amžiaus? Kaip SAS keičiasi su amžiumi vyrams ir moterims?

Visų minėtų tyrimų modelis toks: tiriamas vieno ar kelių kokybinių faktorių (galingumo, vaisto dozės, susirgimo, …) poveikis kiekybiniam atsakui

(SAS, DAS, KSGDD, …). Atsako statistinis modelis – tolydusis atsitiktinis dydis, taigi faktoriaus poveikis kiekybiniam atsakui suprantamas kaip poveikis atsako skirstinio vidurkiui. Dviejų populiacijų individų vidurkius galima palyginti naudojant t ar U kriterijus. Tačiau pasitaiko 3 ir daugiau grupių, kuriose individų skaičius nedidelis – 3–5. Todėl kelių populiacijų kiekybinio kintamojo vidurkiams palyginti reikalingi specialūs metodai. Be to, tiriant kelių kokybinių faktorių įtaką atsakui, aktualu konstatuoti ne tik atskiro faktoriaus įtaką, bet ir nustatyti šių faktorių tarpusavio sąveiką bei kiekybiškai ją įvertinti. Visiems minėtiems uždaviniams spręsti naudojama dispersinė analizė (ANOVA).

Dispersinė analizė – tai visuma statistinių metodų, skirtų kiekybinių tyrimų rezultatams, priklausantiems nuo skirtingų vienu metu veikiančių kokybinių faktorių, apdoroti. Dispersinė analizė padeda nustatyti svarbiausius faktorius ir įvertina jų poveikį kiekybiniam atsakui.

Dispersinės analizės taikymo prielaida: atsako skirstinys visiems faktoriaus lygiams (arba kelių faktorių visoms lygių kombinacijoms) yra normalusis su ta pačia dispersija (12.1 a pav.).



12.1 pav. Kintamojo skirstiniai įvairiems faktoriaus lygiams: (a) – su vienodomis dispersijomis; (b) – su skirtingomis dispersijomis

12.2. Vienfaktorė dispersinė analizė

Šiame skyriuje daroma prielaida, kad individus veikia vienas kokybinis faktorius *X*, apibūdinantis kiekybinę arba kokybinę klasifikaciją. Norima nustatyti, ar šis faktorius daro poveikį kiekybiniam atsakui (priklausomam kintamajam) *Y*.

Sakykime, faktorius *X* turi *I* reikšmių arba *I* lygių. Faktoriaus poveikiui tirti iš populiacijų su faktoriaus 1, 2 … *I* lygiu atitinkamai sudaromos kintamojo *Y* $n1$, $n2$ ... nI dydžio imtys (12.1 lentelė). Pažymėkime y_{i1}, y_{i2} … $y_{i,ni}$ – kintamojo *Y* imtis iš populiacijos su *i*-tuoju faktoriaus lygiu. Vienfaktorės ANOVA prielaida – y_{i1}, y_{i2} … $y_{i,ni}$ skirstinys yra normalusis su ta pačia dispersija visiems faktoriaus lygiams (12.1 a pav.). Kitaip tariant, y_{ij}, $j = 1, 2 ... ni$

skirstinys yra normalusis su vidurkiu m_i ir dispersija σ^2. Todėl y_{ij} struktūriškai galima išreikšti taip:

$$y_{ij} = m_i + \varepsilon_{ij},\ i = 1, 2 \ldots I;\ j = 1, 2 \ldots ni; \qquad (12.1)$$

čia m_i – atsako skirstinio iš populiacijos su i-tuoju faktoriaus lygiu vidurkis, ε_{ij} – nepriklausomi normalieji ats. d. su nuliui lygiu vidurkiu ir dispersija σ^2. (12.1) lygybę galime perrašyti taip:

$$y_{ij} = m + \alpha_i + \varepsilon_{ij};$$

čia m – Y skirstinio bendras vidurkis (sujungus visas populiacijas), $\alpha_i = m - m_i$ – i-tojo faktoriaus lygio efektas (poveikis bendram vidurkiui), be to, visų efektų suma lygi 0: $\alpha_1 + \alpha_2 + \ldots + \alpha_I = 0$. Skirstinio vidurkiai (parametrai) m_i nėra žinomi; didžiausio tikėtinumo m_i įvertis yra $\bar{y}_i$ – atsako i-tojo faktoriaus lygio imties vidurkis. m įvertis yra bendras imties vidurkis $\bar{y}$ (12.1 lentelė).

12.1 lentelė. Duomenys vienfaktorei dispersinei analizei

		Vidurkis
1 imtis (1 faktoriaus lygis)	$y_{11}, y_{12} \ldots y_{1n1}$	$\bar{y}_1$
2 imtis (2 faktoriaus lygis)	$y_{21}, y_{22} \ldots y_{2n1}$	$\bar{y}_2$
…. …. …. …. ….	…. …. …. ….	…
I imtis (I faktoriaus lygis)	$y_{I1}, y_{I2} \ldots y_{InI}$	$\bar{y}_I$
Bendras vidurkis		$\bar{y}$

Teiginys „atsakas Y priklauso nuo kokybinio faktoriaus“ suprantamas: „Y skirstinys priklauso nuo faktoriaus lygio“. Faktorius neturės įtakos atsakui, jei Y skirstiniai visiems faktoriaus lygiams bus identiški. Tai nutiks tuomet, kai sutaps visų faktoriaus lygių Y vidurkiai m_j, nes Y dispersijos visiems faktoriaus lygiams yra vienodos, o skirstiniai – normalieji. Todėl norint nustatyti, ar faktorius daro įtaką atsakui, tikrinama tokia nulinė hipotezė:

H_0: „visų populiacijų vidurkiai lygūs (faktorius X įtakos neturi)“ ($m_1 = m_2 = \ldots = m_I$ arba $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$) su alternatyva: „bent vienai vidurkių porai $m_i \neq m_j$“ („bent vienas $\alpha_i \neq 0$“).

Kriterijaus, skirto H_0 tikrinti, statistikos sudarymą iliustruosime pavyzdžiu. Sakykime, tiriamas vaisto, mažinančio SAS, poveikis. Tyrime dalyvauja I ligonių grupių: viena grupė gauna placebo, kitos (I – 1) – skirtingas vaisto

dozes (čia faktorius – vaisto dozė). Sakykime, placebo paskirta $n1$ atsitiktinai parinktų ligonių. Nustatome šių ligonių SAS reikšmes $y_{11}, y_{12} \ldots y_{1,n1}$. Kadangi ligoniai parinkti atsitiktinai, galima tvirtinti, kad tai atsitiktinė atranka iš populiacijos, vartojusios placebo. Pirmą vaisto dozę gavusių ligonių SAS matavimai yra $y_{21}, y_{22} \ldots y_{2,n2}$; tai atranka iš populiacijos ligonių, gaunančių pirmą vaisto dozę. $(I - 1)$-tąją vaisto dozę gaunančių ligonių SAS matavimai yra: $y_{I1}, y_{I2} \ldots y_{I,nI}$. Iš viso turime $N = n1 + n2 + \ldots + nI$ SAS matavimų.

Pateiktame pavyzdyje SAS kitimą sąlygoja du veiksniai – faktoriaus (vaisto dozės) įtaka bei individuali organizmo reakcija į vaistą. Visą duomenų kitimą galima įvertinti visų SAS reikšmių y_{ij} skirtumo nuo bendro vidurkio $\overline{y}$ kvadratų suma:

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{nI} (y_{ij} - \overline{y})^2.$$

Ši suma vadinama visa kvadratų suma (*Total sum of square*, *Total SS*) ir žymima *SST*. Ji išskaidoma į dvi kvadratų sumas:

$$\sum_{i=1}^{I} \sum_{j=1}^{nI} (y_{ij} - \overline{y})^2 = \sum_{i=1}^{I} \sum_{j=1}^{nI} (y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{I} ni(\overline{y}_i - \overline{y})^2. \qquad (12.2)$$

Dešinėje lygybės (12.2) esanti suma $\sum_{i=1}^{I} \sum_{j=1}^{nI} (y_{ij} - \overline{y})^2$ vadinama vidine kvadratų suma (*sum of square within*) ir žymima *SSW*. Ji atspindi duomenų kitimą grupių (faktoriaus lygių) viduje. Suma *SSW* yra atsitiktinis dydis, turintis vidurkį $\sigma^2(N - I)$; *SSW* yra tuo didesnė, kuo didesnė dispersija σ^2. Suma $\sum_{i=1}^{I} ni(\overline{y}_i - \overline{y})^2$ yra tarpgrupinė kvadratų suma (*sum of square between*) ir žymima *SSB*. Ji atspindi kitimą tarp faktoriaus lygių (tarp grupių). Šios sumos vidurkis lygus:

$$(I - 1)\sigma^2 + \sum_{i=1}^{I} ni(m_i - m)^2.$$

Sumos *SST*, *SSW* ir *SSB* yra atsitiktinių dydžių kvadratų sumos; jose yra atitinkamai $(N - 1)$, $(N - I)$ ir $(I - 1)$ nepriklausomų atsitiktinių dydžių. Todėl sakoma, kad sumos *SST*, *SSW* ir *SSB* turi atitinkamai $(N - 1)$, $(N - I)$ ir $(I - 1)$ laisvės laipsnių. Esant teisingai nulinei hipotezei, *SST* vidurkis lygus $(N - 1)\sigma^2$, *SSW* vidurkis – $(N - I)\sigma^2$, o *SSB* vidurkis lygus $(I - 1)\sigma^2$. Formulėmis

$$MSW = SSW/(N - I), \quad MSB = SSB/(I - 1)$$

apibrėžkime vidutinį kvadratą (*mean square*) *MSW* ir *MSB*. *MSW* vidurkis lygus σ^2, o *MSB* vidurkis lygus

$$\sigma^2 + \left(\sum_{i=1}^{I} ni(m_i - m)^2\right)/(I - 1) \qquad (12.3)$$

arba $\sigma^2 + (\sum_{i=1}^{I} ni\alpha_i^2)/(I-1)$. H_0 tvirtinimas, kad populiacijų vidurkiai lygūs, atitinka teiginį, kad dešinysis (12.3) reiškinio dėmuo lygus 0. Todėl nulinei hipotezei (vaisto dozė neturi įtakos SAS kitimui) tikrinti naudojamas F kriterijus su statistika:

$F = MSB / MSW$.

Esant teisingai H_0, statistika F turi Fišerio skirstinį su $(I - 1)$ ir $(N - I)$ laisvės laipsnių. Taigi daroma išvada: jei apskaičiuota F reikšmė viršija $F_{1-\alpha}(I-1, N-I)$, nulinė hipotezė atmetama – tvirtinama, jog vaisto dozė daro reikšmingą poveikį SAS. Priešingu atveju ($F \le F_{1-\alpha}(I-1, N-I)$) nulinei hipotezei neprieštaraujama ir teigiama: remiantis tyrimų rezultatais, nėra pagrindo tvirtinti, jog vaisto dozė daro reikšmingą poveikį SAS; čia $F_{1-\alpha}(I-1, N-I)$ – Fišerio skirstinio su $(I - 1)$ ir $(N - I)$ laisvės laipsnių $1 - \alpha$ lygio kvantilis, α – parinktas reikšmingumo lygmuo, dažniausiai $\alpha = 0{,}05$.

Taikant vienfaktorę ANOVA statistiniu paketu, H_0 tikrinimo rezultatai pateikiami dispersinės analizės lentelėje (12.2 lentelė). Joje yra kvadratų sumos *SST*, *SSW* ir *SSB*, jų laisvės laipsniai, vidutiniai kvadratai *MSW* ir *MSB*, F kriterijaus statistikos ir jo *p* reikšmės. Pagal *p* reikšmę daroma išvada apie faktoriaus poveikį. Jei $p \ge \alpha$ (0,05), nulinei hipotezei neprieštaraujama; jei $p < \alpha$, nulinę hipotezę atmetame, t. y. tvirtiname, jog „faktorius daro reikšmingą poveikį atsakui“ – atsako skirstinio vidurkiai populiacijose, nusakomose faktoriaus lygiais, nėra vienodi. Taip pat statistiniuose paketuose pateikiama populiacijų skirstinio vidurkių m_i įverčiai $\bar{y}_i$ ir standartinio nuokrypio įverčiai.

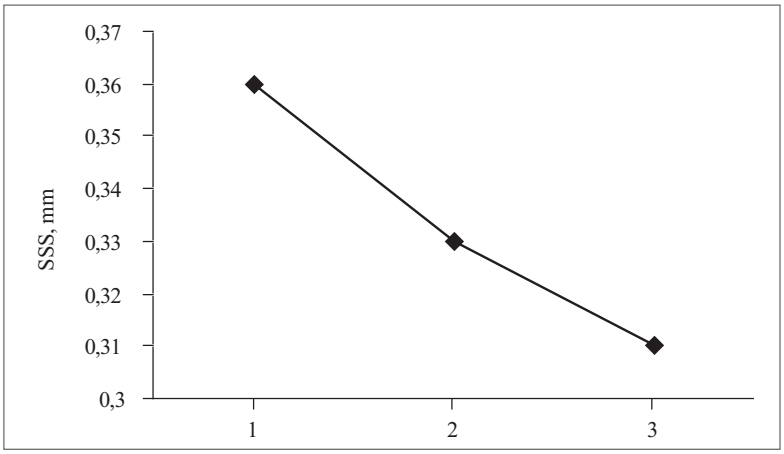
12.2 lentelė. Vienfaktorės dispersinės analizės lentelė

Kitimo šaltinis	Kvadratų suma	Laisvės laipsniai	Vidutiniai kvadratai	*F*	*p*
Tarp grupių	*SSB*	*I – 1*	*MSB*	*F*	*p*
Grupių viduje	*SSW*	*N – I*	*MSW*		
Iš viso	*SST*	*N – 1*			

Nustačius, kad faktorius daro reikšmingą poveikį atsakui, t. y. populiacijų vidurkiai m_i skiriasi, analizuojamas atsako vidurkių kitimas kintant faktoriaus lygiui – nustatoma, kaip faktorius veikia atsaką (12.2 pav). Jei faktorius atspindi kiekybinius pokyčius (yra tvarkos kintamasis), tuomet analizuojama vidurkio trendo dinamika (augimas, mažėjimas) kintant faktoriaus lygiui.

12.1 pavyzdys ([2]). Norima nustatyti, ar kardiomiopatijos rūšis (hipertenzinė, išeminė, idiopatinė dilatacinė) turi įtakos santykiniam miokardo sienelės storiui (SSS). Čia faktorius – kardiomiopatijos rūšis (trys lygiai), atsakas

Y – SSS. Echoskopija atlikta 89 ligoniams: 30 sergantiems hipertenzine, 29 – išemine ir 30 – idiopatine dilatacine kardiomiopatija. Nustatyta, kad sergančių hipertenzine kardiomiopatija SSS vidurkis ir standartinis nuokrypis lygūs 0,358 ir 0,052; sergančių išemine – 0,325 ir 0,054, sergančių idiopatine dilatacine kardiomiopatija – 0,311 ir 0,047 (12.2 pav.). Kaip matyti, visų tirtų ligonių grupių SSS standartiniai nuokrypiai yra panašūs. Poromis lyginant grupių SSS dispersijas F ir Lyvino kriterijais (6.4 skyrius) nustatyta, kad šių kriterijų *p* reikšmės viršijo 0,45; taigi SSS dispersijas galima laikyti lygiomis. Taip pat nė vienas 6.7 skyriuje pateiktas kriterijus nepaneigė SSS normalumo. Taigi dispersinės analizės taikymo prielaidoms prieštaravimų nėra.



12.2 pav. Sergančių hipertenzine (1), išemine (2) ir idiopatine dilatacine (3) kardiomiopatija santykinio sienelės storio (SSS) vidurkiai

Statistiniu paketu apskaičiuotos kvadratų sumos: *SSB* = 0,0348; *SSW* = 0,223; *SST* = 0,2578. *SSB* laisvės laipsnis yra 3 – 1 =2, *SSW*: 89 – 3 = 86, *SST*: 89 – 1 = 88. Atitinkami vidutiniai kvadratai yra lygūs: *MSB* = 0,0174; *MSE* = 0,0026; statistika *F* lygi 6,71, F kriterijaus *p* reikšmė – 0,002. Todėl su 0,01 reikšmingumu galima tvirtinti, kad kardiomiopatijos rūšis reikšmingai sąlygoja SSS dydį.

12.3. Daugybiniai vidurkių palyginimai

Dispersinėje analizėje naudojamu F kriterijumi išsiaiškinama, ar atsako vidurkiai, nustatyti visiems faktoriaus lygiams, statistiškai reikšmingai skiriasi ($m_i \neq m_j$ kažkuriam *i* ir *j*). Tačiau F kriterijus nesuteikia galimybės tarpusavyje palyginti atskirų populiacijų (faktoriaus lygių) vidurkių. Pavyzdžiui, naudodami F kriterijų nustatėme, kad kardiomiopatija (idiopatinė dilataci-

nė, išeminė, hipertenzinė) daro reikšmingą įtaką širdies SSS. Tačiau iš šio fakto negalime daryti išvados apie atskirų kardiopatijos rūšių SSS vidurkių skirtumus.

Dviejų konkrečių populiacijų vidurkiams palyginti naudojamas t kriterijus. Tačiau, poromis lyginant kelių populiacijų vidurkius, tikimybė nustatyti vieną reikšmingą skirtumą didėja augant porų skaičiui. Sakykime, tikrinama hipotezė: „populiacijų A ir B vidurkiai yra lygūs“ su reikšmingumo lygmeniu 0,05. Jei kartu tikrinama ir hipotezė: „populiacijų C ir D vidurkiai yra lygūs“ su tuo pačiu reikšmingumo lygmeniu, tuomet patekimo į šios hipotezės atmetimo sritį tikimybė lygi $1 - (1 - 0{,}05)^2 = 1 - 0{,}95 \cdot 0{,}95 = 0{,}0975$. Lyginant k vidurkių poras su reikšmingumo lygmeniu α, patekimo į hipotezės atmetimo sritį tikimybė lygi $1 - (1 - \alpha)^k$. Todėl reikalingi kriterijai, leidžiantys tarpusavyje palyginti visų populiacijų (faktorių lygių) vidurkius. Šie kriterijai vadinami vidurkių daugybinių palyginimų (*post hoc*) kriterijais. Pateiksime keletą jų.

LSD kriterijus. Visų imčių poros lyginamos naudojant t kriterijų. LSD kriterijus yra pats liberaliausias – dažniausiai randa reikšmingus vidurkių skirtumus.

Bonferoni (*Bonferroni*) kriterijus. Parenkamas kriterijaus reikšmingumo lygmuo α. Visos imčių poros, kurių yra $C = (I - 1)I/2$, lyginamos taikant t kriterijų su reikšmingumo lygmeniu α/C. Bonferoni kriterijus menkai efektyvus, kai imčių yra daug, nes tada labai sumažėja α/C.

Tiuki (*Tukey*) kriterijus. Šis kriterijus grindžiamas stjudentizuoto skirtumo statistika. Sakykime, visų imčių dydžiai vienodi: $n1 = n2 = ... = nI = n$. Norėdami palyginti i-tojo ir j-tojo faktoriaus lygių priklausomo kintamojo vidurkius, skaičiuojame statistiką:

$$Q(i, j) = (\overline{y}_i - \overline{y}_j)/(\sqrt{MSW / n}).$$

Jei abiejų populiacijų vidurkiai vienodi, tuomet atsitiktinis dydis $|Q(i, j)|$ yra stjudentizuotas atstumas su $(I, nI - I)$ laisvės laipsnių (2.5 skyrius). Jei šio kriterijaus p reikšmė mažesnė už α (reikšmingumo lygmenį), tuomet sakoma, kad i-tosios ir j-tosios populiacijų vidurkiai skiriasi.

Tiuki kriterijus – vienas labiausiai taikomų *post hoc* kriterijų. Jis ypač tinkamas, kai faktoriaus lygių yra daug.

Šefės (*Scheffee*) F kriterijus. i-tojo ir j-tojo faktoriaus lygio vidurkių palyginimui skaičiuojama tokia statistika:

$$F_S = \frac{(\overline{y}_i - \overline{y}_j)^2}{MSW(1/n_i + 1/n_i)(I - 1)}.$$

Ši statistika, kai populiacijų vidurkiai yra lygūs, turi F skirstinį su (I – 1, N – I) laisvės laipsnių. Jei F_S statistikos $p < \alpha$, sakoma, kad i-tosios ir j-tosios populiacijų vidurkiai skiriasi.

Šefės kriterijus labai konservatyvus. Jei, remdamiesi šiuo kriterijumi, atmetame hipotezę apie populiacijų vidurkių lygybę, tai tą patį darys ir visi kiti *post hoc* kriterijai.

Statistiniuose paketuose *post hoc* kriterijų taikymo rezultatai dažniausiai pateikiami matrica. Šios matricos stulpeliai atitinka faktoriaus lygius, o i-tosios eilutės ir j-tojo stulpelio susikirtimo gardelėje pateikiama naudojamo kriterijaus p reikšmė. Jei p viršija parinktą reikšmingumo lygmenį, daroma išvada, kad atitinkamų populiacijų vidurkiai skiriasi (atitinkami vidurkiai reikšmingai skiriasi).

12.2 pavyzdys [2]. Pateikiame sergančių hipertenzine (1), išemine (2) ir idiopatine dilatacine (3) kardiomiopatija SSS vidurkių daugybinio palyginimo LSD, Tiuki ir Šefės kriterijais rezultatus (12.3 lentelė).

12.3 lentelė. Sergančių hipertenzine, išemine ir idiopatine dilatacine kardiomiopatija SSS vidurkių daugybinio palyginimo rezultatai

	LSD kriterijus			Tiuki kriterijus			Šefės kriterijus		
Kardiopatija	1	2	3	1	2	3	1	2	3
Hipertenzinė (1)	–	0,014	0,001	–	0,039	0,002	–	0,048	0,003
Išeminė (2)	0,014	–	0,307	0,039	–	0,568	0,048	–	0,592
Id. dilatacinė (3)	0,001	0,307	–	0,002	0,568	–	0,003	0,592	–

Iš 12.3 lentelės matyti, kad sergančių hipertenzine kardiomiopatija SSS vidurkis reikšmingai skyrėsi nuo sergančiųjų išemine ir idiopatine dilatacine. Tai patvirtina visi taikyti kriterijai. Sergančių išemine ir idiopatine dilatacine kardiomiopatija ligonių SSS vidurkių reikšmingo skirtumo nekonstatuojame – visų kriterijų atitinkamos p reikšmės viršijo 0,3 (LSD – 0,307, Tiuki – 0,568, Šefės – 0,592). Pagal vidurkių ir *post hoc* kriterijų p reikšmes galima daryti išvadą, kad sergančių hipertenzine kardiomiopatija SSS vidurkis (0,358) yra reikšmingai didesnis nei sergančiųjų išemine ir idiopatine dilatacine kardiomiopatija (0,325 ir 0,311 atitinkamai).

Iš 12.3 lentelės taip pat matyti, kad, lyginant tuos pačius vidurkius, LSD kriterijaus p reikšmės yra mažiausios, o Šefės – didžiausios. Tai patvirtina teiginį, jog LSD kriterijus liberaliausias, o Šefės – konservatyviausias.

12.4. Kintamųjų vienfaktorėje dispersinėje analizėje ryšio matai

Sakykime, tiriama sergančiųjų IŠL gyvenimo kokybė. Ji vertinama kiekybiniu rodikliu – gyvenimo kokybės (GK) balu. Dispersinės analizės metodu nustatyta, kad GK balo vidurkiui turi įtakos AH laipsnis, amžiaus grupė, susirgimo diagnozė, sergamumas CD. Aptariant šio tyrimo rezultatus, aktualu palyginti minėtų rodiklių įtaką gyvenimo kokybės vidurkiui. Todėl, taikant vienfaktorę ANOVA, kartais svarbu ne tik konstatuoti kiekybinio atsako ir jį veikiančio kokybinio faktoriaus ryšį, bet ir įvertinti šio ryšio stiprumo laipsnį. Šiam tikslui naudojami šie ryšio matai.

Koeficientas η^2. 12.2 skyriuje visą kvadratų sumą išskaidėme į dvi dedamąsias (12.2): $SST = SSB + SSW$. SSB atspindi skirtumą tarp imčių vidurkių: kuo labiau vidurkiai skiriasi, tuo didesnė SSB. Tačiau SSB dydis priklauso ir nuo duomenų matavimo vienetų. Todėl kintamųjų ryšio stiprumui vertinti naudojamas santykinis dydis

$\eta^2 = SSB/SST = SSB/(SSB + SSW)$.

Pagal apibrėžimą η^2 yra tarp nulio ir vieneto. η^2 parodo, kokią duomenų kitimo dalį sąlygoja faktoriaus lygių skirtumai. Dydis $\eta = \sqrt{SSB / SST}$ vadinamas netiesinės koreliacijos koeficientu. Kartais η^2 pateikiamas procentais.

Koeficientas ω^2. Sakykime, visų imčių dydžiai yra vienodi, t. y. $n1 = n2 = ... = nI = n$. Atsitiktinį kitimą atskiruose faktoriaus lygiuose nusako σ^2, faktoriaus įtaką – dydžiai $\alpha_i = m - m_i$. Todėl faktoriaus įtaką vidurkių svyravimams galime vertinti α_i vidutine kvadratine reikšme:

$$\sigma_f^2 = (I)^{-1} \sum_{i=1}^{I} \alpha_i^2. \qquad (12.4)$$

Analogiškai η^2 apibrėžimui, įvedame santykinį faktoriaus įtakos matą:

$$\sigma_f^2/(\sigma^2 + \sigma_f^2). \qquad (12.5)$$

Tiek σ^2, tiek σ_f^2 reikšmių nežinome. 12.2 skyriuje teigėme, kad vidutinių kvadratų MSW ir MSB vidurkiai atitinkamai lygūs σ^2 ir $\sigma^2 + nI\sigma_f^2/(I - 1)$. Todėl σ^2 įverčiu galima laikyti vidutinį kvadratą MSW, o σ_f^2 įverčių – $(I - 1)(MSB - MSW)/(nI)$. Pakeitę (12.5) formulėje σ_f^2 ir σ^2 jų įverčiais, gauname koeficientą ω^2:

$\omega^2 = (SSB - (I - 1)MSW)/(SST + MSW)$.

Ši formulė taikoma ir tuo atveju, kai imčių dydžiai skiriasi.

ω^2 parodo, kokią dalį bendrosios atsako dispersijos paaiškina faktoriaus lygių vidurkių skirtumai. Kuo ω^2 didesnis, tuo labiau duomenų kitimą sąlygoja faktoriaus lygių kitimas.

Tarpklasinės koreliacijos koeficientas (*Intraclass Correlation Coefficient*) *ICC*. Jis apibrėžiamas analogiškai koeficientui ω^2, tik α_i vidutinė kvadratinė reikšmė vertinama taip:

$$\sigma_f^2 = (I-1)^{-1}\sum_{i=1}^{I}\alpha_i^2.$$

Tuomet σ_f^2 įvertis yra $(MSB - MSW)/n$. Įstatę jį ir σ^2 įvertį į (12.5), gauname koeficientą *ICC*, kai imčių faktoriaus lygiuose didumai vienodi:

$$ICC = \frac{MSB - MSW}{MSB + (n-1)MSW} = \frac{F-1}{F+(n-1)};$$ čia $F = MSB/MSW$.

Jei imčių dydžiai skirtingi, *ICC* apibrėžiamas taip:

I*CC* = $(F - 1)/(F + (\overline{n} - 1))$, čia $\overline{n} = I/(1/n1 + 1/n2 + ... + 1/nI)$.

ICC interpretuojamas panašiai kaip ir ω^2.

12.5. Hipotezės apie kelių dispersijų lygybę tikrinimas

Kaip minėta 12.1 skyriuje, dispersinės analizės prielaida yra tokia: atsako dispersijos visuose faktoriaus lygiuose yra vienodos. Todėl prieš naudojant ANOVA, būtina patikrinti hipotezę apie populiacijų dispersijų lygybę. Šiam tikslui naudojami keli kriterijai.

Bartleto (*Bartlett*) kriterijus*. Sakykime, imama *k* imčių iš normalųjį skirstinį turinčių populiacijų. Tikrinama nulinė hipotezė: „visų populiacijų dispersijos lygios“ su alternatyva „bent dviejų populiacijų dispersijos nelygios“. Nulinei hipotezei tikrinti skaičiuojama Bartleto kriterijaus statistika:

$$T = [(N-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2]/(1+\Delta);$$

čia N – imčių dydžių suma, s_i^2 – i-tosios imties dispersija,

$s_p^2 = (N-k)^{-1}\sum_{i=1}^{k}(n_i - 1)s_i^2$ – jungtinės imties dispersija,

$\Delta = [3(k-1)]^{-1}\sum_{i=1}^{k}((n_i-1)^{-1} - (N-k)^{-1})$, n_i – i-tosios imties dydis.

Esant teisingai nulinei hipotezei, statistika T turi asimptotinį χ^2 skirstinį su $(k - 1)$ laisvės laipsnių. Pagal kriterijaus p reikšmę daroma išvada: jei $p < \alpha$ – visų populiacijų dispersijos nėra lygios; jei $p \geq \alpha$ – dispersijų lygybei neprieštaraujame.

Bartleto kriterijus labai jautrus duomenų nenormalumui. Šiuo atžvilgiu alternatyva Bartleto kriterijui yra Lyvino (*Levene*) kriterijus. Šio kriterijaus idėja: imties reikšmės y_{ij} transformuojamos į absoliučius nuokrypius nuo imties vidurkio (skaičiuojami dydžiai $z_{ij} = |y_{ij} - \overline{y}_i|$) ir naudojant F kriterijų tikrinama, ar absoliučių nuokrypių z_{ij} vidurkiai populiacijose nesiskiria. Jei nėra pagrindo tvirtinti, kad z_{ij} vidurkiai populiacijose skiriasi, tuomet ir prielaidai, jog populiacijų dispersijos vienodos, prieštarauti nėra pagrindo.

Statistiniuose paketuose pateikiama Bartleto, Lyvino ar kito kriterijaus, skirto hipotezei apie kelių populiacijų dispersijų lygybę tikrinti, statistikos reikšmė bei atitinkama p reikšmė. Jei $p < 0{,}05$ ($\alpha = 0{,}05$), daroma išvada, kad atskirų faktoriaus lygių atsako populiacijos dispersijos nėra vienodos. Tokiems duomenims apdoroti ANOVA taikytina labai atsargiai. Jei $p \geq 0{,}05$, galima teigti, kad atsako dispersijos atskiruose faktoriaus lygiuose vienodos – taigi ANOVA taikyti galima.

12.6. Dvifaktorė dispersinė analizė

Sakykime, tiriama dviejų kokybinių kintamųjų (faktorių) *A* ir *B* įtaka kiekybiniam atsakui *Y*. Faktorius *A* turi *I* reikšmių arba *I* lygių, faktorius *B* – *J* lygių. Faktorių lygiai gali apibūdinti tiek kiekybinę, tiek kokybinę klasifikaciją. Šių faktorių poveikiui tirti visoms faktorių *A* ir *B* lygių kombinacijoms nustatomos atsako reikšmės (12.4 lentelė). Dėl paprastumo teigiama, kad visų imčių dydžiai vienodi. Kaip ir vienfaktorės dispersinės analizės atveju, daroma prielaida, kad atsako skirstinys visoms faktorių *A* ir *B* lygių kombinacijoms yra normalusis su ta pačia dispersija. Vadinasi, atsako reikšmės, nustatytos esant faktoriaus *A* *i*-tam ir faktoriaus *B* *j*-tam lygiui, turi normalųjį skirstinį su vidurkiu m_{ij} ir dispersija σ^2:

$$y_{ijk} = m_{ij} + \varepsilon_{ij}; \tag{12.6}$$

čia y_{ijk} – *k*-toji imties, nustatytos esant faktoriaus *A* *i*-tam ir faktoriaus *B* *j*-tam lygiui, reikšmė; ε_{ij} – nepriklausomi normalieji ats. dydžiai su nuliui lygiu vidurkiu ir dispersija σ^2. Kadangi atsaką veikia du faktoriai, jo kitimą sąlygoja:

- faktoriaus *A* įtaka;
- faktoriaus *B* įtaka;
- faktorių *A* ir *B* tarpusavio sąveika;
- atsitiktinis kitimas.

Todėl (12.6) struktūrinį modelį galime perrašyti taip:

$$y_{ijk} = m_{ij} + e_{ij} = m + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}; \tag{12.7}$$

čia α_i – faktoriaus *A* *i*-tojo lygio efektas (poveikis bendram vidurkiui), β_j – faktoriaus *B* *j*-tojo lygio efektas, γ_{ij} – faktorių *A* ir *B* *i* ir *j* lygių tarpusavio efektas; be to, visų efektų suma lygi 0. Jei visuose faktoriaus *B* lygiuose faktoriaus *A* poveikis *Y* vidurkiui vienodas (arba faktoriaus *B* poveikis vienodas visuose *A* lygiuose), faktorių *A* ir *B* sąveikos nėra (12.3 a pav.): $\gamma_{ij} = 0$. Jei vidurkio dinamika vieno faktoriaus atžvilgiu priklauso nuo kito faktoriaus lygio, tuomet faktorių tarpusavio sąveika yra (12.3 pav.).

12.4 lentelė. Dvifaktorės dispersinės analizės duomenys

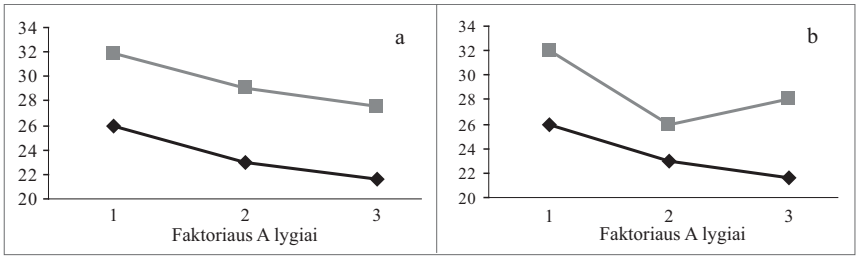
Faktoriaus *A* lygiai	Faktoriaus *B* lygiai				Vidurkis
	1	2	...	J	
1 imtis vidurkis	$y_{111}, \dots y_{11n}$ $\overline{y}_{11}$	$y_{121}, \dots y_{12n}$ $\overline{y}_{12}$	...	$y_{1J1}, \dots y_{1Jn}$ $\overline{y}_{1J}$	$\overline{y}_{1*}$
2 imtis vidurkis	$y_{211}, \dots y_{21n}$ $\overline{y}_{21}$	$y_{221}, \dots y_{22n}$ $\overline{y}_{22}$	...	$y_{2J1}, \dots y_{2Jn}$ $\overline{y}_{2J}$	$\overline{y}_{2*}$
...	...	...	...	...	...
I imtis vidurkis	$y_{I11}, \dots y_{I1n}$ $\overline{y}_{I1}$	$y_{I21}, \dots y_{I2n}$ $\overline{y}_{I2}$	...	$y_{IJ1}, \dots y_{IJn}$ $\overline{y}_{IJ}$	$\overline{y}_{I*}$
Vidurkis	$\overline{y}_{*1}$	$\overline{y}_{*2}$		$\overline{y}_{*J}$	
Bendras vidurkis					$\overline{y}$

Norint nustatyti, ar faktoriai ir jų tarpusavio sąveika daro poveikį atsakui, tikrinamos šios nulinės hipotezės su atitinkamomis alternatyvomis:

H_{0A}: „faktorius *A* neturi įtakos atsakui“ (struktūriniame modelyje (12.7) $\alpha_1 = ... = \alpha_I = 0$) su alternatyva: „faktorius *A* turi įtakos atsakui“ („bent vienas $\alpha_i \neq 0$“);

H_{0B}: „faktorius *B* neturi įtakos atsakui“ (struktūriniame modelyje (12.7) $\beta_1 = ... = \beta_J = 0$) su alternatyva: „faktorius *B* turi įtakos atsakui“ („bent vienas $\beta_i \neq 0$“);

H_{0AB}: „faktorių *A* ir *B* tarpusavio sąveika neturi įtakos“ (struktūriniame modelyje (12.7) $\gamma_{ij} = 0$, $i = 1, 2 ... I$, $j = 1, 2 ... J$) su alternatyva: „faktorių *A* ir *B* tarpusavio sąveika turi įtakos atsakui“ („bent vienas $\gamma_{ij} \neq 0$“).



12.3 pav. Faktorių A ir B tarpusavio sąveikos įtaka vidurkiui: (a) tarpusavio sąveikos nėra; (b) tarpusavio sąveika yra

Šioms hipotezėms tikrinti skirtų statistikų formulėse yra visos kvadratų sumos

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2$$

dedamosios. Kadangi atsako kitimą sąlygoja faktorių *A* ir *B* kitimas, jų tarpusavio sąveika bei atsitiktinis kitimas, todėl *SST* išskaidoma taip: $SST = SSA + SSB + SSAB + SSW$. Čia

$$SSA = nJ\sum_{i=1}^{I}(\bar{y}_{i*} - \bar{y})^2, \quad SSB = nI\sum_{j=1}^{J}(\bar{y}_{*j} - \bar{y})^2,$$

$$SSAB = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij} - \bar{y}_{i*} - \bar{y}_{*j} + \bar{y})^2, \; SSW = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij})^2;$$

$\bar{y}$ – bendras vidurkis; $\bar{y}_{i*}$ – faktoriaus *A* *i*-tojo lygio imties vidurkis; $\bar{y}_{*j}$ – faktoriaus *B* *j*-tojo lygio imties vidurkis; $\bar{y}_{ij}$ – imties su faktoriaus *A* *i*-tuoju, o *B* – *j*-tuoju lygiu vidurkis. *SSA* atspindi faktoriaus *A* įtaką, *SSB* – faktoriaus *B* įtaką, *SSAB* – faktorių tarpusavio sąveikos įtaką. *SSW* yra atsitiktinių paklaidų kvadratų suma. Kvadratų sumose *SST*, *SSA*, *SSB*, *SSAB* ir *SSW* yra atitinkamai $IJn - 1$, $I - 1$, $J - 1$, $(I - 1)(J - 1)$ ir $IJ(n - 1) + 1$ nepriklausomų atsitiktinių dydžių, vadinamų laisvės laipsniais.

Vidutinių kvadratų

$$MSA = SSA/(I - 1), \quad MSB = SSB/(J - 1), \quad MSAB = SSAB/((I - 1)(J - 1)),$$
$$MSW = SSW/(IJ(n - 1)) \tag{12.7}$$

vidurkiai atitinkamai lygūs: $\sigma^2 + Jn\sigma_A^2$, $\sigma^2 + In\sigma_B^2$, $\sigma^2 + n\sigma_{AB}^2$, σ^2; čia $\sigma_A^2, \sigma_B^2, \sigma_{AB}^2$ – faktorių *A*, *B* ir tarpusavio sąveiką atspindintys dydžiai:

$$\sigma_A^2 = (I-1)^{-1}\sum_{i=1}^{I}\alpha_i^2, \; \sigma_B^2 = (J-1)^{-1}\sum_{j=1}^{J}\beta_j^2,$$
$$\sigma_{AB}^2 = (I-1)^{-1}(J-1)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2. \tag{12.8}$$

Esant teisingai nulinei hipotezei H_{0A} ($\alpha_i = 0$ visiems i), *MSA* vidurkis lygus σ^2, nes $\sigma_A^2 = 0$. Todėl hipotezei H_{0A} (faktorius *A* nedaro įtakos *Y* vidurkiui) tikrinti naudojama statistika: $F_A = MSA/MSW$. Esant teisingai H_{0A}, statistika F_A turi F skirstinį su $(I - 1)$ ir $IJ(n - 1)$ laisvės laipsnių.

Esant teisingai hipotezei H_{0B} ($\beta_j = 0$ visiems j), *MSB* vidurkis lygus σ^2, nes $\sigma_B^2 = 0$. Todėl hipotezei H_{0B} (faktorius *B* nedaro įtakos *Y* vidurkiui) tikrinti naudojama statistika: $F_B = MSB/MSW$. Esant teisingai H_{0B}, statistika F_B turi *F* skirstinį su $(J - 1)$ ir $IJ(n - 1)$ laisvės laipsnių.

Jei H_{0AB} ($\gamma_{ij} \equiv 0$) teisinga, *MSAB* vidurkis lygus σ^2, nes $\sigma_{AB}^2 = 0$. Todėl hipotezei H_{0AB} (faktorių sąveika neturi įtakos) tikrinti naudojama statistika: F_{AB} = MSAB/MSW. Esant teisingai H_{0AB}, statistika F_{AB} turi F skirstinį su $(I - 1)(J - 1)$ ir $IJ(n - 1)$ laisvės laipsnių.

Statistiniuose paketuose dvifaktorės ANOVA rezultatai: kvadratų sumos, laisvės laipsniai, vidutiniai kvadratai, F_A, F_B, ir F_{AB} statistikos bei jų *p* reikšmės – pateikiami dispersinės analizės lentelėje (12.5 lentelė). Jei kurios nors *F* statistikos *p* reikšmė mažesnė už α, daroma išvada, jog šis faktorius reikšmingai veikia atsaką.

Ir vienfaktorėje, ir dvifaktorėje ANOVA aktualu tarpusavyje palyginti visų faktorių lygių kombinacijų vidurkius. Tam naudojami tie patys *post hoc* kriterijai. Statistiniuose paketuose daugybinių vidurkių palyginimo rezultatai pateikiami matrica, kurios stulpeliai ir eilutės atitinka faktorių *A* ir *B* visų lygių kombinacijas. *i*-tos ir *j*-tos eilutės susikirtime pateikiama kriterijaus *p* reikšmė. Jei $p < \alpha$, daroma išvada, jog atitinkami vidurkiai statistiškai reikšmingai skiriasi.

Interpretuojant dvifaktorės ANOVA rezultatus (kaip faktoriai veikia atsaką), skaičiuojami atsako vidurkių įverčiai $\bar{y}_{ij}$ bei grafiškai pateikiamas $\bar{y}_{ij}$ kitimas pagal faktorių lygių *i* ir *j* kitimą. Paprastai *X* ašyje atidedamos faktoriaus, turinčio didesnį lygių skaičių, reikšmės, o kito faktoriaus lygių atsako vidurkiai sujungiami tiese (12.3 pav.). Pagal F kriterijaus, daugybinio vidurkių palyginimo *p* reikšmes bei $\bar{y}_{ij}$ kitimą, priklausomai nuo *i* ir *j*, daroma išvada apie tai, kaip faktoriai *A* ir *B* veikia atsaką.

Dvifaktorėje dispersinėje analizėje taip pat vertinamas faktorių ir atsako ryšio stiprumo laipsnis. Kokią dalį duomenų kitimo galima paaiškinti faktoriaus *A*, faktoriaus *B* ir faktorių tarpusavio sąveikos kitimu, parodo, pavydžiui, koeficientai

$$\eta_A^2 = SSA/SST, \eta_B^2 = SSB/SST, \eta_{AB}^2 = SSAB/SST.$$

12.7. Dvifaktorės dispersinės analizės taikymo pavyzdys

12.3 pavyzdys. Kardiologijos instituto Klinikinės kardiologijos laboratorijoje 1998–2002 m. organizuota studija, skirta ligonių, sirgusių ūmiais koronariniais sindromais – Q ir be Q bangos miokardo infarktu (MI) bei nestabilia krūtinės angina (NKA) – būklei tirti. Ligonio sistolinio kraujospūdžio (SAS) priklausomybei nuo koronarinio sindromo ir antsvorio (KMA > 25) tirti taikyta dvifaktorė dispersinė analizė. Čia SAS – atsakas (priklausomas kintamasis); faktorius *A* – koronarinis sindromas, turintis 3 lygius (Q MI, be Q MI, NKA); faktorius *B* – antsvoris, turintis 2 lygius (yra; nėra). Visų faktorių lygių kombinacijų SAS normalumas tikrintas χ^2 kriterijumi. Kadangi χ^2 kriterijaus *p* reikšmė viršijo 0,1 visose 6 grupėse, laikoma, kad visų ligonių grupių SAS skirstiniai yra normalieji. Bartleto kriterijaus *p* reikšmė lygi 0,47; Lyvino kriterijaus – 0,62; taigi galima tvirtinti, kad visų grupių SAS skirstinių dispersijos yra vienodos. Taigi ANOVA taikymo prielaidos teisingos.

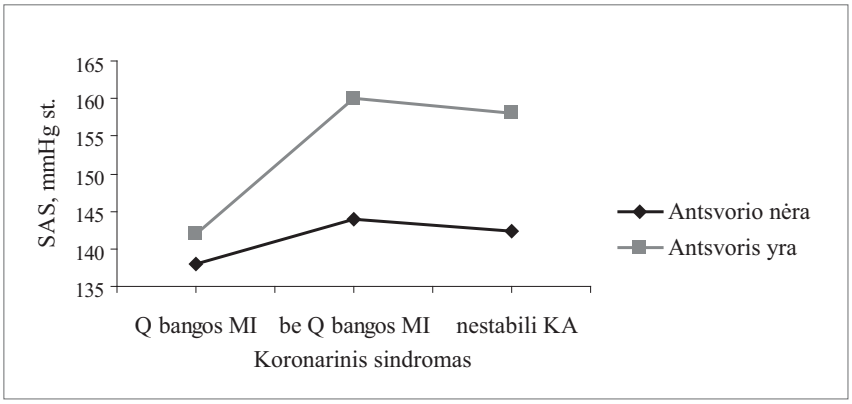
Dvifaktorės dispersinės analizės lentelėje (12.5 lentelė) pateiktos kvadratų sumos, jų laisvės laipsniai, vidutiniai kvadratai, *F* kriterijaus statistikos bei atitinkamos *p* reikšmės. Iš lentelės matyti, kad F kriterijaus, skirto hipotezei apie koronarinio sindromo ir antsvorio įtaką SAS tikrinti, *p* reikšmė mažesnė už 0,001. Todėl galima tvirtinti, kad SAS vidurkis priklauso ir nuo koronarinio sindromo, ir nuo antsvorio. Šių faktorių tarpusavio sąveika neturėjo reikšmingos įtakos SAS, nes statistikos F_{AB} *p* reikšmė lygi 0,14 > 0,05.

12.5 lentelė. Dvifaktorės dispersinės analizės lentelė

Dispersijos šaltinis	Kvadratų suma	L. l.	Vidutinis kvadratas	*F*	*p*
Koronarinis sindromas	*SSA* = 12411,8	2	*MSA* = 6205,9	F_A = 8,45	<0,001
Antsvoris	*SSB* = 11896,7	1	*MSB* = 11896,7	F_B = 16,21	<0,001
(Koronarinis sindromas)* Antsvoris	*SSAB* = 2896,7	2	*MSAB* = 734,1	F_{AB} = 1,97	0,14
Atsitiktinis kitimas	*SSW* = 351646	479			
Iš viso	*SST* = 350238	484			

12.4 pav. pateiktas ligonių, turinčių ir neturinčių antsvorio, SAS vidurkių kitimas pagal koronarinį sindromą. Iš paveikslo matyti, kad turinčiųjų antsvorį SAS vidurkiai yra didesni. Remiantis *post hoc* vidurkių palyginimu LSD kriterijumi galima tvirtinti, kad sergančių Q bangos MI ligonių, turinčių ir neturinčių antsvorio, SAS vidurkiai patikimai nesiskiria (p = 0,25). Tačiau

sergančių be Q MI ir NKA ligonių, turinčių antsvorio, SAS vidurkis buvo reikšmingai didesnis už neturinčių antsvorio ($p < 0,01$). Neturinčių antsvorio ligonių visų koronarinių sindromų SAS vidurkiai reikšmingai nesiskyrė. O turinčių antsvorio ligonių, sergančių Q bangos MI, SAS vidurkis buvo reikšmingai mažesnis, negu sergančių be Q MI ir NKA ($p < 0,005$). Sergančių be Q MI ir NKA SAS vidurkiai reikšmingai nesiskyrė tiek turinčių antsvorio, tiek jo neturinčių grupėje ($p > 0,5$).



12.4 pav. Ligonių, turinčių ir neturinčių antsvorio, SAS vidurkių kitimas pagal koronarinį sindromą

12.8. Kovariancinė analizė*

Kaip minėta, dispersinė analizė nagrinėja kokybinių faktorių įtaką kiekybiniam atsakui. Kovariancinė analizė (ANCOVA) skirta tirti kokybinių faktorių įtaką kiekybiniam atsakui, atsižvelgiant ir į kiekybinių rodiklių (kovariačių, prediktorių) daromą poveikį atsakui.

Pateiksime kovariancinės analizės taikymo pavyzdį. Tirta įvairių rūšių krakmolo (grūdų, bulvių ir t. t.) kokybė. Atlikto eksperimento metu matuotas krakmolo plėvelės stiprumas. Naudojant vienfaktorę dispersinę analizę (faktorius – krakmolo rūšis, atsakas – krakmolo plėvelės stiprumas) nustatyta, kad plėvelės stiprumas statistiškai reikšmingai priklauso nuo krakmolo rūšies. Tačiau plėvelės stiprumą galima paaiškinti jos storumu (kovariate). Pažymėkime y_{ij} – i-tosios krakmolo rūšies j-tasis stiprumo matavimas. Jei šis matavimas buvo nustatytas esant z_{ij} storio plėvelei ir jei tarp plėvelės stiprumo bei storio yra tiesinė priklausomybė, y_{ij} modelis turėtų būti toks:

$$y_{ij} = m_i + \gamma z_{ij} + e_{ij};$$

čia m_i – i-tosios krakmolo rūšies įtakos kiekybinė išraiška, γ – regresijos tarp plėvelės stiprumo ir storio koeficientas, e_{ij} – atsitiktinė paklaida.

Kovariancinės analizės taikymo prielaidos atsako skirstiniui yra tos pačios, kaip ir dispersinėje analizėje: atsako reikšmės nepriklausomos, visuose faktorių lygiuose turinčios normalųjį skirstinį su ta pačia dispersija. Papildomos ANCOVA prielaidos:

- tarp atsako ir kovariatės yra tiesinis ryšys;
- koreliacija tarp atsako ir kovariatės vienoda visuose faktorių lygiuose;
- kovariatės reikšmės nuo faktoriaus lygio nepriklauso.

Todėl struktūrinį kovariancinės analizės modelį su vienu faktoriumi ir viena kovariate galime pateikti taip:

$y_{ij} = m + \alpha_i + \gamma(z_{ij} - z0) + \varepsilon_{ij}$;

čia y_{ij} ir z_{ij} – atsako ir kovariatės j-toji reikšmė esant i-tajam faktoriaus lygiui, m – bendras atsako vidurkis, $z0$ – kovariatės reikšmių vidurkis, α_i – i-tojo faktoriaus lygio efektas, be to, $\alpha_1 + \alpha_2 + ... + \alpha_I = 0$, γ – regresijos tarp Y ir Z koeficientas, ε_{ij} – nepriklausomi normalieji ats. d., $\varepsilon_{ij} \sim N(0,\sigma^2)$.

Kovariancinėje analizėje tikrinamos šios nulinės hipotezės:

H_{0X}: „faktorius neturi įtakos atsakui“ ($\alpha_1 = \alpha_2 = ... = \alpha_I = 0$);
H_{0Z}: „kovariatė neturi įtakos atsakui“ ($\gamma = 0$).

H_{0X} ir H_{0Z} tikrinti naudojamas F kriterijus, kurio statistika sudaryta iš visos kvadratų sumos SST dedamųjų: $SST = SSC + SSB_C + SSW_C$. Čia SSC – kovariatės įtaką atsakui atspindinti kvadratų suma, SSB_C – faktoriaus efektą atsakui be kovariatės įtakos nusakanti suma, SSW_C – atsitiktinį atsako kitimą atspindinti kvadratų suma. Sumos SSC, SSB_C ir SSW_C išreiškiamos per atsako, kovariatės ir ryšio tarp atsako ir kovariatės kitimą (visą ir atskiruose faktoriaus lygiuose) nusakančias dedamąsias:

$$SST_Y = \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{I} n(\overline{y}_i - \overline{y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2 = SSB_Y + SSW_Y,$$

$$SST_Z = \sum_{i=1}^{I}\sum_{j=1}^{n}(z_{ij} - \overline{z})^2 = \sum_{i=1}^{I} n(\overline{z}_i - \overline{z})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n}(z_{ij} - \overline{z}_i)^2 = SSBZ + SSW_Z,$$

$$SST_{ZY} = \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \overline{y})(z_{ij} - \overline{z}) = \sum_{i=1}^{I} n(\overline{y}_i - \overline{y})(\overline{z}_i - \overline{z}) +$$

$$\sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)(z_{ij} - \overline{z}_i) = SSB_{ZY} + SSW_{ZY}.$$

Suma $SSB_C + SSW_C$ lygi likučių kvadratų sumai, skaičiuotai tiesinėje regresijoje tarp Y ir Z. Pertvarkę šią sumą, gauname

$$SSB_C + SSW_C = \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \overline{y} - b(z_{ij} - \overline{z}))^2 = SST_Y - (SST_{ZY})^2 / SST_Z,$$

nes $b = SST_{ZY}/SST_Z$ (10.10 formulė). Taigi $SSB_C + SSW_C$ lygi visai kvadratų sumai, atėmus dydį, vertinantį Y ir Z ryšį. Todėl $SST_Y - (SST_{ZY})^2/SST_Z$ vadinama pagal kovariatę koreguota visa Y kvadratų suma (SST_{Yadj}). Analogiškai apibrėžiama pagal kovariatę koreguota kvadratų suma SSW_{Yadj}, nusakanti kitimą faktoriaus lygiuose:

$SSW_C = SSW_{Yadj} = SSW_Y - (SSW_{ZY})^2/SSW_Z$.

Kovariatės įtakai vertinti skaičiuojamas dydis SST_{ZY}/SST_Z arba SSW_{ZY}/SSW_Z.

H_{0X} tikrinti skaičiuojama statistika $F = (SSB_C/SSW_C)/((I - 1)/((n - 1)I - 1))$, H_{0Z} tikrinti – $F = (SSC/SSW_C)((n - 1)I - 1)$. Statistiniuose paketuose pateikiamos minėtos kvadratų sumos, F statistikos bei atitinkamos p reikšmės. Kaip ir 12.5 skyriuje, pagal p reikšmę daroma išvada apie faktoriaus ir kovariatės įtaką atsakui.

12 skyriaus literatūra

1. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
2. Ragaišytė N. *Sergančiųjų idiopatine dilatacine, išemine ir hipertenzine kardiomiopatija palyginamieji duomenys*. Daktaro disertacija. 2003. Kaunas.
3. Miller J. C., Miller J. N. *Statistics for Analycal Chemistry*. Second ed. 1988. New York: John Wiley & Sons, p. 227.
4. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla. p. 98.
5. Scheffee H. *The Analysis of Variance*. 1999. John Wiley & Sons, p. 477.
6. *ANCOVA apibrėžimas.* Prieiga per internetą: http://bill.psyc.anderson.edu/exdes/ancova.htm.
7. *Apie ANCOVA*. Prieiga per internetą: http://carbon.cudenver.edu/~lsherry/rem/ancova.html.
8. Carey G. *MANOVA*. 1998. Prieiga per internetą: http://www.psych.umn.edu/courses/spring05/federicoc/psy8815/lectures/stats_lecture11_reading.pdf.

13 SKYRIUS

Išgyvenamumo analizė

Klinikinėse studijose, skirtose medikamentiniam ar chirurginiam vėžio gydymui, be kitų ligonio būklę charakterizuojančių rodiklių, fiksuojamas ir laikas nuo ligonio stebėjimo pradžios iki mirties ar susirgimo remisijos. Laikas nuo stebėjimo pradžios (susirgimo, įtraukimo į studiją, operacijos, gydymo pradžios) iki tam tikro baigties taško (*outpoint*), kaip antai remisija, mirtis, komplikacija, fiksuojamas ir kituose biomedicinos tyrimuose. Šio pobūdžio duomenys registruojami tiriant nuodingo preparato įtaką graužikų žūčiai, analizuojant persirgusių MI ar operuotų ligonių išgyvenimo trukmę ir t. t.

Analizės metodų, skirtų laiko iki baigties taško analizei ir modeliavimui, visuma vadinama išgyvenamumo analize (*survival analysis*). Išgyvenamumo duomenys suprantami kaip stebėjimo iki bet kurio baigties taško (nebūtinai mirties) duomenys. Analizuojant šio pobūdžio duomenis, terminas „gyvavimo laikas“ (*failure time*) suprantamas kaip laikas nuo stebėjimo pradžios iki baigties įvykio (*output*).

13.1. Išgyvenamumo tyrimo pavyzdžiai

13.1 pavyzdys ([6]). Tirtas ligonių, sergančių glioblastoma, išgyvenamumas po chirurginės operacijos. Šio tyrimo tikslas – nustatyti cheminės terapijos tikslingumą: ar ligoniai, kuriems taikyta chemoterapija, išgyvena ilgiau, ar ne. 20 ligonių po chirurginės operacijos taikyta chemoterapija, 15 – netaikyta. Ši 15 ligonių grupė buvo kontrolinė. Ligoniai stebėti tam tikrą laiką. Dalis ligonių stebėjimo laikotarpiu mirė, dalis išgyveno. Ligonių stebėjimo laikas savaitėmis, įvykis (mirė = 1, išgyveno = 0) ir poveikis (cheminė terapija taikyta, netaikyta) pateikti 13.1 lentelėje.

13.2 pavyzdys ([6]). Tirtas toksinių medžiagų poveikis pelėms. Eksperimento tikslas – nustatyti, ar yra ryšys tarp toksinio preparato rūšies ir pelių išgyvenamumo. Dėl to 29 pelės buvo paveiktos A preparatu, 34 pelės – B preparatu. Pelės buvo stebimos 25 dienas, fiksuotas jų išgyvenimo laikas dienomis. Po eksperimento abiejose grupėse liko po 3 gyvas peles.

13.3 pavyzdys ([6]). Tirtas ligonių, sergančių leukemija, remisijos laikas. Tyrimo tikslas – nustatyti 6-mercaptorino (6-mp) efektą ligonių remisijai. Buvo fiksuota 21 ligonio, gydyto 6-mercaptorinu, ir 21 ligonio, gavusio placebo, remisijos laikai. Tyrimo duomenys pateikti 13.1 lentelėje (baigties taškas: 1 – įvyko remisija, 0 – remisijos nesulaukta).

13.1 lentelė. 13.1 pav. ir 13.3 pav. tyrimų duomenys išgyvenamumui ir remisijai vertinti ([6])

Glioblastoma sergančiųjų išgyvenamumo duomenys			Leukemija sergančiųjų remisijos duomenys		
Savaitės	Baigties taškas	Poveikis	Savaitės	Baigties taškas	Poveikis
1	1	kontrol.	6	1	6-mp
2	1	kontrol.	6	1	6-mp
5	1	kontrol.	6	1	6-mp
7	1	kontrol.	7	1	6-mp
13	1	kontrol.	10	1	6-mp
22	1	kontrol.	13	1	6-mp
24	1	kontrol.	16	1	6-mp
54	1	kontrol.	22	1	6-mp
7	0	kontrol.	23	1	6-mp
11	0	kontrol.	6	0	6-mp
19	0	kontrol.	9	0	6-mp
22	0	kontrol.	10	0	6-mp
30	0	kontrol.	11	0	6-mp
35	0	kontrol.	17	0	6-mp
39	0	kontrol.	19	0	6-mp
1	1	chem.	20	0	6-mp
2	1	chem.	25	0	6-mp
8	1	chem.	32	0	6-mp
10	1	chem.	32	0	6-mp
15	1	chem.	34	0	6-mp
19	1	chem.	35	0	6-mp
26	1	chem.	1	1	placebo
28	1	chem.	1	1	placebo

Glioblastoma sergančiųjų išgyvenamumo duomenys			Leukemija sergančiųjų remisijos duomenys		
Savaitės	Baigties taškas	Poveikis	Savaitės	Baigties taškas	Poveikis
33	1	chem.	2	1	placebo
36	1	chem.	2	1	placebo
39	1	chem.	3	1	placebo
44	1	chem.	4	1	placebo
2	0	chem.	4	1	placebo
9	0	chem.	5	1	placebo
13	0	chem.	5	1	placebo
22	0	chem.	8	1	placebo
25	0	chem.	8	1	placebo
36	0	chem.	8	1	placebo
43	0	chem.	8	1	placebo
45	0	chem.	11	1	placebo
–	–	–	11	1	placebo
–	–	–	12	1	placebo
–	–	–	12	1	placebo
–	–	–	15	1	placebo
–	–	–	17	1	placebo
–	–	–	22	1	placebo
–	–	–	23	1	placebo

Iš 13.1 lentelės matyti, kad tiek glioblastoma, tiek leukemija sergantys ligoniai stebėti nevienodą laiko tarpą – atskiri ligoniai stebėti net iki 45 savaičių. Dalis glioblastoma sirgusių ligonių stebėjimo metu mirė: chemoterapijos grupėje – 12 ligonių, kontrolinėje – 8. Jų išgyvenimo po operacijos laikas žinomas. Kitų ligonių mirties stebėjimo metu neužfiksuota – jų išgyvenimo po chirurginės operacijos laikas nėra tiksliai žinomas; žinoma tik tiek, kad jis ilgesnis už ligonio stebėjimo laiką. Analogiški ir leukemija sergančių ligonių laiko iki remisijos duomenys.

Analizuojant 13.1–13.3 pavyzdžių duomenis, aktualu:

- įvertinti tam tikros populiacijos – individų kontingento (pvz., sergančių leukemija) – išgyvenamumą;
- palyginti dviejų individų populiacijų išgyvenamumą.

Kaip minėta, duomenys apie išgyvenimo laiką yra specifiniai – žinomas ne visų tikslus išgyvenimo laikas. Todėl būtinas populiacijos (tiriamos grupės) išgyvenamumo įvertis bei kriterijai populiacijų išgyvenamumui palyginti.

13.2. Išgyvenamumo duomenys. Cenzūravimas

Analizuojant n dydžio imties iš tam tikros populiacijos (ligonių kontingento) išgyvenamumą, kiekvieną i-tąjį individą stebime C_i laiką (13.1 pav.). Daliai individų stebėjimo metu įvyko baigties taškas (mirė, susirgo), todėl žinomi tikslūs jų gyvavimo laikai $U_1, U_2 \ldots U_k$. Likusiems $n - k$ individams stebėjimo laikotarpiu baigties taškas neįvyko – tam tikru momentu jie dingo iš stebėjimo (išvyko, nustojo lankytis pas gydytoją ir t. t.) arba visą stebėjimo laiką jų baigties taško nesulaukta. Tolesnį šių individų gyvavimą galima tik prognozuoti. Tokie neišsamūs stebėjimo duomenys vadinami cenzūruotais. Cenzūruotų individų stebėjimo laikai yra žinomi ir lygūs C_1, C_2 ... C_{n-k}. Apibendrinant išgyvenamumo duomenis, galima tvirtinti, kad tyrimo metu nustatome dvimačio ats. dydžio reikšmes:

$(t_1, I_1), (t_2, I_2) \ldots (t_n, I_n)$;

čia I – cenzūravimo indikatorius; $I = 1$, jei žinomas laikas nuo stebėjimo pradžios iki baigties taško, $I = 0$ – cenzūruotas stebėjimas; žinomas tik laikas, iki kurio baigties taškas neįvyko. t_j apibrėžiami taip:

$$t_j = \begin{cases} U_j, I_j = 1, \\ C_j, I_j = 0. \end{cases} \qquad (13.1)$$

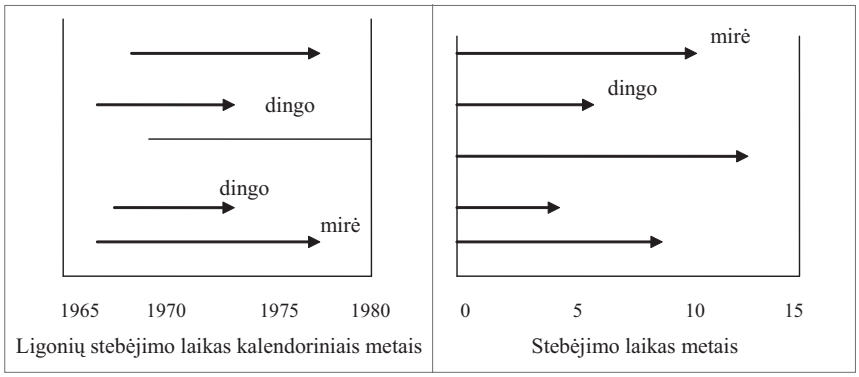
Galimi šie cenzūravimo tipai:

- I tipas: visi individai stebimi vienodą laiką, cenzūravimas atliekamas eksperimento pabaigoje, t. y. $C_j \equiv C$.
- II tipas: visi individai stebimi vienodą laiko tarpą. Cenzūravimas atliekamas tik pasiekus tam tikrą fiksuotą baigties taškų skaičių. Šiuo atveju cenzūravimo laikas yra atsitiktinis.
- III tipas (progresyvus cenzūravimas): individai į studiją įtraukiami skirtingu laiku (13.2 pav.). Cenzūravimas vyksta studijai baigiantis – visiems individams tuo pačiu (neatsitiktiniu) laiku.
- Atsitiktinis cenzūravimas: cenzūravimo laikai C_1, C_2 ... C_n yra atsitiktiniai dydžiai, nepriklausantys nuo gyvavimo laikų U_1, U_2, ..., U_n.

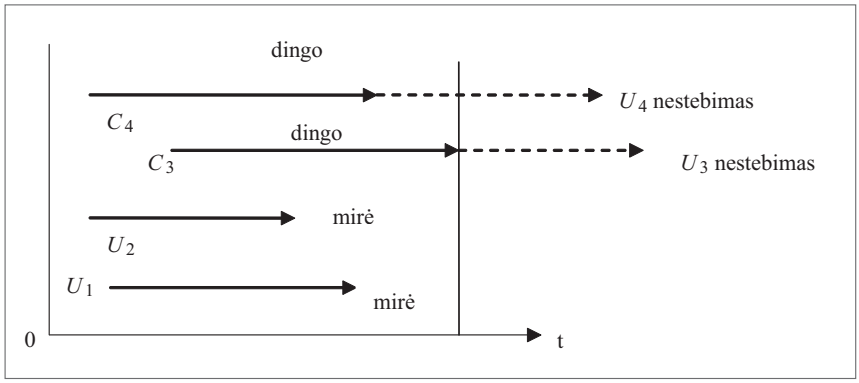
Dažniausiai naudojamas III tipo cenzūravimas. I tipo cenzūravimas yra atskiras III tipo atvejis. Šiame skyriuje apsiribosime III tipo (progresyviu) cenzūravimu.

13.1 skyriaus 13.1 ir 13.3 pavyzdžiuose naudotas III tipo cenzūravimas (ligoniai stebėti skirtingą laiką), 13.2 pavyzdyje – I tipo cenzūravimas (pelės stebėtos vienodą laiką). Nagrinėjant ligonių, sergančių glioblastoma, išgyvenamumą (13.1 pvz., 13.1 lentelė), chemoterapijos grupėje 12 ligonių nustatyti tikslūs išgyvenimo (gyvavimo) laikai (nuo 1 iki 44 savaičių), 8 reikš-

mės buvo cenzūruotos – pateikti tik stebėjimo laikai C_j (ligoniai stebėti nuo 2 iki 45 savaičių). Kontrolinėje grupėje nustatyti 8 tikslūs išgyvenimo laikai, 9 reikšmės cenzūruotos. 13.2 pavyzdyje abiejose pelių grupėse yra po tris cenzūruotas reikšmes. Ligonių, sergančių leukemija, 6-mp grupėje nustatyti 9 tikslūs išgyvenimo laikai (laikas iki remisijos), 12 reikšmių buvo cenzūruotos, placebo grupėje visi laikai iki remisijos – tikslūs (13.1 lentelė).



13.1 pav. Ligonių stebėjimo laikas



13.2 pav. Progresyvus cenzūravimas

13.3. Išgyvenamumo funkcija, rizikos funkcija

Šiame skyriuje pateikiamos sąvokos, reikalingos populiacijos išgyvenamumui apibūdinti.

Sakykime, T – populiacijos individo išgyvenimo nuo stebėjimo pradžios iki baigties taško laikas (gyvavimo laikas). T priklauso nuo populiacijos charakteristikų, individualių organizmo savybių ir daugelio kitų priežasčių,

todėl jį pagrįstai galima laikyti atsitiktiniu dydžiu. Šio atsitiktinio dydžio skirstinio funkciją $P\{T \le t\}$ pažymėkime $F(t)$, tankį $p(t)$. Pagal apibrėžimą (2.4 skyrius), skirstinio funkcija $F(t) = P\{T \le t\}$ yra tikimybė neišgyventi daugiau nei t (laiko vienetų), o funkcija

$S(t) = 1 - F(t) = P\{T > t\}$

yra tikimybė išgyventi daugiau nei t (laiko vienetų). Funkcija $S(t)$ vadinama populiacijos **išgyvenamumo funkcija**. $S(t)$ kinta tarp nulio ir vieneto ir yra nedidėjanti.

Medicininiu požiūriu aktualu įvertinti baigties taško (mirties, susirgimo) riziką – baigties taško atsiradimo tikimybę t momentu su sąlyga, kad iki t momento baigties taško nebuvo. Ši rizika bendru atveju yra laiko t funkcija. Ją žymėsime $h(t)$. Pagal apibrėžimą $h(t)$ lygi:

$$\begin{aligned} h(t) &= \lim_{\Delta \to 0} \{P\{\text{baigties taškas laikotarpiu } (t, t+\Delta] \mid T > t\}/\Delta\} \\ &= \lim_{\Delta \to 0} \{P\{\text{baigties taškas laikotarpiu } (t, t+\Delta],\ T > t\}/(\Delta P\{T > t\})\} \\ &= \lim_{\Delta \to 0} (P\{T > t\} - P\{T > t + \Delta\})/(\Delta P\{T > t\}) \\ &= \lim_{\Delta \to 0} (S(t) - S(t+\Delta))/(\Delta S(t)) = d/dt(F(t))/S(t) = p(t)/S(t). \end{aligned}$$

Remiantis tankio $p(t)$ ir $S(t)$ apibrėžimu, turime

$$h(t) = p(t)/S(t) = [d(-S(t))/dt]/S(t) = d(-\ln(S(t))/dt. \qquad (13.2)$$

Rizikos funkcijos $h(t)$ kitimo pobūdis priklauso nuo baigties taško apibrėžimo bei individo charakteristikų. Pavyzdžiui, mirties rizika priklauso nuo individo amžiaus, biologinių rodiklių (paveldėjimo, susirgimų) bei kitų veiksnių. Pensinio amžiaus individų mirties rizikos funkcija $h(t)$ didėja, didėjant t (13.3 pav.) – nes rizika numirti senatvėje su amžiumi didėja. Nelaimingų atsitikimų ar retų susirgimų rizika yra pastovi (13.3 pav.). Mirties po operacijos rizikos funkcija $h(t)$ – mažėjanti funkcija. Mirties nuo gyvenimo pradžios iki senyvo amžiaus rizikos funkcija yra vonios formos: kūdikystės laikotarpiu ji mažėja, vėliau iki tam tikro amžiaus išlieka pastovi, o senatvėje didėja (13.3 pav.).

Suminiam rizikos vertinimui naudojama suminė rizikos funkcija $H(t)$:

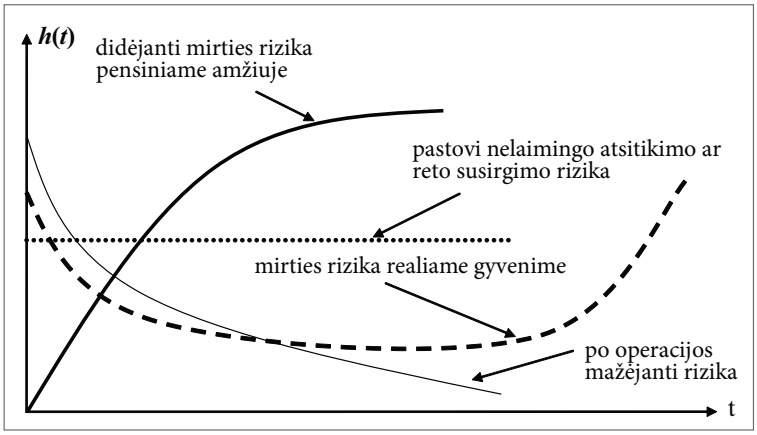
$$H(t) = \int_0^t h(t)dt.$$

Naudojantis (13.2) formule, $S(t)$ galime išreikšti:

$$S(t) = \exp(-\int_0^t h(t)dt) = \exp(-H(t)). \qquad (13.3)$$

Išgyvenamumo modelio pavyzdys. Sakykime, ats. d. T skirstinys yra eksponentinis su parametru $1/\theta$. Tuomet $F(t) = 1 - e^{-t/\theta}$, $t \ge 0$, tankis

$p(t) = (1/\theta)e^{-t/\theta}$, $t \geq 0$. Išgyvenamumo funkcija yra lygi: $S(t) = e^{-t/\theta}$, $t \geq 0$, $h(t) = 1/\theta$, $H(t) = t/\theta$; taigi šiuo atveju baigties taško atsiradimo (mirties) rizika yra pastovi ir lygi skirstinio parametrui $1/\theta$.



13.3 pav. Rizikos kitimas

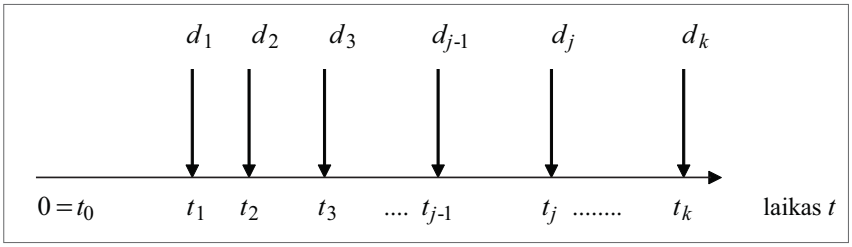
Konkrečios populiacijos išgyvenamumo funkcija vertinama naudojant parametrinį arba neparametrinį išgyvenamumo funkcijos modelį. Naudojant parametrinį metodą, daroma prielaida, jog populiacijos išgyvenamumo funkcija yra tam tikro tipo (sakykime, ekponentinio, Veibulo), tačiau su nežinomais parametrais, kuriuos reikia įvertinti. Dažniausiai naudojamas išgyvenamumo funkcijos neparametrinis įvertis – Kaplano–Mejerio kreivė.

13.4. Išgyvenamumo funkcijos neparametrinis įvertis. Kaplano–Mejerio (*Kaplan–Meier*) kreivė

Sakykime, visiems individams nustatytas gyvavimo laikas – laikas nuo stebėjimo pradžios iki baigties taško (mirties) u_i (cenzūravimo nėra). Tokiu atveju išgyvenamumo funkcijos $S(t)$ įvertis $\hat{S}(t)$ lygus $\hat{S}(t) = n_t / n$; čia n_t – individų, išgyvenusių iki momento t, skaičius, n – populiacijos ar imties individų skaičius. Pažymėkime $t_0 = 0$, $0 < t_1 < ... < t_{j-1} < ... < t_k$ – skirtingi baigties taško atsiradimo laikai, išdėstyti didėjimo tvarka, k – skirtingų baigties taško atsiradimo reikšmių skaičius, d_j – individų, kuriems t_j momentu fiksuotas baigties taškas (mirtis), skaičius (13.4 pav.). Tuomet $n_t = n - (d_1 + d_2 + ... + d_j)$, kai $t_j \leq t < t_{j+1}$, o

$$\hat{S}(t) = \frac{n-d_1}{n} \times \frac{n-d_1-d_2}{n-d_1} \times ... \times \frac{n-d_1-d_2-...-d_j}{n-d_1-d_2-...-d_{j-1}} = (1-\frac{d_1}{r_1})(1-\frac{d_2}{r_2})...(1-\frac{d_j}{r_j}),$$

kai $t_j \le t < t_{j+1}$; j = 1..., k – 1; čia r_j – išgyvenusių iki momento t_j skaičius. Pagal apibrėžimą, $\hat{S}(t) = 1$, kai $0 \le t < t_1$.



13.4 pav. Išgyvenamumo duomenys be cenzūravimo

Tačiau, kaip minėta, individai stebimi tik tam tikrą laiką, nes gana ilgai stebėti ne visada įmanoma. Cenzūruotiems duomenims išgyvenamumo funkcijos $S(t)$ dažniausiai naudojamas įvertis – Kaplano–Mejerio kreivė $\hat{S}(t)$.

Kaplano–Mejerio kreivės sudarymas. Sakykime, t_j – laikai, apibrėžti pagal 13.1 formulę (baigties įvykio ar individo pasitraukimo iš studijos laikas), išdėstyti didėjimo tvarka: $0 < t_1 < ... < t_j < ... < t_k$, k – skirtingų t_j reikšmių skaičius, $t_0 = 0$. Momentinė rizika įvykti baigčiai (numirti) laikotarpiu tarp dviejų gretimų mirties laikų $[t_j, t_{j+1})$ yra:

$q(t_j) = P\{$baigties taškas laikotarpiu $[t_j, t_{j+1})$ / baigties taško iki t_j nebuvo$\}$.

Šios funkcijos įvertis $\hat{q}(t_j)$ yra:

$\hat{q}(t_j) = \{$baigties taškų $[t_j, t_{j+1})$ laikotarpiu skaičius / išgyvenusių iki t_j skaičius$\}$.

Pagal apibrėžimą,

$S(t) = P\{T > t_j\}P\{T > t,\ T > t_j\} = P\{$išgyventi iki $t_j\}P\{$išgyventi laikotarpiu $[t_j, t)\} = S(t_j)\,(1 - q(t)) = S(t_{j-1})\,(1 - q(t_j))\,(1 - q(t)) = (1 - q(t_1))\,...\,(1 - q(t_j))\,(1 - q(t))$.

Pažymėkime:

d_j – baigčių momentu t_j skaičius;
c_j – cenzūruotų reikšmių momentu t_j skaičius;
r_j – individų, stebėtų iki momento t_j, skaičius.

Kadangi žinoma, jog iki momento t_j tikrai buvo r_j išgyvenusių individų, todėl $q(t_j)$ įvertis pagal apibrėžimą lygus $\hat{q}(t_j) = d_j / r_j$, o $S(t)$ įvertis (KM įvertis) lygus:

$\hat{S}(t) = (1 - d_1/r_1)(1 - d_2/r_2)...(1 - d_j/r_j)$, kai t kinta intervale $[t_j, t_{j+1})$, $j = 1, 2\ ...\ k$.

Pagal apibrėžimą, $\hat{S}(t) = 1$, kai $0 \le t < t_1$.

Rizikos funkcijos $h(t)$ įvertis yra:

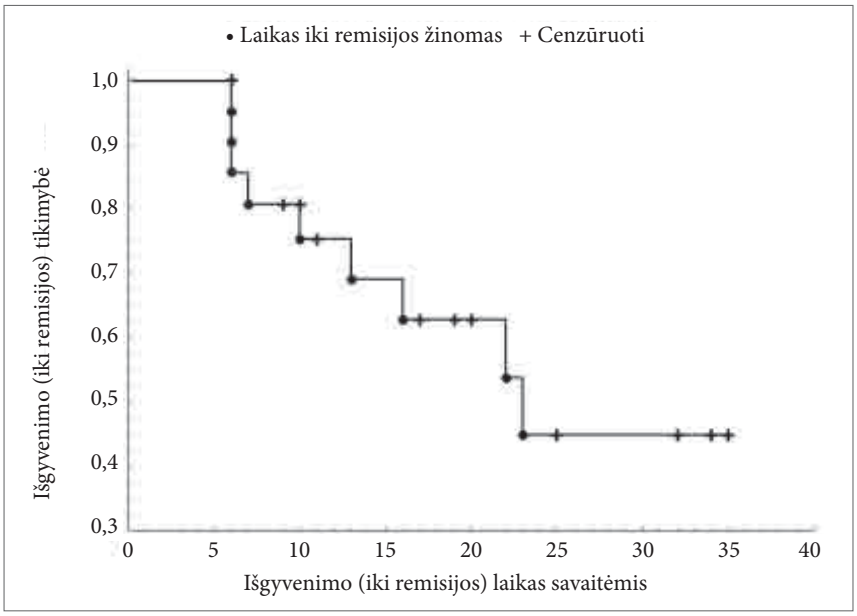
$\hat{h}(t) = d_j / (r_j (t_{j+1} - t_j))$, kai t kinta intervale $[t_j, t_{j+1})$.

Paskaičiuosime ligonių, sergančių leukemija ir gydytų preparatu 6-mp (13.1 lentelė), remisijos laiko Kaplano–Mejerio įvertį. Skirtingi remisijos laikai t_j (savaitėmis) yra: $t_0 = 0$, $t_1 = 6$, $t_2 = 7$ ... $t_{15} = 34$, $t_{16} = 35$. d_j, c_j, r_j bei $\hat{S}(t)$ skaičiavimo duomenys pateikti 13.2 lentelėje. Remiantis joje pateiktu išgyvenamumo funkcijos Kaplano–Mejerio įverčiu, galima teigti, kad ligonių, sergančių leukemija ir gydytų preparatu 6-mp, tikimybė remisijai neatsirasti per 10 savaičių po gydymo yra apie 0,75; per 15 savaičių – 0,69; per 20 savaičių – 0,63.

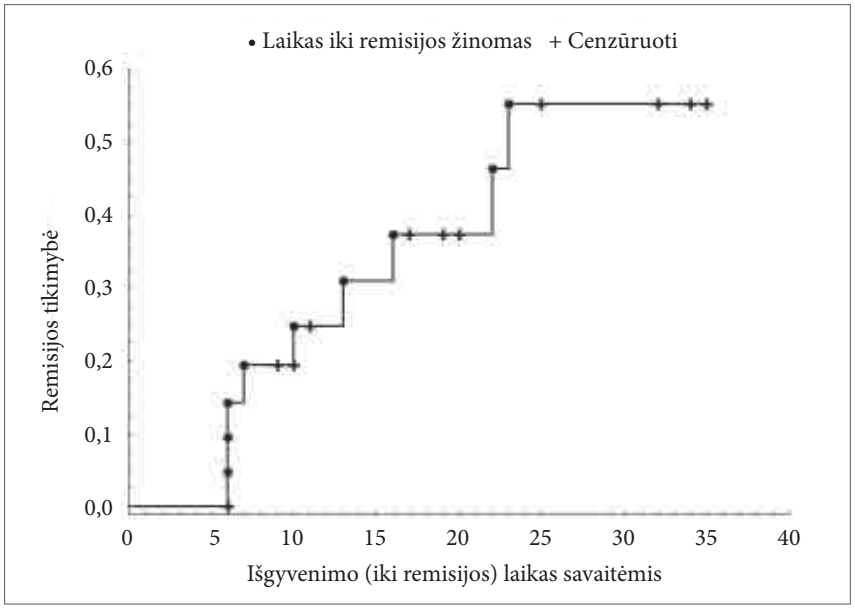
13.2 lentelė. Ligonių, sergančių leukemija ir gydytų preparatu 6-mp, išgyvenamumo funkcijos vertinimas. Pagal apibrėžimą, $\hat{S}(t) = 1$, kai $0 \le t < t_1 = 6$

j	t_j	d_j	c_j	r_j	d_j/r_j	$\hat{S}(t)$	t intervalas
0	0	0	0	21	0	1	[0; 6)
1	6	3	1	21	3/21	0,8571	[6; 7)
2	7	1	0	17	1/17	0,8067	[7; 9)
3	9	0	1	16	0	0,8067	[9; 10)
4	10	1	1	15	1/15	0,7529	[10;11)
5	11	0	1	13	0	0,7529	[11; 13)
6	13	1	0	12	1/12	0,6902	[13; 16)
7	16	1	0	11	1/11	0,6275	[16; 17)
8	17	0	1	10	0	0,6275	[17; 19)
9	19	0	1	9	0	0,6275	[19; 20)
10	20	0	1	8	0	0,6275	[20; 22)
11	22	1	0	7	1/7	0,5378	[22; 23)
12	23	1	0	6	1/6	0,4482	[23; 25)
13	25	0	1	5	0	0,4482	[25; 32)
14	32	0	2	3	0	0,4482	[32; 34)
15	34	0	1	2	0	0,4482	[34; 35)
16	35	0	1	1	0	0,4482	≥ 35

13.5 pav. pateiktas ligonių, sergančių leukemija ir gydytų preparatu 6-mp, išgyvenamumo funkcijos (Kaplano–Mejerio kreivės) grafikas. Rutuliukais pažymėti visiškai žinomi gyvavimo laikai, pliusais – cenzūruoti. Analizuojant populiacijos išgyvenamumą, grafiškai pateikiama ne tik išgyvenamumo funkcija, bet ir baigties taško (mirties ar remisijos) tikimybės intervale $[0, t]$ funkcija $F(t) = 1 - S(t)$, tiksliau, $F(t)$ įvertis $\hat{F}(t) = 1 - \hat{S}(t)$ (13.6 pav.).



13.5 pav. Ligonių, sergančių leukemija ir gydytų preparatu 6-mp, Kaplano–Mejerio išgyvenamumo kreivė



13.6 pav. Ligonių, sergančių leukemija ir gydytų preparatu 6-mp, remisijos tikimybės grafikas

Iš 13.2 lentelės bei 13.5 pav. matyti, kad Kaplano–Mejerio kreivė yra laiptuota funkcija, turinti šuolius tik baigties taškuose $t_j = u_j$, t. y. taškuose t_j, kuriuose stebėtas įvykis (mirtis, remisija ...). Augant n, $\hat{S}(t)$ neartėja į 0.

Kadangi d_j, c_j, r_j – atsitiktiniai dydžiai, tai ir išgyvenamumo funkcijos įvertis $\hat{S}(t)$ yra atsitiktinis dydis. Todėl skaičiuojami $\hat{S}(t)$ pasikliautinieji intervalai, tikrinamos hipotezės apie išgyvenamumo funkcijų lygybę ir t. t.

$\hat{S}(t_j)$ dispersijos įvertis lygus $\Delta^2(t_j) = (\hat{S}(t_j))^2\{d_1/(r_1(r_1 - d_1)) + ... + d_j/(r_j(r_j - d_j))\}$, standartinė paklaida (3.4 skyrius) – $\Delta(t_j)$. $S(t)$ pasikliautinasis intervalas taške $t = t_j$ skaičiuojamas: $\hat{S}(t_j) \pm z_{(\alpha+P)/2}\,\Delta(t_j))$.

Kartais išgyvenamumo funkcija vertinama Flemingo–Haringtono metodu. Šis metodas remiasi tuo, kad $S(t)$ išreiškiama per suminę rizikos funkciją (13.3). Todėl naudodamiesi $H(t)$ apibrėžimu ir $h(t)$ įverčiu, gauname F-H įvertį

$$\hat{S}(t_j) = \exp(-\hat{H}(t)) = \exp(-\sum_{i=1}^{j} \hat{h}(t_j)\,(t_j - t_{j-1})) = \exp(-\sum_{i=1}^{j} d_j / r_j),$$

kai t kinta intervale $[t_j, t_{j+1})$. Šis įvertis efektyvesnis tolydžiam t. Jei duomenys grupuoti, geriau naudoti Kaplano–Mejerio įvertį. Didelių imčių atveju abu metodai lygiaverčiai.

13.5. Dviejų išgyvenamumo funkcijų palyginimas

Norint atsakyti į klausimą, ar toksinės medžiagos vienodai veikia peles (13.2 pavyzdys) arba ar cheminės terapijos poveikis pagerina ligonių išgyvenamumą, reikia tikrinti nulinę hipotezę apie abiejų populiacijų (kontingentų) išgyvenamumo funkcijų $S_0(t)$ ir $S_1(t)$ lygybę:

H_0: $S_0(t) = S_1(t)$ (išgyvenamumo funkcijos tapačios).

Alternatyviosios hipotezės gali būti formuluojamos įvairiai:

H_1: $S_0(t) \neq S_1(t)$; H_2: $S_0(t) < S_1(t)$; H_3: $S_0(t) > S_1(t)$, kai $t < t_0$.

Tikrindami H_0, apsiribosime H_1 alternatyva. H_0 tikrinti pateiksime keletą kriterijų.

Logranginis kriterijus (Mantelio–Henzelio (*Mantel–Haenszel*) kriterijus); (*Logrank test*). Logranginio kriterijaus, kaip ir daugelio neparametrinių kriterijų, pagrindas – stebėtų ir tikėtinų dažnių palyginimas.

Šį kriterijų iliustruosime tirdami preparato 6-mp efektą leukemija sergančių ligonių remisijos laikui (13.1 lentelės duomenys). Tikrinsime nulinę hipotezę, kad abiejų ligonių grupių – 6-mp ir placebo – išgyvenamumo funkcijos vienodos su alternatyva H_1 (išgyvenamumo funkcijos nėra vienodos).

Pažymėkime: $u_0 = 0$, u_1 ... u_k – skirtingi, didėjančia tvarka surašyti sergančių leukemija (tiek 6-mp, tiek placebo grupių) remisijos laikai. Kiekvienam intervalui $[u_j, u_{j+1})$ 6-mp ir placebo grupėse remisijos dažnius galime pateikti 2×2 lentele:

	Remisija		Iš viso
	Įvyko	Neįvyko	
6-mp	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
Placebo	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Iš viso	d_j	$r_j - d_j$	r_j

Čia r_j – ligonių be remisijos iki momento u_j skaičius, d_j – įvykių (remisijų) skaičius momentu u_j. Indeksas 0 – 6-mp, 1 – placebo grupė.

Jei nulinė hipotezė $S_0(t) = S_1(t)$ yra teisinga, remisijos tikimybė $[t_j, t_{j+1})$ laikotarpiu abiejose ligonių grupėse yra vienoda, o remisijų skaičius proporcingas grupės ligonių skaičiui r_{0j} ir r_{1j}. Todėl tikėtini remisijos dažniai 6-mp ir placebo grupėse atitinkamai yra:

$e_{0j} = d_j r_{0j}/r_j$, $e_{1j} = d_j r_{1j}/r_j$.

Šiuos tikėtinus dažnius būtina palyginti su stebėtais d_{0j} ir d_{1j}. Esant fiksuotiems d_j bei r_{0j}, r_{1j}, d_{0j} skirstinys yra hipergeometrinis su parametrais (r_j, r_{0j}, d_j) (2.6 skyrius). Todėl d_{0j} vidurkis lygus $Ed_{0j} = r_{0j}d_j/r_j = e_{0j}$, o dispersija – $Dd_{0j} = (r_{0j}d_j/r_j)(1 - d_j/r_j)(r_j - r_{0j})/(r_j - 1)$.

Skaičiuosime statistiką U:

$U = (d_{01} - e_{01}) + (d_{02} - e_{02}) + ... + (d_{0k} - e_{0k})$ = (stebėtas remisijų skaičius – tikėtinas remisijų skaičius).

Ats. d. U vidurkis ir dispersija lygūs (dydžiai r_j, r_{0j}, d_j fiksuoti):

$$\mathrm{EU} = 0,\ \mathrm{DU} = \sum_{j=1}^{k} (r_{0j} d_j / r_j)(1 - d_j / r_j)(r_j - r_{0j})/(r_j - 1). \qquad (13.4)$$

Logranginio kriterijaus statistika T lygi U^2/DU. Ats. dydžio $U/\sqrt{DU}$ asimptotinis skirstinys yra standartinis normalusis, o T asimptotinis skirstinys – χ^2 su 1 laisvės laipsniu. Logranginio kriterijaus skaičiavimo pavyzdys 6-mp ir placebo išgyvenamumo funkcijoms palyginti pateiktas 13.3 lentelėje. Joje c_{0j} ir c_{1j} – cenzūruotų reikšmių intervale $[u_{j-1}, u_j)$ skaičius atitinkamai 6-mp ir placebo grupėse.

Šiame pavyzdyje DU = 6,2537, logranginio kriterijaus statistika lygi: $T = (9 - 19{,}25)^2/DU = 16{,}8$, kriterijaus p reikšmė – 0,00004. Todėl galime

daryti išvadą, kad 6-mp ir placebo ligonių grupių išgyvenamumo funkcijos reikšmingai skiriasi: 6-mp preparatas – efektyvus vaistas, stabdantis vėžio remisiją.

13.3 lentelė. Logranginio kriterijaus skaičiavimas

t_j	r_{0j}	r_{1j}	c_{0j}	c_{1j}	d_{0j}	d_{1j}	e_{0j}	e_{1j}	$d_{0j} - e_{0j}$	$d_{1j} - e_{1j}$
1	21	21	0	0	0	2	(21/40) × 2	(21/40) × 2	–1	1
2	21	19	0	0	0	0	(21/40) × 2	(19/40) × 2	–1,05	1,05
3	21	17	0	0	0	1	21/38	17/38	–0,55	0,55
4	21	16	0	0	0	2	(21/37) × 2	(16/37) × 2	–1,14	1,14
5	21	14	0	0	0	2	(21/35) × 2	(14/35) × 2	–1,20	1,20
6	21	12	0	0	3	0	(21/33) × 3	(12/33) × 3	1,09	–1,09
7	17	12	1	0	1	0	17/29	12/29	0,41	–0,41
8	16	12	0	0	0	4	(16/28) × 4	(12/28) × 4	–2,29	2,29
10	15	8	1	0	1	0	15/23	8/23	0,35	–0,35
11	13	8	1	0	0	2	(13/21) × 2	(8/21) × 2	–1,24	1,24
12	12	6	1	0	0	2	(12/18) × 2	(6/18) × 2	–1,33	1,33
13	12	4	0	0	1	0	12/16	4/16	0,25	–0,25
15	11	4	0	0	0	1	11/15	4/15	–0,73	0,73
16	11	3	0	0	1	0	11/14	3/14	0,21	0,21
17	10	3	0	0	0	1	10/13	3/13	–0,77	0,77
22	7	2	3	0	1	1	(7/9) × 2	(2/9) × 2	–0,56	0,56
23	6	1	0	0	1	1	(6/7) × 2	(1/7) × 2	–0,71	0,71
Iš viso					9	21	19,26	10,74	–10,26	10,26

Svertinis logranginis kriterijus. Šio kriterijaus statistika lygi:

$$U_w = \sum_{j=1}^{k} w_j (d_{0j} - e_{0j}); \tag{13.5}$$

čia w_j – neneigiami svoriai. Tam tikra logranginių kriterijų šeima naudoja Flemingo–Haringtono svorius:

$$w_j = (\hat{S}(t_j))^{\alpha} (1 - \hat{S}(t_j))^{\gamma}, \alpha \geq 0, \gamma \geq 0; \tag{13.6}$$

čia $\hat{S}(t_j)$ – abiejų sujungtų grupių išgyvenamumo funkcijos Kaplano–Mejerio įvertis.

Svorių w_j (13.6) efektą galime įvertinti taip:

- $\alpha = 0$, $\gamma = 0$: svoriai vienodi;
- $\alpha > 0$, $\gamma > 0$: svoriai didesni *t* intervalo viduryje;
- $\alpha > 0$, $\gamma = 0$: svoriai didesni *t* intervalo pradžioje;
- $\alpha = 0$, $\gamma > 0$: svoriai didesni *t* intervalo gale.

Esant teisingai nulinei hipotezei, statistikos U_w asimptotinis skirstinys yra normalusis su nuliniu vidurkiu. U_w dispersijos įvertis yra:

$$\hat{D}U_w = \sum_{j=1}^{k} w_j^2 D(d_{0j}) =$$

$$\sum_{j=1}^{k} w_j^2 (r_{0j} d_j / r_j)(1 - d_j / r_j)(r_j - r_{0j}) / (r_j - 1).$$

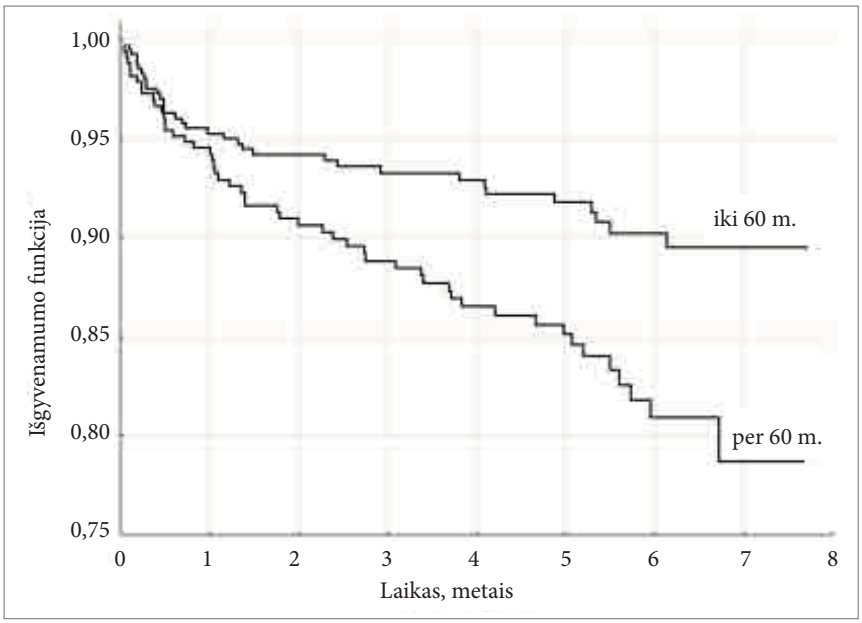
Nulinei hipotezei tikrinti naudojamo kriterijaus statistika $T_w = U_w^2 / \hat{D}U_w$. Jei abiejų grupių išgyvenamumo funkcijos vienodos, tuomet T_w asimptotinis skirstinys yra χ^2 su 1 laisvės laipsniu.

Statistiniuose paketuose naudojami šie svertinio logranginio kriterijaus atvejai:

- Gehano apibendrintas Vilkoksono kriterijus: $w_j = r_j$;
- Taronės–Varo (*Tarone–Ware*) kriterijus: $w_j = \sqrt{r_j}$; w_j apibrėžti (13.6) formule;
- Mantelio–Henzelio arba logranginis kriterijus: $\alpha = 0$, $\gamma = 0$;
- Flemingo–Haringtono kriterijus: $\alpha = -0{,}5$, $\gamma = 0$;
- Peto–Peto–Vilkoksono kriterijus: $\alpha = 1$, $\gamma = 0$.

Statistiniuose paketuose pateikiamos šių kriterijų statistikų bei jų *p* reikšmės. Jei $p < \alpha$ (α – reikšmingumo lygmuo), tvirtiname, kad išgyvenamumo funkcijos nėra lygios. Jei $p \geq \alpha$, išgyvenamumo funkcijų lygybei neprieštaraujama.

13.5 pavyzdys. Lygintos šešiasdešimtmečių ir jaunesnų bei vyresnių nei 60 m. ligonių, persirgusių ūmiais koronariniais sindromais, išgyvenamumo kreivės (13.7 pav.). Šioms kreivėms palyginti skaičiuotos logranginio, Gehano–Vilkoksono bei Peto–Peto–Vilkoksono kriterijaus statistikos reikšmės. Šių kriterijų *p* reikšmės atitinkamai lygios 0,0036; 0,0105 ir 0,0042. Taigi galime tvirtinti, kad iki 60 m. ir vyresnių nei 60 m. ligonių išgyvenamumo funkcijos reikšmingai skiriasi. Iš 13.7 pav. pateiktų Kaplano–Mejerio kreivių galima daryti išvadą, kad maždaug 3 metus išgyvena apie 89 % vyresnių kaip 60 m. amžiaus ligonių, persirgusių ŪKS, ir apie 94 % ne vyresnių kaip 60 m. ligonių. Analogiškai, 5 metus išgyvena apie 85 % vyresnių kaip 60 m. amžiaus ligonių bei 93 % ne vyresnių kaip 60 m. ligonių, persirgusių ŪKS.



13.7 pav. 60 m. ir jaunesnių bei vyresnių kaip 60 m. ligonių, persirgusių ūmiais koronariniais sindromais, išgyvenamumo kreivės

13.6. Kelių išgyvenamumo funkcijų palyginimas

Sakykime, kelerius metus stebimi IŠL sergantys ligoniai. Norima nustatyti, ar jų išgyvenamumas priklauso nuo susirgimo: Q bangos MI, be Q bangos MI, NKA, stabiliosios KA. Tokiu atveju būtina palyginti daugiau nei dviejų ligonių grupių išgyvenamumo funkcijas. Šiam tikslui naudojamas logranginis kriterijus.

Logranginis kriterijus kelioms išgyvenamumo funkcijoms palyginti. Tarkime, tiriama G, $G > 2$, poveikio grupių. Šių grupių išgyvenamumo funkcijos atitinkamai yra $S_1(t)$, $S_2(t)$, ..., $S_G(t)$. Tikrinama nulinė hipotezė:

H_0: $S_1(t) \equiv S_2(t) \equiv ... \equiv S_G(t)$ (grupių išgyvenamumo funkcijos tapačios) su alternatyva: „bent dvi išgyvenamumo funkcijos nėra tapačios“.

Nulinei hipotezei tikrinti naudojamas logranginis kriterijus. Jis skaičiuojamas kaip ir 2 grupių atveju. Pažymėkime: $u_0 = 0$, u_1 ... u_k – didėjančia tvarka surašyti skirtingi visų individų baigties taškų (mirties, remisijos, ...) laikai. Kiekvienam intervalui $[u_j, u_{j+1})$ baigties taškų dažnius grupėse galima pateikti $G \times 2$ porine dažnių lentele:

Grupė	Įvykis (mirtis, susirgimas)		Iš viso
	įvyko	neįvyko	
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
2	d_{2j}	$r_{2j} - d_{2j}$	r_{2j}
...			
G	d_{Gj}	$r_{Gj} - d_{Gj}$	r_{Gj}
Iš viso	d_j	$r_j - d_j$	r_j

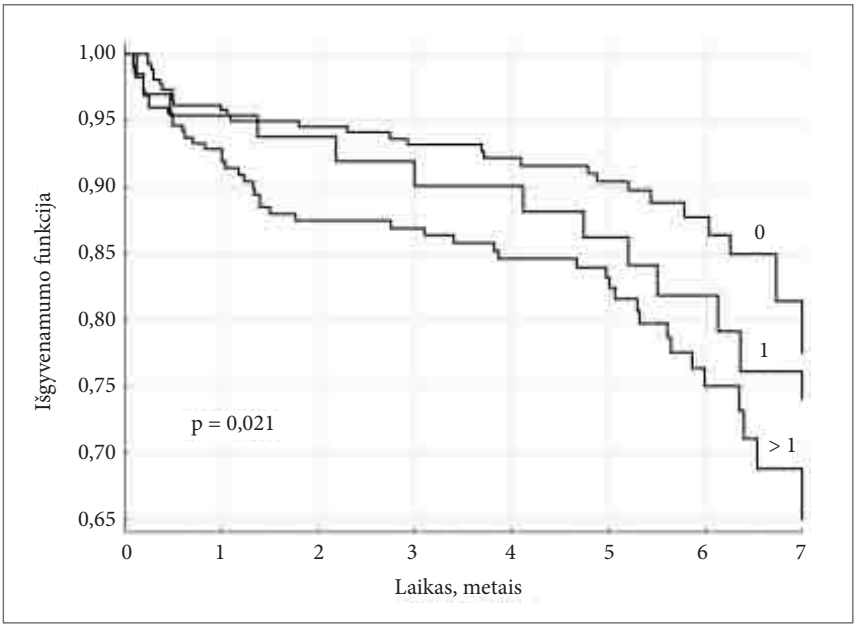
Jei išgyvenamumo funkcijos visose grupėse tapačios, tada kiekvienam intervalui $[u_j, u_{j+1})$ $(G - 1)$-mačio ats. vektoriaus $\boldsymbol{D}_j = (d_{1j}, d_{2j}, ..., d_{(G-1)j})$ skirstinys yra daugiamatis hipergeometrinis. Logranginė statistika U apibrėžiama naudojant daugiamačio atsitiktinio vektoriaus $\boldsymbol{D}_j$ skaitines charakteristikas (16.2 skyrius) – vidurkių vektorių $\boldsymbol{E}_j = (e_{1j}, ..., e_{(G-1)j})$ ir kovariacijų matrica V_j. U išreiškiama naudojant veiksmus su vektoriais ir matrica (15.1 skyrius):

$$U = (\boldsymbol{O} - \boldsymbol{E})^T V^{-1} (\boldsymbol{O} - \boldsymbol{E}), \text{ čia } \boldsymbol{O} = \sum\nolimits_j \boldsymbol{D}_j, \quad \boldsymbol{E} = \sum\nolimits_j \boldsymbol{E}_j, \quad V = \sum\nolimits_j V_j.$$

Esant teisingai H_0, statistikos U asimptotinis skirstinys yra χ^2 su $G - 1$ laisvės laipsnių.

Statistiniuose paketuose pateikiama statistikos U reikšmė bei atitinkama p reikšmė. Jei $p < \alpha$ (α – reikšmingumo lygmuo), galime tvirtinti, kad G išgyvenamumo funkcijų tarpusavyje nėra lygios. Jei $p \geq \alpha$, išgyvenamumo funkcijų lygybei neprieštaraujama.

13.6 pavyzdys. Priklausomai nuo vainikinių arterijų (VA) stenozės daugiau kaip 70 % buvimo, tirtas ligonių, persirgusių ūmiais koronariniais sindromais, išgyvenamumas. 13.8 pav. pateiktos ligonių, neturinčių VA stenozės, turinčių 1 VA stenozę ir turinčių daugiau nei vieną VA stenozę, išgyvenamumo kreivės. Logranginio kriterijaus statistikos reikšmė lygi 8,63, atitinkama p reikšmė – 0,021. Taigi galima teigti, kad ligonių, turinčių skirtingą pažeistų VA skaičių, išgyvenamumo funkcijos nėra vienodos.



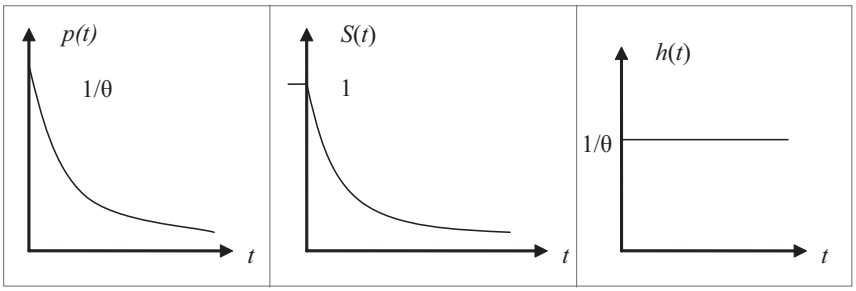
13.8 pav. Ligonių, neturinčių VA stenozės (0), turinčių 1 VA stenozę (1) ir turinčių daugiau nei vienos VA stenozę (>1), išgyvenamumo kreivės

13.7. Išgyvenamumo funkcijos parametriniai modeliai

Kaip minėta 13.3 skyriuje, naudojamas populiacijos išgyvenamumo funkcijos parametrinis ir neparametrinis įvertis. Neparametrinis Kaplano–Mejerio įvertis pateiktas 13.4 skyriuje.

Norint išgyvenamumo funkciją ne tik įvertinti, bet ir prognozuoti, naudojamas parametrinis išgyvenamumo funkcijos modelis: laikoma, kad gyvavimo laiko T skirstinys priklauso tam tikrai parametrinių skirstinių šeimai $P(\theta)$. Pavyzdžiui, T skirstinys yra eksponentinis su nežinomu parametru $1/\theta$, $\theta > 0$. Čia $P(\theta)$ – visi eksponentiniai skirstiniai. Pateiksime keletą dažniausiai naudojamų parametrinių išgyvenamumo funkcijos modelių.

Eksponentinis skirstinys. T skirstinio tankis $p(t)$ lygus $(\theta)^{-1}\exp(-t/\theta)$; čia $\theta > 0$. Tuomet išgyvenamumo funkcija $S(t) = \exp(-t/\theta)$, rizikos funkcija $h(t) = 1/\theta$, suminė rizikos funkcija $H(t) = t/\theta$ (13.9 pav.). Eksponentinio skirstinio atveju rizikos funkcija yra pastovus dydis, lygus $1/\theta$. T vidurkis lygus θ, T dispersija lygi θ^2.

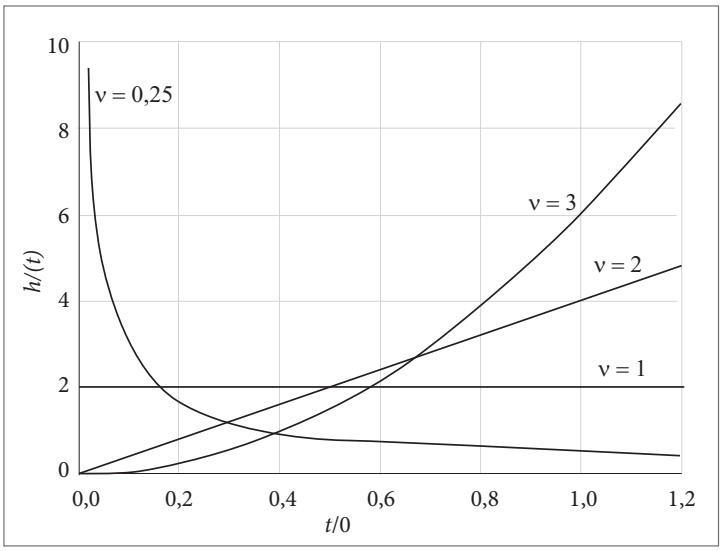


13.9 pav. Eksponentinio skirstinio tankis, išgyvenamumo funkcija, rizikos funkcija

Veibulo skirstinys. T tankis, išgyvenamumo funkcija bei rizikos funkcija yra lygūs:

$p(t) = \nu(\theta)^{-\nu} t^{\nu-1} \exp(-(t/\theta)^{\nu})$; $S(t) = \exp(-(t/\theta)^{\nu})$; $h(t) = \nu(\theta)^{-\nu} t^{\nu-1}$;

čia $\theta > 0$, $\nu > 0$ – skirstinio parametrai. Kai $\nu > 1$, rizikos funkcija yra didėjanti. Esant $\nu = 1$, Veibulo skirstinio tankis sutampa su eksponentinio – rizika pastovi. Kai $\nu < 1$, rizikos funkcija yra mažėjanti (13.10 pav.).



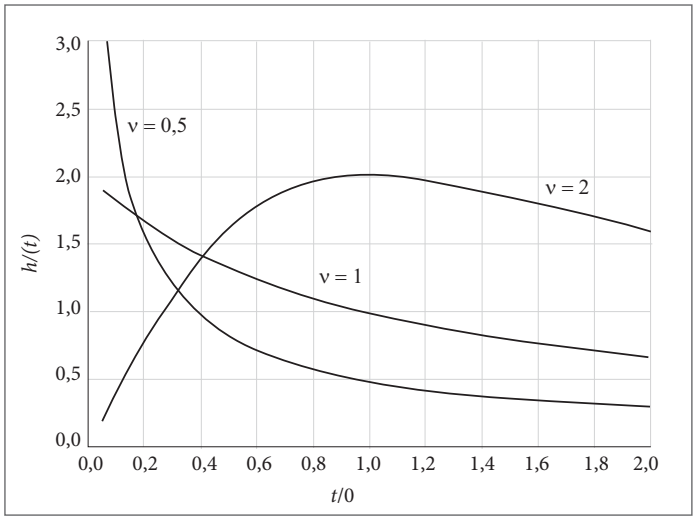
13.10 pav. Veibulo skirstinio rizikos funkcija ($\theta = 2$)

Loglogistinis skirstinys. Šio skirstinio tankis, išgyvenamumo funkcija bei rizikos funkcija lygūs:

$p(t) = \nu(\theta)^{-\nu} t^{\nu-1}(1 + (t/\theta)^{\nu})^{-2}$; $S(t) = 1/(1 + (t/\theta)^{\nu})$;
$h(t) = \nu(\theta)^{-\nu} t^{\nu-1}(1 + (t/\theta)^{\nu})^{-1}$;

čia θ > 0, ν > 0 – skirstinio parametrai. Kai ν > 1, rizikos funkcija yra ∩ formos – iš pradžių didėja, o nuo tam tikro *t* mažėja. Kai ν ≤ 1, rizikos funkcija yra mažėjanti (13.11 pav.).

Parametrinis išgyvenamumo modelis parenkamas pagal rizikos funkcijos kitimo pobūdį. Jei yra pagrindo teigti, kad įvykio rizika pastovi, naudojamas eksponentinis išgyvenamumo funkcijos modelis. Jei kintant *t* rizika didėja ar mažėja, tuomet daroma prielaida, kad išgyvenamumas turi Veibulo ar kitą skirstinį.



13.11 pav. Loglogistinio skirstinio rizikos funkcija (θ = 2)

Išgyvenamumo funkcijos parametrų vertinimas*. Parinkus parametrinį išgyvenamumo funkcijos modelį, būtina įvertinti nežinomus parametrus. Nežinomi parametrai vertinami didžiausio tikėtinumo metodu (3.2 skyrius).

Užrašysime tikėtinumo funkciją cenzūruotiems duomenims. Turime atsitiktinių stebėjimų seką (t_1, δ_1), (t_2, δ_2) ... (t_n, δ_n); čia δ_i – cenzūravimo indikatorius, t_i apibrėžti (13.1) formule. Sakykime, kiekvieno individo gyvavimo laikas T yra ats. d. su tankiu $p(t, \boldsymbol{\theta})$; čia $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_k)$ – nežinomų parametrų vektorius. Jei $\delta_i = 1$, t. y. t_i – stebėtas gyvavimo laikas, tuomet t_i tikėtinumo funkcija lygi $p(t_i, \boldsymbol{\theta})$. Jei $\delta_i = 0$, t. y. t_i – individo stebėjimo laikas, tuomet tik žinoma, kad šio individo gyvavimo laikas viršys t_i. Tokiu atveju t_i tikėtinumo funkcija lygi $S(t_i, \boldsymbol{\theta})$. Bet kurios poros (t_i, δ_i) tikėtinumo funkcija lygi:

$$(p(t_i,\theta))^{\delta_i}(S(t_i,\theta))^{1-\delta_i}.$$

Imties (t_1, δ_1), (t_2, δ_2) ... (t_n, δ_n) tikėtinumo funkcija lygi:

$$L(\theta) = \prod_{i=1}^{n} (p(t_i,\theta))^{\delta_i} (S(t_i,\theta))^{1-\delta_i} =$$
$$= \prod_{i=1}^{n} (S(t_i,\theta)h(t_i,\theta))^{\delta_i} (S(t_i,\theta))^{1-\delta_i} =$$
$$= \prod_{i=1}^{n} (h(t_i,\theta))^{\delta_i} S(t_i,\theta).$$

Log–tikėtinumo funkcija gaunama logaritmuojant $L(\boldsymbol{\theta})$:

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \{\delta_i \ln h(t_i,\theta) + \ln S(t_i,\theta)\}. \qquad (13.7)$$

$l(\boldsymbol{\theta})$ dalines išvestines θ_i atžvilgiu prilyginę 0 ir išsprendę lygčių sistemą θ_i atžvilgiu, gauname nežinomų parametrų $\boldsymbol{\theta}$ įverčius $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$. $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$ kovariacijų matricos įvertis lygus Fišerio informacijos matricos įverčio atvirkštinei matricai (3.5 skyrius).

13.7 pavyzdys. Sakykime, T skirstinys yra eksponentinis su nežinomu parametru θ:

$S(t, \theta) = \exp(-t/\theta)$, $h(t) = 1/\theta$.

Didžiausio tikėtinumo metodu rasime θ įvertį. Tikėtinumo funkcijos logaritmas (13.7) lygus:

$$l(\theta) = \sum_{i=1}^{n} \{\delta_i \ln(1/\theta) - t_i/\theta\}.$$

$l(\theta)$ išvestinę prilyginę 0, gauname:

$$\partial l(\theta)/\partial\theta = -\frac{1}{\theta}\sum_{i=1}^{n}\delta_i + \frac{1}{\theta^2}\sum_{i=1}^{n} t_i = 0,$$

$$\theta = (\sum_{i=1}^{n} t_i / \sum_{i=1}^{n}\delta_i) = \tau_n/\Delta_n;$$

čia Δ_n – baigties taškų (įvykių) skaičius, τ_n – gyvavimo ir stebėjimo laikų suma.

Nustatysime $D(\hat{\theta} - \theta)$ įvertį. Šiuo atveju Fišerio informacijos matrica yra vienas skaičius:

$$-\partial^2 \ln l(\theta)/\partial\theta^2 = -\Delta_n/\theta^2 + 2\tau_n/\tau^3,$$

o jos įvertis

$$-\frac{\partial^2 \ln l(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} = -\frac{\Delta_n^3}{\tau_n^2} + \frac{2\Delta_n^3}{\tau_n^2} = \frac{\Delta_n}{\hat{\theta}_n^2}.$$

$D(\hat{\theta} - \theta)$ įvertis lygus $\tau_n^2 / \Delta_n^3 = \hat{\theta}^2 / \Delta_n$.

Remiantis didžiausio tikėtinumo metodu gautų įverčių savybėmis, $\hat{\theta}$ skirstinys yra normalusis su vidurkiu θ ir dispersija θ^2/Δ_n. Todėl galima užrašyti θ pasikliautinąjį intervalą: $\hat{\theta} \pm z_{(1+P)/2}\, \hat{\theta}/\sqrt{\Delta_n}$; čia P – patikimumas, z_α – $N(0, 1)$ skirstinio α lygio kvantilis.

Dviejų parametrinių išgyvenamumo funkcijų palyginimas. Jei daroma prielaida, kad dviejų ligonių grupių išgyvenamumo funkcijos $S_0(t) = S(t, \boldsymbol{\theta}_0)$ ir $S_1(t) = S(t, \boldsymbol{\theta}_1)$ priklauso tam tikrai skirstinių klasei, tuomet $S_0(t) = S_1(t)$ tada ir tik tada, kai $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$; t. y. vietoje hipotezės $S_0(t) = S_1(t)$ pakanka patikrinti parametrinę H_0: $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$. Alternatyva $S_0(t) \neq S_1(t)$ analogiškai keičiasi į alternatyvą: $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$. H_0 tikrinti naudojami standartizuoto santykio (5.12) tipo kriterijai.

13.8. Regresiniai išgyvenamumo modeliai

Ankstesniuose skyriuose nagrinėjome populiacijos išgyvenamumą. Tačiau konkretaus populiacijos individo išgyvenamumas priklauso ir nuo individo rodiklių. Pavyzdžiui, asmenų, persirgusių miokardo infarktu (MI), išgyvenamumas priklauso nuo ligonio amžiaus, MI sunkumo, koronarų stenozės ir kitų rodiklių. Todėl individo išgyvenamumą tikslinga laikyti atsaku į jį veikiančius neatsitiktinius faktorius.

Pažymėkime $T_1, T_2 \ldots T_n$ – individų gyvavimo laikai. Kiekvienam individui nustatytas p faktorių (kovariačių) reikšmių vektorius $\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)} \ldots x_i^{(p)})$. i-tojo individo gyvavimo laikui T_i modeliuoti tikslinga naudoti regresinius modelius: T_i skirstinį laikyti kovariatės $\boldsymbol{x}_i$ funkcija. Pateiksime keletą regresinių išgyvenamumo modelių.

Eksponentinės regresijos modelis. Daroma prielaida, kad i-tojo individo gyvavimo laikas T_i turi eksponentinį skirstinį su parametru θ_i, priklausančiu nuo kovariatės $\boldsymbol{x}_i$ reikšmės:

$$ET_i = (1/\theta_i) = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)}$$

arba $1/\theta_i = \exp(\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)})$.

Nežinomi modelio parametrai β_1, β_2 ... β_p vertinami dažniausiai didžiausio tikėtinumo metodu.

Beilio–Meikhemo (*Bailey–Makeham*) modelis. Daroma prielaida, kad *i*-tojo individo rizikos funkcija lygi $h_i(t) = \alpha(\boldsymbol{x}_i)\exp(-\gamma(\boldsymbol{x}_i)t) + \delta(\boldsymbol{x}_i)$; čia parametrai α, γ, δ – kovariačių funkcijos:

$$\alpha(\boldsymbol{x}_i) = \exp(\alpha_0 + \alpha_1 x_i^{(1)} + \alpha_2 x_i^{(2)} + \ldots + \alpha_p x_i^{(p)}),$$

$$\gamma(\boldsymbol{x}_i) = \exp(\gamma_0 + \gamma_1 x_i^{(1)} + \gamma_2 x_i^{(2)} + \ldots + \gamma_p x_i^{(p)}),$$

$$\delta(\boldsymbol{x}_i) = \exp(\delta_0 + \delta_1 x_i^{(1)} + \delta_2 x_i^{(2)} + \ldots + \delta_p x_i^{(p)}).$$

Pagreitintas išgyvenamumo modelis (*accelerate failure model*). Šiame modelyje daroma tokia prielaida T_i skirstiniui:

$$\log(T_i) = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i; \tag{13.8}$$

čia ε_i – nepriklausomi ats. d., turintys tą patį skirstinį, β_1, β_2 ... β_p – nežinomi regresijos koeficientai. Dydis $\beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)}$ rodo kovariačių poveikį išgyvenamumui: jam didėjant, didėja išgyvenamumo laiko vidurkis. *j*-tosios kovariatės poveikis išgyvenamumui pasireiškia pagal β_j ženklą: jei $\beta_j > 0$, didėjant $x_i^{(j)}$, individo išgyvenimo vidurkis didėja; jei $\beta_j < 0$, didėjant $x_i^{(j)}$, individo išgyvenimo vidurkis mažėja, esant fiksuotoms kitų kovariačių reikšmėms.

Modelį (13.8) galima perrašyti taip:

$$T_i = \exp(\beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)})\eta_i;$$

čia $\eta_i = \exp(\varepsilon_i)$, $i = 1, 2 \ldots n$. η_i skirstinys vadinamas baziniu išgyvenamumo skirstiniu (*baseline survival distribution*). Jei ε_i skirstinys yra normalusis, tuomet η_i skirstinys lognormalusis. Daugelyje praktinių taikymų laikoma, kad η_i skirstinys yra Veibulo.

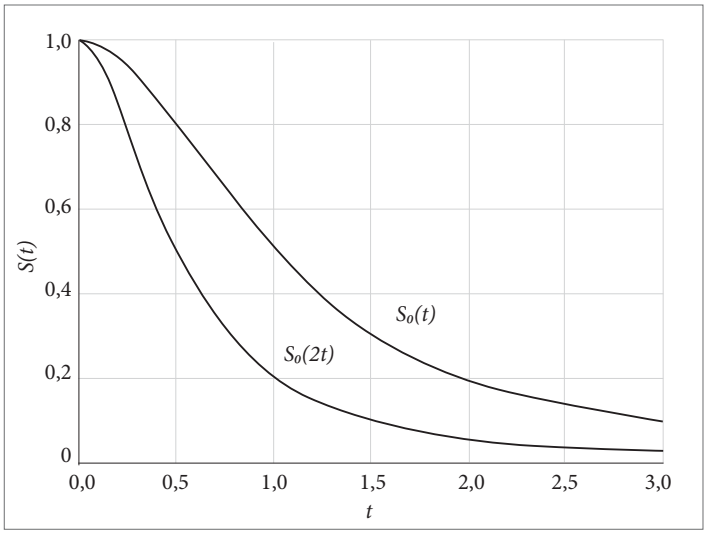
Pagreitintame išgyvenamumo modelyje *i*-tojo individo išgyvenamumo funkcija $S_i(t)$ lygi:

$$S_i(t) = S_0(\exp(-(\beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)})t));$$

čia $S_0(t)$ – bazinio skirstinio išgyvenamumo funkcija. Taigi šiame modelyje kovariačių efektas pasireiškia poveikiu laiko ašies masteliui (13.12 pav.).

Jei žinomas tikslus bazinis skirstinys, pavyzdžiui, Veibulo, tuomet žinoma T_i tankio ir išgyvenamumo funkcijos parametrinė išraiška. Įstačius šias iš-

raiškas į imties $(t_1, \delta_1) \ldots (t_n, \delta_n)$ tikėtinumo funkciją (13.7), gautas išvestines nežinomų parametrų atžvilgiu prilyginus 0 ir išsprendus šią lygčių sistemą, gaunami regresijos koeficientų įverčiai.



13.12 pav. Kovariačių efektas išgyvenamumo funkcijai

13.9. Proporcingos rizikos modelis

Analizuojant išgyvenamumą, aktualu įvertinti kovariatės įtaką gyvavimo laikui, neatsižvelgus į pasirinktą parametrinį modelį. Anksčiau nagrinėtuose regresiniuose modeliuose tai padaryti sudėtinga. Todėl labai populiarus pusiau parametrinis išgyvenamumo modelis, vadinamas **proporcingos rizikos** (*proportional hazard*), arba Kokso (*Cox*), modeliu.

Sakykime, x – kovariatės (faktoriaus) X, lemiančios individo sveikatos būklę, reikšmė. Proporcingos rizikos (Kokso) modelio prielaida – individo, turinčio kovariatės reikšmę x, rizikos funkcija $h(t; x)$ lygi:

$$h(t; x) = h_0(t)\exp(\beta x); \tag{13.9}$$

čia β – regresijos koeficientas, $h_0(t)$ – neneigiama t funkcija, vadinama **bazine rizikos funkcija** (*baseline hazard*). $h_0(t)$ parametrinė išraiška nėra žinoma; $h_0(t)$ interpretuojama kaip individo su nuline kovariatės reikšme rizikos funkcija.

Kokso modelyje (13.9) naudojama neparametrinė funkcija $h_0(t)$, tačiau jame yra ir parametras – regresijos koeficientas β. Todėl šis modelis vadina-

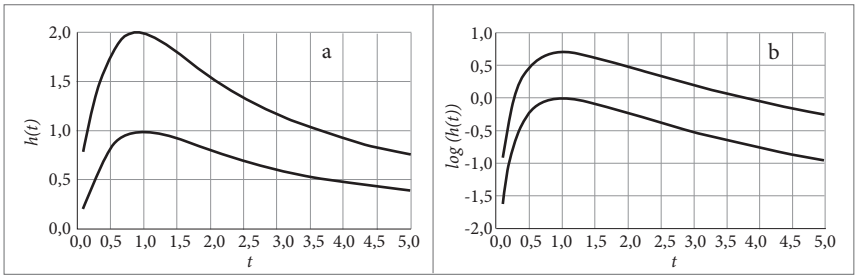
mas pusiau parametriniu. Kadangi individų su kovariatės reikšmėmis x_1 ir x_2 rizikos funkcijų santykis

$$\frac{h(t;x_1)}{h(t;x_2)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta(x_1-x_2)} \tag{13.10}$$

nuo t nepriklauso (13.13 a pav.), todėl šis modelis vadinamas proporcingos rizikos (**PR**) modeliu. Logaritmuodami (13.10) reiškinį gauname, kad

$\log(h(t; x_1)) - \log(h(t; x_2)) = \beta(x_1 - x_2)$.

Tai reiškia, kad PR modelyje logaritmuotos rizikos funkcijos skiriasi tik pastoviu dydžiu (13.13 b pav.).



13.13 pav. Individų su skirtingomis kovariatės reikšmėmis rizikos funkcijos (a) ir logaritmuotos rizikos funkcijos (b)

Regresijos koeficientas β PR modelyje interpretuojamas taip. Pagal apibrėžimą,

$h(t; x + 1) = h_0(t)\exp(\beta(x + 1)) = h(t; x)\exp(\beta)$,

arba $\exp(\beta) = h(t; x + 1)/h(t; x)$. Taigi $\exp(\beta)$ parodo, kiek kartų padidėja rizikos funkcija, kovariatės reikšmei padidėjus 1. Jei $\beta > 0$, didėjant x, individo rizika didėja; o jei $\beta < 0$, didėjant x, rizika mažėja. Dydis $\exp(\beta)$ vadinamas **pavojaus rizika (*hazard ratio*).**

Sakykime, kovariatė X yra kokybinis kintamasis, pavyzdžiui, $x = 1$ – individas serga CD, $x = 0$ – individas neserga CD. Tuomet, pagal (13.9) apibrėžimą, nesergančių CD rizikos funkcija lygi $h_0(t)$, o sergančių CD rizikos funkcija lygi $h_0(t)\exp(\beta)$. Šiuo atveju dydis $\exp(\beta)$ parodo, kiek kartų sergančių CD rizikos funkcija didesnė nei nesergančių CD.

Pagal (13.3) apibrėžimą, individo su kovariatės reikšme x išgyvenamumo funkcija $S(t; x)$ lygi:

$$S(t;x) = \exp(-\int_0^t h_0(u)\exp(\beta x)du)) =$$
$$= \exp(-\int_0^t h_0(u)du)\exp(\beta x) = (S_0(t))^{\exp(\beta x)}; \quad (13.11)$$

čia $S_0(t)$ bazinė išgyvenamumo funkcija: $S_0(t) = \exp(-\int_0^t h_0(u)du))$.

Remiantis (13.11) formule, galima tvirtinti, kad kai $\beta > 0$, tai didėjant x, individo išgyvenamumas prastėja, nes $S_0(t) < 1$, o jei $\beta < 0$, tai didėjant x, individo išgyvenamumas gerėja. Du kartus logaritmuodami (13.11) išraišką, gauname:

$$\log(-\log(S(t;x))) = \log(-\log(S_0(t))) + \beta x, \quad (13.12)$$

taigi proporcingos rizikos modelyje funkcijos $\log(-\log(S(t;x)))$ visoms x reikšmėms skiriasi tik pastoviu dydžiu.

Daugiamatis Kokso modelis. Sakykime, $x^{(1)}$, $x^{(2)}$ ... $x^{(p)}$ – individo kovariačių reikšmės, $\boldsymbol{x} = (x^{(1)}, x^{(2)} \ldots x^{(p)})$ – individo kovariačių vektorius. Laikome, kad šio individo rizikos funkcija $h(t;\boldsymbol{x})$ lygi:

$$h(t;\boldsymbol{x}) = h_0(t)\exp(\beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_p x^{(p)});$$

čia $\beta_1, \beta_2 \ldots \beta_p$ – regresijos koeficientai, $h_0(t)$ – bazinė rizikos funkcija. Daugiamačio modelio parametrai $\beta_1, \beta_2 \ldots \beta_p$ interpretuojami taip: dydis $\exp(\beta_i)$ parodo, kiek kartų padidėja individo rizikos funkcija, i-tosios kovariatės reikšmei padidėjus 1, o kitoms kovariatėms nekintant. Todėl $\exp(\beta_i)$ galima laikyti i-tosios kovariatės pavojaus rizika, standartizuota likusių kovariačių atžvilgiu; $\exp(\beta_i)$ vadinama **koreguota (standartizuota) pavojaus rizika**. Dydis $\exp(\beta)$, nustatytas vienmačiame Kokso modelyje (13.9), rodo rodiklio keliamą pavojaus riziką neatsižvelgiant į kitų rodiklių įtaką – $\exp(\beta)$ vadinama **izoliuota pavojaus rizika**.

Kaip ir vienmačiame modelyje,

$$S(t;\boldsymbol{x}) = (S_0(t))^{\exp(\beta_1 x^{(1)} + \ldots + \beta_p x^{(p)})}.$$

Tiek vienmačio, tiek daugiamačio Kokso modelio parametrai $\beta_1, \beta_2 \ldots \beta_p$ vertinami minimizuojant dalinę tikėtinumo funkciją [2, 3, 6]. Nustačius $\beta_1, \beta_2 \ldots \beta_p$ įverčius, gaunamas $h_0(t)$ bei $S_0(t)$ neparametriniai įverčiai. Visi šie skaičiavimai atliekami programiniais paketais STATISTICA, SPSS, SAS,

S–plus. Programose, skirtose išgyvenamumo duomenims apdoroti, pateikiami regresijos koeficientų įverčiai, jų standartinės paklaidos bei kriterijaus, skirto β_i reikšmingumui tikrinti, p reikšmės. Statistiniuose paketuose taip pat pateikiami $h_0(t)$ bei $S_0(t)$ įverčiai taškuose $t_1 < t_2 < ... < t_k$; čia t_i – skirtingi gyvavimo ar stebėjimo laikai. Ar Kokso modelis reikšmingai mažina imties tikėtinumo funkciją, nustatoma pagal tikėtinumų santykio kriterijaus p reikšmę: jei $p < 0,05$ – PR modelis reikšmingai mažina tikėtinumo funkciją; jei $p \geq 0,05$ – kovariatę įtraukus į PR modelį, tikėtinumo funkcija reikšmingai nesumažėja. Todėl PR arba, Kokso, modelį tokių išgyvenamumo duomenų analizei taikyti netikslinga.

Interpretuojant Kokso modelio rezultatus, svarbu ir pavojaus rizikos $\exp(\beta_i)$ patikimumas, todėl statistiniuose paketuose pateikiami dydžiai $\exp(\beta_i)$ su pasikliautinaisiais intervalais. Jei $\exp(\beta_i)$ pasikliautinojo intervalo ribos didesnės nei 1, i-toji kovariatė didėdama reikšmingai didina baigties taško riziką. Jei $\exp(\beta_i)$ pasikliautinojo intervalo ribos mažesnės nei 1, i-tajai kovariatei didėjant, rizika reikšmingai mažėja. Jei į $\exp(\beta_i)$ pasikliautinąjį intervalą patenka 1, i-toji kovariatė išgyvenamumui reikšmingos įtakos neturi.

Naudojant įvairius kovariačių rinkinius, galima sudaryti gana daug daugiamačių PR modelių. Optimaliu laikytinas modelis, turintis mažiausią tikėtinumų santykio kriterijaus p reikšmę bei visus reikšmingus β_i. Šis modelis sudaromas analogiškai daugialypės regresijos modeliui. Pirmiausia atrenkami kintamieji (kovariatės), turintys įtakos išgyvenamumui. Tai nustatoma remiantis vienmačiu Kokso modeliu. Jei koeficientas β reikšmingai skiriasi nuo 0, kintamasis turi įtakos išgyvenamumui ir jį galima naudoti daugiamačiui modeliui sudaryti. Naudojant atrinktus informatyvius rodiklius, optimalus daugiamatis Kokso modelis sudaromas analogiškai daugialypės tiesinės regresijos modeliui, tik kintamieji įtraukiami ar pašalinami remiantis koeficientų reikšmingumo kriterijaus p reikšme.

Proporcingos rizikos modelio prielaidų tikrinimas. Ar kovariatės įtaką išgyvenamumui gerai atspindi Kokso modelis, nustatoma tikrinant PR modelio prielaidas (13.9) ir (13.12). Tam naudojami grafiniai ir statistiniai kriterijai. Jei kovariatė – kokybinis kintamasis, įgyjantis reikšmes $x_1 ... x_l$, tikrinant PR prielaidą – santykis $h(t; x_i)/h(t; x_j)$ nuo t nepriklauso – naudojami šie grafiniai metodai:

- Išgyvenamumo kreivė, įvertinta Kokso modeliu, $S(t; x_i)$ lyginama su Kaplano–Mejerio išgyvenamumo kreive. Jei Kokso modelio prielaida teisinga, $S(t; x_i)$ neturėtų labai skirtis nuo Kaplano–Mejerio kreivės, nustatytos individams su kovariatės reikšme x_i.

- Log(–log ($S(t; x_i)$)) grafiko analizė; tikrinama, ar šios kreivės, nustatytos skirtingoms x_i reikšmėms Kaplano–Mejerio metodu, skiriasi tik poslinkiu.

Jei kovariatė – kiekybinis kintamasis, tuomet Kokso modelio prielaidai tikrinti skaičiuojami Šionfildo likučiai (*Schoenfeld residuals*):

$$r_j = \sum_{T_i=t_j} \delta_i (x_i - \bar{x}(t_j)),$$

$j = 1, 2 \ldots k$; čia $0 < t_1 < t_2 < \ldots < t_k$ – skirtingi išgyvenimo ar cenzūravimo laikai, išdėstyti didėjimo tvarka, k – skirtingų laikų skaičius, T_i – i-tojo individo gyvavimo ir stebėjimo laikas, δ_i – cenzūravimo indikatorius, x_i – kovariatės reikšmė,

$$\bar{x}(t_j) = \frac{\sum_{t_i \geq t_j} x_i e^{\beta x_i}}{\sum_{t_i \geq t_j} e^{\beta x_i}}.$$

Jei individo rizika didėja proporcingai x, tuomet r_j, $j = 1 \ldots k$ yra atsitiktiniai, apie 0 svyruojantys dydžiai, be to, didėjant t, r_j artėja į 0.

13.10. Proporcingos rizikos modelio taikymas ligonių, persirgusių ūmiais koronariniais sindromais, išgyvenamumo analizei

13.8 pavyzdys. Kardiologijos instituto Klinikinės kardiologijos laboratorijoje 2001–2004 m. analizuotas 725 ligonių, 1997–2002 m. persirgusių ūmiais koronariniais sindromais (Q bangos MI, be Q bangos MI, nestabiliąja krūtinės angina), išgyvenamumas po ūmaus koronarinio sindromo ([7]). Stebėjimo laikotarpiu fiksuotos 85 kardiovaskulinės mirtys; likę 640 ligoniai stebėti nuo 1 iki 9 metų. Šiam ligonių kontingentui proporcingos rizikos metodu:

- nustatyti rodikliai, turintys reikšmingą įtaką išgyvenamumui;
- sudarytas kompleksinis modelis, skirtas individo mirties rizikai vertinti.

Rodikliai, turintys reikšmingą įtaką išgyvenamumui, nustatyti vienmačiu Kokso modeliu. Šių rodiklių koeficientų β įverčiai, kriterijaus, skirto β reikšmingumui tikrinti, p reikšmės, pavojaus rizika su pasikliautinaisiais intervalais pateikta 13.4 lentelėje.

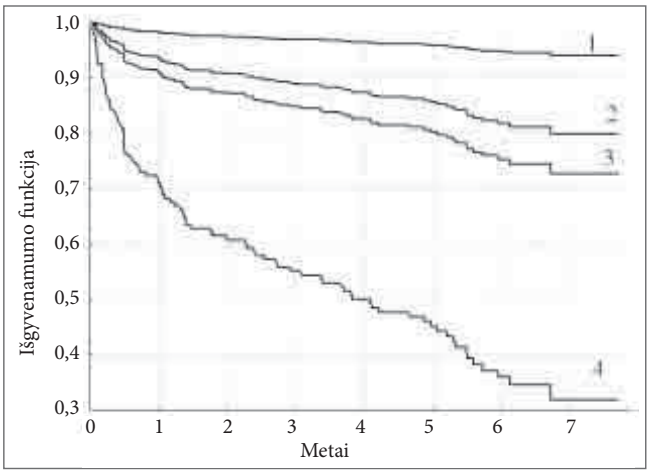
13.4 lentelė. Kokso modelio koeficientų β įverčiai, jų p reikšmė, izoliuota pavojaus rizika e^{β} su pasikliautinaisiais intervalais (PI) (IF – išstūmimo frakcija, DF – diastolinė funkcija)

Rodiklis	β įvertis	*p*	e^{β}	e^{β} PI
Amžius (dešimtmečiais)	0,415	< 0,001	1,514	1,20– 1,91
Išplitęs infarktas (yra, nėra)	0,984	< 0,001	2,674	1,59–4,50
UŠN klasė	0,696	< 0,001	2,006	1,43–2,81
Neatlikta miokardo revaskuliarizacija	0,619	0,01	1,858	1,16–2,98
IF ≤ 25	0,927	0,029	2,528	1,10–5,80
Sutrikusi DF (yra, nėra)	0,646	0,011	1,908	1,16–3,13
Mitalinio vožtuvo nesandarumas (yra, nėra)	1,056	0,01	2,876	1,66–4,97
VA stenozių skaičius	0,293	0,035	1,34	1,02–1,76

Sudarant kompleksinį modelį, skirtą individo mirties rizikai vertinti, visi 13.4 lentelėje pateikti rodikliai įtraukti į daugiamatį Kokso modelį. Iš šio modelio po vieną pašalinus rodiklius su nereikšmingais koeficientais β, gaunamas optimalus daugiamatis Kokso modelis, skirtas įvertinti ligonio išgyvenamumui pagal jo būklės rodiklius. Daugiamačio Kokso modelio koeficientų β įverčiai, jų *p* reikšmės, standartizuota pavojaus rizika su pasikliautinaisiais intervalais pateikta 13.5 lentelėje. Iš jos matyti, kad į daugiamatį modelį įtrauktas amžius, UŠN klasė, infarkto išplitimas, miokardo revaskuliarizacijos atlikimas bei diastolinė funkcija. 13.14 pav. pateiktos išgyvenamumo funkcijos, apskaičiuotos šiuo daugiamačiu Kokso modeliu.

13.5 lentelė. Daugiamačio Kokso modelio koeficientų β įverčiai, jų p reikšmės, standartizuota pavojaus rizika e^{β} su pasikliautinaisiais intervalais

Rodiklis	β įvertis	*p*	e^{β}	e^{β} PI	Rizikos balas
Amžius (dešimtmečiais)	0,385	0,002	1,47	1,16–1,87	1
Išplitęs infarktas	0,824	0,005	2,28	1,29–4,03	3
UŠN klasė	0,553	0,003	1,74	1,20–2,52	2
Neatlikta miokardo revaskuliarizacija	0,762	0,003	2,14	1,30–3,54	3
Sutrikusi diastolinė funkcija (yra, nėra)	0,594	0,023	1,81	1,09–3,02	2



13.14 pav. Išgyvenamumo funkcijos, apskaičiuotos daugiamačiu Kokso modeliu įvairioms ligonių rodiklių kombinacijoms:

1. amžius = 50 m.; USN = 2; neišplitęs MI; atlikta miokardo revaskuliarizacija; DF nesutrikusi;
2. amžius = 50 m.; USN = 3; neišplitęs MI; neatlikta miokardo revaskuliarizacija; DF nesutrikusi;
3. amžius = 60 m.; USN = 2; išplitęs MI; neatlikta miokardo revaskuliarizacija; DF nesutrikusi;
4. amžius = 80 m.; USN = 3; išplitęs MI; neatlikta miokardo revaskuliarizacija; DF nesutrikusi

Individui nepalankaus įvykio pavojaus rizikai vertinti skaičiuojamas rizikos balas. Jis sudaromas remiantis daugiamačio Kokso modelio koeficientų reikšmėmis, analogiškai rizikos balui logistinėje regresijoje. Į Kokso modelį įtrauktų kintamųjų reikšmės dauginamos iš svorių, proporcingų dydžiams e^{b_j}, po to gautos reikšmės sudedamos. 13.5 lentelėje pateikti į daugiamatį Kokso modelį įtrauktų kintamųjų rizikos svoriai.

13.11. Kiti išgyvenamumo modeliai

Proporcingos rizikos modelio apibendrinimas. Jei Kokso modelio sąlyga nėra patenkinta, naudojami bendresni negu (13.9) modeliai.

- **Kokso modelis su kintamomis kovariatėmis**. Sakykime, $x_i(t)$ – i-tojo individo kovariatės reikšmė, nustatyta t momentu. Daroma prielaida, kad i-tojo individo rizikos funkcija lygi:

$$h_i(t) = h_0(t)e^{\beta(t)x_j};$$

čia β – regresijos koeficientas, $h_0(t)$ – bazinė rizikos funkcija. Realiuose tyrimuose dažniausiai pasitaiko tokio tipo nuo t priklausančios kovariatės:

$$x(t) = \begin{cases} 0, 0 < t < z_i, \\ 1, \ t \geq z_i. \end{cases}$$

Pavyzdžiui, tokia kovariate nurodoma operacija, atlikta praėjus z_i laiko nuo stebėjimo pradžios.

- **Neproporcingos rizikos modelis**:

$$h_i(t) = h_0(t)e^{\beta(t)x_j};$$

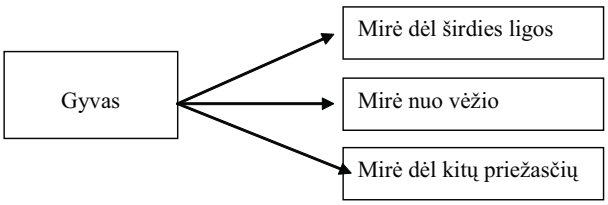
čia $\beta(t)$ – regresijos funkcija. Dažniausiai naudojami modeliai su atskirais $\beta(t)$ atvejais: $\beta(t) = \beta + \alpha Q(t)$; čia $Q(t)$ – nepriklausanti nuo parametrų funkcija. $\beta(t)$ forma vertinama remiantis Šionfildo likučiais.

- **Stratifikuotas Kokso modelis.** Sakykime, individai pagal tam tikrą rodiklį (ligoninę, kurioje gydėsi, susirgimo rūšį ir pan.) padalyti į k stratų (grupių). Daroma prielaida, kad proporcingos rizikos modelis galioja kiekvienam stratui su individualia bazine rizikos funkcija $h_{0j}(t)$ ir bendru regresijos koeficientu β. Taigi j-tojo strato individo su kovariatės reikšme x rizikos funkcija lygi:

$$h(t; x) = h_{0j}(t)\exp(\beta x), j = 1 \ldots k.$$

Statistiniais paketais (SAS, SPSS) apskaičiuojami β ir bazinių rizikos funkcijų neparametriniai įverčiai.

- **Konkuruojančios rizikos modelis** (*Competing risk model*). Iki šiol nagrinėdami išgyvenamumo modelius laikėme, kad baigties įvykis apibūdinamas viena reikšme – mirė, susirgo ir pan. Tačiau baigties įvykis gali įgyti ir daugiau reikšmių. Pavyzdžiui, mirti galima dėl širdies ligos, vėžio ar kitų priežasčių (13.15 pav.). Šiuo atveju turime 3 baigties įvykio reikšmes. Šis išgyvenamumo modelis vadinamas konkuruojančios rizikos modeliu.



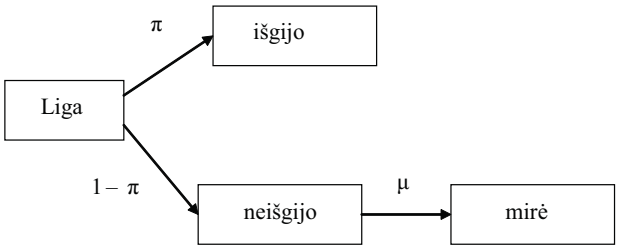
13.15 pav. Konkuruojančios rizikos modelio pavyzdys

Esant keliems, sakykime, m baigties įvykio variantams, išgyvenamumo duomenų struktūra tokia: fiksuojamas i-tojo individo gyvavimo ar stebėjimo

laikas T_i bei baigties taško indikatoriai $D_{i1} \ldots D_{im}$. Iš šių indikatorių 1 gali būti lygus tik vienas – priklausomai nuo baigties taško reikšmės (pvz., mirties priežasties). Jei stebėjimo metu baigties taškas individui nefiksuojamas, visi D_{ij} yra lygūs 0. Ši informacija gali būti pateikta ir 2 kintamaisiais (T_i, S_i); čia T_i – išgyvenimo ar stebėjimo laikas, S_i – baigties taško indikatorius, sakykime, $S_i = j$, jei $D_{ij} = 1$, ir $S_i = 0$, jei visi $D_{ij} = 0$.

Analizuojant tokio pobūdžio duomenis, vertinama išgyvenamumo ir rizikos funkcija j-tajai baigties įvykio reikšmei, sudaromi šių funkcijų regresiniai modeliai ir kt.

Išgijimo modelis (*cure model*). Sakykime, analizuojamas jaunų individų, susirgusiųjų tam tikra liga, išgyvenamumas. Dalis individų „išgyja“ – jų gyvenimo laikas, galima sakyti, sąlygojamas amžiaus, panašiai kaip sveikų individų. Likusių individų, pavyzdžiui, sergančių leukemija ir nepasveikusių, gyvenimo laikas gerokai trumpesnis nei „išgijusiųjų“. Gana ilgai stebint šią jauną populiaciją, praktiškai fiksuosime tik „neišgijusių“ individų baigties taškus, o išgijusių individų baigties taškai paaiškės tik jiems sulaukus brandaus amžiaus. Minėtos populiacijos išgyvenamumo funkcija, didėjant t, į nulį neartėja – ji artėja prie tam tikro teigiamo skaičiaus π; čia π interpretuojamas kaip „išgijusiųjų“ proporcija populiacijoje arba tikimybė išgyti (13.16 pav.). Parametrinės išgyvenamumo funkcijos, pavyzdžiui, eksponentinė, Veibulo, artėja į nulį, augant t. Todėl populiacijos su išgijimu išgyvenamumui modeliuoti naudojami pakoreguoti modeliai.



13.16 pav. Išgijimo modelis. Populiacijos išgyvenamumo funkcija $S_c(t)$ vertinama modeliu $S_c(t) = \pi + (1 - \pi)exp(-\mu t)$, π – tikimybė išgyti, μ – mirtingumo rizika neišgijusiems

Mišinio modelis (*mixture model*) – daroma prielaida, kad tiriamą populiaciją sudaro dvi subpopuliacijos: „išgijusių“ ir „neišgijusių“. Išgijimo tikimybė lygi π, neišgijusių individų išgyvenamumo funkcija lygi $W(t)$. Tiriamos populiacijos išgyvenamumo funkcija $S_c(t)$ vertinama tokiu modeliu:

$$S_c(t) = \pi + (1 - \pi)W(t). \qquad (13.13)$$

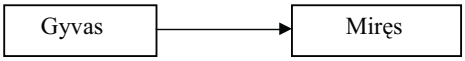
Kai $t = 0$, $S_c(t) = \pi + (1 - \pi)W(0) = 1$. Augant t, $W(t) \to 0$, ir $S_c(t) \to \pi$. $W(t)$ modeliuojama eksponentine, Veibulo ar kita parametrine funkcija. Taip pat sudaromi regresiniai 13.13 modeliai – π vertinama pagal individo sveikatos būklės rodiklius (kovariates) $x_1, x_2 \ldots$. Pavyzdžiui, π priklausomybė nuo kovariačių vertinama logistine funkcija: $\pi = 1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2))$; čia $\beta_0, \beta_1, \beta_2$ – regresinio modelio parametrai.

Ne mišinio modelis (*non-mixture model*):

$$S_c(t) = \pi^{(1-W(t))};$$

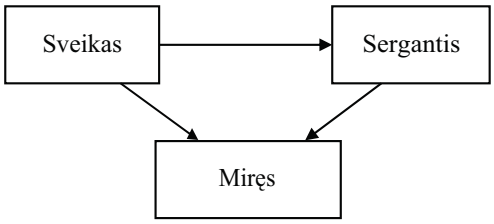
čia $0 < \pi < 1$ – išgijimo tikimybė, $W(t)$ – parametrinė išgyvenamumo funkcija. Kai $t = 0$, $W(0) = 1$ ir $S_c(t) = \pi^0 = 1$. Augant t, $W(t) \to 0$, $S_c(t) \to \pi$. Šiame modelyje π taip pat laikoma kovariačių funkcija.

Kelių būsenų modeliai (*multi-state models*). Individo gyvenime vyksta pokyčiai; kai kurie jų susiję su sveikatos būkle: atsiradusi liga, operacija, pooperacinė komplikacija ir t. t. Po kai kurių įvykių, pavyzdžiui, susirgus arterine hipertenzija, diabetu, miokardo infarktu, organizme atsiranda kokybinių pokyčių – galima sakyti, kad šiems įvykiams vykstant organizmas pereina iš vienos būsenos į kitą. Permainoms modeliuoti naudojami kelių būsenų modeliai. Jie vaizduojami grafu, rodyklėmis nurodant galimus perėjimus iš vienos būsenos į kitą. 13.1–13.10 skyriuose nagrinėtas dviejų būsenų modelis:

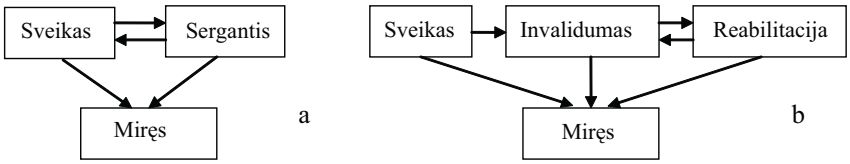


Pateiksime keletą kelių būsenų modelių pavyzdžių.

Sergamumo–mirties modelis (*ilness-death model*). Individas charakterizuojamas 3 būsenomis: sveikas, sergantis, miręs (13.17 pav.): galimas perėjimas iš vienos būsenos į kitą nurodytas rodyklėmis. Taip pat naudojami sergamumo–mirties modeliai su atsistatymu (*reactivation*) (13.18 a pav.) ir su atsistatymu (reabilitacija) kaip atskira būsena (13.18 b pav.).

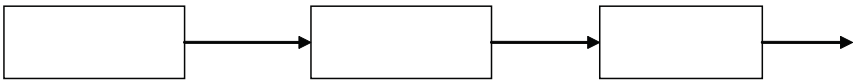


13.17 pav. Sergamumo–mirties modelis



13.18 pav. Sergamumo–mirties modeliai su atsistatymu (a); su reabilitacija kaip atskira būsena (b)

Pasikartojančių įvykių modelis (*recurrent events model*). Šiame modelyje perėjimas iš vienos būsenos į kitą vyks taip:



Tokiu modeliu gali būti iliustruojamas sergančiųjų cukriniu diabetu albinumo baltymo kitimas: norma → mikroalbinurija → makroalbinurija.

Kelių būsenų modeliai aprašomi atsitiktiniais procesais. Naudojami modeliai, kai laikas t yra diskretusis ir tolydusis. Perėjimas iš vienos būsenos į kitą nusakomas perėjimo intensyvumu (perėjimo rizika) – tikimybe, kad iki t momento esant i būsenoje, momentu t įvyks perėjimas į būseną j. Pavyzdžiui, 13.17 pav. pateiktą modelį apibūdina funkcijos: $\lambda(t)$ – rizika numirti sveikam, $\varphi(t)$ – rizika susirgti sveikam ir $\mu(t, t_1)$ – rizika numirti sergančiam, kai susirgta momentu t_1. Funkcijos λ, φ ir μ gali būti pateiktos parametriniu pavidalu; nežinomi parametrai vertinami didžiausio tikėtinumo metodu. Taip pat 13.17 pav. modelis pertvarkomas į modelį su kintama kovariate ([3]).

13 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Bagdonavičius V., Nikulin M. *Accelerated Life Models. Modeling and Statistical Analysis*. 2001. Capman & Hal, 334 p.
3. Hoogaard P. *Analysis of Multivariate Survival Data*. 2000. Berlin: Springer – Verlag, 542 p.
4. Kleinbaum D. G., Klein M. *Survival Analysis*. A self-learning text. Second ed. 2005. Berlin: Springer, 590 p.
5. Machin D., Cheng Y. B., Parmar M. K. B. *Survival Analysis. A Practical Approach*. Second ed. 2006. Wiley, 266 p.
6. Therneau T. M., Grambsch P. M. *Modeling Survival Data: Extending the Cox Model*. 2000. Berlin: Verlag, 345 p.

7. *Išeminės širdies ligos sindromų išeitys ir ligonių išgyvenimas 5–10 metų laikotarpiu*. Kardiologijos institutas, Klinikinės kardiologijos lab. ataskaita. 2005.
8. *Išgyvenamumas: pavyzdžiai, cenzūravimas. Kaplano–Mejerio kreivė, parametriniai modeliai, Kokso modelis, išgyvenamumo funkcijų palyginimas.* Prieiga per internetą: http://yates.coph.usf.edu/~yiliang/surv/index.
9. *Medical Statistics: Survival Analysis.* Course Notes, N. Fieller. 2002. Prieiga per internetą: http://www.shef.ac.uk/nickfieller/tampere/hsurvival.pdf.

14 SKYRIUS

Duomenų surinkimo ir analizės kokybės tyrimas

Analizuojant medikų sukauptą medžiagą, pastebimas didelis duomenų kitimas ir dėl to galimos klaidingos išvados. Tam tikri statistiniai metodai leidžia įvertinti duomenų kitimą, kartais sąlygotą matavimo klaidų, testo kokybę bei palengvina diagnostikos duomenų interpretavimą.

14.1. Klaidos kiekybinių duomenų analizėje

Atlikdami laboratorijoje eksperimentą, kuriuo nustatomos kiekybinio rodiklio reikšmės, susiduriame su dviejų rūšių klaidomis: sistemine ir atsitiktine. Atsitiktinė klaida – tai nuokrypis nuo vidutinės reikšmės, atsirandantis dėl daugelio pašalinių faktorių poveikio, kurį įvardijame atsitiktiniu. Sisteminė klaida daroma tuomet, kai gautų reikšmių vidurkis skiriasi nuo tikrosios reikšmės.

Aiškindamiesi šias sąvokas, analizuokime realią situaciją laboratorijoje. Keturi studentai (A, B, C, D) individualiai atlieka titravimą: jie turi išsiskirti 10 ml tirpalo. Kiekvienas studentas atlieka titravimą 5 kartus. Gauti rezultatai pateikti 14.1 lentelėje.

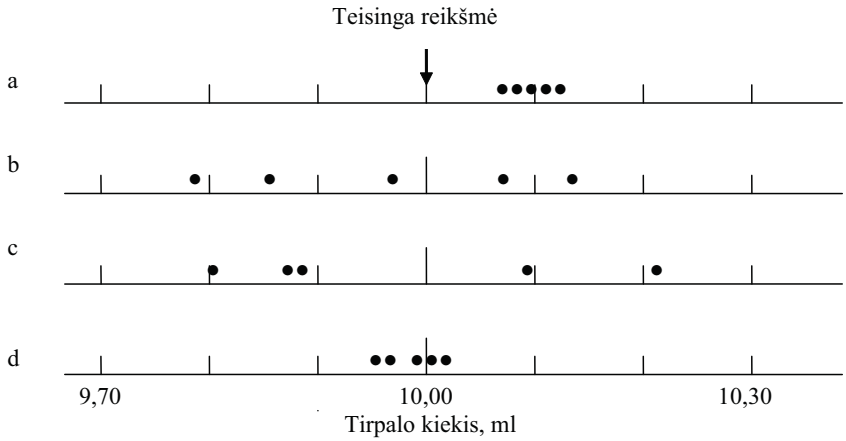
A studento rezultatus galima apibūdinti dviem aspektais. Pirma, gautos reikšmės labai mažai skiriasi viena nuo kitos – jos kinta nuo 10,08 iki 10,12 ml; taigi atkartojimas yra didelis. Antra, visos gautos reikšmės viršija tikrąją reikšmę – 10 ml. Vadinasi, gauta sisteminė paklaida. Matome, kad, atlikdamas eksperimentą, studentas darė dviejų rūšių klaidas. Jo titravimo rezultatai išsibarstę aplink vidurkį – 10,1 ml. Taigi yra atsitiktinės klaidos. Reikšmių nuo vidurkio skirtumas apibūdina eksperimento **patikimumą**,

arba atkartojamumą (*precision, reproducibility, reliability*). A studento atsitiktinės klaidos labai mažos (neviršija 0,02 ml), todėl galima sakyti, kad rezultatai labai patikimi. Tačiau daryta sisteminių klaidų – visi rezultatai viršija tikrą reikšmę. Sisteminės klaidos buvimas ar nebuvimas apibūdinamas eksperimento **tikslumu** (*accuracy*): kiek gautos reikšmės artimos tikrai reikšmei. A studento gauti rezultatai nėra tikslūs. B studento rezultatai prieštaringi A studento rezultatams. Penkių matavimų vidurkis – 10,01 ml – labai nedaug skiriasi nuo tikrosios reikšmės. Taigi gauti rezultatai yra tikslūs. Tačiau nustatytų reikšmių diapazonas – nuo 9,8 ml iki 10,21 ml (14.1 lentelė) – gana platus; atsitiktinės paklaidos svyruoja iki 0,2 ml. Šiuo atveju eksperimento rezultatai nepatikimi. C studento rezultatai nėra nei patikimi (gautos reikšmės svyruoja nuo 9,69 iki 10,19 ml), nei tikslūs (vidurkis 9,9 ml). D studento rezultatai, patikimi (svyruoja nuo 9,97 iki 10,04 ml) ir tikslūs (vidurkis 10,01 ml). Skirtumas tarp tikslumo ir patikimumo pateiktas 14.1–14.2 pav.

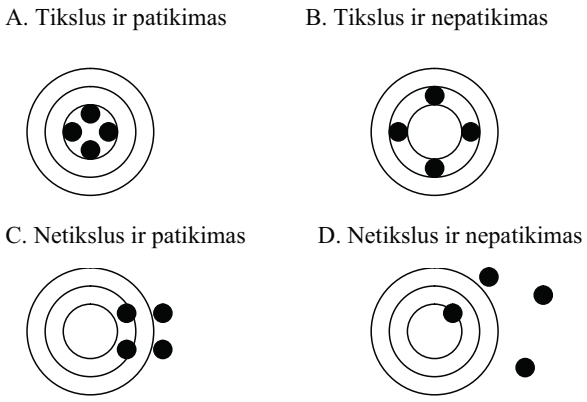
Duomenų patikimumas vertinamas kitimo rodikliais – dispersija, standartiniu nuokrypiu, standartine paklaida, interkvartiliniu pločiu. Atlikto tyrimo tikslumas nustatomas t kriterijumi vienai imčiai (5.3 skyrius) lyginant duomenų vidurkį su tikrąja reikšme (norma).

14.1 lentelė. Atsitiktinės ir sisteminės paklaidos

Studentas	Rezultatas (ml)	Rezultatų apibūdinimas
A	10,08 10,11 10,10 10,12 10,09	patikimi netikslūs
B	9,88 10,14 10,02 9,80 10,21	nepatikimi tikslūs
C	10,19 9,79 9,69 10,05 9,78	nepatikimi netikslūs
D	10,04 9,98 10,02 9,97 10,04	patikimi tikslūs



14.1 pav. Tikslumas ir patikimumas: grafiškai pateikti 14.1 lentelės duomenys. (a) duomenys patikimi, bet netikslūs; (b) nepatikimi, bet tikslūs; (c) nepatikimi ir netikslūs; (d) patikimi ir tikslūs



14.2 pav. Tikslumo ir patikimumo kombinacijos, gautos nustatant kiekybinį kintamąjį

14.2. Diagnozavimo klaidos

Vienas medikams-tyrėjams keliamų uždavinių – sudaryti testus susirgimams diagnozuoti. Būtini ir diagnostinio testo gerumo kriterijai bei kriterijai skirtingais diagnostiniais testais gautiems rezultatams palyginti. Diagnostinio testo gerumas turi būti vertinamas atsižvelgiant į jo taikymo rezultatus tiek sergantiems, tiek nesergantiems asmenims.

Šiame skyriuje terminas „tikimybė“ suprantamas kaip atitinkamos tikimybės ar sąlyginės tikimybės įvertis.

Taikant medicininį testą susirgimui diagnozuoti, įmanoma padaryti dviejų rūšių klaidas. Galima diagnozuoti asmeniui susirgimą (teigiamas testo rezultatas), nors iš tikrųjų jis neserga. Tai I rūšies, arba alfa, klaida. Ši klaida dar vadinama neteisingai teigiama klaida (*false-positive error*). Asmeniui galima nediagnozuoti susirgimo (neigiamas testo rezultatas), nors jis iš tikrųjų serga – tai II rūšies, arba beta, klaida. Ši klaida dar vadinama neteisingai neigiama klaida (*false-negative error*). Diagnostinio testo teigiamas (diagnozuojantis susirgimą) rezultatas, kai asmuo neserga, vadinamas neteisingai teigiamu rezultatu (*false-positive result*). Testo neigiamas (atmetantis susirgimą) rezultatas, kai asmuo iš tikrųjų serga, vadinamas neteisingai neigiamu rezultatu (*false-negative result*).

Diagnostinio testo rezultatus dažnai veikia susirgimo fazė, asmens biologinės savybės. Pavyzdžiui, TBC testas yra tuberkulino bakterijų odos mėginys. Laikoma, kad šis testas duoda teigiamą rezultatą (įtariama TBC), jei paraudimo spindulys viršija 5 mm. Šis testas daug tikslesnis viduriniuoju TBC periodu nei akstyvuoju ar vėlyvuoju. Pavyzdžiui, ankstyvuoju TBC periodu nespėja pasigaminti tuberkulino antigenų, todėl šiuo laiku dažniau pasitaiko neteisingai neigiami rezultatai. Odos mėginys TBC diagnozuoti gali duoti neteisingai teigiamą rezultatą, jei asmuo jautrus mėginio bakterijoms.

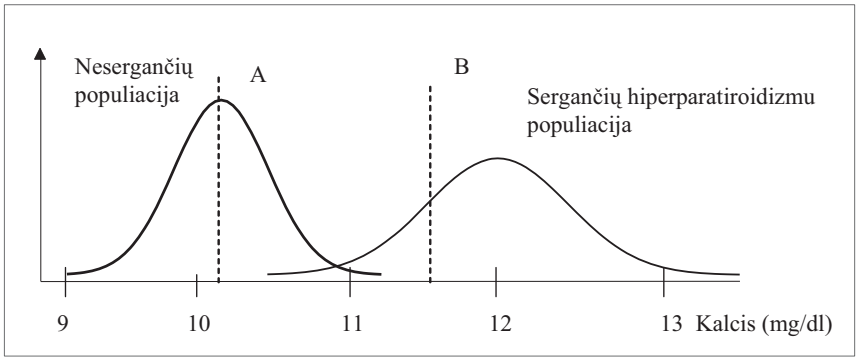
Sumažinus paraudimo skersmens ribą, kurią viršijus nustatomas teigiamas testo rezultatas, sakykime, nuo 5 mm iki 3 mm, padidėtų neteisingai teigiamų rezultatų, bet sumažėtų neteisingai neigiamų. Atvirkščiai, išplėtus paraudimo skersmens ribą už 5 mm, sumažėtų neteisingai teigiamų rezultatų, bet padidėtų neteisingai neigiamų.

14.3. Testo sudarymas. Kritinė reikšmė

Neteisingai teigiamų ir neteisingai neigiamų rezultatų pasitaiko visuose testuose, kuriuose naudojamos kiekybinio kintamojo reikšmės. Tai iliustruosime pavyzdžiu apie kalcio koncentracijos kraujyje testą, skirtą hiperparatiroidizmui (HPT) nustatyti.

Hiperparatiroidizmas – kalcio metabolizmo susirgimas. Šia liga sergančių asmenų kraujyje padidėjusi kalcio koncentracija. Žinoma, sergančių HPT kalcio kiekis kraujyje priklauso ir nuo individualių organizmo savybių, todėl jį galima laikyti atsitiktiniu dydžiu, turinčiu tam tikrą vidurkį ir dispersi-

ją (14.3 pav.). Normali kalcio koncentracija svyruoja nuo 8,5 iki 10,5 mg/dl. Be abejo, ji gali padidėti ir dėl kitų susirgimų. Asmenų, nesergančių HPT, kalcio koncentracija kraujyje taip pat yra ats. dydis su vidurkiu, mažesniu nei sergančių HPT kalcio koncentracijos vidurkis (14.3 pav.).



14.3 pav. Asmenų, nesergančių ir sergančių HPT, kalcio koncentracijos kraujyje skirstinys (mg/dl)

Jei kalcio koncentracija labai maža (žemiau už A ribą, 14.3 pav.), beveik neįmanoma, kad pacientas sirgs HPT. Jei kalcio koncentracija labai didelė (pvz., virš taško B, 14.3 pav.), visiškai tikėtina, kad asmens kalcio metabolizmas nenormalus, jis serga būtent HPT. Tačiau jei kalcio koncentracija yra tarp taškų A ir B, tikėtina, kad hiperparatiroidizmas gali ir būti, ir nebūti. Kuo kalcio koncentracija arčiau taško B, tuo labiau tikėtinas HPT; kuo arčiau taško A, tuo labiau tikėtina, kad šio susirgimo nėra.

Testas HPT diagnozuoti turėtų būti toks: jei kalcio koncentracija viršija *a* reikšmę, testo rezultatas laikomas teigiamu, priešingu atveju – neigiamu. Taškas *a*, naudojamas testui apibrėžti, vadinamas kritiniu tašku (*cutoff point*). I ir II rūšies klaidų tikimybė priklauso nuo taško *a* parinkimo (žr. 5.3 skyriuje).

Daugelyje laboratorijų nustatyta normali kalcio koncentracija – iki 11 mg/dl. Todėl kalcio testas hiperparatiroidizmui diagnozuoti rekomenduojamas toks: jei kalcio koncentracija viršija 11 mg/dl, testas teigiamas, priešingu atveju – neigiamas.

Sudaryti testą hiperparatiroidizmui ar kitam susirgimui diagnozuoti yra diskriminantinės analizės uždavinys. Taškas *a* gali būti nustatomas diskriminantinės analizės metodais, įvedus nuostolių, gautų dėl neteisingo diagnozavimo, funkcijas (15.5 skyrius).

14.4. Jautrumas ir specifiškumas

Kaip minėta 14.3 skyriuje, parinkus skirtingus kritinius taškus, galimi skirtingi susirgimo diagnostikos testai. Diagnostiniams testams palyginti naudojamos jautrumo, specifiškumo ir prognostinio dydžio sąvokos. Pagal šiuos dydžius parenkamas testui optimalus kritinis taškas. Jautrumas (*sensivity*), specifiškumas (*specifity*) ir prognostinis dydis (*predictive value*) taip pat padeda palyginti testus, atliekamus skirtingais tyrimais (pvz., radiologiniu, biocheminiu, echoskopiniu ir kt.). Minėtos charakteristikos gali būti naudojamos ir diskriminantinėje analizėje klasifikavimo į 2 klases kokybei vertinti.

Testo taikymo rezultatai pateikiami 2×2 lentele (14.2 lentelė). Pirmas stulpelis skirtas sergantiems individams, iš jų *a* – nustatytas teisingas teigiamas rezultatas, *c* – neteisingai neigiamas. Antras stulpelis skirtas nesergantiems individams, iš jų *b* – gautas neteisingai teigiamas rezultatas, *d* – teisingai neigiamas rezultatas. Dydis $(a + c)/(a + b + c + d)$ vadinamas studijos sergamumu (*prevalence*).

Jautrumas $a/(a + c)$ parodo testo gebėjimą diagnozuoti susirgimą, jei asmuo iš tikrųjų serga; tai tikimybė, kad sergančio asmens testo rezultatas yra teigiamas. Jautrumas menkas, jei *a* yra gerokai mažesnis už *c*. Neteisingai neigiama klaida (*false-negative error rate*) lygi $c/(a + c)$ – tai tikimybė, kad sergančio asmens diagnozė neteisinga.

Specifiškumas $d/(b + d)$ parodo testo gebėjimą nustatyti, jog susirgimo nėra, jeigu jo iš tikrųjų nėra; tai tikimybė (tikimybės įvertis), kad nesergančiam asmeniui bus nustatyta teisinga diagnozė. Testas nėra specifinis, jei *d* mažesnis už *b*. Dydis $b/(b + d)$ yra neteisingai teigiama klaida (*false-positive error rate*) – tai tikimybė, kad nesergančiam individui gali būti neteisingai nustatyta diagnozė.

14.2 lentelė. Standartinė 2×2 porinė dažnių lentelė, sudaroma lyginant testo rezultatus su faktiniais duomenimis apie sergamumą

		SERGAMUMAS		Iš viso
		Serga	Neserga	
TESTO REZULTATAS	Teigiamas	a	b	a + b
	Neigiamas	c	d	c + d
	Iš viso	*a + c*	*b + d*	*a + b + c + d*

Lentelės paaiškinimai:
a = sergantys su teisingai teigiamu testo rezultatu;
b = nesergantys su neteisingai teigiamu testo rezultatu;
c = sergantys su neteisingai neigiamu testo rezultatu;
d = nesergantys su teisingai neigiamu testo rezultatu;

$a + b$ = visų individų teigiamas testo rezultatas;
$c + d$ = visų individų neigiamas testo rezultatas;
$a + c$ = visi sergantys individai;
$b + d$ = visi nesergantys individai;
$a + b + c + d$ = visi studijoje dalyvavę individai;
$a/(a + c)$ = jautrumas;
$d/(b + d)$ = specifiškumas;
$b/(b + d)$ = neteisingai teigiama klaida (alfa klaida, I rūšies klaida);
$c/(a + c)$ = neteisingai neigiama klaida (beta klaida, II rūšies klaida);
$a/(a + b)$ = teigiamas prognostinis dydis;
$d/(c + d)$ = neigiamas prognostinis dydis;
$[a/(a + c)]/[b/(b + d)]$ = teigiamas tikėtinumo santykis (LR_+);
$[c/(a + c)]/[d/(b + d)]$ = neigiamas tikėtinumo santykis (LR_-);
$(a + c)/(a + b + c + d)$ = sergamumas.

Jautrumas, specifiškumas bei neteisingai teigiama ir neteisingai neigiama klaida nepriklauso nuo sergamumo (jie yra sąlyginės tikimybės). Šie dydžiai taip pat pateikiami procentais.

Jautrumo ir specifiškumo skaičiavimo pavyzdys pateiktas 14.3 lentelėje [4]. 80 asmenų atliktas kalcio testas hiperparatiroidizmui diagnozuoti. Jei kalcio koncentracija viršija 11 mg/dl, testo rezultatas – teigiamas, priešingu atveju – neigiamas. Iš 80 asmenų 20 sirgo HPT. Iš šių 20 asmenų testu 12 buvo diagnozuotas hiperparatiroidizmas. Todėl testo jautrumas yra 12/20 = 60 %, neteisingai neigiama klaida 8/20 = 40 %. Iš 60 nesergančių asmenų 57 buvo nustatyti neigiami testo rezultatai, taigi specifiškumas – 95 % (57/60). Neteisingai teigiama klaida nustatyta 5 % nesergančių (3/60) asmenų.

14.3 lentelė. Kalcio testo rezultatai sergamumui hiperparatiroidizmu diagnozuoti

		SERGAMUMAS		Iš viso
		Serga	Neserga	
Kalcio koncentracija	Didelė	12	3	15
	Maža	8	57	65
	Iš viso	20	60	80

Skaičiavimai:
12/20 = 65 % = jautrumas;
57/60 = 95 % = specifiškumas;
3/60 = 5 % = neteisingai teigiama klaida (alfa klaida, I rūšies klaida);
8/20 = 40 % = neteisingai neigiama klaida (beta klaida, II rūšies klaida);
12/15 = 80 % = teigiamas prognostinis dydis;
57/65 = 88 % = neigiamas prognostinis dydis;
[12/20]/[3/60] =12,0 = teigiamas tikėtinumo santykis (LR_+);
[8/20]/[57/60] = 0,42 = neigiamas tikėtinumo santykis (LR_-);
20/80 = 25 % = sergamumas.

Jautrumas ir specifiškumas neatsako į du klinicistui svarbius klausimus: jei ligonio testo rezultatas teigiamas, kokia tikimybė, kad jis tikrai serga? Ir analogiškai – jei ligonio testo rezultatai neigiami, kokia tikimybė, kad jis iš tikrųjų neserga?

Jautrumas ir specifiškumas į šiuos klausimus atsakymo nepateikia. Todėl skaičiuojamas teigiamas prognostinis dydis (*positive predictive value*) $a/(a + b)$. Tai sąlyginė tikimybė, kad asmuo, kurio testo rezultatas teigiamas, iš tikrųjų serga. Analogiškai skaičiuojamas neigiamas prognostinis dydis (*negative predictive value*) $d/(c + d)$ – tikimybė, kad asmuo, kurio testo rezultatas neigiamas, iš tikrųjų neserga.

14.3 lentelėje teigiamas prognostinis dydis lygus 80 % (12/15), neigiamas prognostinis dydis yra 88 % (57/65).

Testo kokybei vertinti taip pat naudojami tikėtinumo santykiai. Teigiamas tikėtinumo santykis LR_+ (*likelihood ratio positive*) – jautrumo ir neteisingai teigiamos klaidos (1 – specifiškumas) santykis: LR_+ = jautrumas/(1 – specifiškumas) = $[a/(a + c)]/[b/(b + d)]$. LR_+ dažnai interpretuojamas taip: LR_+ yra santykis to, ko klinicistas nori iš testo (jautrumo), su tuo, ko klinicistas nenori (neteisingai teigiamos klaidos). Kuo LR_+ didesnis, tuo testas geresnis. LR_+ nuo sergamumo nepriklauso. Gero testo LR_+ turi būti didesnis už 1.

Analogiškai apibrėžiamas neigiamas tikėtinumo santykis LR_- (*likelihood ratio negative*). Jis lygus neteisingai neigiamos klaidos (1 – jautrumas) ir specifiškumo santykiui: LR_- = (1 – jautrumas)/specifiškumas = $[c/(a + c)]/[d/(b + d)]$. Todėl sakoma, kad LR_- yra santykis to, ko klinicistas nenori, su tuo, ko klinicistas nori. Kuo LR_- artimesnis 0, tuo testas geresnis.

Remiantis 14.3 lentelėje pateiktais duomenimis, teigiamas testo tikėtinumo santykis yra 12 ([12/20]/[3/60]), neigiamas tikėtinumo santykis – 0,42 ([8/20]/[57/60]).

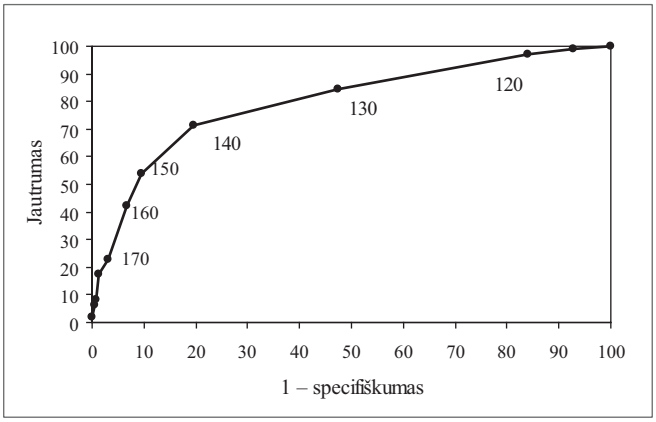
14.5. ROC (*Received Operating Characteristic*) kreivė

Sudarant susirgimo testą pagal kiekybinio rodiklio (kalcio, gliukozės koncentracijos ir kt.) reikšmes, optimalų kritinį tašką reikia parinkti taip, kad tiek jautrumas, tiek specifiškumas būtų kuo didesni. Pavyzdžiui, kalcio kiekio testu diagnozuojant HPT: didinant kritinę reikšmę, mažėja jautrumas, bet didėja specifiškumas; atvirkščiai, mažinant kritinę reikšmę, didėja jautrumas, bet mažėja specifiškumas. Todėl, parenkant optimalų kritinį tašką, būtina turėti kelių galimų kritinių taškų jautrumo ir specifiškumo rodiklius.

Kritinio taško „gerumui“ vertinti ir kelių rodiklių diagnozavimo kokybei palyginti naudojama ROC kreivė. Ji konstruojama taip: kiekviena rodiklio reikšmė x_i naudojama kaip kritinis taškas ir procentais skaičiuojamas jo jautrumas j_i bei specifiškumas s_i. Po to plokštumoje kiekvienam i dedamas taškas su koordinatėmis: Y ašyje atidedamas jautrumas j_i, X ašyje atidedamas 1 – specifiškumas ($1 - s_i$). Visi gretimi taškai sujungiami. Gauta kreivė vadinama rodiklio ROC kreive.

14.2 pavyzdys. 14.4 pav. pateikta ROC kreivė, apibūdinanti ligonių, persirgusių MI, vieno SAS matavimo tinkamumą arterinei hipertenzijai (AH) diagnozuoti. ROC kreivė sudaryta naudojant 958 ligonių, persirgusių MI, duomenis; 739 ligoniai sirgo AH. SAS matuotas 10 mmHg tikslumu; reikšmės kito nuo 100 iki 210.

Pagal vieną SAS matavimą ligoniui diagnozuojama AH, jei SAS reikšmė lygi arba didesnė už kritinę reikšmę SAS_0: t. y. kai $SAS \geq SAS_0$. AH dažnis, jautrumas bei specifiškumas įvairioms SAS_0 reikšmėms pateiktas 14.4 lentelėje. Iš jos matyti, kad, pasirinkus $SAS_0 = 100$, visiems 958 ligoniams diagnozuojama AH; t. y. testo jautrumas – 100 %, specifiškumas – 0 %, 1 – specifiškumas lygus 100 %. Šią kritinę reikšmę ROC kreivėje (14.4 pav.) atitinka taškas su koordinatėmis (100, 100). Iš 14.4 lentelės matyti, kad 311 (42,1 %) sergančių AH ligonių SAS buvo 160 mmHg ir didesnis, 204 (93,2 %) nesergančių AH SAS buvo mažesnis nei 160 mmHg. Todėl, parinkę kritine reikšme $SAS_0 = 160$, gauname jautrumą 311/739 = 42,1 % ir 1 – specifiškumą = 1 – 207/219 = 6,8 %. Šią kritinę reikšmę ROC kreivėje atitinka taškas (6,8; 42,1). Esant kritinei reikšmei $SAS_0 = 210$, jautrumas lygus 2 %, o 1 – specifiškumas = 0 %. Šią kritinę reikšmę ROC kreivėje atitinka taškas (0; 2).



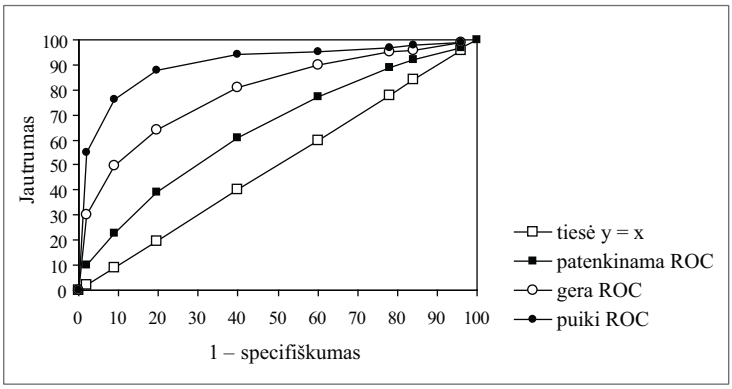
14.4 pav. Vieno SAS matavimo tinkamumą arterinei hipertenzijai diagnozuoti atspindinti ROC kreivė

14.4 lentelė. AH dažnis, jautrumas bei specifiškumas įvairioms kritinėms SAS reikšmėms

Kritinė reikšmė	AH yra	AH nėra	Jautrumas, %	Specifiškumas, %	1 – specifiškumas (%)
100	8	16	739/739=100	0/219=0	100
110	15	19	731/739=98,9	16/219=7,3	92,7
120	91	80	716/739=96,9	35/219=16	84
130	97	61	625/739=84,6	115/219=52,5	47,5
140	130	22	528/739=71,4	176/219=80,4	19,6
150	87	6	398/739=53,9	198/219=90,4	9,6
160	144	8	311/739=42,1	204/219=93,2	6,8
170	38	4	167/739=22,6	212/219=96,8	3,2
180	68	1	129/739=17,5	216/219=98,6	1,4
190	14	1	61/739=8,3	217/219=99,1	0,9
200	32	1	47/739=6,4	218/219=99,5	0,5
210	15	0	15/739=2	219/219=100	0
Iš viso	739	219	–	–	–

ROC kreivė naudojama ir kiekybinio rodiklio diagnozavimo kokybei vertinti. Diagnostinis testas tuo geresnis, kuo didesnis jautrumas ir specifiškumas (arba mažesnis 1 – specifiškumas). Esant tam pačiam specifiškumui, geresnis testas bus jautresnis už prastesnį testą – t. y. geresne testavimo geba pasižyminčio rodiklio ROC kreivė bus virš prastesne testavimo geba pasižyminčios ROC kreivės (14.5 pav.). Todėl rodiklio testavimo kokybė vertinama plotu po ROC kreive – vadinamąja statistika *C*. Kuo šis plotas artimesnis 1, tuo rodiklis tinkamesnis diagnozei. Jei rodiklis testuoti netinkamas, tuomet bet kurio kritinio taško jautrumas ir specifiškumas bus artimi 50 %. Tokio rodiklio ROC kreivė artima tiesei $y = x$ (15.5 pav.). Jei statistika *C* artima 0,5 (plotui po tiese $y = x$, $0 \le x \le 1$), rodiklis diagnozei netinkamas. Todėl, analizuojant rodiklių testavimo gebą, vertinamas statistikos C patikimumas – pasikliautinieji intervalai arba tikimybė, kad statistika *C* viršija 0,5.

Plotas po 14.4 pav. ROC kreive lygus 0,799; 95 % pasikliautinasis intervalas [0,768; 0,831]. Plotas po 14.5 pav. pateikta puikia ROC kreive lygus 0,93, po gera ROC kreive – apie 0,8, po patenkinama ROC kreive – apie 0,66.



14.5 pav. Įvairia testavimo geba pasižyminčių rodiklių ROC kreivės

14.6. Patikimumo ir suderinamumo (*agreement*) vertinimas

Atliekant tiek klinikinius, tiek kitus tyrimus (pvz., fiksuojant širdies ūžesius, karkalų plaučiuose buvimą ir kt.) aktualu įvertinti tyrimo patikimumą, t. y. ar visi tyrėjai nustato tą patį. Jei tyrimo metu nustatoma dvinario kintamojo (serga, neserga; testo rezultatas teigiamas, neigiamas) reikšmė, tai dviejų tyrėjų rezultatus, gautus ištyrus tuos pačius ligonius, galima pateikti 2×2 porine dažnių lentele (14.5 lentelė). Joje skaičiai *a* ir *d* rodo, kad rezultatai sutampa (abu tyrėjai gavo vienodus rezultatus), o skaičiai *b* ir *c* rodo, kad rezultatai nesutampa. Vienas testo rezultatų atitikimo matų – sutapimų procentas $100(a + d)/(a+b+c+d)$. Tačiau sutapimų procentas:

- nėra susijęs su dvinario kintamojo reikšmių dažniu (pvz., sergamumu ar testo teigiamo rezultato tikimybe);
- neparodo abiejų tyrėjų rezultatų nesuderinamumo, t. y. ar abiejų tyrėjų teigiamo ir neigiamo testo rezultato vertinimo tikimybės vienodos;
- gautas sutapimų procentas nerodo skirtumo tarp atsitiktinio sutapimų procento, t. y. sutapimų procento, įvertinto darant prielaidą, kad abu tyrėjai testo tyrimą atlieka nepriklausomai.

Dviejų tyrėjų rezultatų atitikčiai vertinti naudojamas koeficientas κ (*kapa*). *Kapa* palygina faktinį sutapimų dažnį su atsitiktiniu (tikėtinu) sutapimų dažniu, įvertintu darant prielaidą, kad abu tyrėjai vertina nepriklausomai. Tikėtini teigiamų ir neigiamų rezultatų sutapimų dažniai skaičiuojami taip pat, kaip ir tikėtini dažniai porinėje dažnių lentelėje (7.3 skyrius). I ir II tyrėjo teigiamo testo vertinimo tikimybių įverčiai (7.2 skyrius) atitinkamai lygūs $(a+c)/N$ ir $(a+b)/N$ (14.5 lentelė). Jei tyrėjai vertina individus nepriklausomai, tada tikimybė (tiksliau, tikimybės įvertis), kad abiejų tyrėjų testo

rezultatai teigiami, lygi $(a + c)(a + b)/N^2$, tikėtinas teigiamo testo rezultatų sutapimo dažnis lygus $(a + c)(a + b)/N$. Analogiškai nustatoma, kad neigiamo testo rezultatų tikėtinas sutapimo dažnis lygus $(c + d)(b + d)/N$. Sudėjus šiuos dažnius, nustatomas bendras tikėtinas testo rezultatų sutapimų skaičius A_c (14.5 lentelė).

14.5 lentelė. Standartinė 2×2 porinė dažnių lentelė, sudaroma lyginant dviejų tyrėjų testo rezultatus

		I tyrėjas		
		Teigiamas	Neigiamas	Iš viso
II tyrėjas	Teigiamas	a	b	a + b
	Neigiamas	c	d	c + d
	Iš viso	$a + c$	$b + d$	$N = a + b + c + d$

Lentelės paaiškinimai:
a = abiejų tyrėjų testo rezultatas teigiamas;
b = I tyrėjo testo rezultatas neigiamas, II teigiamas;
c = I tyrėjo testo rezultatas teigiamas, II neigiamas;
d = abiejų tyrėjų testo rezultatas neigiamas;

$a + d$ = sutapimų skaičius (A_0);
$a + b + c + d$ = didžiausias galimas sutapimų skaičius (N);
$(a + d)/(a + b + c + d)$ = sutapimų procentas;
$[(a + b)(a + c)]/(a + b + c + d)$ = tikėtinas (atsitiktinis) teigiamų testo rezultatų sutapimų skaičius;
$[(c + d)(b + d)]/(a + b + c + d)$ = tikėtinas (atsitiktinis) neigiamų testo rezultatų sutapimų skaičius;
$[(a + b)(a + c) + (c + d)(b + d)]/(a + b + c + d)$ = tikėtinas testo rezultatų sutapimų skaičius (A_c);
$(A_0 - A_c)/(N - A_c) = \kappa$ (*kapa*).

Faktinį sutapimų skaičių $a + d$ pažymėkime A_0. Skirtumas tarp faktinio ir atsitiktinai gaunamo sutapimų skaičiaus lygus $A_0 - A_c$. Jis parodo, kiek padaugėjo sutapimų, palyginti su atsitiktinai gaunamu sutapimų skaičiumi. Didžiausias galimas skirtumas tarp faktinio ir atsitiktinio sutapimų skaičiaus lygus $N - A_c$. Koeficientas *kapa* lygus

$$\kappa = \frac{A_0 - A_c}{N - A_c}.$$

Jis lygina sutapimų skaičiaus padidėjimą su maksimaliai galimu. *Kapa* kinta nuo –1 (rezultatai visiškai nesutampa) iki 1 (rezultatai visiškai sutampa). Jei κ artimas 0, sutapimų skaičius nedaug skiriasi nuo atsitiktinio. *Kapa*

dažniausiai pateikiamas procentais. Pagal κ reikšmę rezultatų sutapimas (tyrimo patikimumas) vertinamas taip: jei *kapa* neviršija 20 % – sutapimas nereikšmingas; nuo 20 % iki 40 % – sutapimas minimalus; nuo 40 % iki 60 % – sutapimas pakankamas; nuo 60 % iki 80 % – sutapimas geras; daugiau kaip 80 % – sutapimas puikus.

Pateiksime *kapa* skaičiavimo pavyzdį.

14.3 pavyzdys [4, 96 p.]. Du kardiologai tą pačią dieną tyrė 100 pacientų širdies veiklą ir nustatinėjo, ar yra širdies ūžesys. 7 pacientams I kardiologas nustatė, kad ūžesio nėra, II kardiologas – kad ūžesys yra. 3 pacientams I kardiologas nustatė, kad ūžesys yra, II kardiologas – kad ūžesio nėra. 30 pacientų abu kardiologai konstatavo ūžesio buvimą, o 60 – ūžesio nebuvimą (14.6 lentelė).

Iš 14.6 lentelės matyti, kad abiejų kardiologų tyrimo rezultatai sutampa 90 % pacientų. Atsitiktinis sutapimų skaičius turėtų būti apie 54. Koeficientas *kapa* lygus 78 %. Taigi galime tvirtinti, kad abiejų kardiologų širdies ūžesio nustatymo sutapimas yra geras.

14.6 lentelė. Dviejų kardiologų širdies ūžesio nustatymo rezultatai

		I kardiologas		
		Ūžesys yra	Ūžesio nėra	Iš viso
II kardiologas	Ūžesys yra	30	7	37
	Ūžesio nėra	3	60	63
	Iš viso	33	67	100

Skaičiavimai:
30 + 60 = 90 – sutapimų skaičius (A_0);
30 + 7 + 3 + 60 = 100 – didžiausias galimas sutapimų skaičius (N);
90/100 – 90 % = sutapimų procentas;
[(30 + 7)(30 + 3)]/100 = 37×33/100 = 12,2 – tikėtinas teigiamų testo rezultatų sutapimų skaičius;
[(3 + 60)(7 + 60)]/100 = 63×67/100 = 42,2 – tikėtinas neigiamų testo rezultatų sutapimų skaičius;
12,2 + 42,2 = 54,4 – tikėtinas testo rezultatų sutapimų skaičius (A_c);
(90 – 54,4)/(100 – 54,4) = 35,6/45 = 0,78 = 78 % κ (*kapa*).

14.7. Kiekybinio rodiklio imties dydžio nustatymas

Naujo vaisto arba gydymo metodo efektyvumui tirti rengiamos įvairios ligonių studijos. Organizuojant ligonių studiją, būtina pagrįstai įvertinti jos apimtį, pavyzdžiui, nustatyti, kiek ligonių turi būti placebo ar poveikio (*treatment*) grupėje. Studijos apimtis prognozuojama pagal tyrimo pobūdį.

Analizuojant kiekybinius rodiklius, medikų uždaviniai, performuluoti statistikos požiūriu, dažniausiai yra tokie:

- įvertinti vidurkį norimu patikimumu;
- nustatyti imties dydį, reikalingą kartotinių tyrimų vidurkių reikšmingam skirtumui konstatuoti;
- nustatyti imties dydį, reikalingą dviejų grupių vidurkių reikšmingam skirtumui konstatuoti.

Pateiksime prognozuojamą imties dydį kiekvienam minėtam atvejui.

Imties dydis, reikalingas įvertinti vidurkį patikimumu δ. Tam tikro ligonių kontingento (pvz., sergančių MI, CD...) kiekybinio rodiklio (pvz., SAS, ŠSD...) reikšmių visumą atspindi imties x_1, x_2 ... x_n vidurkis $\bar{x}$. Kaip patikimai $\bar{x}$ įvertina populiacijos vidurkį, nurodo vidurkio standartinė paklaida $SE = s/\sqrt{n}$. Ji yra standartinio nuokrypio įvertis. *SE* dydis priklauso nuo duomenų kitimo – imties standartinio nuokrypio *s* ir imties dydžio *n*. Kai *s* yra fiksuotas, *SE* būna tuo mažesnė, kuo didesnis *n*. Todėl norint įvertinti vidurkį tam tikru patikimumu, t. y. kad *SE* neviršytų dydžio δ, reikia parinkti *n*, ne mažesnį už

$N = (s/\delta)^2$.

Imties dydžio nustatymas kartotinių tyrimų atveju. Analizuodami vaisto poveikį tam tikram ligonio būklę atspindinčiam rodikliui, pavyzdžiui, SAS, atliekami tyrimai: tam pačiam ligoniui prieš ir po gydymo nustatoma rodiklio reikšmė. Po to lyginami rodiklio reikšmių prieš ir po gydymo vidurkiai (6.3 skyrius). Darant prielaidą, kad rodiklio skirstinys yra normalusis, šiems kartotiniams vidurkiams palyginti naudojamas kartotinių imčių t kriterijus su statistika (6.6): $t = \bar{d}\sqrt{n} / s_d$; čia $\bar{d}$ – matavimų skirtumo vidurkis, s_d – skirtumų standartinis nuokrypis, *n* – tirtų ligonių skaičius. Vidurkių skirtumas statistiškai reikšmingai skirsis nuo 0 su reikšmingumo lygmeniu α (arba patikimumu $P = 1 - \alpha$), jei statistikos reikšmė tenkins sąlygą: $|t| = |\bar{d}|\sqrt{n}/s_d > t_{1-\alpha/2}(n-1)$; čia $t_{1-\alpha/2}(n-1)$ – Stjudento skirstinio su $n-1$ laisvės laipsniu $1-\alpha/2$ lygio kvantilis. Tai bus tuo atveju, kai $n > (t_{1-\alpha/2}(n-1)s_d/\bar{d})^2$. Pastarojoje formulėje $t_{1-\alpha/2}(n-1)$ galima keisti į standartinio normaliojo skirstinio kvantilį $z_{1-\alpha/2}$, kadangi $t_p(n-1)$ nedaug skiriasi nuo z_p, kai *n* viršija 10. Norint konstatuoti, kad kartotinių matavimų vidurkių skirtumas reikšmingai skiriasi nuo 0, būtinas imties dydis, ne mažesnis už

$N = (z_{1-\alpha/2}s_d/\bar{d})^2$.

Iš *N* formulės matome: kuo didesnė matavimų skirtumų dispersija s_d^2, tuo didesnis turi būti imties dydis. Esant standartiniam reikšmingumo lygme-

niui α = 0,05, $z_{1-\alpha/2} = 1{,}96$, $(z_{1-\alpha/2})^2 = 3{,}84$. Sumažinę reikšmingumo lygmenį iki 0,01, gauname $(z_{1-\alpha/2})^2 = (z_{0{,}995})^2 = (2{,}58)^2 = 6{,}66$. Tai 73 % daugiau nei $(z_{0{,}975})^2$. Taigi, sumažinus I rūšies klaidos tikimybę nuo 5 % iki 1 %, reikiamas imties dydis padvigubėja.

14.4 pavyzdys [4, 162 p.]. Tiriant antihipertenzinio vaisto poveikį, studijos pradžioje ir pabaigoje išmatuotas SAS. Nustatyta, kad studijos pabaigoje SAS vidurkis sumažėjo 10 mmHg ($\overline{d} = 10$). SAS skirtumo standartinis nuokrypis $s_d = 15$ mmHg. Vidurkių skirtumą vertinsime patikimumu $P = 0{,}95$, t. y. α = 0,05. Prognozuojamas imties dydis yra:

$N = (z_{1-\alpha/2}\, s_d / \overline{d}\,)^2 = (1{,}96 \times 15/10)^2 = 384 \times 225/100 = 8{,}64.$

Taigi SAS vidurkio reikšmingam sumažėjimui konstatuoti užtenka 9 ligonių.

Imties dydis, reikalingas dviejų grupių vidurkių reikšmingam skirtumui konstatuoti. Sakykime, abiejų ligonių grupių kiekybinio kintamojo reikšmės x_i ir y_i yra normalieji ats. dydžiai su ta pačia dispersija ir abi grupės yra vienodo dydžio. Tuomet šių ligonių grupių vidurkiams palyginti naudojamas nepriklausomų imčių t kriterijus su statistikos reikšme (6.4 skyrius):

$$t = \frac{(\overline{x} - \overline{y})}{s_p\sqrt{2/n}} = \frac{\overline{d}}{s_p\sqrt{2/n}};$$

čia $\overline{x}$ ir $\overline{y}$ – imčių vidurkiai, n – imties dydis, s_p^2 – imčių jungtinė dispersija. Jei $|t| > t_{1-\alpha/2}(2n-2)$, su reikšmingumu α galima tvirtinti, kad skirstinių vidurkiai nėra lygūs. Taigi n turi būti toks, kad

$$\frac{\overline{d}\sqrt{n}}{\sqrt{2}s_p} > t_{1-\alpha/2}(2n-2) \text{ arba } n > \frac{2t_{1-\alpha/2}^2(2n-2)s_p^2}{\overline{d}^2}.$$

Pastarojoje formulėje $t_{1-\alpha/2}(2n-2)$ galima keisti į $z_{1-\alpha/2}$. Norint konstatuoti reikšmingą vidurkių skirtumą, kiekvienoje grupėje būtina turėti imtį, ne mažesnę už

$$N = \frac{2(z_{1-\alpha/2}s_p)^2}{(\overline{d})^2}. \tag{14.1}$$

Prognozuojant imties dydį pagal (14.1) formulę, atsižvelgiama tik į I rūšies klaidos – atmesti teisingą hipotezę – tikimybę. Standartinė leidžiama II rūšies klaidos tikimybė – β = 0,2. Atsižvelgiant į abiejų rūšių klaidų tikimybes, vietoj (14.1) formulės reikėtų naudoti tokią:

$$N = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 s_p^2}{(\overline{d})^2}. \tag{14.2}$$

14.5 pavyzdys [4, 162 p.]. Antihipertenzinio vaisto efektyvumui tirti organizuota randomizuota studija: atrinkti ligoniai atsitiktinai suskirstyti po lygiai į 2 grupes. Viena grupė vartojo antihipertenzinio vaisto, kita – kontrolinė – placebo. Studijos pabaigoje nustatyta, kad vartojusių vaisto SAS vidurkis buvo 10 mmHg mažesnis nei kontrolinės grupės ($\bar{d} = 10$). SAS reikšmių jungtinė dispersija lygi $s_p^2 = 225$. Laikome, kad $\alpha = 0{,}05$, $z_{1-\alpha/2} = 1{,}96$. Tuomet prognozuojamas grupės dydis lygus:

$$N = \frac{2(z_{1-\alpha/2}s_p)^2}{(\bar{d})^2} = \frac{2(1{,}96)^2 225}{100} = 17{,}28.$$

Taigi kiekvienoje grupėje turėtų būti po 18 ligonių.

Atsižvelgiant į daromą antros rūšies klaidą, prognozuojamas imties dydis skaičiuojamas pagal (14.2) formulę. Laikome, kad $\beta = 0{,}2$, $z_{1-\beta} = 0{,}84$. Tuomet

$$N = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 s_p^2}{(\bar{d})^2} = \frac{(1{,}96 + 0{,}84)2 \times 225}{100} = 35{,}28,$$

taigi kiekvienoje grupėje turėtų būti dvigubai daugiau ligonių – po 36.

14.8. Kokybinio rodiklio imties dydžio nustatymas

Imties dydis, reikalingas tikimybei įvertinti duotu patikimumu. Sakykime, turime dvinarį kintamąjį (pvz., (mirė, išgyveno); (serga, neserga)). Šio kintamojo statistinis modelis – atsitiktinis dydis, įgyjantis reikšmę 1 ir 0 su tikimybėmis atitinkamai π ir $(1 - \pi)$, $0 < \pi < 1$. Tikimybė $\pi = P\{X = 1\}$ vertinama proporcija $p = n_s/n$; čia n_s – sergančiųjų skaičius, n – tirtų individų skaičius. Gauto įverčio patikimumui vertinti skaičiuojama p standartinė paklaida $SE(p) = \sqrt{p(1-p)/n}$. Norint įvertinti tikimybę π tam tikru patikimumu, sakykime, $SE(p) \le \delta$, reikia parinkti n, ne mažesnį už

$N = p(1 - p)/\delta^2$.

Iš formulės matyti, kad imties dydis priklauso nuo p ir nuo δ. Dydžio $p(1 - p)$ didžiausia reikšmė 0,25 pasiekiama, kai $p = 0{,}5$. Jei $p = 0{,}1$, tuomet $N = 0{,}1 \times 0{,}9/\delta^2 = 0{,}09/\delta^2$. Kai $p = 0{,}5$, $N = 0{,}25/\delta^2$, taigi vertinant $SE(p)$ tuo pačiu patikimumu, kai $p = 0{,}5$, reikia 2,8 karto didesnės imties, nei esant $p = 0{,}1$.

Imties dydis, reikalingas dviejų tikimybių reikšmingam skirtumui konstatuoti. Tirtas įvykio (sėkmės) pasirodymas dviejose vienodo dydžio ligonių grupėse. Sakykime, poveikio grupėje įvykis (sėkmė) pasirodė n_1 ligonių, kontrolinėje grupėje – n_2 ligonių (15.6 lentelė).

15.6 lentelė. Sėkmės ir nesėkmės pasirodymo dažnis poveikio ir kontrolinėje grupėse

	Sėkmė	Nesėkmė	Iš viso
Poveikis	n_1	$n - n_1$	n
Kontrolė	n_2	$n - n_2$	n

Pažymėkime π_1 ir π_2 – tikimybes, kad sėkmė pasirodė atitinkamai poveikio ir kontrolinės grupės ligoniui. Šių tikimybių įverčiai yra proporcijos $p_1 = n_1 / n$ ir $p_2 = n_2 / n$. Norėdami įvertinti, kiek reikia ligonių grupėse, kad p_1 ir p_2 reikšmingai skirtųsi ($\pi_1 \neq \pi_2$), sudarysime kriterijaus, skirto hipotezei H_0: $\pi_1 = \pi_2$ tikrinti, statistiką. Dideliam n atsitiktiniai dydžiai n_1 ir n_2 turi asimptotinį normalųjį skirstinį su parametrais ($n\pi_1$, $n\pi_1(1 - \pi_1)$) ir ($n\pi_2$, $n\pi_2(1 - \pi_2)$). Remiantis didžiųjų skaičių dėsniu ir centrine ribine teorema, galima tvirtinti, kad statistikos

$$z = \frac{\sqrt{n}(p_1 - p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}$$

skirstinys, esant $\pi_1 = \pi_2$, yra standartinis normalusis. Nulinė hipotezė atmetama ir teigiama, kad sėkmės pasirodymo tikimybės grupėse skiriasi, jei $|z| > z_{1-\alpha/2}$; čia α – reikšmingumo lygmuo. Pertvarkę pastarąją nelygybę, gauname: norint konstatuoti sėkmės tikimybių reikšmingą skirtumą, kiekvienoje grupėje turi būti ne mažiau kaip

$$N = \frac{(p_1(1-p_1) + p_2(1-p_2))\, z^2_{1-\alpha/2}}{(p_1 - p_2)^2}$$

ligonių. Atsižvelgiant į daromą antros rūšies klaidą, prognozuojamas ligonių skaičius grupėje turi būti ne mažesnis kaip:

$$N = \frac{(p_1(1-p_1) + p_2(1-p_2))(z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_2)^2}.$$

14.6 pavyzdys [4, 164 p.]. Vaisto nuo vėžio efektyvumui tirti organizuota randomizuota studija. Stebimi ligoniai atsitiktinai suskirstyti į 2 lygias grupes. Vienos grupės (kontrolinės) ligoniai gydyti standartiškai, kitos (eksperimentinės) grupės ligoniai – nauju vaistu nuo vėžio. Po 5 metų nustatyta, kad eksperimentinėje grupėje išgyveno 60 % ($p_E = 0{,}6$), kontrolinėje grupėje – 50 % ($p_K = 0{,}6$) ligonių. Nustatysime, kiek ligonių turi būti grupėse, kad išgyvenimo tikimybės reikšmingai skirtųsi ($\alpha = 0{,}05$).

Turime: $p_1 = 0{,}6$; $p_2 = 0{,}5$, $p_1 - p_2 = 0{,}1$; $p_1(1-p_1) + p_2(1-p_2) = 0{,}6 \times 0{,}4 + 0{,}5 \times 0{,}5 = 0{,}49$; $N = 0{,}49 \times (z_{0{,}975}/0{,}1)^2 = 0{,}49 \times 3{,}84/0{,}01 = 168{,}2$. Taigi kiekvienoje grupėje turėtų būti bent po 169 ligonius.

Jei eksperimentinės ir kontrolinės grupių išgyvenusių ligonių procentinis skirtumas būtų didesnis, pavyzdžiui, $p_1 = p_E = 0{,}7$, o $p_2 = p_K = 0{,}6$, tuomet $p_1(1 - p_1) = 0{,}21$, $p_2(1 - p_2) = 0{,}25$; $N = 0{,}46 \times 3{,}84/0{,}04 = 44{,}1$, ir kiekvienoje grupėje užtektų 45 ligonių.

14 skyriaus literatūra

1. Armitage P., Berry G., Matthews J. N. S. *Statistical Methods in Medical Research*. 2002. Fourth ed., Blackwell Science, p. 817.
2. Grabauskas V. J., Misevičienė I., Padaiga Ž. ir kt. *Fundamentinė epidemiologija*. 2003. Kaunas: KMU, 144 p.
3. Feinstein A. R. *Principles of Medical Statistics*. 2001. Chapman & Hall, p. 701.
4. Jekel J. F., Elmore J. G., Katz D. L. *Epidemiology, Biostatistics and Preventive Medicine*. 1996. London: Saunders, p. 297.
5. Miller J. C., Miller J. N. *Statistics for Analycal Chemistry*. Second ed. 1988. New York: John Wiley & Sons, p. 227.
6. Sapagovas J., Šaferis V., Jurėnienė K., Jurkonienė R., Šimatonienė V., Šimoliūnienė R. *Statistikos ir informatikos pagrindai*. 2008. Kaunas: KMU leidykla, p. 98.

15 SKYRIUS

Daugiamačių duomenų modeliai. Diskriminantinė analizė

Analizuojant surinktus duomenis, vis dažniau naudojamas ne vienas rodiklis, o rodiklių kompleksas – daugiamatis kintamasis. Todėl reikia turėti šio daugiamačio kintamojo statistinį modelį – daugiamatį skirstinį. Taip pat būtini kelių populiacijų daugiamačių vidurkių palyginimo metodai bei daugiamačių kintamųjų ryšio vertinimo rodikliai. Analizuojant daugiamačius duomenis, tenka mažinti ir matavimų skaičių: pavyzdžiui, 10 rodiklių pakeisti dviem išvestiniais kintamaisiais taip, kad prarastume kuo mažiau informacijos. Visoms šioms reikmėms ir skiriami daugiamačiai statistikos metodai.

Analizuojant daugiamačius duomenis, būtina susipažinti su matematikos priemonėmis, skirtomis juos apdoroti – matricų algebra, daugiamačiu normaliuoju skirstiniu.

15.1. Matricos ir vektoriai

Analizuodami daugiamačius duomenis, susiduriame su vektoriaus ir matricos sąvoka. Priminsime, kad matrica vadinama skaičių, išdėstytų stačiakampėje lentelėje su m eilučių ir n stulpelių, visuma. Šią matricą vadiname $m \times n$ matrica. Matricos paprastai žymimos didžiosiomis raidėmis, pavyzdžiui:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots. & \ldots. & \ldots. \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix};$$

čia a_{ij}, $i = 1, 2 \ldots m$; $j = 1, 2 \ldots n$ – matricos elementas, esantis i-tosios eilutės ir j-tojo stulpelio susikirtime. Glaustai matrica žymima:

$A = (a_{ij})$, $i = 1, 2 \ldots m$; $j = 1, 2 \ldots n$.

Matrica pateikiami daugiamačio kintamojo imties duomenys arba kelių imčių duomenys, pavyzdžiui, dispersinėje ar diskriminantinėje analizėje. Matricą, turinčią vieną eilutę ($m = 1$), vadiname **vektoriumi-eilute**. Matricą, turinčią vieną stulpelį ($n = 1$), vadiname **vektoriumi-stulpeliu**. Matricoje A sukeitus eilutes ir stulpelius vietomis, gauname **transponuotą matricą:**

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

Analogiškai transponuotas vektorius-eilutė tampa vektoriumi-stulpeliu:

$$\boldsymbol{r} = (r_1 \dots r_n), \qquad \boldsymbol{r}^T = \begin{pmatrix} r_1 \\ \dots \\ r_n \end{pmatrix}.$$

Matrica vadiname **kvadratine**, jei jos eilučių ir stulpelių skaičius yra vienodas ($m = n$). Kvadratinę matricą, kurioje tik diagonaliniai elementai nelygūs nuliui, vadiname **diagonaline**:

$$T = \begin{pmatrix} t_1 & 0 & 0 & \dots & 0 \\ 0 & t_2 & 0 & \dots & 0 \\ 0 & 0 & t_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & t_n \end{pmatrix}.$$

Diagonalinė matrica, kurios diagonalėje yra vienetai, vadiname **vienetine**:

$$I = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Kvadratinė matrica A vadinama simetrine, jei, sukeitus eilutes ir stulpelius vietomis, ji nepakis, t. y. $A = A^T$ arba $a_{ij} = a_{ji}$ visiems i ir j.

Veiksmai su matricomis atliekami pagal tam tikras taisykles. Matricų A ir B suma yra matrica C, kurios elementas c_{ij} lygus a_{ij} ir b_{ij} sumai. Analogiškai: jei $C = A - B$, tai $c_{ij} = a_{ij} - b_{ij}$; jei $C = kA$, tai $c_{ij} = ka_{ij}$, $i = 1, 2 \dots m$, $j = 1, 2 \dots n$. Jei $C = A \times B$, tai $c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}$.

Jei vieną matricos eilutę (stulpelį) galima išreikšti kitų matricos eilučių (stulpelių) tiesine kombinacija, tuomet sakoma, kad ši eilutė (stulpelis) yra tiesiškai priklausoma nuo kitų matricos eilučių (stulpelių). Tokia eilutė ar stulpelis nepateikia papildomos informacijos; tą pačią skaitinę informaciją galima pateikti mažesnės apimties matrica. Didžiausias tiesiškai nepriklausomų eilučių (stulpelių) skaičius vadinamas matricos **rangu**. Sakykime, A – 4×4 kvadratinė matrica; jos rangas lygus 2. Tuomet visą matricos A skaitinę informaciją galima pateikti 2×2 matrica. Matricos rangui nustatyti vartojama determinanto sąvoka. Kvadratinei matricai A pagal tam tikrą taisyklę priskiriamas skaičius, vadinamas matricos **determinantu**, žymimu $|A|$ arba det$A$. 2×2 matricos determinantas skaičiuojamas taip:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \Delta = \det A = a_{11}a_{22} - a_{12}\,a_{21}.$$

Determinanto skaičiavimo formulės pateikiamos ir didesnės apimties matricai.

Jei $n \times n$ matricos determinantas lygus 0, tokia matrica vadinama **išsigimusia.** Determinantas lygus 0 tik tada, kai jo eilutės ar stulpeliai yra tiesiškai priklausomi.

Kvadratinei neišsigimusiai matricai A apibrėžiama **atvirkštinė matrica** A^{-1} taip: $A^{-1}A = I$. 2×2 matricos atveju atvirkštinės matricos formulė yra tokia:

$$A^{-1} = \begin{pmatrix} a_{22}/\Delta & -a_{12}/\Delta \\ -a_{21}/\Delta & a_{11}/\Delta \end{pmatrix}.$$

Matricos A diagonalinių elementų suma vadinama matricos A **pėdsaku**: $\mathrm{tr}(A) = a_{11} + ... + a_{nn}$.

Sakykime, A – kvadratinė $n \times n$ matrica, $\boldsymbol{x}$ – n ilgio vektorius-stulpelis. Skaliarinė funkcija $Q(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$ vadinama **kvadratine forma**. Ji skaičiuojama pagal formulę:

$$Q(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j;$$

čia x_i ir a_{ij} – vektoriaus-stulpelio ir matricos A elementai. Sakoma, kad matrica A yra teigiamai apibrėžta, jei bet kuriems $x_1, x_2 \ldots x_n$ $Q(\boldsymbol{x}) \geq 0$.

Kvadratinė matrica C vadinama **ortogonalia**, jei $C^T C = I$. Jei A yra simetrinė matrica, tuomet visada galima rasti tokią ortogonalią matricą C, kad matrica $C^{-1}AC^T$ būtų diagonalinė:

$$C^{-1}AC^{T} = T = \begin{pmatrix} \lambda_1 0 \dots 0 \\ 0 \lambda_2 \dots 0 \\ 0 \ 0 \dots \lambda_n \end{pmatrix}.$$

Skaičiai $\lambda_1, \lambda_2 \dots \lambda_n$ priklauso tik nuo matricos A ir vadinami jos **tikriniais skaičiais**. Tikriniai skaičiai yra lygties $|A - \lambda I| = 0$ šaknys. Be to, $\det A = \lambda_1\lambda_2 \dots \lambda_n$.

Lygties $A\boldsymbol{x}_i = \lambda_i\boldsymbol{x}_i$ sprendinys $\boldsymbol{x}_i = (x_{1i}, x_{2i} \dots x_{ni})$ vadinamas matricos A tikriniu vektoriumi, atitinkančiu tikrinį skaičių λ_i, $i = 1, 2 \dots n$.

15.2. Daugiamačiai atsitiktiniai vektoriai. Daugiamatis normalusis skirstinys*

Kaip minėta 2.8 skyriuje, daugiamačiu atsitiktiniu vektoriumi (ats. v.) vadinamas vektorius $\boldsymbol{X} = (X^{(1)}, X^{(2)} \dots X^{(p)})$, kurio komponentės yra atsitiktiniai dydžiai; čia p – vektoriaus $\boldsymbol{X}$ matavimų skaičius, $X^{(1)}, X^{(2)} \dots X^{(p)}$ – atsitiktiniai dydžiai, vadinami vektoriaus $\boldsymbol{X}$ koordinatėmis, arba komponentėmis. Pateiksime atsitiktinių vektorių skaitines charakteristikas.

Atsitiktinio vektoriaus reikšmių „sankaupa“ p-matėje erdvėje – vidurkių vektorius $\boldsymbol{m} = (EX^{(1)}, EX^{(2)} \dots EX^{(p)})$. Ats. v. koordinačių kitimą ir jų tarpusavio priklausomybę charakterizuoja kovariacijų matrica $V = \mathrm{cov}(\boldsymbol{X}) = (v_{ij})$, $i = 1, 2 \dots p$; $j = 1, 2 \dots p$; čia v_{ij} – kovariacijų matricos elementas, esantis i-tos eilutės ir j-tojo stulpelio susikirtime: $v_{ij} = \mathrm{cov}(X^{(i)}, X^{(j)})$ – kovariacija tarp ats. d. $X^{(i)}$ ir $X^{(j)}$. Kovariacijų matricos diagonalėje yra ats. d. $X^{(i)}$ dispersijos, nes, pagal apibrėžimą, $v_{ii} = DX^{(i)}$. Kovariacijų matrica simetrinė: $v_{ij} = v_{ji}$, nes $\mathrm{cov}(X^{(i)}, X^{(j)}) = \mathrm{cov}(X^{(j)}, X^{(i)})$. Todėl kovariacijų matricos determinantas $\det V$ yra neneigiamas. Jei $\det V = 0$, tuomet kažkurios vektoriaus $\boldsymbol{X}$ koordinatės yra tiesiškai priklausomos. Jei ats. vektoriaus koordinatės yra nepriklausomos, tuomet kovariacijų matrica – diagonalinė. Daugiamačio ats. v. $(X^{(1)}, X^{(2)} \dots X^{(p)})$ koordinačių tarpusavio tiesinio ryšio analizei naudojama koreliacijų matrica: (ρ_{ij}), $i = 1 \dots p$; $j = 1 \dots p$; čia $\rho_{ij} = \rho(X^{(i)}, X^{(j)})$ – koreliacijos koeficientas tarp $X^{(i)}$ ir $X^{(j)}$. Koreliacijų matrica yra simetrinė; jos diagonalėje yra vienetai (nes $\rho_{ij} = \rho_{ji}$ bei $\rho_{ii} = \rho(X^{(i)}, X^{(i)}) = 1$). Jei ats. v. koordinatės nepriklausomos, tuomet koreliacijų matrica yra vienetinė.

Teoriniuose ir taikomuose statistikos darbuose dažniausiai naudojamas ats. v. modelis – daugiamatis normalusis skirstinys.

Daugiamatis normalusis skirstinys. Sakoma, kad ats. v. $\boldsymbol{X} = (X^{(1)}, X^{(2)} \dots X^{(p)})$ skirstinys yra p-matis normalusis su parametrais $(\boldsymbol{m}, V)$, jei ats. v. $\boldsymbol{X}$ tankis lygus:

$p(\boldsymbol{x}) = (\sqrt{2\pi})^{-p}(\det V)^{-1/2}\exp\{-(1/2)(\boldsymbol{x} - \boldsymbol{m})V^{-1}(\boldsymbol{x} - \boldsymbol{m})^T\}$;

čia $\boldsymbol{m} = (m_1, m_2 \ldots m_p)$, $\boldsymbol{x} = (x_1, x_2 \ldots x_p)$, V – simetrinė teigiamai apibrėžta matrica. Daugiamačio normaliojo skirstinio parametrų prasmė: $\boldsymbol{m}$ yra ats. v. $\boldsymbol{X}$ vidurkių vektorius ($m_i = EX^{(i)}$), o V yra ats. v. $\boldsymbol{X}$ kovariacijų matrica. Kovariacijų matricos determinantas apibūdina normaliojo ats. v. kitimą. Vienmačio normaliojo ats. d. $X \sim N(m, \sigma^2)$ kitimą apibūdina jo dispersija, nes

$P\{|X - m| \le t\} = P\{|(X - m)/\sigma| \le t/\sigma\} = P\{|z| \le t/\sigma\} = P(\sigma)$;

čia $z \sim N(0, 1)$. Analogiškai, normaliam ats. vektoriui $\boldsymbol{X}$: $P\{(\boldsymbol{X} - \boldsymbol{m}) \in A\} = P(\sqrt{\det V})$. Taigi kovariacijų matricos determinantas yra daugiamatis dispersijos analogas.

Normaliojo ats. v. kitimą taip pat charakterizuoja kovariacijų matricos pėdsakas – ats. v. $X^{(i)}$ dispersijų suma: $\mathrm{tr}(V) = v_{11} + v_{22} + \ldots + v_{pp} = DX^{(1)} + \ldots + DX^{(p)}$ bei kovariacijų (ar koreliacijų) matricos tikriniai skaičiai $\lambda_1, \lambda_2 \ldots \lambda_p$, nes $\det V = \lambda_1 \lambda_2 \ldots \lambda_p$. Plačiau apie daugiamatį normalųjį skirstinį pateikta vadovėlyje [3].

Daugiamačių duomenų standartizavimas. Sakykime, ats. d. X turi normalųjį skirstinį su vidurkiu m ir dispersija σ^2. Tuomet ats. d. $Y = (X - m)/\sigma$ skirstinys yra standartinis normalusis, o $Y^2 = (X - m)^2(1/\sigma^2)$ skirstinys – χ^2 su 1 laisvės laipsniu. Sakykime, $\boldsymbol{X} \sim N(\mathbf{m}, V)$; analogiškai ats. v. $\boldsymbol{X}$ standartizuosime. Daugiamačiams duomenims vietoj skirtumo $X - m$ turime vektorių $\boldsymbol{X} - \boldsymbol{m}$, vietoj σ^2 – matricą V, o vietoj $1/\sigma^2$ pagal veiksmų su matricomis taisykles – atvirkštinę matricą V^{-1}. Sandauga $(X - m)^2(1/\sigma^2)$ daugiamačių duomenų atveju keičiasi į kvadratinę formą:

$D^2 = (\boldsymbol{X} - \boldsymbol{m})V^{-1}(\boldsymbol{X} - \boldsymbol{m})^T$.

Ats. d. D^2 vadinamas **Mahalanobiso atstumu**. D^2 skirstinys yra χ^2 su p laisvės laipsnių. Taigi D^2 – standartinio normaliojo ats. dydžio kvadrato daugiamatis analogas.

Daugiamačių duomenų statistinis modelis. Daugiamatis normalusis skirstinys – dažniausiai naudojamas daugiamačio kiekybinio kintamojo statistinis modelis. Atskirų p kintamųjų normalumas dažnai leidžia laikyti p kintamųjų jungtinį skirstinį normaliuoju.

Daugiamačio kintamojo imties skaitinės charakteristikos. Sakykime, $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n, \boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)} \ldots x_i^{(p)})$ – daugiamačio kintamojo imtis. Vektorių $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n$ sankaupą p-matėje erdvėje charakterizuoja vidurkių vektorius $\bar{\boldsymbol{x}} = (\bar{x}^{(1)}, \bar{x}^{(2)} \ldots \bar{x}^{(n)})$; čia $\bar{x}^{(j)} = (1/n)\sum_{i=1}^{n} x_i^{(j)}$ – j-tosios koordinatės vidurkis. Atskirų koordinačių kitimą apibūdina jų dispersijos s_j^2. Ryšį tarp duomenų

vektoriaus *r*-tos ir *c*-tos komponenčių vertinti skaičiuojamas kovariacijos įvertis:

$$\hat{v}_{rc} = (1/(n-1))\sum_{i=1}^{n}(x_i^{(r)} - \bar{x}^{(r)})(x_i^{(c)} - \bar{x}^{(c)}),\ r = 1 \ldots p;\ c = 1 \ldots p;\ \hat{v}_{ii} = s_i^2.$$

Informacija apie daugiamačių duomenų ryšį ir kitimą pateikiama (imties) kovariacijų matrica $\hat{V} = (\hat{v}_{ij})$. Matricos $\hat{V}$ elementai priklauso nuo koordinačių mastelio, todėl ryšiui tarp atskirų koordinačių vertinti pateikiama koreliacijų matrica $R = (r_{ij})$, $i, j = 1 \ldots p$:

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix};$$

čia $r_{ij} = \hat{v}_{ij}/(s_i s_j)$ – imties koreliacijos koeficientas tarp *i*-tosios ir *j*-tosios koordinačių; $r_{ii} = 1$, $i = 1 \ldots p$. Tiek kovariacijų, tiek koreliacijų matricos yra simetrinės ir teigiamai apibrėžtos.

Sakykime, $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n$ yra daugiamačio kintamojo imtis – ats. vektoriai, turintys normalųjį skirstinį su vidurkių vektoriumi $\boldsymbol{m}$ ir kovariacijų matrica *V*. Šie parametrai nežinomi, jie vertinami naudojant imties reikšmes. $\boldsymbol{m}$ įvertis yra imties vidurkių vektorius $\bar{\boldsymbol{x}}$, matricos *V* – imties kovariacijų matrica $\hat{V}$.

Išskirtims nustatyti skaičiuojamas empirinis Mahalanobiso atstumas:

$$D_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\hat{V}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

(vietoj dydžio $(x_i - \bar{x})/s$). Statistikos $F_i = \dfrac{(n-p)}{p(n^2-1)} D_i^2$ skirstinys yra Fišerio su $(p, n - p)$ laisvės laipsnių. Todėl jei apskaičiuota F_i reikšmė viršija atitinkamą Fišerio skirstinio kvantilį, šią reikšmę laikome išskirtimi.

15.3. Hipotezių tikrinimas daugiamačių duomenų atveju*

Hipotezė apie vidurkių vektoriaus lygybę konkrečiam vektoriui. 5.6 skyriuje pateiktas t kriterijus vienai imčiai – kriterijus, skirtas tikrinti hipotezę apie imties $x_1, x_2 \ldots x_n$, turinčios normalųjį skirstinį, populiacijos vidurkio *m* lygybę normai m_0. Šio kriterijaus statistika lygi $t = \sqrt{n}(\bar{x} - m_0)/s$; čia $\bar{x}$ ir *s* – imties vidurkis ir standartinis nuokrypis. Daugiamačių duomenų atveju hipotezei apie populiacijos vidurkių vektoriaus $\boldsymbol{m}$ tapatybę konkrečiam vektoriui $\boldsymbol{m}_0$ (H_0: $\boldsymbol{m} = \boldsymbol{m}_0$, alternatyva H_A: $\boldsymbol{m} \neq \boldsymbol{m}_0$) tikrinti naudojamas Hotelingo (*Hotelling*) kriterijus su statistika $T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{m}_0)\hat{V}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{m}_0)^T$. T^2 yra t^2 statistikos daugiamatis analogas – $(\bar{x} - m_0)$ keičiama į $(\bar{\boldsymbol{x}} - \boldsymbol{m}_0)$, $(1/s^2)$ – į $\hat{V}^{-1}$.

Statistiniuose paketuose pateikiama Hotelingo kriterijaus p reikšmė. Jei $p < \alpha$ (α – reikšmingumo lygmuo), tvirtiname, kad $\boldsymbol{m} \neq \boldsymbol{m}_0$, o jei $p \geq \alpha$ – tvirtinimui, kad „populiacijos vidurkių vektorius sutampa su normos vektoriumi“, neprieštaraujame.

Hipotezė apie dviejų populiacijų daugiamačių vidurkių vektorių lygybę. Sakykime, $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n$ ir $\boldsymbol{y}_1, \boldsymbol{y}_2 \ldots \boldsymbol{y}_m$ – imtys iš X ir Y populiacijų – nepriklausomi ats. v., turintys daugiamatį normalųjį skirstinį su vidurkių vektoriais $\boldsymbol{m}_1$ (X populiacijos) ir $\boldsymbol{m}_2$ (Y populiacijos) bei ta pačia kovariacijų matrica V. Nulinei hipotezei H_0: $\boldsymbol{m}_1 = \boldsymbol{m}_2$ su alternatyva H_A: $\boldsymbol{m}_1 \neq \boldsymbol{m}_2$ tikrinti skaičiuojama statistika:

$T^2 = (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})\hat{V}_u^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^T$; čia $\hat{V}_u = \hat{V}_p(1/n + 1/m)$, $\hat{V}_p$ – jungtinė kovariacijų matrica. Kai $p = 1$, $\hat{V}_u = s_p^2(1/n + 1/m)$ ir statistika T^2 lygi nepriklausomų imčių t kriterijaus statistikos kvadratui:

$$\frac{(\bar{x} - \bar{y})^2}{s_p^2(1/n + 1/m)}.$$

Nulinė hipotezė ar alternatyva pasirenkamos pagal kriterijaus p reikšmę taip pat, kaip ir tikrinant hipotezę apie vidurkių vektoriaus lygybę normos vektoriui.

Kelių normalųjį skirstinį turinčių populiacijų daugiamačių vidurkių palyginimas (MANOVA). Kelių populiacijų, apibūdinamų kokybinio faktoriaus reikšmėmis (lygiais), vienmačio kiekybinio atsako Y vidurkiams palyginti skirta vienfaktorė dispersinė analizė. Joje F kriterijaus, skirto hipotezei apie kelių vidurkių lygybę tikrinti, statistika gaunama skaidant visą kvadratų sumą $SST = \sum_{i=1}^{I}\sum_{j=1}^{nI}(y_{ij} - \bar{y})^2$ į dvi dedamąsias: SSW ir SSB. Daugiamačio kintamojo atveju duomenų kitimą apibūdina visa tarpusavio sandaugų matrica $T = (t_{rc})$:

$$t_{rc} = \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij}^{(r)} - \bar{y}^{(r)})(y_{ij}^{(c)} - \bar{y}^{(c)}),\ r, c = 1 \ldots p;$$

čia $\boldsymbol{y}_{ij} = (y_{ij}^{(1)} \ldots y_{ij}^{(p)})$ – daugiamačio atsako reikšmės, nustatytos esant faktoriaus i-tajam lygiui; $\bar{\boldsymbol{y}} = (\bar{y}^{(1)} \ldots \bar{y}^{(p)})$ – atsako bendrų vidurkių vektorius. Duomenų kitimą grupių viduje atspindi vidinė tarpusavio sandaugų matrica $W = (w_{rc})$:

$$w_{rc} = \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij}^{(r)} - \bar{y}_i^{(r)})(y_{ij}^{(c)} - \bar{y}_i^{(c)}),\ r, c = 1 \ldots p;$$

čia $\bar{\boldsymbol{y}}_i = (\bar{y}_i^{(1)} \ldots \bar{y}_i^{(p)})$ – atsako vidurkių vektorius, nustatytas esant faktoriaus i-tajam lygiui. Matrica $B = T - W$ vadinama tarpgrupine tarpusavio sandaugų matrica.

Hipotezei apie kelių vienmačių vidurkių lygybę tikrinti naudojama statistika $F = SSW(I - 1)/[I(n - 1)SSB]$. Daugiamačiams vidurkiams palyginti naudojamos kelios statistikos, išreiškiamos matricų B ir W funkcijomis:

- matricos $W^{-1}B$ didžiausias tikrinis skaičius (*Roy's largest root*);
- matricos $B(B + W)^{-1}$ pėdsakas (*Pillai trance*);

$\Lambda = \frac{|W|}{|B+W|} = \frac{|W|}{|T|}$ – Vilkso (*Wilks*) *lambda*.

Statistiniuose paketuose pateikiamos šių statistikų reikšmės ir atitinkamos p reikšmės. Jei $p < \alpha$ (α – reikšmingumo lygmuo), tvirtiname, kad kelių populiacijų daugiamačiai vidurkiai nėra lygūs. Jei $p \geq \alpha$ – daugiamačių vidurkių lygybei neprieštaraujame.

15.4. Diskriminantinės analizės samprata ir objektai

Diskriminantinė analizė – visuma metodų, įgalinančių sudaryti taisyklę naujiems individams klasifikuoti, remiantis pradine individų klasifikacija. Diskriminantinės analizės esmę paaiškinsime dviem pavyzdžiais.

15.1 pavyzdys. Universiteto priėmimo komisijos tikslas – pagal egzaminų rezultatus kandidatą į studentus priskirti sėkmingai baigusiųjų universitetą (priimti į universitetą) ar nebaigusiųjų universiteto (nepriimti į universitetą) grupei. Priimdama sprendimą, komisija remiasi ankstesnių metų studentų egzaminų rezultatais.

15.2 pavyzdys. Pagal tam tikrus simptomus gydytojas privalo nustatyti, kuriuo iš K susirgimų pacientas serga. Nustatydamas susirgimą, gydytojas remiasi ankstesnių tyrimų duomenimis.

Diskriminantinės analizės apibrėžimą formalizuosime.

Analizuojamos kelios populiacijos, kurių individai apibūdinami vienu ar keliais rodikliais – atsitiktiniais dydžiais ar vektoriais. Iš kiekvienos populiacijos (ar generalinės visumos) atrenkame po vienmačio ar daugiamačio kintamojo imtį – nustatytas individo reikšmes. Diskriminantinės analizės uždavinys – remiantis šia atranka, sudaryti taisyklę, leidžiančią atrankos elementą (individą) priskirti vienai iš kelių populiacijų (klasių, grupių). Priskyrimo taisyklė sudaroma naudojantis imties elementų savybėmis ir, žinoma, pirmine klasifikacija. Diskriminantinė analizė dar vadinama klasifikacija su apmokymu.

Diskriminantinė analizė taikoma įvairiose srityse. Archeologijoje ir teismo medicinoje pagal rastų kaulų matmenis nustatoma individo lytis, rasė, įvertinamas amžius. Tam naudojamos funkcijos, kuriomis pagal antropometri-

nius rodiklius individas priskiriamas vienai ar kitai rasei ar lyčiai. Naudodami ligonių tyrimo duomenimis, medikai diagnozuoja susirgimą.

Diskriminantinėje analizėje nustatomas fiksuotas klasių (populiacijų) skaičius k. Klasėse atsitiktinai atrenkame po n_1, n_2 ... n_k individų ir kiekvienam individui nustatome p kintamųjų $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ reikšmių. Kintamieji gali būti tiek kiekybiniai, tiek kokybiniai. i-tosios klasės j-tojo individo p-mačio kintamojo $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(p)})$ reikšmes pažymėkime $\boldsymbol{x}_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)} \ldots x_{ij}^{(p)})$. Diskriminantinėje analizėje naudojamų duomenų struktūra pateikta 15.1 lentelėje.

Diskriminantinės analizės tikslas – remiantis pradine individų klasifikacija, t. y. reikšmėmis $\boldsymbol{x}_{ij}$, $j = 1, 2 \ldots n_j$, $i = 1, 2 \ldots k$, sudaryti taisyklę, leidžiančią klasifikuoti naujus individus. Klasifikavimo taisyklė sudaroma pagal duomenų statistinį modelį.

15.1 lentelė. Diskriminantinės analizės duomenys

Klasė	Kintamieji			
	$x^{(1)}$	$x^{(2)}$	...	$x^{(p)}$
1	$x_{11}^{(1)}, x_{12}^{(1)} \ldots x_{1n1}^{(1)}$	$x_{11}^{(2)}, x_{12}^{(2)} \ldots x_{1n1}^{(2)}$	...	$x_{11}^{(p)}, x_{12}^{(p)} \ldots x_{1n1}^{(p)}$
2	$x_{21}^{(1)}, x_{22}^{(1)} \ldots x_{2n2}^{(1)}$	$x_{21}^{(2)}, x_{22}^{(2)} \ldots x_{2n2}^{(2)}$	...	$x_{21}^{(p)}, x_{22}^{(p)} \ldots x_{2n2}^{(p)}$
...	....	....	...	....
k	$x_{k1}^{(1)}, x_{k2}^{(1)} \ldots x_{knk}^{(1)}$	$x_{k1}^{(2)}, x_{k2}^{(2)} \ldots x_{knk}^{(2)}$	...	$x_{21}^{(p)}, x_{22}^{(p)} \ldots x_{2n2}^{(p)}$

15.5. Klasifikavimas minimizuojant klaidingos klasifikacijos nuostolius. Bajeso klasifikacija

Sakykime, turime dvi populiacijas Π_1 ir Π_2 (arba dvi klases I ir II). Šių populiacijų daugiamačiai kintamieji – atsitiktiniai vektoriai, charakterizuojami skirstinio tankiais arba tikimybinėmis funkcijomis $f_1(\boldsymbol{x})$ ir $f_2(\boldsymbol{x})$, čia $\boldsymbol{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ – p matavimų vektorius.

Didžiausio tikėtinumo klasifikacijos taisyklė: vektorių $\boldsymbol{x}$ priskiriame I klasei, jei

$$f_1(\boldsymbol{x}) \geq f_2(\boldsymbol{x}), \tag{15.1}$$

ir priskiriame II klasei, jei $f_1(\boldsymbol{x}) < f_2(\boldsymbol{x})$.

Pateiksime klasifikacijos taisyklę, pagal kurią p matavimų vektorių priskirsime vienai iš dviejų klasių taip, kad klaidingos klasifikacijos nuostoliai būtų

minimalūs. Pažymėkime $C(i|j)$ – patiriami nuostoliai, jei j-tosios populiacijos elementą priskiriame i-tajai populiacijai (klasei). Klasifikavimo nuostolių struktūra pateikta 15.2 lentelėje. Analogiškai pažymėkime $p(i|j)$ – tikimybę, kad j-tosios klasės elementą priskirsime i-tajai klasei; π_i – tikimybė, kad sujungus abi klases (populiacijas) ir iš jų mišinio atsitiktinai parinktas ats. v. $\boldsymbol{X}$ priklausys i-tajai klasei. Dviejų klasių atveju $\pi_1 + \pi_2 = 1$ (nes elementas $\boldsymbol{X}$ priklauso arba I, arba II klasei). Tikimybės π_i nustatomos remiantis pirminiais tyrimais. Pavyzdžiui, 15.1 pavyzdyje pagal ankstesnių metų duomenis žinome, kad universitetą sėkmingai baigė 2/3 įstojusių; taigi $\pi_1 = 2/3$, $\pi_2 = 1/3$. 15.2 pavyzdyje, klasifikuojant į sergančius liga A ir liga B, žinoma, kad liga A serga apie 20 %, liga B – 80 % atsiųstų pacientų; taigi šiuo atveju galima laikyti, kad $\pi_1 = 0{,}2$, o $\pi_2 = 0{,}8$. Tikimybės π_i vadinamos apriorinėmis (žinomomis prieš eksperimentą).

15.2 lentelė. Klasifikavimo nuostolių struktūra

		Klasifikacija	
		Π_1	Π_2
Teisingas priskyrimas	Π_1	0	$C(2\|1)$
	Π_2	$C(1\|2)$	0

Klasifikavimo taisyklę būtina sudaryti taip, kad patirti nuostoliai dėl klaidingos klasifikacijos būtų minimalūs. Klaidingos klasifikacijos nuostoliai lygūs:

$C(2|1)p(2|1)\pi_1 + C(1|2)p(1|2)\pi_2$.

Klaidingos klasifikacijos tikimybės $p(i|j)$ nustatomos naudojantis didžiausio tikėtinumo taisykle (15.1). Todėl minimali klaidinga klasifikacija gaunama tada, kai vektorius $\boldsymbol{x}$ priskiriamas I klasei, esant

$$f_1(\boldsymbol{x})C(2|1)\pi_1 \geq f_2(\boldsymbol{x})C(1|2)\pi_2, \qquad (15.2)$$

ir priskiriama II klasei, kai $f_1(\boldsymbol{x})C(2|1)\pi_1 < f_2(\boldsymbol{x})C(1|2)\pi_2$.

Praktiškai (15.2) taisyklė realizuojama pakeitus $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$, π_1 ir π_2 jų įvertčiais, gautais remiantis pradine elementų klasifikacija. Funkcijų $f_1(\boldsymbol{x})$ ir $f_2(\boldsymbol{x})$ įvertčiai priklauso nuo duomenų skirstinio modelio; kaip taikyti (15.2) taisyklę konkrečiau, pateikta 15.6–15.8 skyriuose.

Bajeso klasifikacija. Turime k populiacijų (klasių) Π_1, Π_2 ... Π_k. Daroma prielaida, kad populiacijų elementai yra atsitiktiniai vektoriai (atskiru atveju – ats. dydžiai) su tankiais arba tikimybėmis $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ ... $f_k(\boldsymbol{x})$. Tarkime, π_i – apriorinė tikimybė, t. y. tikimybė, kad atsitiktinai parinktas elementas priklauso i-tajai populiacijai; $\pi_1 + \pi_2 + ... + \pi_k = 1$. Pažymėkime $P(\Pi_i|\boldsymbol{x})$ –

sąlyginę tikimybę, kad elementas $\boldsymbol{x}$ priklauso i-tajai klasei. Remdamiesi Bajeso teorema (2.3 skyrius), turime

$$P(\Pi_i \mid x) = \frac{P(\Pi_i)P(x \mid \Pi_i)}{\sum_{i=1}^{k} P(\Pi_i)P(x \mid \Pi_i)};$$

čia $P(\Pi_i) = \pi_i$ – tikimybė, kad atsitiktinai parinktas elementas $\boldsymbol{x}$ priklauso i-tajai klasei, $P(\boldsymbol{x}| \Pi_i)$ – sąlyginė tikimybė, kad i-tosios klasės elementas lygus $\boldsymbol{x}$. $P(\boldsymbol{x}| \Pi_i)$ galime vertinti tankio ar tikimybės $f_i(\boldsymbol{x})$ reikšme. Todėl

$$P(\Pi_i \mid x) = \frac{\pi_i f_i(x)}{\sum_{i=1}^{k} \pi_i f_i(x)}.$$

Elementą $\boldsymbol{x}$ tikslinga priskirti klasei, kurios $P(\Pi_i|\boldsymbol{x})$ (arba $\pi_i f_i(\boldsymbol{x})$) yra didžiausia.

Bajeso klasifikacijos taisyklė: elementas $\boldsymbol{x}$ priskiriamas j-tajai klasei, jei

$$\pi_j f_j(\boldsymbol{x}) = \max_i \pi_i f_i(\boldsymbol{x}). \quad (15.3)$$

(15.3) išraišką galima perrašyti:

$$\pi_j f_j(\boldsymbol{x}) \geq \pi_i f_i(\boldsymbol{x}) \text{ visiems } i \neq j. \quad (15.4)$$

Praktiškai taikant (15.4) taisyklę, $f_j(\boldsymbol{x})$ vertinama remiantis pradinės klasifikacijos duomenimis. Jei daroma prielaida, kad $f_j(\boldsymbol{x})$ – žinomo pavidalo tankio parametrinė funkcija (pvz., normalioji), tuomet naudojant imties reikšmes įvertinami nežinomi $f_j(\boldsymbol{x})$ parametrai. Analogiškai elgiamasi, jei $f_j(\boldsymbol{x})$ – žinomo pavidalo tikimybė. Jei $f_j(\boldsymbol{x})$ parametrinė išraiška (formulė) nėra žinoma, $f_j(\boldsymbol{x})$ vertinama neparametriniais metodais.

15.6. Klasifikacija, kai duomenų skirstiniai yra normalieji

Vienmačio kintamojo atvejis. Naudodami bendresnes (15.3) ir (15.4) formules, sudarysime taisyklę normalųjį skirstinį turinčių kintamųjų klasifikacijai. Sakykime, k – klasių skaičius, o i-tosios klasės kintamojo skirstinys yra normalusis su vidurkiu m_i ir bendra visoms klasėms dispersija σ^2. Pagal apibrėžimą

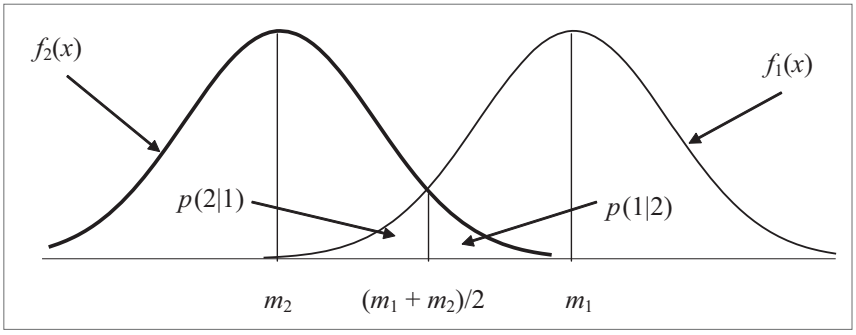
$$f_i(x) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-\frac{(x - m_i)^2}{2\sigma^2}\right), \; i = 1, 2 \ldots k.$$

Įstatę šią išraišką į (15.4) formulę ir atitinkamai ją pertvarkę, gauname klasifikavimo taisyklę: elementą x priskiriame j-tajai klasei, jei

$$m_j x/\sigma^2 + \ln\pi_j - 0{,}5(m_j/\sigma)^2 \geq m_i x/\sigma^2 + \ln\pi_i - 0{,}5(m_i/\sigma)^2 \text{ visiems } i \neq j. \quad (15.5)$$

Praktikoje klasifikuojant normaliuosius ats. dydžius, klasių skirstinio vidurkiai m_i ir bendra dispersija nėra žinomi. m_i ir σ^2 įvertinami naudojant pradinės klasifikacijos duomenis. Tuomet vietoj m_i į (15.5) formulę dedame i-tosios klasės vidurkį $\bar{x}_i$, o vietoj σ^2 – bendros dispersijos įvertį (12.5 skyrius).

Išsamiau panagrinėsime dviejų klasių atvejį. Sakykime, $m_1 > m_2$ ir $\pi_1 = \pi_2$. Populiacijų skirstinių tankiai ir klasifikacijos klaidų tikimybės pateiktos 15.1 pav. Elementas x priskiriamas I klasei, jei $x \geq (m_1 + m_2)/2$. Klaidingo klasifikavimo tikimybės $P(2|1)$ ir $P(1|2)$ lygios $\Phi(-\delta/2)$ ($\pi_1 = \pi_2$); čia $\delta^2 = (m_1 - m_2)^2/\sigma^2$, $\Phi(z)$ – standartinio normaliojo ats. d. skirstinio funkcija. Matome, kad klaidingos klasifikacijos tikimybės priklauso nuo vidurkių skirtumo ir dispersijos. Jei kintamojo vidurkiai m_1 ir m_2 skiriasi nedaug, klaidingos klasifikacijos tikimybės artimos ½ ir šį kintamąjį naudoti klasifikacijai netikslinga.



15.1 pav. Populiacijų tankiai ir klasifikacijos klaidų tikimybės normaliuoju atveju

Klasifikuojant realius duomenis, klasių vidurkiai m_1 ir m_2 ($m_1 > m_2$) įvertinami imčių vidurkiais $\bar{x}_1$ ir $\bar{x}_2$; individas, priskiriamas I klasei, jei jo kintamojo reikšmė viršija $(\bar{x}_1 + \bar{x}_2)/2$. Priešingu atveju individas priskiriamas II klasei.

Kintamąjį tikslinga naudoti klasifikacijai, jei imčių vidurkiai reikšmingai skiriasi. Tai galima patikrinti t kriterijumi (6.4 skyrius). Klasifikuojant į kelias klases, kintamąjį tikslinga naudoti tik tuomet, kai klasių skirstinių vidurkiai nėra lygūs. Ši hipotezė tikrinama F kriterijumi, naudojamu dispersinėje analizėje (12.1 skyrius), arba Kruskalio–Voliso (*Kruskal–Wallis*) kriterijumi (6.9 skyrius).

Klasifikacijai naudojant vieną kintamąjį, klaidingos klasifikacijos tikimybės gali būti nemažos. Todėl klasifikacijai naudojama keletas kintamųjų; naudotini tik tie kintamieji, kurių vidurkiai imtyse reikšmingai skiriasi su $\alpha \leq 0{,}1$. Toks kintamasis laikomas informatyviu klasifikacijai.

Daugiamačio kintamojo atvejis. Turime k populiacijų (klasių). Kiekvienos populiacijos elementai yra p-mačiai ats. vektoriai $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(p)})$, turintys normalųjį skirstinį. Daroma prielaida, kad vektoriaus $\boldsymbol{X}$ koordinatės nėra tiesiškai priklausomos ir visų populiacijų kovariacijų matricos vienodos.

Funkcijos $f_i(\boldsymbol{x})$ yra daugiamačio normaliojo skirstinio su vienodomis kovariacijų matricomis tankiai; įstačius jų išraišką į (15.4) formulę, sudaroma Fišerio klasifikacijos taisyklė. j-tajai klasei skaičiuojama Fišerio klasifikacijos funkcija $D_j(\boldsymbol{x})$:

$$D_j(\boldsymbol{x}) = b_{j0} + b_{j1}x^{(1)} + b_{j2}x^{(2)} + \ldots + b_{jp}x^{(p)}, \qquad (15.6)$$

$j = 1, 2 \ldots k$. Koeficientai b_{j0}, b_{j1}, b_{j2} ... b_{jp} priklauso nuo j-tosios klasės kintamųjų vidurkių vektoriaus $\bar{\boldsymbol{x}}_j$, apriorinės tikimybės π_j ir bendros kovariacijų matricos įverčio $\hat{V}$:

$$D_j(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_j \hat{V}^{-1})\boldsymbol{x}^{\mathrm{T}} - 0{,}5\, \bar{\boldsymbol{x}}_j \hat{V}^{-1} \bar{\boldsymbol{x}}_j^{T} + \ln \pi_j ;$$

čia T – vektoriaus transponavimo ženklas. Klasifikavimo taisyklė: elementas $\boldsymbol{x}$ priskiriamas tai klasei, kurios Fišerio klasifikacijos funkcija (15.6) yra didžiausia.

$\bar{\boldsymbol{x}}_j$ ir bendra kovariacijų matrica $\hat{V}$ apskaičiuojama naudojant pradinės klasifikacijos duomenis. Skaičiuojama statistiniais paketais: duomenų byloje nurodomi kintamieji, atitinkantys vektoriaus $\boldsymbol{X}$ koordinates $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$, bei kintamasis, koduojantis klases. Apriorinės tikimybės π_j gali būti proporcingos pradinių imčių dydžiui, vienodos visoms klasėms bei laisvai parenkamos. Statistiniai paketai (SPSS, STATISTICA, SAS) pateikia Fišerio diskriminantinių funkcijų koeficientus b_{j0}, b_{j1}, b_{j2} ... b_{jp}, klasifikavimo rezultatų palyginimą su faktine klasifikacija bei kiekvieno individo klasifikaciją ir klasifikacijos tikimybes $P(\Pi_j|\boldsymbol{x})$, normaliojo skirstinio atveju išreiškiamas per Fišerio diskriminantines funkcijas:

$$P(\Pi_j \mid x) = \frac{\pi_j \exp(D_j(x))}{\sum_{i=1}^{k} \pi_i \exp(D_i(x))} .$$

Statistiniai paketai taip pat pateikia pradinę kiekvieno individo klasifikaciją.

Klasifikacijai naudojami tik informatyvūs kintamieji, tačiau, sudarius Fišerio klasifikacijos funkcijas, būtina įsitikinti, ar reikalingi visi kintamieji. Gali būti, kad diskriminantinės funkcijos koeficientai prie $X^{(j)}$ b_{ij}, $i = 1 \ldots k$, nuo 0 reikšmingai nesiskiria, t. y. į diskriminantinę funkciją įtraukus $X^{(j)}$, klasifikacija nepagerės. Koeficientų b_{ij}, $i = 1 \ldots k$, reikšmingumui tikrinti skaičiuojama Vilkso (*Wilks*) statistika $\Lambda_j = w_{jj}/t_{jj}$, čia w_{jj} ir t_{jj} – vidinės

tarpusavio sandaugos ir visos tarpusavio sandaugos matricos (15.3 skyrius) diagonaliniai elementai. Kuo Λ_j artimesnė 1, tuo $X^{(j)}$ naudingesnis klasifikacijai. Vieno kintamojo atveju Vilkso statistika λ lygi: $\lambda = SSW/SST$; čia *SSW* – tarpgrupinė kvadratų suma, *SST* – visa kvadratų suma dispersinėje analizėje, *SSW* ir *SST* apskaičiuotos naudojant pirminės klasifikacijos duomenis. Statistiniuose paketuose pateikiama Vilkso statistika ir jos *p* reikšmė. Pagal šią *p* reikšmę daroma išvada, ar kintamasis pagerina klasifikaciją: jei kintamojo $X^{(j)}$ Vilkso statistikos Λ_j *p* reikšmė mažesnė už α, galima tvirtinti, kad $X^{(j)}$ klasifikacijos funkcijai reikalingas. Jei $p \geq \alpha$, daroma išvada, kad koeficientas b_{ij} nuo 0 reikšmingai nesiskiria.

Naudojant daugiamatį kintamąjį, galima sudaryti gana daug skirtingų klasifikavimo funkcijų. Optimali klasifikacijos funkcija sudaroma naudojant žingsninę procedūrą, kaip ir daugialypėje regresijoje, tik čia vietoj F kriterijaus *p* reikšmės naudojama Vilkso statistikos *p* reikšmė.

15.7. Klasifikacija taikant logistinę ir polinominę regresiją

Logistinės regresijos naudojimas klasifikacijai. Turime daugiamačio kintamojo $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(p)})$ imtis iš 2 populiacijų; $X^{(j)}$ gali būti tiek kiekybinis, tiek kokybinis. Abiejų populiacijų individams apibrėžkime dvinarį kintamąjį *Y*: $Y = 1$, jei individas iš I populiacijos; $Y = 0$, jei individas iš II populiacijos. Laikydami, kad *Y* yra atsitiktinis dydis, o $X^{(1)}, X^{(2)} \ldots X^{(p)}$ – neatsitiktiniai, tikimybę $P\{Y = 1\}$ įvertiname logistiniu modeliu (11.6):

$$P\{Y = 1|\boldsymbol{x}\} = \frac{\exp(g(\boldsymbol{x}))}{1+\exp(g(\boldsymbol{x}))} \text{ ir } P\{Y = 0|\boldsymbol{x}\} = \frac{1}{1+\exp(g(\boldsymbol{x}))};$$

čia $g(\boldsymbol{x}) = b_0 + b_1x^{(1)} + \ldots + b_px^{(p)}$.

Tikimybės $P\{Y = 1|\boldsymbol{x}\}$ ir $P\{Y = 0|\boldsymbol{x}\}$ yra sąlyginių tikimybių, kad individas su reikšme $\boldsymbol{x}$ atitinkamai priklauso I ir II klasei, įvertis. Todėl individą su reikšme $\boldsymbol{x}$ priskirsime I klasei, jei $P\{Y = 1|\boldsymbol{x}\} \geq P\{Y = 0|\boldsymbol{x}\}$ arba $P\{Y = 1|\boldsymbol{x}\} \geq ½$. Jei dėl neteisingos klasifikacijos patirti nuostoliai $C(2|1)$ ir $C(1|2)$ yra skirtingi, tada individas su reikšme $\boldsymbol{x}$ priskiriamas I klasei, kai $C(2|1)P\{Y = 1|\boldsymbol{x}\} \geq C(1|2)P\{Y = 0|\boldsymbol{x}\}$, bei priskiriamas II klasei, jei ši nelygybė nėra teisinga.

Polinominės regresijos naudojimas klasifikacijai. Turime daugiamačio kintamojo $\boldsymbol{X} = (X^{(1)}, X^{(2)} \ldots X^{(p)})$ imtis iš *k* populiacijų; $X^{(j)}$ gali būti tiek kiekybinis, tiek kokybinis. Visų populiacijų individams apibrėžkime kokybinį kintamąjį *Y*, lygų populiacijos eilės numeriui –1; taigi *Y* įgyja reikšmes 0, 1 ... $k - 1$. Laikydami, kad *Y* yra atsitiktinis dydis, o $X^{(1)}, X^{(2)} \ldots X^{(p)}$ reikšmės – determinuotos, polinomine regresija įvertiname tikimybes

$P\{Y = 0|\boldsymbol{x}\}$, $P\{Y = 1|\boldsymbol{x}\}$ ... $P\{Y = k - 1|\boldsymbol{x}\}$. Individą su reikšme $\boldsymbol{x}$ ir priskiriame tai klasei, kurią atitinkanti tikimybė didžiausia.

15.8. Diskriminantinės analizės taikymo pavyzdžiai

15.3 pavyzdys. N. Ragaišytė [7] pateikia ligonių, sergančių idiopatine dilatacine (ID), išemine (IŠ) ir hipertenzine (H) kardiomiopatija, klinikinių, elektrokardiografinių, echoskopijos ir perfuzijos tyrimų palyginamuosius duomenis. Tyrime dalyvavo 22 ID, 29 IŠ ir 24 H kardiomiopatija sergantys ligoniai. Naudojant jų tyrimų duomenis, sudarytos Fišerio diskriminantinės funkcijos bei polinominės regresijos tikimybės. Pagal echoskopijos bei perfuzijos rodiklių reikšmes, jomis ligoniai klasifikuojami į sergančius minėtomis kardiomiopatijos rūšimis. Apriorinės tikimybės yra vienodos ir lygios 1/3.

Fišerio klasifikacija. Kintamieji, informatyvūs ligoniams klasifikuoti į 3 minėtas klases, Vilkso statistika, jos p reikšmė bei Fišerio klasifikacijos funkcijų koeficientai pateikti 15.3 lentelėje.

15.3 lentelė. Ligonių klasifikavimo į 3 kardiomiopatijos rūšis rodikliai: Vilkso statistika Λ, jos p reikšmė bei Fišerio klasifikacijos funkcijų koeficientai

Rodiklis	Λ	p	Fišerio klasifikacijos funkcijų koeficientai		
			ID	IŠ	H
Amžius (amz)	0,19	0,009	0,61	0,76	0,74
Tarpskilvelinė pertvara (TSP)	0,22	< 0,001	7,71	8,93	9,43
Sienų judėjimo indeksas dešiniojoje šakoje (SJIDŠ)	0,20	< 0,001	4,22	2,09	0,37
Sienų judėjimo indeksas priekinėje tarpskilvelinėje šakoje (SJIPT)	0,25	< 0,001	14,85	20,83	20,66
Perfuzijos sutrikimo laipsnis dešiniojoje šakoje (PSLDŠ)	0,19	0,004	3,74	5,90	4,61
Perfuzijos sutrikimo plotas priekinėje tarpskilvelinėje šakoje (PSPPT)	0,24	< 0,001	0,77	1,96	–0,40
Konstanta	–	–	–77,43	–112,89	–105,11

Fišerio klasifikacijos funkcijos (15.6) apibrėžiamos taip:

D_1= –77,43 + 0,61×amz + 7,71×TSP + 4,22×SJIDŠ + 14,85×SJIPT + 3,74×PSLDŠ + 0,77×PSPPT;

D_2 = –112,9 + 0,76×amz + 8,93×TSP + 2,09×SJIDŠ + 20,83×SJIPT + 5,9×PSLDŠ + 1,96×PSPPT;

D_3 =–105,1 + 0,74×amz + 9,43×TSP + 0,37×SJIDŠ + 20,66×SJIPT + 4,61×PSLDŠ – 0,4×PSPPT.

Fišerio klasifikacija atliekama taip: naudojant ligonio rodiklius amz, TSP, SJIDŠ, SJIPT, PSLDŠ, PSPPT, apskaičiuojama D_1, D_2, D_3 ir nustatoma didžiausia iš jų. Jei D_1 didesnė už D_2 ir D_3, ligonį priskiriame sergantiems ID kardiomiopatija. Jei didžiausia D_2 – ligonį priskiriame sergantiems išemine, o jei didžiausia D_3 – sergantiems hipertenzine kardiomiopatija.

Klasifikacija polinomine regresija. Kintamieji, įtraukti į daugiamatį polinominės regresijos modelį, koeficientų β_{ij} įverčiai, Valdo statistikos p reikšmė pateikti 15.4 lentelėje.

15.4 lentelė. Ligonių klasifikavimo į 3 kardiomiopatijos rūšis polinominės regresijos modelio rodikliai

Rodiklis	Idiopatinė dilatacinė		Išeminė	
	β	*p*	β	*p*
Amžius (amz)	0,14	0,045	0,137	0,032
Sienų judėjimo indeksas dešiniojoje šakoje (SJIDŠ)	–3,472	0,041	–4,418	0,006
Sienų judėjimo indeksas (SJIPT) priekinėje tarpskilvelinėje šakoje	7,52	0,006	7,216	0,004
Perfuzijos sutrikimo laipsnis dešiniojoje šakoje (PSLDŠ)	2,235	0,038	–0,12	0,86
Perfuzijos sutrikimo plotas dešiniojoje šakoje (PSPDŠ)	1,882	0,055	0,843	0,28
Perfuzijos sutrikimo plotas (PSPPT) priekinėje tarpskilvelinėje šakoje	1,405	0,048	–0,766	0,166
Konstanta	–27,6	–	–11,9	–

Naudodami apskaičiuotus koeficientų įverčius, įvertiname kiekvienos kardiomiopatijos tikimybę:

P{Idiopatinė dilatacinė} = $\exp(g_1)/(1 + \exp(g_1) + \exp(g_2))$;

P{Išeminė} = $\exp(g_2)/(1 + \exp(g_1) + \exp(g_2))$; P{Hipertenzinė} = $1/(1 + \exp(g_1) + \exp(g_2))$;

g_1 = –27,6 + 0,14×amz – 3,472×SJIDŠ + 7,72×SJIPT + 2,235×PSLDŠ + 1,882×PSPDŠ + 1,405×PSPPT;

g_2 = –11,9 + 0,137×amz – 4,418×SJIDŠ + 7,216×SJIPT – 0,12×PSLDŠ + 0,843×PSPDŠ – 0,766×PSPPT.

Ligonio susirgimas priskiriamas tai kardiomiopatijos rūšiai, kurią atitinkanti tikimybė yra didžiausia.

15.5 lentelėje pateikti ligonių klasifikavimo į sergančiuosius ID, IŠ ir H kardiomiopatija rezultatai, gauti Fišerio ir polinominės regresijos metodu.

15.5 lentelė. Ligonių klasifikavimo į sergančiuosius ID, IŠ ir H kardiomiopatija rezultatai, gauti Fišerio ir polinominės regresijos metodu

Kardiomiopatija	Priskirta			Teisingas klasifikacijos %
	ID	IŠ	H	
	Fišerio klasifikacija			
ID	20	2	0	90,9
IŠ	0	28	1	96,6
H	0	4	24	83,3
Iš viso				90,7
	Klasifikacija polinomine regresija			
ID	19	1	2	86,4
IŠ	1	27	1	93,1
H	3	3	19	75,0
Iš viso				85,3

Iš 15.5 lentelės matyti, kad Fišerio klasifikacijos metodu gautas didesnis teisingos klasifikacijos procentas nei polinominės regresijos metodu.

Logistinės regresijos taikymas klasifikacijai. Šiuo metodu minėti ligoniai klasifikuoti į dvi klases: turinčius 50 % ir didesnę VA stenozę bei neturinčius VA stenozės. Vienmatės logistinės regresijos metodu nustatyta, kad VA stenozės tikimybei turėjo įtakos ligonio amžius, krūtinės angina, perfuzijos sutrikimo laipsnis ir perfuzijos sutrikimo plotas visose VA šakose. Naudojant šiuos kintamuosius, žingsniniu metodu sudarytas daugiamatės logistinės regresijos modelis, skirtas įvertinti ligonio VA susiaurėjimo ≥ 50 % tikimybei:

$$P\{\text{ligonio VA susiaurėjimas} \geq 50\,\%\} = \exp(13{,}35 + 2{,}55\times\text{PSLDŠ} + 1{,}37\times\text{PSPDŠ} + 1{,}73\times\text{PSPT}\}/(1 + \exp(13{,}35 + 2{,}55\times\text{PSLDŠ} + 1{,}37\times\text{PSPDŠ} + 1{,}73\times\text{PSPT}\}),$$

$G = 61{,}7$; $p < 0{,}0001$.

Ligoniui prognozuojamas VA susiaurėjimas, kai $P > 0{,}5$. Iš 30 tirtų ligonių, turinčių VA susiaurėjimą, 27 tikimybė P\{ligonio VA susiaurėjimas ≥ 50 %\} viršijo 0,5; iš 47 ligonių, neturinčių VA susiaurėjimo, 43 ligonių P buvo mažesnė už 0,5. Taigi teisingas klasifikavimo procentas – 90,91 %. Prognozuodami VA susiaurėjimą logistiniu modeliu, gavome 90 % jautrumą, 91,5 % specifiškumą, 87,1 % teigiamą prognostinę vertę.

15 skyriaus literatūra

1. Аффи А., Эйзен С. *Статистический анализ*. Подход с использованием ЭВМ. 1982. Москва: Мир, 488 с.
2. Айвазиан С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. *Прикладная статистика. Класификация и снижение размерности*. 1989. Москва: Финансы и статистика, 607 с.
3. Bagdonavičius V., Kruopis J. *Matematinė statistika*. I dalis. 2007. Vilnius, 359 p.
4. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
5. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
6. Manly B. F. *Multivariate Statistical Methods*. Second edition. 1994, 215 p.
7. Ragaišytė N. *Sergančiųjų idiopatine dilatacine, išemine ir hipertenzine kardiomiopatija palyginamieji duomenys*. Daktaro disertacija. 2003. Kaunas.
8. Prieiga per internetą: http://www.psych.umn.edu/courses/spring05/federicoc/psy8815/lectures/stats_lecture11_reading.pdf.
   Carey G. MANOVA. 1998.

16 SKYRIUS

Kiti daugiamačiai statistikos metodai

16.1. Pagrindinių komponenčių analizė

Aplinkos taršai ištirti dvidešimtyje miestų paimti dirvožemio mėginiai. Juose nustatyta 11 cheminių junginių koncentracija (μg/g). Tarp daugelio šių rodiklių stebėtos reikšmingos koreliacijos. Tolesnei duomenų analizei bei rezultatų interpretavimui vietoj šių 11 patogu turėti 2 ar 3 rodiklius, atspindinčius visą gautų duomenų kaitą. Šiam tikslui naudotinas vienas daugiamačių statistikos metodų – pagrindinių komponenčių analizė.

Sakykime, individą charakterizuoja daug koreliuotų rodiklių $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$, turinčių bendrą daugiamatį normalųjį skirstinį su nuliniu vidurkių vektoriumi. Pagrindinių komponenčių analizės tikslas – taip sudaryti naujus kintamuosius $Z^{(1)}$, $Z^{(2)}$ ... $Z^{(p)}$, kad:

- $Z^{(1)}$, $Z^{(2)}$ ... $Z^{(p)}$ būtų nekoreliuoti;
- $Z^{(1)}$, $Z^{(2)}$ ... $Z^{(p)}$ būtų surikiuoti pagal dispersijų dydį: $Z^{(1)}$ dispersija didžiausia, $Z^{(2)}$ dispersija pagal dydį antra ir t. t. – t. y. $DZ^{(1)} \geq DZ^{(2)} \geq ... \geq DZ^{(p)}$;
- $DX^{(1)} + DX^{(2)} + ... + DX^{(p)} = DZ^{(1)} + DZ^{(2)} + ... + DZ^{(p)}$.

Kintamieji $Z^{(i)}$, $i = 1, 2 ... p$ vadinami pagrindinėmis komponentėmis. $Z^{(i)}$ išreiškiamas kintamųjų $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ tiesine kombinacija:

$Z^{(i)} = a_{i1}X^{(1)} + a_{i2}X^{(2)} + \ldots + a_{ip}X^{(p)}$, $i = 1, 2 \ldots p$;

čia koeficientai $a_{i1}, a_{i2} \ldots a_{ip}$ tenkina sąlygą $a_{i1}^2 + a_{i2}^2 + ... + a_{p}^2 = 1$, $i = 1, 2 ... p$ bei parinkti taip, kad $Z^{(1)}$ ... $Z^{(p)}$ būtų nekoreliuoti ir išdėstyti dispersijų $Z^{(i)}$ didėjimo tvarka.

Nustatyta, kad koeficientai a_{ij} priklauso tik nuo $X^{(1)}$ ... $X^{(p)}$ kovariacijų matricos V. $Z^{(1)}$, $Z^{(2)}$ ... $Z^{(p)}$ dispersijos yra lygios kovariacijų matricos V tikriniams skaičiams λ_1, λ_2 ... λ_p, surikiuotiems mažėjimo tvarka: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$. i-toji pagrindinė komponentė sąlygoja $100\lambda_i/(\lambda_1 + \lambda_2 + ... + \lambda_p)$ procentų bendro duomenų kitimo. Koeficientų vektorius (a_{i1}, a_{i2} … a_{ip}) yra matricos V tikrinis vektorius, atitinkantis tikrinį skaičių λ_i. Pirmoji pagrindinė komponentė rodo, kokia $X^{(1)}$ ... $X^{(p)}$ tiesinė kombinacija labiausiai kinta. Paskutinės pagrindinės komponentės dažnai kinta gana nežymiai. Todėl bendrą duomenų kitimą charakterizuojanti p kintamųjų dispersijų suma $DX^{(1)} + DX^{(2)} + ... + DX^{(p)} = \lambda_1 + \lambda_2 + ... + \lambda_p$ nedaug skiriasi nuo $\lambda_1 + \lambda_2 + ... + \lambda_q - q$ pirmųjų pagrindinių komponenčių dispersijų sumos; čia $q < p$. Taigi tą patį duomenų kitimą paaiškina mažesnis kintamųjų skaičius q (vietoje p). Kintamieji $Z^{(1)}$ ... $Z^{(p)}$ nėra koreliuoti – todėl šiuos rodiklius patogu naudoti regresinėje analizėje kaip faktorius. Be to, individų tarpusavio išsidėstymą galima pateikti grafiškai pirmomis 2 ar 3 pagrindinėmis komponentėmis – plokštumoje ar erdvėje.

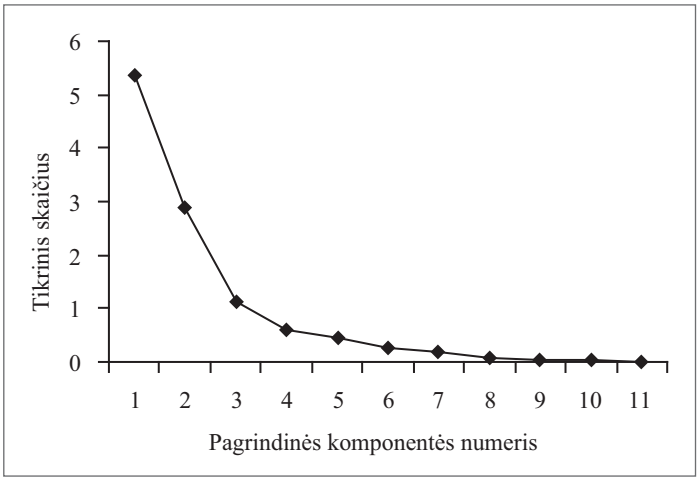
Pagrindinių komponenčių metodas efektyviausias tada, kai visi $X^{(j)}$ yra tos pačios fizikinės prigimties ir yra matuojami tais pačiais vienetais. Jei dydžių $X^{(j)}$ fizikinė prasmė skirtinga (pvz., ilgis ir laikas), tuomet tikslinga $X^{(j)}$ normuoti – dalyti iš standartinio nuokrypio. Taip išvengiama mastelių skirtumo. Normuotų dydžių kovariacijų matrica virsta koreliacijų matrica.

Praktiškai pagrindinės komponentės skaičiuojamos taip: sąlyga, kad $X^{(j)}$ turi lygų nuliui vidurkį, pasiekiama kintamuosius centruojant – vietoj reikšmių $x_i^{(j)}$ naudojamos reikšmės $\tilde{x}_i^{(j)} = x_i^{(j)} - \bar{x}^{(j)}$; čia $\bar{x}^{(j)}$ – kintamojo $X^{(j)}$ imties vidurkis. Jei $X^{(j)}$ nėra to paties mastelio, reikšmės $x_i^{(j)}$, be centravimo, dar ir normuojamos, t. y. pakeičiamos $\tilde{x}_i^{(j)} = (x_i^{(j)} - \bar{x}^{(j)})/s_j$; čia s_j – kintamojo $X^{(j)}$ imties standartinis nuokrypis. Po to statistiniu paketu apskaičiuojama sudarytų kintamųjų kovariacijų ar koreliacijų matrica ir jos tikriniai skaičiai λ_1, λ_2 ... λ_p. Kiekvienam tikriniam skaičiui nustatomas tikrinis vektorius, randami koeficientai a_{ij} bei apskaičiuojamos pagrindinės komponentės $Z^{(j)}$. Pagrindinių komponenčių, skirtų tolesnei analizei, skaičius q nustatomas pagal tai, kokią dalį duomenų kitimo norima paaiškinti pagrindinėmis komponentėmis, t. y. kokią dalį $DX^{(1)} + DX^{(2)} + ... + DX^{(p)}$ sudaro $\lambda_1 + \lambda_2 + ... + \lambda_q$. 16.1 lentelėje ([6, 209 p.) pateikta 11 cheminių junginių, nustatytų dirvožemio mėginiuose, koncentracijos koreliacijų matricos tikriniai skaičiai ir kiekvienos pagrindinės komponentės (PC) sąlygojamas kitimo procentas. Matome, kad pirmoji PC sąlygoja 48,7 %, antroji – 26,2 %, trečioji – 10,2 % bendro kitimo; pirmos trys pagrindinės komponentės sąlygoja 85,2 % bendro duomenų kitimo.

16.1 lentelė. 11 cheminių junginių koncentracijos koreliacijų matricos tikriniai skaičiai ir kiekvienos pagrindinės komponentės (PC) sąlygojamas kitimo procentas

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Tikriniai skaičiai	5,36	2,887	1,125	0,612	0,440	0,261	0,179	0,06	0,046	0,026	0,003
Kitimas (%)	48,7	26,2	10,2	5,6	4,0	2,4	1,6	0,5	0,4	0,2	0,0
Suminis kitimas (%)	48,7	75,0	85,2	90,8	94,8	97,1	98,8	99,3	99,7	100	100

Iš 16.1 lentelės matome, kad paskutinės penkios PC sąlygoja mažiau nei 3 % duomenų kitimo. PC skaičius, reikalingas paaiškinti visam duomenų kitimui, nustatomas pagal tikrinių skaičių grafiką (*scree plot*). Šis grafikas brėžiamas taip: *X* ašyje atitedamas PC eilės numeris, *Y* ašyje – atitinkamas tikrinis skaičius; gauti taškai sujungiami (16.1 pav.). Iš grafiko nustatoma, nuo kurio tikrinio skaičiaus kreivė pradeda kisti tiesiškai. Šis skaičius ir nurodo, kiek pagrindinių komponenčių naudotina. 16.1 pav. nuo ketvirto tikrinio skaičiaus kreivė mažėja tiesiškai, todėl šiuo atveju duomenų kitimui paaiškinti naudotinos keturios PC.



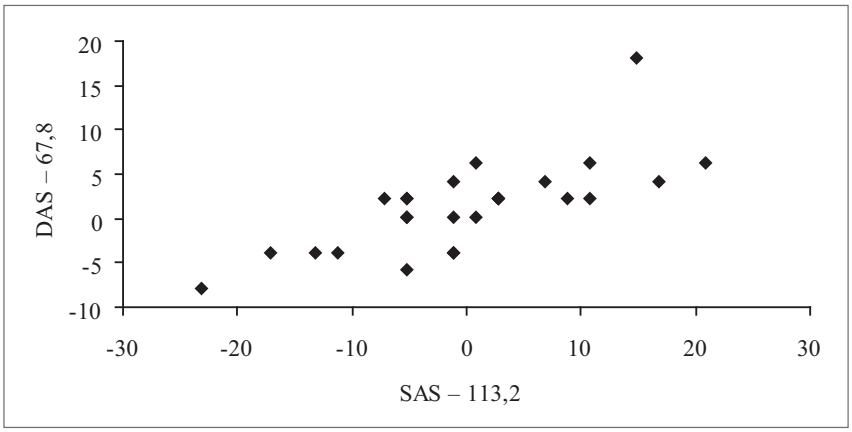
16.1 pav. 11 cheminių junginių koncentracijos koreliacijų matricos tikrinių skaičių grafikas

Pateiksime pagrindinių komponenčių skaičiavimo pavyzdį.

16.1 pavyzdys. 3.4 lentelėje pateiktos 26 jaunų sveikų suaugusių asmenų SAS ir DAS reikšmės (mmHg st.). Šių rodiklių skaidos diagrama pateikta 9.1 pav. Sakykime, (SAS, DAS) skirstinys yra dvimatis normalusis. Centruotų SAS ir

DAS reikšmių SAS1 = SAS – 113,2 ir DAS1 = DAS – 67,8 skaidos diagrama pateikta 16.2 pav. (SAS, DAS) kovariacijų matrica lygi:

$$V = \begin{pmatrix} 105{,}67 & 38{,}7 \\ 38{,}7 & 26{,}3 \end{pmatrix}$$



16.2 pav. Centruotų SAS ir DAS reikšmių skaidos diagrama

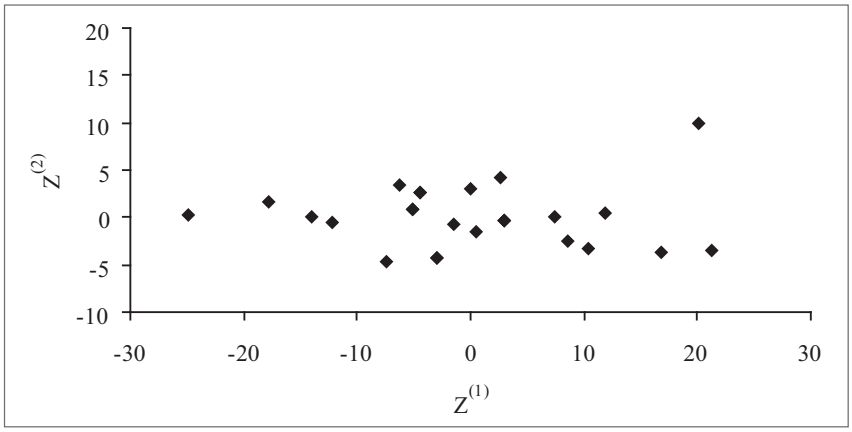
SAS ir DAS bendrą kitimą vertinanti dispersijų suma lygi 105,67 + 26,3 = 131,97. Kovariacijų matricos tikriniai skaičiai lygūs: $\lambda_1 = 121{,}42$ ir $\lambda_2 = 10{,}55$. $\lambda_1 + \lambda_2 = 131{,}97$; λ_1 sudaro 92 %, λ_2 – 8 % bendro kitimo. Šiuos tikrinius skaičius atitinkantys tikriniai vektoriai atitinkamai lygūs (0,926; 0,377) ir (–0,377; 0,926). Todėl (SAS1, DAS1) pagrindinės komponentės lygios:

$Z^{(1)} = 0{,}926$ SAS1 + 0,377 DAS1; $Z^{(2)} = -0{,}377$ SAS1 + 0,926 DAS1.

Pagrindinių komponenčių skaidos diagrama pateikta 16.3 pav. Iš jos matome, kad $Z^{(1)}$ ir $Z^{(2)}$ nėra koreliuoti. Pirmoji pagrindinė komponentė sąlygoja 92 %, antroji – 8 % bendro (SAS, DAS) kitimo.

Geometriškai dvimačių duomenų atveju $Z^{(1)}$ ir $Z^{(2)}$ galime interpretuoti taip: koordinačių ašys taip pasukamos, kad $Z^{(1)}$ ašis eitų per pagrindinę išsibarstymo elipsoido ašį. Kuo koreliacijos koeficientas tarp $X^{(1)}$ ir $X^{(2)}$ artimesnis ± 1, tuo mažiau $Z^{(2)}$ kinta.

Jei kintamieji normuojami, t. y. pagrindinės komponentės nustatomos naudojant koreliacijų matricos tikrinius skaičius ir vektorius, tada koeficientai a_{ij} yra koreliacijos koeficientai tarp $X^{(i)}$ ir $Z^{(j)}$. Pagal a_{1j}, a_{2j} ... a_{pj} reikšmes pateikiama pagrindinės komponentės $Z^{(j)}$ interpretacija. Laikoma, kad $Z^{(j)}$ susijusi su tais kintamaisiais $X^{(i)}$, kuriuos atitinkantys koeficientai a_{ij} viršija 0,4.



16.3 pav. Pagrindinių komponenčių skaidos diagrama

16.2. Faktorinė analizė

Analizuojant didelį kiekį (nuo kelių dešimčių iki kelių šimtų) rodiklių, atspindinčių individo sveikatos būklę, atsiranda poreikis šiuos duomenis susisteminti – sudaryti keletą išvestinių rodiklių, atspindinčių vieną ar kitą organizmo veiklos aspektą. Pavyzdžiui, tiriant individo nusiskundimus, žinoma, kad vieni jų susiję su širdies–kraujagyslių sistemos būkle, kiti – su nervų sistemos būkle ir t. t. Tarp rodiklių, nusakančių tą pačią organizmo patologiją, pavyzdžiui, širdies veiklos nepakankamumą (*Kilip* klasė, išstūmimo frakcija ir pan.) stebimas statistinis ryšys. Kiekybinių rodiklių atveju tai – reikšminga koreliacija, tvarkos kintamųjų atveju – reikšmingas kontingencijos koeficientas. Vienas statistinių metodų, skirtų paaiškinti šį ryšio buvimą, yra faktorinė analizė.

Faktorinės analizės idėja – kintamųjų reikšmes sąlygoja keli nestebimi (latentiniai) faktoriai, atspindintys tam tikrą organizmo veiklos aspektą. Nestebimų faktorių buvimas paaiškina statistinį ryšį tarp kintamųjų.

Faktorinė analizė padeda:

- nustatyti šių nestebimų faktorių skaičių;
- faktorius paaiškinti ir interpretuoti;
- įvertinti individo faktorių reikšmes.

Duomenims analizuoti naudojami įvairūs faktorinės analizės metodai. Pateiksime ortogonalių faktorių tikimybinį modelį.

Sakykime, $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ – kiekybiniai kintamieji, turintys bendrą daugiamatį normalųjį skirstinį. Kintamieji $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ yra koreliuoti. Vienas

metodų, paaiškinančių šias koreliacijas, yra prielaida, kad kintamojo $X^{(j)}$, $j = 1 \ldots p$, reikšmę sąlygoja keli nestebimi, nekoreliuoti, vienetines dispersijas turintys faktoriai $f_1, f_2, .., f_q$, $q < p$. Struktūriškai šį modelį galima išreikšti taip:

$$\begin{aligned} X^{(1)} &= c_{11}f_1 + c_{12}f_2 + \ldots + c_{1q}f_q + e_1; \\ X^{(2)} &= c_{21}f_1 + c_{22}f_2 + \ldots + c_{2q}f_q + e_2; \\ \ldots\ \ldots\ &\ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ X^{(p)} &= c_{p1}f_1 + c_{p2}f_2 + \ldots + c_{pq}f_q + e_p; \end{aligned} \qquad (16.1)$$

arba matricos pavidalu: $\boldsymbol{X} = C\boldsymbol{F} + \boldsymbol{E}$;

čia $f_1, f_2 \ldots f_q$ – nestebimi (latentiniai) faktoriai, $e_1, e_2 \ldots e_p$ – nepriklausomi normalieji ats. d., c_{ij} – faktorių svoriai, $\boldsymbol{X}$, $\boldsymbol{F}$, $\boldsymbol{E}$ – kintamųjų $X^{(i)}$, faktorių f_i ir e_i vektorius-stulpelis, C – $p \times q$ svorių matrica. Šiame faktorinės analizės modelyje daroma prielaida, kad ats. d. $f_1 \ldots f_q$ nekoreliuoti, turi lygius 0 vidurkius ir vienetines dispersijas; e_i ir f_j taip pat nekoreliuoti.

Remiantis šiuo $X^{(j)}$ struktūriniu modeliu, vektoriaus ($X^{(1)}$, $X^{(2)} \ldots X^{(p)}$) kovariacijų matricos $V = (v_{ij})$ elementai taip išreiškiami faktorių svoriais:

$$v_{ij} = c_{i1}c_{j1} + c_{i2}c_{j2} + \ldots + c_{iq}c_{jq},\ i \neq j, \quad v_{ii} = c_{i1}^2 + c_{i2}^2 + \ldots + c_{iq}^2 + \tau_i, \qquad (16.2)$$

$i = 1, 2 \ldots p$; čia $\tau_i - e_i$ dispersija. Matome, kad šiame faktorinės analizės modelyje ats. d. $X^{(i)}$ dispersija v_{ii} išskaidyta į dvi dalis:

- dalį, sąlygotą nestebimų faktorių $f_1, f_2, \ldots, f_q$ – $h_i^2 = c_{i1}^2 + c_{i2}^2 + \ldots + c_{iq}^2$;
- dalį, sąlygotą atsitiktinio kitimo τ_i.

Dydis h_i^2 vadinamas kintamojo $X^{(i)}$ bendrumu (*communality*), dydis τ_i – specifiškumu (*specific variance*). Matrica, kurios diagonaliniai elementai yra h_i^2, $i = 1, 2 \ldots n$, o likusieji – kovariacijų matricos V elementai, vadinama redukuota kovariacijų matrica. Pagal τ_i reikšmes sprendžiama, ar išskirtas pakankamas faktorių kiekis ir ar korektiškai įvertinti faktorių svoriai.

Sistema (16.2) turi be galo daug sprendinių, kai $q > (p - 1)/2$. Jei $q = (p - 1)/2$, sistema (16.2) turi vienintelį sprendinį. Jei $q < (p - 1)/2$, (16.2) sistema gali sprendinių ir neturėti, nes joje lygčių yra daugiau nei nežinomųjų (lygčių $p(p + 1)/2$, nežinomųjų – $(q + 1)p$: nežinoma $q \times p$ faktorių svorių ir p ats. d. e_i dispersijų).

Išraiškoje (16.1) faktoriai $f_1, f_2 \ldots f_q$ nėra vieninteliai. Sakykime, Q – bet kuri $q \times q$ ortogonali matrica. Tuomet (16.1) išraišką pertvarkius matricos pavidalu, turime:

$$\boldsymbol{X} = C\boldsymbol{F} + \boldsymbol{E} = CQQ^T\boldsymbol{F} + \boldsymbol{E} = (CQ)(Q^T\boldsymbol{F}) + \boldsymbol{E} = C^*\boldsymbol{F}^* + \boldsymbol{E},$$

čia C^* – nauja svorių matrica, $\boldsymbol{F}^*$ – naujas faktorių vektorius. Matrica C^* tenkina (16.2) lygčių sistemą, nes $C^*C^{*T} = CQQ^TC^T = CC^T$. Transformacija $\boldsymbol{F}^* = Q^T\boldsymbol{F}$ vadinama faktorių pasukimu. Dviejų faktorių atveju pasukamos koordinačių ašys plokštumoje:

$$f_1^* = q_{11}f_1 + q_{21}f_2;$$
$$f_2^* = q_{12}f_1 + q_{22}f_2.$$

Faktorinei analizei naudotinus duomenis rekomenduojama standartizuoti – vietoj $x_i^{(j)}$ naudoti $z_i^{(j)} = (x_i^{(j)} - \overline{x}^{(j)})/s_j$. Tuomet faktorių svoriams vertinti naudojama $X^{(1)}, X^{(2)} \ldots X^{(p)}$ koreliacijų matrica.

Duomenims apdoroti taikant faktorinę analizę, skiriami šie analizės etapai:

1. Tikrinama, ar duomenys tinka faktorinei analizei.
2. Parenkamas faktorių skaičius.
3. Nustatomi pradiniai faktorių svoriai.
4. Faktoriai pasukami.
5. Faktoriai interpretuojami.
6. Skaičiuojamos faktorių reikšmės.

Faktorinė analizė atliekama statistiniu paketu, dažniausiai STATISTICA ar SPSS. Aptarsime statistikos metodus, naudojamus atskirais faktorinės analizės etapais.

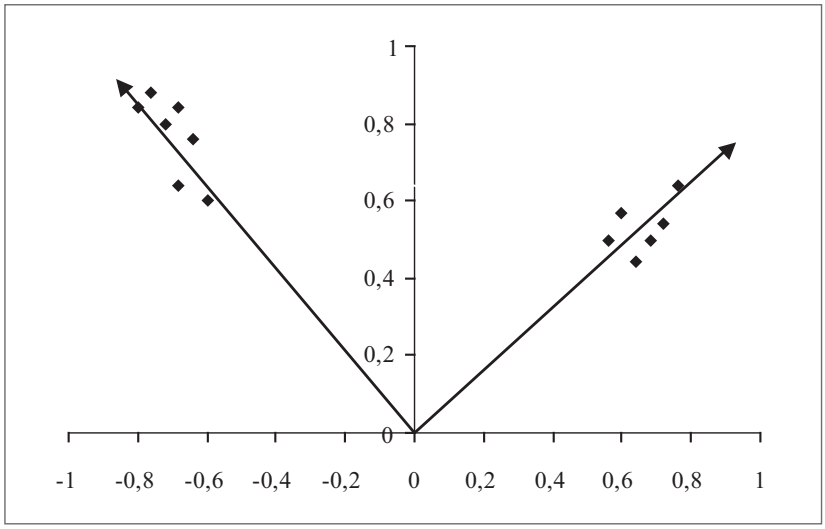
1. Taikant faktorinę analizę būtina, kad duomenys būtų koreliuoti. Duomenų tinkamumas vertinamas Kaizerio–Mejerio–Olkino (*Kaiser–Meyer–Olkin*) (KMO) matu, išreiškiamu per $X^{(i)}$ ir $X^{(j)}$ koreliacijas. Jei KMO > 0,8, faktorinė analizė tinka gerai; jei 0,6 < KMO ≤ 0,8 – tinka pakankamai; jei KMO ≤ 0,6 – faktorinės analizės geriau netaikyti.

2. Faktorių skaičius parenkamas pagal duomenų pobūdį arba faktorių imama tiek, kiek koreliacijos tikrinių skaičių viršija 1. Faktorių skaičiui nustatyti taip pat naudojama tikėtinumų santykio statistika ([3]).

3. Pradiniam faktorių svorių vertinimui dažniausiai naudojamas pagrindinių komponenčių metodas – pradiniais faktoriais laikomos atitinkamai normuotos pirmosios pagrindinės komponentės. Statistiniuose paketuose nurodomi ir kiti pradiniai svorių vertinimo metodai.

4. Šiuo etapu atliekant ortogonalią transformaciją (pasukant koordinačių ašis) randamas faktorių svorių variantas, geriausiai tinkantis jiems interpretuoti. Tai daryti bus lengviau, jei dalis prie kiekvieno faktoriaus esančių svorių reikšmių bus artimos 0, dalis – absoliučiu dydžiu artimos 1 (16.4 pav.) Populiariausias koordinačių pasukimo metodas – VARIMAX. Jo esmė –

koordinačių ašys sukamos taip, kad svorių kvadratų suma prie kiekvieno faktoriaus būtų didžiausia.



16.4 pav. Koordinačių ašių pasukimas

5. Faktoriai interpretuojami taip: peržiūrimi svoriai c_{1i}, c_{2i} ... c_{pi}, esantys prie kiekvieno faktoriaus. *j*-tasis faktorius laikomas susijusiu su tais kintamaisiais, prie kurių jo svoriai c_{ij} absoliučiu dydžiu viršija 0,4 (kartais 0,6). Faktorių interpretacija gana subjektyvi.

6. Faktorių reikšmės vertinamos regresiniais metodais.

Faktorinė analizė buvo plačiai taikyta tiriant Persų įlankos karo, vykusio 1991 m., dalyvių nusiskundimus [1, 4, 7]. Pavyzdžiui, [7] analizuojant nusiskundimus (atsakymus į 19 klausimų), išskirti 3 faktoriai: kognityvinis-psichologinis, jutimų suvokimo sutrikimų (disestezijos) ir vestibuliarinių sutrikimų. Pirmo faktoriaus svoriai, viršijantys 0,6, buvo prie kintamųjų, apibūdinančių nusiskundimus: keistas spindėjimas, depresija, sunkumas susikaupti, nuovargis, nuotaikų kaita, atminties sutrikimai. Šie nusiskundimai – psichologinio pobūdžio, todėl pirmas faktorius ir pavadintas kognityviniu-psichologiniu. Antro faktoriaus svoriai, viršijantys 0,6, buvo prie kintamųjų, nurodančių nusiskundimą sustingimu ir spengimu ausyse; šis faktorius pavadintas disestezijos faktoriumi. Trečio faktoriaus svoriai, viršijantys 0,6, buvo prie nusiskundimų galvos svaigimu ir pusiausvyros sutrikimu. Todėl šis faktorius pavadintas vestibuliarinių sutrikimų faktoriumi.

16.3. Atitikimų analizė

Atitikimų analizė (*correspondence analysis*) skirta grafiškai pateikti porinę dažnių lentelę, įvedus keletą atitikties tarp eilučių ir stulpelių matų. Sakykime, porine dažnių lentele pateikta rūkymo priklausomybė nuo užimamų pareigų (16.2 lentelė) ([9]).

16.2 lentelė. Rūkymo priklausomybė nuo užimamų pareigų

Pareigos	Rūkymo kategorijos				Iš viso
	Nerūko	Rūko retai	Rūko vidutiniškai	Rūko dažnai	
Vyr. specialistas	4	2	3	2	11
Jaun. specialistas	4	3	7	4	18
Vyr. tarnautojas	25	10	12	4	51
Jaun. tarnautojas	18	24	33	13	58
Sekretorė	10	6	7	2	25
Iš viso	61	45	62	25	193

Remiantis surinktais duomenimis, norima įvertinti, kaip skiriasi vienas ar kitas pareigas užimančių darbuotojų rūkymas. 16.2 lentelėje rūkymas pateiktas 4 matavimų kintamuoju. Jei įvairių pareigybių keturmatį rūkymo vektorių transformuotume į dvimatį ar trimatį vektorių, atskirų pareigybių išsidėstymą rūkymo atžvilgiu galėtume pateikti plokštumoje ar erdvėje. Gautą grafiką būtų galima atitinkamai interpretuoti.

Ryšio tarp eilučių ir stulpelių reikšmių alternatyvus matas – χ^2 kriterijaus statistikos išskaidymas (χ^2 *decomposition*). 7.3 skyriuje ryšys tarp dviejų kokybinių rodiklių nustatomas statistika

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}};$$

čia r ir c – atitinkamai kintamojo X ir Y reikšmių skaičius, n_{ij} – individų, turinčių kintamojo X i-tąją ir kintamojo Y j-tąją reikšmę, skaičius, E_{ij} – tikėtini dažniai porinėje dažnių lentelėje. E_{ij} skaičiuojami: $E_{ij} = n_{i+}n_{+j}/n$; čia n_{i+} – skaičius individų, turinčių kintamojo X i-tąją reikšmę, n_{+j} – skaičius individų, turinčių kintamojo Y j-tąją reikšmę, n – individų skaičius porinėje dažnių lentelėje. Jei X ir Y yra nepriklausomi, tada χ^2 statistika turi asimptotinį χ^2 skirstinį su $(r - 1)(c - 1)$ laisvės laipsnių.

χ^2 statistikos skaidymas – tai $r \times c$ matricos C su elementais

$$c_{ij} = (n_{ij} - E_{ij}) / E_{ij}^{1/2}$$

sudarymas. Pažymėkime: R – matricos C rangas; $R \le \min((r-1), (c-1))$; $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_R$ – matricos CC^T tikriniai skaičiai, išdėstyti mažėjimo tvarka. Pagal veiksmų su matricomis taisykles,

$$tr(CC^T) = \lambda_1 + \lambda_2 + ... + \lambda_R = \sum_{i=1}^{r}\sum_{j=1}^{c} c_{ij}^2 = \chi^2.$$

Daugelyje realių duomenų pirmųjų dviejų tikrinių skaičių suma $\lambda_1 + \lambda_2$ nedaug skiriasi nuo χ^2. Tokiu atveju pagrindinė informacija apie kintamųjų X ir Y ryšį yra matricos CC^T ir C^TC tikriniuose vektoriuose, atitinkančiuose du didžiausius tikrinius skaičius λ_1 ir λ_2. Pažymėkime $\boldsymbol{\gamma}_1$ ir $\boldsymbol{\gamma}_2$ – matricos CC^T, o $\boldsymbol{\delta}_1$ ir $\boldsymbol{\delta}_2$ – atitinkamai matricos C^TC pirmieji tikriniai vektoriai. Vektoriai $\boldsymbol{\delta}_1$ ir $\boldsymbol{\delta}_2$ atspindi Y (stulpelių) kitimą, $\boldsymbol{\gamma}_1$ ir $\boldsymbol{\gamma}_2$ – vektoriaus X (eilučių) kitimą.

Norint grafiškai pateikti matricos eilutes, skaičiuojamas koordinačių vektorius:

$$\boldsymbol{r}_k = A^{-1/2}\sqrt{\lambda_k}\gamma_k,\ k = 1, 2;$$

čia A – diagonalinė $r \times r$ matrica su diagonaliniais elementais n_{i+}. i-tąją matricos eilutę (kintamojo X i-tąją reikšmę) koordinačių plokštumoje atitinka taškas (r_{1i}, r_{2i}). Analogiškai – norint grafiškai pateikti matricos stulpelius, skaičiuojamos koordinatės:

$$\boldsymbol{s}_k = B^{-1/2}\sqrt{\lambda_k}\delta_k,\ k = 1, 2;$$

čia B – diagonalinė $c \times c$ matrica su diagonaliniais elementais n_{+j}. j-tąjį matricos stulpelį (kintamojo Y j-tąją reikšmę) koordinačių plokštumoje vaizduoja taškas (s_{1i}, s_{2i}).

16.3 lentelėje pateikti matricos CC^T, sudarytos pagal rūkymo priklausomybės nuo užimamų pareigų duomenis (16.2 lentelė), tikriniai skaičiai ($\chi^2 = 16{,}44$).

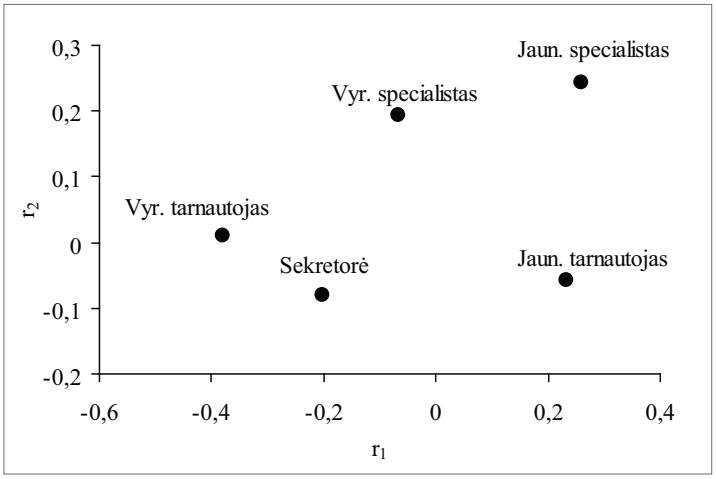
16.3 lentelė. Porinės dažnių lentelės, sudarytos pagal rūkymo priklausomybės nuo užimamų pareigų duomenis, tikriniai skaičiai ($\chi^2 = 16{,}44$)

Nr.	Tikrinis skaičius λ	χ^2 procentas	Suminis χ^2 procentas
1.	14,43	87,76	87,76
2.	1,93	11,75	99,51
3.	0,08	0,49	100,00

Vektoriai $\boldsymbol{r}_1$ ir $\boldsymbol{r}_2$, apskaičiuoti pagal 16.2 lentelės duomenis, lygūs:

$\boldsymbol{r}_1$ = (–0,066; 0,259; –0,381; 0,233; –0,201) ir $\boldsymbol{r}_2$ = (0,194; 0,243; 0,011; –0,058; –0,079).

Įvairių pareigybių darbuotojų išsidėstymas rūkymo atžvilgiu pateiktas 16.5 pav. Vyriausiųjų specialistų rūkymą atitinka taškas (–0,066; 0,194), jaunesniųjų specialistų – (0,259; 0,243) ir t. t.



16.5 pav. Įvairių pareigybių darbuotojų išsidėstymas rūkymo atžvilgiu

16.4. Kanoninė koreliacija

Daugiamačius duomenis pagal jų prigimtį dažnai galima dalyti į dvi grupes. Kanoninė koreliacija skirta ryšiui tarp šių kintamųjų grupių vertinti.

Pateiksime pavyzdį, kaip taikoma kanoninė koreliacija [5, 10 skyrius]. Tirtas drugelių *Euphydryas editha* populiacijų, gyvenančių Kalifornijos ir Oregono (JAV) valstijose, 4 genų pasiskirstymas. 16 kolonijų fiksuoti šie klimato rodikliai: aukštis virš jūros lygio, metinis kritulių kiekis, metinė maksimali ir minimali temperatūra. Nustatytas kiekvienos kolonijos 4 genų dažnis procentais. Norėta įvertinti ryšį tarp klimato rodiklių ir genų pasiskirstymo.

Kanoninė koreliacija yra daugialypės regresijos apibendrinimas. Daugialypėje regresijoje ieškoma faktorių $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ tiesinės kombinacijos, kuri stipriausiai koreliuotų su atsaku Y. Kanoninės koreliacinės analizės tikslas – nustatyti kuo glaudesnį ryšį tarp dviejų kintamųjų grupių tiesinių funkcijų.

Sakykime, dvi kintamųjų grupės $X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$ ir $Y^{(1)}$, $Y^{(2)}$ ... $Y^{(q)}$ apibūdina skirtingos prigimties duomenis, be to, $\boldsymbol{X}$ ir $\boldsymbol{Y}$ skirstiniai yra daugiamačiai normalieji. Kanoninės koreliacinės analizės tikslas – nustatyti kanoninius kintamuosius:

$$U_1 = a_{11}X^{(1)} + a_{12}X^{(2)} + ... + a_{1p}X^{(p)},$$
$$... \quad .... \quad .... \quad ... \quad ... \quad ... \quad .... \quad ... \quad ...$$
$$U_r = a_{r1}X^{(1)} + a_{r2}X^{(2)} + ... + a_{rp}X^{(p)},$$

ir

$$V_1 = b_{11}Y^{(1)} + b_{12}Y^{(2)} + ... + b_{1q}Y^{(q)},$$
$$... \quad .... \quad .... \quad ... \quad ... \quad ... \quad ... \quad .... \quad ...$$
$$V_r = b_{r1}Y^{(1)} + b_{r2}Y^{(2)} + ... + b_{rq}Y^{(q)};$$

čia $r = \min(p, q)$. Šie kanoniniai kintamieji sudaromi taip, kad koreliacija tarp U_1 ir V_1 būtų didžiausia, koreliacija tarp U_2 ir V_2 būtų antra pagal dydį, be to, U_2 ir V_2 nekoreliuotų su U_1 ir V_1, ir t. t.

Kanoninei koreliacijai praktiškai taikyti reikalingi dydžiai U_1 ir V_1, t. y. reikalingi svoriai a_{11} … a_{1p} ir b_{11} … b_{1q}. Pažymėkime A – vektoriaus ($X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$) koreliacijų matrica, B – ($Y^{(1)}$, $Y^{(2)}$ ... $Y^{(q)}$) koreliacijų matrica, C – $p \times q$ matrica, susidedanti iš koreliacijos koeficientų tarp ($X^{(1)}$, $X^{(2)}$ ... $X^{(p)}$) ir ($Y^{(1)}$, $Y^{(2)}$ ... $Y^{(q)}$). Svoriai a_{11} … a_{1p} ir b_{11} … b_{1q} nustatomi taip: skaičiuojamas matricos $B^{-1}C^TA^{-1}C$ didžiausias tikrinis skaičius, jį atitinkantis tikrinis vektorius $\boldsymbol{b}$ bei vektorius $\boldsymbol{a} = A^{-1}C\boldsymbol{b}$. Tuomet vektoriaus $\boldsymbol{a}$ ir $\boldsymbol{b}$ koordinatės ir bus koeficientai a_{11} … a_{1p} ir b_{11} … b_{1q}.

Pateiksime kanoninius kintamuosius, skirtus ryšiui tarp drugelių *Euphydryas editha* genotipo ir populiacijos klimato rodiklių vertinti. Pažymėkime $X^{(1)}$ – aukštį virš jūros lygio, $X^{(2)}$ – metinį kritulių kiekį, $X^{(3)}$ ir $X^{(4)}$ – maksimalią ir minimalią metinę temperatūrą, $Y^{(1)}$ – nuo 0,4 iki 0,6 mobilumo genų dažnį, $Y^{(2)}$ – 0,8 mobilumo genų dažnį, $Y^{(3)}$ – 1,00 mobilumo genų dažnį, $Y^{(4)}$ – 1,16 mobilumo genų dažnį. Didžiausias tikrinis skaičius lygus 0,0743, kanoniniai dydžiai lygūs:

$U_1 = -0{,}09X^{(1)} - 0{,}29X^{(2)} + 0{,}48X^{(3)} + 0{,}29X^{(4)}$; $V_1 = 0{,}54Y^{(1)} + 0{,}42Y^{(2)} - 0{,}1Y^{(3)} + 0{,}82Y^{(4)}$.

Koreliacijos koeficientas tarp U_1 ir V_1 lygus 0,862.

U_1 teigiamai koreliuoja su $X^{(3)}$ ir $X^{(4)}$ (r = 0,9 ir 0,92) ir neigiamai su $X^{(1)}$ ir $X^{(2)}$ (r = –0,92 ir –0,77). Todėl U_1 galima interpretuoti kaip aukštos temperatūros ir žemo aukščio virš jūros lygio bei nedidelio kritulių kiekio matą. V_1 koreliacija su $Y^{(1)}$ lygi 0,38, su $Y^{(2)}$ – 0,74, su $Y^{(3)}$ – –0,96 ir su $Y^{(4)}$ – 0,49. Taigi V_1 aiškiai rodo 1,00 mobilumo geno stoką. U_1 ir V_1 ryšį galima interpretuoti taip: populiacijoje, kurios aplinkoje vyrauja didelis skirtumas tarp temperatūros ir aukščio virš jūros lygio, stebimas mažas 1,00 mobilumo geno dažnis. Populiacijos, kuri gyvena aukštesniame lygyje virš jūros ir žemesnėje aplinkos temperatūroje, stebimas didesnis 1,00 mobilumo genų dažnis.

16 skyriaus literatūra

1. Bourdette D. N., McCauley L. A., Barkhuizen A. et. al. Symptom factor analysis, clinical findings, and functional status in a population-based case control study of Gulf War unexplained illness. 2001. *Journal of Occupational and Environmental Medicine*, 43, p. 1026–1040.
2. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai*. II dalis. 2002. Vilnius: TEV, 272 p.
3. Hardle W., Similar L. *Applied Multivariate Statistical Analysis*. 2003. Prieiga per internetą: http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf.
4. Knoke J. D., Smith T. C., Gray G. C. et al. Factor analysis of self-reported symptoms: does it identify a Gulf War syndrome? 2000. *Am.J. Epidemiol*, 152, p. 379–388.
5. Manly B. F. *Multivariate Statistical Methods*. Second edition. 1994, 215 p.
6. Townend J. *Practical Statistics for Environmental and Biological Scientists*. 2006. New York. John Wiley & Sons, 276 p.
7. Shapiro S. E., Lasarev M. R., McCauley L. Factor Analysis of Gulf War Illness: What Does It Add to Our Unterstanding of Possible Health Effects of Deployment. 2002. *Am. J. Epidemiol*, 156, p. 578–585.
8. *Faktorinė analizė*. Prieiga per internetą: http://www.spss.com/tech/stat/algorithms/7.5/factor.pdf.
9. *Atitikimų analizė*. Prieiga per internetą: http://www.statsoft.com/textbook/stcoran.html.

Lentelės

1 lentelė. Standartinio normaliojo skirstinio α lygmens kvantiliai z_α

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
0,61	0,279	0,72	0,583	0,83	0,954	0,94	1,555	0,978	2,014	0,989	2,290
0,62	0,305	0,73	0,613	0,84	0,994	0,95	1,645	0,979	2,034	0,990	2,326
0,63	0,332	0,74	0,643	0,85	1,036	0,96	1,751	0,980	2,054	0,991	2,366
0,64	0,358	0,75	0,674	0,86	1,080	0,97	1,881	0,981	2,075	0,992	2,409
0,65	0,385	0,76	0,706	0,87	1,126	0,971	1,896	0,982	2,097	0,993	2,457
0,66	0,412	0,77	0,739	0,88	1,175	0,972	1,911	0,983	2,120	0,994	2,512
0,67	0,440	0,78	0,772	0,89	1,227	0,973	1,927	0,984	2,144	0,995	2,576
0,68	0,468	0,79	0,806	0,90	1,282	0,974	1,943	0,985	2,170	0,996	2,652
0,69	0,496	0,80	0,842	0,91	1,341	0,975	1,960	0,986	2,197	0,997	2,748
0,70	0,524	0,81	0,878	0,92	1,405	0,976	1,977	0,987	2,226	0,998	2,878
0,71	0,553	0,82	0,915	0,93	1,476	0,977	1,995	0,988	2,257	0,999	3,090

2 lentelė. χ^2 skirstinio α lygmens kvantiliai $\chi^2_\alpha(n)$

n\α	0,01	0,025	0,05	0,95	0,975	0,99
1	0,0002	0,0010	0,0039	3,841	5,024	6,635
2	0,0201	0,0506	0,1026	5,991	7,378	9,210
3	0,115	0,216	0,352	7,815	9,348	11,345
4	0,297	0,484	0,711	9,488	11,143	13,277
5	0,554	0,831	1,145	11,070	12,832	15,086
6	0,872	1,237	1,635	12,592	14,449	16,812
7	1,239	1,690	2,167	14,067	16,013	18,475
8	1,647	2,180	2,733	15,507	17,535	20,090
9	2,088	2,700	3,325	16,919	19,023	21,666
10	2,558	3,247	3,940	18,307	20,483	23,209
11	3,053	3,816	4,575	19,675	21,920	24,725
12	3,571	4,404	5,226	21,026	23,337	26,217
13	4,107	5,009	5,892	22,362	24,736	27,688
14	4,660	5,629	6,571	23,685	26,119	29,141

n\α	0,01	0,025	0,05	0,95	0,975	0,99
15	5,229	6,262	7,261	24,996	27,488	30,578
16	5,812	6,908	7,962	26,296	28,845	32,000
17	6,408	7,564	8,672	27,587	30,191	33,409
18	7,015	8,231	9,390	28,869	31,526	34,805
19	7,633	8,907	10,117	30,144	32,852	36,191
20	8,260	9,591	10,851	31,410	34,170	37,566
25	11,524	13,120	14,611	37,652	40,646	44,314
30	14,953	16,791	18,493	43,773	46,979	50,892
40	22,164	24,433	26,509	55,758	59,342	63,691
50	29,707	32,357	34,764	67,505	71,420	76,154
60	37,485	40,482	43,188	79,082	83,298	88,379
80	53,540	57,153	60,391	101,879	106,629	112,329
100	70,065	74,222	77,929	124,342	129,561	135,807

3 lentelė. Stjudento skirstinio α lygmens kvantiliai $t_{\alpha}(n)$

n\α	0,9	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878

n\α	0,9	0,95	0,975	0,99	0,995
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

4 lentelė. Fišerio skirstinio 0,95 lygmens kvantiliai

n\m	1	2	3	4	5	6	7	8	9	10	15	20
1	161,5	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	246	248
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,43	19,45
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,703	8,660
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,858	5,803
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,619	4,558
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	3,938	3,874
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,511	3,445
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,218	3,150
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,006	2,936
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,845	2,774
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,719	2,646
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,617	2,544
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,533	2,459
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,463	2,388
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,403	2,328

n\m	1	2	3	4	5	6	7	8	9	10	15	20
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,352	2,276
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,308	2,230
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,269	2,191
19	14,38	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,234	2,155
20	14,35	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,203	2,124
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,015	1,932
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	1,924	1,839
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,836	1,748

5 lentelė. Vilkoksono kriterijaus statistikos α lygio kvantiliai (t(α, n) + t(1 – α, n) = n(n + 1)/2)

n\α	0,025	0,05	0,95	0,975
5		0	15	
6	0	2	19	21
7	2	3	25	26
8	3	5	31	33
9	5	8	37	40
10	8	10	45	47
11	10	13	53	56
12	13	17	61	65
13	17	21	70	74
14	21	25	80	84
15	25	30	90	95

6 lentelė. Binominio skirstinio su p = 0,5 α lygio kvantiliai (B(α, n) + B(1 – α, n) = n)

n\α	0,025	0,05	0,95	0,975
5		1	4	
6	1	1	5	5
7	1	1	5	6
8	1	2	6	7
9	2	2	6	7
10	2	2	7	8
11	2	3	8	9
12	3	3	9	9

n\α	0,025	0,05	0,95	0,975
13	3	4	9	10
14	3	4	10	11
15	4	4	11	11
16	4	5	11	12
17	5	5	12	12
18	5	6	12	13
19	5	6	13	13
20	6	6	14	14
21	6	7	14	15
22	6	7	15	16
23	7	8	15	16
24	7	8	16	17
25	8	8	17	17
26	8	9	17	18
27	8	9	18	19
28	9	10	18	19
29	9	10	19	20
30	10	11	19	20

7 lentelė. U kriterijaus statistikos U skirstinio 0,05 lygio kvantiliai w(0,05, n, m); w(0,95, n, m) = n×m – w(0,05, n, m)

*n**m*	2	3	4	5	6	7	8	9	10	11	12
2				0	0	0	1	1	1	1	2
3		0	0	1	2	2	3	3	4	5	5
4		0	1	2	3	4	5	6	7	8	9
5	0	1	2	4	5	6	8	9	11	12	13
6	0	2	3	5	7	8	10	12	14	16	17
7	0	2	4	6	8	11	13	15	17	19	21
8	1	3	5	8	10	13	15	18	20	23	26
9	1	3	6	9	12	15	18	21	24	27	30
10	1	4	7	11	14	17	20	24	27	31	34
11	1	5	8	12	16	19	23	27	31	34	38
12	2	5	9	13	17	21	26	30	34	38	42

8 lentelė. Pirsono koreliacijos koeficiento tikslaus skirstinio kvantiliai. Tikslus skirstinys yra simetriškas, todėl jo kvantiliams teisinga: $x_{\alpha} = -x_{1-\alpha}$.

n	Lygis		*n*	Lygis	
	0,95	0,975	–	0,95	0,975
3	0,988	0,997	12	0,497	0,576
4	0,900	0,950	13	0,476	0,553
5	0,805	0,878	14	0,457	0,532
6	0,729	0,811	15	0,441	0,514
7	0,669	0,754	16	0,426	0,497
8	0,621	0,707	17	0,412	0,482
9	0,582	0,666	18	0,400	0,468
10	0,549	0,632	19	0,389	0,456
11	0,521	0,602	20	0,378	0,444

9 lentelė. Spirmeno koeficiento statistikos S tikslaus skirstinio $\widetilde{S}$ *kvantiliai. 0,025 ir 0,05 lygio kvantiliai gaunami atėmus atitinkamai 0,975 ir 0,95 lygio kvantilių reikšmes iš* $n(n^2 - 1)/3$

n	Lygis		*n*	Lygis	
–	0,95	0,975	–	0,95	0,975
4	18	–	8	132	138
5	36	38	9	188	196
6	62	64	10	248	268
7	94	98	–	–	–

10 lentelė. Kendalo koeficiento statistikos K tikslaus skirstinio $\widetilde{K}$ *kvantiliai. 0,025 ir 0,05 lygio kvantiliai lygūs atitinkamai 0,975 ir 0,95 lygio kvantilių reikšmėms su minuso ženklu, nes* $\widetilde{K}$ *skirstinys simetriškas*

n	Lygis		*n*	Lygis	
–	0,95	0,975	–	0,95	0,975
4	4	–	8	14	18
5	6	10	9	16	18
6	9	11	10	19	19
7	11	13	–	–	¬

Dalykinė rodyklė

Venclovienė, Jonė
Ve-118 Statistiniai metodai medicinoje, vadovėlis / Jonė Venclovienė. – Kaunas: Vytauto Didžiojo universitetas, 2010. – 344 p.: iliustr.

ISBN 978-9955-12-558-7

Vadovėlyje pateikti ir pavyzdžiais iliustruoti statistikos metodai, skirti medicinos duomenų analizei. Be standartinių statistikos metodų, vadovėlyje pateikti medikams aktualūs metodai – išgyvenamumo analizė, duomenų kokybės, rizikos vertinimas. Pateikti statistikos metodai gali būti taikomi ir kitų biomedicinos sričių duomenims apdoroti. Vadovėlyje paliesti ir kai kurie paskutiniais dešimtmečiais populiarėjantys apibendrinti tiesiniai modeliai, daugiapakopiai modeliai, Puasono regresija, atitikimų analizė, kanoninė koreliacija, todėl jis gali būti naudingas ir taikomosios matematikos specialybių studentams.

UDK 311:61(075.8)

Jonė Venclovienė
STATISTINIAI METODAI MEDICINOJE
Bendrasis vadovėlis aukštosioms mokykloms

Redaktorė Renata Endzelytė
Korektorė Simona Grušaitė
Maketuotoja Janina Baranavičienė

Išleido Vytauto Didžiojo universiteto leidykla
S. Daukanto g. 27, LT-44249 Kaunas