Vydas Čekanavičius
Gediminas Murauskas

STATISTIKA 1

IR JOS TAIKYMAI

Vydas Čekanavičius
Gediminas Murauskas

1

STATISTIKA
IR JOS TAIKYMAI

TEV
VILNIUS 2001

UDK 311(075.8)
Če82

Lietuvos Respublikos švietimo ir mokslo ministerijos Aukštųjų mokyklų bendrųjų vadovėlių leidybos komisijos rekomenduota 2000 06 19 Nr. 05A-35

Darbo vadovas *Elmundas Žalys*

Redaktorė *Zita Manstavičienė*

Programinė įranga: *Tadeuš Šeibak*

Kompiuterinė grafika: *Edita Tatarinavičiūtė*

Kompiuterinis maketavimas: *Aldona Žalienė*

Gamybos vadovas *Algimantas Paškevičius*

Kalbos redaktorė *Diana Gustienė*

Korektorė *Birutė Laurinskienė*

ISBN 9986-546-93-1

TURINYS

PRATARMĖ

Šiuolaikinė statistika – tai mokslas apie informacijos rinkimą, sisteminimą, analizavimą ir interpretavimą. Daugumai žmonių žodis „statistika" primena gausybę skaičių, diagramų bei rodiklių su išsamiais komentarais. Apskritai vyrauja gana skeptiškas požiūris ir į pateikiamų duomenų patikimumą, ir į komentarus – ne veltui statistika vadinama „trečiąja melo rūšimi". Nedidelei daliai visuomenės (ypač tiems, kurie susidūrė su matematinės statistikos vadovėliais) statistika – tai sudėtinga matematinė disciplina su daugybe sunkiai suvokiamų ir neaišku kaip su realiu pasauliu susijusių formulių.

Autorių nuomone, šis vadovėlis turėtų padėti atsikratyti abiejų stereotipų. Jame bandoma parodyti, kad statistika – tai ne vien sausas faktų aprašymas, kad ja galima remtis tiriant sudėtingus gyvenimo reiškinius; beje, tam pakanka itin saikingo „matematizavimo". Be abejo, tai tik pradinis įvadinis kursas. Jį išklausęs studentas galės pats pasirinkti, ar toliau nagrinėti sudėtingesnius modelius (tam reikės daugiau matematikos žinių).

Kad ir kokia graži bei logiška būtų teorija, vis tiek ateina laikas, kai tenka ją pagrįsti pavyzdžiais. Tuomet atliekamas eksperimentas ir renkami duomenys. Po to prireikia statistikos – duomenims apdoroti, išvadoms daryti, eksperimento rezultato sąlygotoms naujoms hipotezėms formuluoti. Statistinių tyrimų neišvengia medikai, norėdami nustatyti vaistų efektyvumą ar pagrįsti naują gydymo metodiką. Be jų neapsieina ir sociologai, kai reikia prognozuoti rinkimų rezultatus ar nustatyti populiariausią televizijos laidą. Be statistinių tyrimų psichologams sunku nustatyti, kas lemia žmogaus gebėjimą prisitaikyti prie naujos aplinkos. Statistinius tyrimus atlieka socialiniai darbuotojai, ieškodami esminių skirtumų tarp normalių ir asocialių šeimų vaikų. Kai edukologams prireikia palyginti kelių mokomųjų programų efektyvumą, jie taip pat taiko statistiką. O ekonomistai statistiką taiko nuolat – tiek esamai padėčiai įvertinti, tiek ekonomikos ateičiai prognozuoti.

Šis vadovėlis skiriamas visiems, kurie savo tyrimams taiko ar taikys statistinę analizę – net matematikams (pradiniam susipažinimui su statistika).

Įdėmiai perskaičiusiam šį vadovėlį jau turėtų pakakti žinių: 1) išmokti sisteminti duomenis; 2) gebėti parinkti tinkamą savo tiriamai problemai statistinį modelį; 3) apskaičiuoti visas reikiamas jo charakteristikas; 4) teisingai atlikti statistinę analizę ir interpretuoti atsakymus; 5) suprasti ir kritiškai įvertinti profesinėje literatūroje pateikiamus statistinius tyrimus. Trumpai tariant, įgyti veiksmingą įrankį praktiniams skaičiavimams, taip pat suvokti bendruosius teorinius principus, leidžiančius tinkamai tą įrankį taikyti.

Kokio matematinio pasirengimo reikalauja šis vadovėlis? Didžiajai daliai teksto suvokti pakanka vidurinės mokyklos kurso. Pateikiamus sudėtingesnius matematinių faktų įrodymus gali įveikti išklausiusieji vieno semestro aukštosios matematikos kursą.

Tikslai lėmė ir vadovėlio struktūrą. Pirmoji dalis skirta aprašomajai statistikai – duomenų sisteminimui ir jų pateikimui. Antrojoje dalyje supažindinama su tikimybių teorijos elementais. Viena vertus, šis vadovėlis *nėra* tikimybių teorijos vadovėlis, taigi pateikiami tik gerokai adaptuoti tikimybių teorijos faktai. Antra vertus, nesusipažinus su pagrindiniais tikimybių teorijos teiginiais, neįmanoma iki galo suprasti ir statistinių modelių. Tik suvokus modelį bei jo ribotumo lygį, galima teisingai interpretuoti rezultatus jų nefetišizuojant.

Trečioji, didžiausioji, vadovėlio dalis skirta statistinėms išvadoms. Nežinomų parametrų įverčiai, daugybė statistinių modelių, interpretacijos problemos – visa tai rasite šioje dalyje. Ypač stengtasi pabrėžti praktinį statistinių išvadų taikymą. Pavyzdžiai turėtų padėti įgyti praktinių įgūdžių. Pateikiamų statistinių kriterijų aibė pakankamai didelė, tad skaitytojai gali juos atsirinkti pagal dominančią sritį. Be abejo, rašydami šį vadovėlį nesistengėme „aprėpti neaprėpiamo“ – viename vadovėlyje galima aprašyti nemažai statistinių metodų bei modelių, tačiau toli gražu ne visus. Sudėtingesnius statistinius metodus bei modelius pateiksime antrojoje knygoje. Skaitytojui, ieškančiam papildomos informacijos, padės literatūros sąrašas bei autorių parengtas tinklalapis *www.ts.vu.lt*.

Kiekvieno skyriaus pabaigoje yra uždavinių. Stengėmės, kad jie atspindėtų kuo įvairesnius modelių taikymo atvejus. Vadovėlio pabaigoje pateiktos pagrindinės statistinės lentelės, reikalingos uždaviniams spręsti.

Šiuolaikinėje statistikoje dažniausiai susiduriama su dideliais duomenų masyvais, todėl skaičiavimams naudojami įvairūs statistikos paketai – SAS, SPlus, SPSS, STATISTICA ir pan. Dalį statistinių operacijų atlieka ir tokia populiari programa kaip EXCEL. Tikėtina, kad šio vadovėlio skaitytojas irgi naudosis kuriuo nors statistinių programų paketu. Kadangi visuose paketuose rezultatų pateikimo principai yra panašūs, mes apsiribojome SPSS programėlių rezultatų fragmentais ir jų analize.

Kiekvienos dalies ir skyriaus pabaigoje pateikiami raktiniai žodžiai.

Paveikslėliu išskirtos statistikų folkloro ir ironiškos citatos:

Ankstesniais laikais neturėta statistikos, todėl tekdavo remtis melu.

S. Likokas

Pastabos ir faktai knygoje išskirti taip:

*Žodis **statistika** kilo iš italų **stata** – valstybė. **Statista** – žmogus, tvarkantis valstybės reikalus.*

Kadangi lietuviškoji statistikos terminologija dar tebekuriama, skaitytojui gali kilti sunkumų nagrinėjant tarptautines publikacijas bei statistikos paketus. Todėl vadovėlio pabaigoje pateikiame vartojamų terminų anglų-lietuvių kalbų žodynėlį.

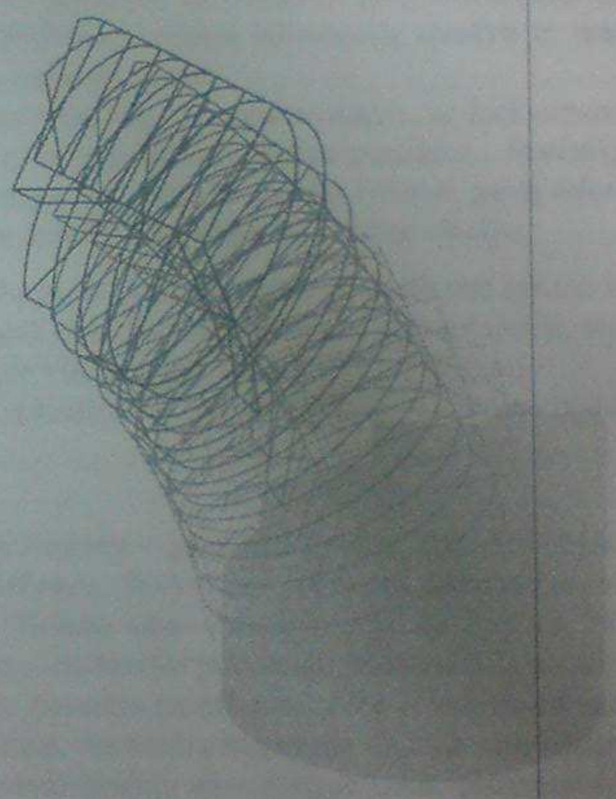
Knyga skirstoma į dalis. Cituojant skyrelį, esantį toje pačioje dalyje, nurodomas tik jo numeris (pvz., 5), kitais atvejais nurodoma ir dalis (pvz., II.1 žymi antros dalies 1 skyrelį). Trečioje dalyje dar yra ir skyriai, todėl nuoroda 5.1 reiškia 5 skyriaus 1 skyrelį.

Rašydami šį vadovėlį daug naudingų patarimų bei kritinių pastabų sulaukėme iš V. Kazakevičiaus, D. Krapavickaitės, N. Kligienės ir G. Kasnauskienės. R. Leipus leido pasinaudoti savo tikimybių teorijos rankraščiu ir jo iliustracijomis. Rengiant vadovėlį spaudai, nemažą techninę pagalbą suteikė T. Šeibakas. Knygoje pasinaudojome G. C. Ramseyerio ir J. Verhageno statistikų folkloro pavyzdžių tinklalapiais.

Rankraščio kalbą sutvarkė Z. Manstavičienė, tekstą piešiniais pagyvino E. Tatarinavičiūtė. Visiems jiems nuoširdžiai dėkojame.

Vydas Čekanavičius, Gediminas Murauskas

ĮVADAS: PRADINĖS SĄVOKOS



O kas kuria teoriją, neturėdamas duomenų, daro didžiulę klaidą.

A. Konanas Doilis. *Skandalas Bohemijoje*

Fizikas, chemikas ir statistikas laukia rektoriaus priimamajame. Tuo metu kambario kampe užsiliepsnoja šiukšlių dėžė. Fizikas sušunka: „Reikia skubiai atšaldyti visas medžiagas iki žemesnės nei jų degimo temperatūros". Chemikas paprieštarauja: „Reikia pasirūpinti, kad ugnis nebegautų deguonies". Statistikas puola padeginėti kitų kambario kampų. „Ką darai?" – sušunka fizikas su chemiku. Į tai statistikas atsako: „Tuojau įsitikinsime, kuris iš jūsų teisus, bet eksperimentui reikia didesnės imties".

Šiuolaikinė *statistika* – tai mokslas apie informacijos rinkimą, sisteminimą, analizavimą ir interpretavimą. Išskiriamos trys statistikos dalys: 1) duomenų rinkimas, 2) aprašomoji statistika, nagrinėjanti duomenų sisteminimo metodus, 3) statistinės išvados – analizės ir interpretavimo metodai.

Štai keletas statistinio tyrimo pavyzdžių:

1. Reikia nustatyti, kuri Vilniaus gyventojų dalis ruošiasi dalyvauti artėjančiuose rinkimuose. Pirmiausia statistikas apibrėžia kriterijus, kuriais remiantis galima nustatyti, kokie gyventojai laikomi vilniečiais rinkėjais. Tuomet, atsitiktinai parinkęs n rinkėjų, statistikas sužino jų ketinimus, o remdamasis gautais rezultatais, daro išvadas apie visus Vilniaus rinkėjus.
2. Edukologai nori išsiaiškinti, kuri mokomoji programa yra efektyvesnė. Reika nustatyti žmonių, kuriems bus skiriamos mokomosios programos, aibę; parinkti jos dalį tyrimams; atlikti stebėjimus; surinktą informaciją aprašyti ir, rezultatus apibendrinus, parinkti efektyvesnę programą.
3. Alaus darykla gamina dviejų rūšių alų. Norima nustatyti, ar kiekvienos rūšies dalis visoje alaus daryklos gamyboje atitinka rinkos poreikius. Statistikas turi nusakyti daryklos alaus vartotojų aibę, parinkti dalį jos tyrimui, gautą informaciją susisteminti, ją išanalizuoti ir pateikti daryklai savo rekomendacijas.

Visiems šiems pavyzdžiams būdinga tai, kad iš pradžių pasirenkama tyrimo objektų aibė, po to surenkama informacija apie objektus arba jų dalį, ji apdorojama ir, remiantis gautais rezultatais, daroma išvada apie vieną ar kitą tyrimo objektų požymį.

Šioje dalyje aptarsime pradines statistikos sąvokas bei duomenų rinkimo būdus.

1. Populiacija ir imtis

Pirmasis bet kokio statistinio tyrimo žingsnis – pasirinkti tiriamą aibę. Jos objektai turi vieną ar keletą tyrėją dominančių požymių. Pavyzdžiui, sociologą domina pensinio amžiaus žmonių balsavimo prioritetai. Tiriama aibė – pensinio amžiaus žmonės. Požymis, kuris šiame tyrime domina sociologą, – balsavimo prioritetai. Medikai tiria naujos skausmą malšinančios priemonės šalutinio poveikio simptomus. Aibė – visi vaisto vartotojai, požymis – šalutinio poveikio simptomai. Mokesčių inspekciją domina visų uždarųjų akcinių bendrovių metinis pelnas. Tyrimo objektų aibė – visos UAB, požymis – metinis pelnas. Statistinių tyrimų nagrinėjama objektų aibė dar vadinama *populiacija*.

Kartais populiacijos sąvoka esti gana abstrakti. Tarkime, ekonomistas nori įvertinti konkrečių akcijų kainos svyravimus praeitais metais. Nesunku apibrėžti tiriamą požymį – tai akcijų kaina. Objektas, turintis šį požymį, – akcijos. Populiaciją šiuo atveju sudaro tiriamos akcijos pirmą prekybos dieną, antrą prekybos dieną ir pan. Taigi šiame pavyzdyje populiacijos objektus skiria atstumas laike, o ne erdvėje. Analogiškai meteorologui populiacija gali būti visa (prabėgusių ir būsimų) vasario mėnesio 16 dienų aibė. Šiuo atveju populiacija yra begalinė.

Norint įvertinti tiriamą objektų požymį, reikia mokėti jį išmatuoti. Tinkamai parinkti *matavimo priemones* – tai viena iš svarbesnių eksperimento planavimo problemų, kurią turėtų spręsti atitinkamos srities specialistai kartu su statistikais, tačiau jos šiame vadovėlyje nenagrinėjame. Taigi kalbėdami apie intelekto koeficientą nenagrinėjame, ar naudotas klausimynas tinkamas (*validus*). Kalbėdami apie infliacijos didumą, mes neaptariame, kaip ji yra matuojama (atrodo, net ekonomistams šiuo klausimu ne visada pavyksta susitarti).

Mokėdami išmatuoti požymį, galime kelti klausimą apie jo reikšmių paplitimą visoje populiacijoje. Tiksliau šį uždavinį galima suformuluoti taip:

Bendriausias statistikos uždavinys – nustatyti tiriamų požymių reikšmių dažnių pasiskirstymus populiacijoje.

Tegul tiriamasis požymis yra VU studentų ūgis. Žinodami kiekvieno VU studento ūgį, galime pasakyti, kiek VU studentų yra aukštesnių nei 200 cm, keliolikos studentų ūgis yra nuo 150 cm iki 160 cm ir pan. Pasak statistikų, žinomas požymio ūgis reikšmių dažnių pasiskirstymas VU studentų populiacijoje. Žinodami šį skirstinį, galime atsakyti į visus klausimus apie VU studentų ūgį, pavyzdžiui, koks yra aukščiausio (žemiausio) VU studento ūgis.

Paprastai atliekant realius tyrimus požymių reikšmių pasiskirstymo dažniai populiacijoje tiksliai nežinomi. Be to, dažnai visos populiacijos ištirti neįmanoma. Paminėsime tris pagrindines priežastis, dėl kurių visa populiacija tiriama retai. Pirma, tokiam tyrimui reikia daug laiko. Antra, jis brangus. Trečia, dažnai neįmanoma išvardyti visų populiacijos elementų (pvz., kai populiacija begalinė). Taigi dažniausiai tiriama tik populiacijos dalis, vadinamoji *imtis*.

Populiacija – objektų, kurių požymiai tiriami, aibė.
Imtis – tai populiacijos dalis, naudojama statistiniam tyrimui.

Imčių pavyzdžiai: trisdešimt atsitiktinai parinktų VU studentų; tūkstantis apklausoje dalyvavusių rinkėjų; septynios mokesčių inspekcijos patikrintos firmos ir pan. Imties elementų skaičių vadinsime *imties didumu*. Imties elementų tiriamo požymio reikšmes vadinsime *duomenimis*, arba *duomenų aibe*.

Populiacijos elementai į imtį gali būti atrenkami įvairiais būdais. Šiame vadovėlyje nagrinėjamos tik visiškai atsitiktinės imtys. Galimi imčių sudarymo būdai aptariami šios dalies antrame skyrelyje.

Vienas iš svarbiausių reikalavimų – imtys turi būti reprezentatyvios. Imtis *reprezentatyvi*, jei ji teisingai atspindi tiriamo požymio galimų reikšmių populiacijoje proporcijas.

Būtent reprezentatyvumas lemia, ar ištyrus imtį galima padaryti patikimas išvadas apie visą populiaciją.

Akivaizdu, kad imties reprezentatyvumas glaudžiai susijęs su imties didumu. Jeigu imtis apima beveik visą populiaciją, tai ji labai reprezentatyvi. Kuri populiacijos dalis pateko į imtį, nurodo imties koeficientas K. Jis apibrėžiamas tik baigtinėms populiacijoms

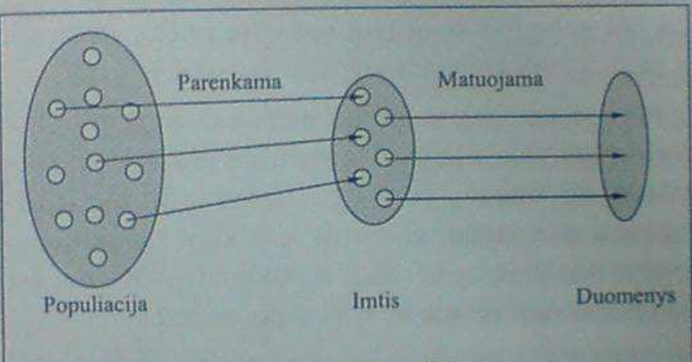
$$K = (n/N) \cdot 100\%,$$

čia n – imties didumas, N – populiacijos didumas.

Didinant imtį, galima padaryti patikimesnes išvadas, bet taip būna ne visada. Retai naudojamos labai didelės imtys, nes panašaus patikimumo informaciją galima gauti ir iš vidutinio didumo imčių. Be to, reprezentatyvumas priklauso ne tik nuo imties didumo, bet ir nuo jos sudarymo metodo.

Viena iš didžiausių imčių buvo sudaryta 1954 metais, tiriant Salko priešpoliomielitinę vakciną. Ją sudarė beveik du milijonai JAV pradinių klasių mokinių.

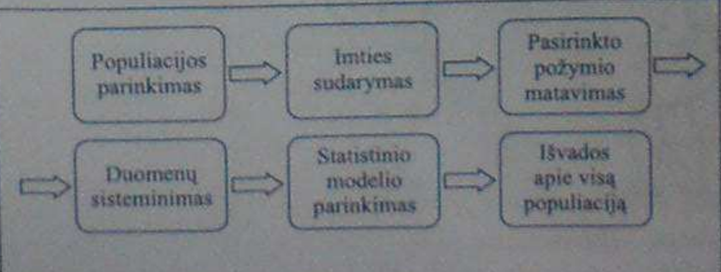
Po to, kai populiacija apibrėžta, iš jos parinkta imtis ir rastos tiriamo požymio reikšmės, gauti duomenys sisteminami, pavyzdžiui, nustatomas vyrų VU studentų imtyje skaičius, didžiausias tikrintų firmų pelnas ir pan. Susisteminti duomenys – informacijos „koncentratas". Jie ne tik leidžia lengviau suvokti tiriamą situaciją, bet ir padeda iškelti vieną ar kitą naują hipotezę. Duomenų sisteminimo ir jų pateikimo būdams nagrinėti skirta visa pirmoji šio vadovėlio dalis.



1 pav. Populiacija, imtis, duomenų aibė

Duomenis susisteminus, atliekama sudėtingesnė statistinė analizė. Kodėl? Į imtį elementai dažniausiai parenkami atsitiktinai. Taigi ir visi gauti duomenys bei rezultatai tam tikra prasme yra atsitiktiniai. Žinoma, tas atsitiktinumas nėra labai didelis, tačiau jį reikia skaitiškai įvertinti ir į gautą įvertį atsižvelgti. Todėl parenkamas duomenis atitinkantis matematinis modelis. Galų gale, atsižvelgus į atsitiktinumą, padaroma išvada apie visą populiaciją.

Dažnai (ypač matematinės statistikos vadovėliuose) imtimi vadinama ne kokį nors požymį turinčių *objektų* aibė, o gautų požymio reikšmių aibė (duomenų aibė). Tuomet populiacija yra visų požymio reikšmių aibė. Ateityje mes taip pat kartais žodį *imtis* vartosime kaip *duomenų aibės* sinonimą. Manome, kad skaitytojams tai nesukels problemų.



2 pav. Būdinga statistinio tyrimo eiga

2. Imčių sudarymo būdai

Populiacijos elementai tyrimui parenkami ne bet kaip, o iš anksto pasirinktu imties sudarymo būdu. Atliekant statistinius tyrimus, domina išvados apie visą populiaciją, todėl galima teigti, kad svarbiausia kiekvienos imties savybė yra jos reprezentatyvumas. Tačiau praktiniams tyrimams svarbi ir imties sudarymo kaina, elementų atrankos į imtį paprastumas, dažnai ir laikas, kurio prireikia imčiai sudaryti.

Parenkant kiekvienos imties elementus, egzistuoja tam tikras atsitiktinumas. Tačiau kartais šis atsitiktinumas visiškai subjektyvus, jo įtakos imties sudarymui neįmanoma išmatuoti. Taip sudarytos imtys vadinamos *netikimybinėmis* imtimis. Kitoms imtims atsitiktinumas yra griežtai apibrėžtas – kiekvieno elemento galimybę priklausyti imčiai nusako tam tikra tikimybė. Tokios imtys vadinamos *tikimybinėmis*. Trumpai aptarsime būdingiausius imčių pavyzdžius.

2.1. Netikimybinės imtys

Ekspertinė imtis. Elementai į imtį įtraukiami atsižvelgus į ekspertų nuomonę. Subjektyvumo čia tiek daug, kad negalima net palyginti kelių taip sudarytų imčių. Tokios imtys nereprezentatyvios ir jų rezultatų neįmanoma apibendrinti visai populiacijai.

Kvotinė imtis. Atsižvelgus į populiacijos sandarą, iš anksto numatomos populiacijos dalių kvotos. Pavyzdžiui, tiriant įvairių tautybių Lietuvos gyventojų požiūrį į ekonominę situaciją tariama, kad imtį turi sudaryti 80 lietuvių, 10 rusų, 7 lenkai ir 3 baltarusiai. Kai kiekvienos kvotos elementai parenkami atsitiktinai, gaunama atsitiktinė sluoksninė imtis (žr. 2.2 skyrelį). Imtis, kurioje kvotos užpildomos ne visai atsitiktinai, vadinama kvotine imtimi. Šiuo atveju atsiranda sunkumų darant išvadą apie visą populiaciją.

Proginė imtis. Į imtį įtraukiami pirmi pasitaikę populiacijos elementai. Sudarant tokią imtį, daug lemia atsitiktinumas, kurio negalima aprašyti paprastais matematiniais modeliais ir kaip nors įvertinti. Tokia imtis visiškai nereprezentatyvi.

Deja proginės imtys gana dažnai pasitaiko. Būtent joms taikant tikimybinių imčių tyrimo metodus ir padaroma daugiausia klaidų. Būdingas tokios klaidos pavyzdys – mokytojas apklausia savo auklėjamosios klasės mokinius, o išvadas daro apie daug didesnę mokinių aibę (viso miesto, rajono ar net šalies mokinius). Iš proginės imties rezultatų neįmanoma padaryti jokių statistinių išvadų apie visą populiaciją.

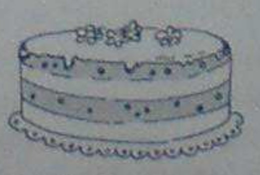
2.2. Tikimybinės imtys

Paprastoji atsitiktinė imtis. Jeigu visų populiacijos elementų galimybės priklausyti imčiai visiškai vienodos, tai turime *paprastąją atsitiktinę grąžintinę imtį*. Tokiai imčiai sudaryti būtinai naudojami visi populiacijos elementai. Dažnai jie atrenkami pagal atsitiktinių skaičių lenteles. Jeigu elementų galimybės priklausyti imčiai nevienodos, bet žinomos ir nusakomos tikimybėmis, tai turime nelygių tikimybių atsitiktinę imtį (*NTAI*).

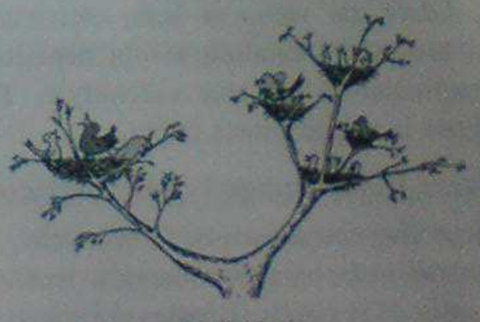
Sistemingoji imtis. Sistemingosios imties sudarymo principas yra gana paprastas: 1) atsižvelgus į populiacijos didumą ir numatomą pačios imties didumą, parenkamas išrinkimo žingsnis, 2) visi elementai išrikiuojami į eilę, 3) iš kelių pirmųjų elementų atsitiktinai imamas vienas, 4) pasirinktu žingsniu atrenkami visi likę elementai. Būdingas sistemingosios imties sudarymo pavyzdys.

Iš 100 įmonių abėcėlinio sąrašo numačius patikrinimui išrinkti 10, iš pirmo dešimtuko atsitiktinai parenkama pirmoji, po to imamos iš eilės kas dešimta. Tarkime, kad pirma buvo trečioji pagal sąrašą įmonė. Tuomet į imtį paklius ir tryliktoji, dvidešimt trečioji, trisdešimt trečioji ir pan.

Sluoksninė imtis. Visa populiacija suskirstoma į sluoksnius (*stratus*). Kiekvienam sluoksniui taikomas paprastosios atsitiktinės grąžintinės imties sudarymo būdas. Sluoksninė imtis leidžia surinkti informaciją apie kiekvieną populiacijos sluoksnį. Be to, šią informaciją galima apibendrinti visai populiacijai. Tiesa, tuomet reikia atsižvelgti į kiekvieno sluoksnio užimamą visoje populiacijoje dalį. Sluoksninės imties pranašumas yra tai, kad be papildomų lėšų galima atlikti kelis tyrimus (ir visos populiacijos, ir atskirų sluoksnių). Kita vertus, ne visada lengva visą populiaciją suskaidyti į sluoksnius. Tarkime, norime sudaryti visus Lietuvos gyventojus atitinkančią sluoksninę imtį. Priklausomai nuo tyrimo tikslų gali tekti atsižvelgti į lytį, tautybę, gyvenamąją vietą, amžių ir pan. Savo ruožtu kyla klausimų, kaip skirstyti į amžiaus grupes, kaip suprasti gyvenamąją vietą (tik miestą ir kaimą ar miestą, miestelį, kaimą) ir pan. Be to, reikės žinoti, kokią visos Lietuvos gyventojų populiacijos dalį (kiek procentų) sudaro atitinkamo sluoksnio atstovai. Jei kriterijų, pagal kuriuos apibrėžiami sluoksniai, yra labai daug, tai tokią imtį sudaryti ne ką lengviau nei paprastąją atsitiktinę imtį. Bendras reikalavimas sluoksninėms imtims: populiacija turi būti nevienalytė (heterogeniška) sluoksnių atžvilgiu ir vienalytė (homogeniška) sluoksnių viduje.



Sluoksninė imtis



Lizdinė imtis

Lizdinė imtis. Visa populiacija pagal tam tikrą požymį suskirstoma į panašias grupes – lizdus (*klasterius*). Iš visų lizdų aibės paprastosios atsitiktinės imties būdu atrenkama dalis. Į imtį pakliūna *visi* atrinktųjų lizdų elementai. Pateiksime būdingą lizdinės imties sudarymo pavyzdį.

Tiriamas Vilniaus rajono penktųjų klasių mokinių pažangumas. Iš sąrašo visiškai atsitiktinai parenkama dalis klasių. Tyrime dalyvauja visi parinktųjų klasių mokiniai. Jie ir sudaro lizdinę imtį. Tai tikimybinė imtis, todėl rezultatus, gautus naudojant lizdines imtis, galima apibendrinti visai populiacijai. Tiesa, tam šio vadovėlio metodai netinka. Bendras reikalavimas lizdinėms imtims: populiacija lizdų atžvilgiu turi būti homogeniška, o lizdų viduje heterogeniška.

Be abejo, yra ir kitų imčių sudarymo būdų, pavyzdžiui, vadinamosios *daugiapakopės* imtys, kurios gaunamos įvairiai derinant jau minėtus imčių sudarymo metodus.

Paprastosios atsitiktinės imtys. Šis vadovėlis skirtas *paprastosioms atsitiktinėms grąžintinėms imtims* nagrinėti. Todėl jas aptarsime išsamiau.

> *Paprastoji atsitiktinė grąžintinė imtis:*
> *kiekvienu* imties sudarymo momentu *visų* populiacijos elementų galimybės patekti į imtį yra *vienodos*.

Tokia atsitiktinė imtis gaunama, kai yra *grąžintiniai* ėmimai. Apibrėžime aiškiai pasakyta, kad kiekvieno populiacijos elemento galimybė patekti į imtį yra tokia pat. Klasikinė vienodas galimybes teikianti situacija yra tokia: 1) visi populiacijos elementai sunumeruojami, 2) elementų numeriai užrašomi ant rutulių, 3) rutuliai sudedami į dėžę ir gerai sumaišomi, 4) ištraukiamas pirmas pasitaikęs rutulys, 5) elementas, atitinkantis ant rutulio užrašytą numerį, įtraukiamas į imtį. Tarkime, taip nustatėme pirmąjį imties elementą. Kiek netikėtas antrasis apibrėžimo reikalavimas, kad vienodų galimybių sąlyga galiotų kiekvienu imties sudarymo momentu. Taigi ji turi galioti ir tam elementui, kuris jau įtrauktas į imtį. Todėl ištrauktąjį rutulį turėtume grąžinti į dėžę ir tik po to vėl traukti kitą. Jeigu rutulio negrąžintume, tai galėtume iš esmės pakeisti lygių galimybių sąlygą. Pateiksime pavyzdį, kad tokia situacija įmanoma.

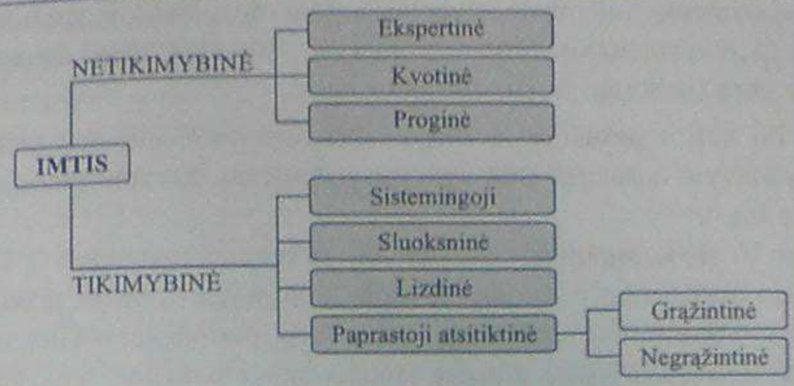
Dėžėje yra dešimt rutulių su loterijos bilietų numeriais, iš kurių tik vienas laimingas. Tarkime, kad laimingasis ištrauktas traukiant pirmąkart. Aišku, kad galimybės ištraukti laimingą numerį prieš pirmą traukimą ir po jo (kai to numerio jau nebėra) iš esmės skiriasi.

Paprastoji atsitiktinė imtis, kai elementai negrąžinami, vadinama *negrąžintine*. Jai nagrinėti yra specialių metodų, tačiau šiame vadovėlyje jie neaptariami. Praktiškai beveik visada imtys sudaromos be grąžinimų, bet taikomi grąžintinių imčių tyrimo metodai. Jeigu populiacija *didelė*, tai vieno ar kelių elementų pašalinimas tiriamam reiškiniui įtakos neturi, taigi tuo atveju esminių klaidų nepadaroma. Būtent tokios didelių populiacijų situacijos nagrinėjamos ir šiame vadovėlyje. Todėl toliau jau neakcentuosime skirtumo tarp grąžintinės ir negrąžintinės imčių.

2.3. Atsitiktinė ir sistemingoji imties paklaidos

Imtys neapima visos populiacijos. Dažniausiai, naudodamiesi imties duomenų aibe, skaitiškai įvertiname nežinomą populiacijos parametrą. Skirtumas tarp tikrosios parametro reikšmės ir apskaičiuotosios iš imties duomenų vadinamas imties paklaida. Tarkime, pasirinkome tikimybinį imties sudarymo būdą. Elementai į tokias imtis patenka atsitiktinai. Todėl tuo pačiu metodu sudarę kelias vienodo dydžio tikimybines imtis, galime tikėtis, kad jų elementai nesutaps. Toks imties *kintamumas* sąlygoja ir *atsitiktinę* rezultatų paklaidą.

Būtent su šio atsitiktinumo įvertinimais ir susidursime tolesniuose skyriuose. Atsitiktinė paklaida priklauso nuo imties didumo. Kuo didesnė imtis, tuo mažesnė atsitiktinė



3 pav. Imtys

paklaida. Tačiau šis mažėjimas nėra tolygus, o pasiekus tam tikrą imties didumą, atsitiktinė paklaida pradeda mažėti labai lėtai.

Yra ir kitokių imties paklaidų, iškreipiančių rezultatus, kartu ir statistines išvadas. Gerai žinomi atvejai, kai priešrinkiminių apklausų rezultatais pagrįstos prognozės per pačius rinkimus visiškai nepasitvirtina. Taip atsitinka dėl paklaidos, susijusios su „matavimo instrumento" netobulumu, t. y. *sistemingosios* paklaidos. Svarbiausias sistemingosios paklaidos šaltinis – atsakymų stoka (pvz., daug neužpildytų anketų, respondentai atsisako dalyvauti tyrime).

Socialiniuose moksluose nustatant klausimo „neatsakymo laipsnį", naudojamasi vadinamuoju atsakymo lygiu.

$$\textit{Atsakymo lygis} = \frac{\text{atsakiusiųjų respondentų skaičius}}{\text{visų parinktų respondentų skaičius}}.$$

Jeigu atsakymų lygiai maži, tai galima įtarti, kad netinkamai sudarytas klausimynas, blogai parinkta imtis ir kt. Kiti sistemingosios paklaidos šaltiniai: 1) respondentai meluoja, 2) duomenų rinkėjo asmenybė turi įtakos rezultatams, 3) klausimai nevienareikšmiškai suformuluoti, t. y. respondentai skirtingai juos supranta, 4) respondentai nežino, ko jų klausia, 5) klausimas lemia norimą atsakymą (respondentas gal ir mano kitaip, bet jam tiesiog nepatogu tai parodyti).

Sistemingoji paklaida nepriklauso nuo imties didumo. Ji išsamiau šiame vadovėlyje nenagrinėjama.

2.4. Eksperimentų imtys

Iki šiol kalbėjome apie vienos imties[1] sudarymo metodus, tačiau praktiškai tyrimams naudojamos ir kelios imtys. Dažniausiai norima rasti tiriamo reiškinio priklausomybę nuo kokio nors kito (ne atsitiktinio) reiškinio. Pavyzdžiui, norima nustatyti dietos efektyvumą (svorio sumažėjimo priklausomybę nuo pasirinktos dietos), kelių mokomųjų programų efektyvumą (žinių testo priklausomybę nuo taikytos mokomosios programos), alaus rūšies ir perkamumo ryšį ir pan. Tuomet tradiciškai sudaromos kelios imtys, o tiriamasis

[1] Toliau atsitiktinę imtį suprasime kaip paprastąją atsitiktinę grąžintinę imtį.

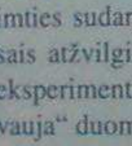
požymis matuojamas eksperimento nulemtose situacijose. Manoma, kad visi gauti matavimų skirtumai priklauso tik nuo situacijos. Organizuojant eksperimentą, imtys derinamos tarpusavyje. Paminėsime porą tradicinių imčių derinimo būdų.

Lygiagrečiosios imtys. Tai kelios nesusijusios imtys. Nei atskirų imčių, nei kiekvienos imties elementai tarpusavyje nesusiję. Būdingas eksperimentas, kuriam naudojamos lygiagrečiosios imtys, yra toks:

Iš pradžių atsitiktinai iš visos populiacijos tiriamieji elementai parenkami į kelias grupes. Reikalaujama, kad matuojamo požymio atžvilgiu šios pradinės imtys praktiškai nesiskirtų. Po to viena grupė (kontrolinė) dalyvauja viename eksperimente, o kitos (tiriamosios) – kituose. Po eksperimento visose grupėse išmatuojamas požymis. Pavyzdžiui, dvi metų pradžioje vienodo pajėgumo klasės metus mokosi pagal skirtingas programas. Metų gale, specialiu testu patikrinę kiekvieno mokinio mokymosi rezultatus, gausime dvi lygiagrečias nepriklausomas duomenų aibes.

Porinės imtys. Tai dvi imtys, kurių elementai nesusiję (nepriklausomi), bet kiekvienas pirmos imties elementas turi savo (priklausomą) „porininką" antrojoje imtyje. Būdingas eksperimentas: imties elementus tiriame skirtingais laiko momentais, t. y. dvi imtis skiria tik atstumas laike. Pavyzdžiui, viena imtis – žmonių grupė iki dietos, o antroji – tie patys žmonės po dietos. Dažnai porinės imtys sudaromos tuo pačiu laiko momentu. Tuomet pirmajai, atsitiktinei imčiai, sudaroma jos „dvynė" imtis. Kiekvienam pirmosios imties elementui parenkamas kaip nors su juo susijęs „porininkas". Pavyzdžiui, tiriame sutuoktinių poras. Vieną imtį sudaro vyrai, o kitą – jų žmonos. Be abejo, priklausomybė tarp poros elementų gali būti ir labai stipri, ir vidutinė. Vienas iš būdingiausių porinės imties sudarymo būdų yra toks: kiekvienam pirmosios imties elementui parenkamas visais atžvilgiais analogiškas antrosios imties atstovas. Abi imtys dalyvauja skirtinguose eksperimentuose. Po to abiejose imtyse išmatuojamas tiriamasis požymis. Tyrime „dalyvauja" duomenų aibės reikšmių poros. Pavyzdžiui, tiriamas kiekvieno žmogaus kairiosios ir dešiniosios rankų stiprumas. Porinės imties sudarymo principus galima taikyti ir trims, keturioms ar daugiau imčių.

Viena iš problemų, su kuria susiduriama atliekant eksperimentus, – tai kontrolinės ir eksperimentinės grupių skirtumai ne tik dėl tiriamojo (kontroliuojamojo), bet ir kokio nors kito požymio įtakos. Skaičiuojant tokio dalyko įvertinti neįmanoma.



Pakviesti statistiką, kai eksperimentas jau atliktas, gali reikšti ne ką kita kaip prašymą atlikti pomirtinį skrodimą: jis galbūt galės pasakyti, kodėl eksperimentas nepasisekė.

R. A. Fišeris

Be tyrimų, kurių metu daromas eksperimentas (t. y. pats tyrėjas kontroliuoja reiškinius, darydamas įtaką eksperimento rezultatams), yra ir *stebimieji* tyrimai. Pavyzdžiui, apklausa dėl mirties bausmės panaikinimo.

3. Kintamieji

3.1. Kintamojo sąvoka

Duomenų analizės metodo parinkimas labai priklauso nuo jų prigimties. Paaiškinsime, kaip klasifikuojami duomenys. Tam mums reikia aptarti kintamojo sąvoką. Populiacijos kartu ir imties, elementus vienija tiriamasis požymis. Matuodami šį požymį, gauname tam tikrą dydį, kuris kinta kartu su imties nariais. Šį dydį ir vadinsime *kintamuoju*. Imties duomenų aibė – tai ne kas kita kaip kintamojo reikšmių aibė (visų galimų kintamojo reikšmių poaibis). Išmatavę visą populiaciją, gautume visas kintamojo reikšmes. Tarkime, norime sužinoti, kokia yra VU studentų tautinė sudėtis. Šiuo atveju VU studentai – tiriamoji populiacija, tautybė – kintamasis, pasirinktų studentų (imties elementų) tautybės – duomenų aibės reikšmės (kintamojo „tautybė" realizacijos).

Pagal matuojamo reiškinio prigimtį kintamieji skirstomi į *kiekybinius* ir *kokybinius*.

Kiekybinio kintamojo reikšmė – tai atsakymas, *kiek* tiriamo požymio turi populiacijos elementas, tuo tarpu kokybiniai kintamieji nusako dydžius, kurių neįmanoma įvertinti skaičiais. Kiekybinių kintamųjų pavyzdžiai – laikas, aukštis, šeimos gausumas, metinė infliacija, sesijos pažymių vidurkis ir pan. Kokybinių kintamųjų pavyzdžiai – lytis, rasė, kraujo grupė, automobilio markė ir pan. Anksčiau pateiktame pavyzdyje studentų tautybė – kokybinis kintamasis. Beje, kokybinio kintamojo reikšmes koduojant skaitmenimis, kintamojo prasmė nesikeičia. Pavyzdžiui, kintamojo „lytis" reikšmes galima koduoti pagal tokią taisyklę: „vyras" = 1, „moteris" = 2.

Kiekybiniai kintamieji savo ruožtu yra skirstomi į tolydžiuosius ir diskrečiuosius. Kiekybinis kintamasis yra vadinamas *tolydžiuoju*, jei jo reikšmių skirtumas gali būti kiek norima mažas. Kiekybinis kintamasis, kurio reikšmių skirtumas yra ne mažesnis už tam tam tikrą minimalų pokytį, vadinamas *diskrečiuoju* kintamuoju. Laikas, masė, aukštis – tolydieji kintamieji; šeimos gausumas, korektūros klaidų skaičius, banko klientų skaičius – diskretieji kintamieji. Kokybinių ir kiekybinių kintamųjų analizės metodai traktuojami skirtingai. Pavyzdžiui, kokybiniai kintamieji negali būti sudedami, dauginami, vidurkinami.

3.2. Matavimų skalės

Duomenys, kartu ir kintamieji, yra klasifikuojami ir atsižvelgiant į naudotą matavimų skalę. Panagrinėkime skaitmenį 6 tokiose situacijose:

1. Jonuko sportinių marškinėlių numeris yra 6.
2. Jonukas mokosi VI klasėje.
3. Jonukas žaidė lauke esant 6°C temperatūrai.
4. Jonukas užaugino 6 cm ilgio agurką.

Keturios skirtingos situacijos iliustruoja skirtingą informacijos lygį, priklausomą nuo naudotos skalės. Pirmuoju atveju Jonuką pagal marškinėlių numerį galime tik priskirti vienai ar kitai grupei. Pavyzdžiui, kai žaidžiamas futbolas, galėtume sakyti, kad Jonukas yra gynėjas (jei numeriai 2-as, 3-as, 4-as, 5-as, 6-as priskiriami gynėjų grupei). Antruoju atveju turime daugiau informacijos. Jonukas mokosi VI klasėje (priskiriamas šeštokų grupei), taigi jis mokosi aukštesnėje nei V ir žemesnėje nei VII klasėje. Trečiuoju atveju galime sakyti, kad temperatūra lauke keliais laipsniais žemesnė (aukštesnė) už ankstesnę, t. y. galime įvertinti kiekybinius skirtumus (ko negalima daryti pirmoje ir antroje situacijose!). Ketvirtuoju atveju galime Jonuko išaugintą agurką lyginti su kitų vaikų išaugintaisiais, imdami ilgių santykį (ko negalima daryti ankstesnėse situacijose!). Šios keturios situacijos demonstruoja keturių skirtingų matavimo skalių naudojimo pavyzdžius.

> Yra keturios kintamųjų matavimo skalės:
> 1) *pavadinimų*, 2) *rangų*, 3) *intervalų*, 4) *santykių*.

Trumpai aptarsime kiekvieną iš jų.

3.3. Pavadinimų skalė

Pavadinimų skalė dar vadinama *nominaliąja*, arba *klasifikacine*, skale. Pagal kintamojo reikšmes, gautas naudojant pavadinimų skalę, imties objektus galima tik klasifikuoti, t. y. priskirti vienai ar kitai grupei. Objektams arba objektų klasėms priskiriami simboliai (kategorijų kodai). Jie tarpusavyje nepalyginami (net ir tuo atveju, kai yra skaičiai). Matuotų pagal šią skalę kintamųjų reikšmėms aritmetinės operacijos neturi prasmės.

> Kintamieji, kurie matuojami pavadinimų skalėje, vadinami *nominaliaisiais* kintamaisiais.

Nominaliųjų kintamųjų pavyzdžiai: futbolo marškinėlių numeris, telefono numeris, pašto indeksas, lytis, tautybė, kraujo grupė.

Matuojant nominalųjį kintamąjį, reikia turėti omenyje, kad:

1) kiekvienas imties elementas turi turėti jam tinkamą kategoriją (pvz., nustatant tautybę nepakanka kategorijų „lietuvis", „rusas", „lenkas" – turi būti dar bent viena kategorija, tarkime, „kita");

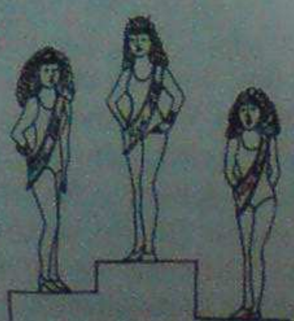
2) visos kategorijos turi aiškiai skirtis (pvz., kintamojo „šeiminė padėtis" kategorijos „vedęs", „nevedęs", „našlys", „kiti" nėra aiškiai diferencijuotos, nes išsiskyręs respondentas gali pakliūti ir į kategoriją „nevedęs", ir į kategoriją „kiti").

3.4. Rangų skalė

Rangų skalė dar vadinama *tvarkos* skale. Ši skalė naudojama tada, kai statistikas gali nustatyti objektų tiriamo požymio skirtumus ir pagal tai objektus išrikiuoti į eilę. Kintamieji, matuojami rangų skalėje, vadinami *ranginiais* kintamaisiais. Pagal ranginių kintamųjų reikšmes objektus galima ne tik skirstyti į klases, bet ir jas sutvarkyti (tvarkos savybė). Pavyzdžiui, bėgimo dalyviams pagal sugaištą distancijai nubėgti laiką skiriamos vietos. Šiuo atveju vieta yra rangas.

Ar didesnis skaičius atitinka didesnį požymio kiekį, priklauso nuo skaičių – rangų priskyrimo taisyklės. Šie skaičiai tarpusavyje gali būti lyginami tik eiliškumui nustatyti.

Kiti ranginių kintamųjų pavyzdžiai: pedagoginiai mokslo vardai, mokymosi lygis. Galima manyti, kad gražuolė, užėmusi grožio konkurse pirmąją vietą, yra gražesnė už trečiosios vietos laimėtoją, bet negalime teigti, kad ji triskart gražesnė.



Kartais ranginių kintamųjų reikšmės įvardijamos kaip tam tikros kategorijos. Pavyzdžiui, mokslo laipsnis – viena vertus, parodo vietą (laiptelį), kurią mokslininkas užima mokslininkų hierarchijoje, antra vertus, nusako tam tikrą kategoriją.

Ranginiai ir nominalieji kintamieji vadinami kategoriniais.

Sąvokos „kokybinis" ir „kategorinis" vartojamos kaip sinonimai.

3.5. Intervalų skalė

Matavimams naudojant intervalų skalę, objektus galima ne tik klasifikuoti, tvarkyti, bet ir kiekybiškai įvertinti skirtumus. Intervaliniai duomenys visada *skaitiniai*. Skirtumas (intervalas) tarp dviejų kintamojo reikšmių rodo, *kiek* daugiau (mažiau) matuojamojo požymio yra viename elemente, palyginti su kitu elementu.

Nulinis taškas intervalų skalėje yra laisvai parenkamas. Dviejų šios skalės intervalų santykis nepriklauso nei nuo matavimo vienetų, nei nuo nulinio taško. Imkime tokį pavyzdį.

Temperatūra matuojama pagal Celsijaus arba Farenheito skalę (abi yra intervalų skalės). Šių dviejų skalių nulinis taškas ir matavimo vienetai skiriasi, tačiau abi jos pateikia vienodą informacijos kiekį. Iš tiesų vienos ir kitos skalės parodymus sieja tiesinis ryšys

$$F = \frac{9}{5}C + 32;$$

čia F – laipsnių skaičius Farenheito skalėje, C – laipsnių skaičius Celsijaus skalėje. Keletas matavimų dviejose skalėse pateikta 1 lentelėje.

1 lentelė. Celsijaus ir Farenheito skalių palyginimas

Celsijaus (°C)	Farenheito (°F)
0	32
10	50
30	86
100	212

Paėmę dviejų intervalų santykį Celsijaus skalėje ir atitinkamų intervalų santykį Farenheito skalėje, gausime tą patį skaičių. Pavyzdžiui, $(30 - 10)/(10 - 0) = 2$ – Celsijaus skalėje, $(86 - 50)/(50 - 32) = 2$ – Farenheito skalėje.

Intervalinių kintamųjų pavyzdžiai: temperatūros matavimai, kalendorinis laikas, intelekto koeficiento vertinimas. Būdingas intervalų skalės pavyzdys yra metų skaičiavimas – nuo Romos įkūrimo, nuo Kristaus gimimo, nuo Mahometo išvykimo į Mediną. Intervalinius duomenis galime sudėti, atimti, dauginti, dalyti iš skaičiaus (taigi ir vidurkinti).

3.6. Santykių skalė

Ši skalė skiriasi nuo intervalų skalės tik tuo, kad joje yra apibrėžta absoliuti atskaitos pradžia. Šioje skalėje yra *absoliutusis nulis*, t. y. nulinis taškas, rodantis, kad tiriamojo požymio nėra. Taigi rezultatai visuomet neneigiami skaičiai. Skaičių, gautų matuojant požymius, santykis parodo kiekybinį matuojamojo požymio santykį. Santykių skalėje jis nuo matavimų vienetų nepriklauso. Pavyzdžiui, dviejų žmonių svorių santykis nepriklauso nuo to, ar svoriai buvo matuoti gramais, ar svarais. Santykių skalės kintamieji: ūgis, svoris, amžius, atlyginimas, kaina, laikas nuo eksperimento pradžios ir pan.

3.7. Kintamieji ir skalės

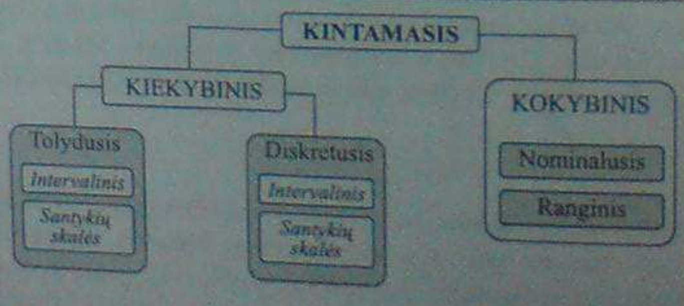
Kiekybiniams kintamiesiems matuoti naudojama intervalų arba santykių skalė. Kokybiniams kintamiesiems matuoti naudojama pavadinimų arba rangų skalė. Iš kiekybinių tolydžiųjų kintamųjų galima gauti ranginius, o iš šių – nominaliuosius kintamuosius, tačiau reikia žinoti, kad taip prarandama dalis informacijos. Pavyzdžiui, tegul kintamasis yra ūgis. Renkame informaciją apie studentų vaikinų ūgį. Mums tereikia žinoti, ar vaikinas žemaūgis (sąlygiškai laikykime jį mažesniu už 165 cm), ar vidutinio ūgio ([165; 180]), ar aukštaūgis (per 180 cm). Šiuo atveju matavimams naudojame rangų skalę. Duomenų aibė atrodo taip:

Šiaulys	Vidutinio ūgio
Jonauskas	Žemaūgis
Žukauskas	Aukštaūgis
...	...

Apklaustųjų studentų ūgio vidurkiui skaičiuoti tokia lentelė nebetiktų.

Kintamuosius galima klasifikuoti pagal įgyjamų reikšmių skaičių. Kintamasis, įgyjantis tik dvi reikšmes, vadinamas *dvireikšmiu*, arba binariuoju, kintamuoju. Statistinių išvadų teorijoje kokybiniams kintamiesiems taikomi neparametriniai kriterijai, o kiekybiniams – parametriniai ir neparametriniai.

Dažnai dėl tam tikrų priežasčių neįmanoma išmatuoti kai kurių vieno ar kelių objektų kintamųjų reikšmių. Tai vadinamosios *trūkstamosios reikšmės* (praleistieji stebėjimai). Atliekant skaičiavimus ne visi tokio objekto duomenys atmetami, tik reikia žinoti, kaip elgtis atsižvelgiant į naudojamų programų paketų ypatybes. Be to, kartais trūkstamos reikšmės (arba jų skaičius) irgi yra informatyvios. Pavyzdžiui, paprašius įvertinti politiko populiarumą 10 balų skalėje, 1% apklaustųjų jį įvertino dešimtukais, o 99% nurodė, kad tokio politiko nežino (t. y. turime net 99% trūkstamų reikšmių). Taigi ar šis politikas gali laikyti save populiariu?



4 pav. Kintamieji

2 lentelė. Skalės ir leistini veiksmai

Skalės	Leistini veiksmai	Skaičius
Pavadinimų	Elementų, patekusių į kiekvieną kategoriją, skaičiaus radimas	Kategorijos kodas
Rangų	Elementų, turinčių konkretų rangą, skaičiaus radimas. Rangų palyginimas (santykiai „daugiau“, „mažiau“)	Rangas
Intervalų	Sudėtis, atimtis, daugyba, dalyba iš skaičiaus	Kintamojo reikšmė
Santykių	Visos matematinės operacijos	Kintamojo reikšmė

atsitiktinė paklaida
diskretusis kintamasis
duomenų aibė
dvireikšmis kintamasis
ekspertinė imtis
intervalinis kintamasis
kategorinis kintamasis
kiekybinis kintamasis
kokybinis kintamasis
kvotinė imtis
lizdinė imtis
nominalusis kintamasis
paprastoji atsitiktinė imtis
populiacija
proginė imtis
ranginis kintamasis
santykių skalė
sisteminė imtis
sistemingoji paklaida
sluoksninė imtis
tikimybinė imtis
tolydusis kintamasis
trūkstamoji reikšmė

UŽDAVINIAI

1. Tiriamas Lietuvos paauglių požiūris į kontraceptinių priemonių vartojimą. Kas sudarys populiaciją, o kas imtį?

2. Vienos firmos vadovybė, norėdama būsimiems savo programuotojams nustatyti patrauklų atlyginimą, kreipėsi pagalbos į statistikus. Kaip organizuoti tyrimą: ką laikyti populiacija, kaip išrinkti imtį?

3. Lietuvių tautosakoje pasakojama apie poną, kuris nenorėjo samdyti berno Andriaus dėl to, kad jo vardas negražus. Tuomet Andrius pasiūlė atlikti statistinį tyrimą. „Dabar visi Andriai“, – kalbėjo bernas. – „Nueikit į bažnyčią sekmadienį ir sušukit ‘Andriau’, – visi atsisuks.“ Ponas taip ir padarė: 1) atsitiktinai pasirinko bažnyčią, 2) sušuko „Andriau“, 3) nusprendė, kad visi Andriai. Kokią imtį sudarė ponas? Kokia paklaida lėmė neteisingą pono sprendimą?

4. Kurie iš pateiktų teiginių pagrįsti aprašomosios statistikos, o kurie statistinių išvadų metodais:
   a) daugiau kaip 70% Lietuvos gyventojų pasisako prieš lito pakeitimą euru;
   b) per pirmą ketvirtį infliacija sudarė 2%;
   c) „Vilniaus trauktinė“ per metus pagamino 200 tūkst. dekalitrų degtinės;
   d) per praėjusį mėnesį prezidento reitingas išaugo 5% ir šiuo metu reitingų sąrašuose jis pirmauja;
   e) dienomis, kai „Žalgiris“ žaidžia Eurolygos rungtynes, baruose išgeriamo alaus kiekis patrigubėja.

5. Kurie kintamieji diskretieji, o kurie tolydieji:
   a) vaikų skaičius šeimoje,
   b) vidutinė TV žiūrėjimo per savaitę trukmė,
   c) benzino kiekis bake,
   d) sesijos pažymių vidurkis,
   e) dešimties žmonių vidutinis ūgis.

6. Kurie kintamieji kokybiniai, o kurie kiekybiniai:
   a) vaikų skaičius šeimoje,
   b) komunaliniai mokesčiai (litais),
   c) pritarimas mirties bausmei,
   d) maksimalus automobilio greitis,
   e) grietinės indelio dydis,
   f) akių spalva,
   g) gimimo metai.

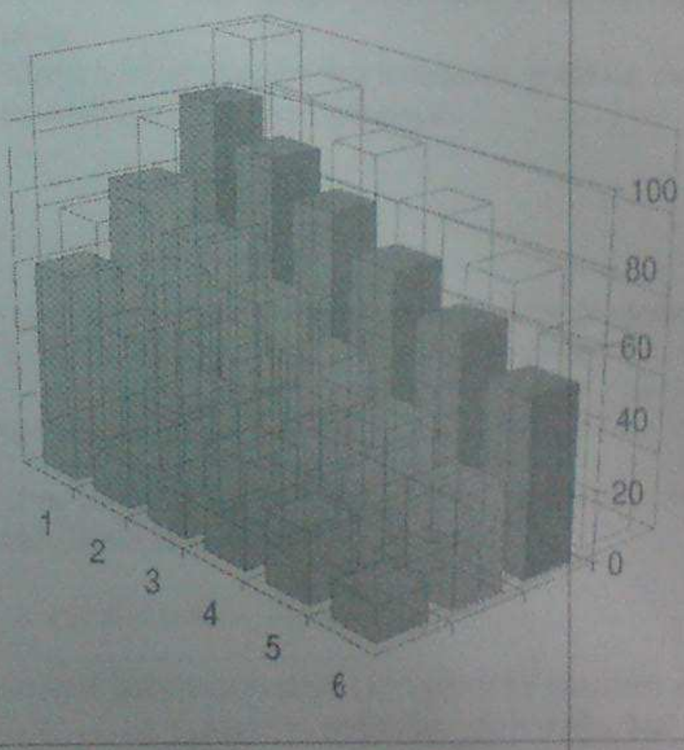
7. Pagal kokią skalę matuoti kintamieji:
   a) akių spalva,
   b) telefono numeris,
   c) batų numeris,
   d) religija,
   e) metų, praleistų mokymosi įstaigose, skaičius,
   f) regos stiprumas.

8. Pagal kokią skalę matuoti kintamieji, jeigu klausimyne pateikti tokie klausimai:
   a) kaip vertinate savo galimybes po studijų gauti gerai apmokamą darbą: geros – vidutinės – menkos,
   b) gimimo vieta,
   c) amžius,
   d) kuriam socialiniam sluoksniui priklausote: aukštesniajam – vidutiniam – žemesniajam,
   e) kuriam socialiniam sluoksniui priklausote: aukštesniajam – vidutiniam – žemesniajam – nežinau – netikiu skirstymais į socialinius sluoksnius,
   f) dėstytojai per paskaitas neturi teisės kritiškai vertinti vyriausybės veiklos: sutinku – nesutinku,
   g) ar dažnai darote namų darbus: visuomet – dažniau darau, nei nedarau – dažniau nedarau, nei darau – niekada,
   h) įstojimas į Europos Sąjungą yra savaiminis gėris: visiškai sutinku – iš dalies sutinku – nesutinku,
   i) prezidento politikai: visuomet pritariu – beveik visuomet pritariu – beveik visuomet nepritariu – visuomet nepritariu.

9. Kodėl negalima manyti, kad kategoriniai kintamieji sudaryti teisingai (pasirinkite vieną):
   a) Madona – tai: a) Dievo motina, b) dainininkė.
   b) Šią savaitę darbuotojas neatėjo į darbą: pirmadienį – antradienį – trečiadienį – ketvirtadienį – penktadienį. Kaip pakeisti šį kintamąjį, kad jis taptų teisingas?
   c) Kokią mašiną norėtumėte vairuoti: *Mercedes* – *BMW* – *Audi* – tanką – kitą?

1

APRAŠOMOJI STATISTIKA



Turime būti atsargūs ir nemaišyti duomenų su abstrakcijomis, kurias naudojame jiems tirti.

V. Džeimsas

Vidutiniškai gabus žmogus gali ramintis tuo, kad pusė žmonijos nė kiek ne gabesnė už jį.

Aprašomoji statistika – tai duomenų sisteminimo ir grafinio vaizdavimo metodai. Aprašomosios statistikos metodų taikymas yra labai svarbus statistinio uždavinio sprendimo etapas. Dažnai išsamus surinktos informacijos aprašymas bei duomenų grafikai leidžia daryti pagrįstas išvadas apie visos populiacijos nagrinėjamus požymius.

Vienas iš didžiausių aprašomosios statistikos privalumų yra tai, kad ji leidžia koncentruotai užrašyti informaciją, esančią dideliuose duomenų masyvuose. Todėl aprašomoji statistika gali būti taikoma ir visos populiacijos duomenims apdoroti.

> Jeigu skaičiuojant naudojami visos populiacijos duomenys, tai rezultatas vadinamas *populiacijos parametru.*
>
> Jeigu skaičiuojant naudojami imties duomenys, tai rezultatas vadinamas *imties statistika.*

Šioje knygoje populiacijos parametrai žymimi graikiškomis raidėmis, o imties statistikos – lotyniškomis.

Iš pradžių susipažinsime su pagrindiniais vieno kintamojo duomenų aprašymo etapais ir aprašomosios statistikos sąvokomis.

Aprašomojoje statistikoje stebėjimo reikšmės pateikiamos lentelėmis, grafikais, dažnių skirstiniais arba charakteristikomis, susijusiomis su šiais skirstiniais.

1. Duomenų grupavimas

1.1. Įvadinės pastabos

Tarkime, kad stebimas tam tikras kintamasis. Populiaciją laikysime turinčia N elementų. Atsitiktinai išrinkę n elementų, sudarome kintamojo reikšmių *statistinę* eilutę:

$$x_1, x_2, \ldots, x_n, \quad n \leqslant N. \tag{$*$}$$

Toliau duomenų aibę visada laikysime paprastosios atsitiktinės imties stebėjimo rezultatą.

Išdėstyta nemažėjimo tvarka kiekybinio kintamojo duomenų eilutė

$$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \cdots \leqslant x_{(n)}$$

vadinama *variacine* eilute. Skliaustuose pažymėtą skaičių (j) vadinsime elemento *eilės numeriu*, o reikšmę $x_{(j)}$, $(j = 1, 2, \ldots, n)$ – pozicine statistika. Akivaizdu, kad $x_{(j)}$ nebūtinai sutampa su x_j. Be to, žymėsime $x_{\min} = x_{(1)}$, $x_{\max} = x_{(n)}$. Pavyzdžiui, vertiname 7 studentų IQ (intelekto koeficientą)[1]. Gauname tokius duomenis:

$$100,\ 150,\ 120,\ 98,\ 100,\ 130,\ 95.$$

[1] *Intelligence quotient* – specialiu testu nustatomas rodiklis, nusakantis žmogaus protinių gebėjimų lygį.

Variacinė eilutė yra šitokia:

$$95,\ 98,\ 100,\ 100,\ 120,\ 130,\ 150.$$

Be to,

$$x_{\min} = x_{(1)} = 95, \qquad x_{\max} = x_{(7)} = 150, \qquad x_{(5)} = 120.$$

Tarkime, vienu metu tiriami keli, sakykime, k kintamųjų. Tokiu atveju, tirdami n objektų, gauname $n \times k$ eilės *duomenų matricą*

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nm} \end{pmatrix}.$$

Dažniausiai tokia matrica išdėstoma šitaip: eilutėmis žymimi objektai, stulpeliais – kintamieji. Atskira eilutė vadinama *stebėjimu* (realizacija).

1.2. Dažnių lentelės

Statistinėje eilutėje kintamojo X reikšmės gali kartotis. Tegul (∗) statistinėje eilutėje yra k skirtingų reikšmių. Tarkime, kad skirtingosios reikšmės yra $x_1, x_2, \dots, x_k$. Galima suskaičiuoti, kiek kartų statistinėje eilutėje pasikartojo kiekviena reikšmė, ir rasti, kurią visų stebėjimų dalį ji sudarė. Sakykime, kad stebima reikšmė x_j pasikartojo f_j kartų. Tuomet $f_1 + f_2 + \cdots + f_k = n$, o x_j statistinėje eilutėje sudaro f_j/n dalį visų stebėjimų.

Kintamojo reikšmės *dažnis* f_j – tai skaičius, nusakantis, kiek kartų reikšmė x_j pasikartojo statistinėje eilutėje.

Kintamojo reikšmės *santykinis dažnis* f_j/n – tai skaičius, nusakantis, kurią statistinės eilutės dalį sudaro x_j.

Skaičiuojami kiekybinių ir kokybinių kintamųjų dažniai ir santykiniai dažniai. Jei stebimasis kintamasis įgyja nedaug skirtingų reikšmių, tai duomenys surašomi į dažnių arba santykinių dažnių lenteles. Taip pateiktą informaciją daug lengviau suvokti bei pastebėti įvairias duomenų aibės savybes (pvz., dažniausiai pasikartojančią reikšmę, mažiausiąją reikšmę).

Katės uoslė 14 kartų geresnė už žmogaus. Biblijoje šuo paminėtas 14 kartų, liūtas – 89 kartus, katė – nė karto.
Kas dešimtas žmogus gyvena saloje. Vien Indonezijoje gyvena apie 200 mln. žmonių.
Pasaulyje yra dukart daugiau dviračių nei automobilių. 50% visų dviračių yra Kinijoje.

Duomenims sisteminti naudojami ir *sukauptieji* bei *santykiniai sukauptieji dažniai*.

1.1 lentelė. Dažniai

Reikšmė	x_1	x_2	x_3	...	x_k
Santykinis dažnis	f_1/n	f_2/n	f_3/n	...	f_k/n
Sukauptasis santykinis dažnis	f_1/n	$(f_1+f_2)/n$	$(f_1+f_2+f_3)/n$	...	$(f_1+\cdots+f_k)/n = 1$

Vietoje santykinių dažnių galima rašyti procentus. Kadangi vienas populiacijos procentas yra šimtoji jos dalis, tai šiuo atveju ta pati informacija pateikiama tiktai kita forma.

Santykinių dažnių lentelė dar vadinama kintamojo *dažnių* (empiriniu) *skirstiniu.*

1.1 pavyzdys. Tarkime, kad 50 studentų išreiškė savo požiūrį į „muilo operas":

labai patinka	patinka	neturiu nuomonės	nepatinka
labai patinka	patinka	patinka	neturiu nuomonės
nepatinka	neturiu nuomonės	neturiu nuomonės	neturiu nuomonės
nepatinka	nepatinka	labai nepatinka	labai nepatinka
labai patinka	nepatinka	labai nepatinka	labai patinka
patinka	neturiu nuomonės	patinka	neturiu nuomonės
labai patinka	patinka	labai patinka	patinka
neturiu nuomonės	patinka	patinka	neturiu nuomonės
labai patinka	labai patinka	patinka	neturiu nuomonės
neturiu nuomonės	labai patinka	labai nepatinka	labai nepatinka
patinka	neturiu nuomonės	patinka	neturiu nuomonės
patinka	neturiu nuomonės	patinka	patinka
neturiu nuomonės	labai nepatinka		

1.2 lentelė. Požiūris į „muilo operas"

Reikšmė	Dažnis	Sukauptasis dažnis	Santykinis dažnis	Sukauptasis santykinis dažnis
Labai patinka	9	9	0,18 (9:50)	0,18
Patinka	15	24	0,30 (15:50)	0,48
Neturiu nuomonės	15	39	0,30	0,78
Nepatinka	5	44	0,10	0,88
Labai nepatinka	6	50	0,12	1,00
	50			

Be abejo, iš taip pateiktų duomenų sunkoka suprasti, koks požiūris vyrauja. Tuo tarpu dažniai ir sukauptieji dažniai informatyvesni (1.2 lentelė).

Taigi labai nedaug studentų turi nepalankią nuomonę apie „muilo operas". Iš 1.2 lentelės matyti, kam gali būti naudojami sukauptieji dažniai. Pavyzdžiui, 24 studentams tokios operos patinka arba labai patinka, o 0,78 (78%) tirtųjų studentų nėra prieš jas nusistatę.

1.2 pavyzdys. Šeši šimtai atsitiktinai parinktų piliečių nurodė savo tikėjimą. SPSS[1] programų paketu rastas dažnių skirstinys parodytas 1.1 paveiksle.

Atkreipiame dėmesį į gautosios lentelės visų reikšmių procentų (*percent*) ir galiojančių reikšmių procentų (*valid percent*) stulpelius. SPSS programoje trūkstamos reikšmės pateikiamos kaip atskira reikšmių kategorija. Todėl iš pradžių nurodomi ir trūkstamų reikšmių procentai. Pateiktame pavyzdyje trūko 5 atsakymų ir tai sudarė 0,8% visų duomenų.

Pastaba. Programų paketų rezultatai (procentais) dažnai apvalinami, todėl jų suma gali keliomis šimtosiomis skirtis nuo 100%.

[1] SPSS – angl. *Statistical Package for Social Sciences.*

TIKĖJIMAS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Katalik.	380	63,3	63,9	63,9
	Protest.	2	,3	,3	64,2
	Stačiat.	29	4,8	4,9	69,1
	Kitas	93	15,5	15,6	84,7
	Joks	91	15,2	15,3	100
	Total	595	99,2	100	
Missing	System Missing	5	,8		
	Total	5	,8		
Total		600	100		

1.1 pav. SPSS paketu gautas dažnių skirstinys

Statistinius tyrimus (gyventojų surašymus) atliko jau senovės egiptiečiai. Tačiau pirmasis statistinis tyrimas, turėjęs didžiulės įtakos statistikos mokslo vystymuisi, buvo 1662 metais pasirodęs Londono gyventojų mirties priežasčių aprašymas. Dž. Grauntas (1620–1674) ištyrė 20 metų mirčių registracijų įrašus[2] *ir pateikė dažnių lenteles. Apie tuometinį mirtingumo lygį liudija tokie skaičiai: iš 100 naujagimių 6 metų sulaukdavo* 64, 16 *metų* – 40, 26 *metų* – 25, 36 *metų* – 16, 46 *metų* – 10, 56 *metų* – 6, 66 *metų* – 3.

Kiekybiniams kintamiesiems galima apibrėžti ne tik dažnių skirstinį, bet ir dažnių (empirinę) pasiskirstymo funkciją.

Dažnių (empirinė) pasiskirstymo funkcija

$$F(x) = \frac{\text{stebėjimų, ne didesnių už } x, \text{ skaičius}}{n}, \quad -\infty < x < \infty.$$

Dažnių pasiskirstymo funkcija atspindi visą sukauptąjį santykinį dažnį iki x. Iš $F(x)$ apibrėžimo išplaukia, kad $0 \leqslant F(x) \leqslant 1$. Kai kuriems tyrimams naudojamas $F(x)$ papildinys iki 1 – vadinamoji *garantijų* funkcija.

Garantijų funkcija $G(x) = 1 - F(x)$.

Garantijų funkcija parodo, kokią dalį sudaro imties reikšmės, didesnės už x. Pavyzdžiui, $G(3) = 0{,}2$ reiškia, kad 20% duomenų aibės stebėjimų viršija 3. Garantijų funkcija taikoma *meteorologijoje*. Daugiamečiai stebėjimai leidžia sudaryti temperatūrų, kritulių ir pan. garantijų funkcijas. Naudodami garantijos funkciją, galime rasti, kiek procentų liepos mėnesių vidutinė temperatūra buvo didesnė už 23°C. Su garantijų funkcija susijęs ir toks teiginys: 99% iš visų stebėtų dienų iškritusių kritulių kiekis buvo ne mažesnis už 25 mm.

Santykiniai dažniai nusako, kurią duomenų dalį sudaro kiekviena reikšmė. Tačiau kartais reikšmių dažnius norima palyginti tarpusavyje, t. y. apskaičiuoti f_j/f_i.

Pavyzdžiui, jau nagrinėtoje imtyje vienam protestantui tenka 190 katalikų. Dažnių santykį nusako ir tokios frazės: „filologų yra penkis kartus mažiau nei filologių“, „praėjusį sezoną rinktinė laimėdavo triskart dažniau, nei pralaimėdavo“. Apibrėžėme *vieno* dažnių

[1] J. Grount, *Observations Made Upon the Bills of Mortality*, 1662.

skirstinio dažnių santykį. Šio santykio nereikia maišyti su *kelių* skirtingų duomenų aibių dažnių santykiu. Pavyzdžiui, demografiniams tyrimams naudojamas gimimų ir mirčių santykis (trims gimimams tenka dvi mirtys ir pan.), bet jis nėra vieno skirstinio dažnių santykis.

1.3. Grupuotieji duomenys

Kai turime daug tolydžiojo kintamojo stebėjimų, dažnių lentelė tampa nebeinformatyvi – joje yra labai daug skirtingų reikšmių. Kartu dingsta didžiausias dažnių lentelės pranašumas, nes informacija nebekoncentruojama. Be to, kai kurie stebėjimai gali labai mažai skirtis tarpusavyje. Tokius duomenis reikia grupuoti. Prieš tai reikia nustatyti: 1) grupavimo intervalų skaičių, 2) jų plotį, 3) intervalų kraštinius taškus.

Dažniausiai pasirenkama nuo 5 iki 15 intervalų. Jeigu duomenų aibė gana simetriška, tai intervalų skaičių patariama rinkti pagal tokią taisyklę:

$$k = 1 + 3{,}222 \cdot \log_{10} n,$$

čia k – intervalų skaičius, n – duomenų aibės didumas.

Grupavimo intervalų ilgiai yra vienodi, intervalai *nesikerta*, kiekviena kintamojo reikšmė patenka tik į vieną intervalą. Kuo grupavimo intervalų skaičius didesnis, tuo mažiau informacijos prarandame.

Pažymėkime i-ąjį grupavimo intervalą $(c_{i-1}, c_i]$. Grupuodami imame atvirus iš kairės intervalus. Žinoma, galima imti ir atvirus iš dešinės intervalus. Svarbu tik kiekvienam duomeniui vienareikšmiškai parinkti tinkamą intervalą. Kartais net reikalaujama intervalų galus parinkti taip, kad jie nesutaptų su jokiu duomenų aibės elementu. Tuomet nebesvarbu, kuris intervalo galas atviras, o kuris uždaras, nes visi vėlesni skaičiavimai sutampa. Lentelėje 1.3 pateiktas f_i reikšmių, patekusių į $(c_{i-1}, c_i]$, dažnis.

Sisteminant kiekybinius duomenis, labai svarbi yra *empirinio tankio* funkcija. Grupuotų duomenų empirinė tankio funkcija

$$f_n(x) = \begin{cases} 0, & \text{kai } x < c_0, \\ f_j/(nh), & \text{kai } c_{j-1} \leqslant x < c_j, \\ 0, & \text{kai } x \geqslant c_k. \end{cases}$$

Čia h – intervalo ilgis (visiems vienodas).

Duomenis sugrupavus, dingsta informacija apie konkrečią kiekvieno duomens reikšmę. Todėl į konkretų intervalą pakliuvusiems duomenims apibūdinti imamas *intervalo vidurys*.

Grupuotų kiekybinių duomenų dažnių pasiskirstymo funkcija apibrėžiama kaip ir negrupuotųjų, tiktai *visi* patekę į intervalą $[c_{j-1}, c_j)$ duomenys yra laikomi lygiais $(c_{j-1} + c_j)/2$, t. y. intervalo viduriui.

1.3 lentelė. Intervalinių dažnių lentelė

Intervalas	$(c_0, c_1]$	$(c_1, c_2]$	$(c_2, c_3]$	...	$(c_{k-1}, c_k]$
Dažnis	f_1	f_2	f_3	...	f_k

1.4 lentelė

Amžius	Dažnis	Vidurys
18–19	10	18,5
20–21	17	20,5
22–23	35	22,5
...	...	...

1.5 lentelė

Tradicinis būdas			Alternatyvus būdas		
Amžius	Dažnis	Vidurys	Amžius	Dažnis	Vidurys
(17,5; 19,5]	10	18,5	[18, 20)	10	19
(19,5; 21,5]	17	20,5	[20, 22)	17	21
(21,5; 23,5]	35	22,5	[22, 23)	35	22
...	...	...	...	...	...

Praktiškai dažnai susiduriama su intervalais, tarp kurių yra „tarpų". Pavyzdžiui, surašinėjant [?] statant banko klientų amžių, gauta 1.4 duomenų lentelė.

Mes be vargo priskyrėme kiekvieną klientą atitinkamo amžiaus grupei. Tačiau, norėdami tokią lentelę pavaizduoti grafiškai, susidursime su problemomis. Mat atsiras tarpai tarp skaičių 19 ir 20, 21 ir 22, ir pan. Tokie tarpai negeidautini. Todėl aprašomojoje statistikoje intervalų ribas įprasta praplėsti taip, kad jie susiliestų. Tradicinis tokio praplėtimo metodas – prie kiekvieno intervalo krašto pridėti pusę „tarpo". Be abejo, intervalų ribas galima praplėsti ir atsižvelgiant į matavimų prigimtį. Pateiktame pavyzdyje pirmai grupei priskyrėme visus, kuriems ne mažiau kaip 18 metų, bet *mažiau* kaip 20 metų, ir pan. Todėl intervalų ribas galime praplėsti taip, kaip parodyta 1.5 lentelėje.

Tradicinis būdas *visuomet* išsaugo tuos pačius vidurinius intervalo taškus.

1.4. Dažnių skirstinio grafikai

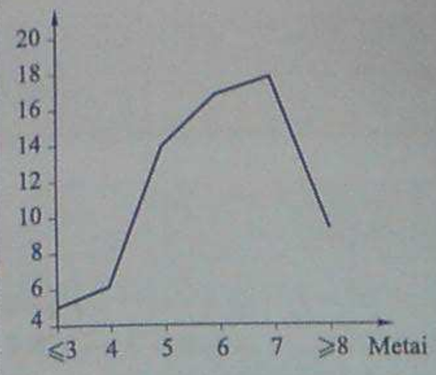
Paprasčiausias dažnių skirstinį iliustruojantis grafikas yra *dažnių daugiakampis*. Dažnių daugiakampis gaunamas Dekarto koordinatėse atidėtas dažnių reikšmes sujungus atkarpomis.

1.3 pavyzdys. Septyniasdešimt automobilių savininkų nurodė, prieš kiek metų pagamintas jų automobilis. Duomenys pateikti 1.6 lentelėje.

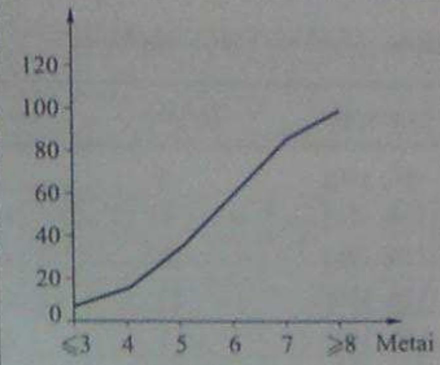
1.6 lentelė. Automobilių naudojimo trukmė

Metai	$\leqslant 3$	4	5	6	7	$\geqslant 8$
Automobilių skaičius	5	6	14	17	18	10

Nesunku nubraižyti šią lentelę atitinkantį dažnių daugiakampį ir sukauptųjų dažnių laužtę (žr. 1.2 ir 1.3 pav.).



1.2 pav. Automobilių „amžiaus“ dažnių daugiakampis



1.3 pav. Automobilių „amžiaus“ sukauptųjų dažnių laužtė

Kartais braižomas ir santykinių dažnių daugiakampis. Taip pat galima nubraižyti sukauptųjų dažnių ar sukauptųjų santykinių dažnių laužtę. Dažniausiai braižoma *sukauptųjų santykinių dažnių laužtė* ar sukauptųjų procentų (kai santykiniai grafikai išreiškiami procentais) laužtė. Abiejų šių grafikų forma identiška, nes vienintelis skirtumas yra kitoks ordinačių ašies mastelis – vienetą atitinka 100%.

Kaip ir daugiakampyje, atidėti sukauptieji procentai sujungiami atkarpomis.

Grupuotiems duomenims dažniausiai braižoma histograma.

Empirinės grupuotų duomenų tankio funkcijos grafikas vadinamas *histograma*.

Histograma braižoma taip:

1) Ox ašyje atidedami grupavimo intervalai,

2) kiekviename intervale braižomas stačiakampis, kurio aukštinė proporcinga pakliuvusių į intervalą reikšmių skaičiui (f_j/n). Šiaip jau reikalaujama, kad visų stačiakampių plotų suma būtų lygi 1. Šis reikalavimas esminis tikimybinei interpretacijai, tačiau nelabai svarbus grafiko formai. Mat ordinačių (Oy) ašies mastelis vis tiek skiriasi nuo Ox ašies mastelio (kitu atveju būtų sunkoka ką nors įžiūrėti). Pirmos lentelės duomenų histograma braižoma taip: intervale $(c_0, c_1]$ braižomas stačiakampis, kurio aukštinė yra $f_1/(nh)$, intervale $(c_1, c_2]$ – stačiakampis, kurio aukštinė yra $f_2/(nh)$ ir pan.

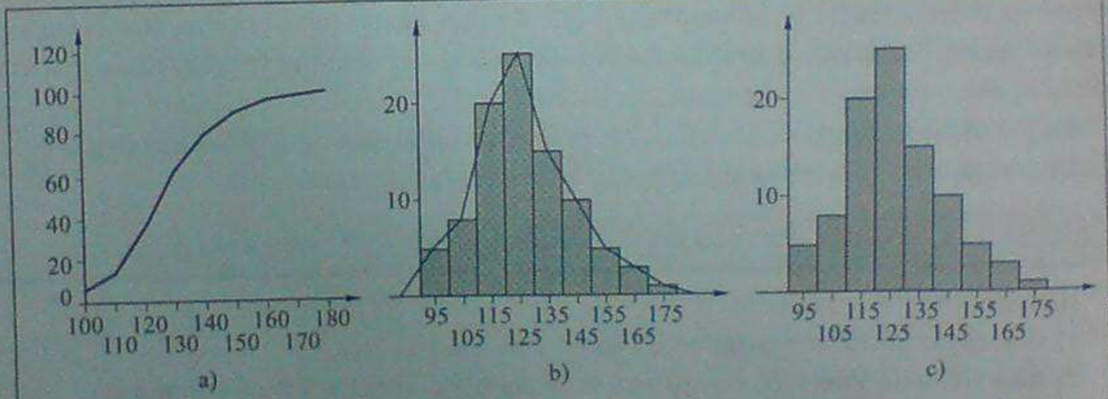
Galima nubraižyti grupuotų duomenų dažnių daugiakampį. Jį gautume histogramoje sujungę stačiakampių viršutinių briaunų vidurio taškus. Dažnių daugiakampis irgi atskleidžia empirinio tankio elgesį. Dažnių daugiakampį galima braižyti ir be histogramos. Tam užtenka visus į konkretų intervalą patekusius taškus prilyginti intervalo viduriui ir braižyti įprastą dažnių daugiakampio grafiką. Panašiai braižoma ir sukauptųjų dažnių laužtė. Tačiau šiuo atveju visi į intervalą patekę taškai prilyginami dešiniajai kraštinei intervalo reikšmei. Dažnių daugiakampiui imami viduriniai grupuotų duomenų intervalų taškai; sukauptųjų dažnių laužtei imami kraštiniai intervalų taškai.

1.4 pavyzdys. Grupei ligonių buvo matuojamas sistolinis kraujo spaudimas. Gautus duomenis sugrupavę (ir praplėtę intervalus), gauname dažnių skirstinį, kuris parodytas 1.7 lentelėje.

Sukauptųjų (procentinių) dažnių laužtė, dažnių daugiakampis ir histograma parodyti 1.4 paveiksle.

1.7 lentelė. Sistolinis kraujo spaudimas

Intervalas	Dažnis	Intervalo vidurys
(90, 100]	5	95
(100, 110]	8	105
(110, 120]	20	115
(120, 130]	25	125
(130, 140]	15	135
(140, 150]	10	145
(150, 160]	5	155
(160, 170]	3	165
(170, 180]	1	175



1.4 pav. a) sukauptųjų dažnių laužtė; b) dažnių daugiakampis ir histograma; c) histograma

2. Duomenų padėties charakteristikos

Pagrindinės duomenų padėties charakteristikos yra vidurkis, moda ir mediana, apibūdinantys duomenų „centrą“, bei kvantiliai. Visos charakteristikos, išskyrus modą, skaičiuojamos tik kiekybiniams duomenims.

2.1. Vidurkis

Vidurkis – tai taškas, kuris vidutiniškai artimiausias visiems statistinės eilutės elementams. Skaičiuojamas tik kiekybinių duomenų vidurkis. Tarkime, kad M koks nors skaičius. Atstumą tarp M ir statistinės eilutės elementų matuojame taip:

$$f(M) = (x_1 - M)^2 + (x_2 - M)^2 + \cdots + (x_n - M)^2.$$

Funkcija $f(M)$ pasiekia minimumą taške $(x_1+x_2+\cdots+x_n)/n$. Pastarąjį skaičių vadiname *imties vidurkiu (vidurkiu, aritmetiniu vidurkiu, empiriniu*[1] *vidurkiu)* ir žymime $\overline{x}$. Taigi vidurkis yra ne kas kita kaip visų statistinės eilutės elementų suma, padalyta iš jų skaičiaus. Analogiškai apibrėžiamas ir populiacijos vidurkis.

Imties vidurkis $$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j. \qquad (1.1)$$

Populiacijos vidurkis $$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j. \qquad (1.2)$$

Grupuotų duomenų vidurkiui skaičiuoti pasirenkami viduriniai intervalų taškai. Tegul intervalai ir dažniai apibrėžti 1.1 lentele. Pirmo intervalo vidurio taškas $x_1^* = c_0 + h/2 = (c_0 + c_1)/2$ (čia h – intervalo ilgis), j-ojo $x_j^* = (c_{j-1} + c_j)/2$.

Imties vidurkis

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j^* f_j = \sum_{j=1}^{n} x_j^* \frac{f_j}{n}. \qquad (1.3)$$

Pirmoji (1.3) formulės lygybė yra ne kas kita kaip vidurkio apibrėžimas. Antroji lygybė išplaukia iš sandaugos distributyvumo. Ji parodo, kaip vidurkiui skaičiuoti gali būti panaudoti santykiniai dažniai.

1.5 pavyzdys. Dirbančių studentų atlyginimai (Lt per mėn.) pateikti 1.8 lentelėje. Atlyginimo vidurkis

$$\overline{x} = (400 \cdot 7 + 500 \cdot 10 + 600 \cdot 13 + 700 \cdot 12 + 800 \cdot 10 + 900 \cdot 8)/60 = 653{,}333.$$

1.8 lentelė

Atlyginimas	Studentų skaičius	
400	7	7
500	10	17
600	13	30
700	12	
800	10	
900	8	

60

1.6 pavyzdys. Vidutinis sistolinis 1.4 pavyzdyje aprašytų ligonių kraujo spaudimas

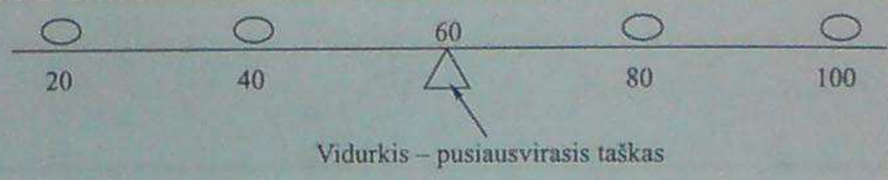
$$\overline{x} = (95 \cdot 5 + 105 \cdot 8 + \cdots + 175 \cdot 1)/92 = 126{,}739\ldots$$

Vidurkis yra labiausiai paplitusi duomenų padėties charakteristika – skaičiuojamas vidutinis atlyginimas, vidutinis energijos sunaudojimas, sesijos pažymių vidurkis, vidutinė

[1] Visos šioje dalyje nagrinėjamos charakteristikos yra empirinės. Dėl paprastumo žodis „empirinis“ praleidžiamas.

poinsultinės reabilitacijos trukmė ir pan. Vidurkis dažnai vadinamas vidutine stebėjimų reikšme.

Kartais patogu vidurkį įsivaizduoti kaip tam tikrą pusiausvirąjį tašką. Tegul duomenų aibė yra (20; 40; 80; 100). *Tuomet vidurkis lygus* 60.



1.5 pav.

Jei statistinėje eilutėje yra kelios labai išsiskiriančios iš kitų stebėjimų reikšmės (labai didelės arba mažos), vidurkis nėra itin geras matas, nes neatspindi to, kas būdinga daugumai stebėjimų. Pavyzdžiui, smarkiai pakėlus gamyklos direktoriaus atlyginimą, pakils ir vidutinis gamyklos darbuotojų atlyginimas, nors kitų darbuotojų atlyginimai ir nepasikeis.

Kartais vidurkį sunkoka interpretuoti. Tokiu atveju geriau naudoti kitas skaitines charakteristikas.

Dauguma žmonių turi didesnį nei vidutinis kojų skaičių. Iš tikrųjų tikėtina, kad tarp 4 milijonų Lietuvos gyventojų yra 1000 vienakojų. Taigi vidutinis kojų skaičius: $(3\,999\,000 \cdot 2 + 1000 \cdot 1)/4\,000\,000 = 1{,}9975$. *Bet dauguma turi dvi kojas, o* $2 > 1{,}9975$.

Vidurkio savybės:

1 *Pasinaikinimo efektas*:

$$\sum_{j=1}^{n}(x_j - \overline{x}) = 0.$$

2 *Daugyba iš konstantos*. Visas stebėjimo reikšmes padauginus iš to paties skaičiaus, gautasis aritmetinis vidurkis taip pat bus padaugintas iš šio skaičiaus:

$$\frac{1}{n}\sum_{j=1}^{n}(Cx_j) = \frac{C}{n}\sum_{j=1}^{n}x_j = C\overline{x}.$$

3 *Postūmis*. Pridėjus (arba atėmus) prie kiekvieno stebėjimo tam tikrą skaičių, vidurkis padidės (sumažės) tokiu pat skaičiumi:

$$\frac{1}{n}\sum_{j=1}^{n}(x_j + C) = \overline{x} + C.$$

Tarkime, turime m duomenų aibės poaibių. Pirmojo poaibio duomenų (n_1 stebėjimų) aritmetinis vidurkis yra $\overline{x}_1, \ldots, m$-ojo ($n_m$ stebėjimų) – $\overline{x}_m$.

Tuomet bendrasis aritmetinis vidurkis apskaičiuojamas šitaip:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_m\overline{x}_m}{n_1 + n_2 + \cdots + n_m}. \tag{1.4}$$

1.7 pavyzdys. Pirmoje grupėje mokosi 20 studentų, o antrojoje – 25 studentai. Pirmos grupės sesijos pažymių vidurkis 7,5 balo, antrosios – 8 balai. Tada bendrasis abiejų grupių sesijos pažymių vidurkis yra

$$\overline{x} = \frac{20 \cdot 7{,}5 + 25 \cdot 8}{20 + 25} = 7{,}777\ldots$$

Kelių grupių pažymių vidurkių vidurkis su bendruoju vidurkiu sutaps tik tuomet, kai visose grupėse bus tiek pat studentų.

2.2. *Moda*

Jau matėme, kad vidurkis ne visada yra pati tinkamiausia charakteristika. Tarkime, kad ufonautas paprašė duomenų apie vidutinį žmogų. Ir gavo atsakymą, kad vidutiniškai žmogus turi: kojų – 1,99...; akių – 1,99...; galūnės pirštų – 4,88... ir pan. Kaip manote, *kokį* žmogų įsivaizduos ufonautas? Šiuo atveju informatyviau būtų aprašyti tipišką žmogų. Tipiškiausia nagrinėjamos duomenų aibės reikšmė yra imties moda Mo. Moda – tai dažniausiai duomenų aibėje pasikartojusi reikšmė. Pavyzdžiui, duomenų aibės 1; 1; 2; 3; 4; 5 moda $Mo = 1$.

Jeigu visos reikšmės statistinėje eilutėje pasikartoja vienodai dažnai, sakoma, kad pasiskirstymas modos neturi. Pavyzdžiui, duomenų aibė 2,3; 2,3; 3,8; 3,8; 4,5; 4,5 modos neturi.

Jeigu kelių *gretimų* variacinės eilutės reikšmių dažnis vienodas ir yra didesnis negu bet kurių kitų reikšmių dažnis, tai moda yra šių reikšmių vidurkis. Pavyzdžiui, duomenų aibės 0; 1; 1; 2; 2; 2; 3; 3; 3; 4 moda $Mo = (2+3)/2 = 2{,}5$. Dažnių skirstinys, turintis vieną modą, vadinamas *unimodiniu* skirstiniu.

Jeigu dvi negretimos variacinės eilutės reikšmės pasikartoja vienodu dažniu ir jis didesnis negu bet kurių kitų reikšmių, tai egzistuoja dvi modos ir sakoma, kad dažnių skirstinys *bimodinis*. Pavyzdžiui, statistinė eilutė 10; 11; 11; 11; 12; 13; 14; 14; 14; 17 turi dvi modas – 11 ir 14. Jeigu negretimų vienodo dažnio variacinės eilutės narių yra daugiau nei du, modų taip pat yra daugiau. Toks dažnių skirstinys vadinamas *multimodiniu*.

Galima skaičiuoti tiek kiekybinių, tiek ir kokybinių duomenų modą.

Grupuotų duomenų moda apytiksliai lygi intervalo, į kurį pateko daugiausia duomenų, vidurinei reikšmei, 1.4 pavyzdžio $Mo = 125$. Tikslesnę modos skaičiavimo formulę galima rasti [3] knygelėje.

2.3. Mediana

Tarkime, kad turime variacinę eilutę

$$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \cdots \leqslant x_{(n)}.$$

Imties mediana Md yra skaičius, už kurį 50% variacinės eilutės reikšmių yra ne didesnės ir 50% ne mažesnės. Taigi mediana – tai skaičius, perskiriantis variacinę eilutę į dvi maždaug lygias dalis. Tikslesnis medianos apibrėžimas yra toks:

> Jeigu stebėjimų skaičius n nelyginis, tai mediana yra variacinės eilutės *reikšmė*, atitinkanti $(n+1)/2$ poziciją.
>
> Jeigu n lyginis, tai mediana yra variacinės eilutės *reikšmių*, atitinkančių pozicijas $(n/2)$ ir $(n/2)+1$, aritmetinis vidurkis.

Taigi

$$Md = \begin{cases} x_{((n+1)/2)}, & \text{kai } n \text{ – nelyginis}, \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{kai } n \text{ – lyginis}. \end{cases} \tag{1.5}$$

1.8 pavyzdys. Pavyzdžio apie studentų algas mediana $Md = (x_{(30)} + x_{(31)})/2 = (600 + 700)/2 = 650$.

Pastaba. Norime dar kartą atkreipti dėmesį, kad mediana dalija pusiau *variacinę* eilutę, t. y. jau sutvarkytus duomenis, o ne *statistinę* eilutę, t. y. pradinius duomenis.

Kaip ir aritmetinis vidurkis, mediana charakterizuoja duomenų centrą. Paprastai ją patariama naudotis, kai duomenų aibėje yra išskirčių. Išskirtis – tai tokia duomenų aibės reikšmė, kuri yra nenatūraliai didesnė ar mažesnė už kitas reikšmes. Tikslesnį išskirties apibrėžimą pateiksime vėliau.

Panagrinėkime dvi duomenų aibes: 10; 20; 30; 40 ir 10; 20; 30; 100. Abiejų aibių medianos yra lygios skaičiui 25, tačiau pirmosios aibės vidurkis yra 25 , o antrosios – 40. Aišku, kad antruoju atveju vidurkis nėra tinkama centro charakteristika, nes viską lemia vienintelė didelė reikšmė – išskirtis 100.

Mediana dažniausiai naudojama ranginiams duomenims ir intervaliniams – santykiniams duomenims, kuriuose yra išskirčių. Pavyzdžiui, butų kainoms (vieni butai naujuose rajonuose, kiti senamiestyje), įmonių metiniam pelnui (daug mažų įmonių ir viena įmonė – gigantė) skaičiuoti.

Paprasčiausias būdas rasti grupuotų duomenų medianą – visas į intervalą patekusias reikšmes pakeisti vidurinėmis intervalo reikšmėmis ir pritaikyti (1.5) formulę.

1.9 pavyzdys. Pateikto 1.4 pavyzdžio (žr. p. 32) apie sistolinį kraujo spaudimą $Md = (x_{(46)} + x_{(47)})/2 = (125 + 125)/2 = 125$.

Kartais grupuotų duomenų medianą patariama skaičiuoti atsižvelgiant į sukauptuosius dažnius (žr. [3]).

Empiriškai nustatyta, kad tolydaus kintamojo stebėjimams

$$\overline{x} - Mo \simeq 3(\overline{x} - Md).$$

Vadinasi, dažniausiai mediana yra tarp aritmetinio vidurkio ir modos. Be abejo, dažniausiai tai dar ne visuomet. Galima sukonstruoti statistinių eilučių, kurių mediana mažesnė ir už vidurkį, ir už modą, ir pan., tačiau tokios imtys praktiškai pasitaiko retai.

Kai dažnių pasiskirstymas simetrinis ir unimodalus, tai $\overline{x} = Md = Mo$.

Ir vidurkis, ir moda, ir mediana yra duomenų centro charakteristikos. Kokią charakteristiką geriau naudoti, priklauso nuo tyrimo tikslų.

1.10 pavyzdys. Nedidelėje firmoje dirbančiųjų pareigos ir alga nurodytos 1.9 lentelėje. SPSS paketu gauti rezultatai pateikti 1.6 paveiksle.

1.9 lentelė. Pareigos ir alga (Lt)

Prezidentas	10 000
Viceprezidentas	8000
Buhalteris	6000
Programuotojas A	2100
Programuotojas B	2000
Programuotojas C	1900
Programuotojas D	1800
Vairuotojas	750
Valytoja	750

STATISTICS

	N		Mean	Median	Mode
	Valid	Missing			
Alga	9	0	3700,00	2000,00	750,00

1.6 pav.

Vidutinė firmos darbuotojų alga (vidurkis) $\overline{x} = 3700$. Taigi vidutinė yra tokia alga, kai visi darbuotojai sudeda savo algas į krūvą ir po to pasidalija po lygiai. Moda $Mo = 750$ yra alga, kurią gauna daugiausia firmos darbuotojų (vairuotojas ir valytoja). Mediana $Md = 2000$ yra visų algų, išrikiuotų pagal didumą, viduryje.

2.4. *Kvantiliai*

Reikšmė, dalijanti variacinę eilutę į $q \times 100$ ir $(1-q) \times 100$ procentinių dalių, vadinama q-osios $(0 < q < 1)$ eilės *kvantiliu*. Kvantiliui skaičiuoti galima naudotis tokia procedūra:

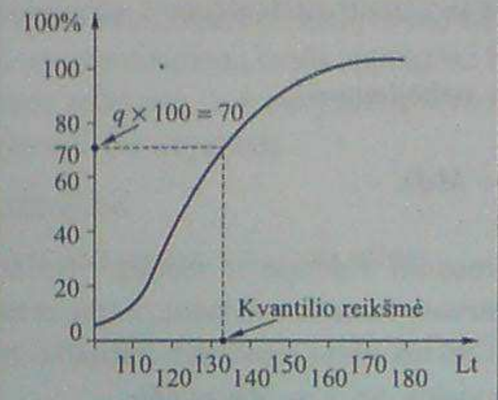
1 Randamas indeksas i:

$$i = q \cdot n.$$

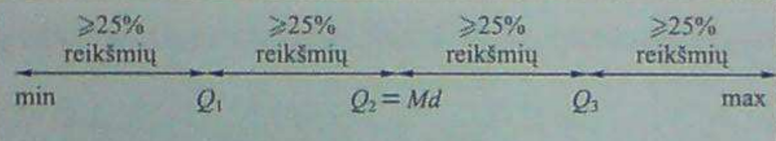
2 Jeigu i nėra sveikasis skaičius, tai imama sveikoji jo dalis $[i]$. Ieškomasis kvantilis yra $[i]+1$ variacinės eilutės narys, t. y. $x_{([i]+1)}$.

3 Jeigu i yra sveikasis skaičius, tai ieškomasis kvantilis – $(x_{(i)} + x_{(i+1)})/2$.

1.11 pavyzdys. Tarkime, norime rasti 1.1 lentelės duomenų 20% $(q = 0{,}2)$ kvantilį. Tuomet $i = 0{,}2 \cdot 9 = 1{,}8$. Kadangi i nėra sveikasis skaičius, tai imame jo sveikąją dalį $[i] = 1$. Ieškomasis kvantilis yra $x_{(1+1)} = x_{(2)} = 750$.



1.7 pav. Kvantilio skaičiavimas ($q = 0,7$)



1.8 pav. Kvartiliai

Dar vienas būdas kvantiliams skaičiuoti yra naudojantis procentine sukauptųjų dažnių laužte. Kaip tai padaryti, matyti iš 1.7 paveikslo.

Nėra vienos kvantilių skaičiavimo metodikos. Tiksliausias (ir sudėtingiausias) yra metodas, kai naudojamasi sukauptųjų dažnių laužte (šį metodą galima užrašyti ir formulėmis). Laimei, praktiškai, kai duomenų yra pakankamai, skirtingais metodais rasti kvantiliai mažai skiriasi, todėl galima taikyti anksčiau pateiktą skaičiavimo algoritmą.

Kvantiliai, dalijantys variacinę eilutę į keturias maždaug lygias dalis, vadinami *kvartiliais*. Jie žymimi Q_1, Q_2, Q_3.

Vienas iš kvartilių radimo metodų nusakomas taip (žr. 1.8 pav.): Q_2 sutampa su mediana ir dalija imtį į dvi dalis; tuomet Q_1 yra apatinės dalies mediana, o Q_3 yra viršutinės dalies mediana. Dažnai dydis $(Q_1 + Q_3)/2$ naudojamas kaip viena iš duomenų sankaupos (centro) charakteristikų, t. y. kaip alternatyva vidurkiui, medianai arba modai.

Variacinę eilutę galima dalyti į daugiau dalių. Skaičiai, dalijantys (suskirstantys) variacinę eilutę į 100 maždaug vienodų dalių, vadinami *procentiliais*. Taigi Md yra 50% kvantilis, Q_1 – 25% kvantilis, Q_3 – 75% kvantilis.

2.5. Nupjautieji vidurkiai

Vidurkis labai „jautrus“ išskirtims. Todėl praktiškai kartais taikomi ir nupjautasis bei triskaitis vidurkiai. Nupjautasis vidurkis skaičiuojamas atmetus tam tikrą procentą mažiausių ir didžiausių duomenų aibės reikšmių. Pavyzdžiui, 50% nupjautasis vidurkis yra variacinės eilutės reikšmių vidurkis, suskaičiuotas atmetus po 25% didžiausių ir mažiausių reikšmių. Nupjautasis vidurkis taikytinas, kai duomenų aibės reikšmių pasiskirstymas yra asimetriškas. Jis skaičiuojamas ir tuomet, kai ekstremalios reikšmės yra nepatikimos. Nupjautasis vidurkis dažnai naudojamas vertinant sportininkų pasirodymus (šuolių į vandenį, dailiojo čiuožimo ir pan.), kai norima sumažinti ekstremalių balų įtaką, kurią lemia šališki teisėjai. Aritmetinis vidurkis yra 0% nupjautasis vidurkis.

Triskaitis vidurkis skaičiuojamas pagal tokią formulę:

$$\text{Triskaitis vidurkis} = \frac{Q_1 + 2Md + Q_3}{4}.$$

Ši charakteristika beveik tokia pat „nejautri“ ekstremalioms reikšmėms kaip ir mediana, todėl naudojama asimetrinių reikšmių skirstiniams. Ji nėra efektyvi, kai skirstinys normalusis (žr. 5).

3. Duomenų sklaidos charakteristikos

Palyginkime dviejų firmų programuotojų atlyginimus. Pirmojoje firmoje dirba penki programuotojai, per mėnesį uždirbantys 1000; 2000; 3000; 5000 ir 9000 Lt. Atlyginimų vidurkis – 5000 Lt. Antrojoje firmoje dirba penki programuotojai, uždirbantys 5000; 5000; 5000; 5000 ir 5000 Lt; atlyginimų vidurkis – 5000 Lt. Taigi abiejų firmų programuotojų *vidutiniai* atlyginimai sutampa. Tačiau matome, kad pirmojoje firmoje yra ir didelių, ir mažų atlyginimų, o antrojoje visi atlyginimai vienodi.

Šiame skyrelyje aptarsime skaitines charakteristikas, leidžiančias įvertinti duomenų sklaidą.

Duomenų sklaida nėra tas pats, kas duomenų įvairovė. Sklaidos charakteristikos parodo, kiek duomenys skiriasi. Dešimties atlyginimų (kurie visi skirtingi, bet skiriasi ne daugiau kaip vienu litu) duomenų aibė daug mažiau išsklidusi nei trys, bet šimtais tūkstančių litų besiskiriantys atlyginimai.

Pagrindinės sklaidos charakteristikos yra duomenų aibės plotis, vidutinis nuokrypis, dispersija, standartinis nuokrypis, kvartilių skirtumas ir kitimo koeficientas. Tai – kiekybinių kintamųjų charakteristikos. Skyrelio pabaigoje pateikiama ir viena kokybinių duomenų sklaidos charakteristika.

3.1. Dispersija

Imties dispersija parodo duomenų sklaidą apie vidurkį.

Imties dispersija $$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2. \tag{1.6}$$

Populiacijos dispersija $$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2. \tag{1.7}$$

Skyrelio pradžioje minėtame pavyzdyje pirmosios firmos programuotojų atlyginimų dispersija:

$$\frac{1}{5}\big((1000 - 5000)^2 + (2000 - 5000)^2 + \cdots + (10\,000 - 5000)^2\big) = 14\,000\,000.$$

Antrosios firmos atlyginimų dispersija:

$$\frac{1}{5}\big((5000 - 5000)^2 + (5000 - 5000)^2 + \cdots\big) = 0.$$

Taigi pirmosios firmos atlyginimų dispersija (ir sklaida) didelė, o antrosios maža.

Dispersija – viena iš populiariausių sklaidos charakteristikų. Jos privalumas yra tas, kad atsižvelgiama į visus duomenis ir pateikiamas *vidutinis* (dalijame iš $(n-1)$ arba N) skirtumų nuo vidurkio kvadratas. Dispersija plačiai naudojama lyginant kelių duomenų aibių sklaidas.

Apibrėždami imties dispersiją dalijame iš $(n-1)$, o ne iš n. Tolesniuose skyreliuose bus parodyta, kad tokia formulė labiau tinka, kai stebėjimams pasirenkami matematiniai modeliai (ypač, kai duomenų yra nedaug).

Iš apibrėžimo akivaizdu, kad dispersija visuomet neneigiama. Be to, dispersija lygi nuliui tik tuo atveju, kai visi stebėjimai lygūs. Dispersijos savybės:

1 *Inertiškumas postūmiui.* Pridėjus (arba atėmus) prie kiekvieno stebėjimo tą patį skaičių, dispersija nesikeičia:

$$s^2(x+C) = \frac{1}{n-1}\sum_{j=1}^{n}\left((x_j+C)-\overline{x}-C\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j-\overline{x}\right)^2 = s^2(x).$$

2 *Daugyba iš konstantos.* Visas stebėjimo reikšmes padauginus iš to paties skaičiaus, pradinių reikšmių dispersija yra dauginama iš šio skaičiaus kvadrato:

$$s^2(Cx) = \frac{1}{n-1}\sum_{j=1}^{n}\left(Cx_j-C\overline{x}\right)^2 = C^2\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j-\overline{x}\right)^2 = C^2s^2(x).$$

Kodėl apibrėžiant dispersiją skirtumai nuo vidurkio keliami kvadratu? Todėl, kad atitinkamos sumos nesusiprastintų. Pavyzdžiui, duomenų aibės 50; 60; 70; 80; 90 vidurkis yra 70, o dispersija lygi 250. Tuo tarpu susumavę skirtumus nuo vidurkio be kvadratų, gautume: $(50-70)+(60-70)+(70-70)+(80-70)+(90-70)=0$, taigi jokios sklaidos neužfiksuotume.

Skaičiuojant grupuotų duomenų dispersiją, taikoma tokia formulė:

$$s_h^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j\left(x_j^*-\overline{x}\right)^2 - \frac{h^2}{12}, \tag{1.8}$$

čia h – grupavimo intervalo ilgis, f_j – j-ojo grupavimo intervalo dažnis, x_j^* – j-ojo grupavimo intervalo vidurio taškas, o vidurkis $\overline{x}$ skaičiuojamas pagal (1.3) formulę. Dėmuo $-h^2/12$ vadinamas Šepardo[1] pataisa.

1.12 pavyzdys. Raskime 1.4 pavyzdyje aprašytų ligonių sistolinio kraujo spaudimo dispersiją:

$$s_h^2 = \frac{1}{91}\left(5(95-126{,}739)^2 + 8(105-126{,}739)^2 + \cdots\right) - 100/12 = 355{,}084.$$

[1] William Fleetwood Sheppard (1863–1936) – anglų statistikas.

Skaičiuoti s^2 pagal (1.6) formulę nelabai patogu. Patogiau naudotis tokiu (1.6) formulės variantu:

$$s^2 = \frac{1}{n-1}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right) - \frac{n}{n-1}(\overline{x})^2. \tag{1.9}$$

1.13 pavyzdys. Apskaičiuokime 1.5 pavyzdžio studentų atlyginimų dispersiją:

$$s^2 = \frac{1}{59}(400^2 \cdot 7 + 500^2 \cdot 10 + 600^2 \cdot 13 + 700^2 \cdot 12 + 800^2 \cdot 10 + 900^2 \cdot 8)$$
$$- \frac{60}{59}(653{,}333)^2 = 24\,565.$$

3.2. Standartinis nuokrypis

Standartinis nuokrypis yra dažniausiai taikomas sklaidos matas. Jis gaunamas ištraukus kvadratinę šaknį iš dispersijos.

$$\textit{Imties standartinis nuokrypis } s = \sqrt{s^2}. \tag{1.10}$$

$$\textit{Populiacijos standartinis nuokrypis } \sigma = \sqrt{\sigma^2}. \tag{1.11}$$

Skyrelio pradžioje minėtos pirmosios firmos programuotojų atlyginimų standartinis nuokrypis yra $\sqrt{14000000} = 3741{,}657$ Lt. Studentų atlyginimų (1.5 ir 1.13 pavyzdžiai) standartinis nuokrypis lygus 156,73 Lt. Sistolinio kraujo spaudimo standartinis nuokrypis (1.4 ir 1.12 pavyzdžiai) lygus 18,84.

Kaip ir dispersija, standartinis nuokrypis parodo vidutinę duomenų sklaidą apie vidurkį. Ką išlošiame pereidami nuo dispersijos prie standartinio nuokrypio? Visų pirma pastebėsime, kad standartinis nuokrypis matuojamas *tokiais pačiais* vienetais kaip ir patys duomenys. Jeigu kalbame apie atlyginimus, tai ir duomenys, ir vidurkis, ir standartinis nuokrypis matuojami litais. Tuo tarpu dispersijos matavimo vienetai būtų litai kvadratu. Šiuo atžvilgiu standartinį nuokrypį lengviau interpretuoti ir lyginti su duomenimis. Kita svarbi standartinio nuokrypio naudojimo priežastis yra duomenų koncentracijos apie vidurkį tiesioginė priklausomybė nuo standartinio nuokrypio dydžio (žr. 5 ir 7).

3.3. Kitimo koeficientas

Kitimo (variacijos) koeficientas skaičiuojamas *tik* santykių skalės kintamiesiems, turintiems teigiamus vidurkius:

$$\overline{x} > 0. \tag{1.12}$$

Kitimo koeficientas yra *bedimensis* dydis. Jis naudojamas lyginant skirtingų duomenų aibių sklaidas.

$$\textit{Populiacijos kitimo koeficientas } CV = \frac{\sigma}{\mu}.$$
$$\textit{Imties kitimo koeficientas } cv = \frac{s}{\overline{x}}. \tag{1.13}$$

$$\textit{Procentinis populiacijos kitimo koeficientas } CVP = \frac{\sigma}{\mu}100\%.$$

$$\textit{Procentinis imties kitimo koeficientas } cvp = \frac{s}{\overline{x}}100\%. \tag{1.14}$$

1.14 pavyzdys. Svarbi akcijų charakteristika yra kainos stabilumas. Tarkime, tris mėnesius stebėjus akcijų kainų kitimą, buvo nustatyta vidutinė firmos A akcijų kaina – 200 Lt ir jų standartinis nuokrypis – 40 Lt. Firmos B vidutinė akcijų kaina – 48 Lt, standartinis nuokrypis – 12 Lt.

Firmos A akcijų kainos sklaida didesnė nei firmos B. Tačiau labai skiriasi patys akcijų kainų vidurkiai. Galima apskaičiuoti abiejų firmų akcijų kainų kitimą vidurkių atžvilgiu:

Firmos A
$$cvp = \frac{40}{200}100\% = 20\%.$$

Firmos B
$$cvp = \frac{12}{48}100\% = 25\%.$$

Taigi vidurkio atžvilgiu firmos A akcijos stabilesnės už firmos B akcijas.

Kitimo koeficientas taikomas ir lyginant skirtingais vienetais matuotų duomenų aibių sklaidą.

3.4. *Kitos sklaidos charakteristikos*

Paprasčiausia sklaidos charakteristika yra duomenų aibės plotis.

$$\textit{Duomenų aibės plotis} = x_{(n)} - x_{(1)} = x_{\max} - x_{\min}.$$

Duomenų aibės plotis labai jautrus išskirtims. Todėl dažniau skaičiuojamas kvartilių skirtumas (IQR).

$$IQR = Q_3 - Q_1.$$

Duomenų aibės pločiui ir kvartilių skirtumui skaičiuoti reikia tik kelių duomenų aibės reikšmių. Visų reikšmių prireikia vidutiniam nuokrypiui d rasti.

$$d = \frac{1}{n}\sum_{j=1}^{n}|x_j - \overline{x}|.$$

Vidutinis nuokrypis matuojamas tais pačiais vienetais kaip ir duomenys. Tačiau jis ne toks patogus naudoti kaip dispersija ar standartinis nuokrypis.

3.5. *Kokybinės įvairovės indeksas*

Paminėsime dar vieną sklaidos matą – kokybinės įvairovės indeksą IQV, kuris taikomas kategoriniams kintamiesiems.

$$IQV = \frac{k\left(n^2 - \left(f_1^2 + f_2^2 + \cdots + f_k^2\right)\right)}{n^2(k-1)}, \tag{1.15}$$

čia k – kategorijų skaičius, n – stebėjimų skaičius, f_j – j-osios kategorijos stebėjimų skaičius (j-osios kategorijos dažnis). Kokybinės įvairovės indeksas kinta nuo 0 (nėra reikšmių sklaidos) iki 1 (maksimali reikšmių sklaida).

1.15 pavyzdys. Tarkime, turime informaciją apie trijų rajonų gyventojų tautybę. Ji pateikta 1.10 lentelėje.

1.10 lentelė. Tautinė rajonų sudėtis

Tautybė	Rajonas A	Rajonas B	Rajonas C
Lietuviai	900	600	300
Lenkai	0	200	300
Rusai	0	100	300

A: $n = 900 + 0 + 0$
$k = 3$ $f_1^2 = 900^2$
$f_2^2 = 0$
$f_3^2 = 0$.

Pirmojo rajono

k n^2

$$IQV_1 = \frac{3(900^2 - (900^2 + 0^2 + 0^2))}{900^2 \cdot 2} = 0.$$

Antrojo rajono

$(600 + 200 + 100)^2$

$$IQV_2 = \frac{3(900^2 - (600^2 + 200^2 + 100^2))}{900^2 \cdot 2} = \frac{1\,200\,000}{1\,620\,000} = 0{,}74.$$

n^2

Trečiojo rajono

$$IQV_3 = \frac{3(900^2 - (300^2 + 300^2 + 300^2))}{900^2 \cdot 2} = 1.$$

Pirmas rajonas tautiniu požiūriu yra vienalytis. Antrojo rajono $IQV_2 = 0{,}74$, todėl jo tautinė gyventojų įvairovė gana didelė, o trečiojo – didžiausia, t. y. jame visų tautybių žmonių gyvena po lygiai.

4. Dažnių skirstinių formos charakteristikos – histogramos formos charakteristikos!

Šios charakteristikos skaičiuojamos tik tada, kai duomenis galima grupuoti, t. y. turint kiekybinius dažniausiai tolydžiojo kintamojo stebėjimo duomenis. Dažnių skirstinio formos (dažnių daugiakampio) formos charakteristikos. Jos charakteristikos – tai histogramos (dažnių daugiakampio) formos charakteristikos. Jos yra dvi – ekscesas ir asimetrijos koeficientas. Jiems apibrėžti reikia centrinio empirinio momento sąvokos.

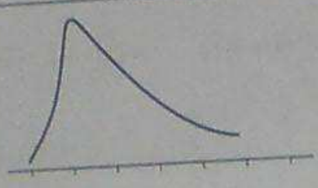
Centriniu empiriniu j-osios eilės momentu (žymimu m_j) vadinamas

$$m_j = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^j.$$

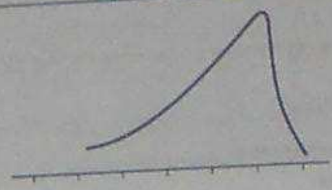
Formos charakteristikos yra centrinių momentų ir standartinio nuokrypio funkcijos.

Imties asimetrijos koeficientas $g_1 = \dfrac{m_3}{s^3}$.

Asimetrijos koeficientas yra histogramos simetrijos matas. Jeigu $g_1 > 0$, tai histogramos asimetrija teigiama (dešinioji), jeigu $g_1 < 0$, asimetrija neigiama (kairioji). Histograma simetriška, kai $g_1 = 0$. Beje, jeigu $g_1 > 0$ ($g_1 < 0$), tai $\overline{x} > Md$ ($\overline{x} < Md$).



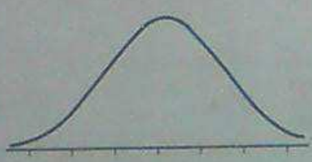
1.9 pav. Teigiama asimetrija



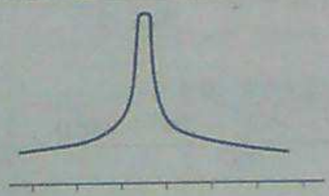
1.10 pav. Neigiama asimetrija

Eksceso koeficientas yra histogramos lėkštumo matas.

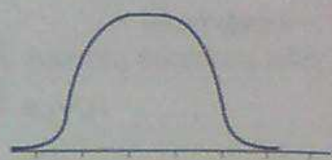
Imties eksceso koeficientas $g_2 = \dfrac{m_4}{s^4} - 3.$



1.11 pav. Simetriškas grafikas su nuliniu ekscesu



1.12 pav. Teigiamas ekscesas



1.13 pav. Neigiamas ekscesas

Jeigu $g_2 > 0$ – histograma lėkšta, t. y. duomenų sklaida apie vidurkį yra didesnė nei normaliosios kreivės atveju. Jeigu $g_2 < 0$ – histograma smaila, t. y. duomenų sklaida apie vidurkį yra mažesnė nei normaliosios kreivės. Jeigu $g_2 = 0$, tai sklaida apie vidurkį tokia pati kaip ir normaliosios kreivės.

Asimetrijos ir eksceso koeficientai yra panašumo į normaliąją kreivę matai. Kas gi ta normalioji kreivė, su kuria lyginama histograma?

Histograma dažnai yra varpo formos. Statistikas skirstinį priderina prie tam tikro šablono – duomenų matematinio modelio. Labiausiai paplitęs normalusis modelis. Skaičiavimams naudodami normalųjį skirstinį, daug naujo sužinome apie visą *populiaciją*. Išsamiau normalioji kreivė aptariama 5 skyrelyje.

5. Normalioji kreivė

Empiriškai nustatyta, kad daugelis histogramų yra panašios į funkcijos

$$\varphi_{\overline{x},s}(x) = \frac{1}{\sqrt{2\pi}\,s} \exp\left\{-\frac{(x-\overline{x})^2}{2s^2}\right\} \tag{1.16}$$

grafiką.

Funkcijos $\varphi_{\overline{x},s}(x)$ grafikas vadinamas ***normaliąja*** (arba Gauso) kreive. Teorinis ir praktinis jos vaidmuo statistikoje milžiniškas. Išsamiau su normaliąja kreive susipažinsime trečiojoje vadovėlio dalyje. Dabar tik paminėsime, kad didelė statistinių išvadų dalis grindžiama nagrinėjamų duomenų histogramos keitimu funkcija $\varphi_{\overline{x},s}(x)$.

Paminėsime keletą $\varphi_{\overline{x},s}(x)$ savybių:

1 $\varphi_{\overline{x},s}(x)$ grafikas yra varpo formos ir visas juo apribotas plotas lygus vienetui.

2 $\varphi_{\overline{x},s}(x)$ grafikas yra simetriškas $\overline{x}$ atžvilgiu.

3 $\varphi_{\overline{x},s}(x)$ yra apibrėžta su visais $-\infty < x < \infty$, bet toli nuo vidurkio funkcijos reikšmės labai mažos.

4 Intervale $(\overline{x} - ks, \overline{x} + ks)$ plotas, apribotas $\varphi_{\overline{x},s}(x)$ grafiku, priklauso nuo k, bet *nepriklauso* nuo $\overline{x}$ ir s.

Ketvirtoji savybė leidžia palyginti skirtingų duomenų normaliąsias kreives. Pavyzdžiui, intervale $(\overline{x} - s, \overline{x} + s)$ funkcijos $\varphi_{\overline{x},s}(x)$ grafiku apribotas plotas lygus 0,6826... (nepriklausomai nuo to, koks buvo vidurkis ir koks standartinis nuokrypis).

Į kairę ir dešinę nuo vidurkio atidėjus du standartinius nuokrypius, gautu intervalu ir normaliąja kreive apribotas plotas lygus 0,9544... Ketvirtosios savybės taikymas empiriniams duomenims vadinamas empirine taisykle.

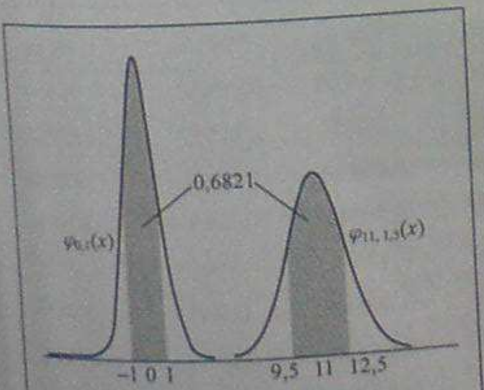
Empirinė taisyklė

Jeigu duomenų histograma yra varpo formos, tai:

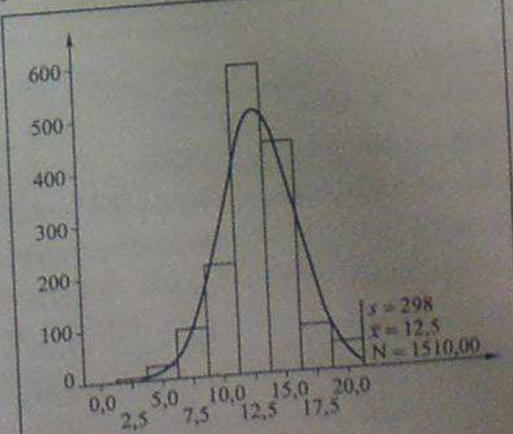
- apytiksliai 68% visų duomenų patenka į intervalą $(\overline{x} - s, \overline{x} + s)$;
- apytiksliai 95% visų duomenų patenka į intervalą $(\overline{x} - 2s, \overline{x} + 2s)$;
- beveik visi duomenys patenka į intervalą $(\overline{x} - 3s, \overline{x} + 3s)$.

1.16 pavyzdys. Tam tikrų paslaugų firma turi daugelio ketvirčių informaciją apie nepatenkintų klientų skundus. Ketvirčių skundų vidurkis ir standartinis nuokrypis iš esmės stabilizavosi. Histograma mažai skiriasi nuo normaliosios kreivės. Jeigu kurį nors ketvirtį skundų skaičius vidurkį viršija daugiau nei dviem standartiniais nuokrypiais (empirinė taisyklė teigia, kad tai labai mažai tikėtina), firma imasi tyrimo, ar nepablogėjo paslaugų kokybė.

Kaip jau minėjome, išvadų statistikoje labai svarbu nustatyti, ar galima duomenų histogramą keisti funkcija $\varphi_{\overline{x},s}(x)$. Tikslesni šios problemos sprendimo metodai aptariami statistinių išvadų dalyje, dabar tik pabrėšime, kad statistiniais paketais viename grafike galima nubraižyti ir histogramą, ir normaliąją kreivę (žr. 1.15 pav.).



1.14 pav. Dvi normaliosios kreivės ir empirinė taisyklė



1.15 pav. Histograma ir normalioji kreivė

6. Standartizuotosios reikšmės ir išskirtys

Svarbu ne tik konkreti stebėjimo reikšmė, bet ir jos padėtis duomenų aibėje. Įprastinė procedūra – lyginti stebėjimą su vidurkiu. Pavyzdžiui, studentų grupės statistikos egzamino pažymių vidurkis 6,5 balo, o Remigijus gavo 6 balus. Kad Remigijus „nesublizgėjo" aišku, tačiau vienaip jį vertinsime žinodami, kad ir kiti gavo nuo 6 iki 7 balų, ir kitaip, žinodami, kad buvo daug 9 ir 10 ir nemažai 5 bei 6. Taigi vertinant svarbi ir duomenų sklaida. Dar sudėtingiau lyginti kelias duomenų aibes. Tarkime, pirmos grupės studentai per egzaminą sprendė 100 uždavinių, o antrosios – 80 uždavinių. Kaip nustatyti, kas iš grupės pasirodė geriau: Vytas, išsprendęs 90 iš 100, ar Rimas, išsprendęs 75 iš 80. Viena vertus, Rimas išsprendė daugiau uždavinių. Kita vertus, gali būti taip, kad Vytas *vienintelis* grupėje išsprendė daugiau nei 20 uždavinių (ir daug daugiau – tikras lyderis), o Rimas *vienintelis* nesugebėjo išspręsti visų 80 uždavinių (tikras atsilikėlis). Taigi reikia metodikos rezultatų grupėse svarbai palyginti. Vienas iš būdų tą padaryti – rezultatus standartizuoti.

6.1. Standartizuotosios z reikšmės

Labiausiai paplitęs standartizavimas – z reikšmių skaičiavimas. Tarkime, turime duomenų aibę $x_1, x_2, \ldots, x_n$. Tuomet z reikšmė skaičiuojama pagal formulę

$$z_i = \frac{x_i - \overline{x}}{s}.$$

Standartizavę duomenis, gauname naują duomenų aibę $z_1, z_2, \ldots, z_n$, kurios vidurkis *visada* lygus 0, o standartinis nuokrypis *visada* lygus 1: $\overline{z} = 0$, $s_z = 1$.

Teigiama standartizuotoji reikšmė parodo geresnį nei vidurkis rezultatą, neigiama – blogesnį. Standartizuotosios reikšmės gaunamos tiesiškai transformuojant duomenis.

Populiacijos z reikšmės apibrėžiamos empirinį vidurkį keičiant μ ir standartinį nuokrypį – σ, t. y. $z_i = (x_i - \mu)/\sigma$.

1.11 lentelė

Produktas	Suvartotų produktų kiekis (kg)			z reikšmės		
	1994 m.	1995 m.	1996 m.	1994 m.	1995 m.	1996 m.
Mėsa ir jos produktai	50,00	52,00	50,00	−0,34760	−0,35683	−0,36596
Pienas ir jo produktai	291,00	238,00	247,00	2,36251	2,11557	2,15578
Duona ir grūdų produktai	135,00	136,00	135,00	0,60824	0,75974	0,72210
Bulvės	99,00	127,00	127,00	0,20341	0,64011	0,61970
Daržovės, arbūzai	65,00	65,00	65,00	−0,17892	−0,189403	−0,17395
Vaisiai ir uogos	45,00	48,00	40,00	−0,40383	−0,41000	−0,49379
Cukrus	22,70	22,20	22,40	−0,65460	−0,75294	−0,71926
Aliejus ir margarinas	10,40	11,50	10,90	−0,79292	−0,89517	−0,86647
Žuvis ir jos produktai	10,10	9,90	10,00	−0,79629	−0,91644	−0,87799

Net ir skirtingų duomenų aibių z reikšmes galima lyginti tarpusavyje. Tarkime, Vyto z reikšmė lygi 1,2, o Rimo – 1,1. Taigi *savo* grupėje Vytas pasirodė geriau nei Rimas savojoje.

1.17 pavyzdys. Lietuvos 1994–1996 metų vidutiniškai vieno gyventojo suvartotų produktų kiekis (kg) ir z reikšmės surašytos 1.11 lentelėje. Matome, kad nors visais metais daržovių buvo suvartota tiek pat (65 kg), tačiau santykinis daržovių žmonių racione kiekis didžiausias buvo 1996 metais. (Duomenys paimti iš metinio statistikos biuletenio „Ūkininkų ūkių veikla 1996 metais".)

Be z reikšmių, naudojamos ir kitos tiesinės duomenų transformacijos. Pavyzdžiui, T reikšmė: $T_i = 10z_i + 50$.

T reikšmių vidurkis lygus 50, o standartinis nuokrypis – 10. Žinoma, šie skaičiai nėra kuo nors ypatingi, naudojamos ir kitos transformacijos, pavyzdžiui, $100z_i + 500$ (vidurkis 500, standartinis nuokrypis 100) ir pan. Visos gautos naudojant panašias transformacijas reikšmės turi *tiek pat* informacijos kiek ir z reikšmės. Kam tuomet jų reikia? Tam, kad vartotojui būtų lengviau jas suvokti. Žmogui patogiau operuoti sveikais neneigiamais skaičiais. Palyginkite dvi, tiek pat informacijos turinčias, frazes:

1. Jūsų sūnaus fizikos žinių testo z reikšmė lygi $-2{,}2$. Klasės vidurkis 0.
2. Jūsų sūnaus fizikos žinių testo z reikšmė 22. Klasės vidurkis 50.

Yra ir netiesinių transformacijų. Pavyzdžiui, edukologijoje paplitusios Rašo kreivės. Jų šiame vadovėlyje neaptarsime.

6.2. Išskirtys

Dažnai, ypač kai duomenų histograma yra varpo formos (artima tų duomenų normaliajai kreivei, žr. 5), duomenų aibės išskirtys nustatomos naudojant z reikšmes. Tuomet remiantis *empirine taisykle*, galima teigti, kad apytiksliai 68% visų z reikšmių patenka į intervalą $(-1, 1)$ ir apytiksliai 95% visų reikšmių patenka į intervalą $(-2, 2)$, o beveik visos z reikšmės – į intervalą $(-3, 3)$. Taigi duomuo, kurio z reikšmė absoliučiuoju didumu didesnė už 2 ar 3, tam tikra prasme jau yra išskirtinis. Toks požiūris gali būti taikomas praktiškai. Pavyzdžiui, mokiniai, kurių žinių kompleksinio testo z reikšmė mažesnė už -2, reikalauja papildomo dėmesio.

Išvados apie z reikšmes tinka tik pakankamai „tirštiems" duomenims (jų turi būti ne mažiau kaip 10), kurių histogramos panašios į normaliąją kreivę.

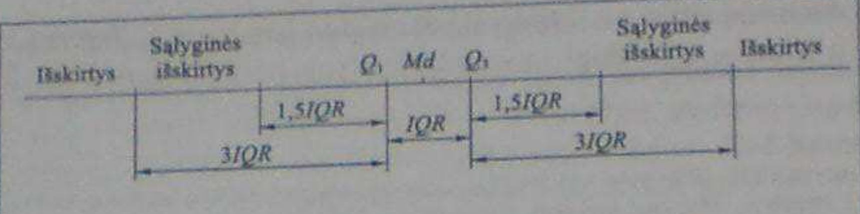
Sąlygine išskirtimi vadinamas *duomuo*, kurio z reikšmė absoliučiuoju didumu didesnė už 2, bet mažesnė už 3.

Išskirtimi vadinamas *duomuo*, kurio z reikšmė absoliučiuoju didumu didesnė už 3.

Ne visuomet duomenų histograma yra artima normaliajai kreivei (pavyzdžiui, dažnių skirstinys asimetriškas). Todėl yra dar vienas būdas išskirtims apibrėžti (žr. 1.16 pav.). Jis universalesnis ta prasme, kad nesusijęs su empirine taisykle.

Sąlygine išskirtimi vadinamas *duomuo*, priklausantis intervalui $[Q_1 - 3IQR,\ Q_1 - 1{,}5IQR)$ arba $(Q_3 + 1{,}5IQR,\ Q_3 + 3IQR]$.

Išskirtimi vadinamas *duomuo*, mažesnis už $Q_1 - 3IQR$ arba didesnis už $Q_3 + 3IQR$.



1.16 pav. Išskirtys

1.18 pavyzdys. Tarkime, žinome penkiolikos studentų ūgį (cm): 154; 160; 172; 175; 176; 179; 180; 180; 190; 198; 215; 165; 170; 171; 172. Naudodamiesi 2.4 skyrelio nurodymais, randame:

$$Q_1 = x_{(4)} = 170, \qquad Q_3 = x_{(12)} = 180, \qquad IQR = 180 - 170 = 10,$$
$$1{,}5IQR = 15, \qquad 3IQR = 30, \qquad Q_1 - 1{,}5IQR = 155,$$
$$Q_1 - 3IQR = 140, \qquad Q_3 + 1{,}5IQR = 195, \qquad Q_3 + 3IQR = 210.$$

Taigi sąlyginės išskirtys yra ūgiai, pakliuvę į intervalą [140, 155) arba į (195, 210]. Tokių duomenų yra du: 154 ir 198. Išskirtys turėtų būti ūgiai, mažesni už 140 arba didesni už 210. Iš tikrųjų yra tik viena išskirtis – 215.

7. Čebyšovo taisyklė

Čebyšovo[1] taisyklė sieja duomenis su jų vidurkiu ir standartiniu nuokrypiu.

> Ne mažiau kaip $1 - 1/k^2$ visų stebėjimų patenka į intervalą $(\overline{x} - ks, \overline{x} + ks)$.

Atkreipiame dėmesį, kad $k > 0$ nebūtinai sveikasis skaičius.

Svarbiausias Čebyšovo taisyklės privalumas – jos universalumas. Čebyšovo taisyklė nepriklauso nuo dažnių pasiskirstymo. Tereikalaujama, kad stebėjimai būtų kiekybiniai. Čebyšovo taisyklė parodo $\overline{x}$ ir s svarbą aprašomojoje statistikoje. Ją galima suformuluoti ir visai populiacijai, tam tereikia $\overline{x}$ pakeisti μ, o s – σ.

Čebyšovo taisyklė yra Čebyšovo teoremos (žr. III.20) išvada. Tikslesnė, bet nepatogi yra tokia taisyklės formuluotė: ne mažiau kaip $1 - 1/k^2$ visų stebėjimų pakliūva į intervalą $(\overline{x} - ks\sqrt{(n-1)/n}, \overline{x} + ks\sqrt{(n-1)/n})$.

Suformuluosime dvi svarbias Čebyšovo taisyklės išvadas.

> Ne mažiau kaip 75% visų stebėjimų pakliūva į intervalą $(\overline{x} - 2s, \overline{x} + 2s)$.
> Ne mažiau kaip 88% visų stebėjimų pakliūva į intervalą $(\overline{x} - 3s, \overline{x} + 3s)$.

Palyginę Čebyšovo taisyklę su empirine taisykle (žr. 5), matome, kad ji daug nuosaikesnė. Taigi kai duomenų histograma panaši į normaliąją kreivę, tikslingiau taikyti empirinę taisyklę. Tačiau Čebyšovo taisyklė universalesnė ir galioja net tada, kai empirinė taisyklė netinka.

[1] Pafnutij Čebyšev (1821–1894) – rusų matematikas.

Čebyševo taisyklė

1.19 pavyzdys. Pateiksime vieną Čebyšovo taisyklės taikymo pavyzdį. Tarkime, kad buvo įvertintas 200 jaunų žmonių intelekto koeficientas (IQ). Visų rezultatų vidurkis $\overline{x} = 110$ balų, o standartinis nuokrypis $s = 5$ balai. Kiek žmonių gavo IQ įvertinimus, didesnius už 95 balus, bet mažesnius už 125 balus? Pastebime, kad $95 = 110 - 15 = \overline{x} - 3s$ ir $125 = \overline{x} + 3s$. Todėl iš Čebyšovo taisyklės išplaukia, kad ne mažiau 88% IQ rezultatų pateko į norimą intervalą. Taigi minėto dydžio IQ įvertinimus gavo ne mažiau kaip $200 \cdot 88/100 = 176$ žmonės.

8. Poriniai stebėjimai

Ankstesniuose skyreliuose aptarėme aprašomosios statistikos metodus, taikomus tuomet, kai duomenų aibę sudaro vieno kintamojo matavimų reikšmės. Tačiau praktiškai dažnai susiduriame su kelių kintamųjų matavimais. Pavyzdžiui, sociologiniams tyrimams naudojami klausimynai, apimantys dešimtis klausimų apie amžių, išsilavinimą, pajamas, politikų vertinimą ir pan. Medicinoje dažna duomenų aibė – ligų istorijos, kuriose yra informacijos apie ligonio kraują, šlapimą, persirgtas ligas. Ekonomikoje duomenų aibė gali būti ūkinių bendrovių veiklos rodiklių visuma. Žinoma, iš pradžių kiekvieno kintamojo reikšmes reikia susisteminti. Vieno kintamojo aprašomoji statistika leidžia pastebėti atskiro kintamojo savybes. Tačiau ji bejėgė, kai statistiką domina, ar tarp kintamųjų yra priklausomybė; ar, žinant vieno kintamojo reikšmes, galima įvertinti (prognozuoti) kito kintamojo reikšmes. Pavyzdžiui, norima nustatyti, ar yra ryšys tarp darbo stažo ir atlyginimo, tarp vaistų vartojimo trukmės ir pooperacinių komplikacijų, tarp kritulių kiekio ir vidutinės liepos mėnesio temperatūros, ar galima universiteto studentų požiūrį į „muilo operas“ prognozuoti priklausomai nuo jų IQ ir pan.

Išsamius atsakymus į šiuos klausimus galima gauti taikant statistinių išvadų koreliacinės ir regresinės analizės metodus. Su jais susipažinsime vėliau. Šiame skyrelyje aptarsime porinių stebėjimų aprašomosios statistikos metodus, padedančius pastebėti kintamųjų *porų* savybes.

Kategorinių ar ranginių kintamųjų porinius stebėjimus galime surašyti į porinę dažnių (sąveikos) 1.12 lentelę. Joje f_{ij} – reikšmių poros (x_i, y_j) dažnis, t. y. skaičius, nusakantis, kiek kartų (x_i, y_j) pasikartojo duomenų aibėje.

Paprastai sudarant porinę dažnių lentelę statistinių programų paketu, langeliuose esti daugiau informacijos. Prisiminkime 1.1 pavyzdį. Tarkime, kad žinome ne tik patį požiūrį į „muilo operas“, bet ir atsakinėjusių studentų lytį.

1.12 lentelė. Porinė dažnių lentelė

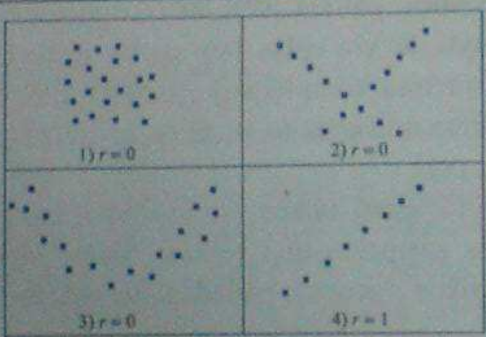
	y_1	y_2	$\cdots$	y_r
x_1	f_{11}	f_{12}	$\cdots$	f_{1r}
x_2	f_{21}	f_{22}	$\cdots$	f_{2r}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_s	f_{s1}	f_{s2}	$\cdots$	f_{sr}

Panagrinėkime SPSS paketu gautą porinę dažnių lentelę (1.17 pav.).

LYTIS * MUILAS Crosstabulation

			MUILAS					
			labai patinka	patinka	neturiu nuomonės	nepatinka	labai nepatinka	Total
LYTIS	vyras	Count	1	2	6	7	8	24
		% within LYTIS	4.2%	8.3%	25.0%	29.2%	33.3%	100.0%
		% within MUILAS	12.5%	18.2%	50.0%	77.8%	80.0%	48.0%
	moteris	Count	7	9	6	2	2	26
		% within LYTIS	26.9%	34.6%	23.1%	7.7%	7.7%	100.0%
		% within MUILAS	87.5%	81.8%	50.0%	22.2%	20.0%	52.0%
Total		Count	8	11	12	9	10	50
		% within LYTIS	16.0%	22.0%	24.0%	18.0%	20.0%	100.0%
		% within MUILAS	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

1.17 pav. Požiūrio į „muilo operas" dažnių lentelė



1.18 pav. Sklaidos diagramos

Be porinių dažnių (*count*), lentelėje dar pateikiami eilučių (*row*), stulpelių (*col.*) ir visumos (*table*) procentai. Matome, kad 87,5% apklaustųjų, kuriems labai patinka „muilo operos", sudaro studentės; 23,1% apklaustųjų studenčių apie jas neturi nuomonės. Matome, kad studentės „muilo operas" vertina palankiau. Tačiau, ar galima sakyti, kad yra priklausomybė tarp studentų lyties ir požiūrio į „muilo operas"? Ar ši priklausomybė statistiškai reikšminga? Į šiuos klausimus atsako statistinių išvadų metodai, aprašyti III vadovėlio dalyje. Čia tik paminėsime, kad spausdinant porinių dažnių lentelę dažnai kartu spausdinamas empirinis ryšio stiprumo koeficientas, kurio reikšmė yra tolesnių statistiko veiksmų indikatorius. Tokių koeficientų yra net keletas. Nagrinėto pavyzdžio Pirsono[1] *sąveikos* koeficientas $C = 0{,}475$, o Kramero[2] koreliacijos koeficientas $V = 0{,}54$. Kuo koeficientai didesni (kuo arčiau 1), tuo priklausomybė stipresnė. Nagrinėto pavyzdžio atveju požiūris į „muilo operas" ir lytis yra priklausomi. Išsamiau priklausomybės matas aptartas III.1.6 skyrelyje.

[1] Karl Pearson (1857–1936) – anglų statistikas.

[2] Harold Cramér (1893–1985) – švedų statistikas.

Jei stebimi tolydieji kintamieji, tai prieš sudarant porinių dažnių lentelę duomenys grupuojami. Taip prarandama dalis informacijos. Todėl dažnai vietoje porinių dažnių lentelės braižoma negrupuotų duomenų sklaidos diagrama. Joje reikšmių (x_j, y_j) pora žymima kvadratėliu. Sklaidos diagramų pavyzdžiai, iliustruojantys galimas situacijas, pateikti 1.18 paveiksle.

Iš paveikslo matyti, kad 1) ir 2) atvejais tarp kintamųjų X ir Y priklausomybės nėra; 3) yra netiesinė priklausomybė; 4) – tiesinė priklausomybė. Aišku, kad šios išvados preliminarios, grafikai tik „sufleruoja" galimas išvadas. Kas yra tiesinė (netiesinė) priklausomybė? Kas yra priklausomybės matas? Tiesinė priklausomybė parodo, kad kintamuosius sieja tiesinis funkcinis ryšys: $y = a + bx$. Tiesinės priklausomybės pavyzdys yra ryšys tarp Farenheito ir Celsijaus temperatūrų skalių. Aišku, kad duomenų aibėse tokia situacija praktiškai nepasitaiko. Galima kalbėti tik apie funkcijas, kurios gana tiksliai (tam tikro kriterijaus prasme) aprašo x ir y sąveiką. Šios funkcijos vadinamos regresijos lygtimis, o jų grafikai – regresijos kreivėmis. Regresijos lygčiai sudaryti naudojamas *mažiausiųjų kvadratų* metodas.

1.20 pavyzdys. Tarkime, turime tokius porinius x ir y stebėjimus:

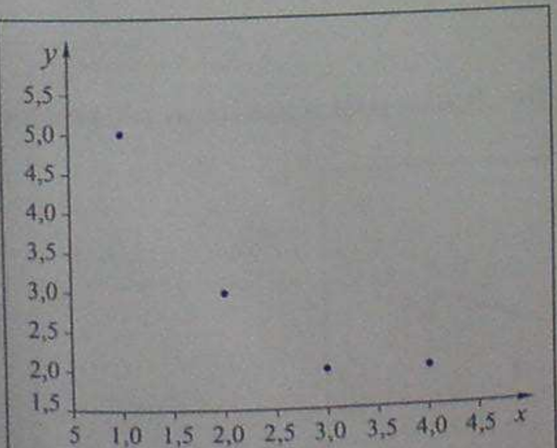
x	3	2	4	1
y	2	3	2	5

Sklaidos diagrama pateikta 1.19 paveiksle.
Lygtis $\widehat{y} = a + bx = -2 + 2x$ geriausiai iš visų tiesinių lygčių aprašo x ir y sąveiką, t. y. dydis

$$\sum e_i^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2 = (y_1 - \widehat{y}_1)^2 + \cdots + (y_4 + \widehat{y}_4)^2$$

yra minimalus.

Taigi $\widehat{y} = -2 + 2x$ yra regresijos lygtis (jos grafikas – regresijos kreivė), $a = -2$ yra laisvasis narys, $b = 2$ – krypties koeficientas.



1.19 pav.

Vienas iš tiesinės sąveikos stiprumo įverčių yra empirinis koreliacijos (vadinamasis Pirsono koreliacijos) koeficientas:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{s_x s_y}; \qquad (1.17)$$

čia s_x ir s_y atitinkamai x ir y stebėjimų standartiniai nuokrypiai.

Pirsono koreliacijos koeficientas turi tokias savybes:

1 Visus x_i (y_i) padauginus iš konstantos, koeficiento r reikšmė nepasikeičia.

2 $r = 1$, kai visi taškai (x_i, y_i) yra tiesėje, kurios krypties koeficientas teigiamas.

3 $r = -1$, kai visi taškai (x_i, y_i) yra tiesėje, kurios krypties koeficientas neigiamas.

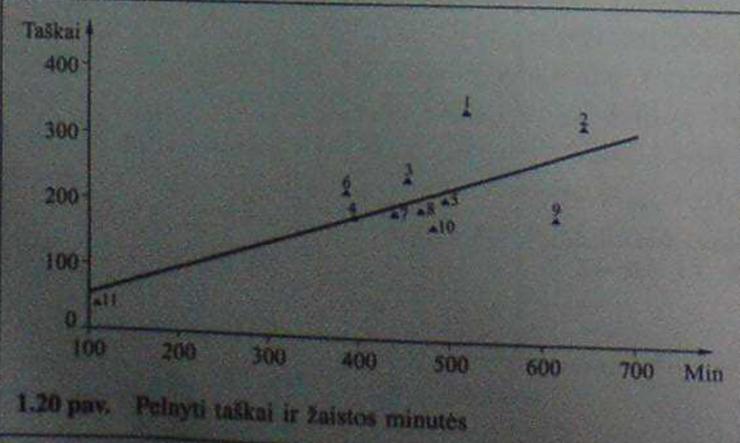
4 $r = 0$, jei kintamieji yra tiesiškai nepriklausomi (žr. II.16).

1.21 pavyzdys. Panagrinėkime Kauno „Žalgirio" krepšinio komandos žaistų minučių ir pelnytų taškų LKL 98–99 metais priklausomybę. Duomenys pateikti 1.13 lentelėje.

1.13 lentelė. Taškai ir žaistos minutės

Vardas, pavardė	Žaista minučių	Pelnyta taškų
G. Zidekas	517	338
A. Bowie	644	319
S. Štombergas	453	237
T. Edney	392	179
D. Adomaitis	494	209
K. Šeštokas	386	218
E. Žukauskas	439	186
M. Žukauskas	468	193
D. Maskoliūnas	614	185
T. Masiulis	481	168
G. Gustas	109	39

Empirinis Pirsono koreliacijos koeficientas $r = 0{,}761$. Sklaidos grafikas pateikiamas 1.20 paveiksle.



1.20 pav. Pelnyti taškai ir žaistos minutės

Regresijos lygtis $y = 11{,}4966 + 0{,}4292x$, čia x – žaistų minučių skaičius, o y – pelnytų taškų skaičius. Taigi aprašomosios statistikos prasme galima teigti, kad žalgiriečių pelnytų taškų skaičiui turėjo įtakos žaistos minutės, o regresijos lygtis aprašo šios įtakos tiesinę tendenciją. Reikia pabrėžti, kad negalime teigti, jog gauta lygtis tinka pelnytų taškų prognozei arba jog pelnyti taškai tiesiškai susiję su žaistų minučių skaičiumi. Tai *išvadų* teorijos prerogatyva. Būtent išvadų teorijoje sprendžiama apie modelių tinkamumą, patikimumą ir pan.

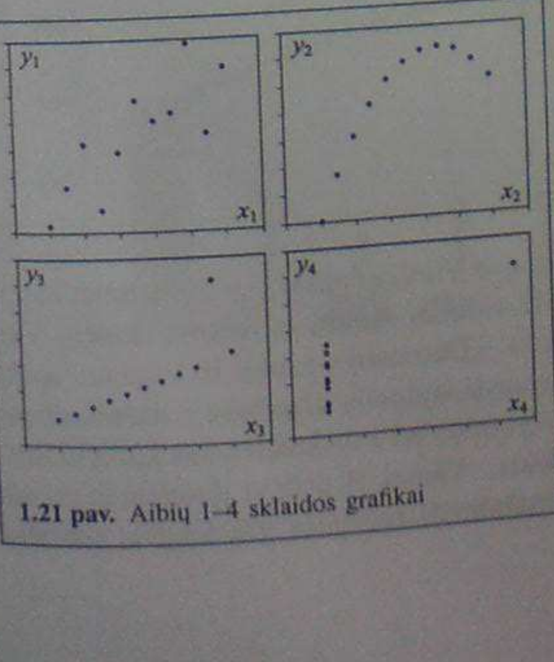
9. Grafinis stebėjimų vaizdavimas

Grafikas yra vaizdinė priemonė glaustai tiek pradiniams duomenims, tiek ir analizės rezultatams pateikti. Grafiniai elementai lengvai dekoduojami (suvokiami), todėl grafikas suteikia daugiau informacijos nei „pliki" skaičiai. Tai taikoma žiniasklaidoje, nes vaizdinė informacija geriau suvokiama. Dažnai grafikai pateikiami kartu su dažnių lentelėmis, kurias galima laikyti skaičių suvestinėmis arba duomenų „koncentratais".

Panagrinėkime keturias duomenų aibes, pateiktas 1.14 lentelėje. Tiesinės regresinės analizės rezultatai yra identiški šioms duomenų aibėms: tie patys vidurkiai, dispersijos,

1.14 lentelė

Aibė 1		Aibė 2		Aibė 3		Aibė 4	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	80,4	10	91,4	10	74,6	8	65,8
8	69,5	8	81,4	8	67,7	8	57,6
13	75,8	13	87,4	13	127,4	8	77,1
9	88,1	9	87,7	9	71,1	8	88,4
11	83,3	11	92,6	11	78,1	8	84,7
14	99,6	14	81,0	14	88,4	8	70,4
6	72,4	6	61,3	6	60,8	8	52,5
4	42,6	4	31,0	4	53,9	19	125,0
12	108,4	12	91,3	12	81,5	8	55,6
7	48,2	7	72,6	7	64,2	8	79,1
5	56,8	5	47,4	5	57,3	8	68,9



1.21 pav. Aibių 1–4 sklaidos grafikai krypties koeficientai, laisvieji nariai. Ar pastebite esminius skirtumus tarp šių duomenų aibių? Ar jos tinkamos tiesinei regresijai?

Pažvelkime į duomenų aibių sklaidos grafikus (žr. 1.21 pav.). Čia skirtumai akivaizdūs. Pirmoji duomenų aibė – būdingi regresijos duomenys, t. y. įžiūrima kintamųjų x_1 ir y_1 priklausomybė. Antrojoje aibėje y yra x funkcija, bet ji netiesinė, nėra ir paklaidos. Trečiosios aibės kintamuosius x ir y sieja tiesinė priklausomybė, išskyrus vieną matavimų porą. Ketvirtojoje aibėje yra tik dvi skirtingos x reikšmės. Ar pastebėjote šiuos skirtumus prieš pažvelgdami į grafikus?

Taigi dažnai grafikas atskleidžia naujas ir sunkiai kitais metodais įvertinamas duomenų savybes. Šio skyriaus tikslas – ne pateikti „receptus", kaip braižyti grafikus, o nurodyti pagrindinius požymius, charakterizuojančius gerą grafiką, pasiaiškinti, kaip pasinaudoti įvairia grafine technika, perteikiant specifinę informaciją.

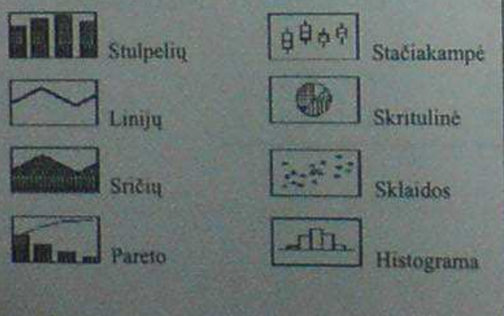
Visi grafikai – tai užkoduota informacija apie duomenis. Grafiko kokybė daugiausia priklauso nuo mūsų gebėjimų vizualiai tą informaciją dekoduoti. Jei negalime to padaryti, tai grafikas yra bevertis. Tipiniai reikalavimai spausdinamiems grafikams:

Aiškumas – grafikai turi būti suvokiami be papildomų aprašymų.

Skiriamoji galia – kiekvienas grafiko elementas turi būti lengvai įžiūrimas.

Kopijuojamumas – grafiko kopija (pvz., nespalvota) turi likti informatyvi.

Nėra paprasta „išgauti" visą informaciją, esančią duomenyse. Negalima tikėtis pateikti jos visos vienu grafiku (tuo labiau pirmuoju). Grafinis duomenų tyrimas yra interaktyvus procesas. Tenka braižyti daug grafikų, iš kurių kiekvienas padeda atskleisti (aprėpti) vis daugiau informacijos apie duomenis. Tačiau negalima grafiko perkrauti, reikia atsiminti žinomą posakį: „per medžius turi būti matomas miškas". Grafikų įvairovė yra labai didelė, todėl nėra galimybės visų jų aprašyti. Būdingiausi grafikų tipai pateikiami 1.22 paveiksle. Toliau aprašysime dažniausiai naudojamus grafikus.

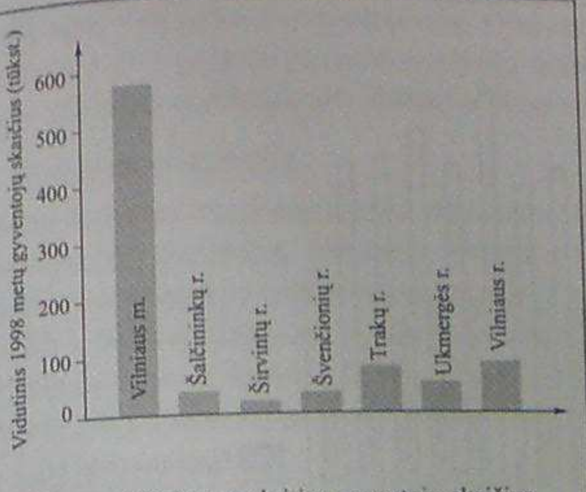


1.22 pav. Diagramos

9.1. Stulpelių diagrama

Pradėkime nuo pavyzdžio. Turime duomenis apie Vilniaus apskrities 1998 metų kiekvieno mėnesio gyventojų skaičių. Informacija apie vidutinį metinį gyventojų skaičių Vilniaus apskrityje 1998 metais pateikta 1.23 paveiksle. (Duomenys paimti iš Vilniaus apskrities statistikos valdybos tinklalapio.) Tai paprasčiausia stulpelių diagrama – dažnių diagrama.

Dažnių diagramos tinka, kai stebime kokybinius arba diskrečiuosius kiekybinius kintamuosius. Čia dažnį atitinka stulpelio aukštis. Vienas iš dažnių diagramos privalumų yra tai, kad išskirtys, moda, minimali ir maksimali reikšmės yra lengvai pastebimos.



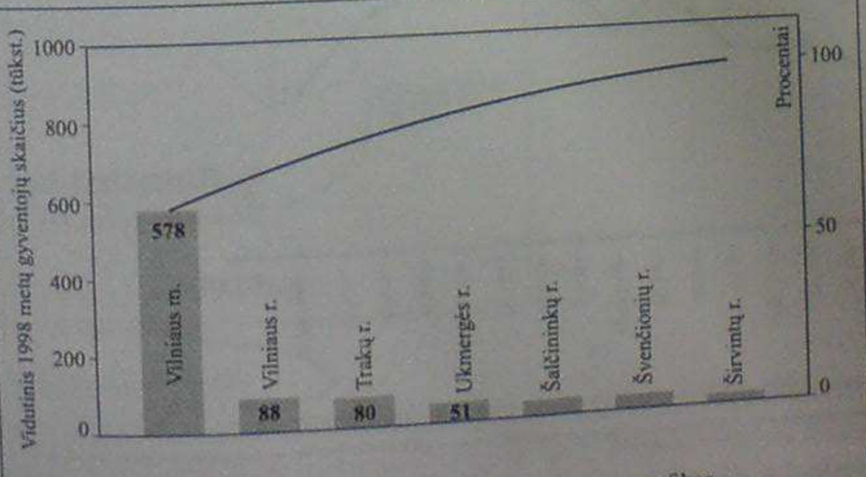
1.23 pav. Vilniaus apskrities gyventojų skaičius

Braižomos horizontaliosios normaliųjų kintamųjų dažnių diagramos ir vertikaliosios kitų duomenų diagramos.

Braižant stulpelių diagramą, paprastai laikomasi tokių taisyklių:

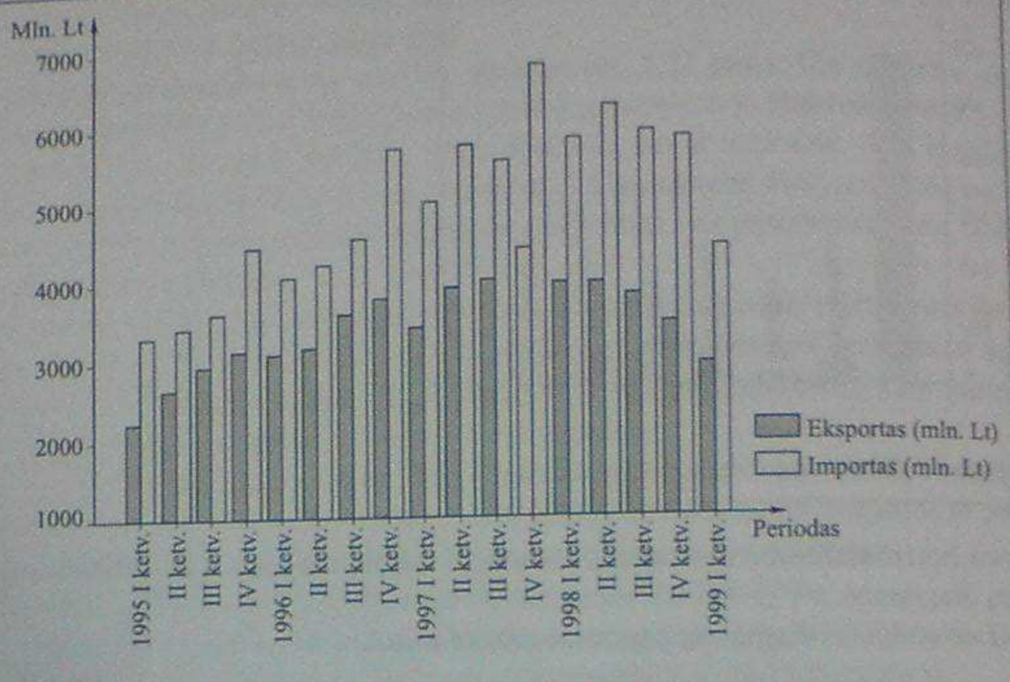
1) visi stulpeliai turi būti to paties pločio,
2) tarpai tarp stulpelių turi būti ne mažesni kaip pusė ir ne didesni kaip visas stulpelio plotis.

Yra daug efektyvių būdų, kaip naudojant stulpelių diagramas „išryškinti" duomenų aibės savybes. Viena iš stulpelių diagramos atmainų – Pareto[1] grafikas. Koordinačių (Ox) ašyje atidedamos kategorijos (pradedant nuo dažniausiai pasikartojančios). Be to, brėžiama sukauptųjų procentų kreivė. Palyginkite 1.23 ir 1.24 paveikslus. Juose pavaizduoti tie patys duomenys, tačiau antrasis perteikia daugiau informacijos. Iškart matome, kad Vilniuje gyvena daugiau kaip 50% apskrities gyventojų, o Švenčionių rajone gyvena mažiausiai.

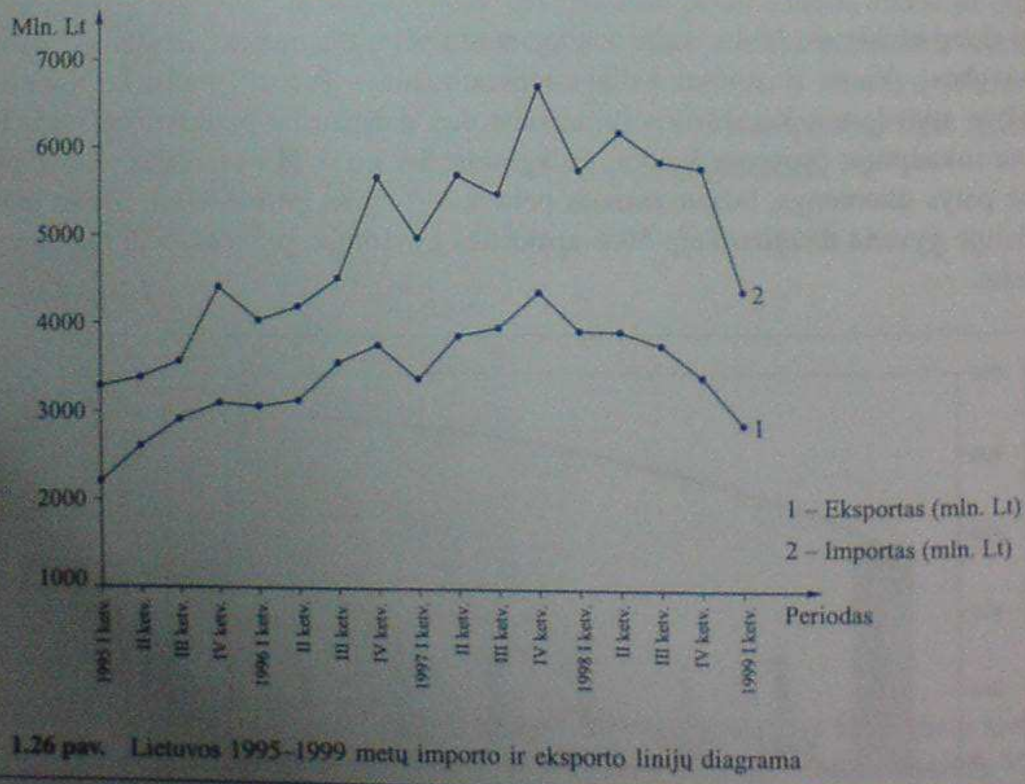


1.24 pav. Vilniaus apskrities gyventojų skaičiaus Pareto grafikas

[1] Vilfredo Pareto (1848–1923) – italų ekonomistas ir sociologas.



1.25 pav. Lietuvos 1995–1999 metų importo ir eksporto stulpelių diagrama



1.26 pav. Lietuvos 1995–1999 metų importo ir eksporto linijų diagrama

Kalbant apie stulpelių diagramas, reiktų paminėti grupuotą stulpelių diagramą. Ji naudojama norint palyginti kelių duomenų aibių reikšmes ar charakteristikas. Stulpelių diagrama, vaizduojanti Lietuvos 1995–1999 metų eksporto ir importo apimtis, pateikta 1.25 paveiksle. (Duomenys paimti iš Statistikos departamento tinklalapio.) Turime 17 porų reikšmių (importas ir eksportas nuo 1995 metų). Beje, šiems duomenims vaizduoti labiau tinka 1.26 paveiksle pavaizduota linijų diagrama, nes geriau perteikia kitimą laike. Be to, ji yra kompaktiškesnė nei stulpelių diagrama, kai reikia pavaizduoti daug taškų.

9.2. *Skritulinė diagrama*

Skritulinė diagrama naudojama kaip alternatyva stulpelių diagramai. Ji perteikia tą pačią informaciją, tik kita forma. Skritulys atitinka visą populiaciją (100%), o išpjovos – kategorijas proporcingai jų santykiniam dažniui. Pavyzdžiui, tarkime, santykinis dažnis yra 0,25 (25%). Kadangi išpjovos plotas yra proporcingas išpjovos centriniam kampui, tai šį dažnį atitinka išpjova, kurios kampas yra $25 \cdot 360/100 = 90°$.

Norint pavaizduoti išsamią skritulinę diagramą, reikia žinoti šias taisykles:

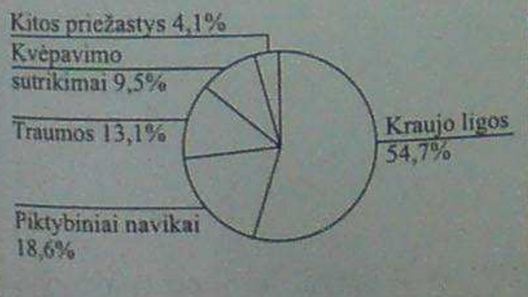
1. Diagramos pavadinimas ir populiacijos didumas N yra nurodomi po diagrama.

2. Diagramoje išpjovos išdėstomos mažėjimo tvarka pagal laikrodžio rodyklę pradedant 12-ąja pozicija. Šios taisyklės nepaisoma, kai tame pačiame grafike yra dvi skritulinės diagramos (siekiame jas palyginti).

3. Skritulinė diagrama yra per daug marga ir nevaizdi, jei kategorijų skaičius didesnis už 5 arba mažiausia išpjova yra mažesnė nei 3% (10,8°) viso skritulio.

4. Nerekomenduojama naudoti trimatės skritulinės diagramos, nes toliau esančios išpjovos atrodo mažesnės.

Skritulinė diagrama naudojama tik visumos dalims vaizduoti. Ji labai populiari biudžeto sandarai vaizduoti. Skritulinė diagrama, kurioje yra Lietuvos Respublikos sveikatos apsaugos ministerijos informacija apie 1998 metais mirusiųjų mirties priežastis, pavaizduota 1.27 paveiksle.



1.27 pav. Mirties priežastys

9.3. *Diagrama medis*

Sugrupuota dažnių lentelė turi trūkumą – grupavimo metu pradinė informacija prarandama. Diagrama medis leidžia šio trūkumo išvengti. Kaip ši diagrama yra sudaroma? Jei skaičius turi du ar daugiau skaitmenų, tada jį galima išskaidyti į *šaką* ir *lapą*. *Šaka* yra pirmasis skaitmuo (pirmieji skaitmenys), *lapas* – paskutinis skaitmuo (paskutiniai skaitmenys). Pavyzdžiui, skaičių 367 galima išskaidyti dviem būdais: 1) 3|67, čia 3 – *šaka*, 67 – *lapas*; 2) 36|7, čia 36 – *šaka*, 7 – *lapas*. Išskaidytus tokiu būdu duomenis nesunku pateikti grafiškai. Sudarykime diagramą medį duomenų aibės, kurioje yra 25 studentų svoriai (kg), pateikti 1.15 lentelėje.

1.15 lentelė. Studentų svoriai

78	67	65	87	75
65	71	54	94	64
84	82	81	68	85
76	89	98	59	57
79	65	59	80	67

1 žingsnis. Šakos skaitmenis išdėstome vertikaliai. Brūkšniu atskiriame *šaką* nuo *lapų*:

5|
6|
7|
8|
9|

2 žingsnis. Kiekvieną *lapą* atidedame į dešinę nuo savojo skaitmens *šakoje*. Kadangi pirmasis skaičius yra 78, skaičiaus 7 dešinėje parašome 8:

5|
6|
7| 8
8|
9|

Tęsdami šį procesą, sudarome tokią diagramą:

5| 9 7 4 9
6| 4 5 7 5 7 8 5
7| 8 6 1 9 5
8| 5 4 2 9 7 1 0
9| 8 4

Lapų tvarka neturi reikšmės. Bet jei lapai yra išdėstyti didėjimo tvarka, gauname sutvarkytą diagramą. Pažvelgę į diagramą medį, nesunkiai pastebėsime, kad:

1) didžiausias svoris yra 98 kg;
2) mažiausias svoris yra 54 kg;
3) svoriai kinta nuo 54 iki 98 kg;
4) sveriančių daugiau kaip 90 kg yra mažiausiai;
5) daugiausia yra studentų, kurių svoris yra nuo 60 iki 70 ir nuo 80 iki 90 kg;
6) *lapų* skaičius nurodo, kiek reikšmių patenka į atitinkamą intervalą.

Pastaba. Kai *lapų* labai daug, *šaką* skaidome į kelias dalis. Pavyzdžiui, mūsų nagrinėtąją diagramą medį galima užrašyti ir taip (kartu 1.28 paveiksle pateikiama studentų svorių diagrama medis, sudaryta SPSS paketu):

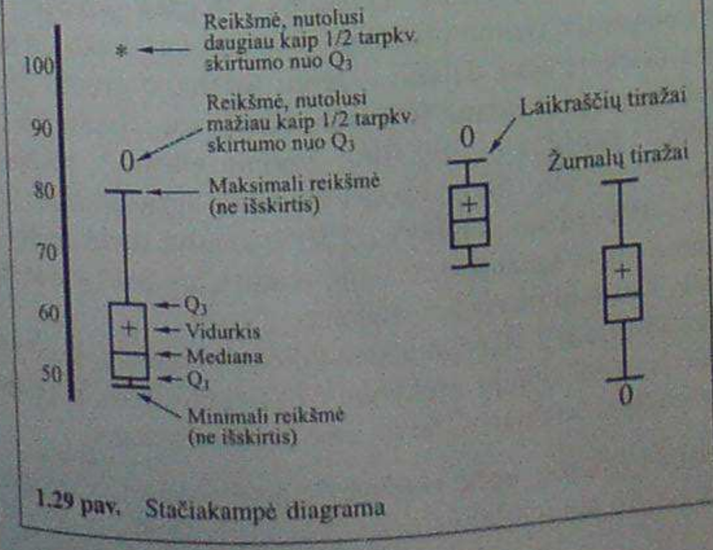
```
5*| 4
5·| 7 9 9
6*| 4
6·| 5 5 5 7 7 8
7*| 1
7·| 5 6 8 9
8*| 0 1 2 4
8·| 5 7 9
9*| 4
9·| 8
```

```
Stem-and-Leaf Plot

 Frequency    Stem & Leaf

     1,00        5 .  4
     3,00        5 .  799
     1,00        6 .  4
     6,00        6 .  555778
     1,00        7 .  1
     4,00        7 .  5689
     4,00        8 .  0124
     3,00        8 .  579
     1,00        9 .  4
     1,00        9 .  8

 Stem width:      10,00
 Each leaf          1 case(s)
```

1.28 pav. SPSS paketu gauta diagrama medis

9.4. *Stačiakampė diagrama*

Stačiakampė diagrama parodo grafinį penkiaskaitės suvestinės vaizdą (min, Q_1, Md, Q_3, max). Stačiakampėje diagramoje yra „dėžė“ – stačiakampis, braižomas nuo pirmojo kvartilio Q_1 iki trečiojo kvartilio Q_3, padalytas brūkšniu į dvi dalis ties mediana Md. (Kartais diagramose pliusu pažymimas vidurkio taškas.) Nuo stačiakampio šono brėžiami „ūsai“, besitęsiantys iki paskutinės neišsiskiriančios duomenų aibės reikšmės ir didžiausios neišsiskiriančios duomenų aibės reikšmės. Išskirčių (sąlyginių išskirčių) reikšmės pažymimos specialiais simboliais. Tipiška duomenų aibės stačiakampė diagrama pavaizduota 1.29 paveiksle. Stačiakampės diagramos leidžia palyginti keleto imčių duomenis. Tarkime, kad 1.29 paveikslo dešinėje pateiktos laikraščių ir žurnalų tiražų stačiakampės diagramos.



1.29 pav. Stačiakampė diagrama

Iš jų matyti, kad laikraščių tiražai yra ne tik beveik visi didesni už žurnalų tiražus, bet ir daug labiau koncentruoti.

Stačiakampės diagramos ypač patogios dviejų ar daugiau duomenų aibių charakteristikoms palyginti.

10. Trečioji melo rūšis

Žinomas posakis skelbia, kad yra melas, didelis melas ir statistika. Statistika tampa trečiąja melo rūšimi, kai: 1) neteisingai parenkamas matematinis modelis; 2) neteisingai interpretuojami gautieji rezultatai; 3) duomenys pateikiami taip, kad nuslepiama tikroji situacija.

Posakis apie statistiką, kaip melo rūšį, priskiriamas Markui Tvenui. Jis netiksliai citavo Dizraelį, teigusį, kad yra melas, didelis melas ir bažnytinė statistika.

Modelio parinkimo ir interpretacijos problemos smulkiau aptariamos III šio vadovėlio dalyje. Šiame skyrelyje kalbėsime, kaip pateikiami aprašomosios statistikos rezultatai.

Pradėsime nuo procentų ir dalių. Procentai pateiktini *kartu* su bendruoju stebėjimų skaičiumi. Procentas tereiškia šimtąją visumos dalį, todėl, nežinodami visumos dydžio, galime gauti iškreiptą vaizdą. Palyginkime tris teiginius:

1. Dr. J. Ankauskas yra vadovavęs dviem magistro darbams. Vienas magistro darbas buvo apgintas, kitas – ne.
2. Kas antras dr. J. Ankausko magistrantas darbo neapsigynė.
3. Penkiasdešimt procentų dr. J. Ankausko magistrantų neapsigynė magistro darbų.

Nors visi šie teiginiai teisingi, tačiau dalys ir procentai be nuorodos, kad magistrantų buvo tik du, sukelia įspūdį, kad daktaras yra nekoks vadovas.

Kalbėdamas mitinge, kandidatas į prezidentus pareiškė, kad jam atėjus į valdžią visi gaus didesnes už vidutinę algas. Ir niekas nenusijuokė...

Metodas procentais užmaskuoti absoliučiųjų dydžių menkumą yra gana paplitęs. Dažniausiai bendrasis skaičius paminimas tik tyrimo pradžioje (tarytum tarp kitko), o toliau operuojama tik procentais arba (kiek rečiau) dalimis. Jeigu tiriamoje grupėje tik 10 žmonių, tai kiekvieno iš jų stebėjimai jau sudaro 10% visų stebėjimų. Todėl frazę „iš dešimties tirtų žmonių du mano...“ pakeitus į „20% tirtų žmonių mano...“, sudaromas klaidingas įspūdis, kad „manančių“ yra daug. Beje, kartais procentai suapvalinami ir tuomet galimos kuriozinės situacijos. Pavyzdžiui, teiginys „iš 50 apklaustų studentų net 7% mano, kad sesijas, kaip studentams stresą sukeliantį reiškinį, reiktų uždrausti“ yra nelabai aiškus. Mat apskaičiavę gautume, kad taip mano 3,5 studento.

Procentai geriau nei dalys ar absoliutieji skaičiai atspindi kelių proporcijų skirtumus. Tarkime, įmonės vadovui norisi įvertinti dviejų prekybinių padalinių darbą ir (galbūt) vieną iš jų premijuoti. Štai trys visiškai analogiški faktai apie padalinių darbą:

1. Pirmame padalinyje dirba 14 žmonių, kurie pardavė 77 automobilius, o antrajame – 26 žmonės, kurie pardavė 98 automobilius.

2. Abiejuose padaliniuose dirba 40 žmonių, kurie pardavė 175 automobilius. Be to, pirmame padalinyje dirba 7/20 visų darbuotojų, kurie pardavė 11/25 visų automobilių. Antrajame dirba 13/20 visų darbuotojų, kurie pardavė 14/25 visų automobilių.
3. Abiejuose padaliniuose dirba 40 žmonių, kurie pardavė 175 automobilius. Be to, pirmame padalinyje dirba 35% visų darbuotojų, kurie pardavė 44% visų automobilių. Antrajame dirba 65% visų darbuotojų, kurie pardavė 56% visų automobilių.

Šiuo atveju procentai padeda iškart suvokti, kad pirmasis padalinys dirba geriau.

Pagrindinės manipuliacijos grafiškai vaizduojant duomenis vyksta keičiant ašių mastelius. Statistikoje nedaug grafikų, kuriuose vienetas Ox ašyje atitinka vienetą Oy ašyje. Būna, kad Ox ašies viena padala žymi 1 kilometrą, o Oy ašies viena padala žymi 0,00001 milimetro dalį. Padalų dydžiui nustatyti nėra griežtų taisyklių, todėl, parenkant šį dydį, galima skaitytojui sukelti norimą įspūdį.

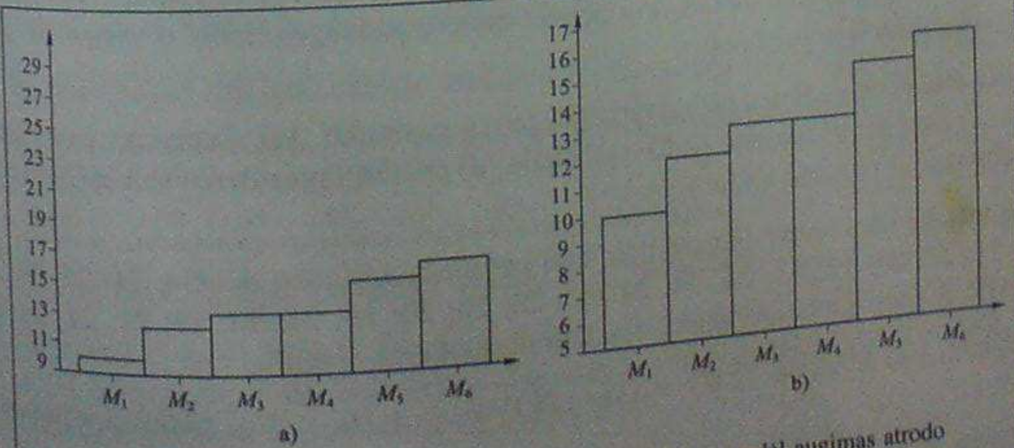
Dalydami Oy ašį į daugiau padalų, sumažiname vizualų stulpelių aukštį, arba išlyginame regimuosius kreivės svyravimus. Be to, Oy ašies padalas galima atidėti ne nuo 0, o nuo kokio nors kito skaičiaus. Taip galima užmaskuoti įmonės nuostolius, savižudybių daugėjimo tempus, vertybinių popierių nestabilumą.

Visi čia pateikti grafikai gauti SPSS paketu.

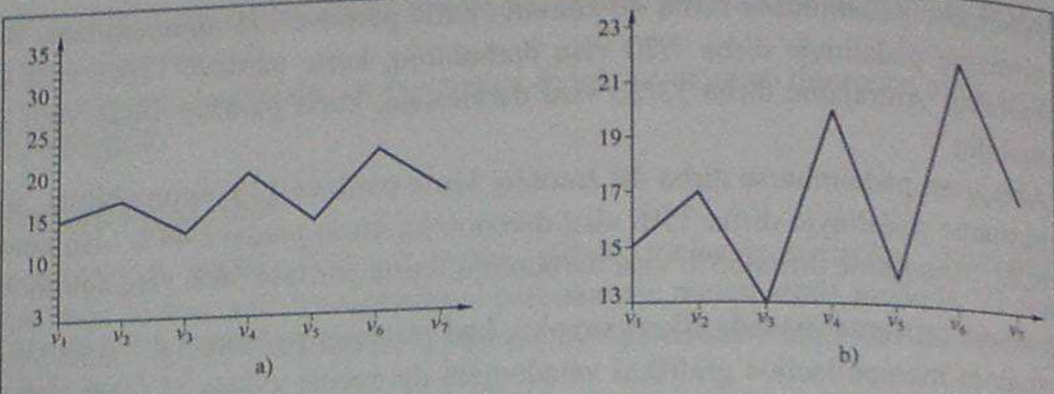
Tų pačių duomenų dvi stulpelių diagramos pateiktos 1.30 paveiksle. Stulpeliai gali simbolizuoti įmonės nuostolius, infliaciją, savižudybių skaičių, pajamų augimą ir pan. Laimėjimų didumas arba mažumas dar labiau išryškėja dėl to, kad tarp stulpelių nėra tarpų.

Tą patį efektą galima gauti ir braižant kreives. Tų pačių duomenų kreivės nubrėžtos 1.31 paveiksle. Dėl didesnio skaičiaus padalų Oy ašyje pirmos kreivės kitimas atrodo stabilesnis negu antrosios. Nagrinėjant vertybinius popierius ir jų kainų kitimą, iš tokių grafikų galima spręsti apie kainų stabilumą (verta pirkti) arba nestabilumą (rizikinga juos pirkti).

Žinoma, padalų skaičių galima didinti (mažinti) ir Ox ašyje.



1.30 pav. a) Oy ašis padalyta į daug padalų ir prasideda nuo 9, todėl augimas atrodo menkas; b) Oy ašis padalyta į mažai padalų ir prasideda nuo 5, todėl augimas atrodo didelis



1.31 pav. a) Oy ašis padalyta į daug padalų, todėl vizualiai laužtė atrodo lygesnė;
b) Oy ašis padalyta į mažai padalų, todėl vizualiai laužtė atrodo kampuotesnė

asimetrijos koeficientas
Čebyšovo taisyklė
dažnis
dažnių daugiakampis
dažnių skirstinys
diagrama medis
dispersija
eksceso koeficientas
empirinė taisyklė
garantijų funkcija
histograma
išskirtis
kitimo koeficientas
kokybinės įvairovės indeksas
kvartilis
kvartilių skirtumas
mediana
moda
normalioji kreivė
Pareto diagrama
Pirsono koreliacija
santykinis dažnis
skritulinė diagrama
stačiakampė diagrama
standartinis nuokrypis
stulpelių diagrama
sukauptųjų dažnių laužtė
vidurkis
z reikšmė

UŽDAVINIAI

1. Anksčiau pateiktame statistikų folkloro pavyzdyje kandidatas pasakė nesąmonę. Įrodykite. Situacija „visi gaus ne mažesnes už vidutinę algas“ galima. Ką ji reiškia?

2. Skyriaus pradžioje pateiktas folkloro pavyzdys apie vidutiniškai gabų žmogų yra teisingas, jeigu vidutiniškai gabus suprantamas kaip gabumų mediana. Ar liks šis teiginys teisingas, jeigu vidutiniai gabumai bus suprantami kaip gabumų vidurkis?

3. Darbdavys pareiškė, kad jo įmonėje moterys tikrai nediskriminuojamos. Pavyzdžiui, palyginti su praėjusiais metais, jo įmonėje moterų padaugėjo 50%, o vyrų – tik 10%. Sukritikuokite šį argumentą.

4. Kokią duomenų padėties charakteristiką reiktų pasirinkti, kai duomenys gauti nustačius: a) tautybę; b) šeiminę padėtį; c) amžių; d) požiūrį į prezidento veiklą („pritariu“, „nepritariu“, „nežinau tokio“).

5. Visą mėnesį parduotuvė fiksavo parduotų per dieną ausinukų skaičių: 24; 25; 23; 26; 25; 21; 27; 24; 22; 23; 24; 24; 24; 25; 28; 20; 30; 19; 24; 23; 19; 21; 24; 24; 25. Raskite dažnių skirstinį, vidurkį, standartinį nuokrypį, išskirtis.

6. Įrodykite, kad funkcija $f(M) = (x_1 - M)^2 + \cdots + (x_n - M)^2$ pasiekia minimumą taške $\bar{x}$.

7. Įrodykite vidurkio 1 savybę (pasinaikinimo efektą).

8. Paaiškinkite lygybę (1.4) formulėje.

9. Sugalvokite aibę duomenų, kurių $Md < \overline{x} < Mo$.

10. Turime duomenis apie 30 žmonių per vakarą kazino pralоštų pinigų sumas (Lt): 200; 385; 330; 275; 340; 125; 259; 432; 375; 362; 252; 309; 238; 284; 130; 310; 335; 254; 335; 202; 390; 381; 305; 455; 305; 520; 405; 516; 425; 448. Raskite $\overline{x}$, Md, Mo, Q_1, Q_3.

11. Turime duomenis: 30; 80; 50; 40 ir x. Raskite x, jeigu žinoma, kad moda, mediana ir vidurkis sutampa.

12. Įrodykite dispersijos savybes.

13. Įrodykite (1.9) formulę.

14. Sukonstruokite imtį, kurios $\overline{x} = 1$, $s^2 = 8$.

15. Turime statistikos testo rezultatus: 52; 54; 57; 49; 63; 54; 38; 46; 49; 33; 43; 40; 29; 43; 60; 69; 54; 64; 41; 63; 44; 55; 58; 55; 41; 37; 49; 36; 43; 36; 44; 35; 54; 57; 55; 56; 56; 56; 41; 49; 63; 41; 46; 45; 55; 45; 49; 47; 37; 62; 48; 44; 45; 48; 62; 56; 57; 62; 40; 46; 53; 58; 63; 62; 47; 39; 33; 58; 46; 68; 62; 57; 55; 54; 60. Sugrupuokite duomenis ($h = 5$). Nubraižykite histogramą. Raskite grupuotų duomenų Mo ir Md.

16. Ką galima pasakyti apie imtį, jei $s = 0$?

17. Turime duomenis: 8; 8; 26; 10; 8; 8; 8; 18; 8; 14; 10; 10; 6; 14; 14. Kaip pasikeis s (sumažės ar padidės) 26 pakeitus 20?

18. Apklausus dvi grupes paaiškėjo, kad pirmoje grupėje yra 30 katalikų, 20 protestantų, 5 budistai ir 10 kitokio tikėjimo atstovų. Antroje grupėje – 20 katalikų, 25 protestantai, 10 budistų ir 10 kitokių. Kuri grupė tikėjimo prasme homogeniškesnė?

19. Kaip turint duomenis rasti t reikšmes, neskaičiuojant z reikšmių? Užrašykite formulę.

20. Žinoma, kad $\overline{x} = 7$, $s = 1$. Raskite x_j, jeigu $z_j = 2x_j$.

21. Duomenys: 3, 4, 5, x_4. Raskite x_4, jeigu $z_4 = 2$, $s = 1$.

22. Turime $x_3 = 5$, $\overline{x} = 7$. Ar gali būti $z_3 = 3$?

23. Banko klerkai 100 balų skalėje nurodė savo pasitenkinimo darbu ir atlyginimu lygį:

Darbas	72	90	84	85	71	88	72	88	77
Alga	57	65	53	55	45	49	60	45	60

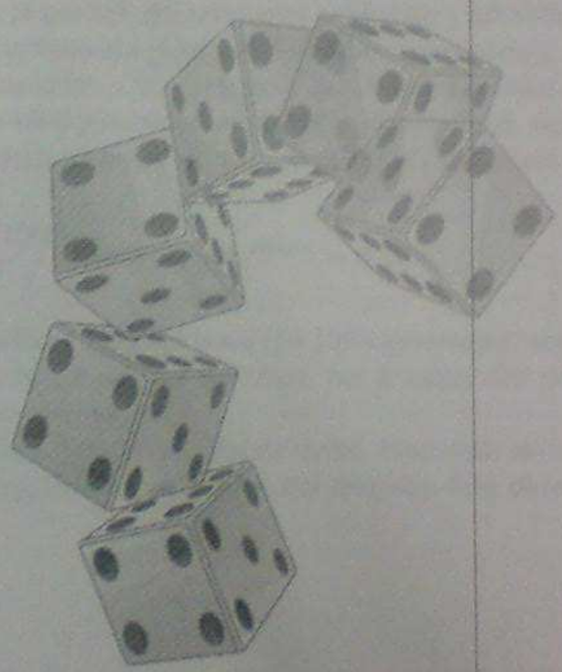
Nubraižykite stebėjimų sklaidos grafiką. Pakomentuokite.

24. Duomenys: 4,87; 4,68; 4,80; 4,76; 4,77; 4,52; 4,65; 4,90; 4,78; 4,77; 4,45; 4,72; 4,79; 4,85; 6,39; 4,49. Sudarykite diagramą medį ir penkiaskaitę suvestinę. Ar yra išskirčių? Ar duomenys pasiskirstę simetriškai?

25. Duomenys: 5,19; 6,27; 4,47; 2,86; 2,20; 4,00; 3,62; 7,15; 3,13; 9,97; 6,25; 4,71; 7,53; 5,76; 1,43; 5,94; 3,65; 4,52; 8,45; 3,45. Suapvalinkite iki vieno skaitmens po kablelio. Sudarykite diagramą medį ir penkiaskaitę suvestinę.

2

TIKIMYBIŲ TEORIJOS ELEMENTAI



Kaip mes drįstame kalbėti apie atsitiktinumo dėsnius?
Argi atsitiktinumas nėra dėsnio antitezė?

B. Raselas

Studentas į statistikos egzaminą atėjo visiškai nepasiruošęs. Kadangi į kiekvieną klausimą reikėjo pasirinkti vieną iš dviejų atsakymų, studentas nusprendė tai daryti mesdamas monetą. Profesorius dvi valandas stebėjo, kaip studentas mėto monetą ir užrašinėja atsakymus. Kiti jau seniai išėjo, o studentas vis dar mėto monetą. Neiškentęs profesorius prieina ir sako: „Matau, kad jūs egzaminui nesiruošėte ir atsakymą renkatės atsitiktinai. Kodėl gi tai trunka tiek ilgai?" Studentas: „Netrukdykit, netrukdykit – atsakymus tikrinu."

Pirmoje dalyje susipažinome, kaip aprašyti pagrindines duomenų aibės savybes. Tačiau duomenų aibė yra tik dalis tiriamos populiacijos. Pavyzdžiui, turime duomenis apie 1000 šeimų pajamas. Aprašomosios statistikos metodais galime apskaičiuoti *tik šių* šeimų vidutines pajamas. Norint padaryti pagrįstas išvadas apie *visų* šalies šeimų vidutines pajamas, aprašomosios statistikos metodų neužteks. Teks taikyti matematinius (tikimybių teorijos) modelius bei metodus. Šią knygos dalį galime vadinti statistinių išvadų įvadu. Statistinių išvadų metodai remiasi tikimybių teorijos aksiomomis bei modeliais. Norėdami geriau suprasti šios dalies tikslus, panagrinėkime tokį pavyzdį.

Per Tautogalos mero rinkimus nugalėtojas surinko 62% rinkėjų balsų. Apklausus po pusmečio 1000 atsitiktinai parinktų pilnamečių Tautogalos gyventojų, paaiškėjo, kad tik 600 iš jų pritaria mero veiklai. Ar tai reiškia, kad mero populiarumas sumažėjo?

Mero veiklai pritaria tik 60% apklaustųjų, tačiau kito tūkstančio gyventojų šis procentas gali būti ir didesnis. Aišku, kad nei teigiamas, nei neigiamas atsakymas į suformuluotą klausimą negali būti pateiktas su šimtaprocentine garantija. Statistikas samprotauja taip:

Tarkime, mero populiarumas nepasikeitė, t. y. jį kaip ir per rinkimus remia 62% gyventojų. Tada iš atsitiktinai parinkto 1000 gyventojų, pritariančių mero veiklai, skaičius X bus atsitiktinis dydis, turintis vadinamąjį binominį skirstinį, o tikimybė, kad jis neviršija 600, lygi

$$P(X \leqslant 600) = \sum_{i=0}^{600} \binom{1000}{i} (0{,}62)^i (0{,}38)^{1000-i} \approx 0{,}10.$$

Šį rezultatą galima interpretuoti taip: atlikus daug apklausų (po 1000 gyventojų), net 10% atvejų rezultatai bus mažiau palankūs merui nei per rinkimus, net ir esant tam pačiam 62% populiarumui. Taigi merui nėra ko nerimauti.

Skaitytojui gali kilti daug klausimų – kas tas atsitiktinis dydis, binominis skirstinys ir tikimybė. Šioje knygos dalyje susipažinsime su šiomis bei daugeliu kitų tikimybių teorijos sąvokų.

1. Atsitiktiniai įvykiai

1.1. Elementarieji įvykiai

Eksperimentą, kurio metu gali būti keletas atsitiktinių baigčių, vadinsime tikimybiniu. Atliekant tikimybinį eksperimentą, negalima iš anksto pasakyti, kuri iš galimų baigčių įvyks. Tikimybinių eksperimentų ir jų galimų baigčių pavyzdžiai pateikti 2.1 lentelėje.

Atsitiktinės eksperimento baigtys vadinamos *atsitiktiniais įvykiais*. Įvykiai, kurie smulkiau neskaidytini, vadinami *elementariaisiais*. Pateiktoje 2.1 lentelėje „skaičius", „herbas", „1 taškas", „2 taškai" ir pan., „6 taškai" yra elementarieji. Įvykius „tūzas", „ne tūzas" galima laikyti elementariaisiais tik tuomet, jeigu darant eksperimentą (traukiant kortą), mūsų nedomina galimybė ištraukti kitas kortas. Priešingu atveju elementarieji įvykiai yra: „kryžių tūzas", „pikų dama" ir pan. – iš viso 24 elementarieji įvykiai. Taigi elementarieji įvykiai tam tikra prasme yra bazinės, smulkiausios eksperimento baigtys.

2.1 lentelė. Atsitiktinės eksperimentų baigtys

Eksperimentas	Įvykis
Metama moneta	Skaičius, herbas
Metamas kauliukas	1, 2, 3, 4, 5, 6 akutės
Iš 24 kortų traukiama viena	Tūzas, ne tūzas
Tikrinamas gaminys	Brokuotas, geras
Krepšininkas meta baudą	Pataiko, nepataiko

Visų elementariųjų įvykių aibė Ω vadinama *elementariųjų įvykių erdve*.

Elementarieji įvykiai žymimi simboliais $\omega_1, \omega_2, \ldots, \omega_n$, o elementariųjų įvykių erdvė – simboliu Ω. Aišku, kad $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Pavyzdžiui, 2.1 lentelėje minėtų eksperimentų:

$$\Omega = \{\text{skaičius, herbas}\}, \qquad \Omega = \{1, 2, 3, 4, 5, 6\}, \qquad \Omega = \{\text{tūzas, ne tūzas}\},$$

$$\Omega = \{\text{geras, brokuotas}\}, \qquad \Omega = \{\text{pataiko, nepataiko}\}.$$

1.2. Įvykiai ir veiksmai su jais

Eksperimento baigtį galima nusakyti ne vien elementariaisiais įvykiais. Pavyzdžiui, metant kauliuką, galima baigtis (atsitiktinis įvykis) yra „atsivers lyginis skaičius akučių". Elementariųjų įvykių erdvės įvykiai (juos žymėsime $A, B, \ldots$) yra ne kas kita kaip Ω poaibiai, sudaryti iš elementariųjų įvykių. Pavyzdžiui, įvykis $A = \{$lyginis atsivertusių akučių skaičius$\} = \{2, 4, 6\}$. Kadangi Ω irgi sudaro elementarieji įvykiai (Ω yra savo pačios poaibis), tai į Ω galima pažiūrėti kaip į tam tikrą įvykį, kuris visuomet įvyksta. Įvykis, kuris įvyksta, įvykus bet kuriai eksperimento baigčiai, vadinamas *būtinuoju* įvykiu. Kaip priešprieša būtinajam įvykiui apibrėžiamas *negalimasis* įvykis $\varnothing$, t. y. toks, kuris atliekant eksperimentą įvykti negali. Pavyzdžiui, metant kauliuką, įvykis $\{$nelyginis atsivertusių akučių skaičius yra didesnis už 5$\}$ yra negalimas. Žymėsime: Ω – būtinąjį įvykį, $\varnothing$ – negalimąjį įvykį.

Vieni įvykiai gali būti kitų įvykių dalys.

Įvykis A yra įvykio B *dalis*, t. y. $A \subset B$, jeigu įvykus įvykiui A, įvyksta ir įvykis B.

Elementariųjų įvykių erdvėje A yra B dalis, jeigu kiekvienas elementarusis įvykis, įeinantis į įvykį A, įeina ir į įvykį B (todėl kartais dar sakome, kad A yra B *poaibis*). Sąryšio $A \subset B$ prasmė pavaizduota 2.1 paveiksle, a), t. y. vadinamojoje Veno[1] diagramoje. Tarkime, metamas kauliukas. Tuomet įvykis $A = \{3\}$ yra įvykio $B = \{$atsivers nelyginis skaičius akučių$\}$ dalis. Kiekvienas įvykis A yra būtinojo įvykio Ω dalis, t. y. $A \subset \Omega$.

Nors kiekvienas įvykis yra būtinojo įvykio poaibis, tačiau bendruoju tikimybių teorijos atveju (Ω struktūra daug sudėtingesnė nei baigtinė elementariųjų įvykių aibė) ne kiekvienas aibės Ω poaibis laikomas įvykiu.

[1] John Venn (1834–1923) – anglų logikas.

Du įvykius vadiname *lygiais*, t. y. $A = B$, jeigu A yra B dalis, o B yra A dalis.

Elementariųjų įvykių erdvėje $A = B$, jeigu juos sudarančios elementariųjų įvykių aibės sutampa. Pavyzdžiui, metant kauliuką, įvykiai $A = \{2, 4, 6\}$ ir $B = \{$lyginis atsivertusių akučių skaičius$\}$ yra lygūs.

Apibrėšime kai kuriuos atsitiktinių įvykių ryšius.

Įvykių A ir B *sąjunga* $A \cup B$ vadiname įvykį, kai įvyksta *bent vienas* iš įvykių A ir B.

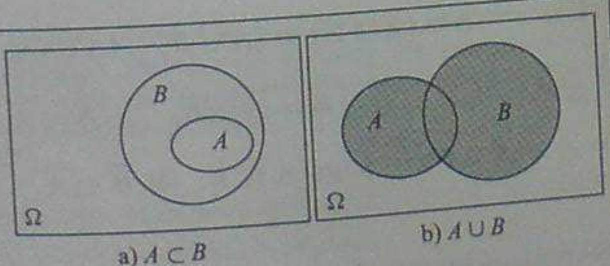
Elementariųjų įvykių erdvėje $A \cup B$ žymi įvykį, sudarytą iš elementariųjų įvykių, priklausančių bent vienam iš įvykių A ir B. Įvykių sąjungos Veno diagrama pavaizduota 2.1 paveiksle, b). Šnekamojoje kalboje įvykių $A \cup B$ sąjungą atitinka teiginys: įvyks A *arba* B. Tegul metant kauliuką $A = \{1, 2\}$, $B = \{2, 3\}$. Tuomet $A \cup B = \{1, 2, 3\}$.

Atkreipiame dėmesį, kad atliekant veiksmus su atsitiktiniais įvykiais pasikartojantys elementai nedvigubinami. Pavyzdžiui, jeigu $\omega_1 \in A$ ir $\omega_1 \in B$, tai tik vienas ω_1 priklausys $A \cup B$.

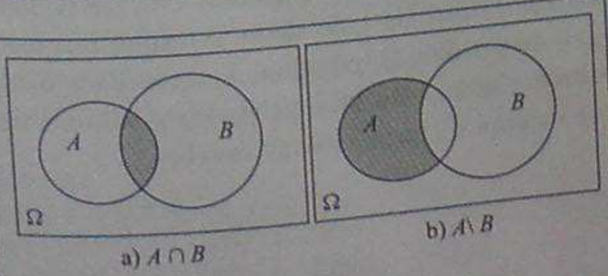
Įvykių A ir B *sankirta* $A \cap B$ vadiname įvykį, kai kartu įvyksta abu įvykiai A ir B.

Elementariųjų įvykių erdvėje $A \cap B$ yra įvykis, sudarytas iš visų *bendrųjų* elementariųjų įvykių. Sankirtos Veno diagrama pavaizduota 2.2 paveiksle, a). Šnekamojoje kalboje įvykių $A \cap B$ sankirtą atitinka teiginys: įvyks A *ir* B. Tegul metant kauliuką $A = \{1, 2\}$, $B = \{2, 3\}$. Tuomet $A \cap B = \{2\}$.

Panašiai apibrėžiama ir didesnio įvykių skaičiaus sąjunga bei sankirta.



2.1 pav. Įvykio dalis ir įvykių sąjunga



2.2 pav. Įvykių sankirta ir įvykių skirtumas

> Įvykiai A ir B vadinami *nesutaikomaisiais* $A \cap B = \varnothing$, jeigu jie negali įvykti kartu.

Metant kauliuką, $A = \{1, 2\}$ ir $B = \{3, 4\}$ yra nesutaikomi. Nesunku įsitikinti, kad bet koks įvykis A ir negalimas įvykis $\varnothing$ yra nesutaikomi.

> Dviejų įvykių A ir B *skirtumu* $A \setminus B$ vadinamas įvykis, kai įvyksta A, o B neįvyksta.

Elementariųjų įvykių erdvėje $A \setminus B$ žymi įvykį, sudarytą iš tų A elementų, kurie nepriklauso B. $A \setminus B$ Veno diagrama pavaizduota 2.2 paveiksle, b). Atkreipiame dėmesį, kad įvykiai $A \setminus B$ ir $B \setminus A$ vienas kitu neišsireiškiami. Tarkime, metant kauliuką, $A = \{1, 2\}$, $B = \{2, 3\}$. Tuomet $A \setminus B = \{1\}$, o $B \setminus A = \{3\}$.

> Įvykis $\overline{A} = \Omega \setminus A$ vadinamas *priešinguoju* įvykiui A.

Metant monetą, įvykis $A = \{\text{herbas}\}$ yra priešingas įvykiui $\overline{A} = \{\text{skaičius}\}$. Metant kauliuką, $B = \{\text{atsivertusių akučių} > 3\}$, $\overline{B} = \{\text{atsivertusių akučių} \leqslant 3\}$. Įvykis ir jo priešingas įvykis yra nesutaikomi.

Suformuluosime kai kurias veiksmų su įvykiais savybes. Tegul A, B ir C yra bet kokie įvykiai.

$$A \cup A = A, \quad A \cap A = A, \quad A \subset \Omega, \quad A \cup \Omega = \Omega, \quad A \cap \Omega = A,$$

$$A \cup \varnothing = A, \quad A \cap \varnothing = \varnothing, \quad A \setminus \Omega = \varnothing, \quad \varnothing \setminus A = \varnothing, \quad A \setminus A = \varnothing,$$

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C), \quad A \subset A \cup B,$$

$$A \cap B = B \cap A, \quad (A \cap B) \cap C = A \cap (B \cap C), \quad A \cap B \subset A,$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

$$\overline{\Omega} = \varnothing, \quad \overline{\varnothing} = \Omega, \quad \overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}, \quad \overline{\overline{A}} = A,$$

$$A \cup \overline{A} = \Omega, \quad A \setminus B = A \setminus (A \cap B); \quad \text{jeigu } A \subset B, \text{ tai } \overline{B} \subset \overline{A}.$$

1.3. Įvykių σ algebra

Tikimybių teorijoje vartojama ir abstrakti elementariųjų įvykių erdvės Ω sąvoka. Bendruoju atveju Ω gali būti bet kuri netuščia aibė ir turėti be galo daug reikšmių. Ką tuomet laikyti atsitiktiniais įvykiais? Jeigu Ω turi tik baigtinį skaičių elementų, tai įvykiais galima laikyti visus jos poaibius. Tačiau kai Ω turi be galo daug skirtingų elementų, tai begalinės jų sąjungos bei sankirtos gali būti tokie Ω poaibiai, kad, įvedant tikimybės sąvoką, kils dideli matematiniai sunkumai. Todėl atsitiktiniais įvykiais laikoma tik tam tikra, pakankamai „turtinga" Ω poaibių sistema $\mathbf{F}$, turinti tokias savybes:

1 $\Omega \in \mathbf{F}$,

2 $A \in \mathbf{F} \Rightarrow \overline{A} \in \mathbf{F}$, 3 $A_1, A_2, \ldots \in \mathbf{F} \Rightarrow A_1 \cup A_2 \cup \cdots \in \mathbf{F}.$

Poaibių sistema **F** vadinama atsitiktinių įvykių σ (sigma) algebra. Jai tinka visi anksčiau pateikti apibrėžimai ir savybės. Šiame vadovėlyje neakcentuosime σ algebros vaidmens. Kalbėdami apie atsitiktinius įvykius, turėsime omenyje, kad jie priklauso tam tikrai σ algebrai.

2. Statistinis ir klasikinis tikimybės apibrėžimai

Kai atliekant eksperimentą galimos kelios skirtingos jo baigtys, pageidautina įvertinti kiekvienos baigties galimybę įvykti. Tikimybė ir yra atsitiktinės baigties įvykio galimybės *skaitinis* matas. Susipažinsime su dviem tradiciniais tikimybės apibrėžimais.

2.1. Statistinis (dažnių) tikimybės apibrėžimas

Daug kartų kartodami eksperimentą, stebime, kaip dažnai konkreti eksperimento baigtis pasikartoja. Jeigu didinant eksperimentų skaičių šis dažnis stabilizuojasi ties kokiu nors skaičiumi, tai pastarąjį ir laikome tiriamos baigties (įvykio) statistine tikimybe. Tarkime, kad eksperimentas gali pasisekti arba ne (2 baigtys). Atlikus n eksperimentų, m kartų jis pasisekė. Tuomet sėkmių santykinis dažnis

$$\nu_n = \frac{m}{n}. \tag{2.1}$$

m – sėkmių sk.
n – eksperimentų

Tarkime, kad n didėjant ν_n stabilizuojasi ties skaičiumi p ($\lim_{n\to\infty} \nu_n = p$). Tuomet p vadinama atliekamo eksperimento sėkmės tikimybe. Būtina, kad: a) visi eksperimentai vyktų visiškai vienodomis sąlygomis, b) vieno eksperimento rezultatai neturėtų įtakos kito eksperimento rezultatams (eksperimentai būtų nepriklausomi).

2.1 pavyzdys. Norėdami nustatyti herbo atsivertimo tikimybę, monetą mėtė daugelis žymių matematikų – G. Biufonas, K. Pirsonas, Dž. Kerichas. Kai kurie jų rezultatai pateikti 2.2 lentelėje. Matome, kad herbo atsivertimo dažnis stabilizuojasi ties 0,5, todėl herbo atsivertimo tikimybė, metant *vieną* kartą, yra 0,5. Žinoma, jeigu moneta būtų nesimetriška (pvz., jubiliejinė su itin solidžiu jubiliato atvaizdu), herbo atsivertimo tikimybė būtų kita.

Statistinis tikimybės apibrėžimas retai vartojamas tikimybių teorijoje, tačiau nepakeičiamas *interpretuojant* tikimybes. Pavyzdžiui, teiginys „brokuotos detalės pagaminimo tikimybė yra 0,1“ interpretuojamas taip: gaminant daug detalių, viena iš dešimties bus brokuota, arba tikėtina, kad bus 10% broko. Šnekamojoje kalboje sakoma „20% tikimybė“ ir pan. Iš (2.1) matyti, kad tikimybė yra skaičius tarp 0 ir 1, todėl tokį pasakymą reikia suprasti kaip 0,20.

Būdingi statistinės tikimybės taikymo pavyzdžiai: Vatikane 100% gyventojų nevedę – tikimybė, kad atsitiktinai parinktas Vatikano gyventojas bus nevedęs, yra lygi 1; 65%

2.2 lentelė

Mesta kartų	1000	4040	12 000	24 000
Atsivertė herbas	511	2048	6019	12 012
Dažnis	0,511	0,5069	0,50158...	0,5005

tam tikros rūšies operacijų sėkmingos – tikimybė, kad operacija bus sėkminga, yra lygi 0,65.

Vienas profesorius, šiaip jau labai atsargus vairuotojas, sankryžas pralėkdavo visu greičiu. Savo vairavimo manierą jis aiškino taip: „Statistika liudija, kad avarijos dažniausiai įvyksta sankryžose. Todėl aš stengiuosi jose neužsibūti“.

Su statistiniu tikimybės apibrėžimu nereikia painioti *subjektyviosios* tikimybės, kuri tėra subjektyvus kalbančiojo nuomonės apie galimą reiškinį atspindys. Kai kandidatas į prezidentus sako: „Manau, kad mano pergalės rinkimuose tikimybė 90% (0,9)“, tai tereiškia kandidato (optimistinę) nuomonę apie jo galimybes. Niekas nekartos rinkimų daugybę kartų ir neskaičiuos, kiek kartų kandidatas laimėjo.

2.2. *Klasikinis tikimybės apibrėžimas*

Klasikinis tikimybės apibrėžimas formuluojamas tik baigtinėms elementariųjų įvykių aibėms Ω, kurios visi elementarieji įvykiai vienodai galimi. Pavyzdžiui, metant simetrišką monetą, įvykiai „atsivers herbas“ ir „atsivers skaičius“ vienodai galimi. Traukiant kortą iš 24 kortų kaladės, įvykiai „tūzas“ ir „ne tūzas“ nėra vienodai galimi (antrasis įvykis labiau tikėtinas). Įvykio A tikimybe laikysime įvykį A sudarančių elementariųjų įvykių ir visų elementariųjų įvykių skaičių santykį. Tarkime, kad A yra atsitiktinis įvykis. Taigi

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}, \qquad A = \{w_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\},$$

čia $\omega_1, \ldots, \omega_n$ yra *vienodai galimi* elementarieji įvykiai, o įvykį A sudaro kuri nors šių elementariųjų įvykių dalis. Indeksai $i_1, i_2, \ldots, i_k$ yra $1, 2, \ldots, n$ poaibis. Klasikiniam tikimybės apibrėžimui nesvarbu, kurie konkretūs elementarieji įvykiai sudaro įvykį A, tačiau svarbu, *kiek* jų yra (mūsų atveju jų yra k). Įvykio A tikimybė apibrėžiama taip:

$$P(A) = \frac{k}{n}. \tag{2.2}$$

Taigi klasikinis tikimybės apibrėžimas yra toks:

$$P(A) = \frac{A \text{ sudarančių elementariųjų įvykių skaičius}}{\text{visų elementariųjų įvykių skaičius}}.$$

Iš klasikinio tikimybės apibrėžimo išplaukia, kad tikimybė yra neneigiamas skaičius, neviršijantis vieneto.

2.2 pavyzdys. Metame simetrišką monetą. Vienodai galimi du elementarūs įvykiai: h – atsivers herbas, s – atsivers skaičius. Raskime įvykio $A = \{\text{atsivers herbas}\}$ tikimybę. Taigi

$$\Omega = \{h, s\}, \qquad A = \{h\}.$$

Iš viso yra 2 vienodai galimi elementarūs įvykiai, o įvykį A sudaro 1 elementarusis įvykis. Todėl ieškomoji tikimybė

$$P(A) = \frac{1}{2} = 0{,}5.$$

Nesunku patikrinti, kad tokia pat ir skaičiaus atsivertimo tikimybė.

2.3 pavyzdys. Metame simetrišką kauliuką. Raskime įvykio $A =$ {atsivertusių akučių ne mažiau kaip 3} tikimybę. Šiuo atveju

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \qquad A = \{3, 4, 5, 6\},$$

$$P(A) = \frac{4}{6} = \frac{2}{3}.$$

2.4 pavyzdys. Iš 24 kortų kaladės (4 rūšys po 6 kortas) traukiama korta. Kokia tikimybė ištraukti tūzą? Elementarieji įvykiai „tūzas" ir „ne tūzas" nėra vienodai galimi. Todėl, norint taikyti klasikinį tikimybės apibrėžimą, reikia imti sudėtingesnį Ω.

Nesunku suprasti, kad Ω, apimanti visas kortas ($\Omega =$ {ištrauktas kryžių tūzas, pikų dama, ...}, iš viso 24 elementarūs įvykiai), jau sudaryta iš vienodai galimų elementariųjų įvykių. Įvykį „tūzas" sudaro 4 elementarūs įvykiai (kryžių tūzas, būgnų tūzas, pikų tūzas, čirvų tūzas). Trumpai tariant, galima ištraukti bet kurią iš 24 kortų, o mus tenkina bet kuris iš 4 tūzų. Todėl ieškomoji tikimybė yra $4/24 = 1/6$.

3. Klasikinės tikimybės taikymas uždaviniams spręsti

3.1. Kelios kombinatorikos formulės

Norėdami sužinoti, kiek elementariųjų įvykių sudaro A ir Ω, neišsiversime be kombinatorikos formulių. Susipažinsime su kai kuriomis iš jų. Kombinatorikos formules pateiksime kaip atsakymus į standartinius klausimus.

1 Junginiai, gaunami n objektų išrikiavus į eilę, vadinami kėliniais. Jų skaičius $n!$

> Keliais skirtingais būdais n objektų galima išrikiuoti į eilę?
>
> *Atsakymas*: $n! = 1 \cdot 2 \cdot 3 \cdots n$.

Klausimo variantai: Kiek skirtingų eilių galima sudaryti iš tų pačių n objektų? Keliais skirtingais būdais ant n kėdžių galima susodinti n žmonių?

Atsakymo pagrindimas: Pirmasis objektas gali užimti n pozicijų, antrajam lieka $(n-1)$ pozicija, trečiajam – $(n-2)$ ir pan., paskutiniam – 1.

2.5 pavyzdys. Kiek skirtingų frazių galima sudaryti iš trijų žodžių: *dėstytojas, visada, teisus*?
Atsakymas: $3! = 1 \cdot 2 \cdot 3 = 6$.

2 Junginiai, gauti iš n objektų išrinkus k skirtingų atsižvelgiant į jų išrinkimo tvarką, vadinami gretiniais be pasikartojimo. Jų skaičius $A_n^k = n(n-1)\cdots(n-k+1)$.

> Keliais skirtingais būdais iš n objektų galima išrinkti k objektų (pakliuvimo į išrinktųjų grupę eilė svarbi)?
>
> *Atsakymas*: $A_n^k = \dfrac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1)$.

Patekimo į atrinktųjų sąrašą eilė svarbi (grupė (direktorius Jonas, pavaduotojas Petras) skiriasi nuo grupės (direktorius Petras, pavaduotojas Jonas)).

Atsakymo pagrindimas: Į pirmą vietą pretenduoja n objektų, į antrąją – $(n-1)$ objektas ir pan., į k-ąją – $(n-k+1)$ objektas.

2.6 pavyzdys. Iš 50 radijo konkurso dalyvių laiškų atsitiktinai renkami 3. Pirmo išrinkto laiško autorius gaus automobilį, antrojo – kelionę į Prahą, trečiojo – marškinėlius su užrašu „Klausau ir myliu". (Taigi išrinktiesiems *svarbu*, kokia tvarka jie renkami.) Kiek skirtingų laimėtojų trejetų galima sudaryti?

Atsakymas: $A_{50}^3 = 50 \cdot 49 \cdot 48 = 117\,600$.

3 Junginiai, gauti iš n objektų išrinkus k skirtingų neatsižvelgiant į jų išrinkimo tvarką, vadinami deriniais be pasikartojimo. Jų skaičius $\binom{n}{k}$.

Keliais skirtingais būdais iš n objektų galima išrinkti k objektų (pakliuvimo į išrinktųjų grupę eilė nesvarbi)?

Atsakymas: $$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

Pakliuvimo į grupę eilės nesvarba reiškia, kad svarbi tik atrinktosios grupės sudėtis, bet nesvarbu, ar objektas į grupę pateko pirmas ar antras (grupė (Jonas, Petras) nesiskiria nuo grupės (Petras, Jonas)). Kartais dar vartojamas simbolis C_n^k.

Atsakymo pagrindimas: Turime A_n^k grupių, kurių sudarymo tvarka svarbi. Turėdami k elementų, iš jų galime sudaryti $k!$ grupių, besiskiriančių tik sudarymo tvarka. Taigi „skirtingų grupių, kurių sudarymo tvarka svarbi, skaičius" = „skirtingai sudarytų grupių skaičius" × $k!$. Arba $A_n^k = \binom{n}{k}k!$.

Iš apibrėžimo matyti, kad

$$\binom{n}{k} = \binom{n}{n-k}.$$

Skaičiuoti derinius patogiausia užrašius vardiklyje $k!$ sandauga $1 \cdot 2 \cdot 3 \cdots k$, o skaitiklyje tiek pat narių sandaugą, pradedant skaičiumi n. Pavyzdžiui:

$$\binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} = 120, \qquad \binom{17}{15} = \binom{17}{17-15} = \binom{17}{2} = \frac{17 \cdot 16}{1 \cdot 2} = 136.$$

2.7 pavyzdys. Iš 30 TV konkurso nugalėtojų reikia sudaryti 25 žmonių grupę, kuri vyks ilsėtis į Bahamus. Kiek skirtingų grupių galima sudaryti?

Atsakymas:

$$\binom{30}{25} = \binom{30}{5} = \frac{30 \cdot 29 \cdot 28 \cdot 27 \cdot 26}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 142\,506.$$

3.2. Uždavinių pavyzdžiai

1 Tarkime, dėžėje yra a baltų ir b juodų rutulių. Atsitiktinai ištraukiame vieną rutulį. Kokia tikimybė, kad jis baltas?

Sprendimas. Elementariųjų įvykių erdvę sudaro visi rutuliai, jų yra $a+b$. Įvykį $A = \{$rutulys baltas$\}$ sudaro visi balti rutuliai. Jų yra a. Todėl pagal (2.2) formulę

$$P(A) = \frac{a}{a+b}. \tag{2.3}$$

2 Dėžėje yra a baltų ir b juodų rutulių. Atsitiktinai ištraukiame 2 rutulius. Kokia tikimybė, kad jie skirtingų spalvų?

Sprendimas. Elementariųjų įvykių erdvė yra visos galimos rutulių poros. Tokių porų yra $\binom{a+b}{2}$. Ieškomą atsitiktinį įvykį (pažymėkime jį A) sudaro visos skirtingų spalvų poros. Jų yra ab. *Atsakymas:*

$$P(A) = \frac{ab}{\binom{a+b}{2}}.$$

3 Dėžėje yra a baltų ir b juodų rutulių. Atsitiktinai ištraukiame m rutulių. Kokia tikimybė, kad tarp ištrauktųjų bus k baltų ir $m-k$ juodų rutulių?

Sprendimas. Elementariųjų įvykių erdvė yra visi galimi rinkiniai po m rutulių. Tokių rinkinių yra $\binom{a+b}{m}$. Ieškomą atsitiktinį įvykį (pažymėkime jį A) tenkina bet koks k baltų rutulių rinkinys (tokių rinkinių yra $\binom{a}{k}$) ir bet koks $m-k$ juodų rutulių rinkinys (tokių rinkinių yra $\binom{b}{m-k}$). *Atsakymas:*

$$P(A) = \frac{\binom{a}{k}\binom{b}{m-k}}{\binom{a+b}{m}}.$$

4 Žaidimo automatas vienodai galimai generuoja triženklius skaičius (nuo 000 iki 999). Kokia tikimybė, kad jis sugeneruos 888?

Sprendimas. Iš viso gali sugeneruoti 1000 skirtingų skaičių. Tenkina vienas. *Atsakymas:* $1/1000 = 0{,}001$.

5 Iš 24 kortų kaladės atsitiktinai ištraukiame 3 kortas. Kokia tikimybė, kad į trejetuką pateko vienas tūzas, viena dama ir vienas karalius?

Sprendimas. Iš viso galima sudaryti $\binom{24}{3}$ skirtingų trejetų. Tenkina bet kuris iš 4 tūzų, bet kuri iš 4 damų ir bet kuris iš 4 karalių. Tokių rinkinių 4^3. *Atsakymas:*

$$\frac{4^3}{\binom{24}{3}} = \frac{4^3 \cdot 3!}{24 \cdot 23 \cdot 22} = 0{,}03162\ldots$$

6 Būrėja paima tūzą, damą ir karalių ir leidžia klientei atsitiktinai šias kortas išdėlioti. Kokia tikimybė, kad jas atvertus klientė išvys tūzą, karalių, damą?

Sprendimas. Turime 3 objektus. Iš jų galima sudaryti $3! = 6$ skirtingas eiles. Tenkina tik 1 variantas. *Atsakymas:* $1/6$.

7 Tegul tenkinamos ankstesnio uždavinio sąlygos. Kokia tikimybė, kad vidurinė korta bus dama?

Sprendimas. Tenkina variantai: „tūzas, dama, karalius" ir „karalius, dama, tūzas". *Atsakymas:* $2/3! = 1/3$.

8 Iš vienodą rezultatą individualiose lenktynėse pasiekusių penkiolikos dviratininkų (Jono, Fransua, Ibrahimo ir pan.) atsitiktinai parenkami etapo nugalėtojai. Kokia tikimybė, kad Jonui atiteks pirmoji vieta, o Fransua ir Ibrahimas pasidalys antrąja ir trečiąja vietomis?

Sprendimas. Iš 15 dviratininkų galima sudaryti (tvarka svarbi!) A_{15}^3 trejetų. Norimą įvykį tenkina dvi situacijos: (Jonas, Fransua, Ibrahimas) ir (Jonas, Ibrahimas, Fransua). *Atsakymas:*

$$\frac{2}{A_{15}^3} = \frac{2}{15 \cdot 14 \cdot 13} = 0{,}00073\ldots$$

3.3. Kortos, monetos, kauliukai, rutuliai

Beveik visuose tikimybių teorijos uždavinynuose yra daugybė uždavinių, susijusių su kortomis, monetų mėtymu, kauliukais ir rutuliais. Taip yra ne todėl, kad tikimybininkams trūktų fantazijos ar jie jaustų silpnybę azartiniams žaidimams.

Istoriškai kai kurie tikimybiniai uždaviniai atsirado sprendžiant lošimo problemas (galima paminėti P. Ferma[1] ir B. Paskalio[2] susirašinėjimą).

Tokie uždaviniai pateikiami todėl, kad tai vaizdūs ir lengvai suvokiami uždavinių *modeliai*. Visai nesunku prikurti uždavinių, kuriuose po žodžių apvalkalu slypi viena iš paprastų, rutuliais, kortomis ar pan. objektais jau aprašytų schemų. Tarkime, turime tokį uždavinį (dėl tęstinumo, pažymėkime jį numeriu 9).

9 Druskininkų miške neeilinėje tarybos sesijoje 150 ežių svarstė tokį nutarimo projektą: „Kad nebūtų pažeista būtinoji gintis, užpultas ežys turi:

a) garsiai įspėti užpuoliką, kad durs;
b) bakstelėti spygliais į orą;
c) durti užpuolikui."

Per apklausą paaiškėjo, kad 30 ežių pritaria visiems nutarimo punktams; 60 ežių pritaria tik punktui c); 10 ežių (pacifistų) pritaria tik punktui a); o likę ežiai pritaria tik punktams a) ir c). Žurnalistas iš Vakarų Europos paklausė vieno sesijos dalyvio nuomonės apie projektą. Kokia tikimybė, kad tas dalyvis buvo ežys pacifistas?

Sprendimas. Nesunku atspėti, kokia schema glūdi šioje sąlygoje. Įsivaizduokime ežius, susirietusius į kamuoliukus (rutulius). Yra 10 ežių rutulių pacifistų (tarkime, nusidažiusių spyglius baltai) ir $150 - 10 = 140$ pilkų ežių rutulių. Taigi 10 baltų ir 140 pilkų rutulių. Atsitiktinai ištraukiame vieną (paimame interviu). Kokia tikimybė, kad jis baltas (pacifistas)? Schema atitinka nagrinėtąją 1 uždavinyje. *Atsakymas:*

$$\frac{10}{140 + 10} = \frac{10}{150} = 0{,}066666\ldots$$

Kitos informacijos neprireikė.

---

[1] Pierre de Fermat (1601–1665) – prancūzų matematikas.
[2] Blaise Pascal (1623–1662) – prancūzų matematikas.

4. Bendrasis tikimybės apibrėžimas

Labai retai bandymus galima kartoti daugybę kartų, todėl statistinis tikimybės apibrėžimas ne itin tinka bendrajai tikimybių teorijai. Klasikinis tikimybės apibrėžimas netinka, jeigu Ω nėra baigtinis arba elementarieji įvykiai nėra vienodai galimi. Todėl reikia bendresnio tikimybės apibrėžimo.

Tikimybe vadiname funkciją $P : \Omega \to [0, 1]$, kuri kiekvienam atsitiktiniam įvykiui A priskiria skaičių $P(A)$ ir:

1) $0 \leqslant P(A) \leqslant 1$,

2) $P(\Omega) = 1$,

3) $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots$, jeigu $A_i \cap A_j = \varnothing$, $i \neq j$.

Matome, kad bendrasis tikimybės apibrėžimas nenusako konkretaus tikimybės skaičiavimo metodo. Tai tik bendrosios savybės, kurias tenkina daug modelių – ir kauliuko mėtymas, ir rutulių traukimas, ir korektūros klaidų ieškojimas. Klasikinė tikimybė tenkina visas bendrojo apibrėžimo savybes (įrodykite). Griežtesnis tikimybės apibrėžimas reikalautų sąlygos, kad visi įvykiai imami tik iš tam tikros σ algebros.

Nesvarbu, kokį modelį tiriame, jo realizuota tikimybė tenkina bendrąjį tikimybės apibrėžimą. Taip pat galioja ir iš bendrojo tikimybės apibrėžimo išplaukiančios savybės, kurias pateikiame šiame skyrelyje.

1 $P(\varnothing) = 0$.

Įrodymas. $\varnothing \cap \varnothing = \varnothing$ ir $\varnothing \cup \varnothing = \varnothing$, todėl pagal tikimybės apibrėžimo 3) sąlygą

$$P(\varnothing) = P(\varnothing \cup \varnothing \cup \varnothing \ldots) = P(\varnothing) + P(\varnothing) + \cdots. \tag{2.4}$$

Nulis yra vienintelis skaičius, su kuriuo teisinga ši lygybė.

2 $P(A \setminus B) = P(A) - P(A \cap B)$.

Įrodymas. Nesunku nustatyti (ypač iš Veno diagramos), kad $A = (A \setminus B) \cup (A \cap B)$. Be to, šie įvykiai nesutaikomi. Todėl pagal tikimybės apibrėžimo 3) sąlygą

$$P(A) = P(A \setminus B) + P(A \cap B).$$

3 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Įrodymas. $A \cup B = (A \setminus B) \cup B$ ir šie įvykiai nesutaikomi. Todėl pagal tikimybės apibrėžimo 3) sąlygą

$$P(A \cup B) = P(A \setminus B) + P(B).$$

4 Jeigu $A \subset B$, tai $P(A) \leqslant P(B)$.

Įrodymas. $A \cap B = A$, todėl pagal 2) savybę

$$P(B) = P(B \setminus A) + P(A).$$

Bet kuri tikimybė neneigiama, todėl $P(B \setminus A) \geqslant 0$.

5 $P(A) = P(A \cap B) + P(A \cap \overline{B})$.

Įrodymas. Faktiškai tai ta pati 2) savybė, tik užrašyta sankirtomis.

6 Bet kokiam įvykiui A teisinga lygybė

$$P(\overline{A}) = 1 - P(A). \tag{2.5}$$

Įrodymas. $\overline{A} = \Omega \setminus A$ ir $A \cap \Omega = A$, todėl pagal 2) savybę ir tikimybės apibrėžimo 2) sąlygą

$$P(\overline{A}) = P(\Omega \setminus A) = P(\Omega) - P(\Omega \cap A) = 1 - P(A).$$

Ši priešingo įvykio tikimybės skaičiavimo formulė yra labai svarbi uždaviniams spręsti. Pateiksime vieną pavyzdį.

2.8 pavyzdys. Tarkime, kad grupėje yra a mėlynakių, b juodaakių, c rudaakių ir d žaliaakių. Išrenkame keturis atstovus. Kokia tikimybė, kad bent dviejų akys tos pačios spalvos?

Sprendimas. Ieškomą įvykį pažymėkime A. Matome, kad jis apima labai daug elementariųjų įvykių ((visi mėlynakiai), (2 mėlynakiai, 2 juodaakiai) ir pan.). Visus juos suskaičiuoti gana sudėtinga. Pažiūrėkime, kas sudaro priešingą įvykį $\overline{A}$. Nesunku įsitikinti, kad A = {visų akys skirtingų spalvų} = {1 mėlynakis, 1 juodaakis, 1 rudaakis, 1 žaliaakis}. Modifikavę 3 uždavinio iš 3.2 skyrelio sprendimą, gauname

$$P(\overline{A}) = \frac{\binom{a}{1}\binom{b}{1}\binom{c}{1}\binom{d}{1}}{\binom{a+b+c+d}{4}} = \frac{abcd}{\binom{a+b+c+d}{4}}.$$

Atsakymas:

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{abcd}{\binom{a+b+c+d}{4}}.$$

5. Sąlyginė tikimybė

Dažnai galimybė įvykti vienam įvykiui priklauso nuo to, ar įvyksta kitas įvykis. Tarkime, norime rasti įvykio A tikimybę, žinodami, kad įvyko įvykis B. Tokia tikimybė vadinama įvykio *A sąlygine* tikimybe ir žymima $P(A|B)$ (skaitoma „tikimybė, kad įvyks A su sąlyga, kad įvyko B", arba „įvyks A, jeigu įvyko B"). Jeigu B nėra negalimas įvykis ($P(B) > 0$), tai sąlyginę tikimybę galima apibrėžti besąlyginėmis tikimybėmis.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.6}$$

2.9 pavyzdys. Tarkime, kad parinkta 100 žmonių grupė. Gauti duomenys pateikti 2.3 lentelėje.

2.3 lentelė.

Šeiminė padėtis	Lytis		Iš viso
	Vyr.	Mot.	
Vedęs (ištekėjusi)	30	40	70
Nevedęs (netekėjusi)	10	20	30
Iš viso	40	60	100

Tegul $A = \{$apklaustas vedęs respondentas$\}$, $B = \{$apklaustas vyras$\}$. Tuomet $\overline{A} = \{$apklaustas nevedęs respondentas$\}$, $\overline{B} = \{$apklausta moteris$\}$.

Tarkime, norime rasti tikimybę, kad apklaustasis yra vedęs, jeigu žinoma, kad jis vyras, t. y. $P(A|B)$. Pasinaudoję 2.3 lentele, randame

$$P(A \cap B) = P(\text{vedęs ir vyras}) = 30/100 = 0{,}3,$$

$$P(B) = P(\text{vyras}) = 40/100 = 0{,}4.$$

Todėl ieškomoji tikimybė

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}3}{0{,}4} = 0{,}75.$$

Suformuluosime keletą sąlyginės tikimybės savybių. Tegul A, B yra bet kokie atsitiktiniai įvykiai, o $P(D) > 0$. Tuomet:

1) $P(\Omega|D) = 1$.

2) Jeigu $A \cap B = \varnothing$, tai $P(A \cup B|D) = P(A|D) + P(B|D)$.

Pirmoji savybė išplaukia tiesiog iš (2.6) formulės ir to, kad bet koks įvykis yra Ω dalis. Antroji savybė išplaukia iš (2.6) ir įvykių sąjungos bei sankirtos savybių (žr. 1 skyrių).

Sąlyginės tikimybės (2.6) formulę galima užrašyti ir taip:

$$P(A \cap B) = P(A|B)P(B). \tag{2.7}$$

Kadangi $A \cap B = B \cap A$, tai analogiškai gauname tokią formulę:

$$P(A \cap B) = P(A)P(B|A). \tag{2.8}$$

Gautos (2.7) ir (2.8) formulės galioja ir tuomet, kai A arba B yra negalimas įvykis (tada abi pusės lygios nuliui). Formulė (2.8) dar vadinama *tikimybių daugybos teorema*. Ja remiantis, tikimybę, kad įvyks du įvykiai, galima išskaidyti į dvi dalis – tikimybę, kad įvyks vienas įvykis, ir tikimybę, kad įvyks antrasis, jeigu pirmasis jau įvyko.

2.10 pavyzdys. Pagaminta detalė tikrinama du kartus. Tikimybė, kad tikrinant pirmą kartą brokas nebus pastebėtas, yra 0,05, antrąjį – 0,01. Kokia tikimybė, kad bloga detalė nebus išbrokuota?

Atsakymas: $0{,}05 \cdot 0{,}01 = 0{,}0005$.

2.11 pavyzdys. Norėdami pradėti naujas ekologiškos medžioklės tradicijas, dviejų valstybių prezidentai medžioja balionėlį. Pirmasis šauna svečias (jo taiklumas 80%), antrasis – šeimininkas (jo taiklumas 70%). Kokia tikimybė, kad balionėlis bus sumedžiotas, jeigu prezidentai turi tik po vieną šovinį?

Sprendimas. Pažymėkime: A = {šaudamas į balionėlį, pataiko svečias}, B = {šaudamas į balionėlį, pataiko šeimininkas}. Esminis sprendimo momentas yra tas, kad svečiui pataikius šeimininkui nebėra į ką šauti. Todėl ieškomoji tikimybė:

$$\begin{aligned} &P(\text{balionėlis bus sumedžiotas}) \\ &= P((\text{pataiko svečias}) \text{ arba } (\text{svečias nepataiko, o pataiko šeimininkas})) \\ &= P(A \cup (\overline{A} \cap B)) = P(A) + P(\overline{A} \cap B) = P(A) + P(\overline{A})P(B|\overline{A}) \\ &= 0{,}7 + (1 - 0{,}7)0{,}8 = 0{,}94. \end{aligned}$$

2.12 pavyzdys. Vienas politologas pasakė, kad 10% politikų yra sąžiningi ir protingi, 10% sąžiningi ir kvaili, 20% nesąžiningi ir kvaili, 60% nesąžiningi ir protingi. Kokia tikimybė, kad atsitiktinai parinktas politikas yra sąžiningas, jeigu žinoma, kad jis laikytinas protingu?

Sprendimas. Pažymime A = {sąžiningas}, B = {protingas}. Tada gauname

$$P(A \cap B) = 0{,}1, \qquad P(A \cap \overline{B}) = 0{,}1, \qquad P(\overline{A} \cap \overline{B}) = 0{,}2, \qquad P(\overline{A} \cap B) = 0{,}6.$$

Ieškomoji tikimybė

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B \cap A) + P(B \cap \overline{A})} = \frac{0{,}1}{0{,}1 + 0{,}6} = 1/7 = 0{,}142857\ldots$$

Vardiklyje pritaikėme 4 skyrelio 5 savybę.

6. Nepriklausomieji įvykiai

Prisiminkime 2.9 pavyzdį. Kaip matyti, $P(A|B) = 0{,}75$, $P(A) = 0{,}7$. Tikimybė, kad apklaustas respondentas yra vedęs, priklauso nuo to, ar žinoma respondento lytis. Natūralu tokius įvykius vadinti priklausomais. Bet kokie įvykiai A ir B vadinami *priklausomaisiais*, jeigu $P(A|B) \neq P(A)$. Analogiškai apibrėžiamas įvykių nepriklausomumas.

Įvykiai A ir B *nepriklausomi*, jeigu $P(A|B) = P(A)$. (2.9)

Atsižvelgę į (2.8), gauname kitą įvykių priklausomumo apibrėžimą.

Įvykiai A ir B *nepriklausomi*, jeigu $P(A \cap B) = P(A)P(B)$. (2.10)

Įvykiai A ir B *priklausomi*, jeigu $P(A \cap B) \neq P(A)P(B)$. (2.11)

Nesunku įsitikinti, kad bet koks įvykis ir būtinasis įvykis Ω arba negalimas įvykis $\varnothing$ yra nepriklausomi.

Įvykių nepriklausomumas – tai viena iš fundamentaliųjų tikimybių teorijos sąvokų. Nereikia jos painioti su nesutaikomumu. Nepriklausomi įvykiai nebūtinai nesutaikomi.

Dažnai įvykių nepriklausomumas pastebimas intuityviai. Tačiau intuicija gali ir suklaidinti.

2.13 pavyzdys. Metame kauliuką. Įvykis $A = \{$atsivertė lyginis skaičius akučių$\}$, įvykis $B = \{$atsivertė ne mažiau kaip 4 akutės$\}$. Ar A ir B nepriklausomi?

Sprendimas. Patikriname (2.10) sąlygą:

$$A = \{2, 4, 6\}, \qquad B = \{4, 5, 6\}, \qquad A \cap B = \{4, 6\},$$

$$P(A) = \frac{3}{6} = \frac{1}{2}, \qquad P(B) = \frac{3}{6} = \frac{1}{2}, \qquad P(A \cap B) = \frac{2}{6} = \frac{1}{3},$$

$$P(A)P(B) = \frac{1}{2} \cdot \frac{1}{2} \neq \frac{1}{3} = P(A \cap B).$$

Taigi A ir B priklausomi įvykiai.

2.14 pavyzdys. Ankstesnio pavyzdžio sąlygoje truputį pakeisime įvykį B. Tegul A lieka toks pat, o $B = \{$atsivertė daugiau negu 4 akutės$\}$. Tuomet

$$A = \{2, 4, 6\}, \qquad B = \{5, 6\}, \qquad A \cap B = \{6\},$$

$$P(A) = \frac{3}{6} = \frac{1}{2}, \qquad P(B) = \frac{2}{6} = \frac{1}{3}, \qquad P(A \cap B) = \frac{1}{6},$$

$$P(A)P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} = P(A \cap B).$$

Taigi įvykiai A ir B nepriklausomi.

Vienas profesorius ilgą laiką neskraidė lėktuvais. Jis sakydavo: „Tikimybė, kad lėktuve bus bomba, yra 0,000001. Man tai per daug rizikinga." Po kiek laiko kolegos pastebėjo, kad profesorius į konferenciją atskrido lėktuvu. Paklaustas, kodėl pakeitė savo įpročius, profesorius paaiškino: „Tikimybė, kad lėktuve bus dvi bombos, yra 0,000001 · 0,000001. Tokia rizika man jau priimtina. Todėl dabar visur vežiojuosi savo bombą."

Įvykių nepriklausomumo sąvoką galima apibendrinti ir esant didesniam įvykių skaičiui. Įvykiai A, B ir C vadinami nepriklausomaisiais, jeigu

$$P(A \cap B) = P(A)P(B), \qquad P(A \cap C) = P(A)P(C), \tag{2.12}$$

$$P(B \cap C) = P(B)P(C), \tag{2.13}$$

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Šie reikalavimai visai natūralūs. Parodysime, kad jų sumažinti negalima, nes iš (2.12) neišplaukia (2.13), o iš (2.13) neišplaukia (2.12).

2.15 pavyzdys. Dėžėje yra keturi rutuliai – baltas, juodas, raudonas ir geltonas, t.y. $\Omega = \{b, j, r, g\}$. Ištraukiame rutulį ir pažiūrime, kokia jo spalva. Tegul $A = \{b, j\}$, $B = \{b, g\}$, $C = \{b, r\}$. Tuomet

$$P(A) = P(B) = P(C) = 2/4 = 1/2, \qquad P(A \cap B \cap C) = P(\{b\}) = 1/4,$$

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = P(\{b\}) = 1/4.$$

Matome, kad tenkinama (2.12) sąlyga, bet netenkinama (2.13) sąlyga.

2.16 pavyzdys. Dėžėje yra 24 rutuliai, ant kurių užrašyti skaičiai nuo 1 iki 24, t.y. $\Omega = \{1, 2, \dots, 24\}$. Atsitiktinai ištraukiame vieną rutulį. Tegul $A = \{1, 2, \dots, 12\}$, $B = \{1, 13, 14, \dots, 19\}$, $C = \{1, 20, 21, 22, 23, 24\}$. Tada

$$P(A) = 12/24 = 1/2, \qquad P(B) = 8/24 = 1/3, \qquad P(C) = 6/24 = 1/4,$$

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = P(\{1\}) = 1/24,$$

$$P(A \cap B \cap C) = 1/24 = P(A)P(B)P(C).$$

Matome, kad (2.13) sąlyga tenkinama, bet (2.12) – ne. Tęsdami įvykių nepriklausomumo apibrėžimą keturiems ir daugiau įvykių, turėsime formuluoti vis daugiau sąlygų – turi būti nepriklausomos visos įvykių poros, visi trejetai it t.t.

7. Pilnosios tikimybės formulė

Tarkime, kad gamykla 40% gaminių pagamina per pirmą pusmetį, 60% – per antrąjį. Pirmą pusmetį brokuotų gaminių būna 2%, antrąjį – 3%. Atsitiktinai pasirenkame vieną gamyklos gaminį. Kokia tikimybė, kad jis brokuotas?

Atsakymas aiškus, jeigu žinoma, kurį pusmetį gaminys buvo gamintas. Tačiau kaip išspręsti šį uždavinį to nežinant? Atsakymą padės rasti *pilnosios tikimybės formulė*.

Pilnosios tikimybės formulė

Tegul:

1) $H_1 \cup H_2 \cup \ldots = \Omega$,

2) $H_i \cap H_j = \varnothing, \quad i \neq j$.

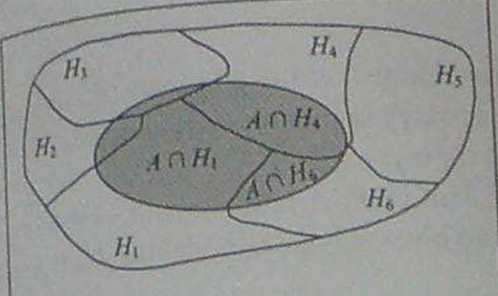
Tuomet

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

Pilnosios tikimybės formulė teigia, kad užtenka žinoti įvykio A tikimybę, esant sąlygoms $H_1, H_2, \dots$, ir tų sąlygų susidarymo tikimybes. Abu reikalavimai labai svarbūs: sąlygos $H_1, H_2, \dots$ turi apimti visas įmanomas situacijas, be to, jos visos turi būti poromis nesutaikomos.

Įrodysime pilnosios tikimybės formulę. Pasinaudoję 1.2 skyrelio formulėmis ir 1) reikalavimu, gauname

$$A = A \cap \Omega = A \cap (H_1 \cup H_2 \cup \ldots) = (A \cap H_1) \cup (A \cap H_2) \cup (A \cap H_3) \cup \ldots \tag{2.14}$$



2.3 pav. Pilnosios tikimybės Veno diagrama

Kadangi pagal 2) reikalavimą $H_1, H_2, \ldots$ tarpusavyje nesutaikomi, tai nesutaikomos ir jų dalys $(A \cap H_1), (A \cap H_2), \ldots$ Iš tikrųjų

$$(A \cap H_1) \cap (A \cap H_2) = A \cap (H_1 \cap H_2) = A \cap \varnothing = \varnothing$$

ir pan. Todėl pagal bendrojo tikimybės apibrėžimo 3) sąlygą

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + P(A \cap H_3) + \cdots. \tag{2.15}$$

Bet iš tikimybių daugybos teoremos (žr. (2.8)) išplaukia, kad

$$P(A \cap H_1) = P(H_1)P(A|H_1), \qquad P(A \cap H_2) = P(H_2)P(A|H_2), \ldots$$

Įstatę šias lygybes į (2.15), gauname pilnosios tikimybės formulę.

Geometrinė pilnosios tikimybės formulės (Veno diagramos) interpretacija pavaizduota 2.3 paveiksle. Iš tiesų plotas A susideda iš tos A dalies, kuri priklauso H_1, iš A dalies, kuri priklauso H_2, ir pan., t. y. $A = (A \cap H_1) \cup (A \cap H_2) \cup \ldots$

2.17 pavyzdys. Išspręsime skyrelio pradžioje suformuluotą uždavinį. Kadangi viskas priklauso nuo gamybos laiko, tai atitinkamai ir parinksime sąlygas H_1, H_2:

$$H_1 = \{\text{gaminta pirmąjį pusmetį}\}, \qquad P(H_1) = 0{,}40$$
$$H_2 = \{\text{gaminta antrąjį pusmetį}\}, \qquad P(H_2) = 0{,}60.$$

Tegul $A = \{\text{gaminys brokuotas}\}$. Tuomet uždavinio sąlygos teigia, kad

$$P(A|H_1) = 0{,}02, \qquad P(A|H_2) = 0{,}03.$$

Pritaikę pilnosios tikimybės formulę, gauname

$$P(A) = P(H_1)P(A|H_1) + P(\overset{H_2}{\cancel{A}})P(A|H_2) = 0{,}4 \cdot 0{,}02 + 0{,}6 \cdot 0{,}03 = 0{,}026.$$

2.18 pavyzdys. Žvejys turi tris pamėgtas žūklės vietas. Žvejo katinas ėda žuvį po 80% žvejo apsilankymų pirmoje žūklės vietoje, po 70% apsilankymų antrojoje ir po 75% apsilankymų trečiojoje vietoje. Į kurią žūklės vietą patraukti, žvejys sprendžia mesdamas kauliuką. Jei atsiverčia 6 akutės, jis eina į pirmą vietą, jei 5 – į antrąją, kitais atvejais eina į trečiąją vietą. Kokia tikimybė, kad po ateinančios žūklės žvejo katinas gaus žuvies?

Sprendimas. Ieškomą įvykį pažymime simboliu A. Tegul

$$H_1 = \{\text{žvejos 1-oje vietoje}\}, \quad H_2 = \{\text{žvejos 2-oje vietoje}\}, \quad H_3 = \{\text{žvejos 3-oje vietoje}\}.$$

Tuomet

$$P(A|H_1) = 0{,}80, \qquad P(A|H_2) = 0{,}70, \qquad P(A|H_3) = 0{,}75,$$
$$P(H_1) = P(\{\text{atsivers 6 akutės}\}) = 1/6, \qquad P(H_2) = 1/6, \qquad P(H_3) = 4/6.$$

Atsakymas:

$$P(A) = 0{,}8 \cdot \frac{1}{6} + 0{,}7 \cdot \frac{1}{6} + 0{,}75 \cdot \frac{4}{6} = 0{,}75.$$

8. Bajeso formulė

Bajeso[1] formulė galioja, esant toms pačioms prielaidoms kaip ir pilnosios tikimybės formulė. Kaip ir anksčiau, $H_1, H_2, \ldots$ žymi galimas sąlygas įvykiui A įvykti. Mes žinome pradines įvykių $H_1, H_2, \ldots$ įvykimo (*apriorines*) tikimybes $P(H_1)$, $P(H_2)$, … – t. y. tikimybes, kad sąlygos $H_1, H_2, \ldots$ galioja *prieš* vykdant įvykį A realizuojantį eksperimentą. Tarkime, kad įvykis A įvyko. Kokia tikimybė, kad buvo sąlygų komplektas H_j? Atsakymą duoda Bajeso formulė.

Bajeso formulė

Tegul:

1) $H_1 \cup H_2 \cup \ldots = \Omega$,
2) $H_i \cap H_j = \varnothing, \quad i \neq j$.

Tuomet

$$P(H_j|A) = \frac{P(H_j)P(A|H_j)}{P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots}.$$

Tikimybės $P(H_1|A)$, $P(H_2|A)$, … vadinamos *aposteriorinėmis* tikimybėmis. Jos skiriasi nuo apriorinių tikimybių tuo, kad atsižvelgta į naują informaciją (A įvyko).

Bajeso formulės *įrodymas* nesudėtingas. Iš sąlyginės tikimybės apibrėžimo ir tikimybių daugybos teoremos išplaukia

$$P(H_j|A) = \frac{P(H_j \cap A)}{P(A)} = \frac{P(H_j)P(A|H_j)}{P(A)}.$$

Belieka vardiklyje užrašyti pilnosios tikimybės formulę.

2.19 pavyzdys. Krepšinio komanda „Kablys" 55% rungtynių žaidžia išvykoje, 45% – namuose. Žaisdama išvykoje, komanda laimi 60% rungtynių, žaisdama namuose – 80% rungtynių. Komanda laimėjo rungtynes. Kokia tikimybė, kad žaidė namuose?

Sprendimas. Natūraliai išsiskiria dvi komandos žaidimo sąlygos – išvyka ir namai. Pažymime

$$H_1 = \{\text{žaidė išvykoje}\}, \qquad H_2 = \{\text{žaidė namuose}\}.$$

Tegul $A = \{\text{komanda laimėjo}\}$. Tuomet

$$P(H_1) = 0{,}55, \qquad P(H_2) = 0{,}45, \qquad P(A|H_1) = 0{,}60, \qquad P(A|H_2) = 0{,}80.$$

Mus domina $P(H_2|A)$. Pritaikę Bajeso formulę, gauname

$$P(H_2|A) = \frac{0{,}80 \cdot 0{,}45}{0{,}55 \cdot 0{,}60 + 0{,}45 \cdot 0{,}80} = 0{,}54545454\ldots$$

2.20 pavyzdys. Grupėje yra 10 studentų, mokančių taikyti statistinius metodus, ir 15 – nemokančių (bet vis tiek taikančių). Sociologiniam tyrimui atlikti sudaroma dviejų studentų grupė. Jeigu abu studentai

---

[1] Thomas Bayes (1702–1761) – anglų matematikas.

mokės taikyti statistinius metodus, tyrimas bus atliktas sėkmingai; jeigu abu nemokės, tyrimas žlugs; jeigu vienas mokės, o kitas ne – galimybės, kad jis bus sėkmingas ar nesėkmingas, vienodos. Studentų tyrimas buvo sėkmingas. Kokia tikimybė, kad grupę sudarė vienas išmanantis statistikos taikymą studentas, o kitas ne?

Sprendimas. Pažymime $A = \{\text{tyrimas sėkmingas}\}$, $H_1 = \{\text{abu moka taikyti}\}$, $H_2 = \{\text{vienas moka, kitas ne}\}$, $H_3 = \{\text{abu nemoka}\}$. Tuomet

$$P(A|H_1) = 1, \qquad P(A|H_2) = 0{,}5, \qquad P(A|H_3) = 0.$$

Prisiminę trečio skyriaus formules, užrašome

$$P(H_1) = \frac{\binom{10}{2}}{\binom{25}{2}} = 0{,}15, \qquad P(H_2) = \frac{\binom{10}{1}\binom{15}{1}}{\binom{25}{2}} = 0{,}5.$$

Todėl

$$P(H_2|A) = \frac{0{,}5 \cdot 0{,}5}{0{,}15 + 0{,}25 + 0} = 0{,}625.$$

9. Bernulio schema ir jos apibendrinimas

Bernulio[1] eksperimentų schema nusakoma taip: eksperimentą atlikus vieną kartą, jo sėkmės tikimybė lygi p. Atliekame n nepriklausomų eksperimentų. Kokia tikimybė, kad eksperimentas pavyks k kartų? Atsakymas į šį klausimą toks:

$$P\,(\text{iš } n \text{ bandymų } k \text{ sėkmingų}) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.16}$$

Įrodysime šią formulę. Pasinaudoję priešingo įvykio tikimybe, gauname, kad bandymą atliekant vieną kartą, jis nepavyks su tikimybe $1 - p$. Sėkmingą bandymą pažymėkime raide S, o nesėkmingą N ($P(S) = p$, $P(N) = 1 - p$). Ieškomoji tikimybė

$$\begin{aligned}
P(SSS\ldots SNNN\ldots N &\cup SSS\ldots SNSNNN\ldots \cup \ldots) \\
&= P(SSS\ldots SNNN\ldots N) + P(SSS\ldots SNSNNN\ldots) + \cdots \\
&= P(S)P(S)\ldots P(S)P(N)\ldots P(N) \\
&= P(S)P(S)\ldots P(S)P(N)\ldots P(N) + \cdots \\
&\quad + P(S)P(S)\ldots P(S)P(N)P(S)P(N)\ldots P(N) + \cdots + p^k(1-p)^{n-k}. \\
&= p^k(1-p)^{n-k} + p^k(1-p)^{n-k} + \cdots + p^k(1-p)^{n-k}.
\end{aligned}$$

Kiekvieno palankaus įvykio tikimybė ta pati ir lygi $p^k(1-p)^{n-k}$. Kiek yra tokių įvykių? Jų yra tiek, kiek ir būdų iš n eksperimentų gauti k sėkmingų, t. y. $\binom{n}{k}$ (žr. 3 skyrių). Taigi (2.16) įrodėme.

2.21 pavyzdys. Studentas gavo 10 klausimų klausimyną. Atsakymą į kiekvieną klausimą reikia parinkti iš 4 galimų variantų, iš kurių tik vienas teisingas. Testas parašytas runomis pietinių zulusų dialektu. Studentas atsakymą renkasi atsitiktinai. Kokia tikimybė, kad jis teisingai atsakys į 3 klausimus?

Sprendimas. Studentas 10 kartų kartoja bandymą – atsitiktinai renka atsakymą iš 4 galimų variantų. Vieno bandymo sėkmės tikimybė $1/4 = 0{,}25$. Ieškomoji tikimybė:

$$\binom{10}{3}(0{,}25)^3(1-0{,}25)^7 = 0{,}25028\ldots$$

n=10
k=3 p=0,25

[1] Jakob Bernoulli (1654–1705) – šveicarų matematikas.

2.22 pavyzdys. Krepšininkas pataiko 80% baudos metimų. Kokia tikimybė, kad jis pataikys bent vieną metimą iš penkiolikos?

Sprendimas. Vieno metimo pataikymo tikimybė 0,8. Ieškomoji tikimybė

$$P(\text{bent vieną}) = P(\text{vieną}) + P(\text{du}) + P(\text{tris}) + \cdots + P(\text{penkiolika})$$
$$= \binom{15}{1}(0,8)(0,2)^{14} + \binom{15}{2}(0,8)^2(0,2)^{13} + \cdots + \binom{15}{15}(0,8)^{15}(0,2)^0.$$

Tačiau toks sprendimo metodas neracionalus. Paprasčiau būtų pereiti prie priešingo įvykio.

$$P(\text{bent vieną}) = 1 - P(\text{nė vieno}) = 1 - \binom{15}{0}(0,8)^0(0,2)^{15} = 1 - (0,2)^{15} = 0,999\ldots$$

Bernulio schemoje vienas bandymas galėjo tik pavykti arba nepavykti. Tokią schemą galima apibendrinti ir esant k skirtingų baigčių. Tarkime, kad vieną kartą darant bandymą baigčių tikimybės yra $p_1, p_2, \ldots, p_k$ (žinoma $p_1 + p_2 + \cdots + p_k = 1$). Tuomet tikimybė, kad po n bandymų bus m_1 pirmųjų baigčių, m_2 antrųjų baigčių, ..., m_k – k-ųjų baigčių ($m_1 + \cdots + m_k = n$), yra lygi

$$P(m_1, m_2, \ldots, m_k) = \frac{n!}{m_1!m_2!\ldots m_k!} p^{m_1} p^{m_2} \ldots p^{m_k}. \tag{2.17}$$

2.23 pavyzdys. Iš penkių fizikos, trijų filologijos ir septynių ekonomikos studentų kartais atsitiktinai parenkamas atstovas bendrauti su rektoriaus svečiais. Kokia tikimybė, kad su paskutiniais keturiais rektoriaus svečiais bendravo 1 fizikas, 1 filologas ir 2 ekonomistai?

Sprendimas. Vieną kartą renkant, tikimybės išrinkti fiziką, filologą arba ekonomistą lygios:

$$p_1 = 5/(5+3+7) = 1/3, \qquad p_2 = 1/5, \qquad p_3 = 7/15.$$

Be to, $m_1 = 1$, $m_2 = 1$, $m_3 = 2$. *Atsakymas:*

$$P(1,1,2) = \frac{4!}{1!1!2!} \cdot \frac{1}{3} \cdot \frac{1}{5} \cdot \left(\frac{7}{15}\right)^2 = 0,1742222\ldots$$

10. „Geometrinė“ tikimybė

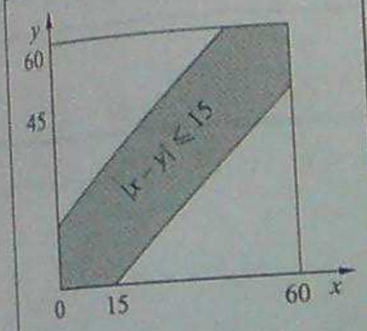
Trumpai susipažinsime su geometrinės tikimybės sąvoka. Turime geometrinę figūrą Ω. Geometrinė figūra A yra Ω poaibis. Galimybės pasirinkti bet kurį Ω tašką yra vienodos. Kokia tikimybė, kad atsitiktinai parinktas taškas priklausys figūrai A? Ši tikimybė nusakoma figūrų plotų santykiu:

$$P(\text{taškas priklauso } A) = \frac{\text{plotas } A}{\text{plotas } \Omega}.$$

Matome, kad svarbus tik figūrų plotas, bet ne jų forma. Analogiškai „geometrinė“ tikimybė apibrėžiama trimatėje erdvėje (gausime tūrių santykį). Dažnai „geometrinę“ tikimybę galima panaudoti sprendžiant uždavinius, kurie iš pirmo žvilgsnio nieko bendro su geometrija neturi.

2.24 pavyzdys. Kasparas ir Audronė susitarė susitikti tarp 12 ir 13 val. prie F. Zapos biusto. Pirmasis atėjęs laukia 15 minučių arba (jei atėjo prieš pat 13 val.) iki 13 valandos. Kokia tikimybė, kad Kasparas ir Audronė susitiks?

Sprendimas. Tariame, kad ir Kasparo ir Audronės galimybės atvykti bet kuriuo laiku tarp 12 ir 13 valandos vienodos. Kasparo atėjimo laiką pažymėkime x, o Audronės – y (dėl paprastumo tarkime, kad tai minutės, praėjusios po 12 val.). Taigi visi įmanomi Kasparo ir Audronės atėjimo laikai aprašomi pora (x, y), čia x ir y įgyja visas galimas reikšmes nuo 0 iki 60. Dekarto koordinačių sistemoje tai atitiktų kvadratą (žr. 2.4 pav.). Kasparas susitiks su Audrone, jeigu $|x - y| \leqslant 15$. Pastarąją sąlygą tenkina užbrūkšniuota 2.4 paveikslo dalis. Ieškomoji tikimybė yra užbrūkšniuotos dalies ploto santykis su viso kvadrato plotu. Atlikę nesudėtingą geometrinį skaičiavimą, gauname, kad ieškomoji tikimybė yra 0,4375.



2.4 pav. Atėjimo koordinatės

11. Atsitiktiniai dydžiai

Apibrėždami atsitiktinius įvykius, kalbėjome apie eksperimentus, turinčius keletą baigčių. Praktiškai beveik visada susiduriame su skaitiniais stebimojo dydžio matavimais, t. y. su kokio nors atsitiktinio dydžio reikšmėmis.

Atsitiktinis dydis – tai funkcija $X\colon \Omega \to \mathbb{R}$.

Taigi atsitiktinis dydis nusako taisyklę, pagal kurią kiekvienam atsitiktiniam įvykiui priskiriama skaitinė reikšmė. Eksperimentų ir atsitiktinių dydžių pavyzdžiai pateikti 2.4 lentelėje. Atsitiktinis dydis – tai tam tikras matematinis modelis, o tas pats matematinis modelis gali tikti daugeliui situacijų (eksperimentų). Vėliau matysime, kad statistiniams tyrimams naudojama palyginti nedaug skirtingų atsitiktinių dydžių tipų. Taigi nors atsitiktiniai dydžiai ir labiau matematizuoti už atsitiktinius įvykius, juos išnagrinėjus skirtingiems eksperimentams nebereikės kurti atskirų teorijų. Be to, atsitiktinio dydžio reikšmės yra skaičiai, o su skaičiais (skirtingai negu su įvykiais) galimos aritmetinės operacijos.

Iš pateiktos 2.4 lentelės matyti, kad X konkrečią reikšmę įgyja su tokia tikimybe, su kokia įvyksta atitinkamas atsitiktinis įvykis. Tačiau iškart galima apibrėžti atsitiktinio dydžio įgyjamas reikšmes ir jų įgijimo tikimybes (vadinamąjį skirstinį), nebandant atsitiktinio dydžio sieti su kokiu nors eksperimentu. Atsitiktinio dydžio *skirstinys* – tai atsitiktinio dydžio įgyjamos reikšmės ir jų įgijimo tikimybės.

Pavyzdžiui, herbų skaičiaus skirstinys nusakomas taip: atsitiktinis dydis X reikšmę 0 įgyja su tikimybe 0,5 ir reikšmę 1 – su tikimybe 0,5.

Atsitiktinio dydžio skirstinį galima nusakyti ir specialia – *pasiskirstymo funkcija*.

2.4 lentelė. Eksperimentai ir atsitiktiniai dydžiai

Eksperimentas	Įvykis	Atsitiktinis dydis X	Galimos X reikšmės
Metama moneta	Skaičius, herbas	Herbų skaičius	0, 1
Metamas kauliukas	1, 2, 3, 4, 5, 6 akutės	Akučių skaičius	1, 2, 3, 4, 5, 6
100 gaminių kontrolė	Visi geri, 1 blogas ir pan.	Blogų gaminių skaičius	0, 1, 2, ..., 100
Matuojamas pieštukas	< 20 cm, > 15 cm ir pan.	Ilgis cm	[13, 20]
Krepšininkas meta baudą	Pataiko, nepataiko	Pataikymų skaičius	0, 1
Sveriamas naujagimis	> 3 kg, < 5 kg ir pan.	Svoris kg	[2, 5]

Atsitiktinio dydžio X *pasiskirstymo funkcija* $F(x) = P(X \leqslant x)$.

Pasiskirstymo funkcija yra apibrėžta visoje skaičių tiesėje. Iš tikimybių savybių išplaukia, kad:

1) $0 \leqslant F(x) \leqslant 1$.
2) $\lim\limits_{x\to-\infty} F(x) = 0$, $\lim\limits_{x\to\infty} F(x) = 1$.
3) $F(x)$ nemažėjanti, t. y. $F(x_1) \leqslant F(x_2)$, kai $x_1 < x_2$.
4) $F(x)$ tolydi iš dešinės.
5) $P(a < X \leqslant b) = F(b) - F(a)$.

Pasiskirstymo funkcijų pavyzdžiai pateikti 2.5 paveiksle.

Iš pasiskirstymo funkcijos savybių aišku, kad atsitiktinio dydžio skirstinį visiškai nusako jo pasiskirstymo funkcija. Tačiau yra ir kitų skirstinio apibrėžimo būdų. Su jais susipažinsime 12 skyriuje. Atsitiktiniai dydžiai gali būti nepriklausomi.

Atsitiktiniai dydžiai X ir Y yra *nepriklausomi*, jeigu bet kokiems realių skaičių aibės poaibiams B_1 ir B_2 $P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$.

Trys ir daugiau nepriklausomų atsitiktinių dydžių apibrėžiami analogiškai, remiantis (2.12) ir (2.13) formulėmis. Į kiekvieną skaičių (konstantą) C galima žiūrėti kaip į tam tikrą išsigimusį atsitiktinį dydį. Pažymėtina, kad *bet koks atsitiktinis dydis X ir bet kokia konstanta C yra nepriklausomi*.

Galima apibrėžti ir daugiamačius atsitiktinius dydžius (vektorius). Pavyzdžiui, matuodami žmogaus ūgį ir svorį, gausime atsitiktinį vektorių (X, Y). Tuomet galime kalbėti

2.5 pav. Pasiskirstymo funkcijų pavyzdžiai apie jungtines tikimybes $P(X = a, Y = b)$. Be to, jeigu X ir Y nepriklausomi, tai $P(X = a, Y = b) = P(X = a)P(Y = b)$.

Jeigu X yra atsitiktinis dydis, tai atsitiktinis dydis bus ir jo tolydžioji funkcija. Pavyzdžiui, X^4, $3X + 15$, $\ln(|X| + 1)$. Pažymėtina, kad jeigu X ir Y yra nepriklausomi, tai nepriklausomi dydžiai yra ir jų tolydžiosios funkcijos. Pavyzdžiui, jeigu X ir Y nepriklausomi, tai nepriklausomi ir X^2 bei $2Y$.

12. Diskretieji ir tolydieji atsitiktiniai dydžiai

Šiame vadovėlyje aptarsime dviejų tipų atsitiktinius dydžius: a) diskrečiuosius, b) absoliučiai tolydžiuosius.

12.1. *Diskretieji atsitiktiniai dydžiai*

Metant kauliuką, traukiant kortą, skaičiuojant restorano lankytojus, susiduriama su baigtine galimų reikšmių aibe. Matematiniams modeliams kartais naudinga apibrėžti ir begalines reikšmių aibes. Skaiti reikšmių aibė – tai begalinė aibė, kurios elementus galima sunumeruoti (pvz., visų natūraliųjų skaičių aibė).

Atsitiktinis dydis X, įgyjantis baigtinę arba skaičią reikšmių aibę, vadinamas *diskrečiuoju*.

Be jau minėtųjų, diskretieji dydžiai yra: 1) per savaitę parduotų skutimosi peiliukų skaičius, 2) politinių skandalų per metus skaičius, 3) razinų bandelėje skaičius, 4) klientų kirpykloje sekmadienį skaičius ir pan. Diskrečiojo atsitiktinio dydžio pasiskirstymo funkcija trūki, jos grafiko pavyzdys pateikas 2.5 paveiksle, b).

Diskretusis atsitiktinis dydis X įgyja reikšmes $x_1, x_2, \dots$ su tikimybėmis $p_1, p_2, \dots$ Jo skirstinį patogiausia aprašyti lentele:

X	x_1	x_2	x_3	$\dots$
P	p_1	p_2	p_3	$\dots$

Žinoma, $p_1 + p_2 + \cdots = 1$, $p_1 \geqslant 0$, $p_2 \geqslant 0, \dots$

Tikimybė, kad atsitiktinis dydis įgis reikšmes iš aibės B, skaičiuojama taip: randami tie x, kurie patenka į B, ir jų tikimybės sudedamos:

$$P(X \in B) = \sum_{j:\, x_j \in B} p_j. \tag{2.18}$$

2.25 pavyzdys. Metame kauliuką. X yra atsivertusių akučių skaičius. Tuomet X skirstinys yra toks:

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

$P(X = 1) = 1/6$, $P(2 < X \leqslant 4) = P(X = 3) + P(X = 4) = 1/6 + 1/6 = 1/3$, $P(X > 6) = 0$,

$P(X \in [0, 2)) = P(X = 1) = 1/6$.

2.26 pavyzdys. Tarkime, kad atsitiktinio dydžio X skirstinys yra toks:

X	-7	0	16
P	0,3	0,2	0,5

Tuomet $P(X < 0) = P(X = -7) = 0{,}3$ ir pan.

Šiame pavyzdyje nenurodėme, koks bandymas realizuoja atsitiktinį dydį. Jau minėjome, kad tą patį skirstinį gali turėti ir daugiau atsitiktinių dydžių. Maža to, turint skirstinį, dažniausiai nesunku sugalvoti jį realizuojantį eksperimentą. Pavyzdžiui, tarkime, kad dėžėje yra 10 sunumeruotų rutulių. Lošimo taisyklės tokios: jei ištraukto rutulio numeris ne didesnis už 3, lošėjas sumoka kazino 7 Lt; jeigu numeris lygus 4, nei lošėjas, nei kazino nieko nemoka; kitais atvejais kazino sumoka lošėjui 16 Lt. Atsitiktinis dydis X – lošėjo pajamos iš vieno lošimo. Akivaizdu, kad X turi 2.26 pavyzdyje nurodytą skirstinį. Atkreipiame dėmesį, kad X gali *įgyti neigiamas reikšmes*, bet reikšmių įgijimo *tikimybės* visuomet neneigiamos.

2.27 pavyzdys. Tarkime, kad X skirstinys yra:

X	1	2	3	…
P	$1/2$	$1/2^2$	$1/2^3$	…

Šiuo atveju įgyjamų reikšmių aibė yra begalinė. Teoriškai šį skirstinį realizuoja toks bandymas: mėtome simetrišką monetą tol, kol iškrinta herbas. Atsitiktinis dydis – metimų skaičius.

Žinant X skirstinį, nesunku rasti ir $Y = f(X)$ skirstinį. Tam užtenka: a) X skirstinio lentelėje x_i pakeisti $f(x_i)$ (tikimybės lieka tos pačios!); b) gautojoje lentelėje palikti tik skirtingas $f(x_i)$ reikšmes, sudedant atitinkamas tikimybes.

2.28 pavyzdys. Tegul X turi skirstinį

X	-1	0	1
P	0,2	0,3	0,5

Tarkime, kad $Y = X^2 + 1$. Tuomet pagalbinė lentelė ir atsakymas atitinkamai yra:

2	1	2
0,2	0,3	0,5

Y	1	2
P	0,3	0,2 + 0,3

Pasinaudojome tuo, kad $f(-1) = (-1)^2 + 1 = 2$, $f(0) = 0$, $f(1) = 2$. Nesunku suprasti, kodėl atitinkamos tikimybės sudedamos. Mat Y įgyja reikšmę 2, jeigu X įgyja -1 arba 1.

Panašiai randamas $Z = f(X, Y)$, čia X ir Y *nepriklausomi* atsitiktiniai dydžiai, skirstinys. Pirmu etapu sudaroma pagalbinė lentelė, imant *visas* įmanomas X ir Y reikšmių poras (dydžiai nepriklausomi, todėl atitinkamos tikimybės sudauginamos!). Po to paliekamos tik skirtingos reikšmės, o jų atitinkamos tikimybės sudedamos.

2.29 pavyzdys. Tegul $Z = X^2 + Y$, čia X ir Y skirstiniai:

X	-1	0	1
P	0,4	0,5	0,1

Y	1	2
P	0,8	0,2

① užrašom

Pagalbinė lentelė yra tokia:

② imam visas galimas X ir Y poras

2	1	2	3	2	3
$0{,}4 \cdot 0{,}8$	$0{,}5 \cdot 0{,}8$	$0{,}1 \cdot 0{,}8$	$0{,}4 \cdot 0{,}2$	$0{,}5 \cdot 0{,}2$	$0{,}1 \cdot 0{,}2$

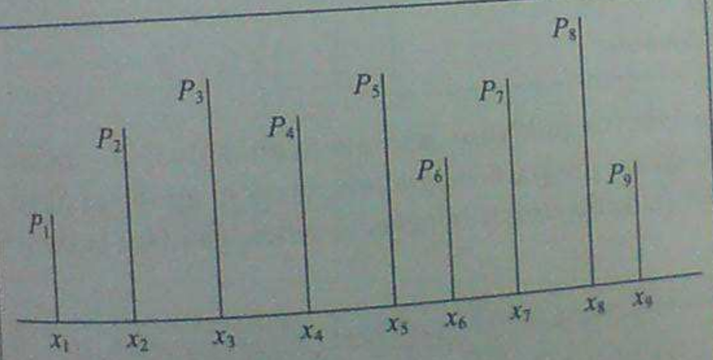
③ tikimybes dauginam

Atsakymas:

Z	1	2	3
P	0,4	0,5	0,1

④ palikom unikalias tik reikšmes, P sudedam

Galima nubraižyti diskrečiojo atsitiktinio dydžio skirstinio daugiakampį arba jo tikimybes atidėti tiesėje (žr. 2.6 pav.).



2.6 pav. Diskrečiojo atsitiktinio dydžio tikimybės

12.2. *Absoliučiai tolydūs atsitiktiniai dydžiai*

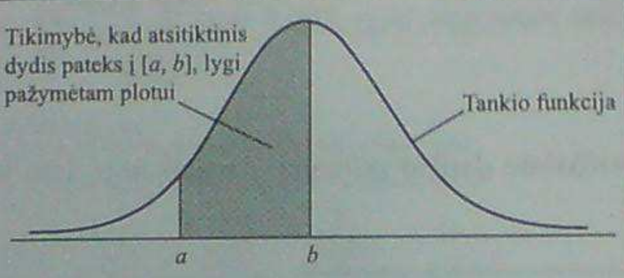
Kalbėdami apie eksperimentus ir jų generuojamus atsitiktinius dydžius, matėme, kad pieštuko ilgis gali įgyti visas reikšmes iš intervalo [13, 20]. Panaši situacija būna ir gaminant detales; matuojant laiką nuo policijos iškvietimo iki jos atvykimo; matuojant naftos gręžinio gylį; sveriant parduotuvėje pirktus produktus. Visais šiais atvejais įgyjamų reikšmių yra be galo daug (jų net negalima sunumeruoti), todėl prasminga kalbėti tik apie reikšmių priklausymo tam tikram intervalui tikimybę. Pavyzdžiui, apie detales, kurių ilgis 2,1–2,2 mm; apie sekundžių intervalą [60, 3600] ir pan. Šiame skyrelyje aptarsime dydžius, kurių patekimo į intervalą tikimybės skaičiuojamos naudojant *tankio* funkcijas – absoliučiai tolydžiuosius atsitiktinius dydžius.

Atsitiktinis dydis X, kurio patekimo į intervalą $[a, b]$ tikimybė skaičiuojama pagal formulę

$$P(a \leqslant X \leqslant b) = \int_a^b p(x)\,\mathrm{d}x, \text{ čia } p(x) \geqslant 0, \tag{2.19}$$

vadinamas *absoliučiai tolydžiuoju* dydžiu. Funkcija $p(x)$ vadinama *tankiu*.

Kaip tankio funkcija naudojama tikimybėms skaičiuoti? Jeigu X tankis yra $p(x)$, tai tikimybė, kad X paklius į intervalą $[a, b]$, yra lygi plotui, kurį apriboja intervalas ir $p(x)$ grafikas. Grafiškai tikimybės skaičiavimas pavaizduotas 2.7 paveiksle. Pastebėsime, kad apibrėžimo formulė išlieka teisinga ir tuo atveju, kai negriežtas nelygybes pakeičiame griežtomis. Taip yra todėl, kad bet kokiam taškui a ir absoliučiai tolydžiam dydžiui X $P(X = a) = 0$.



2.7 pav. Tankio funkcija ir tikimybės

Konkreti tankio funkcijos išraiška priklauso nuo atsitiktinio dydžio. Tačiau visos tankio funkcijos tenkina dvi savybes: jos yra neneigiamos; visas jų apribotas plotas lygus vienetui. Pasirodo, kad bet kuri funkcija, turinti minėtas savybes, gali būti laikoma tankio funkcija.

Tankio funkcijos $p(x)$ savybės:

1) $p(x) \geqslant 0$,

2) $\displaystyle\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$.

Tankio funkcijų pavyzdžiai:

a) $p(x) = \begin{cases} \exp\{-x\}, & \text{kai } x > 0, \\ 0, & \text{kai } x \leqslant 0. \end{cases}$ b) $p(x) = \begin{cases} 1, & \text{kai } 0 \leqslant x \leqslant 1, \\ 0 & \text{kitur.} \end{cases}$

Absoliučiai tolydžiojo atsitiktinio dydžio pasiskirstymo funkcijos pavyzdys pateiktas 2.5 paveiksle, a). Absoliučiai tolydžiojo atsitiktinio dydžio pasiskirstymo funkcija ir tankis yra susiję:

1) $F'(x) = p(x)$,

2) $\displaystyle F(x) = \int_{-\infty}^{x} p(u)\,\mathrm{d}u$.

2.30 pavyzdys. Tarkime, kad X tankis yra $p(x) = \exp\{-x\}$, kai $x > 0$, ir $p(x) = 0$, kai $x \leqslant 0$. Raskime $P(-1 < X \leqslant 7)$. Remdamiesi (2.19) formule, gauname

$$\begin{aligned} P(-1 < X \leqslant 7) &= \int_{-1}^{7} p(x)\,\mathrm{d}x = \int_{-1}^{0} p(x)\,\mathrm{d}x + \int_{0}^{7} p(x)\,\mathrm{d}x \\ &= \int_{-1}^{0} 0\,\mathrm{d}x + \int_{0}^{7} \mathrm{e}^{-x}\,\mathrm{d}x = 0 - \mathrm{e}^{-x}\Big|_0^7 = 1 - \mathrm{e}^{-7} = 0{,}999\ldots \end{aligned} \tag{2.20}$$

Atkreipiame dėmesį, kad čia, atsižvelgdami į $p(x)$ reikšmes, integralą išskaidėme į dvi dalis.

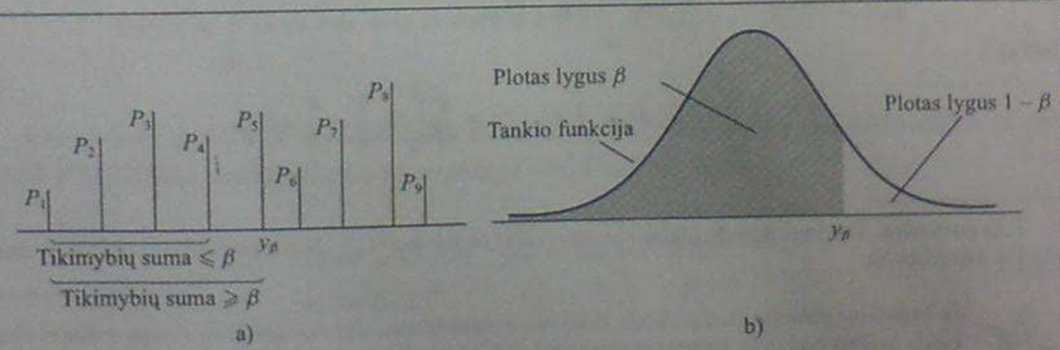
13. Kvantiliai

Tegul $0 < \beta < 1$. Atsitiktinio dydžio β lygmens kvantiliu (β *kvantiliu*) vadinsime skaičių y_β, tenkinantį nelygybes:

$$P(X < y_\beta) \leqslant \beta \leqslant P(X \leqslant y_\beta). \tag{2.21}$$

Diskrečiojo atsitiktinio dydžio kvantilis – tai tokia X reikšmė y_β, kuriai: a) visa tikimybių reikšmių suma į kairę nuo jos yra mažesnė už β; b) prie tų tikimybių reikšmių sumos pridėjus y_β įgijimo tikimybę, ji tampa ne mažesnė už β.

Absoliučiai tolydžiųjų atsitiktinių dydžių kvantilius galima apibrėžti lygybėmis. Tolydaus atsitiktinio dydžio X, kurio tankis $p(x)$, $0 < \beta < 1$, β lygmens *kvantiliu* vadinsime tokį skaičių y_β, su kuriuo $P(X \leqslant y_\beta) = \beta$. Taigi β lygmens kvantilis yra toks skaičius, už kurį ne didesnes reikšmes X įgyja su tikimybe β.



2.8 pav. Diskrečiojo (a) ir tolydžiojo (b) atsitiktinių dydžių kvantiliai

14. Atsitiktinio dydžio vidurkis

Aprašomojoje statistikoje jau nagrinėjome *empirinį* (aritmetinį) vidurkį. Šiame skyrelyje apibrėšime *teorinį*, t. y. atsitiktinio dydžio, vidurkį dviem specialiais – diskrečiųjų dydžių ir absoliučiai tolydžiųjų dydžių – atvejais.

Tarkime, kad X turi skirstinį

X	x_1	x_2	x_3	$\dots$
P	p_1	p_2	p_3	$\dots$

Tuomet X vidurkis $\mathbf{E}X$ yra atsitiktinio dydžio X reikšmių ir jų įgijimo tikimybių sandaugų suma.

$$\mathbf{E}X = x_1p_1 + x_2p_2 + x_3p_3 + \cdots \tag{2.22}$$

Jeigu X tankis yra $p(x)$, tai $\mathbf{E}X$ apibrėžiamas kaip integralas.

$$\mathbf{E}X = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x. \tag{2.23}$$

2.31 pavyzdys. Tegul X yra ištrauktų tūzų skaičius, traukiant vieną kortą iš 24. Rasime X skirstinį ir vidurkį:

X	0	1
P	5/6	1/6

$\mathbf{E}X = 0 \cdot 5/6 + 1 \cdot 1/6 = 1/6.$

2.32 pavyzdys. Tarkime, kad X tankis yra

$$p(x) = \begin{cases} 1, & \text{kai } x \in [2, 3], \\ 0 & \text{kitur.} \end{cases}$$

Tuomet

$$\mathbf{E}X = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \int_2^3 x \cdot 1\,\mathrm{d}x = \frac{x^2}{2}\bigg|_2^3 = \frac{9}{2} - \frac{4}{2} = 2{,}5.$$

2.33 pavyzdys. Tarkime, kad X tankis yra $p(x) = (1/\pi)(1 + x^2)^{-1}$. Nesunku įsitikinti, kad šiuo atveju vidurkis neegzistuoja.

Su begaliniu vidurkiu susijęs Sankt Peterburgo paradoksas. Paprasčiausia jo versija – Sankt Peterburgo lošimo namuose moneta metama tol, kol atsiverčia herbas. Jeigu herbas atsiverčia metant n-ąjį kartą, tai žaidėjas išlošia 2^n rublių. Koks vidutinis išlošis taip lošiant? Pasirodo, kad taip lošti labai apsimoka. Nesunku įsitikinti, kad vidutinis išlošis yra $2(1/2) + 4(1/4) + 8(1/8) + \cdots = 1 + 1 + 1 + \cdots = \infty$ rublių, t. y. vidutinis išlošis begalinis.

Sufomuluosime pagrindines vidurkių savybes:

1 Konstantos vidurkis lygus pačiai konstantai: $\mathbf{E}C = C$.

2 Konstantą galima iškelti prieš vidurkio ženklą: $\mathbf{E}CX = C\mathbf{E}X$.

3 Sumos vidurkis lygus vidurkių sumai: $\mathbf{E}(X+Y) = \mathbf{E}X + \mathbf{E}Y$.

4 Jeigu X ir Y *nepriklausomi*, tai $\mathbf{E}XY = \mathbf{E}X\mathbf{E}Y$.,

5 Jeigu $a \leqslant X \leqslant b$, tai $a \leqslant \mathbf{E}X \leqslant b$.

6 $|\mathbf{E}X| \leqslant \mathbf{E}|X|$.

Iš vidurkio apibrėžimo matyti, kad vidurkis yra skaičius. Jis gali būti ir teigiamas, ir neigiamas, ir didesnis už vienetą, ir mažesnis. Vidurkis pažymi vidutinę atsitiktinio dydžio reikšmę.

Galima apibrėžti ir atsitiktinio dydžio funkcijos vidurkį. Pateiktų diskrečiojo ir tolydžiojo atsitiktinių dydžių atvejais funkcijos vidurkis atitinkamai yra:

$$\mathbf{E}f(X) = f(x_1)p_1 + f(x_2)p_2 + \cdots, \qquad \mathbf{E}f(X) = \int_{-\infty}^{\infty} f(x)p(x)\,\mathrm{d}x. \tag{2.24}$$

2.34 pavyzdys. Tegul

$$p(x) = \begin{cases} 2x, & \text{kai } 0 \leqslant x \leqslant 1, \\ 0 & \text{kitur.} \end{cases}$$

Rasime $\mathbf{E}\sin X$. Remiantis (2.24) formule,

$$\mathbf{E}\sin X = \int_0^1 2x\sin x\,\mathrm{d}x = (-2x\cos x + 2\sin x)\Big|_0^1 = -2\cos 1 + 2\sin 1 = 1{,}96\ldots$$

Atsitiktinio dydžio X k-osios eilės *momentu, centriniu momentu, absoliučiuoju momentu* ir *centriniu absoliučiuoju momentu* atitinkamai vadinsime:

$$\begin{aligned} \nu_k &= \mathbf{E}X^k, & \mu_k &= \mathbf{E}(X - \mathbf{E}X)^k, \\ \alpha_k &= \mathbf{E}|X|^k, & \beta_k &= \mathbf{E}|X - \mathbf{E}X|^k. \end{aligned} \tag{2.25}$$

Centrinius momentus galima išreikšti paprastaisiais, ir atvirkščiai. Pavyzdžiui, $\mu_1 = 0$, $\mu_2 = \nu_2 - \nu_1^2$, $\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$, $\alpha_2 = \mu_2 + \alpha_1^2$, $\alpha_3 = \mu_3 + 3\mu_2\alpha_1 + \alpha_1^3, \ldots$

Be vidurkio, galima apibrėžti ir kitus teorinius empirinių charakteristikų analogus – modą, medianą, asimetrijos koeficientą $\gamma_1 = \mu_3/\mu_2^{3/2}$, eksceso koeficientą $\gamma_2 = \mu_4/\mu_2^2 - 3$ ir pan. Pavyzdžiui, diskrečiojo atsitiktinio dydžio teorinė moda yra ta skirstinio reikšmė, kurios įgijimo tikimybė didžiausia. Tolydžiojo atsitiktinio dydžio moda yra ta reikšmė, kur tankis pasiekia maksimumą. Skirstinys, kuris turi tik vieną modą, vadinamas *unimodaliuoju*.

15. Atsitiktinio dydžio dispersija

Vidurkis parodo vidutinę atsitiktinio dydžio reikšmę. Dispersija aprašo jo sklaidą apie vidurkį. Dispersija yra ne kas kita, kaip antrasis centrinis momentas (žr. 14).

Atsitiktinio dydžio X *dispersija* $\mathbf{D}X = \mathbf{E}(X - \mathbf{E}X)^2$. (2.26)

Skaičiavimams patogiau naudotis formule

$$\mathbf{D}X = \mathbf{E}X^2 - (\mathbf{E}X)^2. \tag{2.27}$$

Atkreipiame dėmesį, kad pirmas dėmuo yra *kvadrato* vidurkis, o antrasis – *vidurkio* kvadratas. Diskrečiųjų ir absoliučiai tolydžiųjų dydžių dispersija skaičiuojama pagal formules:

$$\mathbf{D}X = x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 + \cdots - (x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots)^2, \tag{2.28}$$

ir

$$\mathbf{D}X = \int_{-\infty}^{\infty} x^2 p(x)\,\mathrm{d}x - \left(\int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x\right)^2. \tag{2.29}$$

2.35 pavyzdys. Tarkime, kad X turi tokį skirstinį:

X	-1	0	1
P	$0{,}3$	$0{,}4$	$0{,}3$

$DX = EX^2 - (EX)^2$

Raskime $\mathbf{E}X$ ir $\mathbf{D}X$. Pagal apibrėžimą $\mathbf{E}X = (-1)\cdot 0{,}3 + 0\cdot 0{,}4 + 1\cdot 0{,}3 = 0$, $\mathbf{D}X = (-1)^2\cdot 0{,}3 + 0^2\cdot 0{,}4 + (1)^2\cdot 0{,}3 - 0^2 = 0{,}6$.

2.36 pavyzdys. Tarkime, kad X tankis toks kaip 2.32 pavyzdyje. Tuomet

$$\mathbf{D}X = \int_2^3 x^2\,\mathrm{d}x - (2{,}5)^2 = \frac{x^3}{3}\Big|_2^3 - \frac{25}{4} = \frac{19}{3} - \frac{25}{4} = \frac{1}{12}.$$

Dispersija yra skaičius. Yra atsitiktinių dydžių, kurie dispersijų neturi. Suformuluosime atsitiktinių dydžių, turinčių baigtines dispersijas, kai kurias dispersijos savybes:

1 Dispersija visuomet neneigiama: $\mathbf{D}X \geqslant 0$.

2 Konstantos dispersija lygi nuliui: $\mathbf{D}C = 0$.

3 Konstantą pakėlus kvadratu galima iškelti prieš dispersijos ženklą: $\mathbf{D}CX = C^2\mathbf{D}X$.

4 $\mathbf{D}(X + Y) = \mathbf{D}X + \mathbf{D}Y + 2\mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y)$.

5 Jeigu X ir Y *nepriklausomi*, tai $\mathbf{D}(X+Y) = \mathbf{D}X + \mathbf{D}Y$.

Kvadratinė šaknis iš dispersijos vadinama teoriniu standartiniu nuokrypiu.

Standartinis nuokrypis $= \sqrt{\mathbf{D}X}$.

Standartinis nuokrypis statistikoje naudojamas net dažniau už dispersiją. Taip yra todėl, kad jį lengviau interpretuoti. Pavyzdžiui, jeigu atsitiktinis dydis yra ūgis (matuotas cm), tai standartinio nuokrypio reikšmes irgi galima interpretuoti centimetrais.

16. Kovariacija ir koreliacijos koeficientas

Kovariacija ir koreliacijos koeficientas – tai skaitinės charakteristikos, įvertinančios dviejų atsitiktinių dydžių tiesinę priklausomybę.

Atsitiktinių dydžių X ir Y *kovariacija* $cov(X,Y) = \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y)$.

Skaičiuojant patogiau naudotis formule

$$cov(X,Y) = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y. \tag{2.30}$$

Norint ją įrodyti, užtenka pasinaudoti vidurkio savybėmis:

$$\begin{aligned}\mathbf{E}(X-\mathbf{E}X)(Y-\mathbf{E}Y) &= \mathbf{E}(XY - (\mathbf{E}X)Y - (\mathbf{E}Y)X + (\mathbf{E}X)(\mathbf{E}Y))\\ &= \mathbf{E}XY - \mathbf{E}((\mathbf{E}X)Y) - \mathbf{E}((\mathbf{E}Y)X) + \mathbf{E}(\mathbf{E}X)(\mathbf{E}Y)\\ &= \mathbf{E}XY - (\mathbf{E}X)\mathbf{E}Y - (\mathbf{E}Y)\mathbf{E}X + (\mathbf{E}X)(\mathbf{E}Y) = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y.\end{aligned}$$

Kovariacija yra skaičius, kuris gali būti ir teigiamas, ir neigiamas. Svarbiausios kovariacijos savybės yra šios:

1 Jeigu X ir Y yra *nepriklausomi*, tai $cov(X,Y) = 0$.

2 $|cov(X,Y)| \leqslant \sqrt{\mathbf{D}X\mathbf{D}Y}$.

Pirma savybė išplaukia iš vidurkio savybių, antroji – iš matematikos rezultato, vadinamo Helderio nelygybe.

Du atsitiktiniai dydžiai, kurių kovariacija lygi nuliui, vadinami *nekoreliuotaisiais*. Nepriklausomi dydžiai X ir Y visada nekoreliuoti. Iš tikro pagal 4) vidurkio savybę (žr. 14 skyrelį) kovariacija lygi nuliui. Palyginę kovariacijos apibrėžimą su 4) dispersijos savybe (žr. 15 skyrelį), matome, kad $\mathbf{D}(X+Y) = \mathbf{D}X + \mathbf{D}Y$ tada ir tik tada, kai X ir Y nekoreliuoti.

Nors šnekamojoje kalboje sąvokos „priklausomieji dydžiai" ir „koreliuotieji dydžiai" dažnai vartojamos kaip sinonimai, tai nėra teisinga. Iš kovariacijos apibrėžimo išplaukia: jeigu dydžiai koreliuoja, tai jie yra priklausomi; jeigu dydžiai nekoreliuoja, jie gali būti ir priklausomi, ir nepriklausomi.

2.37 pavyzdys ([7]). Tegul atsitiktiniai dydžiai Z ir X yra nepriklausomi, $\mathbf{E}X = 0$, $\mathbf{E}Z = 0$ ir $Y = ZX$. Akivaizdu, kad X ir Y stipriai priklausomi dydžiai. Tačiau Z ir X^2 yra nepriklausomi, todėl $cov(X,Y) = \mathbf{E}X^2Z - \mathbf{E}X\mathbf{E}Y = \mathbf{E}X^2\mathbf{E}Z - 0\cdot\mathbf{E}Y = 0 - 0 = 0$. Taigi X ir Y nekoreliuoja.

Nors kovariacija ir jautri atsitiktinių dydžių tiesiniam priklausomumui, ji nėra itin patogus priklausomybės matas. Kovariacija rodo ne tik priklausomybės stiprumą, bet ir kokias reikšmes atsitiktiniai dydžiai įgyja – dideles ar mažas. Pavyzdžiui, jeigu X matuotas metrais, tai pavertus metrus centimetrais gaunama visai kita kovariacija, nors iš tiesų kintamųjų X ir Y tarpusavio priklausomybės stiprumas nepasikeitė. Problema kyla ir norint palyginti keletą kovariacijų. Todėl reikia universalesnio nei kovariacija priklausomybės mato, nepriklausančio nuo matavimo vienetų. Toks matas yra *koreliacijos koeficientas*.

Atsitiktinių dydžių X ir Y *koreliacijos koeficientas*

$$\varrho(X,Y) = \frac{cov(X,Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}} = \frac{\mathbf{E}XY - \mathbf{E}X\mathbf{E}Y}{\sqrt{\mathbf{D}X\mathbf{D}Y}}.$$

Koreliacijos koeficiento savybės:

1 Jeigu a ir b yra konstantos, tai $\varrho(aX+b, Y) = \varrho(X,Y)$.

2 Koreliacijos koeficientas yra skaičius, kintantis intervale nuo -1 iki 1: $-1 \leqslant \varrho(X,Y) \leqslant 1$.

3 Koreliacijos koeficientas $\varrho(X,Y) = \pm 1$ tada ir tik tada, kai egzistuoja konstantos $a \neq 0$ ir b tokios, kad $Y = aX + b$.

Jeigu $\varrho(X,Y) = 1$, tai $a > 0$ (didesnius X atitiks didesni Y), jeigu $\varrho(X,Y) = -1$, tai $a < 0$ (didesnius X atitiks mažesni Y). Koreliacijos koeficientas *nematuoja* netiesinės priklausomybės.

2.38 pavyzdys. Turime tokį vektoriaus (X, Y) skirstinį:

$Y \setminus X$	0	1	2
1	0,1	0,2	0,3
2	0,2	0,1	0,1

Čia $P(X=1, Y=2) = 0,1$ ir pan. Apskaičiuosime koreliacijos koeficientą $\varrho(X,Y)$:

$$\mathbf{E}XY = 0\cdot 1\cdot 0,1 + 0\cdot 2\cdot 0,2 + 1\cdot 1\cdot 0,2 + 1\cdot 2\cdot 0,1 + 2\cdot 1\cdot 0,3 + 2\cdot 2\cdot 0,1 = 1,4.$$

Iš bendrojo (X, Y) skirstinio galima rasti marginaliuosius X ir Y skirstinius:

X	0	1	2
P	$0,1+0,2$	$0,2+0,1$	$0,3+0,1$

Y	1	2
P	$0,1+0,2+0,3$	$0,2+0,1+0,1$

$$\mathbf{E}X = 0\cdot 0,3 + 1\cdot 0,3 + 2\cdot 0,4 = 1,1, \qquad \mathbf{E}X^2 = 0\cdot 0,3 + 1\cdot 0,3 + 4\cdot 0,4 = 1,9,$$

$$\mathbf{D}X = 0,69, \qquad \mathbf{E}Y = 1,4, \qquad \mathbf{D}Y = 0,24.$$

Visas gautas reikšmes įstatę į koreliacijos koeficiento formulę, gauname $\varrho(X,Y) = -0,344$. Taigi X ir Y yra silpnai koreliuoti atsitiktiniai dydžiai.

17. Entropijos sąvoka

Intuityviai aišku, kad kai kuriuose atsitiktiniuose dydžiuose atsitiktinumo yra daugiau nei kituose. Palyginkime du atsitiktinius dydžius X ir Y.

X	0	1
P	0,5	0,5

Y	0	1
P	0,001	0,999

Abu jie įgyja tas pačias reikšmes – nulį ir vienetą. Tačiau (prisiminkime statistinį tikimybės apibrėžimą) daug kartų stebint X, nulių ir vienetų bus maždaug po lygiai (reikšmių tikimybės lygios). Tuo tarpu daug kartų stebint Y, beveik visuomet gausime vienetą (nulio tikimybė labai maža). Aišku, kad Y atsitiktinumas skiriasi nuo X atsitiktinumo. Skaitinis matas, parodantis atsitiktinio dydžio atsitiktinumo lygį, vadinamas *entropija*.

Atsitiktinio dydžio X *entropija*

$$H = -\sum_i p_i \ln p_i, \quad \text{jeigu } X \text{ diskretus;}$$

$$H = -\int_{-\infty}^{\infty} p(x) \ln p(x)\,\mathrm{d}x, \quad \text{jeigu } X \text{ absoliučiai tolydus.}$$

Čia $p(x) \ln p(x) = 0$, kai $p(x) = 0$. Apskaičiavę minėtų dydžių X ir Y entropijas, gauname $H_X = 0{,}693$, $H_Y = 0{,}0069$. Jeigu X įgyja reikšmes $x_1, x_2, \dots, x_n$ su tikimybėmis $p_1, p_2, \dots, p_n$, tai X entropija yra didžiausia, kai $p_1 = p_2 = \dots = p_n = 1/n$, o mažiausia, kai viena tikimybė lygi 1, o kitos lygios 0.

18. Diskrečiųjų skirstinių pavyzdžiai

18.1. Binominis skirstinys

Tarkime, atliekant eksperimentą galimos tik dvi baigtys – „sėkmė" ir „nesėkmė". Eksperimento sėkmės tikimybė yra p. Atliekame n nepriklausomų eksperimentų (t. y. turime Bernulio schemą, žr. 9). Sėkmių skaičius yra atsitiktinis dydis, kuris vadinamas *binominiu* atsitiktiniu dydžiu. Jis žymimas $X \sim \mathcal{B}(n, p)$, čia $0 < p < 1$, n – natūralusis skaičius. Binominio dydžio tikimybės nusakomos formule

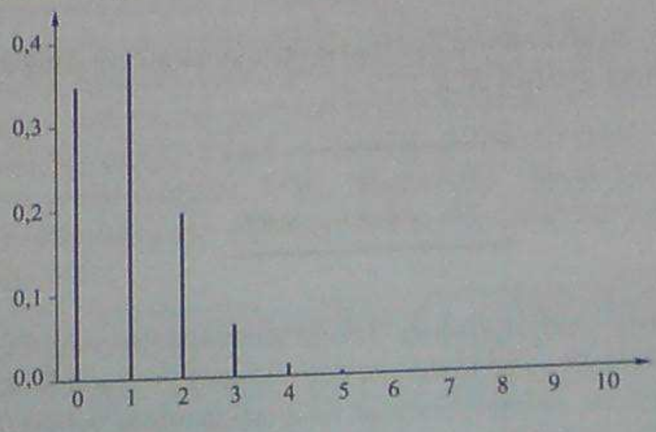
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n. \tag{2.31}$$

Skaitinės charakteristikos:

$$\mathbf{E}X = np, \qquad \mathbf{D}X = np(1-p). \tag{2.32}$$

Binominio dydžio dispersija ne didesnė už vidurkį. Binominio skirstinio tikimybės pavaizduotos 2.9 paveiksle.

Binominį skirstinį galima apibrėžti ir atvejais, kai $p = 0$ arba $p = 1$, tačiau tuomet atsitiktinumo nebelieka ir X virsta konstanta (0 arba n).



2.9 pav. Binominis skirstinys

18.2. Geometrinis skirstinys

Vieną kartą daromo bandymo sėkmės tikimybė $0 < p < 1$. Nepriklausomus bandymus kartojame tol, kol sulaukiame pirmos sėkmės. Atsitiktinis dydis X yra bandymų skaičius iki pirmos sėkmės. Sakysime, kad X turi *geometrinį* skirstinį. Geometrinio skirstinio tikimybės nusakomos formule

$$P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, \ldots . \tag{2.33}$$

Skaitinės charakteristikos:

$$\mathbf{E}X = \frac{1}{p}, \qquad \mathbf{D}X = \frac{1 - p}{p^2}. \tag{2.34}$$

2.39 pavyzdys. Naftos kompanijos atstovai žino, kad tiriamajame rajone 80% gręžinių naftos neturi. Kokia tikimybė rasti naftos gręžiant penktą kartą? Kiek vidutiniškai gręžinių reiks išgręžti, kol bus rasta nafta?

Sprendimas. X – padarytų gręžinių skaičius iki naftos radimo. Atsitiktinis dydis X turi geometrinį skirstinį su parametru $p = 0{,}20$ (jeigu 80% gręžinių naftos nėra, tai 20% – yra). Todėl ieškomoji tikimybė $P(X = 5) = 0{,}2 \cdot (0{,}8)^4 = 0{,}08192$, o vidurkis $\mathbf{E}X = 1/0{,}2 = 5$.

Būdingas geometrinio skirstinio pavyzdys yra minučių, kol klientas laukia eilėje, skaičius. Atkreipiame dėmesį, kad, *parinkdami* vieną ar kitą skirstinį realiai situacijai aprašyti, neišvengiamai darome tam tikrą paklaidą. Pavyzdžiui, geometrinis skirstinys numato situaciją, kad bandymų bus be galo daug. Akivaizdu, kad klientas eilėje nelauks be galo ilgai, o kompanija, ieškodama naftos, apsiribos tik tam tikru skaičiumi gręžinių. Tikslesnis modelio parinkimas reikalautų atsižvelgti į šį baigtinumą. Kodėl taip nedaroma? Geometrinio skirstinio didesnių reikšmių tikimybės tokios mažos, kad „pataisytų" modelių skaitinės charakteristikos beveik nesiskiria, tuo tarpu skaičiavimai ir formulės tampa gerokai gremėzdiškesnės. Ta pati situacija yra ir su kitais atsitiktiniais dydžiais. Visuomet reikia atsiminti, kad, parinkdami geriausiai tiriamąjį reiškinį aprašantį atsitiktinį dydį, pereiname prie *matematinio* modelio, kuris šiek tiek skiriasi nuo tiriamojo reiškinio.

18.3. Puasono skirstinys

Puasono[1] skirstinys dar vadinamas retų įvykių skirstiniu. Jis priklauso nuo vieno parametro $\lambda > 0$ ir žymimas $X \sim \mathcal{P}(\lambda)$. Puasono skirstinio tikimybės nusakomos formule

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, \ldots \tag{2.35}$$

Skaitinės charakteristikos:

$$\mathbf{E}X = \lambda, \qquad \mathbf{D}X = \lambda. \tag{2.36}$$

Puasono dydžio dispersija lygi vidurkiui. Didėjant λ, Puasono skirstinys tampa vis labiau simetriškas. Puasono skirstinys taikomas aprašant korektūros klaidų puslapyje skaičių, razinų bandelėje skaičių, telefono skambučių per valandą skaičių, radioaktyvių dalelių, išspinduliuotų per laiko vienetą, skaičių, gamybinių traumų per mėnesį skaičių, draudimo firmos išmokų per mėnesį skaičių ir pan. Puasono skirstiniu aprašomi modeliai dar vadinami „katastrofų“ modeliais.

Puasono skirstinys buvo pasiūlytas S. D. Puasono 1837 metais, tačiau statistikoje jis buvo panaudotas tik 1898 metais L. Bortkievičiaus (1868–1931) darbe. L. Bortkievičius pateikė statistinius rezultatus, liudijančius, kad Puasono skirstinį turi Prūsijos armijos kareivių, žuvusių įspyrus arkliui, skaičius.

2.40 pavyzdys. Vidutiniškai pirmadieniais į darbą neateina 3 darbuotojai. Kokia tikimybė, kad šį pirmadienį į darbą neateis ne mažiau kaip 2 darbuotojai?

Sprendimas. Stebime $X \sim \mathcal{P}(3)$.

Atsakymas: $P(X \geqslant 2) = 1 - P(X = 0) - P(X = 1) = 1 - e^{-3} - 3e^{-3} = 0{,}80\ldots$

Puasono skirstinys gali būti naudojamas binominiam dydžiui aproksimuoti, kai p yra labai mažas. Tuomet $X \sim \mathcal{B}(n, p)$ keičiamas $Y \sim \mathcal{P}(np)$.

18.4. Hipergeometrinis skirstinys

Turime N objektų, iš kurių M žymėti. Atsitiktinai renkame n objektų. Iš jų žymėtų objektų skaičius yra atsitiktinis dydis, turintis hipergeometrinį skirstinį. Hipergeometrinis skirstinys žymimas $X \sim \mathcal{H}(N, M, n)$. Jo tikimybės nusakomos formule

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \tag{2.37}$$

$$\max(0, M + n - N) \leqslant k \leqslant \min(M, n).$$

Skaitinės charakteristikos:

$$\mathbf{E}X = \frac{nM}{N}, \qquad \mathbf{D}X = \frac{nM(N-M)(N-n)}{N^2(N-1)}. \tag{2.38}$$

---

[1] Simeon Denis Poisson (1781–1840) – prancūzų matematikas.

2.41 pavyzdys. TV loterijoje tarp 60 skaičių yra 20 laimingų. Žaidėjas spėja 15 skaičių. Jeigu atspėja 5 skaičius, gauna x_1 Lt, jei 6 – x_2 Lt ir pan. Tarkime, kad X yra atspėtų laimingųjų skaičius, o Y – laimėta pinigų suma. Nesunku suprasti, kad $X \sim \mathcal{H}(60, 20, 15)$, o tikimybė

$$P(X = k) = \frac{\binom{20}{k}\binom{40}{15-k}}{\binom{60}{15}}.$$

Akivaizdu, kad Y įgyja reikšmes $x_1, x_2, \ldots$ su tikimybėmis $P(Y = x_k) = P(X = k)$. Tačiau Y jau nėra hipergeometrinis atsitiktinis dydis. Pavyzdžiui, vidutinis atspėtų laimingųjų skaičius yra $\mathbf{E}X = 15 \cdot 20/60 = 5$, tuo tarpu vidutinis išlošis $\mathbf{E}Y = x_1 P(Y = x_1) + x_2 P(Y = x_2) + \cdots$.

19. Tolydžiųjų skirstinių pavyzdžiai

19.1. Tolygusis skirstinys

Sakysime, kad X turi *tolygųjį* skirstinį intervale $[a, b]$, jei jo tankis

$$p(x) = \begin{cases} \dfrac{1}{b-a}, & \text{kai } a \leqslant x \leqslant b, \\ 0 & \text{kitur.} \end{cases} \tag{2.39}$$

Skaitinės charakteristikos:

$$\mathbf{E}X = \frac{a+b}{2}, \qquad \mathbf{D}X = \frac{(b-a)^2}{12}. \tag{2.40}$$

19.2. Normalusis skirstinys

Sakysime, kad atsitiktinis dydis X turi *standartinį normalųjį* skirstinį, jei jo tankis

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2}, \quad -\infty < x < \infty. \tag{2.41}$$

Jis žymimas $X \sim \mathcal{N}(0, 1)$. Standartinio normaliojo atsitiktinio dydžio pasiskirstymo funkcija

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\, \mathrm{d}y = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, \mathrm{d}y. \tag{2.42}$$

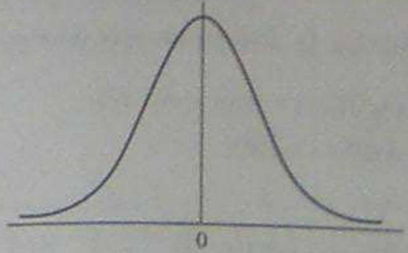
Tankis $\varphi(x)$ simetrinis, todėl

$$\Phi(-x) = 1 - \Phi(x). \tag{2.43}$$

Kvantilis $y_\beta = -y_{1-\beta}$. Standartinio normaliojo skirstinio skaitinės charakteristikos:

$$\mathbf{E}X = 0, \qquad \mathbf{D}X = 1. \tag{2.44}$$

Standartinio normaliojo skirstinio asimetrijos koeficientas ir ekscesas lygūs nuliui: $\gamma_1 = 0$, $\gamma_2 = 0$. Taigi standartinis normalusis skirstinys yra tas skirstinys, su kuriuo dažniausiai lyginami visi kiti. Pasiskirstymo funkcijos $\Phi(x)$ reikšmės pateiktos priedo 1 lentelėje.



2.10 pav. Tankio funkcijos $\varphi(x)$ grafikas

Atsižvelgiant į (2.43) savybę, lentelės sudaromos tik neneigiamiems x. Tankio funkcijos $\varphi(x)$ grafikas pavaizduotas 2.10 paveiksle.

Standartinis normalusis skirstinys yra atskiras bendrojo normaliojo skirstinio atvejis. Sakysime, kad atsitiktinis dydis turi *normalųjį skirstinį su parametrais* $-\infty < \mu < \infty$, $\sigma^2 > 0$, jei jo tankis

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty. \tag{2.45}$$

Jis žymimas $X \sim \mathcal{N}(\mu, \sigma^2)$. Normaliojo skirstinio $X \sim \mathcal{N}(\mu, \sigma^2)$ skaitinės charakteristikos:

$$\mathbf{E}X = \mu, \qquad \mathbf{D}X = \sigma^2, \qquad \sqrt{\mathbf{D}X} = \sigma. \tag{2.46}$$

Normalusis skirstinys dar vadinamas Gauso[1] skirstiniu. Daugelis statistinių išvadų remiasi prielaida, kad stebimas atsitiktinis dydis turi normalųjį skirstinį. Žinoma, sakydami, kad stebimas atsitiktinis dydis turi normalųjį skirstinį, mes tik pritaikome matematinį modelį, ignoruodami tai, kad stebimas dydis paprastai neįgyja be galo didelių arba be galo mažų reikšmių. Normalusis skirstinys gerai aprašo žmonių ūgį, svorį, vidutinę oro temperatūrą, matavimo paklaidas, molekulės judėjimo dujose greitį, vidutinį pelną, intelekto koeficientą.

Viena iš svarbiausių normaliojo skirstinio savybių yra jo ryšys su standartiniu normaliuoju skirstiniu. Jeigu $X \sim \mathcal{N}(\mu, \sigma^2)$, tai $(X-\mu)/\sigma \sim \mathcal{N}(0, 1)$ ir

$$P\left(\frac{X-\mu}{\sigma} \leqslant x\right) = \Phi(x), \tag{2.47}$$

$$P(a \leqslant X \leqslant b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \tag{2.48}$$

$$P(X \leqslant b) = \Phi\left(\frac{b-\mu}{\sigma}\right), \qquad P(X \geqslant a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right); \tag{2.49}$$

čia $\Phi(x)$ yra standartinio normaliojo dydžio pasiskirstymo funkcija. Šiose formulėse negriežtas nelygybes galima pakeisti griežtomis.

[1] Carl Friedrich Gauss (1777–1855) – vokiečių matematikas.

2.42 pavyzdys. Tarkime, kad $X \sim \mathcal{N}(1, 16)$. Raskime $P(|X| > 1)$. Ieškomą tikimybę perrašysime be absoliutinio didumo ir pritaikysime (2.43) ir (2.48) formules.

$$P(|X| > 1) = 1 - P(|X| \leqslant 1) = 1 - P(-1 \leqslant X \leqslant 1) = 1 - \Phi(0) + \Phi(-1/2)$$
$$= 2 - \Phi(0) - \Phi(1/2) = 2 - 0{,}5 - 0{,}6914 = 0{,}8086.$$

Čia pasinaudojome $\Phi(x)$ reikšmių priedo 1 lentele.

Aprašomojoje statistikoje jau nagrinėjome normaliąją kreivę (žr. I.5). Remiantis empirine taisykle, daugelio stebimų atsitiktinių reiškinių skirstiniai yra artimi normaliajam. Ši taisyklė nusako normaliojo skirstinio elgesį. Iš tikro, jei $X \sim \mathcal{N}(\mu, \sigma^2)$, tai

$$P(|X - \mu| < 3\sigma) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 0{,}9974\ldots$$

Analogiški rezultatai gaunami 2σ ir σ.

Statistinėms išvadoms plačiai taikomi keli su normaliuoju susiję skirstiniai – tai χ^2, Stjudento ir Fišerio. Susipažinsime su jais išsamiau.

19.3. χ^2 skirstinys

χ^2 (*chi* kvadratu) skirstinį *su n laisvės laipsnių* turi atsitiktinis dydis

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2, \tag{2.50}$$

čia $X_1, X_2, \ldots, X_n$ yra nepriklausomi, standartinį normalųjį skirstinį turintys atsitiktiniai dydžiai, $X_i \sim \mathcal{N}(0, 1)$. Šio skirstinio tankis yra gana sudėtingo pavidalo, tačiau jo skaitinės charakteristikos paprastos:

$$\mathbf{E}\chi_n^2 = n, \qquad \mathbf{D}\chi_n^2 = 2n. \tag{2.51}$$

χ^2 skirstinio $1 - \alpha$ lygmens kvantiliai (juos žymėsime $\chi_\alpha^2(n)$) pateikti priede knygos pabaigoje. Tarp χ^2 skirstinio ir Puasono skirstinio yra ryšys.

Tegul $X \sim \mathcal{P}(\lambda)$, o χ_{2m}^2 yra *chi kvadratu* skirstinys su $2m$ laisvės laipsnių. Tuomet

$$P(X < m) = P(\chi_{2m}^2 > 2\lambda). \tag{2.52}$$

19.4. Stjudento t skirstinys

Stjudento t skirstinį su n *laisvės laipsnių* turi atsitiktinis dydis

$$t_n = \frac{X}{\sqrt{(X_1^2 + X_2^2 + \cdots + X_n^2)/n}} = \frac{X}{\sqrt{\chi_n^2/n}}; \tag{2.53}$$

čia $X, X_1, X_2, \ldots, X_n$ yra nepriklausomi, standartinį normalųjį skirstinį turintys atsitiktiniai dydžiai, $X, X_i \sim \mathcal{N}(0, 1)$. Skaitinės charakteristikos:

$$\mathbf{E}t_n = 0, \qquad \mathbf{D}t_n = \sqrt{\frac{n}{n-2}}. \tag{2.54}$$

Stjudento (angl. student – studentas) skirstinio pavadinimas kilęs iš jo autoriaus V. Goseto[1] slapyvardžio. Dėstytojo skirstinio nėra.

Kai kurie Stjudento skirstinio $1 - \alpha$ lygmens kvantiliai (juos žymėsime $t_\alpha(n)$) pateikti priedo 3 lentelėje.

[1] William Sealey Gosset (1876–1937) – anglų statistikas.

19.5. Fišerio skirstinys

Fišerio[1] skirstinį su *m ir n laisvės laipsnių* turi atsitiktinis dydis

$$F_{m,n} = \frac{(Y_1^2 + Y_2^2 + \cdots + Y_m^2)/m}{(X_1^2 + X_2^2 + \cdots + X_n^2)/n} = \frac{\chi_m^2/m}{\chi_n^2/n}; \tag{2.55}$$

čia $Y_1, \ldots, Y_m, X_1, \ldots, X_n$ yra nepriklausomi, standartinį normalųjį skirstinį turintys atsitiktiniai dydžiai, $Y_j, X_i \sim \mathcal{N}(0, 1)$. Skaitinės charakteristikos:

$$\mathbf{E}F_{m,n} = \frac{n}{n-2}, \qquad \mathbf{D}F_{m,n} = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)}. \tag{2.56}$$

Fišerio skirstinio $1-\alpha$ lygmens kvantilius (juos žymėsime $F_\alpha(m,n)$) galima rasti priedo 5 lentelėje.

20. Čebyšovo nelygybė

Iš dispersijos apibrėžimo išplaukia, kad kuo mažesnė atsitiktinio dydžio dispersija, tuo daugiau dydžio įgyjamų reikšmių yra arčiau vidurkio. Kartu padidėja ir tikimybė, kad stebima reikšmė bus arti vidurkio. Įvertinti šią tikimybę padeda Čebyšovo nelygybė. Tegul X yra atsitiktinis dydis, turintis baigtinę dispersiją $\mathbf{D}X < \infty$. Tuomet kiekvienam fiksuotam $\varepsilon > 0$ teisinga

Čebyšovo nelygybė

$$P(|X - \mathbf{E}X| > \varepsilon) \leqslant \frac{\mathbf{D}X}{\varepsilon^2}. \tag{2.57}$$

Kartais vietoje (2.57) formulės naudojama jai ekvivalenti forma:

$$P(|X - \mathbf{E}X| \leqslant k\sqrt{\mathbf{D}X}) \geqslant 1 - 1/k^2, \tag{2.58}$$

čia $k > 0$. Čebyšovo nelygybė *universali*. Ji tinka visiems baigtinės dispersijos skirstiniams. Aprašomojoje statistikoje jau naudojomės Čebyšovo taisykle. Ji tėra atskiras (2.57) formulės taikymo pavyzdys. Iš tikrųjų į imties duomenų santykinių dažnių lentelę galima žiūrėti kaip į tam tikrą diskretųjį skirstinį:

x_1	x_2	$\ldots$	x_k
f_1/n	f_2/n	$\ldots$	f_k/n

Tarkime, kad tokį skirstinį turi atsitiktinis dydis X. Tuomet $\mathbf{E}X = x_1 f_1/n + x_2 f_2/n + \cdots + x_k f_k/k = \overline{x}$, $\mathbf{D}X = (x_1 - \overline{x})^2 f_1/n + \cdots + (x_k - \overline{x})^2 f_k/n = (n-1)s^2/n$. Įstatę jų reikšmes į (2.58) formulę, gauname

$$P(|X - \overline{x}| \leqslant ks) \geqslant P\left(|X - \overline{x}| \leqslant \sqrt{\frac{n-1}{n}}ks\right) \geqslant 1 - 1/k^2. \tag{2.59}$$

[1] Ronald Alymer Fisher (1890–1962) – anglų statistikas.

Belieka išsiaiškinti, ką (2.59) formulėje reiškia tikimybė. Iš skirstinio lentelės matyti, kad tai tam tikrų santykinių dažnių suma. Kiekvienas santykinis dažnis rodo, kiek kartų stebėjimas pasitaikė duomenų eilutėje, todėl santykinių dažnių suma – tai dalis duomenų eilutės elementų, tenkinančių nelygybę $|x_i - \bar{x}| \leqslant ks$. Taigi kairioji (2.59) pusė yra tik kitas Čebyšovo taisyklės formulavimas.

2.43 pavyzdys. Simetrišką kauliuką metame 180 kartų. Ką galima pasakyti apie atsivertusių 6 akučių skaičių? Šis skaičius X turi binominį skirstinį $X \sim B(180, 1/6)$. Todėl $\mathbf{E}X = 180/6 = 30$, $\mathbf{D}X = 180 \cdot (1/6) \cdot (5/6) = 25$. Pritaikę (2.58) formulę, kai $k = 4$, gauname $P(|X - 30| \leqslant 20) = P(10 \leqslant X \leqslant 50) \geqslant 15/16 = 0{,}9375$. Taigi su tikimybe, ne mažesne už 0,9375, atsivers nuo 10 iki 50 tokių akučių.

21. Didžiųjų skaičių dėsnis

J. Bernulio pomirtiniame 1713 metais atspausdintame darbe buvo įrodyta pirmoji tikimybių teorijos ribinė teorema. *J. Bernulis* ištyrė atsitiktinių dydžių sumos, padalytos iš jų skaičiaus, elgesį, kai sumuojamųjų dydžių skaičius neaprėžtai didėja. Jo gautasis rezultatas šiuolaikinėje tikimybių teorijoje vadinamas *didžiųjų skaičių dėsniu* (terminą pasiūlė S. D. Puasonas). Šiame skyriuje susipažinsime su dalimi didžiųjų skaičių dėsnio atvejų.

Tegul $X_1, X_2, \ldots, X_n$ yra nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai, kurių vidurkis $\mathbf{E}X_i = \mu$ ir dispersija $\mathbf{D}X_i = \sigma^2$. Tuomet kiekvienam fiksuotam $\varepsilon > 0$ galioja

Didžiųjų skaičių dėsnis

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0, \quad \text{kai } n \to \infty. \tag{2.60}$$

Įrodysime jį. Pažymėkime $\bar{X} = (X_1 + \cdots + X_n)/n$. Pasinaudoję vidurkio ir dispersijos savybėmis, gauname: $\mathbf{E}\bar{X} = (\mu + \mu + \cdots + \mu)/n = \mu$, $\mathbf{D}\bar{X} = (\mathbf{D}X_1 + \cdots + \mathbf{D}X_n)/n^2 = \sigma^2/n$. Parašome Čebyšovo (2.57) nelygybę atsitiktiniam dydžiui $\bar{X}$:

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

Tačiau $\sigma^2/(n\varepsilon^2) \to 0$, kai $n \to \infty$. Taigi (2.60) įrodėme.

Nors mes reikalavome, kad atsitiktiniai dydžiai turėtų baigtinę dispersiją, tačiau didžiųjų skaičių dėsniui galioti užtenka tik baigtinių vidurkių. Tiesa, įrodymas tampa sudėtingesnis.

Didžiųjų skaičių dėsnį nesunku interpretuoti. Kadangi ε gali būti labai mažas, tai didžiųjų skaičių dėsnis iš esmės teigia, kad esant dideliam nepriklausomų vienodai pasiskirsčiusių atsitiktinių dydžių skaičiui labai tikėtina, jog jų aritmetinis vidurkis mažai skirsis nuo tikrojo vidurkio (tikimybė, kad skirtumas bus didesnis už ε, yra maža).

2.44 pavyzdys. Tarkime, kad $X \sim B(n, p)$. Nesunku suprasti, kad $X = X_1 + \cdots + X_n$, čia $X_i \sim B(1, p)$ ir visi X_i nepriklausomi. Pritaikę didžiųjų skaičių dėsnį, gauname

$$\begin{aligned} P\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) &= P\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| < \varepsilon\right) \\ &= 1 - P\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| \geqslant \varepsilon\right) \to 1, \quad n \to \infty. \end{aligned} \tag{2.61}$$

Atsitiktinis dydis X/n yra ne kas kita kaip sėkmingų eksperimentų *santykinis dažnis*. Iš (2.61) formulės matyti, kad eksperimentą kartojant daug kartų santykinis dažnis mažai skiriasi nuo tikimybės p. Tai pagrindžia teorinę statistinę tikimybę.

22. Centrinė ribinė teorema

Empirinė taisyklė (žr. 1.5) skirta varpo formos skirstiniams. Faktiškai ji teigia, kad daugelis atsitiktinių dydžių turi skirstinius, panašius į normalųjį. Šiame skyrelyje suformuluosime teiginį, kad didinant sumuojamų atsitiktinių dydžių skaičių, jų sumų skirstiniai supanašėja su normaliuoju skirstiniu. Visos ribinės teoremos, tiriančios skirstinių artumą normaliajam, vadinamos bendru *centrinės ribinės teoremos* (CRT) vardu. Suformuluosime atskirą CRT atvejį.

Tegul $X_1, X_2, \ldots, X_n$ yra nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai, turintys vidurkius $\mathbf{E}X_i = \mu$ ir dispersijas $\mathbf{D}X_i = \sigma^2 > 0$. Tuomet

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) \to \Phi(x), \qquad \text{kai } n \to \infty. \tag{2.62}$$

Priminsime, kad $\Phi(x)$ yra standartinio normaliojo dydžio pasiskirstymo funkcija. Kaip ir didžiųjų skaičių dėsnis, centrinė ribinė teorema yra universali. Ji galioja ir tolydiesiems, ir diskretiesiems skirstiniams. Pažymėję $S_n = X_1 + X_2 + \cdots + X_n$, dideliems n (2.62) galime užrašyti taip:

$$P\left(\frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} \leqslant x\right) \approx \Phi(x). \tag{2.63}$$

Dažnai CRT galioja ir tuomet, kai S_n nėra nepriklausomų dydžių suma. Tuo paaiškinamas ir empirinės taisyklės (žr. 1.5) fenomenas – daugeliui atsitiktinių dydžių galioja (2.63) formulė.

Aptarsime, kaip CRT taikoma apytiksliams skaičiavimams. Tam (2.63) formulę perrašome taip:

$$P(S_n \leqslant x) \approx \Phi\left(\frac{x - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\right), \tag{2.64}$$

$$P(a \leqslant S_n \leqslant b) \approx \Phi\left(\frac{b - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\right) - \Phi\left(\frac{a - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\right). \tag{2.65}$$

Pagal šias formules galima apytiksliai skaičiuoti binominio, Puasono, χ^2 ir daugelį kitų atsitiktinių dydžių skirstinių, kai jų dispersijos didelės. Todėl minėtų skirstinių lentelės sudaromos tik nedideliems n. Diskretiesiems dydžiams kartais taikomos diskretumo pataisos, pavyzdžiui, jei $X \sim \mathcal{B}(n, p)$, tai

$$P(X \leqslant m) \approx \Phi\left(\frac{m - np + 0{,}5}{\sqrt{np(1-p)}}\right). \tag{2.66}$$

2.45 pavyzdys. Tarkime, gamykloje pagaminama 10% brokuotų gaminių. Kontrolei atrinkta 1000 gaminių. Kokia tikimybė, kad iš jų bus ne mažiau kaip 100 brokuotų? Kadangi brokuotų gaminių skaičiaus skirstinys yra binominis, t. y. $X \sim \mathcal{B}(1000; 0{,}1)$, tai šią tikimybę galima užrašyti tiksliai:

$$P(X \geqslant 100) = 1 - P(X \leqslant 99) = 1 - \left(\binom{1000}{0} 0{,}1^0 0{,}9^{1000} + \right.$$

$$\left. + \binom{1000}{1} 0{,}1^1 0{,}9^{999} + \cdots + \binom{1000}{99} 0{,}1^{99} 0{,}9^{901}\right),$$

tačiau tokią sumą itin sunku suskaičiuoti. Tuo tarpu pritaikę (2.66) formulę, gauname

$$P(X \geqslant 100) = 1 - P(X \leqslant 99) \approx 1 - \Phi\left(\frac{99 - 100 + 0{,}5}{\sqrt{90}}\right) \approx 0{,}49.$$

---

atsitiktinis įvykis	įvykių erdvė	nepriklausomi įvykiai
Bajeso formulė	įvykių sąjunga	nesutaikomi įvykiai
Bernulio schema	įvykių sankirta	normalusis skirstinys
centrinė ribinė teorema	įvykių skirtumas	pasiskirstymo funkcija
Čebyšovo nelygybė	klasikinė tikimybė	pilnosios tikimybės formulė
didžiųjų skaičių dėsnis	koreliacijos koeficientas	priešingasis įvykis
diskretusis dydis	kovariacija	tankis
dispersija	kritinė reikšmė	tikimybė
entropija	kvantilis	Veno diagrama
geometrinė tikimybė	negalimasis įvykis	vidurkis
įvykio dalis	nepriklausomi dydžiai	

UŽDAVINIAI

1. Duomenys apie vienos firmos darbuotojų amžių ir šeiminę padėtį pateikti 2.5 lentelėje. Raskite, kiek darbuotojų patenka į aibes: $A_1 \cap B_4$, $A_2 \cap B_3$, $A_3 \cup B_2$, $A_1 \cup A_3$, $B_1 \cup B_2$, $A_1 \cap (B_1 \cup B_4)$, $(A_1 \cup A_3) \cap B_2$.

2.5 lentelė

		A_1 18–25	A_2 26–35	A_3 36–45	A_4 46–55	A_5 55–70	Iš viso
B_1	Nevedę	20	8	5	1	2	36
B_2	Vedę	5	10	15	10	5	45
B_3	Išsiskyrę	3	6	8	4	2	23
B_4	Našliai	0	1	5	4	4	14
Iš viso		28	25	33	19	13	118

2. Iš dešimties detalių dvi brokuotos. Atsitiktinai imame vieną detalę. Kas yra elementarūs įvykiai?

3. Ministerijoje į du referento postus pretenduoja 5 kandidatai ir 2 kandidatės. Tarkime, kad nugalėtojai parenkami atsitiktinai. Raskite tikimybes, kad: a) bent vienas iš referentų yra vyras, b) tik vienas iš referentų yra vyras.

4. Ant atskirų kortelių užrašytos raidės: k, r, i, a, u, š, ė. Kokia tikimybė jas atsitiktinai išrikiavus į eilę gauti žodį „kriaušė"? Kokia tikimybė iš raidžių a, n, a, n, a, s, a, s sudėti žodį „ananasas"?

5. Žinoma, kad automatinės spynos kodą sudaro keturi skaitmenys. Kokia tikimybė atsitiktinai atspėti nežinomą kodą? Kaip pasikeis tikimybė žinant, kad visi skaitmenys skirtingi?

6. Įrodykite sąlyginės tikimybės 2) savybę (žr. 5).

7. Naujasis kotedžas yra trijų aukštų. Kiekviename aukšte – po du butus. Trys pirkėjai traukia burtus, kam koks butas teks. Kokia tikimybė, kad visi butus gaus skirtinguose aukštuose?

8. Audronė, Vilius ir dar dvidešimt jų draugų atsitiktinai susėda ratu. Kokia tikimybė, kad Audronė ir Vilius yra kaimynai?

9. Įstaigoje 70% darbuotojų turi humanitarinį išsilavinimą. Iš jų 80% yra moterų. Kokia tikimybė, kad atsitiktinai parinktas darbuotojas bus humanitarinį išsilavinimą turinti moteris?

10. Vyksta muitininkų mokymai. Kiekvienam iš dešimties muitininkų tikimybė atrasti mašinoje paslėptą kontrabandinį trilitrį „kaukolinio" yra lygi p. Kiekvienas muitininkas apžiūri jam paskirtas penkiolika mašinų, kuriose paslėpta kontrabanda. Kokia tikimybė, kad bent vieną kartą kontrabanda bus rasta? Kokia tikimybė, kad kiekvienas muitininkas bent kartą ras kontrabandą?

11. Kokia tikimybė, kad, atsitiktinai padalijus 24 kortų kaladę į dvi dalis, abiejose dalyse bus po lygiai juodų ir raudonų kortų?

12. Parduotuvėje buvo stebima, ar pirkėjas pirko naujos rūšies pieną (įvykis A), ar ne. Po to kiekvienas pirkėjas atsakė į klausimą, ar matė per TV naujojo pieno reklamą (įvykis B), ar ne. Duomenys pateikti 2.6 lentelėje. Raskite $P(A)$, $P(B)$, $P(A \cap B)$. Ar A ir B nesutaikomi? Ar A ir B nepriklausomi?

2.6 lentelė

	Matė	Nematė	Iš viso
Pirko	100	60	160
Nepirko	140	200	340
Iš viso	240	260	500

13. Miško prezidentu tapo Vilkas, paskelbęs, kad jam valdant ant eglių augs bananai. Tapęs prezidentu, Vilkas paskelbė įsaką, reikalaujantį kankorėžius vadinti bananais. Šimto narių parlamente kilo ginčas – įvykdė Vilkas rinkiminius pažadus ar ne. Šešiasdešimt humanitarinį išsilavinimą turinčių žvėrių teigė, kad Vilkas teisus, trisdešimt

tiksliųjų mokslų atstovų mano, kad Vilkas neteisus, o likę dešimt žaliųjų parlamentarų siūlė klausimą spręsti metant monetą. Kokia tikimybė, kad atsitiktinai parinktas parlamentaras nuspręs, jog Vilkas teisus?

14. Potencialus investuotojas iš Vakarų Konanas vyksta į Naująją R. Privažiuoja kryžkelę. Kryžkelėje stovi reklaminis akmuo. Ant akmens užrašyta: „Kairėn pasuksi – žmoną rasi (firma 'Vasilisa'). Tiesiai keliausi – turtą įgysi (kazino 'Trys karžygiai'). Dešinėn trauksi – mirtis (uostoma, rūkoma, ryjama, badoma – UAB 'Kaščėjus')". Išsitraukia Konanas iš kišenės žinyną (*New R. in your pocket*), o tenai parašyta: „Prostitucija, nelegalūs lošimai ir narkobiznis yra pašėlusiai pelningos šešėlinės ekonomikos sritys. Tačiau 80% investavusiųjų į prostituciją, 70% investavusiųjų į lošimus ir 90% investavusiųjų į narkobiznį tampa žudiko 'Lakštingala' darbo objektais. O 'Lakštingala' 99% dirba efektyviai". Konanas pamąsto ir nusprendžia mesti burtus. Jeigu iš 24 kortų kaladės ištrauks damą – vyks į kairę, jeigu karalių arba valetą – keliaus tiesiai, jeigu kryžių tūzą arba pikų dešimakę – suks dešinėn. Visais kitais atvejais apsuks „Mustangą" ir vyks investuoti į Baltijos šalis. Kokia tikimybė, kad Konanas liks gyvas?

15. Į vieną karalystę užklydo drakonas ir užsisakė pietums karalaitę. Karalystėje buvo skubiai organizuotas atvirasis konkursas karaliaus žento pareigoms (būtina sąlyga – sumedžioti drakoną). Atsirado pretendentas, kuris drakoną sumedžiojo. Pasičiupo liežuvį ir nuskuodė į rūmus. O ten – dar devyniolika pretendentų. Ir visi su daiktiniais įrodymais – kas su dantimi, kas su ausimi, kas su nagu. Sprendimus toje demokratiškoje karalystėje priimdavo trys asmenys: karalius, karalaitė arba juokdarys. Kuriam metas spręsti, lemdavo mestas kauliukas: jei atsiversdavo 6, 5 arba 4 akutės, spręsdavo karalius, jeigu 3 arba 2 akutės, spręsdavo karalaitė, jeigu 1 – juokdarys. Karalius pretendento paieškai žadėjo naudoti psichologų sukurtus testus (tikimybė, kad nustatys tikrąjį drakono medžiotoją, lygi 0,8). Karalaitė buvo pasiryžusi spręsti, kaip širdis lieps. Širdis karalaitei teisingai liepia 60 atvejų iš šimto. Juokdarys žadėjo problemą spręsti demokratiškai – visus pretendentus išrikiuoti į eilę, užsimerkti ir sviesti kepurę. Į kurį pataikys, tas ir tiks. Žinoma, kad konkursą laimėjo tikrasis drakono medžiotojas. Kokia tikimybė, kad jį parinko juokdarys?

16. Naudotos aparatūros parduotuvėje yra 8 televizoriai, iš jų 4 brokuoti. Pirkėjas renkasi televizorių tol, kol randa nebrokuotą. Raskite patikrintų televizorių skaičiaus skirstinį.

17. Atsitiktinis dydžio tankis

$$p(x) = \begin{cases} ax^2 - 1, & \text{kai } 1 \leqslant x \leqslant 2, \\ 0, & \text{kitur.} \end{cases}$$

Raskite a.

18. Suprastinkite reiškinius $\mathbf{E}(\mathbf{D}X)$, $\mathbf{D}(\mathbf{E}X)$, $\mathbf{E}(\mathbf{E}X)$, $\mathbf{D}(\mathbf{D}X)$.

19. Atsitiktiniai dydžiai X ir Y nepriklausomi, be to, $\mathbf{E}X = 1$, $\mathbf{E}Y = 2$, $\mathbf{D}X = 3$, $\mathbf{D}Y = 2$. Raskite $\mathbf{D}XY$.

20. Įrodykite dispersijos 1)–5) savybes (žr. 15). Įrodykite (2.27) formulę.

21. Įrodykite koreliacijos koeficiento 1) savybę (žr. 16).

22. Raskite koreliaciją tarp X ir Y, jeigu jų bendrasis skirstinys nusakytas 2.7 lentele.

2.7 lentelė

$X \backslash Y$	1	2
0	0,30	0,40
3	0,20	0,10

23. Moneta mėtoma tol, kol du kartus atsiverčia herbas. Raskite metimų skaičiaus skirstinį.

24. Tikimybė, kad baro klientas bus aptarnautas, kiekvieną minutę yra ta pati ir lygi 0,4. Kokia tikimybė, kad klientui teks laukti 4 minutes? Kiek vidutiniškai laukia klientai?

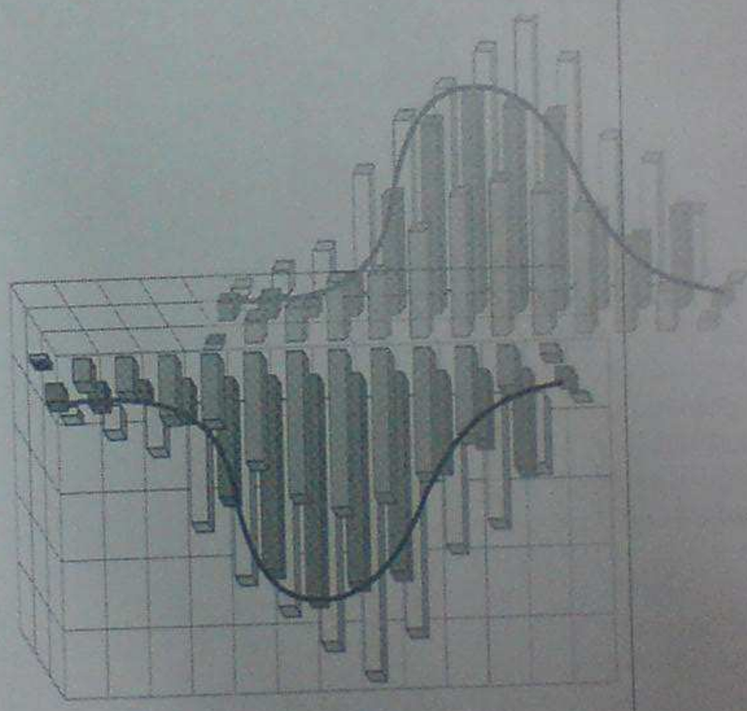
25. Kiek vidutiniškai razinų turi būti bandelėje, kad tikimybė rasti bent vieną raziną atsitiktinai paimtoje bandelėje būtų ne mažesnė kaip 0,9?

26. $X \sim \mathcal{N}(1, 4)$. Raskite $P(|X - 1| > 1)$, $P(|2X - 3| \leqslant 2)$.

27. Raskite Stjudento skirstinio su 30 laisvės laipsnių 97,5% ir 95% lygmens kvantilius.

3

STATISTINĖS IŠVADOS



Geras krikščionis turi saugotis matematikų ir visų tų, kurie skelbia netikras pranašystes.

Šv. Augustinas (354–430)

Statistikas – tai žmogus, kuris nutiesia matematiškai tikslų kelią nuo visiškai nepagrįstų prielaidų iki nepamatuotai apibendrinančių išvadų.

Ankstesnėse šios knygos dalyse susipažinome su pradinėmis statistikos sąvokomis, aprašomosios statistikos metodais ir tikimybių teorijos elementais. Šioje dalyje aprašomi statistinių išvadų, t. y. duomenų analizės ir interpretavimo, metodai.

Tarkime, sociologas nori sužinoti, kiek vidutiniškai minučių užtrunka Vilniaus ir Kauno gyventojai, vykdami į darbą. Sociologas iškelia hipotezę: vidutinė Vilniaus ir Kauno gyventojų vykimo į darbą trukmė ta pati. Po to apklausia 350 atsitiktinai parinktų vilniečių, 250 kauniečių ir gautus duomenis susistemina – apskaičiuoja empirinius vidurkius, dispersijas, nubraižo histogramas. Tarkime, skaičiavimai parodė, kad vykdami į darbą vilniečiai užtrunka vidutiniškai dešimčia minučių ilgiau už kauniečius. Bet gal šis skirtumas yra tik tarp *apklaustųjų* vilniečių ir kauniečių? Gal, apklausus visus vilniečius bei kauniečius, reikšmingo skirtumo nebeliktų? Pabandykime sudaryti tikimybinį uždavinio modelį. Tegul kintamasis X yra vilniečių vykimo į darbą laikas. Tarkime, Ω yra elementariųjų įvykių aibė, kurią sudaro įvykiai $\omega_1, \omega_2, \ldots$, čia ω_i – įvykis, kad buvo parinktas i-asis vilniečių populiacijos atstovas. Tuomet kintamasis X yra atsitiktinis dydis, kurio vidurkis sutampa su vidutiniu visų vilniečių vykimo į darbą laiku. Simboliu Y pažymėję kauniečių vykimo į darbą laiką, gauname tokį tikimybinį modelį: stebimi du atsitiktiniai dydžiai, kurių vidurkius norime palyginti. Taigi šioje dalyje:

populiacija $\longleftrightarrow$ elementariųjų įvykių aibė
kintamasis $\longleftrightarrow$ atsitiktinis dydis

Kintamojo (atsitiktinio dydžio) skirstinys sutampa su jo reikšmių *populiacijoje* santykinių dažnių lentele. Neretai kintamojo skirstinį gerai aproksimuoja žinomo tolydžiojo atsitiktinio dydžio skirstinys. Nagrinėjamojo pavyzdžio atveju sociologas remiasi centrine ribine teorema ir nusprendžia, kad vykimo į darbą trukmės apytikslis skirstinys yra normalusis. Tuomet padaro prielaidą, kad *visų* vilniečių vidutinė vykimo į darbą trukmė $\mathbf{E}X$ lygi vidutinei *visų* kauniečių vykimo į darbą trukmei $\mathbf{E}Y$. Pasinaudojęs Stjudento kriterijumi, sociologas nustato, kad tokia prielaida labai mažai tikėtina. Todėl padaro išvadą, kad vidutinis laikas, kurį užtrunka į darbą vykdami vilniečiai, *statistiškai* reikšmingai skiriasi nuo vidutinio laiko, kurį užtrunka į darbą vykdami kauniečiai.

Kodėl sociologas pasirenka Stjudento kriterijų? Kodėl apklausai pasirenkama po kelis šimtus žmonių? Kiek galima tikėti sociologo išvada? Kuo remdamasis jis atmeta hipotezę? Kodėl išvadoje jis pažymi, kad populiacijų vidurkiai *statistiškai* reikšmingai skiriasi? Į šiuos ir daugelį kitų klausimų pasistengsime atsakyti šioje vadovėlio dalyje.

1. IMTIES SKIRSTINIAI. ĮVERČIAI

Vieno žmogaus mirtis – tragedija. Milijono žmonių mirtis – statistika.

1.1. Imties atsitiktinumas. Statistikos sąvoka

Prisiminkime, kad pagrindinė statistikos problema – remiantis imtimi gauti išvadas apie visą populiaciją. Reikia skirti imties sudarymo *būdą* ir imties *rezultatą* (realizaciją). Iš imties realizacijos aprašymo (dažnių lentelių, histogramų ir pan.) matyti, kaip *gautus* duomenis galima susisteminti. Pirmojoje dalyje nagrinėjome, kaip konkrečiu atveju duomenys aprašomi, nekreipdami dėmesio į atsitiktinumą, kuris atsiranda sudarant imtį. Tuo tarpu imties sudarymo būdas lemia, kokie skaičiai ir kaip dažnai imtyje *gali* pasirodyti. Imtis gaunama n kartų stebint atsitiktinį dydį (matuojant kintamąjį). Imkime pirmąjį imties elementą. Aprašomojoje statistikoje kalbėjome apie konkrečią šio elemento *įgytą* reikšmę. Tačiau ką galima pasakyti apie *būsimą* pirmojo imties elemento reikšmę iki matuojant kintamąjį? Aišku, kad matuodami galime gauti visas tiriamo kintamojo reikšmes. Todėl *prieš* atlikdami konkrečius matavimus, pirmąjį imties elementą galime laikyti atsitiktiniu dydžiu, turinčiu tokį pat kaip ir tiriamas kintamasis skirstinį. Paprastosios atsitiktinės grąžintinės imties atveju tas pat pasakytina ir apie likusius imties elementus. Todėl bendrasis *atsitiktinės* imties modelis užrašomas taip.

Tarkime, kad n kartų matuojame atsitiktinį dydį X. Tuomet atsitiktinę imtį[1] sudaro atsitiktinis vektorius $(X_1, X_2, \ldots, X_n)$, kurio visi atsitiktiniai dydžiai $X_1, X_2, \ldots, X_n$ yra:

a) nepriklausomi,

b) vienodai pasiskirstę ir turintys tą patį kaip ir matuojamasis dydis X skirstinį.

Atsitiktinė imtis $(X_1, \ldots, X_n)$ ir bet kuri jos funkcija yra atsitiktiniai dydžiai, kurių skirstiniai priklauso nuo imties sudarymo būdo.

Dažnai mus domina ne visas stebimojo kintamojo skirstinys, o tik tam tikra jo charakteristika. Statistinėms išvadoms naudojama kokia nors imties duomenų funkcija – *statistika*.

Atsitiktinės imties funkcija $f(X_1, X_2, \ldots, X_n)$ vadinama *statistika*.

Terminas *statistika* – tradicinis. (Tikimės, kad nekils painiavos tarp statistikos – atsitiktinės imties funkcijos ir statistikos – mokslo.) Statistika yra atsitiktinių dydžių funkcija, todėl ji taip pat yra atsitiktinis dydis (vienamatis arba daugiamatis). Taigi galima kalbėti apie *statistikos*, arba vadinamąjį statistikos *imties, skirstinį*.

Pirmojoje šios knygos dalyje nagrinėjome konkrečias atsitiktinės imties ir statistikų realizacijas. Norėdami atskirti statistikos realizaciją (konkretų skaičių) nuo pačios statistikos (atsitiktinio dydžio), realizaciją žymėsime mažosiomis raidėmis, o statistiką – didžiosiomis. Pavyzdžiui, $\overline{X}$ yra statistika, o $\overline{x}$ – realizacija. Kiti statistikų pavyzdžiai: S^2, $X_{(1)}$.

[1] Čia ir toliau atsitiktine imtimi vadiname paprastąją atsitiktinę grąžintinę imtį. Įvade minėjome, kad realių tyrimų imtys retai būna grąžintinės, tačiau didelių populiacijų atveju grąžintinės ir negrąžintinės imčių rezultatų skirtumas toks nedidelis, kad jo galima nepaisyti.

3.1.1 pavyzdys. Firmoje dirba trys buhalteriai: Jonaitis, Ivanauskas ir Klimienė. Jonaičio darbo stažas yra dveji metai, Ivanausko – ketveri, Klimienės – šešeri. Kadrų skyriuje atsitiktinai parenkame vieną asmens bylą ir užsirašome darbo stažą. Ją grąžiname prie kitų bylų ir iš visų vėl parenkame antrąją (grąžintinis ėmimas). Apskaičiuojame užrašytų darbo stažų vidurkį. Galimos imties realizacijos pateikiamos 3.1.1 lentelėje.

3.1.1 lentelė

x_1	x_2	$\overline{x}$	x_1	x_2	$\overline{x}$
2	2	2	4	6	5
2	4	3	6	2	4
2	6	4	6	4	5
4	2	3	6	6	6
4	4	4			

Matome, kad imtis ir jos vidurkis gali įgyti įvairias reikšmes. Susisteminkime jas naudodamiesi santykinių dažnių lentelėmis. Šias lenteles galima traktuoti kaip atsitiktinių dydžių skirstinius. Pirmąjį atsitiktinį dydį žymėsime (X_1, X_2), o antrąjį – $\overline{X}$.

3.1.2 lentelė

(X_1, X_2)	(2, 2)	(2, 4)	(2, 6)	(4, 2)	(4, 4)	(4, 6)	(6, 2)	(6, 4)	(6, 6)
P	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

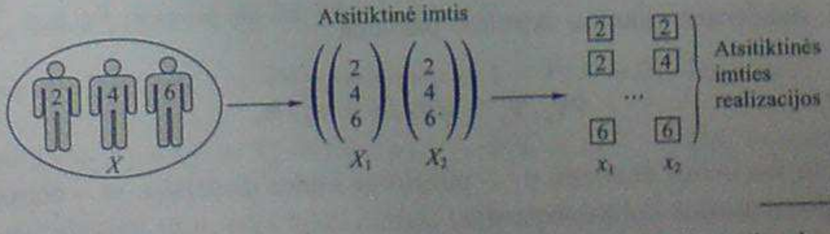
3.1.3 lentelė

$\overline{X}$	2	3	4	5	6
P	1/9	3/9	3/9	2/9	1/9

3.1.4 lentelė

X	2	4	6
P	1/3	1/3	1/3

Paties matuojamojo dydžio (pažymėkime jį X) skirstinys populiacijoje nusakytas 3.1.4 lentele. Tikimybių teorijos terminais turimą situaciją galima aprašyti taip: stebimas atsitiktinis dydis X – darbo stažas; (X_1, X_2) – atsitiktinė imtis, čia X_1, X_2 yra nepriklausomi atsitiktiniai dydžiai, turintys tokius pat skirstinius kaip ir X; $\overline{X} = (X_1 + X_2)/2$ – atsitiktinės imties vidurkis; $(x_1, x_2) = (2, 4)$, $\overline{x} = 3$ – galimos imties ir jos vidurkio realizacijos. Nesunku įsitikinti, kad visų darbuotojų darbo stažo vidurkis lygus 4. Pažymėtina, kad $\mathrm{E}\overline{X} = 4$ ir tai nėra atsitiktinumas.



Sprendžiant konkretų uždavinį, statistinės išvados remiasi tik viena (konkrečia) imties realizacija. Kad ir kokia ta realizacija būtų, pasirinktos sprendimo taisyklės turi tikti. Kitaip tariant, pasirėmę tikimybių teorijos metodais, turime įvertinti statistikos atsitiktinumo lygį ir pasirinkti tokią sprendimo procedūrą, kad klaidingos išvados tikimybė būtų maža.

Statistikos yra atsitiktiniai dydžiai, todėl galima apibrėžti jų vidurkius, dispersijas bei kitas charakteristikas.

Statistikos standartinis nuokrypis vadinamas *standartine statistikos paklaida*.
Vidurkio $\overline{X}$ standartinis nuokrypis vadinamas *standartine vidurkio paklaida*.

Standartinė 3.1.1 pavyzdžio vidurkio paklaida yra $2/\sqrt{3}$. Iš tikrųjų iš 3.1.2 lentelės gauname:

$$\mathbf{D}\overline{X} = (2-4)^2(1/9) + (3-4)^2(2/9) + 3(4-4)^2(3/9) + (5-4)^2(2/9) + (6-4)^2(1/9) = 4/3.$$

Kuo standartinė paklaida mažesnė, tuo statistika labiau koncentruota apie vidurkį.

Tarkime, turime atsitiktinę imtį $(X_1, X_2, \ldots, X_n)$. Be to, $\mathbf{D}X_i = \sigma^2$. Tuomet, pasinaudoję dispersijos savybėmis, gauname

$$\mathbf{D}\overline{X} = \mathbf{D}\frac{X_1 + \cdots + X_n}{n} = \frac{1}{n^2}(\mathbf{D}X_1 + \cdots + \mathbf{D}X_n) = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}.$$

Šiuo atveju standartinė vidurkio paklaida lygi $\sigma/\sqrt{n}$. Didėjant n, ji nyksta. Panašią įtaką stebėjimų skaičius (imties didumas) daro daugeliui statistikų – kuo daugiau imtyje elementų, tuo mažesnė standartinė statistikos paklaida, taigi mažesnė ir klaidingos statistinės išvados tikimybė. Realiai tikrosios standartinės paklaidos dažniausiai neįmanoma rasti, todėl vietoje jos skaičiuojamas empirinis standartinės paklaidos analogas, pavyzdžiui,

$$S_{\overline{X}} = \sqrt{\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n^2(n-1)}}. \tag{3.1.1}$$

Praktiškai tirdami susiduriame su kelių kintamųjų stebėjimais. Pavyzdžiui, lygindami: dviejų vertybinių popierių vidutinį kainų kilimą per tam tikrą laikotarpį; moterų, dirbančių vidurinėse ir aukštosiose mokyklose, procentą; dviejų vaistų efektyvumą ir pan. Visais šiais atvejais svarbus imčių statistikų skirtumas. Tarkime, kad pirmosios imties statistika yra T_1, o antrosios – T_2. Tuomet $\mathbf{D}(T_1 - T_2) = \mathbf{D}T_1 + \mathbf{D}T_2$. Gavome standartinės paklaidos $T_1 - T_2$ ir standartinių paklaidų T_1, T_2 ryšį:

$$\text{std. } (T_1 - T_2) \text{ paklaida } = \sqrt{(\text{std. } T_1 \text{ paklaida})^2 + (\text{std. } T_2 \text{ paklaida})^2}.$$

Pavyzdžiui, standartinė vidurkių skirtumo paklaida

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_1^2/n + \sigma_2^2/m},$$

čia n – pirmosios imties didumas, σ_1^2 – pirmosios imties dispersija, m – antrosios imties didumas, σ_2^2 – antrosios imties dispersija.

1.2. Dažniausiai naudojamų statistikų savybės

Ankstesniame skyrelyje statistiką apibrėžėme kaip imties funkciją. Elementai į imtį parenkami atsitiktinai, todėl statistika yra atsitiktinis dydis. Padarę prielaidas apie matuojamo atsitiktinio dydžio atsitiktinumą (skirstinį), galime gauti daugiau informacijos ir apie statistikos skirstinį, kartu ir apie įvairias duomenų aibės charakteristikas. Pateiksime kelis dažniausiai naudojamų statistikų skirstinius. Analogiškai kaip ir aprašomojoje statistikoje, imties $(X_1, \ldots, X_n)$ vidurkį žymėsime

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}, \tag{3.1.2}$$

dispersiją –

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\left(\overline{X}\right)^2. \tag{3.1.3}$$

1 Stebime $X \sim \mathcal{N}(\mu, \sigma^2)$. Tuomet:

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1), \tag{3.1.4}$$

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \text{ turi } \chi^2 \text{ skirstinį su } n \text{ laisvės laipsnių,} \tag{3.1.5}$$

$$\frac{\overline{X} - \mu}{\sqrt{S^2/n}} \text{ turi Stjudento skirstinį su } (n-1) \text{ laisvės laipsnių,} \tag{3.1.6}$$

$$(n-1)\frac{S^2}{\sigma^2} \text{ turi } \chi^2 \text{ skirstinį su } (n-1) \text{ laisvės laipsnių.} \tag{3.1.7}$$

Formulių (3.1.2)–(3.1.7) įrodymą galima rasti [7] knygoje.

2 Stebime $X \sim \mathcal{P}(\lambda)$. Tuomet

$$S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{P}(n\lambda).$$

Be to,

$$P(S_n < m) = P(\chi^2_{2m} > 2n\lambda), \tag{3.1.8}$$

čia χ^2_{2m} turi χ^2 skirstinį su $2m$ laisvės laipsnių.

3 Stebime $X \sim \mathcal{B}(1, p)$, t. y. $P(X = 1) = p = 1 - P(X = 0)$. Tuomet:

$$S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{B}(n, p), \qquad \mathbf{E}S_n = np, \qquad \mathbf{D}S_n = np(1-p).$$

Jeigu p, palyginti su n, nėra labai mažas, tai galima taikyti centrinę ribinę teoremą ir kalbėti apie *asimptotinį* skirstinį:

$$\frac{S_n - np}{\sqrt{np(1-p)}} = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0, 1). \tag{3.1.9}$$

Jeigu p labai mažas (np – nedidelis skaičius), tai $S_n \approx \mathcal{P}(np)$.

1.3. Taškiniai įverčiai

Dažniausiai mus domina ne visa informacija apie stebimo atsitiktinio dydžio skirstinį, o tik kai kurios skaitinės to skirstinio charakteristikos. Pavyzdžiui: norime sužinoti pirmą kartą tekančių nuotakų amžiaus vidurkį, įvertinti vidutinę švietimo darbuotojų gyvenimo trukmę; įvertinti maksimalų akcijų kurso svyravimą ir pan. Trumpai kalbant, remdamiesi imties duomenimis, norime įvertinti *populiacijos parametrą*. Matematiškai tokią situaciją aprašo *parametrinis* modelis.

> *Parametrinis modelis*: stebimas atsitiktinis dydis X, kurio skirstinys P_θ priklauso nuo nežinomo parametro θ iš aibės Θ. Žymima $X \sim P_\theta$.

Parametras θ gali būti ir labai abstraktus. Pavyzdžiui, galime kalbėti apie visus tolydžiuosius skirstinius, kurių vidurkiai nežinomi. Tačiau dažniausiai parametriniame modelyje yra žinomas stebimo atsitiktinio dydžio skirstinio tipas. Tarkime, norime įvertinti vidutinį skambučių telefono stotyje skaičių sekmadieniais. Matematiškai aprašius šį uždavinį, gaunamas parametrinis modelis, kur X yra telefono skambučių skaičius, turintis Puasono skirstinį $X \sim \mathcal{P}(\lambda)$ su nežinomu parametru $\lambda > 0$ (pagal apibrėžimą λ atitinka θ, o $\Theta = (0, \infty)$).

Parametras θ gali būti ir daugiamatis. Pavyzdžiui, stebėdami normalųjį atsitiktinį dydį, galime nežinoti nei jo vidurkio, nei dispersijos.

Tiriant parametrinius modelius, svarbiausias uždavinys yra įvertinti nežinomą parametrą.

> Statistika, kuri naudojama nežinomam parametrui θ įvertinti, vadinama θ taškiniu *įverčiu* ir žymima $\widehat{\theta}$.

Užuot vartoję terminą *taškinis įvertis*, toliau sakysime tiesiog *įvertis*. Atkreipiame dėmesį, kad:

1. nežinomas parametras θ yra *skaičius*;

2. parametro įvertis $\widehat{\theta}$ yra *atsitiktinis dydis*;

3. įverčio *realizacija* yra skaičius, randamas konkrečiai imties realizacija.

3.1.2 pavyzdys. Tarkime, mus domina maksimalus tam tikros valdininkų kategorijos intelekto koeficientas. Nežinomas parametras θ – tai tikrasis visų tos kategorijos valdininkų IQ maksimumas (konkretus skaičius). Natūralus θ įvertis yra $\widehat{\theta} = X_{(n)}$, t. y. taisyklė, reikalaujanti apytikslę θ reikšmę laikyti didžiausią imties elementą. Dėl atsitiktinės imties prigimties skirtingoms imties realizacijoms įverčio $X_{(n)}$ realizacijos yra skirtingos. Duomenų aibės (100; 100; 120; 100) įverčio realizacija $x_{(4)} = 120$, o duomenų aibės (180; 140; 120; 130) įverčio realizacija $x_{(4)} = 140$.

1.4. Taškinių įverčių klasifikacija

Gerais laikomi tie įverčiai, kurie bet kuriai imties realizacijai mažai skiriasi nuo tikrojo parametro. Aptarsime tris požymius, kurių dažniausiai reikalaujama iš gerų įverčių (nebūtinai visų trijų iškart).

1.4.1. Suderintasis įvertis

Įvertis yra suderintasis, jeigu jo realizacija apskaičiuota didelėms imtims, beveik nesiskiria nuo paties parametro. Tiksliai šis reikalavimas formuluojamas taip: didėjant imčiai, tikimybė, kad įvertis $\widehat{\theta}$ bent kiek reikšmingai (tarkime, per ε) skirsis nuo paties θ, artėja į nulį.

Parametro θ įvertis $\widehat{\theta}$ vadinamas *suderintuoju*, jeigu kiekvienam fiksuotam $\varepsilon > 0$

$$P\big(|\widehat{\theta} - \theta| > \varepsilon\big) \to 0, \quad \text{kai } n \to \infty.$$

3.1.3 pavyzdys. Stebimas atsitiktinis dydis X, kurio vidurkis $\mathbf{E}X = \mu$ nežinomas. Tuomet $\widehat{\mu} = \overline{X}$ yra suderintasis μ įvertis. Iš tikrųjų pagal didžiųjų skaičių dėsnį (žr. II.21),

$$P(|\overline{X} - \mu| > \varepsilon) = P\left(\left|\frac{(X_1 + \cdots + X_n)}{n} - \mu\right| > \varepsilon\right) \to 0, \quad \text{kai } n \to \infty.$$

Atkreipiame dėmesį, kad šiame pavyzdyje nereikėjo jokios informacijos apie X skirstinį.

1.4.2. Nepaslinktasis įvertis

Įverčio nepaslinktumas – viena iš dažniausiai pasitaikančių įverčių savybių. Tam tikra prasme tai reikalavimas, kad įverčio realizacijų nuokrypiai nuo parametro būtų subalansuoti.

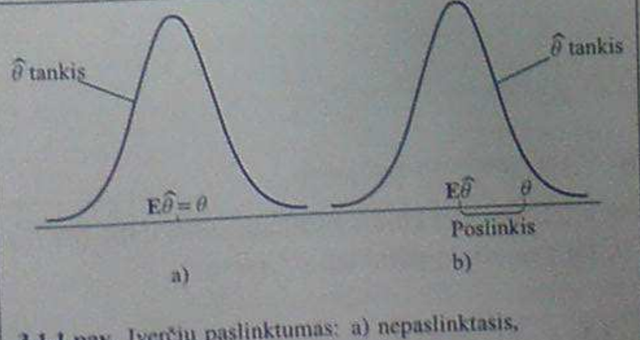
Parametro θ įvertis $\widehat{\theta}$ vadinamas *nepaslinktuoju*, jeigu $\mathbf{E}\widehat{\theta} = \theta$.

Normaliojo atsitiktinio dydžio nepaslinktojo ir paslinktojo įverčių tankiai parodyti 3.1.1 paveiksle, atitinkamai a) ir b).

Matome, kad b) atveju $\mathbf{E}\widehat{\theta} < \theta$, t. y. imties statistika $\widehat{\theta}$ su nemaža tikimybe $P(\widehat{\theta} < \theta)$ įgyja mažesnę už tikrojo parametro θ reikšmę. Baigtinės populiacijos atveju įverčio $\widehat{\theta}$ nepaslinktumas reiškia, kad, paėmus visas skirtingas dydžio n imtis ir suskaičiavus statistikos realizacijas $\widehat{\theta}$, visų šių realizacijų aritmetinis vidurkis lygus θ. Įsitikinsime, kad paprastosios atsitiktinės grąžintinės imties atveju $\overline{X}$ visuomet yra nepaslinktas vidurkio įvertis.

Tarkime, stebime atsitiktinį dydį X, kurio vidurkis $\mathbf{E}X = \mu$. Tegul imtis yra $(X_1, X_2, \ldots, X_n)$. Tuomet

$$\mathbf{E}\overline{X} = \mathbf{E}\frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}(\mathbf{E}X_1 + \mathbf{E}X_2 + \cdots + \mathbf{E}X_n) = \frac{1}{n}n\mu = \mu.$$



3.1.1 pav. Įverčių paslinktumas: a) nepaslinktasis, b) paslinktasis

Šiuo atveju neprireikė jokios papildomos informacijos apie X skirstinį. Taigi $\overline{X}$ yra universalus nepaslinktasis vidurkio įvertis. Šis faktas paaiškina, kodėl 3.1.1 pavyzdyje apie vidutinį darbo stažą įverčio vidurkis sutapo su tikruoju vidurkiu. Yra ir kitokių nepaslinktųjų įverčių, nepriklausančių nuo stebimo atsitiktinio dydžio skirstinio diskretumo, tolydumo ar kitų savybių.

3.1.4 pavyzdys. Įrodysime, kad S^2 yra nepaslinktasis nežinomos dispersijos įvertis. Tarkime, kad stebime X, kurio vidurkis $\mathbf{E}X = \mu$ nežinomas ir dispersija $\mathbf{D}X = \sigma^2$ taip pat nežinoma. Kadangi $X_1, \ldots, X_n$ turi tą patį skirstinį kaip X, tai $\mathbf{E}X_i = \mu$, $\mathbf{D}X = \sigma^2$.

$$\begin{aligned}\mathbf{E}S^2 &= \mathbf{E}\Big(\frac{1}{n-1}\sum_i (X_i - \overline{X})^2\Big) = \frac{1}{n-1}\mathbf{E}\sum_i (X_i - \mu + \mu - \overline{X})^2 \\ &= \frac{1}{n-1}\sum_i \mathbf{E}\big((X_i-\mu)^2 + 2(X_i-\mu)(\mu-\overline{X}) + (\mu-\overline{X})^2\big) \\ &= \frac{1}{n-1}\sum_i \mathbf{E}(X_i-\mu)^2 + \frac{2}{n-1}\mathbf{E}\sum_i (X_i-\mu)(\mu-\overline{X}) + \frac{n}{n-1}\mathbf{E}(\mu-\overline{X})^2.\end{aligned} \tag{3.1.10}$$

Be to,

$$\sum (X_i - \mu)(\mu - \overline{X}) = (\mu - \overline{X})\sum (X_i - \mu) = (\mu - \overline{X})(n\overline{X} - n\mu) = -n(\mu - \overline{X})^2. \tag{3.1.11}$$

Į (3.1.10) įstatę (3.1.11), gauname

$$\mathbf{E}S^2 = \frac{1}{n-1}n\sigma^2 - \frac{n}{n-1}\mathbf{E}(\mu - \overline{X})^2. \tag{3.1.12}$$

Atsitiktiniai dydžiai $X_1, \ldots, X_n$ yra nepriklausomi, nes toks yra paprastosios atsitiktinės imties reikalavimas. Todėl nepriklausomi ir $\mu - X_i$ bei $\mu - X_j$, kai $i \neq j$. Be to,

$$\mathbf{E}(\mu - X_i)(\mu - X_j) = \mathbf{E}(\mu - X_i)\mathbf{E}(\mu - X_j) = 0 \cdot 0 = 0, \quad \text{kai } i \neq j.$$

Taigi

$$\mathbf{E}(X_i - \mu)(\mu - \overline{X}) = -\frac{1}{n}\mathbf{E}(X_i - \mu)^2 + 0 = -\frac{\mathbf{D}X_i}{n} = -\frac{\sigma^2}{n}. \tag{3.1.13}$$

Belieka sutvarkyti $\mathbf{E}(\mu - \overline{X})^2$. Tačiau

$$(\mu - \overline{X})^2 = \frac{1}{n^2}\big((\mu - X_1)^2 + \cdots + (\mu - X_n)^2\big) + \frac{2}{n^2}\big((\mu - X_1)(\mu - X_2) + \cdots\big).$$

Todėl remdamiesi (3.1.13) gauname

$$\mathbf{E}(\mu - \overline{X})^2 = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) + 0 = \frac{\sigma^2}{n}.$$

Įstatę į (3.1.12), gauname $\mathbf{E}S^2 = \sigma^2$. Taigi S^2 yra nepaslinktasis σ^2 įvertis.

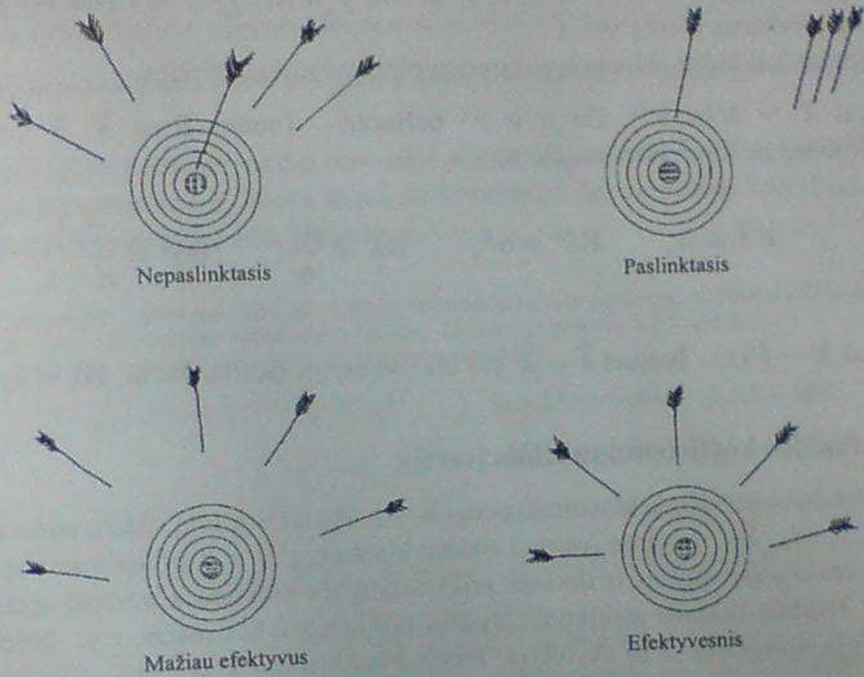
Būtent S^2 nepaslinktumas yra ta priežastis, dėl kurios kvadratų suma dalijama iš $n-1$. Jeigu kvadratų sumą dalytume iš n, tai gautume paslinktąjį σ^2 įvertį:

$$\mathbf{E}\frac{1}{n}\sum_i (X_i - \overline{X})^2 = \sigma^2\Big(1 - \frac{1}{n}\Big).$$

Matome, kad n didėjant $(1 - 1/n)$ artėja prie 1. Tokie įverčiai, kurių vidurkis didėjant imčiai artėja prie vertinamojo parametro ($\mathbf{E}\widehat{\theta} \to \theta$, kai $n \to \infty$), vadinami *asimptotiškai* nepaslinktaisiais.

Nepaslinktumas – įverčio ***vidurkio*** *savybė. Įvertis gali būti nepaslinktas, nors* ***nė viena*** *jo realizacija nesutampa su vertinamuoju parametru.*

Ne visi statistikoje naudojami įverčiai yra nepaslinktieji. Nors S^2 yra nepaslinktasis σ^2 įvertis, S yra paslinktasis standartinio nuokrypio įvertis.



Įverčiai

1.4.3. Efektyvusis įvertis

Tarkime, turime kelis n elementų imties nepaslinktuosius parametro θ įverčius. Natūralu geresniu laikyti tą įvertį, kuris labiau „sukoncentruotas" apie θ, t. y. kurio dispersija mažesnė.

Tegul $\widehat{\theta}_1$ ir $\widehat{\theta}_2$ yra nepaslinktieji θ įverčiai. Tuomet $\widehat{\theta}_1$ yra *efektyvesnis* už $\widehat{\theta}_2$, jeigu $\mathbf{D}\widehat{\theta}_1 < \mathbf{D}\widehat{\theta}_2$.

Tarkime, dukart stebime atsitiktinį dydį X, kurio vidurkis $\mathbf{E}X = \mu$ ir dispersija $\mathbf{D}X = \sigma^2$. Imtis (X_1, X_2). Sudarome du μ įverčius: $\widehat{\mu}_1 = X_1$ ir $\widehat{\mu}_2 = (X_1 + X_2)/2$. Nesunku įsitikinti, kad $\widehat{\mu}_1$ ir $\widehat{\mu}_2$ yra nepaslinktieji μ įverčiai. Tačiau $\mathbf{D}\widehat{\mu}_1 = \mathbf{D}X_1 = \sigma^2$,

$$\mathbf{D}\widehat{\mu}_2 = \frac{1}{4}(\mathbf{D}X_1 + \mathbf{D}X_2) = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2} < \sigma^2,$$

t. y. $\widehat{\mu}_2$ yra efektyvesnis už $\widehat{\mu}_1$.

Nepaslinktasis efektyvusis parametro θ įvertis $\widehat{\theta}$ yra efektyvesnis už visus likusius n elementų imties nepaslinktuosius įverčius.

Kartais efektyvumas suprantamas kiek kitaip. Efektyviuoju įverčiu vadinamas toks įvertis, kurio dispersija pasiekia tikimybinėje Rao–Kramero nelygybėje nurodytą dydį [7]. Jeigu efektyvumas suprantamas Rao–Kramero prasme, tai nepaslinktas mažiausios dispersijos įvertis žymimas NMD

Įverčiai, tiesiškai priklausantys nuo $X_1, X_2, \ldots, X_n$, t. y. $\widehat{K} = a_1X_1 + \cdots + a_nX_n$ ($a_1, \ldots, a_n$ – skaičiai), vadinami tiesiniais. Tiesinis nepaslinktasis įvertis, efektyvesnis už kitus tiesinius įverčius, vadinamas geriausiu tiesiniu nepaslinktuoju įverčiu (angl. BLUE[1]). Iš visų nepaslinktųjų vidurkio įverčių (t. y. iš visų $\sum a_iX_i$, $\sum a_i = 1$ įverčių) geriausias tiesinis nepaslinktasis įvertis yra $\overline{X}$.

Pateiksime kai kurių skirstinių parametrų efektyviuosius įverčius.

1 Tegul $X \sim \mathcal{N}(\mu, \sigma^2)$, čia μ ir σ^2 nežinomi. Tuomet $\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = S^2$ yra efektyvieji μ ir σ^2 įverčiai. Be to,

$$\mathbf{E}\overline{X} = \mu, \qquad \mathbf{E}S^2 = \sigma^2, \qquad \mathbf{D}\overline{X} = \frac{\sigma^2}{n}, \qquad \mathbf{D}S^2 = \frac{2\sigma^4}{n-1}.$$

2 Tegul $X \sim \mathcal{P}(\lambda)$. Tuomet $\widehat{\lambda} = \overline{X}$ yra efektyvusis λ įvertis. Be to, $\mathbf{D}\widehat{\lambda} = \lambda/n$.

1.5. Koreliacijos koeficiento taškinis įvertis

Tarkime, stebime intervalinių kintamųjų porą (X, Y). Atsitiktinę imtį sudaro poros (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$. Norime įvertinti tiesinę kintamųjų X ir Y priklausomybę. Grafinis (X, Y) realizacijų vaizdas ir tiesinės priklausomybės problemos trumpai aptartos I.8 skyrelyje. Teorinis tiesinės atsitiktinių dydžių priklausomybės matas, t. y. koreliacijos koeficientas ϱ, apibrėžtas II.16 skyrelyje. Pirsono koreliacijos koeficiento ϱ įvertis R:

$$R = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{\sqrt{\left(\sum(X_i - \overline{X})^2/(n-1)\right)\left(\sum(Y_i - \overline{Y})^2/(n-1)\right)}}. \tag{3.1.14}$$

Skaičiavimams dažnai naudojama tokia R išraiška:

$$R = \frac{(n-1)\sum X_iY_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{\left((n-1)\sum X_i^2 - \left(\sum X_i\right)^2\right)\left((n-1)\sum Y_i^2 - \left(\sum Y_i\right)^2\right)}}. \tag{3.1.15}$$

Koreliacijos koeficiento įverčio R realizacija r turi tokias savybes:

1 $-1 \leqslant r \leqslant 1$. Kuo r reikšmė absoliučiuoju didumu arčiau 1, tuo tiesinė Y priklausomybė nuo X stipresnė.

[1] Best Linear Unbiased Estimator.

2 Jeigu $r > 0$, tai didėjant X, didėja ir Y. Jeigu $r < 0$, tai didėjant X, Y mažėja.

3 r nepriklauso nuo X ir Y matavimo skalių. Pavyzdžiui, visus x_i padauginus iš 1000, koreliacijos koeficientas r nepasikeičia.

4 Koreliacijos X su Y koeficientas yra lygus koreliacijos Y su X koeficientui.

5 r neparodo *netiesinės* priklausomybės.

6 r priklauso nuo duomenų homogeniškumo. Kuo $x_1, x_2, \ldots, x_n$ arba $y_1, y_2, \ldots, y_n$ vienodesni, tuo r mažesnis. Jeigu visi x_i vienodi, tai $r = 0$.

7 Kuo didesnė imtis, tuo r yra arčiau nežinomo tikrojo koreliacijos koeficiento ϱ.

8 Koreliacijos koeficientas dar nenusako priežastingumo. Pavyzdžiui, peršalusio žmogaus temperatūros ir slogos stiprumo koreliacija dar nereiškia, kad sloga yra temperatūros priežastis, arba atvirkščiai.

3.1.5 pavyzdys. Firma nori įvertinti tiesinę priklausomybę tarp pardavėjų skaičiaus (X) ir parduodamos produkcijos kiekio (Y), matuoto tonomis per mėnesį. Duomenys pateikti 3.1.5 lentelėje.

Apskaičiuosime koreliacijos koeficiento įverčio realizaciją r:

$$\sum x_i = 10 + 13 + \cdots + 32 = 213, \qquad \sum y_i = 130 + 160 + \cdots + 380 = 2621,$$

$$\sum x_i^2 = 10^2 + 13^2 + \cdots + 32^2 = 5037,$$

$$\sum y_i^2 = 130^2 + 160^2 + \cdots + 380^2 = 780\,371,$$

$$\sum x_i y_i = 10 \cdot 130 + 13 \cdot 160 + \cdots + 32 \cdot 380 = 62\,085, \quad n = 10,$$

$$r = \frac{620\,850 - 213 \cdot 2621}{\sqrt{(50\,370 - 45\,369)(7\,803\,710 - 6\,869\,641)}} = 0{,}915.$$

Gavome, kad parduodamos produkcijos kiekis stipriai tiesiškai priklauso nuo pardavėjų skaičiaus. Kadangi koreliacijos koeficientas teigiamas, tai priklausomybė yra tiesioginė – kuo daugiau pardavėjų, tuo daugiau parduodama.

3.1.5 lentelė

Metai	X	Y	Metai	X	Y
90	10	130	95	24	295
91	13	160	96	25	339
92	15	234	97	27	320
93	17	240	98	30	360
94	20	263	99	32	380

Yra nusistovėjusios tam tikros tradicijos, kokią koreliaciją laikyti stipria. Jos taikytinos dideliems n ir pateiktos 3.1.6 lentelėje. Atkreipiame dėmesį, kad stipri koreliacija gali būti ir neigiama (kai r arti -1). Koreliacija silpna, kai r arti nulio.

3.1.6 lentelė. Empiriniai r vertinimai

r reikšmė	Interpretacija
Nuo 0,9 iki 1,0 (nuo −0,9 iki −1,0)	Labai stipri teigiama (neigiama) tiesinė koreliacija
Nuo 0,7 iki 0,9 (nuo −0,7 iki −0,9)	Stipri teigiama (neigiama) tiesinė koreliacija
Nuo 0,5 iki 0,7 (nuo −0,5 iki −0,7)	Vidutinė teigiama (neigiama) tiesinė koreliacija
Nuo 0,3 iki 0,5 (nuo −0,3 iki −0,5)	Silpna teigiama (neigiama) tiesinė koreliacija
Nuo 0,3 iki −0,3	Labai silpna koreliacija arba jokios

1.6. Įverčių sudarymo būdai

Ankstesniuose skyreliuose aprašėme įverčių savybes. Tačiau kaip juos sukonstruoti? Aptarsime du geriausiai žinomus įverčių sudarymo būdus.

1.6.1. Momentų metodas

Tiriamojo atsitiktinio dydžio X skirstinys priklauso nuo nežinomo parametro θ, todėl nuo jo turėtų priklausyti ir momentai. Momentų metodas siūlo sulyginti atsitiktinio dydžio momentus su jų empiriniais atitikmenimis ir sudaryti lygčių sistemą:

$$\mathbf{E}X = \overline{X}, \qquad \mathbf{D}X = S^2 \quad \text{ir t. t.}$$

Lygčių sudaroma tiek, kiek yra nežinomų parametrų (primename, kad bendru atveju $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$). Išsprendę sudarytas lygtis nežinomų parametrų atžvilgiu, gauname $\theta_1, \theta_2, \ldots, \theta_k$ įverčius.

3.1.6 pavyzdys. Stebime $X \sim \mathcal{N}(\mu, \sigma^2)$, kurio μ ir σ^2 nežinomi. Momentų metodu raskime μ ir σ^2 įverčius. Kadangi $\mathbf{E}X = \mu$, $\mathbf{D}X = \sigma^2$, tai iškart gauname

$$\mu = \overline{X}, \qquad \sigma^2 = S^2.$$

Pažymime, kad gavome įverčius (t. y. atsitiktinius dydžius).

Atsakymas: $\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = S^2$.

3.1.7 pavyzdys. Stebime $X \sim \mathcal{B}(100, p)$. Momentų metodu raskime p įvertį. Iš tikimybių teorijos žinome, kad $\mathbf{E}X = 100p$. Todėl

$$100p = \overline{X}, \qquad p = \frac{\overline{X}}{100}.$$

Atsakymas: $\widehat{p} = \overline{X}/100$.

3.1.8 pavyzdys. Stebime $X \sim \mathcal{P}(\lambda)$. Momentų metodu raskime λ įvertį. Žinome, kad $\mathbf{E}X = \lambda$. Todėl $\widehat{\lambda} = \overline{X}$. Tačiau ir $\mathbf{D}X = \lambda$. Todėl galime imti $\widehat{\lambda} = S^2$. Kuris iš dviejų įverčių – $\widehat{\lambda} = \overline{X}$, $\widehat{\lambda} = S^2$ – geresnis? Momentų metodu atsakymo į šį klausimą negauname.

3.1.9 pavyzdys. Stebime tolygiai intervale $[0, \theta]$ pasiskirsčiusį atsitiktinį dydį X. Momentų metodu įvertinkime θ. Tolygiojo atsitiktinio dydžio tankis

$$p(x) = \begin{cases} 1/\theta, & \text{kai } 0 \leqslant x \leqslant \theta, \\ 0, & \text{kitur.} \end{cases} \qquad (3.1.16)$$

Tada vidurkis

$$\mathrm{E}X = \int_0^\theta \frac{x}{\theta}\,\mathrm{d}x = \frac{x^2}{2\theta}\bigg|_0^\theta = \frac{\theta}{2}.$$

Teorinį vidurkį prilyginame empiriniam: $\overline{X} = \theta/2$, $\widehat{\theta} = 2\overline{X}$.

Tarkime, gauta tokia imties realizacija: $(1, 2, 3, 10)$. Tuomet $\theta = 2\overline{X} = 8$. Atrodytų, kad X su nenuline tikimybe turėtų įgyti reikšmes tik iš intervalo $(0, 8)$. Tačiau $x_{(4)} = 10$. Taigi įsitikinome, kad momentų metodas nėra visiškai patikimas.

1.6.2. Didžiausio tikėtinumo metodas

Didžiausio tikėtinumo metodą pirmiausia aptarsime tolydžiųjų atsitiktinių dydžių atveju. Tarkime, stebime atsitiktinį dydį X, kurio tankis $p_\theta(x)$ priklauso nuo nežinomo vienamačio parametro θ. *Tikėtinumo* funkcija sudaroma taip:

$$\mathcal{L}_\theta = p_\theta(X_1)p_\theta(X_2)\cdots p_\theta(X_n). \tag{3.1.17}$$

Taigi tankio funkcijoje vietoje argumento iš eilės įstatome $X_1, \ldots, X_n$. Ieškome tokio θ, kuris maksimizuotų funkciją $\mathcal{L}_\theta$. Dažniausiai tai daroma taip:

1) randame $\ln \mathcal{L}_\theta$;
2) apskaičiuojame $\ln \mathcal{L}_\theta$ išvestinę pagal θ: $(\ln \mathcal{L}_\theta)'$;
3) prilyginame rastą išvestinę nuliui $(\ln \mathcal{L}_\theta)' = 0$ ir gautą lygtį išsprendžiame θ atžvilgiu;
4) gautą rezultatą $\hat{\theta}$ laikome θ įverčiu.

Pastaba. Jeigu $\theta = (\theta_1, \ldots, \theta_k)$, t. y. turime k nežinomų parametrų, tai randamos dalinės $\ln \mathcal{L}_\theta$ išvestinės pagal $\theta_1, \ldots, \theta_k$. Jos prilyginamos nuliui, ir sprendžiama gautoji k lygčių sistema

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_1} = 0, \\ \cdots\cdots\cdots \\ \dfrac{\partial \mathcal{L}}{\partial \theta_k} = 0. \end{cases}$$

Gauti $\widehat{\theta_1}, \ldots, \widehat{\theta_k}$ laikomi ieškomaisiais įverčiais.

3.1.10 pavyzdys. Stebime $X \sim \mathcal{N}(\mu, 1)$. Didžiausio tikėtinumo metodu įvertinkime μ. Nagrinėjamu atveju X tankis

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2}\right\}.$$

Todėl

apsk. ln Lθ

1

$$\begin{aligned}\mathcal{L}_\mu &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_1-\mu)^2}{2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_2-\mu)^2}{2}\right\} \cdots \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_n-\mu)^2}{2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\left((X_1-\mu)^2 + (X_2-\mu)^2 + \cdots + (X_n-\mu)^2\right)\right\}.\end{aligned}$$

Logaritmuojame $\mathcal{L}_\mu$:

$$\ln \mathcal{L}_\mu = -n \ln \sqrt{2\pi} - \frac{1}{2}\left((X_1-\mu)^2 + (X_2-\mu)^2 + \cdots + (X_n-\mu)^2\right).$$

Randame išvestinę pagal μ:

① išvestinė

$$(\ln \mathcal{L}_\mu)' = 0 - \frac{1}{2}\big(2(X_1-\mu)(-1) + 2(X_2-\mu)(-1) + \cdots + 2(X_n-\mu)\big)(-1)$$
$$= (X_1 + X_2 + \cdots + X_n) - n\mu.$$

③ prilyginam 0

Gautą reiškinį prilyginame nuliui ir išsprendžiame μ atžvilgiu.

$$\mu = (X_1 + X_2 + \cdots + X_n)/n = \overline{X}.$$

Taigi ieškomas įvertis $\widehat{\mu} = \overline{X}$.

Kartais tikėtinumo funkcijos maksimumą galima rasti ir be išvestinės.

3.1.11 pavyzdys. Stebime tolygiai intervale $[0, \theta]$ pasiskirsčiusį atsitiktinį dydį X (jo tankis aprašytas (3.1.16) formule). Didžiausio tikėtinumo metodu įvertinkime parametrą θ. Tada tikėtinumo funkcija yra

$$\mathcal{L}_\theta = \begin{cases} 1/\theta^n, & 0 \leqslant X_i \leqslant \theta,\ i = 1, 2, \ldots, n, \\ 0, & \text{kitur.} \end{cases} \tag{3.1.18}$$

Aišku, kad $\mathcal{L}_\theta$ pasiekia maksimumą, kai θ mažiausias, o $0 \leqslant X_i \leqslant \theta$ su visais i. Taigi $0 \leqslant X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)} \leqslant \theta$. Imame $\theta = X_{(n)}$, nes θ mažesnis už $X_{(n)}$ negali būti (tuomet būtų $\mathcal{L}_\theta = 0$). Skirtingai nei skaičiuojant momentų metodu (žr. 3.1.9 pavyzdį), imties reikšmės niekada neviršija intervalo $[0, \theta]$ viršutinio rėžio.

Didžiausio tikėtinumo metodą taikant diskretiems dydžiams, tikėtinumo funkcija sudaroma taip:

$$\mathcal{L}_\theta = P_\theta(X = X_1)P_\theta(X = X_2)\cdots P_\theta(X = X_n).$$

Užrašas $P(X = X_1)$ reiškia, kad tikimybės skaičiavimo formulėje $P(X = a)$ vietoje a formaliai įrašome X_1. Tolesnis tyrimas toks pat kaip ir tolydžiojo atsitiktinio dydžio atveju.

3.1.12 pavyzdys. Stebime $X \sim \mathcal{P}(\lambda)$. Didžiausio tikėtinumo metodu įvertinkime parametrą λ. Žinome, kad

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

Todėl

$$\mathcal{L}_\lambda = \frac{\lambda^{X_1}}{X_1!}e^{-\lambda}\frac{\lambda^{X_2}}{X_2!}e^{-\lambda}\cdots\frac{\lambda^{X_n}}{X_n!}e^{-\lambda} = \frac{\lambda^{X_1+X_2+\cdots+X_n}}{X_1!X_2!\cdots X_n!}e^{-n\lambda},$$
$$\ln \mathcal{L}_\lambda = -\ln X_1!X_2!\cdots X_n! + (X_1 + X_2 + \cdots + X_n)\ln\lambda - n\lambda,$$
$$(\ln \mathcal{L}_\lambda)' = 0 + (X_1 + X_2 + \cdots + X_n)/\lambda - n.$$

Šį reiškinį prilyginę nuliui ir išsprendę λ atžvilgiu, gauname

$$\widehat{\lambda} = (X_1 + \cdots + X_n)/n = \overline{X}.$$

Didžiausio tikėtinumo metodu gaunami geresni taškiniai įverčiai negu momentų metodu. Todėl Puasono skirstinio atveju geriau naudoti įvertį $\widehat{\lambda} = \overline{X}$ (plg. 3.1.8 ir 3.1.9 pavyzdžius). Dauguma didžiausio tikėtinumo metodu gautų įverčių yra paslinktieji.

1.7. Pasikliautinieji intervalai

1.7.1. Pasikliautinojo intervalo sąvoka

Parametrų įverčiai yra atsitiktiniai dydžiai. Jų realizacijos yra išsibarsčiusios apie tikrąją parametro reikšmę. Taikymams svarbu žinoti *intervalą*, kuriam gali priklausyti nežinomas parametras. Tarkime, stebime atsitiktinį dydį, turintį skirstinį P_θ su nežinomu $\theta \in \Theta$. Pasirenkame skaičių $0 < Q < 1$.

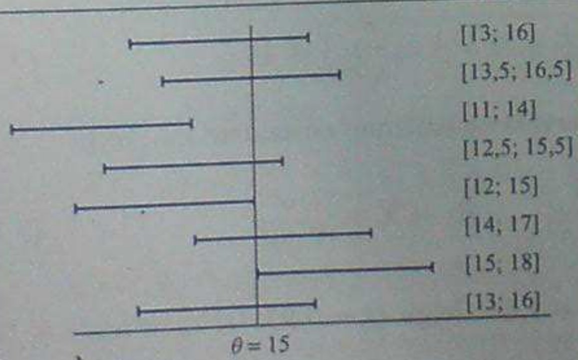
Tegul $\widehat{\theta}_1, \widehat{\theta}_2$ dvi tokios statistikos, kad $P(\widehat{\theta}_1 \leqslant \theta \leqslant \widehat{\theta}_2) = Q$.
Intervalas $[\widehat{\theta}_1, \widehat{\theta}_2]$ vadinamas parametro θ *pasikliautinuoju* intervalu.
Skaičius Q vadinamas *pasikliovimo lygmeniu*.

Tradiciniai pasikliovimo lygmenys $Q = 0{,}9; 0{,}95; 0{,}99$. Pasikliautinojo intervalo rėžiai $\widehat{\theta}_1, \widehat{\theta}_2$ yra statistikos (t. y. atsitiktiniai dydžiai). Panagrinėsime pasikliautinojo intervalo ir jo realizacijų skirtumą. Pasikliautinasis intervalas yra atsitiktinis dydis (vektorius), kurio skirstinys toks, kad tikimybė, jog θ priklausys šiam intervalui, didelė (lygi Q). Pasikliautinojo intervalo realizacijoje (pvz., [10; 20]) atsitiktinumo nebėra. Parametras arba priklauso intervalui [10; 20], arba ne. Pasikliovimo lygmuo Q tik atskleidžia mūsų pasitikėjimą (*pasikliovimą*) intervalo *sudarymo taisyklėmis*. Tarkime, kad $Q = 0{,}95$. Vadinasi, daug kartų taikant $[\widehat{\theta}_1, \widehat{\theta}_2]$ skaičiavimo taisykles skirtingoms imčių realizacijoms, parametras θ priklauso maždaug 95% visų intervalų.

Statistikas neteisingą išvadą padaro 95% pasikliaudamas savo sprendimu.

Parametro θ pasikliautinųjų intervalų realizacijų pavyzdžiai pateikti 3.1.2 paveiksle.

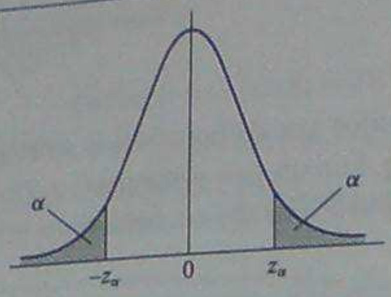
Q · dustas



3.1.2 pav. Pasikliautinųjų intervalų realizacijos

1.7.2. Normaliojo skirstinio vidurkio pasikliautinasis intervalas

Pateiksime pavyzdį, kuris atskleidžia pasikliautinųjų intervalų sudarymo principus. Tarkime, kad $X \sim \mathcal{N}(\mu, \sigma^2)$, čia μ – nežinomas vidurkis, o σ^2 – žinoma dispersija. Sukonstruosime parametro μ pasikliautinąjį intervalą. Normaliai pasiskirsčiusių nepriklausomų



3.1.3 pav. Normaliojo dėsnio tankis

atsitiktinių dydžių suma irgi yra normalusis atsitiktinis dydis. Todėl $X_1 + \cdots X_n \sim \mathcal{N}(n\mu, n\sigma^2)$, t. y.

$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0,1).$$

Tegul $0 < Q < 1$ yra pasikliovimo lygmuo. Pažymime

$$\alpha = \frac{1-Q}{2}. \tag{3.1.19}$$

Tegul z_α žymi standartinio normalaus skirstinio $1-\alpha$ lygmens kvantilį. Kadangi standartinis normalusis skirstinys yra simetriškas nulio atžvilgiu, tai (žr. 3.1.3 pav.)

$$P\left(-z_\alpha \leqslant \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leqslant z_\alpha\right) = Q, \tag{3.1.20}$$

arba

$$P\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = Q. \tag{3.1.21}$$

Palyginę (3.1.21) su pasikliautinojo intervalo apibrėžimu, nustatome $[\widehat{\mu}_1, \widehat{\mu}_2]$:

$$\widehat{\mu}_1 = \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \qquad \widehat{\mu}_2 = \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}. \tag{3.1.22}$$

Pastaba. Jeigu $Q = 0{,}9$, tai $z_\alpha = z_{0.05} = 1{,}64$. Jeigu $Q = 0{,}95$, tai $z_\alpha = z_{0{,}025} = 1{,}96$.

3.1.13 pavyzdys. Išmatuotas 49 studentų cholesterino kraujyje kiekis. Gautas $\overline{x} = 310$. Koks yra cholesterino kraujyje kiekio 95% pasikliautinasis intervalas (pasikliautinojo intervalo realizacija), jeigu $\sigma = 25$?

Sprendimas. Šiuo atveju $z_\alpha = z_{0{,}025} = 1{,}96$. Todėl (dviejų ženklų po kablelio tikslumu)

$$\widehat{\mu}_1 = 310 - 1{,}96 \cdot \frac{25}{7} = 303, \qquad \widehat{\mu}_2 = 310 + 1{,}96 \cdot \frac{25}{7} = 317.$$

Atsakymas: ieškomasis intervalas yra [303; 317].

1.7.3. Dažniausiai skaičiuojami pasikliautinieji intervalai

Dažniausiai pasitaikančių skirstinių parametrų pasikliautinieji intervalai pateikiami 3.1.7 lentelėje. Joje vartojami tokie žymenys:

Q	pasikliovimo lygmuo	z_α	standartinio normaliojo skirstinio $1-\alpha$ lygmens kvantilis
$\overline{X}$	vidurkio įvertis		
S^2	dispersijos įvertis		
$S = \sqrt{S^2}$	standartinis nuokrypis	$t_\alpha(n-1)$	Stjudento skirstinio su $(n-1)$ laisvės laipsnių $1-\alpha$ lygmens kvantilis
$\alpha = (1-Q)/2,$			
$S_n = X_1 + \cdots + X_n,$			
$S_0^2 = \Sigma\,(X_i-\mu)^2/n$		$\chi^2_\alpha(k)$	χ^2 su k laisvės laipsnių $1-\alpha$ lygmens kvantilis

Visus kvantilius galima rasti vadovėlio pabaigoje pateiktose lentelėse ($1-\alpha$ lygmens kvantilį atitinka α lygmens kritinė reikšmė).

Sudarant 3.1.7 lentelę remtasi 1.3 skyreliu. Pavyzdžiui, kai $X \sim \mathcal{N}(\mu,\sigma^2)$, čia μ ir σ^2 – nežinomi, tai pagal (3.1.6) formulę

$$\frac{\overline{X}-\mu}{\sqrt{S^2/n}} \text{ turi Stjudento skirstinį su } (n-1) \text{ laisvės laipsnių.}$$

3.1.7 lentelė. Pasikliautinieji intervalai

populiacija

Skirstinys	Pasikliautinasis intervalas
$X \sim \mathcal{N}(\mu,\sigma^2)$ μ nežinomas, σ^2 žinoma	$\widehat{\mu}_1 = \overline{X} - z_\alpha \dfrac{\sigma}{\sqrt{n}}, \quad \widehat{\mu}_2 = \overline{X} + z_\alpha \dfrac{\sigma}{\sqrt{n}}$
$X \sim \mathcal{N}(\mu,\sigma^2)$ μ žinomas, σ^2 nežinoma	$\widehat{\sigma}_1 = \dfrac{S_0^2 n}{\chi^2_\alpha(n)}, \quad \widehat{\sigma}_2 = \dfrac{S_0^2 n}{\chi^2_{1-\alpha}(n)}$
$X \sim \mathcal{N}(\mu,\sigma^2)$ μ nežinomas, σ^2 nežinoma	$\widehat{\mu}_1 = \overline{X} - t_\alpha(n-1)\dfrac{S}{\sqrt{n}}, \quad \widehat{\mu}_2 = \overline{X} + t_\alpha(n-1)\dfrac{S}{\sqrt{n}}$ $\widehat{\sigma}_1 = \dfrac{S^2(n-1)}{\chi^2_\alpha(n-1)}, \quad \widehat{\sigma}_2 = \dfrac{S^2(n-1)}{\chi^2_{1-\alpha}(n-1)}$
$X \sim \mathcal{P}(\lambda)$ λ nežinomas	$\widehat{\lambda}_1 = \dfrac{1}{2n}\chi^2_{1-\alpha}(2S_n), \quad \widehat{\lambda}_2 = \dfrac{1}{2n}\chi^2_\alpha(2S_n+2)$
$X \sim B(k,p)$ μ nežinomas, σ^2 žinoma (apytikslė formulė atvejui $S_n < (nk-1)/2$)	$\widehat{p}_1 = \dfrac{2\chi^2_{1-\alpha}(2S_n)}{2(2nk-S_n+1)+\chi^2_{1-\alpha}(2S_n)},$ $\widehat{p}_2 = \dfrac{2\chi^2_\alpha(2S_n+2)}{2(2nk-S_n)+\chi^2_\alpha(2S_n+2)}$

Todėl
$$P\left(-t_\alpha(n-1) \leqslant \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \leqslant t_\alpha(n-1)\right) = Q,$$
o tai ekvivalentu tokiai lygybei:
$$P\left(\overline{X} - \frac{S}{\sqrt{n}} t_\alpha(n-1) \leqslant \mu \leqslant \overline{X} + \frac{S}{\sqrt{n}} t_\alpha(n-1)\right) = Q.$$
Šis reiškinys sutampa su pasikliautinojo intervalo apibrėžimu.

Binominiam skirstiniui aproksimuoti galima taikyti normalųjį skirstinį. Kadangi $X_i \sim \mathcal{B}(k, p)$, tai $S_n = X_1 + \cdots + X_n \sim \mathcal{B}(nk, p)$. Pagal CRT
$$\frac{S_n - nkp}{\sqrt{nkp(1-p)}} \approx \mathcal{N}(0, 1).$$
Vardiklyje p pakeitę $\widehat{p}$ ir paėmę $\alpha = (1-Q)/2$, gauname
$$P\left(-z_\alpha \leqslant \frac{S_n - nkp}{\sqrt{nk\widehat{p}(1-\widehat{p})}} \leqslant z_\alpha\right)$$
$$= P\left(\widehat{p} - z_\alpha\sqrt{\frac{\widehat{p}(1-\widehat{p})}{nk}} \leqslant p \leqslant \widehat{p} + z_\alpha\sqrt{\frac{\widehat{p}(1-\widehat{p})}{nk}}\right) \approx Q,$$
t. y.
$$\widehat{p_1} \approx \widehat{p} - z_\alpha\sqrt{\frac{\widehat{p}(1-\widehat{p})}{nk}},$$
$$\widehat{p_2} \approx \widehat{p} + z_\alpha\sqrt{\frac{\widehat{p}(1-\widehat{p})}{nk}}, \quad \text{čia } \widehat{p} = \frac{\overline{X}}{k}. \tag{3.1.23}$$
Nors šios formulės paprastesnės negu 3.1.7 lentelėje, tačiau jos ne tokios tikslios (lentelės formulės gautos naudojant kitus – netikimybinius metodus).

Nagrinėjant 3.1.7 lentelę, į akis krinta skirtingas naudojamų laisvės laipsnių skaičius. Jis priklauso nuo to, kiek parametrų nežinome. Pavyzdžiui, keisdami nežinomą dispersiją σ^2 jos įverčiu S^2, prarandame vieną laisvės laipsnį. Iš tikrųjų į S^2 įeina reiškiniai $(X_1 - \overline{X}), (X_2 - \overline{X}), \ldots, (X_n - \overline{X})$, kurie nėra visiškai nepriklausomi, nes $(X_1 - \overline{X}) + \cdots + (X_n - \overline{X}) = 0$. Taigi laisvai parinkę $X_1 - \overline{X}, \ldots, X_{n-1} - \overline{X}$, jau negalime laisvai parinkti $X_n - \overline{X}$.

3.1.14 pavyzdys. Naujo leidinio dešimtyje puslapių rasta atitinkamai 3; 0; 2; 1; 0; 4; 3; 2; 1; 2 korektūros klaidos. Raskime vidutinio klaidų puslapyje skaičiaus 95% pasikliautinąjį intervalą.

Sprendimas. Stebime $X \sim \mathcal{P}(\lambda)$ su nežinomu λ. Tada $Q = 0{,}95$, $\alpha = 0{,}025$, $S_n = 3 + 0 + \cdots + 2 = 18$, $n = 10$. Iš lentelių randame $\chi^2_{0,975}(36) = 21{,}336$, $\chi^2_{0,025}(38) = 56{,}895$. Todėl
$$\widehat{\lambda}_1 = \frac{1}{20} \cdot 21{,}336 = 1{,}06, \qquad \widehat{\lambda}_2 = \frac{1}{20} \cdot 56{,}895 = 2{,}84.$$

3.1.15 pavyzdys. Vertybinių popierių rinkos ekspertas kiekvieną dieną nustato dvidešimties iš anksto pasirinktų vertybinių popierių buvusį dienos kainų vidurkį. Per mėnesį gauti tokie duomenys: 24; 30; 25; 23; 25; 27; 30; 24; 27; 23; 26; 23; 25; 24; 25; 26; 24; 26; 22; 24; 24; 28. Raskime dienos kainų vidurkio 90% pasikliautinąjį intervalą.

Sprendimas. Turime $\alpha = (1 - 0{,}9)/2 = 0{,}05$, $n = 21$, $\overline{x} = 25{,}0952$, $s^2 = 4{,}390$ ($s = 2{,}093$), $t_{0{,}05}(20) = 1{,}725$. Taigi:

$$\hat{\mu}_1 = 25{,}0952 - 1{,}725 \cdot \frac{2{,}1}{\sqrt{21}} = 24{,}3047,$$

$$\hat{\mu}_2 = 25{,}0952 + 1{,}725 \cdot \frac{2{,}1}{\sqrt{21}} = 25{,}8857.$$

1.8. Imties didumas

Pasikliautinajam intervalui konstruoti naudojami trys tarpusavyje susiję dydžiai: Q, n ir $\hat{\theta}_2 - \hat{\theta}_1$. Žinoma, norisi turėti kuo didesnį pasikliovimo lygmenį, tačiau tuomet labai padidėja intervalo ilgis ir sumažėja informacijos tikslumas. Pavyzdžiui, kažin ar kam reikalinga išvada, kad 100% vidutinis vyrų gyvenimo trukmės pasikliautinasis intervalas yra (0, 130) metų. Kita vertus, labai mažindami pasikliautinojo intervalo ilgį, kartu sumažiname ir pasikliovimo lygmenį. Lygmenys 0,9; 0,95 ir 0,99 yra kompromisinis sprendimas, leidžiantis išsaugoti pakankamą pasikliovimo lygį, kartu garantuojant ne per didelius intervalų ilgius. Didinant imtį, *mažinamas* pasikliautinojo intervalo ilgis. Kuo didesnis n, tuo trumpesnis pasikliautinasis intervalas.

Vienas iš standartinių statistikos uždavinių – nustatyti imties didumą, kai kiti parametrai fiksuoti. Naudojantis 3.1.7 lentele, pasikliautinojo intervalo ilgį galima išreikšti kitais parametrais. Pavyzdžiui, jei $X \sim \mathcal{N}(\mu, \sigma^2)$, μ nežinomas, o σ^2 – žinomas, tai vidurkio pasikliautinojo intervalo ilgis yra $2z_\alpha\sigma/\sqrt{n}$.

3.1.16 pavyzdys. Detalės ilgis (matuojamas mm) $X \sim \mathcal{N}(\mu, 0{,}0001)$. Tarkime, norime, kad 95% μ pasikliautinasis intervalas būtų ne ilgesnis už 0,002 mm. Kokio dydžio imties reikia?

Sprendimas. Kadangi $\sigma^2 = 0{,}0001$ ($\sigma = 0{,}01$) ir nežinome tik μ, tai iš 3.1.7 lentelės randame, kad pasikliautinojo intervalo ilgis $2z_\alpha\sigma/\sqrt{n}$, $\alpha = (1 - 0{,}95)/2 = 0{,}025$, $z_\alpha = 1{,}96$. Taigi klausimą galima suformuluoti taip: koks turi būti n, kad būtų teisinga nelygybė

$$2 \cdot 1{,}96 \cdot \frac{0{,}01}{\sqrt{n}} \leqslant 0{,}002.$$

Išsprendę gauname $n \geqslant 384{,}16$, t. y. užteks $n = 385$ matavimų.

Deja, praktiškai dispersija retai kada būna žinoma. Tuomet naudojamas jos įvertis, gautas kitais tyrimais. Nagrinėjant proporciją, šią problemą galima apeiti. Tarkime, mus domina populiacijos objektai, turintys tam tikrą savybę. Tuomet imties duomenis galima įsivaizduoti kaip vienetų ir nulių seką (1 – jei atrinktas objektas turi savybę, 0 – jei neturi). Matematiškai tai reikštų, kad stebime $X \sim B(1, p)$, čia p yra visi, turintys tą savybę populiacijos objektai (tai išplaukia iš klasikinio tikimybės apibrėžimo). Šiuo atveju taikome normaliąją aproksimaciją ir gauname, kad pasikliautinojo intervalo ilgis $l = 2z_\alpha\sqrt{\overline{X}(1 - \overline{X})/n}$. Kai visą imtį sudaro tik nuliai arba vienetai, tai $\overline{X} \leqslant 1$, todėl galima pasiremti nesunkiai įrodomu matematiniu faktu:

$$\sqrt{\overline{X}(1 - \overline{X})} \leqslant \max_{0 \leqslant y \leqslant 1} \sqrt{y(1 - y)} \leqslant 0{,}5.$$

Taigi, darydami nedidelę paklaidą, galime teigti, kad, tiriant proporciją, pasikliautinojo intervalo ilgis $l \leqslant z_\alpha/\sqrt{n}$.

3.1.17 pavyzdys. Kiek loterijos bilietų reikia patikrinti, kad ne mažesniu kaip 3% tikslumu būtų galima įvertinti laimingų loterijos bilietų procentą? Pasikliovimo lygmuo yra 95%. Q = 95%

Sprendimas. Iš visų bilietų mus domina laimingų dalis. Klausimą galima performuluoti taip: koks turi būti n, kad būtų teisinga nelygybė $|p-\widehat{p}| \leqslant 0{,}03$, t. y. $l \leqslant 0{,}06$? Šiuo atveju $\alpha = 0{,}025$, $z_\alpha = 1{,}96$. Kadangi $l \leqslant 1{,}96/\sqrt{n}$, tai pakaks tokio n, kad $1{,}96/\sqrt{n} \leqslant 0{,}06$. Iš čia

$$\sqrt{n} \geqslant 1{,}96/0{,}06, \quad \text{t. y. } n \geqslant 1067{,}111.$$

Atsakymas: užteks $n = 1068$.

Remiantis imties didumu, uždaviniuose apie proporcijas galima įvertinti procentinį paklaidos rėžį. Panagrinėkime tokį pavyzdį.

3.1.18 pavyzdys. Iš 500 Vilniaus, Kauno, Šiaulių ir Panevėžio vyresnių nei 16 metų amžiaus gyventojų 48,5% atsakė, kad atostogas planuoja praleisti kaime arba sode. (*Veidas*, 2000 06 1–7, Nr. 22). Koks šių rezultatų patikimumas, t. y. koks yra paklaidos rėžis? Apklausta tik 500 didžiųjų miestų gyventojų, todėl turime žinoti, kiek galime *suklysti* sakydami, kad 48,5% *visų* gyventojų ruošiasi atostogauti kaime arba sode.

Tarkime, X yra atostogų vietos pasirinkimas. Tegul X įgyja 1 (jei apklaustasis ruošiasi atostogauti kaime arba sode) arba 0 (priešingu atveju). Tada $X \sim \mathcal{B}(1, p)$, čia p – dalis didžiųjų miestų gyventojų, kurie ruošiasi atostogauti kaime arba sode. Turime 500 binominio kintamojo X dvireikšmių stebėjimų aibę. Kaip ir anksčiau, galime įvertinti pasikliautinojo intervalo ilgį

$$l \leqslant \frac{z_\alpha}{\sqrt{n}} = \frac{1{,}96}{\sqrt{500}} = 0{,}088.$$

Kadangi l išreiškiama vieneto dalimis, tai ilgį galime užrašyti procentais: $l = 8{,}8\%$. Taigi su 95% garantija galime sakyti, kad $p \cdot 100\%$ patenka į $[48{,}5 - 4\%; 48{,}5 + 4\%]$ intervalą (4% – procentinis paklaidos rėžis).

Atkreipiame dėmesį, kad nagrinėtu atveju paklaidos rėžio įvertis *nepriklauso* nuo populiacijos didumo. Jis priklauso tik nuo imties didumo ir gali būti įvertintas dar prieš sudarant imtį.

1.9. Prognozės intervalai

Tarkime, mus domina studentų, laikančių statistikos egzaminą, sistolinis kraujo spaudimas. Tarę, kad kraujo spaudimas turi normalųjį skirstinį, iš imties galime sukonstruoti vidurkio, t. y. vidutinio kraujo spaudimo, pasikliautinąjį intervalą (žr. 3.1.7 lentelę). Tačiau ką daryti, jeigu norime *prognozuoti* konkretaus studento galimų kraujo spaudimo reikšmių aibę? Šiuo atveju konstruojamas prognozės intervalas. Jis šiek tiek didesnis už pasikliautinąjį intervalą. Atkreipiame dėmesį, kad pasikliautinasis intervalas konstruojamas nežinomam *parametrui*, tuo tarpu prognozės intervalas konstruojamas stebimojo dydžio tikėtinai *reikšmei*.

Pateiksime prognozės intervalo pavyzdį. Tegul stebimas $X \sim \mathcal{N}(\mu, \sigma^2)$, kurio μ ir σ^2 nežinomi. Q reikšmingumo lygmens naujo stebėjimo reikšmės prognozės intervalas yra

$$\left[\overline{X} - t_\alpha(n-1)S\sqrt{1+1/n},\ \overline{X} + t_\alpha(n-1)S\sqrt{1+1/n}\right]; \tag{3.1.24}$$

čia $\alpha = (1-Q)/2$, $t_\alpha(n-1)$ – Stjudento skirstinio su $(n-1)$ laisvės laipsnių $1-\alpha$ lygmens kvantilis, n – imties didumas, $S = \sqrt{S^2}$ – standartinis kvadratinis nuokrypis.

657500000

357250000 127694000

3.1.19 pavyzdys. Nekilnojamojo turto agentūra stebėjo vieno rajono standartinių sklypų kainas. Gauta imtis 8000; 6000; 10 000; 5000; 11 500; 7500; 6500; 5900; 8700; 7000; 6500; 8000; 9500; 11 200; 10 600; 8300; 9100; 8700 Lt. Sudarykime būsimos sklypo kainos 90% prognozės intervalą.

Sprendimas. Gauname $n = 18$, $\bar{x} = 8222{,}22$, $s = 1879{,}47$, $t_{0{,}05}(17) = 1{,}74$. Įstatę šias reikšmes į (3.1.24) formulę, gauname, kad ieškomas intervalas yra [4857,9; 11586,5].

didžiausio tikėtinumo įvertis	parametrinis modelis	statistika
efektyvusis įvertis	pasikliautinasis intervalas	taškinis įvertis
momentų įvertis	pasikliovimo lygmuo	tikėtinumo funkcija
nepaslinktasis įvertis	prognozės intervalas	
paklaidos rėžis	suderintasis įvertis	

UŽDAVINIAI

$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2 \quad DX = EX^2 - (EX)^2$

1. Kompiuterinių žaidimų mėgėjas sprendžia, pirkti jam 10 žaidimų kompaktinį diskelį ar ne. Žaidimą jis vertina tik taip: geras arba blogas. Iš dešimties žaidimų jis išbando du. Diskelį pirks, jei patiks abu. Tarkime, kad žaidimų reitingai tokie (1 – geras, 0 – blogas): 1; 0; 0; 1; 1; 1; 1; 1; 1; 0.
   a) Sudarykite imties skirstinį.
   b) Raskite tikimybę, kad diskelis bus nupirktas.
   c) Tegul X yra gerų žaidimų tarp išbandytųjų dviejų skaičius. Raskite jo imties skirstinį.
2. Elektroninių žaislų gamykla teigia, kad vidutinė žaisliuko „Ameba-tabis" gyvavimo trukmė yra 60 mėn., o standartinis nuokrypis – 5 mėn. Tarkime, kad 100 šių žaisliukų buvo padovanota žymiems politikams.
   a) Koks vidutinės dovanotų žaislų gyvavimo trukmės skirstinys? (Parinkite tinkamą matematinį modelį.) $X \sim N(60, 25)$ $\bar{X} \sim N(EX, D\bar{X})$ $\bar{X} \sim N(60, 0{,}25)$
   b) Kokia tikimybė, kad vidutinė dovanotų žaislų gyvavimo trukmė neviršys 55 mėn.? $P(\bar{X} < 55) = 0$
3. Specialios pakuotės vidutinis svoris yra 8 g, o standartinis nuokrypis lygus 2 g. Klientas užsakė didelę partiją pakuočių, tačiau pirks tik tuo atveju, jeigu atsitiktinai parinktų 36 pakuočių vidutinis svoris neviršys 8,3 g. Kokia tikimybė, kad partija bus nupirkta?
4. Stebimas atsitiktinis dydis X, kurio vidurkis $\mathbf{E}X = \mu$ žinomas ir dispersija $\mathbf{D}X = \sigma^2$ nežinoma. Įrodykite, kad $S_0^2 = \sum(X_i - \mu)^2/n$ yra suderintasis ir nepaslinktasis σ^2 įvertis.
5. Atsitiktinio dydžio X tankis $p(x) = \lambda \exp\{-\lambda x\}$, $x \geqslant 0$, čia parametras $\lambda > 0$ nežinomas. Raskite λ įvertį momentų ir didžiausio tikėtinumo metodais.
6. Mokslininkas atliko dešimt serijų eksperimentų. Kiekvieną seriją sudaro 100 kartų metama moneta. Skaičiavo iškritusius herbus ir gavo: (70; 65; 67; 59; 71; 60; 70; 72; 75; 69). Didžiausio tikėtinumo metodu raskite herbo atsivertimo tikimybės įvertį.
7. Stebimas $X \sim \mathcal{N}(\mu, \sigma^2)$, čia μ – žinomas. Didžiausio tikėtinumo metodu raskite σ^2 įvertį.
8. Stebimas $X \sim \mathcal{P}(\lambda)$. Duomenys yra (5; 5; 6; 8; 10). Sukonstruokite parametro λ 95% pasikliautinąjį intervalą.

9. Stebimas $X \sim \mathcal{N}(\mu, \sigma^2)$. Duomenys yra $(-5; -4; -3; -1; -1; 0; 1; 1; 2; 3; 3)$. Sukonstruokite μ ir σ^2 95% pasikliautinuosius intervalus.

10. Stebimas $X \sim \mathcal{N}(\mu, 4)$. Imtyje 100 stebėjimų, $\widehat{\mu}_1 = 1{,}25$; $\widehat{\mu}_2 = 2{,}05$. Koks pasikliovimo lygmuo?

11. Stebimas $X \sim \mathcal{N}(\mu, 4)$. Kiek elementų turi būti imtyje, kad 95% pasikliautinojo intervalo ilgis neviršytų 0,2?

12. Vienas iš staklių išsiderinimo požymių yra gaminamų detalių skersmens pokyčiai. Leistina svyravimų norma yra $\sigma = 1$ mm. Išmatavus 31 detalės skersmenis, gauta $s = 1{,}5$ mm. Raskite skersmens svyravimų dispersijos 95% pasikliautinąjį intervalą ir nuspręskite, ar galima stakles laikyti išsiderinusiomis.

13. Edukologai tyrė dvi studentų populiacijas (atsitiktinai parinko 10 ir 20 studentų). Tirtas raštingumo lygis (testas iki 100 balų). Pirmos imties vidurkis – 80 balų (standartinis nuokrypis 6 balai), o antrosios – 60 balų (standartinis nuokrypis 10 balų). Sukonstruokite 95% procentų pasikliautinąjį intervalą ir pasakykite, ar viena populiacija statistiškai reikšmingai raštingesnė už kitą. Koks abiejų populiacijų vidutinio studentų raštingumo 95% pasikliautinasis intervalas?

14. Atsitiktinai nustatyta, kad iš 100 šalia vairuotojų sėdinčių žmonių 40 nebuvo užsisegę saugos diržų. Raskite keleivių, nesinaudojančių saugos diržais, 90% pasikliautinąjį intervalą.

15. Vienas bulvarinis laikraštis už asmens įžeidimą buvo baustas 700; 500; 1000; 2000; 2500; 900; 800 ir 1000 Lt baudomis. Raskite vidutinės baudos 90% pasikliautinąjį intervalą.

16. Alaus darykla susirūpino automatu, klijuojančiu etiketes. Daryklos vadovai nori įvertinti, koks yra kreivai priklijuotų etikečių procentas. Iš atsitiktinai parinktų 500 butelių 23 etiketės buvo priklijuotos kreivai. Nustatykite kreivų etikečių skaičiaus 99% pasikliautinąjį intervalą.

2. HIPOTEZIŲ TIKRINIMO ĮVADAS

Biologas, matematikas ir statistikas dalyvauja fotosafaryje Afrikoje. Staiga biologas sušunka: „Žiūrėkit, zebrų banda! O tenai, bandos viduryje – baltas zebras! Fantastika! Yra baltų zebrų! Mes išgarsėsime!" Matematikas: „Tiesą sakant, mes tik žinome, kad yra vienas zebras, kurio viena pusė balta." Statistikas: „Vienas zebras – statistiškai nereikšminga. Hipotezė, kad yra baltų zebrų – atmestina."

Ankstesniajame skyriuje nagrinėjome nežinomų parametrų įverčius. Parametro įverčio realizacija – tai apytikslė parametro reikšmė. Tuo tarpu dažnai prireikia populiacijos parametrą *palyginti* su konkrečiu skaičiumi arba kitos populiacijos analogišku parametru. Pavyzdžiui, norime patikrinti, ar vidutinės turgaus prekiautojo daržovėmis pajamos yra 1200 Lt per mėnesį. Tarkime, sužinojome 100 prekiautojų uždarbius ir gavome $\overline{x} = 1250$. Aišku, kad vidutinis *išrinktųjų* prekiautojų uždarbis didesnis už 1200 Lt. Tačiau išvadą mes norime padaryti apie *visą* turgaus prekiautojų populiaciją. Ar prognozuojamojo uždarbio ir imties duomenų 50 Lt skirtumas yra toks didelis, kad prognozę reikėtų atmesti? Turbūt ne. O jeigu tas skirtumas būtų 300 Lt? Turbūt taip. O jeigu 100 Lt? Kaip nuspręsti, ar teisinga prognozė, ar ne? Kadangi statistika $\overline{X}$ yra atsitiktinis dydis, viskas priklauso nuo to, kiek tikėtina, kad $\overline{X}$ įgis vieną ar kitą reikšmę. Spręsdami galime padaryti klaidą – atmesti teisingą hipotezę, kad vidutinis uždarbis yra 1200 Lt, arba priimti šią hipotezę, nors ji ir neteisinga.

Šiame skyriuje aptarsime bendriausius principus, kaip remiantis imtimi priimti sprendimą apie nežinomą populiacijos parametro reikšmę.

2.1. Sąvokos

2.1.1. Hipotezė ir alternatyva

Pateiksime keletą hipotezių apie populiacijos parametrą pavyzdžių:

1 Sporto žurnalistas nori sužinoti, ar šiuo metu vidutinis krepšininkų ūgis yra 202 cm. Pažymėję vidutinį krepšininkų ūgį μ, šią hipotezę galime užrašyti taip: $\mu = 202$.

2 Edukologas nori nustatyti, ar naujoji mokomoji programa efektyvesnė už senąją. Pagal senąją ir naująją programas mokytų vaikų vidutinius žinių tikrinimo testo rezultatus pažymėję atitinkamai μ_1 ir μ_2, šią hipotezę galime užrašyti taip: $\mu_1 < \mu_2$, arba $\mu_1 - \mu_2 < 0$.

3 Sociologas spėja, kad ne mažiau kaip 40% protestantų abejoja žmogaus kilme iš gyvūnų. Protestantų, abejojančių žmogaus kilme iš gyvūnų, dalį pažymėję raide p, šią hipotezę galime užrašyti taip: $p \geqslant 0{,}40$.

4 Politologas nori nustatyti, ar katalikiškose Rytų Europos valstybėse priklausomybė tarp vaikų skaičiaus šeimoje ir motinos išsilavinimo skiriasi nuo analogiškos priklausomybės protestantiškose Rytų Europos valstybėse. Koreliaciją tarp vaikų skaičiaus šeimoje ir motinos mokymosi trukmės (metais) katalikiškose valstybėse pažymėję ϱ_1, o protestantiškose – ϱ_2, šią hipotezę galime užrašyti taip: $\varrho_1 \neq \varrho_2$, arba $\varrho_1 - \varrho_2 \neq 0$.

Bet koks teiginys apie populiacijos parametro(ų) reikšmę(es) vadinamas *parametrine hipoteze*. Statistinę parametrinę hipotezę sudaro du alternatyvūs teiginiai apie galimas parametro reikšmes. Problema formuluojama kaip spėjimas apie galimas parametro θ reikšmes, priklausančias Θ_0, pateikiant alternatyvą, kad θ priklauso Θ_1.

$$\begin{cases} H_0\colon \theta \in \Theta_0, \\ H_1\colon \theta \in \Theta_1. \end{cases}$$

Čia H_0 – parametrinė (*nulinė*) *hipotezė*, o H_1 – *alternatyva* (*alternatyvioji hipotezė*). Vidutinį turgaus prekiautojo mėnesinį uždarbį pažymėjus μ, *statistinę* problemą galima suformuluoti taip:

$$\begin{cases} H_0\colon \mu = 1200, \\ H_1\colon \mu \neq 1200. \end{cases} \tag{3.2.1}$$

Protestantų, abejojančių žmogaus kilme iš gyvūnų, dalį visoje populiacijoje pažymėję simboliu p, gauname tokią statistinę hipotezę:

$$\begin{cases} H_0\colon p \geqslant 0{,}40, \\ H_1\colon p < 0{,}40. \end{cases} \tag{3.2.2}$$

Matome, kad šiuo atveju sociologo spėjimas virto alternatyvia hipoteze. Koks teiginys apie parametrą laikytinas nuline hipoteze, o koks alternatyva, išsamiau aptarsime antrame skyrelyje. Dabar tik pažymėsime, kad *griežtos nelygybės naudojamos tik alternatyvai* H_1. Norint pabrėžti alternatyvos svarbą, dažnai hipotezei H_0 naudojama tik lygybė:

$$\begin{cases} H_0\colon \theta = \theta_0, \\ H_1\colon \theta > \theta_0, \end{cases} \quad \text{o ne} \quad \begin{cases} H_0\colon \theta \leqslant \theta_0, \\ H_1\colon \theta > \theta_0. \end{cases} \tag{3.2.3}$$

Alternatyvos skirstomos į *dvipuses* $\theta \neq \theta_0$ ir *vienpuses* $\theta < \theta_0$ (arba $\theta > \theta_0$). Kurią iš vienpusių alternatyvų pasirinkti, lemia tiriamoji problema. Pavyzdžiui, tirdamas, ar produktas nesuges per du mėnesius, gamintojas domisi, ar jis nesuges anksčiau. Kuo ilgiau produktas nesuges, tuo gamintojas bus labiau patenkintas.

2.1.2. Pirmosios ir antrosios rūšies klaidos bei reikšmingumo lygmuo

Priimdami arba atmesdami hipotezę H_0, galime padaryti dviejų rūšių klaidas. Tradiciškai jos vadinamos *pirmosios* ir *antrosios* rūšies klaidomis.

> *Pirmosios rūšies klaida:* H_0 atmetame, kai ji teisinga.
> *Antrosios rūšies klaida:* H_0 priimame, kai ji klaidinga.

3.2.1 lentelė

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

Taisyklė, pagal kurią iš imties rezultatų darome išvadą apie hipotezės teisingumą ar klaidingumą, vadinama *statistiniu kriterijumi*. Galimos statistinio kriterijaus taikymo baigtys pateiktos 3.2.1 lentelėje.

Kaip sudaromas statistinis kriterijus? Aišku, kad kriterijus yra tuo geresnis, kuo mažesnės abiejų rūšių klaidų tikimybės. Dėl imties atsitiktinumo praktiškai neįmanoma sudaryti kriterijaus, kad abiejų klaidų tikimybės būtų lygios nuliui. Dažniausiai parenkamas mažas teigiamas skaičius α ir nagrinėjami tik tokie kriterijai, kurių pirmosios rūšies klaidos tikimybė lygi α (α vadinamas reikšmingumo lygmeniu).

Kriterijaus *reikšmingumo lygmuo* $\alpha = P(H_0 \text{ atmetame} \mid H_0 \text{ teisinga})$.

Tradicinis nematematinis pavyzdys, motyvuojantis pirmosios rūšies klaidos tikimybės fiksavimą, yra toks: teismas sprendžia, ar teisiamasis kaltas. Hipotezė H_0 – nekaltas (nekaltumo prezumpcija), alternatyva H_1 – kaltas. Pirmosios rūšies klaida – nekaltą pripažinti kaltu. Antrosios rūšies klaida – kaltą išteisinti. Manoma, kad pirmosios rūšies klaida pavojingesnė (beje, žinoma ir kitokių požiūrių). Todėl stengiamasi pirmosios rūšies klaidos tikimybę fiksuoti, t.y. imti ją pakankamai mažą.

Nesunku analogiškai paaiškinti situaciją, kai dėstytojas įtaria (arba ne) studentą per egzaminą nusirašinėjant.

Antrosios rūšies klaida

Tradiciniai reikšmingumo lygmenys yra $\alpha = 0{,}1$; $\alpha = 0{,}05$ ir $\alpha = 0{,}01$. Pavyzdžiui, tegul naudojamas statistinis kriterijus, kurio reikšmingumo lygmuo $\alpha = 0{,}05$. Vadinasi, daug kartų jį taikydami, vidutiniškai penkis kartus iš šimto teisingą hipotezę atmesime.

Atrodo, kad populiarųjį reikšmingumo lygmenį $\alpha = 0{,}05$ pirmasis pradėjo naudoti R. A. Fišeris. Tai įvyko ketvirtajame XX amžiaus dešimtmetyje. Pasirinkti vienokį ar kitokį reikšmingumo lygmenį – susitarimo reikalas. Todėl nereikia suabsoliutinti 0,05 reikšmingumo lygmens svarbos. Jis tik parodo mūsų pasirinktą teisės suklysti laipsnį.

2.1.3. *Kritinė sritis ir kritinė reikšmė*

Kaip sudaromas statistinis kriterijus? Turint konkrečius duomenis, visada galima apskaičiuoti parametro θ įverčio $\hat{\theta}$ realizaciją ir patikrinti, ar ji pateko į aibę Θ_0. Tačiau dėl imties atsitiktinės prigimties $\hat{\theta}$ yra atsitiktinis dydis. Todėl turime atsakyti į klausimą: jeigu įverčio $\hat{\theta}$ realizacija nepatenka į aibę Θ, ar tikėtina, kad taip įvyksta dėl imties atsitiktinumo, ar ne.

Sprendimui priimti dažniausiai naudojama pati statistika $\hat{\theta}$ arba kokia nors jos transformacija. Pažymėkime ją simboliu $T(X_1, X_2, \ldots, X_n)$. Statistika $T(X_1, X_2, \ldots, X_n)$ parenkama taip, kad turėtų žinomą skirstinį, kai H_0 teisinga.

Priimti ar atmesti hipotezę, sprendžiama atsižvelgus į T realizaciją. Jeigu T realiz… patenka į skaičių aibę W, tenkinančią tam tikras sąlygas, hipotezė H_0 atmetama (… alternatyva). Priešingu atveju hipotezė neatmetama. Aibė W vadinama *kritine sritimi*.

Pavyzdžiui, imkime (3.2.1) hipotezę apie vidutinį turgaus prekiautojų uždarbį. S… tistika, įvertinanti vidutinį uždarbį, yra $\overline{X}$. Norint nuspręsti, priimti ar atmesti H_0, r… taip parinkti du skaičius C_1 ir C_2, kad situacijos $\overline{X} < C_1$ arba $\overline{X} > C_2$, galiojant hipotezei H_0: $\mu = 1200$, būtų labai mažai tikėtinos. Tuomet, jeigu imties realizacijos $\overline{x} < C_1$ arba $\overline{x} > C_2$, tai H_0 atmetame. Kritinę sritį sudaro $W = \{(-\infty, C_1) \cup (C_2, \infty)\}$. Rėžių C_1, C_2 parinkimas priklauso ir nuo matuojamojo kintamojo skirstinio, ir nuo imties didumo.

Kai kurioms statistikoms (pvz., $T = (X_1, X_2)$) kritinė sritis gali turėti gana sudėtingą struktūrą (pvz., būti plokštumos poaibiu). Tačiau dažniausiai (kaip ir pavyzdžio apie uždarbį atveju) kritinė sritis W yra intervalas arba intervalų sąjunga $W = (-\infty, C_1) \cup (C_2, \infty)$, $W = (-\infty, C_1)$ ir pan. Skaičiai $C_1, C_2, \ldots$, kurie atskiria kritinę sritį nuo hipotezės neatmetimo srities, vadinami *kritinėmis reikšmėmis*.

Kritinės reikšmės išreiškiamos atitinkamų skirstinių kvantiliais.

> α lygmens kritinė reikšmė yra lygi $1 - \alpha$ kvantiliui.

Statistinėms išvadoms dažniausiai naudojamų skirstinių kritinių reikšmių lentelės pateiktos šio vadovėlio priede. Pavyzdžiui, jose galima rasti Stjudento skirstinio su 3… laisvės laipsnių 0,05 lygmens kritinę reikšmę, Fišerio skirstinio su 3 ir 9 laisvės laipsniais 0,025 lygmens kritinę reikšmę ir pan.

Jeigu statistika T yra absoliučiai tolydus atsitiktinis dydis, tai kritinę sritį W ir reikšmingumo lygmenį α sieja tokia priklausomybė:

$$\alpha = P(T \in W), \quad \theta \in \Theta_0. \tag{3.2.4}$$

Jeigu T diskretusis atsitiktinis dydis (šiame vadovėlyje taip bus retai), tai $\alpha \geqslant P(T \in W)$, $\theta \in \Theta_0$, t.y. sritis W parenkama taip, kad pirmosios rūšies klaidos tikimybė būtų kiek galima arčiau α, bet jo neviršytų.

Taigi hipotezės tikrinimo taisyklės yra tokios:

> Parenkamas reikšmingumo lygmuo α ir jam sudaroma kritinė sritis W.
> Jeigu $T(x_1, x_2, \ldots, x_n) \in W$, hipotezė H_0 *atmetama*.
> Jeigu $T(x_1, x_2, \ldots, x_n) \notin W$, hipotezė H_0 *neatmetama*.
> Pirmosios rūšies klaida padaryta, jeigu $T \in W$, o $\theta \in \Theta_0$.
> Antrosios rūšies klaida padaryta, jeigu $T \notin W$, o $\theta \notin \Theta_0$.

Atkreipiame dėmesį, kad tikrinant hipotezę *iš pradžių* pasirenkamas reikšmingumo lygmuo α, o *po to* parenkama kritinė sritis W, tenkinanti (3.2.4). Be abejo, toks parinkimas nėra vienintelis. Kurią iš galimų kritinių sričių naudoti, lemia mažesnė antrosios rūšies klaida.

2.1.4. Kriterijaus galia

Dažniausiai skaičiuojama ne antrosios rūšies klaidos tikimybė, o jai priešingo įvykio tikimybė – *kriterijaus galia*.

Kriterijaus galia β – tai tikimybė atmesti hipotezę H_0, kai ji klaidinga:
$$\beta = P(H_0 \text{ atmetama} \mid H_0 \text{ klaidinga}) = P(T \in W) \mid \theta \notin \Theta_0).$$

Taigi $\beta = 1 - P$ (antros rūšies klaida). Kriterijaus galia leidžia palyginti du kriterijus, turinčius tą patį reikšmingumo lygmenį α ir taikomus to paties didumo imtims. *Galingesnis* kriterijus yra tas, kurio β didesnis.

Norėdami geriau išsiaiškinti reikšmingumo lygmens α ir kriterijaus galios β ryšį, panagrinėsime paprastą pavyzdį. Tarkime, matuojamojo kintamojo vidurkis gali būti tik 10 arba 16. Tikriname hipotezę, kad vidurkis lygus 10. Pasirinkime penkių stebėjimų imtį ir tarkime, kad statistika $\boxed{T = \bar{X}}$ turi normalųjį skirstinį su žinoma dispersija (pvz., $\boxed{\text{lygia } 4}$). Suformuluojame statistinį uždavinį:

$$\begin{cases} H_0\colon \ \mu = 10, \\ H_1\colon \ \mu = 16. \end{cases}$$

$W=(13,3;\infty)$

Tegul $\alpha = 0{,}05$. Parenkame kritinę sritį $W = (13{,}3; \infty)$. Jeigu H_0 teisinga, tai $T \sim \mathcal{N}(10, 4)$ ir $P(T \in W) = P(T > 13{,}3) = 0{,}05$ (žr. 3.2.1 pav.). Taigi jeigu konkrečiai imties realizacijai statistikos reikšmė $T(x_1, \ldots, x_n) > 13{,}3$, tai hipotezę H_0 atmetame. Kokia šiuo atveju yra antrosios rūšies klaidos tikimybė? Kai H_0 klaidinga, tai $T \sim \mathcal{N}(16, 4)$. Tikimybė, kad tuomet H_0 priimame, lygi

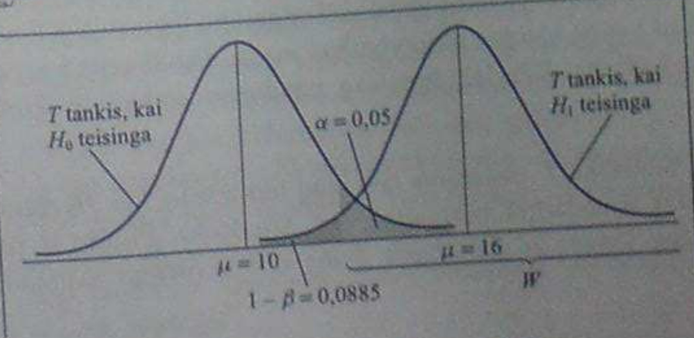
$$P(T \leqslant 13{,}3) = \Phi((13{,}3 - 16)/2) = 0{,}0885.$$

Kriterijaus galia $\beta = 1 - 0{,}0885 = 0{,}9115$. Mes parinkome kritinę sritį $W = (13{,}3; \infty)$. Kadangi vienintelis reikalavimas kritinei sričiai yra $P(T \in W) = 0{,}05$, kai H_0 teisinga, tai akivaizdu, kad kritinių sričių gali būti be galo daug. Tačiau iš 3.2.1 paveikslo matyti, kad bet kokiai kitai kritinei sričiai β yra mažesnis. Taigi parinkome *galingiausią* kriterijų.

Iš 3.2.1 paveikslo matyti, kad mažinant reikšmingumo lygmenį α (pirmosios rūšies klaidos tikimybę) kartu mažėja ir kriterijaus galia β (antrosios rūšies klaidos tikimybė didėja). Be to, kriterijaus galia priklauso ir nuo atstumo tarp nulinės hipotezės ir alternatyvos vidurkių reikšmių. Pavyzdžiui, pasirinkus alternatyvą $\mu = 17$, β būtų didesnis.

Mažėjant tikimybei pirmos rūšies klaidos, vis auga tikimybė klaidos rūšies antros.
Kurią daryti klaidą? Štai klausimas keblus! Tik testas man padės. Jei(gu) galingas bus.

MaDi **18**, 1994 04 14



3.2.1 pav. Pirmosios ir antrosios rūšies klaidų tikimybės

Nagrinėjome pavyzdį, kai imtis penkių stebėjimų. Paminėtina, kad didinant imtį *kriterijaus galia* β dažniausiai *didėja*. Taip atsitinka todėl, kad imčiai didėjant sprendimui naudojamos statistikos dispersija dažniausiai mažėja. Nagrinėto pavyzdžio atveju tankiai tampa „smailesni“, labiau „atsiskyrę“. Taip pat pravartu žinoti alternatyvos kryptį – turint konkrečias n, α ir statistikos T dispersijos reikšmes, kriterijus su vienpuse alternatyva vidurkio reikšmei yra galingesnis už kriterijų su dvipuse alternatyva.

Šiame skyrelyje kalbėjome apie *parametrines* hipotezes. Yra ir *neparametrinių* hipotezių, apibūdinančių tiriamo atsitiktinio dydžio skirstinį, o ne konkrečių parametrų reikšmes. Su neparametrinėmis hipotezėmis dar susidursime ateityje.

2.2. Parametrinio statistinio kriterijaus sudarymo ir taikymo etapai

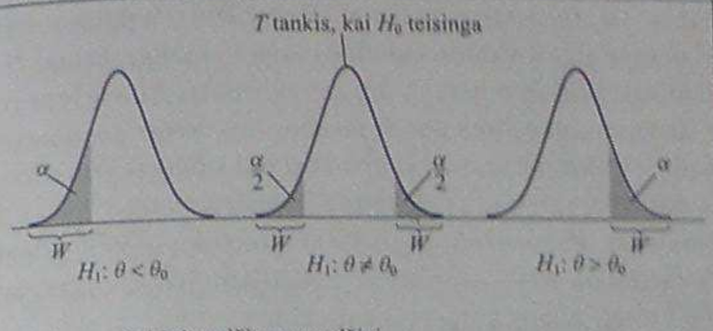
Smulkiau aptarsime parametrinio statistinio kriterijaus sudarymo ir taikymo etapus.

1 *Uždavinio formulavimas.* Tai – ne statistiko uždavinys. Pavyzdžiui: vaistai A yra efektyvesni už vaistus B, kompiuterius naudojančių žmonių regėjimas blogesnis nei jų nenaudojančių, paauglių seksualinis gyvenimas ir erotinių žurnalų skaitymas susiję reiškiniai, vidutinis akcijų kainų svyravimas neviršija 15% jų nominaliosios vertės ir pan.

2 *Tikimybinio modelio parinkimas.* Tai esminis etapas, kuriuo nustatoma, su kokios rūšies atsitiktinumu susiduriame, t. y. koks yra matuojamojo kintamojo tikimybinis skirstinys. Šis skirstinys turi vieną arba kelis nežinomus parametrus. Pasirinkdami tinkamą tikimybinį skirstinį, sąmoningai ignoruojame neatitikimus, kurių tikimybės labai mažos. Pavyzdžiui, nusprendžiame, kad žmogaus svoris turi normalųjį skirstinį, nors iš tikro normalusis skirstinys gali įgyti ir neigiamas reikšmes.

3 *Statistinės hipotezės užrašymas.* Užrašoma hipotezė H_0 ir jos alternatyva H_1. Tradiciškai hipotezė H_0 reiškia, kad nėra skirtumo, o alternatyva H_1 – skirtumas yra. Hipotezės H_0 atmetimas primena įrodymą prieštaros būdu: hipotezę H_0 laikome teisinga ir atmetame ją tik tuomet, kai duomenys rodo, kad ji labai mažai tikėtina. Ką parinkti hipoteze, lemia pirmosios ir antrosios rūšies klaidos. Tarkime, kad norima įsitikinti, ar naujus vaistus gali vartoti nėščios moterys. Šiuo atveju H_0 laikytinas teiginys, kad vaistų vartoti negalima, o alternatyva H_1 – galima. Kodėl? Galime padaryti dvi klaidas: leisti vartoti kenksmingus vaistus (pirmosios rūšies klaida) arba uždrausti vartoti nekenksmingus (antrosios rūšies klaida). Aišku, kad pirmosios rūšies klaida pavojingesnė. Tačiau jos tikimybę (reikšmingumo lygmenį) galime fiksuoti, t. y. užtikrinti, kad tokios klaidos tikimybė bus maža. Pateiktame pavyzdyje hipotezes formulavome žodžiais. Užrašant hipotezę, naudojami stebimojo skirstinio parametrai.

4 *Statistikos parinkimas.* Statistika T sudaroma taip, kad tuo atveju, kai H_0 teisinga, ji turėtų žinomą skirstinį (standartinį normalųjį, χ^2, Stjudento, Fišerio ir pan.). Parenkant statistiką, daug lemia, ar stebimasis dydis turi normalųjį skirstinį, ar neturi.



3.2.2 pav. Kritinių sričių pavyzdžiai

5 *Kritinės srities parinkimas.* Formuluojant hipotezę, kartu pasirenkamas reikšmingumo lygmuo α. Šis pasirinkimas – *subjektyvus* dalykas. Pagal nusistovėjusią tyrimų tradiciją $\alpha = 0{,}05$ arba $\alpha = 0{,}01$. Jeigu H_0 teisinga, tai $T(X_1, \ldots, X_n)$ turi skirstinį, kurio visi parametrai žinomi. Kritinė sritis W parenkama taip, kad, esant teisingai H_0, T pakliuvimo į W tikimybė būtų lygi pasirinktajam α (tolydžiuoju atveju, diskrečiuoju – kuo arčiau α, bet nedidesnė). Kritinė sritis priklauso nuo n (imties dydžio), α (reikšmingumo lygmens), θ_0. Be to, kritinė sritis W parenkama taip, kad kriterijaus galia β būtų didžiausia. Kritinių sričių parinkimo pavyzdžiai pateikti 3.2.2 paveiksle.

6 *Kriterijaus taikymas.* Turėdami konkrečius duomenis, skaičiuojame statistikos $T(X_1, X_2, \ldots, X_n)$ realizaciją $T(x_1, x_2, \ldots, x_n)$. Jeigu ji patenka į kritinę sritį, t. y. $T(x_1, x_2, \ldots, x_n) \in W$, tai H_0 atmetame, priešingu atveju H_0 neatmetame.

7 *Išvadų formulavimas.* Formuluodami išvadas, sakome „statistiškai reikšmingai skiriasi“. Vadinasi, parametro įverčio realizacijos ir spėjamos parametro reikšmės θ_0 skirtumas didelis ir labai mažai tikėtina, kad to priežastis yra imties atsitiktinumas. Pavyzdžiui, norime patikrinti hipotezę, kad vidutinis televizoriaus veikimo iki pirmo gedimo laikas yra 30 mėnesių, o vidutinis 120 stebėtų televizorių veikimo iki pirmo gedimo laikas buvo 29 mėnesiai. Mes nenorime daryti išvados apie *šių* 120 televizorių veikimo laiką. Išvadą norime padaryti apie *visus* tos rūšies televizorius. Jeigu imtis tokia, kad H_0 atmetama, tai sakome, kad vidutinis televizorių veikimo iki gedimo pradžios laikas *statistiškai reikšmingai* skiriasi nuo 30 mėnesių. Tuo tiktai konstatuojame, kad spėjamo (30 mėn.) ir gautojo (29 mėn.) vidurkių skirtumas toks didelis, jog jo beveik negalima paaiškinti imties atsitiktinumu (120 normaliojo atsitiktinio dydžio matavimų). Jeigu imtis tokia, kad H_0 neatmetame, tai sakome, kad gauto ir spėjamo vidurkių skirtumas *statistiškai nereikšmingas*. Mes neteigiame, kad *imties* vidurkis ir spėjamasis nesiskiria; sakome, kad nėra pagrindo manyti, jog *populiacijos* vidurkis skiriasi nuo spėjamojo, o imties ir spėjamojo vidurkių skirtumą galima paaiškinti atsitiktinumu. Primename, kad statistinės išvados daromos su tam tikra tikimybe. Tam tarnauja ir pasirenkamas reikšmingumo lygmuo α. Jeigu H_0 neatmetame, tai dažniausiai išvada taip ir formuluojama: „H_0 atmesti nėra pagrindo“ (tarytum rekomenduojama susilaikyti nuo galutinio sprendimo), o ne „H_0 teisinga“.

Reikia skirti *statistiškai reikšmingą* skirtumą nuo tyrimo prasme *reikšmingo* skirtumo.

Jeigu matavimų skaičius mažas, tai H_0 retai atmetama. Pavyzdžiui, tarkime, ištyrus 5 privačių ir 5 valstybinių darbuotojų algas, gautos vidutinės algos yra atitinkamai 2000 ir 1000 Lt, bet *statistiškai* reikšmingo skirtumo nėra. Nerastas skirtumas šiuo atveju yra per mažų imčių pasekmė – iš 5 stebėjimų negalima daryti išvados apie *visas* populiacijas.

Galima ir priešinga situacija – didelėms imtims *statistiškai* reikšmingas pripažintas net menkiausias skirtumas (tokia statistikų skirstinių prigimtis). Tačiau nebūtinai šis skirtumas iš tikro reikšmingas (tyrimo prasme). Pavyzdžiui, jeigu 2000 aštuntokų ir 3000 aštuntokių amžiaus vidurkiai skiriasi 3 dienomis ir šis skirtumas statistiškai reikšmingas, tai dar nereiškia, kad tos 3 dienos gali turėti įtakos kitiems rezultatams.

Tyrimo išvados hipotezė ir statistinė hipotezė – ne tas pats. Gautą statistinį rezultatą (išvadą apie statistinę hipotezę) reikia performuluoti tiriamam uždaviniui. Pavyzdžiui, sociologas nori nustatyti, ar yra ryšys tarp smurtinių laidų žiūrėjimo (žiūri, nežiūri) ir elgesio agresyvumo (įvertinto balais). Tiriamoji hipotezė: ryšys yra. Statistinė hipotezė H_0: vidutinis žiūrinčiųjų agresyvumas sutampa su vidutiniu nežiūrinčiųjų agresyvumu. Alternatyva H_1: vidutinis žiūrinčiųjų agresyvumas didesnis už vidutinį nežiūrinčiųjų agresyvumą. Taigi tiriamoji hipotezė yra statistinės hipotezės alternatyva. Statistinė išvada gali būti tokia: H_0 atmestina, abiejų grupių vidurkiai statistiškai reikšmingai skiriasi. Tyrimo išvada – reiškiniai susiję.

Statistikas visada teisus! Jis nieko negarantuoja 100%.

Trečioje dalyje visuomet formuluosime 3–7 tyrimo etapus. Pasirinkti modelį ir formuluoti tyrimo išvadas yra kiekvieno vartotojo konkrečiu atveju prerogatyva, tačiau tikimės, kad pateikti pavyzdžiai šį procesą palengvins.

3.2.1 pavyzdys. Daugiametė statistikos dėstymo patirtis dėstytojui G. Murauskui leido suformuluoti tokį pedagoginio neišvengiamumo dėsnį.

Murausko dėsnis: *Dėstyk kaip nori, – vis tiek ne mažiau kaip kas penktas studentas nieko nesupras.*

Dėstytojas V. Čekanavičius egzamino metu įsitikino, kad iš 30 studentų 5 nieko nesuprato. Ar gauti duomenys neprieštarauja Murausko dėsniui?

Parinksime tikimybinį modelį. Į egzaminą galima pažiūrėti kaip į Bernulio schemos realizavimą – kiekvienas iš 30 studentų su tikimybe p kažką supranta, o su tikimybe $(1 - p)$ nieko nesupranta. Taigi 30 kartų stebime kintamąjį $X \sim \mathcal{B}(1, p)$, o p nežinome.

Suformuluosime statistinį uždavinį. Mus domina, ar imties rezultatas $(5/30 = 1/6)$ paneigia Murausko dėsnį. Taigi:

$$\begin{cases} H_0\colon\ p = 1/5, \\ H_1\colon\ p < 1/5. \end{cases}$$

Tegul T yra neišlaikiusių studentų skaičius. Tuomet $T \sim \mathcal{B}(30, p)$. Imkime reikšmingumo lygmenį $\alpha = 0{,}05$. Kritinė sritis W parenkama taip: padaroma prielaida, kad $p = 1/5$ ir surandama aibė W tokia, kad $P(X \in W) \leqslant 0{,}05$. Turimoje imtyje statistikos T realizacija yra 5. Jeigu $5 \in W$, tai hipotezė H_0 atmestina, nes labai mažai tikėtina, kad $T \sim \mathcal{B}(30, 1/5)$ tokią reikšmę įgytų. Jeigu $5 \notin W$, tai tarsime, kad duomenys neleidžia paneigti Murausko dėsnio.

Rasti kritinę sritį W gana sudėtinga. Dažniausiai statistikai T taikoma normalioji aproksimacija (remiamasi CRT). Išsamiau šios procedūros neaptarinėsime, nes toks kriterijus smulkiai nagrinėjamas kitame skyriuje. Ten pat bus atsakyta į 3.2.1 pavyzdžio klausimą.

2.3. Reikšmingumo lygmuo ir p-reikšmė

Tarkime, tikriname hipotezę su vienpuse alternatyva:

$$\begin{cases} H_0\colon \theta = \theta_0, \\ H_1\colon \theta > \theta_0. \end{cases}$$

Tegul sprendimui naudojama statistika $T(X_1, X_2, \ldots, X_n)$ yra tolydi. Tarkime, t^* yra statistikos T realizacija. vektoriu.

> Tikimybė, kad kriterijaus statistika T (tuo atveju, kai H_0 teisinga) ne mažesnė už stebimą realizaciją t^*, vadinama *p-reikšme*.
>
> $p = P(T \geqslant t^*)$, kai H_0 teisinga.

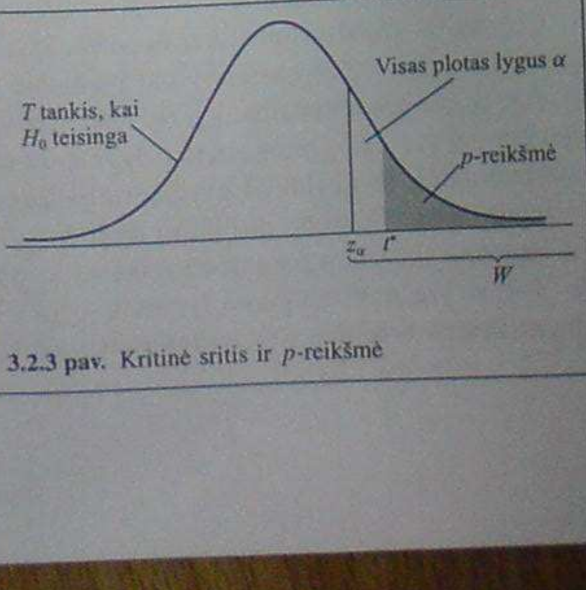
Žinoma, jeigu alternatyva $\theta < \theta_0$ arba $\theta \neq \theta_0$, tai keičiasi ir p-reikšmės apibrėžimas. Pateikėme apibrėžimą tik pačiu paprasčiausiu atveju (žr. 3.2.3 pav.). Kiekvienu atveju p-reikšmė rodo statistikos T tikimybę atsidurti nuo parametro taip toli kaip t^* arba dar toliau (esant prielaidai, kad H_0 teisinga).

Kartais p-reikšmės vadinamos skirstinio *uodegos tikimybėmis*. Tokią p-reikšmę galima panaudoti ir priimant arba atmetant hipotezę. Jeigu (žr. 3.2.3 pav.) skaičius t^* patenka į kritinę sritį W (hipotezę H_0 atmetame), tai $p < \alpha$. Ir atvirkščiai, jeigu t^* nepatektų į W (t. y. t^* būtų į kairę nuo z_α, o H_0 neatmestume), tai $p \geqslant \alpha$. Šis dėsningumas išlieka ir dvipusės alternatyvos atveju. Tuomet tereiktų lyginti dešinės (jei t^* būtų dešinėje) uodegos tikimybę su $\alpha/2$ arba (jei t^* būtų kairėje) kairiosios pusės tikimybę su $\alpha/2$. Dėl patogumo sutarta, kad *dvipusės alternatyvos atveju p-reikšmė apima abi* uodegos tikimybes. Taigi galima suformuluoti bendrą taisyklę, tinkančią visų rūšių nulinėms hipotezėms H_0 bei alternatyvoms H_1.

> Tegul α yra reikšmingumo lygmuo, o p – p-reikšmė.
> Jeigu $p < \alpha$, tai hipotezė H_0 atmetama.
> Jeigu $p \geqslant \alpha$, tai hipotezė H_0 neatmetama.

Pavyzdžiui, jeigu $\alpha = 0{,}05$, o $p = 0{,}015$, tai H_0 atmetama.

Patogiau yra naudotis p-reikšmėmis nei W, nes jos nėra susietos su konkrečiais reikšmingumo lygmenimis. Pavyzdžiui, $p = 0{,}002$ galima lyginti ir su $\alpha = 0{,}05$, ir su $\alpha = 0{,}01$. Tačiau skaičiuoti p-reikšmes pakankamai sudėtinga, todėl paprastai jos skaičiuojamos tik kompiuteriais. Beveik visi statistikos paketai (iš jų ir SPSS) skaičiuoja



3.2.3 pav. Kritinė sritis ir p-reikšmė

p-reikšmes. Dažniausiai jos naudojamos ir statistinėms išvadoms formuluoti – „vidurkis statistiškai reikšmingai skiriasi nuo 1000 ($p < 0{,}01$)“ ir pan.

p-reikšmė dar vadinama *stebimuoju reikšmingumo lygmeniu*, nes tai mažiausias reikšmingumo lygmuo, su kuriuo teisinga H_0 gali būti atmesta *turimiems* duomenims.

2.4. Parametrinių hipotezių ryšys su pasikliautinaisiais intervalais

Pasikliautinieji intervalai glaudžiai susiję su parametrinėmis hipotezėmis. Tarkime, turime parametrinę hipotezę:

$$\begin{cases} H_0\colon\ \theta = \theta_0, \\ H_1\colon\ \theta \neq \theta_0. \end{cases}$$

Tegul naudojamos statistikos T išraiška yra tokia:

$$T = \frac{\widehat{\theta} - \theta_0}{\sqrt{\mathbf{D}\widehat{\theta}}}, \tag{3.2.5}$$

čia $\widehat{\theta}$ yra parametro θ įvertis ir yra tolydusis simetrinis atsitiktinis dydis. Pasirinkime reikšmingumo lygmenį $\alpha = 0{,}05$. Sudarydami kritinę sritį W, randame tokį skaičių $z_{0{,}025}$, kad

$$P(T < -z_{0{,}025}) = 0{,}025, \qquad P(T > z_{0{,}025}) = 0{,}025. \tag{3.2.6}$$

Iš čia randame

$$P(-z_{0{,}025} \leqslant T \leqslant z_{0{,}025}) = 1 - P(T < -z_{0{,}025}) - P(T > z_{0{,}025}) = 0{,}95.$$

Įstatę T apibrėžimą (3.2.5), gauname

$$P\Big(- z_{0{,}025}\sqrt{\mathbf{D}\widehat{\theta}} \leqslant \theta_0 \leqslant z_{0{,}025}\sqrt{\mathbf{D}\widehat{\theta}}\Big) = 0{,}95,$$

o tai yra ne kas kita, kaip pasikliautinojo intervalo apibrėžimas.

Analogiškas ryšys tarp pasikliautinojo intervalo ir *dvipusės* alternatyvos išlieka ir bendruoju atveju – jeigu reikšmingumo lygmuo lygus α, tai hipotezėje naudojamai statistikai T sudarome $(1 - \alpha)$ pasikliautinąjį intervalą. Imties realizacijai randame pasikliautinojo intervalo realizaciją. Jeigu θ_0 *nepatenka* į šį intervalą, tai H_0 atmetame. Priešingu atveju H_0 neatmetame.

Jeigu alternatyva vienpusė, t. y. $H_1\colon\ \theta > \theta_0$, tai H_0 atmetama, kai $\theta_0 > \theta_2^*$, čia θ_2^* yra viršutinis statistikos T pasikliautinojo intervalo su $(1 - 2\alpha)$ pasikliovimo lygmeniu rėžis.

Jeigu alternatyva vienpusė, t. y. $H_1\colon\ \theta < \theta_0$, tai H_0 atmetama, kai $\theta_0 < \theta_1^*$, čia θ_1^* yra apatinis statistikos T pasikliautinojo intervalo su $(1 - 2\alpha)$ pasikliovimo lygmeniu rėžis.

Taigi ar naudojant pasikliautinuosius intervalus, ar tiesiogiai tikrinant hipotezes, vis tiek reikia žinoti statistikos T skirstinį. Tegul turime Puasono atsitiktinį dydį $X \sim \mathcal{P}(\lambda)$. Tarkime, norime patikrinti hipotezę apie Puasono skirstinio parametro reikšmę λ. Sprendimo taisyklės pateiktos 3.2.2 lentelėje. Joje α yra reikšmingumo lygmuo, $\chi^2_\alpha(S)$ – χ^2 skirstinio su S laisvės laipsniais α lygmens kritinė reikšmė, S – imties stebėjimų suma, t. y. $S = x_1 + x_2 + \cdots + x_n$.

3.2.2 lentelė. H_0: $\lambda = \lambda_0$ Puasono dėsniui $X \sim \mathcal{P}(\lambda)$

Alternatyva	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\lambda \neq \lambda_0$	$\lambda_0 < (2n)^{-1}\chi^2_{1-\alpha/2}(2S)$ arba $\lambda_0 > (2n)^{-1}\chi^2_{\alpha/2}(2S+2)$	$(2n)^{-1}\chi^2_{1-\alpha/2}(2S) \leqslant \lambda_0 \leqslant (2n)^{-1}\chi^2_{\alpha/2}(2S+2)$
$\lambda > \lambda_0$	$\lambda_0 < (2n)^{-1}\chi^2_{1-\alpha}(2S)$	$\lambda_0 \geqslant (2n)^{-1}\chi^2_{1-\alpha}(2S)$
$\lambda < \lambda_0$	$\lambda_0 > (2n)^{-1}\chi^2_{\alpha}(2S+2)$	$\lambda_0 \leqslant (2n)^{-1}\chi^2_{\alpha}(2S+2)$

3.2.2 pavyzdys. Receptūroje nurodyta, kad vidutinis razinų bandelėje skaičius turi būti ne mažesnis kaip 5. Patikrinus dešimt bandelių, rasta (5; 4; 6; 5; 3; 6; 7; 2; 3; 1) razinų. Ar esant 0,05 reikšmingumo lygmeniui galima teigti, kad receptūros reikalavimai pažeisti?

Sprendimas. Iš tikimybių teorijos žinome, kad razinų bandelėje skaičius X turi Puasono skirstinį $X \sim \mathcal{P}(\lambda)$. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \lambda \geqslant 5, \\ H_1\colon \lambda < 5. \end{cases}$$

Iš knygos pabaigoje pateiktų lentelių randame

$$\chi^2_{0,05}(2 \times 42 + 2) = \chi^2_{0,05}(86) = 108{,}648.$$

Kadangi $n = 10$, tai $108{,}648/20 = 5{,}4324$. Matome, kad 5 yra mažesnis už šią reikšmę. Taigi hipotezė H_0 neatmetama.

alternatyva	kriterijaus galia	pirmosios rūšies klaida
antrosios rūšies klaida	kritinė sritis	reikšmingumo lygmuo
hipotezė	p-reikšmė	statistinis kriterijus

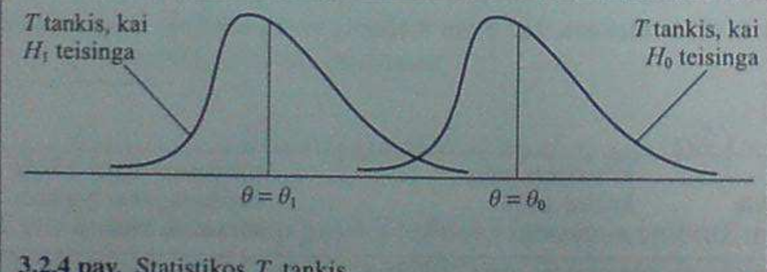
UŽDAVINIAI

1. Kaip pasikeis kriterijaus galia, jeigu vietoje reikšmingumo lygmens $\alpha = 0{,}05$ paimsime $\alpha = 0{,}01$? (Imties didumas fiksuotas.)
2. Kaip reikšmingumo lygmuo α susijęs su pirmosios rūšies klaida? Kaip kriterijaus galia β susijusi su antrosios rūšies klaida?
3. Tegul H_0: $\mu_X = 100$, H_1: $\mu_X \neq 100$, o imtyje yra n stebėjimų. Kada kriterijaus galia β didesnė: kai *tikroji* μ_X reikšmė yra 90 ar kai 75? Kodėl?
4. Kiekvienai hipotezių porai nustatykite nulinę hipotezę ir alternatyvą:

a) $\begin{cases} A\colon \mu > 21, \\ B\colon \mu \leqslant 21, \end{cases}$ b) $\begin{cases} A\colon p = 0{,}6, \\ B\colon p \neq 0{,}6, \end{cases}$ c) $\begin{cases} A\colon \sigma \neq 1{,}2, \\ B\colon \sigma = 1{,}2, \end{cases}$ d) $\begin{cases} A\colon \sigma^2 < 5{,}3, \\ B\colon \sigma^2 \geqslant 5{,}3. \end{cases}$

5. Suformuluokite statistines hipotezes, paaiškinkite, ką reiškia pirmosios ir antrosios rūšies klaidos:
   a) Norime sužinoti, ar šiuolaikiški studentai intelektualesni nei prieš dešimt metų. Turime ankstesnių 120 studentų IQ ir dabartinių 127 studentų IQ.
   b) Lygių galimybių komisija nori patikrinti, ar moterų vadybininkių vidutinis atlyginimas mažesnis nei vyrų vadybininkų.
   c) Norime nustatyti, ar naujas (daug brangesnis už senąjį) vaistas tikrai dukart rečiau sukelia pašalines reakcijas.
   d) Iki remonto kavinę per dieną aplankydavo vidutiniškai 200 klientų. Ar remontas reikšmingai padidino klientų skaičių?
   e) Laikydami specialų reakcijos testą, autobusų vairuotojai padaro vidutiniškai 15 klaidų. Dvidešimties taksistų klaidų vidurkis yra 14,3. Norime nustatyti, ar taksistų reakcija greitesnė už autobusų vairuotojų.
   f) Stomatologinė klinika teigia, kad vidutiniškai klientas jiems sumoka 500 Lt. Buhalteris nori patikrinti, ar paskutinio mėnesio duomenys nerodo, kad klientai vidutiniškai išleidžia daugiau pinigų.
   g) Automobilininkų asociacija nori sužinoti vairuotojų nuomonę, kurios markės automobiliai dažniau genda: „Opel“ ar „Mazda“.

6. Tarkime, tikriname hipotezę apie parametro reikšmę H_0: $\theta = \theta_0$ su vienintele alternatyva H_1: $\theta = \theta_1$. Statistikos T, pagal kurią sprendžiame, priimti ar atmesti hipotezę, tankis pavaizduotas 3.2.4 paveiksle. Raskite tokią kritinę sritį, kad kriterijaus galia būtų didžiausia.



3.2.4 pav. Statistikos T tankis

3. STATISTINĖS IŠVADOS VIENAI IMČIAI

Statistinė analizė – tai mįslingos, dažnai keistos manipuliacijos su eksperimento duomenimis siekiant nuslėpti, kad eksperimentas žmonijai neturi jokios reikšmės. Įprasta skaičiavimams naudoti kompiuterį, nes tai sudaro solidžios analizės įspūdį.

Prisiminkime antrosios dalies įžangos pavyzdį apie Tautogalos mero rinkimus. Parinkę tikimybinį modelį, gavome, kad atlikus daug apklausų po 1000 gyventojų 10% atvejų rezultatai merui mažiau palankūs nei per praėjusius rinkimus, net ir išlikus tam pačiam populiarumo lygiui. Padarėme išvadą, kad merui nėra ko nerimauti.

Minėtą uždavinį galima spręsti ir kitaip. Formuluojame hipotezę, kad merą remia 62% gyventojų. Alternatyva – meras mažiau populiarus. Atsižvelgdami į imties rezultatus, hipotezę priimsime arba atmesime. Kaip spręsti šią ir kitas vienos imties hipotezių tikrinimo problemas, nagrinėjama šiame skyriuje.

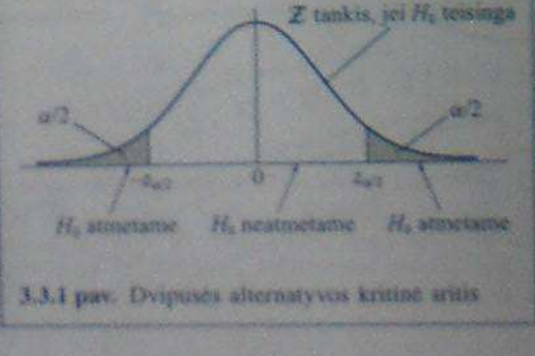
3.1. Hipotezė apie vidurkio lygybę skaičiui, kai dispersija žinoma

Tarkime, kad stebime normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Populiacijos dispersija σ^2 žinoma, o vidurkis μ nežinomas. Reikia patikrinti hipotezę H_0: $\mu = a$, čia a yra fiksuotas skaičius. Norėdami priimti sprendimą, turime fiksuotam reikšmingumo lygmeniui α parinkti tinkamą statistiką ir sukonstruoti kritinę sritį. Pats paprasčiausias nežinomo vidurkio μ įvertis yra statistika $\overline{X}$. Jeigu imties vidurkio realizacijos $\overline{x}$ mažai skiriasi nuo a (atitinkama statistikos reikšmė nepakliūna į kritinę sritį), tai hipotezę H_0 priimame, priešingu atveju hipotezės priimti negalime. Kritinė sritis sudaroma remiantis tuo, kad

$$Z = \frac{\overline{X} - a}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1), \quad \text{kai } \mu = a. \tag{3.3.1}$$

Tarkime, alternatyva H_1: $\mu \neq a$. Tuomet kritinę sritį sudaro aibė $W = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$ (žr. 3.3.1 pav.), čia $z_{\alpha/2}$ yra $\alpha/2$ lygmens standartinio normaliojo atsitiktinio dydžio kritinė reikšmė. Iš tikrųjų pagal kritinės reikšmės apibrėžimą:

$$\begin{aligned} P(\text{atmesti } H_0, \text{ kai } H_0 \text{ teisinga}) &= P(Z \in W, \text{ kai } \mu = a) \\ &= P(Z < -z_{\alpha/2}, \text{ kai } \mu = a) + P(Z > z_{\alpha/2}, \text{ kai } \mu = a) = \alpha/2 + \alpha/2 = \alpha. \end{aligned} \tag{3.3.2}$$



3.3.1 pav. Dvipusės alternatyvos kritinė sritis

Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju. Apibendrindami šiuos pastebėjimus, suformuluosime nagrinėjamojo uždavinio sprendimo etapus:

1 *Duomenys.* Intervalinių duomenų imtis $(x_1, x_2, \ldots, x_n)$ gauta matuojant normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Vidurkis μ – nežinomas, dispersija σ^2 – žinoma.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ \mu = a, \\ H_1\colon\ \mu \neq a. \end{cases} \tag{3.3.3}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$Z = \frac{\overline{x} - a}{\sigma/\sqrt{n}}. \tag{3.3.4}$$

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi μ statistiškai reikšmingai skiriasi nuo a), jeigu $|Z| > z_{\alpha/2}$. Čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leqslant z_{\alpha/2}$.

Pateikiame keletą suapvalintų $z_{\alpha/2}$ reikšmių:

$z_{0,025} = 1{,}96;$ $z_{0,05} = 1{,}64;$ $z_{0,01} = 2{,}326;$ $z_{0,1} = 1{,}281;$ $z_{0,005} = 2{,}575.$

3.3.1 pavyzdys. Sociologas nori nustatyti, ar požiūris į seksualines mažumas pasikeitė per praėjusius 30 metų. Vidutinis 1970 metų nepakantumo testo rezultatas buvo 150 balų, $s = 15$. Kuo didesnė naudojamo testo reikšmė, tuo didesnis nepakantumas. Apklausus 1999 metais 49 atsitiktinai parinktus žmones, paaiškėjo, kad $\overline{x} = 138$. Padaręs prielaidą, kad $\sigma = 15$, ir pasirinkęs reikšmingumo lygmenį $\alpha = 0{,}05$, sociologas suformulavo tokią hipotezę:

$\hat{\mu} = \overline{x} = 150$ $\hat{\sigma} = s = 15$

$$\begin{cases} H_0\colon\ \mu = 150, \\ H_1\colon\ \mu \neq 150. \end{cases}$$

Jei $|Z| > z_{\alpha/2} \to H_0$ atmetam

Tačiau $\sigma = 15$, $\alpha = 0{,}05$

Jei $|Z| \leq z_{\alpha/2} \to H_0$ neatme...

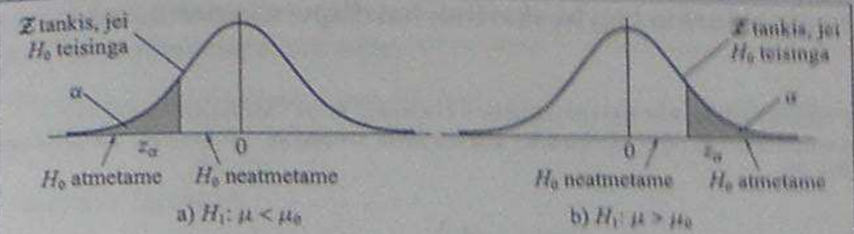
Apskaičiuojame

$\Rightarrow \overline{x} \neq 150$

$$Z = (138 - 150)/(15/\sqrt{49}) = -5{,}6.$$

$|Z| > z$

Kadangi $|Z| = |-5{,}6| = 5{,}6 > 1{,}96 = z_{0,025} = z_{0,05/2}$, tai H_0 *atmetama*. Taigi 1999 metais vidutinis žmonių požiūris į seksualines mažumas *statistiškai* reikšmingai skiriasi nuo 1970 metų požiūrio.



3.3.2 pav. Vienpusių alternatyvų kritinės sritys

Kaip ir ankstesniame skyrelyje, norime atkreipti dėmesį, kad išvadoje nekvestionuojame, ar 138 skiriasi nuo 150 (tai akivaizdu). Mes tik konstatuojame, kad skirtumas tarp šių skaičių toks didelis, kad mažai tikėtina, jog tai įvyko dėl imties atsitiktinumo. Taigi su didele tikimybe galime teigti, kad šis skirtumas būdingas ne tik šiai konkrečiai imčiai, bet ir pačiai tirtai *populiacijai*.

Vienpusėms alternatyvoms naudojama ta pati statistika Z, apibrėžiama (3.3.4) formule. Vienpusei alternatyvai H_1: $\mu < a$ parenkama kritinė sritis $W = (-\infty, -z_\alpha)$, t. y. H_0 atmetama, kai $Z < -z_\alpha$. Vienpusei alternatyvai H_1: $\mu > a$ parenkama kritinė sritis $W = (z_\alpha, \infty)$, t. y. H_0 atmetama, kai $Z > z_\alpha$ (žr. 3.3.2 pav.). Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.1 lentelėje.

3.3.1 lentelė. H_0: $\mu = a$, kai σ^2 žinoma

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\mu \neq a$	$\lvert Z\rvert > z_{\alpha/2}$	$\lvert Z\rvert \leqslant z_{\alpha/2}$
$\mu > a$	$Z > z_\alpha$	$Z \leqslant z_\alpha$
$\mu < a$	$Z < -z_\alpha$	$Z \geqslant -z_\alpha$

H_0 atmetam

H_0 neatmetam

3.3.2 pavyzdys. Kuo lėčiau tirpsta tabletė, tuo efektyviau veikia vaistai. Vidutiniškai tam tikro vaisto tabletė ištirpdavo per 18 min ($\sigma = 3$). Farmakologijos firma, sukūrusi naujas to paties vaisto tabletes, teigia, kad jos tirpsta ilgiau už ankstesnes. Bandymas parodė, kad šešiolikos naujų tablečių vidutinis tirpimo laikas yra 20 minučių. Ar reikšmingumo lygmeniui esant 0,01 galima teigti, kad naujųjų tablečių tirpimo laikas statistiškai reikšmingai ilgesnis už 18 minučių?

Sprendimas. Formuluojame *statistinę* hipotezę:

$$\begin{cases} H_0\colon \mu = 18, \\ H_1\colon \mu > 18. \end{cases} \qquad (3.3.5)$$

$\mu = \bar{X} = 18$ $\sigma = 3$ $n = 16$ $\bar{X} = 20$ $\alpha = 0{,}01$ $H_1 > 18$

$Z = \dfrac{\bar{X} - a}{\sigma/\sqrt{n}}$

Tegul naujų tablečių tirpimo laiko standartinis nuokrypis lygus 3 min ($\sigma^2 = 9$). Randame $Z = (20 - 18)/(3/\sqrt{16}) = 8/3 = 2{,}666\ldots$. Kadangi $Z = 2{,}66 > 2{,}326 = z_{0{,}01}$, tai hipotezę H_0 atmetame. Liko alternatyva H_1: $\mu > 18$. Taigi galime teigti, kad vidutinis naujų tablečių tirpimo laikas statistiškai reikšmingai ilgesnis už 18 minučių. Atkreipiame dėmesį, kad išvadą darome apie *visas* naująsias tabletes. Be to, farmakologijos firmos teiginys tapo statistinės hipotezės *alternatyva*.

$\dfrac{20-18}{3:4} = 2$

3.2. Hipotezė apie vidurkio lygybę skaičiui, kai dispersija nežinoma

Jei reklama teigia, kad 100 km pakanka 9 ℓ benzino, tai jo vidutiniškai reikia 11 ℓ. Jei reklama teigia, kad automobilis 100 km kelio sunaudoja 7 ℓ benzino, tai jo vidutiniškai reikia 9 ℓ.

Tarkime, kad:

žinome, kiek vidutiniškai santuokoje išgyvena Zanzibaro gyventojai, ir norime atsakyti į klausimą, ar lietuviai šiuo aspektu skiriasi nuo zanzibariečių;

reklama teigia, kad laikantis naujos dietos vidutiniškai per mėnesį numetama ne mažiau kaip 3 kg svorio, o konkurencijos tarnyba nori patikrinti, ar reklama nemeluoja;

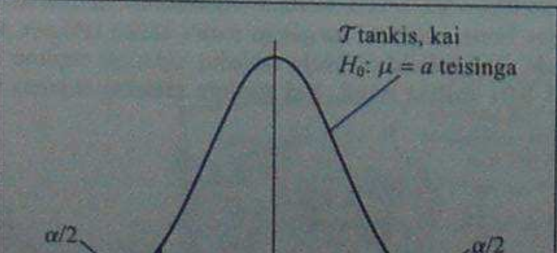
prieš penkerius metus daryti išsamūs tyrimai parodė, kad vidutinis pradinukų matematikos žinių testo įvertinimas yra 70,15 balo (pagal 100 balų skalę), o norime žinoti, ar dabartinių pradinukų žinių įvertinimas pakito.

Visais minėtais atvejais reikia atsakyti į klausimą, ar nežinomas populiacijos vidurkis skiriasi nuo tam tikro skaičiaus. Priešingai nei ankstesniame skyrelyje, populiacijos dispersija σ^2 nežinoma. Statistiniams tyrimams tokia situacija ypač dažna. Nežinoma populiacijos dispersija keičiama jos įverčiu S^2. Tačiau tada nebegalima taikyti ankstesnio skyrelio metodų, nes reikia atsižvelgti į atsitiktinę imties prigimtį ir galimą dispersijos įverčio skirtumą nuo tikrosios populiacijos dispersijos.

Tarkime, stebime normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Populiacijos dispersija σ^2 ir vidurkis μ nežinomi. Norime patikrinti hipotezę H_0: $\mu = a$, čia a fiksuotas skaičius. Kritinė sritis sudaroma remiantis tuo, kad

$$\mathcal{T} = \frac{\overline{X} - a}{\sqrt{S^2/n}}$$

turi Stjudento skirstinį su $(n-1)$ laisvės laipsnių, kai $\mu = a$ (žr. (3.1.6)). Stjudento skirstinys simetriškas nulio atžvilgiu, todėl esant dvipusei alternatyvai H_1: $\mu \neq a$ kritinė sritis yra aibė $W = (-\infty, -t_{\alpha/2}(n-1)) \cup (t_{\alpha/2}(n-1), \infty)$ (žr. 3.3.3 pav.), čia $t_{\alpha/2}(n-1)$ yra Stjudento skirstinio su $(n-1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė.



3.3.3 pav. Kritinė sritis, kai dispersija nežinoma

Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju. Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Intervalinių duomenų imtis $(x_1, x_2, \ldots, x_n)$ gauta matuojant normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Vidurkis μ ir dispersija σ^2 nežinomi.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ \mu = a, \\ H_1\colon\ \mu \neq a. \end{cases} \tag{3.3.6}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$\boxed{t = \frac{\overline{x} - a}{\sqrt{s^2/n}},} \tag{3.3.7}$$

čia $\overline{x}$ yra imties vidurkis, s^2 – imties dispersija, n – imties didumas.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi μ statistiškai reikšmingai skiriasi nuo a), jeigu $|t| > t_{\alpha/2}(n-1)$. Čia $t_{\alpha/2}(n-1)$ yra Stjudento skirstinio su $(n-1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|t| \leqslant t_{\alpha/2}(n-1)$.

Kritines reikšmes $t_{\alpha/2}(n)$ galima rasti priedo 3 lentelėje.

3.3.3 pavyzdys. Edukologas nori sužinoti, ar teisingi dėstytojų skundai, kad kasmet pirmakursiai vis negabesni. Prieš penkerius metus pirmakursių standartinio gabumų testo rezultatų vidurkis buvo 80 balų. Apklausus 25 šių metų pirmakursius, gauta $\overline{x} = 82$, $s^2 = 26$. Tarkime, kad reikšmingumo lygmuo $\alpha = 0{,}05$. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ \mu = 80, \\ H_1\colon\ \mu \neq 80. \end{cases}$$

$|t| > t_{\alpha/2}(n-1)$

$|t| \leqslant t_{0,025}(24)$

Apskaičiuojame

$$t = (82 - 80)/\sqrt{26/25} = 1{,}961.$$

Kadangi $|t| = 1{,}961 \leqslant 2{,}064 = t_{0{,}025}(24)$, tai H_0 *neatmetama*. Taigi nėra pagrindo teigti, kad šiuolaikiniai pirmakursiai gabumais statistiškai reikšmingai skiriasi nuo ankstesnių metų pirmakursių.

Vienpusėms alternatyvoms naudojama ta pati statistikos T realizacija t, apibrėžiama (3.3.7) formule. Vienpusei alternatyvai $H_1\colon \mu < a$ parenkama kritinė sritis $W = (-\infty, -t_\alpha(n-1))$, t. y. H_0 atmetama, kai $t < -t_\alpha(n-1)$. Vienpusei alternatyvai $H_1\colon \mu > a$ parenkama kritinė sritis $W = (t_\alpha(n-1), \infty)$, t. y. H_0 atmetama, kai

3.3.2 lentelė. $H_0\colon \mu = a$, kai σ^2 nežinoma

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\mu \neq a$	$\|t\| > t_{\alpha/2}(n-1)$	$\|t\| \leqslant t_{\alpha/2}(n-1)$
$\mu > a$	$t > t_\alpha(n-1)$	$t \leqslant t_\alpha(n-1)$
$\mu < a$	$t < -t_\alpha(n-1)$	$t \geqslant -t_\alpha(n-1)$

$t > t_\alpha(n-1)$. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.2 lentelėje.

3.3.4 pavyzdys. Švietimo ministerija nori žinoti, ar neakivaizdines studijas renkasi vis jaunesni žmonės. Prieš dešimtmetį vidutinis neakivaizdininkų amžius buvo 35,6 metų. Atsitiktinai parinktų 20 neakivaizdininkų amžius: 29; 38; 20; 24; 30; 32; 40; 34; 33; 32; 30; 28; 37; 35; 34; 39; 28; 40; 35 ir 32 metai. Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$.

Sprendimas. Formuluojame *statistinę* hipotezę:

$\overline{X} = 35{,}6$

$$\begin{cases} H_0\colon\ \mu = 35{,}6, \\ H_1\colon\ \mu < 35{,}6. \end{cases}$$

$\alpha = 0{,}05$

Randame $\overline{x} = 32{,}5$, $s^2 = 27{,}211$, $n = 20$, $t = -2{,}65$. Kadangi $t = -2{,}65 < -1{,}729 = -t_{0,05}(19)$, tai hipotezę H_0 atmetame. Liko alternatyva $H_1\colon \mu < 35{,}6$. Taigi galime teigti, kad vidutinis dabartinių neakivaizdininkų amžius statistiškai reikšmingai mažesnis už 35,6 metų. Atkreipiame dėmesį, kad išvadą darome apie *visus* neakivaizdininkus, o pavyzdžio klausimas tapo statistinės hipotezės *alternatyva*.

Kriterijus, naudojant SPSS paketą. Tarkime, kad reikšmingumo lygmuo yra α, hipotezė $H_0\colon \mu = 0$, o sprendžiant uždavinį gautoji p-reikšmė lygi p. Tuomet:

1 Jeigu $H_1\colon \mu \neq a$, tai H_0 *atmetama*, kai $p < \alpha$. Hipotezė H_0 neatmetama, jeigu $p \geqslant \alpha$.

2 Jeigu $H_1\colon \mu > a$ ir $\overline{x} > a$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} > a$ ir $p \geqslant \alpha$ arba $\overline{x} \leqslant a$.

3 Jeigu $H_1\colon \mu < a$ ir $\overline{x} < a$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} < a$ ir $p \geqslant \alpha$ arba $\overline{x} \geqslant a$.

3.3.5 pavyzdys. Pradėdami naujo bealkoholinio alaus gamybą, alaus daryklos savininkai nori žinoti jo poreikį. Apklausus 30 atsitiktinai parinktų prekybos tinklo „Trys paršeliai“ parduotuvių direktorių, paaiškėjo, kad per metus parduotuvėms reikia: 20; 40; 30; 38; 37; 42; 50; 42; 36; 37; 41; 47; 27; 42; 34; 22; 42; 32;

ONE-SAMPLE STATISTICS

	N	Mean	Std. Deviation	Std.Error Mean
Alus	30	37.60	7.02	1.28

ONE-SAMPLE TEST

	Test Value = 40					
	t	df	Sig (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Alus	-1.873	29	.071	-2.40	-5.02	.22

3.3.4 pav. SPSS rezultatas, kai $H_0\colon \mu = 40$

40; 39; 44; 48; 40; 34; 35; 39; 45; 29; 40 ir 36 tūkst. dekalitrų. Ar gali daryklos savininkas tikėtis, kad viena parduotuvė vidutiniškai sunaudos ne mažiau kaip 40 tūkst. dekalitrų naujojo produkto? ($\alpha = 0{,}05$.)

Sprendimas. Formuojame statistinę hipotezę:

$$\begin{cases} H_0\colon \mu = 40, \\ H_1\colon \mu < 40. \end{cases}$$

$\mu = 40$ $\mu \geq 40$ $t = \frac{\bar{x} - a}{\sqrt{\frac{S^2}{n}}}$

SPSS paketu (*One-sample t test*) gauname, kad $\overline{x} = 37{,}6$, $s^2 = 49{,}28$ ($s = 7{,}02$), $n = 30$. Be to (žr. 3.3.4 pav.), $p = 0{,}071$. Kadangi $\overline{x} = 37{,}6 < 40$ ir $p = 0{,}071 < 0{,}100 = 2 \cdot 0{,}05 = 2\alpha$, tai hipotezę H_0 atmetame. Taigi gavome statistiškai reikšmingą įrodymą, kad vidutiniškai parduotuvei reikia mažiau nei 40 tūkst. dekalitrų produkto. Alaus daryklai, gaminančiai naująjį produktą, teks į tai atsižvelgti.

$\frac{37{,}6 - 40}{\sqrt{\frac{49{,}2}{30}}} \approx \frac{-2{,}4}{1{,}3} = -2{,}1$

3.3. Hipotezė apie dispersijos lygybę skaičiui, kai vidurkis žinomas

Kontroliuojant kokybę, svarbu atsižvelgti į rezultatų sklaidą. Tarkime, gamykla, gamindama 5 colių vinis, pusę vinių pagamino 3 colių, o pusę – 7 colių. Vidutinis vinies ilgis yra 5 coliai, tačiau pirkėjai nebus patenkinti. Dar aktualesnė ši problema vaistų gamyboje – kažin ar kas sutiks vartoti vaistų ampules, kuriose vidutiniškai preparato yra tiek, kiek reikia, tačiau kartais jo yra dukart daugiau, o kartais – perpus mažiau, nei reikia. Abiem minėtais atvejais gaminių kokybę nusako populiacijos dispersija.

Hipotezės apie dispersijos reikšmę tikrinamos tik normaliai pasiskirsčiusiems kintamiesiems. Tarkime, stebime normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu_0, \sigma^2)$. Populiacijos vidurkis μ_0 žinomas, o dispersija σ^2 nežinoma. Norime patikrinti hipotezę H_0: $\sigma^2 = a$, čia a yra fiksuotas skaičius. Kritinė sritis sudaroma remiantis tuo, kad visiems $i = 1, 2, \ldots, n$

$$\frac{X_i - \mu_0}{\sigma} \sim \mathcal{N}(0, 1), \quad \text{kai } \sigma^2 = a.$$

$-t_{0{,}05}(28)$

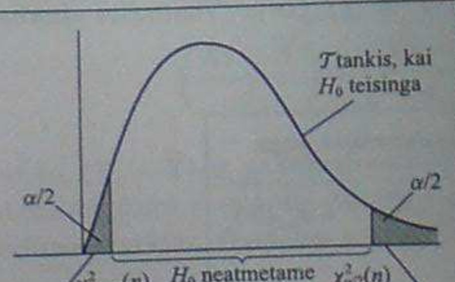
$-2 < -1{,}657$

Todėl statistika

$$\mathcal{T} = \left(\frac{X_1 - \mu_0}{\sigma}\right)^2 + \left(\frac{X_2 - \mu_0}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \mu_0}{\sigma}\right)^2 \tag{3.3.8}$$

neatmetam H. priimam.

turi χ^2 skirstinį su n laisvės laipsnių (žr. II). Kadangi χ^2 skirstinys nėra simetrinis, tai dvipusės alternatyvos H_1: $\sigma^2 \neq a$ kritinę sritį sudaro aibė $W = (-\infty, -\chi^2_{1-\alpha/2}(n)) \cup (\chi^2_{\alpha/2}(n), \infty)$ (žr. 3.3.5 pav.), čia $\chi^2_{1-\alpha/2}(n)$ yra χ^2 skirstinio su n laisvės laipsnių $1-\alpha/2$ lygmens kritinė reikšmė.



3.3.5 pav. Nežinomos dispersijos dvipusės alternatyvos kritinė sritis

Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju. Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Intervalinių duomenų imtis $(x_1, x_2, \ldots, x_n)$ gauta matuojant normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu_0, \sigma^2)$. Vidurkis μ_0 žinomas, dispersija σ^2 nežinoma.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \sigma^2 = a, \\ H_1\colon \sigma^2 \neq a. \end{cases} \tag{3.3.9}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$T = \frac{1}{a}\left((x_1 - \mu_0)^2 + (x_2 - \mu_0)^2 + \cdots + (x_n - \mu_0)^2\right). \tag{3.3.10}$$

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi σ^2 statistiškai reikšmingai skiriasi nuo a), jeigu $T > \chi^2_{\alpha/2}(n)$ arba $T < \chi^2_{1-\alpha/2}(n)$, čia $\chi^2_{\alpha/2}(n)$ ir $\chi^2_{1-\alpha/2}(n)$ yra χ^2 skirstinio su n laisvės laipsnių kritinės reikšmės. Hipotezė H_0 neatmetama, jeigu $\chi^2_{1-\alpha/2}(n) \leqslant T \leqslant \chi^2_{\alpha/2}(n)$.

3.3.6 pavyzdys. Prieš pradėdamas eksperimentą, psichologas nori sudaryti grupes iš populiacijos, kurios vidutinis testo rezultatas būtų 85 balai, o standartinis nuokrypis – 10 balų. Vienos iš sudarytų grupių testo rezultatai yra: 85; 92; 93; 90; 81; 78; 76; 78; 77; 80; 89; 92; 94 (vidurkis – 85 balai). Ar galima manyti, kad ši grupė sudaryta iš populiacijos, kurios $\sigma^2 = 10$, atstovų?

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon \sigma^2 = 10, \\ H_1\colon \sigma^2 \neq 10. \end{cases}$$

$\bar{x} = 85$
$\sigma = 10$

Randame

$$T = \frac{1}{10}\left((85-85)^2 + (92-85)^2 + \cdots\right) = 568/100 = 5{,}68.$$

Kadangi $\chi^2_{0,975}(13) = 5{,}00 < 5{,}68 < 24{,}736 = \chi^2_{0,025}(13)$, tai H_0 *neatmetama.* Taigi galime manyti, kad grupė sudaryta iš populiacijos su norimomis savybėmis atstovų.

Vienpusėms alternatyvoms naudojama ta pati statistikos T realizacija T, apibrėžiama (3.3.10) formule. Vienpusei alternatyvai $H_1\colon \sigma^2 < a$ parenkama kritinė sritis $W = (0, \chi^2_{1-\alpha}(n))$, t. y. H_0 atmetama, kai $T < \chi^2_{1-\alpha}(n)$. Vienpusei alternatyvai

3.3.3 lentelė. $H_0\colon \sigma^2 = a$, kai vidurkis žinomas

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\sigma^2 \neq a$	$T < \chi^2_{1-\alpha/2}(n)$ arba $T > \chi^2_{\alpha/2}(n)$	$\chi^2_{1-\alpha/2}(n) \leqslant T \leqslant \chi^2_{\alpha/2}(n)$
$\sigma^2 > a$	$T > \chi^2_{\alpha}(n)$	$T \leqslant \chi^2_{\alpha}(n)$
$\sigma^2 < a$	$T < \chi^2_{1-\alpha}(n)$	$T \geqslant \chi^2_{1-\alpha}(n)$

$\sigma^2 = 0{,}015.$

H_1: $\sigma^2 > a$ parenkama kritinė sritis $W = (\chi^2_\alpha(n), \infty)$, t. y. H_0 atmetama, kai $T > \chi^2_\alpha(n)$. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.3 lentelėje.

3.3.7 pavyzdys. Vaisvandenių gamykloje naudojamas pilstymo automatas pildo 0,5 ℓ talpos butelius. Nors vidutinis įpilamo į butelį vaisvandenių kiekis yra 0,5 ℓ, gamyklos savininkai susirūpino išpilstomo kiekio sklaida. Leistinas įpilamo kiekio standartinis nuokrypis yra 0,015 litro. Išmatavus 15 butelių turinį, rasta 0,53; 0,52; 0,48; 0,47; 0,50; 0,49; 0,46; 0,51; 0,52; 0,49; 0,53; 0,47; 0,50; 0,50 ir 0,53 ℓ. Ar reikia iš naujo derinti pilstymo automatą? ($\alpha = 0{,}05$.)

Sprendimas. Formuluojame *statistinę* hipotezę:

$$\begin{cases} H_0\colon \sigma^2 = 0{,}015, \\ H_1\colon \sigma^2 > 0{,}015. \end{cases}$$

Randame

$$T = (0{,}015)^{-2}\big((0{,}53 - 0{,}50)^2 + \cdots + (0{,}53 - 0{,}50)^2\big) = 33{,}777.$$

Kadangi $T = 33{,}77 > \chi^2_{0,05}(15) = 25$, tai hipotezė H_0 atmetama. Gavome statistiškai reikšmingą patvirtinimą, kad automatas išsiderino.

3.4. Hipotezė apie dispersijos lygybę skaičiui, kai vidurkis nežinomas

Ankstesniame skyrelyje minėjome, kad daugeliui tyrimų svarbi rezultatų sklaida. Šiame skyrelyje tirsime situaciją, kai stebimojo dydžio vidurkis nežinomas. Pavyzdžiui, dispersija svarbi: nustatant laiką, per kurį po iškvietimo atvyksta greitoji pagalba; vertinant produkto kalorijų kiekį; kontroliuojant gaminamų termometrų tikslumą; pasirenkant stabilios kainos vertybinius popierius ir pan.

Tarkime, stebime normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Populiacijos vidurkis μ ir dispersija σ^2 nežinomi. Norime patikrinti hipotezę H_0: $\sigma^2 = a$, čia a – fiksuotas skaičius. Kritinė sritis sudaroma remiantis tuo, kad statistika

$$T = \left(\frac{X_1 - \overline{X}}{\sigma}\right)^2 + \left(\frac{X_2 - \overline{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \overline{X}}{\sigma}\right)^2 \tag{3.3.11}$$

turi χ^2 skirstinį su $(n-1)$ laisvės laipsnių.

Kodėl, palyginti su (3.3.8), sumažėjo laisvės laipsnių skaičius? Nesunkiai įsitikiname, kad nors patys $X_1, X_2, \ldots, X_n$ nepriklausomi, atsitiktiniai dydžiai $(X_1 - \overline{X})/\sigma$, $(X_2 - \overline{X})/\sigma, \ldots, (X_n - \overline{X})/\sigma$ jau yra priklausomi. Iš tikrųjų:

$$\left(\frac{X_1 - \overline{X}}{\sigma}\right) + \left(\frac{X_2 - \overline{X}}{\sigma}\right) + \cdots + \left(\frac{X_n - \overline{X}}{\sigma}\right) = 0,$$

t. y. vieną $(X_i - \overline{X})/\sigma$ galime išreikšti kitų suma.

Atsižvelgdami į mažesnį laisvės laipsnių skaičių, perrašome ankstesniojo skyrelio sprendimo taisykles. Dvipusės alternatyvos H_1: $\sigma^2 \neq a$ kritinę sritį sudaro aibė

$$W = (-\infty, -\chi^2_{1-\alpha/2}(n-1)) \cup (\chi^2_{\alpha/2}(n-1), \infty),$$

čia $\chi^2_{1-\alpha/2}(n-1)$ yra χ^2 skirstinio su $(n-1)$ laisvės laipsnių $1 - \alpha/2$ lygmens kritinė reikšmė. Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju.

Nagrinėjamojo uždavinio sprendimo etapai konkrečiai imties realizacijai yra tokie:

1 *Duomenys.* Intervalinių duomenų imtis $(x_1, x_2, \ldots, x_n)$ gauta matuojant normalųjį atsitiktinį dydį $X \sim \mathcal{N}(\mu, \sigma^2)$. Vidurkis μ ir dispersija σ^2 nežinomi.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \sigma^2 = a, \\ H_1\colon \sigma^2 \neq a. \end{cases} \tag{3.3.12}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$T = \frac{1}{a}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right) = \boxed{\frac{(n-1)s^2}{a}} \tag{3.3.13}$$

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi σ^2 statistiškai reikšmingai skiriasi nuo a), jeigu $T > \chi^2_{\alpha/2}(n-1)$ arba $T < \chi^2_{1-\alpha/2}(n-1)$, čia $\chi^2_{\alpha/2}(n-1)$ ir $\chi^2_{1-\alpha/2}(n-1)$ yra χ^2 skirstinio su $(n-1)$ laisvės laipsnių kritinės reikšmės. Hipotezė H_0 neatmetama, jeigu $\chi^2_{1-\alpha/2}(n-1) \leqslant T \leqslant \chi^2_{\alpha/2}(n-1)$.

3.3.8 pavyzdys. Taikant naują mokymo metodą 21 studentui, gautas baigiamojo egzamino testo rezultatų standartinis nuokrypis yra 4 balai. Ar galima teigti, kad naujojo mokymo metodo rezultatų sklaida skiriasi nuo senojo metodo rezultatų, jeigu žinoma, kad, taikant ankstesnįjį metodą, rezultatų standartinis nuokrypis buvo 5 balai? ($\alpha = 0{,}01$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon \sigma^2 = \overset{a}{25}, \\ H_1\colon \sigma^2 \neq \underset{a}{25}. \end{cases}$$

n = 21
s = 4
$\sigma^2 = 25$
$\sigma \neq 25$

Randame $T = (21-1)\cdot 4^2/5^2 = 12{,}8$.

Kadangi $\chi^2_{0{,}995}(20) = 7{,}43 < 12{,}8 < 39{,}99 = \chi^2_{0{,}005}(20)$, tai H_0 *neatmetama.* Taigi naujojo ir senojo metodų rezultatų sklaidų skirtumas statistiškai nereikšmingas.

Vienpusėms alternatyvoms naudojama ta pati statistika T, apibrėžiama (3.3.11) formule. Vienpusei alternatyvai $H_1\colon \sigma^2 < a$ parenkama kritinė sritis $W = (0, \chi^2_{1-\alpha}(n-1))$, t. y. H_0 atmetama, kai $T < \chi^2_{1-\alpha}(n-1)$. Vienpusei alternatyvai $H_1\colon \sigma^2 > a$ parenkama kritinė sritis $W = (\chi^2_{\alpha}(n-1), \infty)$, t. y. H_0 atmetama, kai $T > \chi^2_{\alpha}(n-1)$. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.4 lentelėje.

3.3.4 lentelė. H_0: $\sigma^2 = a$, kai vidurkis nežinomas

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\sigma^2 \neq a$	$T < \chi^2_{1-\alpha/2}(n-1)$ arba $T > \chi^2_{\alpha/2}(n-1)$	$\chi^2_{1-\alpha/2}(n-1) \leqslant T \leqslant \chi^2_{\alpha/2}(n-1)$
$\sigma^2 > a$	$T > \chi^2_{\alpha}(n-1)$	$T \leqslant \chi^2_{\alpha}(n-1)$
$\sigma^2 < a$	$T < \chi^2_{1-\alpha}(n-1)$	$T \geqslant \chi^2_{1-\alpha}(n-1)$

3.3.9 pavyzdys. Firma, gaminanti termometrus, teigia, kad termometrų parodymų paklaidų standartinis nuokrypis neviršija $\boxed{0{,}3°C}$. Ištyrus 26 termometrus (palyginus juos su etalonu), rasta, kad visų termometrų parodymų paklaidų standartinis nuokrypis $s = 0{,}4°C$. Ar galima manyti, kad firmos teiginys nepagrįstas? ($\alpha = 0{,}05$.)

Sprendimas. Formuluojame *statistinę* hipotezę (ne standartiniam nuokrypiui, o dispersijai):

$$\begin{cases} H_0\colon \sigma^2 = 0{,}3^2; \\ H_1\colon \sigma^2 > 0{,}3^2. \end{cases}$$

$\sigma^2 = \sigma_0^2$ n = 26, s = 0,4.

Apskaičiuojame $T = 25 \cdot 0{,}16/0{,}09 = 44{,}44\ldots$. Kadangi $T > 37{,}65 = \chi^2_{0{,}05}(25)$, tai hipotezę H_0 atmetame. Gavome, kad duomenų ir firmos nurodytos sklaidos skirtumas yra statistiškai reikšmingas. Todėl firmos teiginys pernelyg optimistinis.

3.5. Hipotezė apie proporciją. Normalioji aproksimacija

– procentai.

Tarkime, kad per rinkimus politinis judėjimas „Rytai–Vakarai" surinko 15% balsų. Praėjus dvejiems metams po rinkimų, judėjimo vadovai nori žinoti, ar rinkėjų nuotaikos nepasikeitė. Apklausus 1000 rinkėjų, paaiškėjo, kad šimtas iš jų balsuotų už judėjimą „Rytai–Vakarai". Ar rinkėjų požiūris į judėjimą pasikeitė? Išvadą norime padaryti apie *visą* rinkėjų populiaciją. Todėl, vertindami ankstesnės rėmėjų dalies (15%) ir imties rėmėjų (100 iš 1000, t. y. 10% skirtumą), turime atsižvelgti į imties atsitiktinumą.

Visų pirma išsiaiškinkime, kokį atsitiktinį dydį stebime. Kiekvienas apklaustasis arba remia judėjimą, arba neremia. Tikimybė, kad atsitiktinai parinktas apklaustasis rems judėjimą, lygi visų remiančiųjų populiacijoje *daliai*. Pažymėkime ją simboliu p. Pavyzdžiui, jeigu populiaciją sudaro 3 000 000 rinkėjų, iš kurių 600 000 judėjimą remia, tai tikimybė p, kad atsitiktinai parinktas rinkėjas yra judėjimo rėmėjas, lygi $600\,000/3\,000\,000 = 0{,}2$. Tegul X yra atsitiktinis dydis, įgyjantis dvi reikšmes: $X = 1$ su tikimybe $P(X = 1) = p$ (kai apklaustas rinkėjas judėjimą remia) arba $X = 0$ su tikimybe $P(X = 0) = 1 - p$ (kai apklaustas rinkėjas judėjimo neremia). Taigi stebime binominį atsitiktinį dydį $X \sim \mathcal{B}(1, p)$ su nežinomu parametru p. Atsitiktinę imtį $(X_1, X_2, \ldots, X_n)$ sudaro nepriklausomi atsitiktiniai dydžiai, turintys tokį pat binominį skirstinį kaip ir X.[1] Atsitiktinių dydžių suma S_n turi binominį skirstinį su parametrais n ir p, t. y. $S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{B}(n, p)$.

Statistiką S_n galima taikyti hipotezėms tikrinti (tai ir daroma mažoms imtims). Tačiau dideliems n sunku apskaičiuoti binominio atsitiktinio dydžio reikšmių tikimybes. Todėl tuo atveju naudojama statistikos S_n aproksimacija. Jeigu spėjama p reikšmė, palyginti su n, *nėra labai mažas* skaičius, taikoma normalioji aproksimacija, t. y. statistika S_n keičiama nedaug nuo jos besiskiriančiu normaliuoju atsitiktiniu dydžiu. Iš centrinės ribinės teoremos (žr. II dalį) išplaukia, kad

$$Z = \frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} \approx \mathcal{N}(0, 1).$$

Kadangi $\mathbf{E}S_n = np$, o $\mathbf{D}S_n = np(1 - p)$, tai perrašome Z taip:

$$\boxed{Z} = \frac{S_n - np}{\sqrt{np(1 - p)}} = \boxed{\frac{\overline{X} - p}{\sqrt{p(1 - p)/n}}}. \tag{3.3.14}$$

[1] Kad taip būtų, ėmimas turi būti grąžintinis. Praktiškai tai retai pasitaiko. Tačiau jeigu rinkėjų populiacija gana didelė, o rėmėjų joje pakankamai daug, galima taikyti šio skyrelio samprotavimus.

Atsitiktinis dydis X įgyja tik dvi reikšmes – 0 ir 1, todėl $\overline{X}$ yra skaičius tarp 0 ir 1, atitinkantis rėmėjų imtyje skaičių. Tai yra ne kas kita kaip p įvertis, todėl labiau priimta vietoje $\overline{X}$ vartoti žymenį $\widehat{p}$. Taigi

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0, 1). \tag{3.3.15}$$

Tarkime, H_0: $p = a$. Jeigu H_0 teisinga, galime pasinaudoti asimptotiniu Z normalumu. Kritinė sritis sudaroma remiantis tomis pačiomis 3.1 skyrelio taisyklėmis. Pavyzdžiui, tegul alternatyva H_1: $p \neq a$. Tuomet kritinę sritį sudaro aibė $W = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$, čia $z_{\alpha/2}$ yra $\alpha/2$ lygmens standartinio normaliojo atsitiktinio dydžio kritinė reikšmė. Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju.

Apibendrindami šiuos pastebėjimus, suformuluosime nagrinėjamojo uždavinio sprendimo etapus:

1 *Duomenys.* Dvireikšmių duomenų aibę sudaro nuliai (matuotos savybės nerasta) ir vienetai (matuota savybė rasta).

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p = a, \\ H_1\colon\ p \neq a. \end{cases} \tag{3.3.16}$$

3 *Kriterijaus statistika.* Apskaičiuojame

sėkmių skaičius

$$Z = \frac{m - na}{\sqrt{na(1-a)}} = \frac{\widehat{p} - a}{\sqrt{a(1-a)/n}}; \tag{3.3.17}$$

čia m yra imties vienetų skaičius, $\widehat{p} = m/n$.

100 − 1000·0,15

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi p statistiškai reikšmingai skiriasi nuo a), jeigu $|Z| > z_{\alpha/2}$, čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leqslant z_{\alpha/2}$.

Kelios dažnai naudojamos z_α reikšmės buvo pateiktos 3.1 skyrelyje (žr. taip pat priedo 2 lentelę).

Pastaba. Nėra vieningos nuomonės, kokioms n ir a reikšmėms normalioji aproksimacija yra pakankamai tiksli. Kartais reikalaujama, kad tarp n ir a galiotų toks ryšys:

$$n \geqslant \max\left(\frac{5}{a}, \frac{5}{1-a}, \frac{25(1-2a)^2}{a(1-a)}\right). \tag{3.3.18}$$

Pavyzdžiui, jeigu $a = 0{,}1$, tai $n \geqslant 178$; jeigu $a = 0{,}5$, tai $n \geqslant 10$. Kartais tereikalaujama, kad $\max(na, n(1-a)) \geqslant 30$.

3.3.10 pavyzdys. Išspręsime skyrelio pradžioje suformuluotą problemą, laikydami $\alpha = 0{,}01$. Statistinė hipotezė:

$$\begin{cases} H_0\colon\ p = \underbrace{0{,}15}_{a}, \\ H_1\colon\ p \neq \underbrace{0{,}15}_{a}. \end{cases}$$

$m = 100$

$n = 1000$

Apskaičiuojame $Z = (100 - 1000 \cdot 0{,}15)/(\sqrt{1000 \cdot 0{,}15(1-0{,}15)}) = -4{,}428$. Kadangi $|Z| = 4{,}428 > 2{,}575 = z_{0,005}$, tai H_0 *atmetama*. Taigi rinkėjų požiūris statistiškai reikšmingai pasikeitė.

Vienpusėms alternatyvoms naudojama ta pati Z, apibrėžiama (3.3.17) formule. Vienpusei alternatyvai H_1: $p < a$ parenkama kritinė sritis $W = (-\infty, -z_\alpha)$, t. y. H_0 atmetama, kai $Z < -z_\alpha$. Vienpusei alternatyvai H_1: $p > a$ parenkama kritinė sritis $W = (z_\alpha, \infty)$, t. y. H_0 atmetama, kai $Z > z_\alpha$. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.5 lentelėje.

3.3.5 lentelė. H_0: $p = a$. Normalioji aproksimacija

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p \neq a$	$\lvert Z\rvert > z_{\alpha/2}$	$\lvert Z\rvert \leqslant z_{\alpha/2}$
$p > a$	$Z > z_\alpha$	$Z \leqslant z_\alpha$
$p < a$	$Z < -z_\alpha$	$Z \geqslant -z_\alpha$

3.3.11 pavyzdys. Prieš pradėdama masinę dietinių „mėsainių su lašinių kvapu" gamybą, užkandinė „Makkauskas" paprašė 100 lankytojų įvertinti naująjį produktą. Teigiamai naująjį produktą įvertino 63 lankytojai. Ar šie duomenys neprieštarauja naujojo mėsainio kūrėjo reklaminiam teiginiui, kad pagamintas produktas patiks bent dviem iš trijų lankytojų? ($\alpha = 0{,}01$.)

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p = 2/3, \\ H_1\colon\ p < 2/3. \end{cases}$$

Apskaičiuojame $Z = (63 - 200/3)/(\sqrt{100(2/3)(1/3)}) = -0{,}777\ldots$. Kadangi $Z = -0{,}777 > -2{,}326 = -z_{0,01}$, tai hipotezės H_0 neatmetame. Imties duomenys neprieštarauja reklaminiam teiginiui.

3.6. Hipotezė apie proporciją. Puasoninė aproksimacija

Kartais žinoma, kad tiriamą savybę turinčių elementų visoje populiacijoje dalis yra labai maža (pvz., 0,1% ir pan.). Tuomet normalioji proporcijos aproksimacija nebetinka ir vietoje jos taikoma puasoninė aproksimacija. Tarkime, kad stebime binominį atsitiktinį dydį $X \sim \mathcal{B}(1, p)$ su nežinomu parametru p. Atsitiktinės imties $(X_1, X_2, \ldots, X_n)$ visi atsitiktiniai dydžiai X_i nepriklausomi ir turi tą patį skirstinį kaip ir X. Imties elementų suma S_n turi binominį skirstinį su parametrais n ir p, t.y. $S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{B}(n, p)$. Mažoms p reikšmėms statistiką S_n galima pakeisti atsitiktiniu dydžiu $Y \sim \mathcal{P}(np)$, turinčiu Puasono skirstinį su parametru np.

Jeigu hipotezė apie parametro reikšmę H_0: $p = a$ teisinga, tai $Y \sim \mathcal{P}(na)$ ir galime kintamajam Y konstruoti kritines sritis. Tačiau šiuo atveju patogiau kriterijų formuluoti p-reikšmėms.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Dvireikšmių duomenų aibę sudaro nuliai (matuotos savybės nerasta) ir vienetai (matuota savybė rasta).

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p = a, \\ H_1\colon\ p \neq a. \end{cases} \tag{3.3.19}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$P(Y \geqslant m) \text{ ir } P(Y \leqslant m), \tag{3.3.20}$$

čia $Y \sim \mathcal{P}(na)$, o m – vienetų imtyje skaičius.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi p statistiškai reikšmingai skiriasi nuo a), jeigu $P(Y \geqslant m) < \alpha/2$ arba $P(Y \leqslant m) < \alpha/2$. Kitais atvejais hipotezė H_0 neatmetama.

3.3.12 pavyzdys. Vienoje valstybėje 0,1% visų žmonių turi polinkį į psichopatiškai agresyvų elgesį savo kaimynų atžvilgiu. Iš 2000 atsitiktinai parinktų kitos valstybės piliečių tokį polinkį turi trys. Ar šiuo aspektu abi valstybės skiriasi? ($\alpha = 0{,}1$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon\ p = 0{,}001, \\ H_1\colon\ p \neq 0{,}001. \end{cases}$$

Kadangi $n = 2000$, o $a = 0{,}001$, tai $Y \sim \mathcal{P}(2)$, kai H_0 teisinga. Todėl

$$P(Y \geqslant 3) = 1 - P(Y \leqslant 2) = 1 - e^{-2} - 2e^{-2} - 2^2e^{-2}/2 = 0{,}3233,$$
$$P(Y \leqslant 3) = e^{-2} + 2e^{-2} + 2^2e^{-2}/2 + 2^3e^{-2}/6 = 0{,}857.$$

Kadangi nė viena iš rastų tikimybių nėra mažesnė už 0,05, tai H_0 neatmetame. Taigi negavome patvirtinimo, kad nagrinėjamu aspektu abi valstybės statistiškai reikšmingai skiriasi.

Vienpusėms alternatyvoms naudojamos tos pačios tikimybės, tik jos lyginamos su α. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.6 lentelėje.

3.3.6 lentelė. H_0: $p = a$. Puasoninė aproksimacija $Y \sim \mathcal{P}(na)$

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p \neq a$	$P(Y \geqslant m) < \alpha/2$ arba $P(Y \leqslant m) < \alpha/2$	$P(Y \geqslant m) \geqslant \alpha/2$ ir $P(Y \leqslant m) \geqslant \alpha/2$
$p > a$	$P(Y \geqslant m) < \alpha$	$P(Y \geqslant m) \geqslant \alpha$
$p < a$	$P(Y \leqslant m) < \alpha$	$P(Y \leqslant m) \geqslant \alpha$

3.3.13 pavyzdys. Tam tikra liga serga 0,05% visos populiacijos. Naujus skiepus išbandė 3000 savanorių. Iš jų susirgo vienas. Ar skiepai statistiškai reikšmingai sumažino riziką susirgti? ($\alpha = 0{,}05$.)

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p = 0{,}0005, \\ H_1\colon\ p < 0{,}0005. \end{cases}$$

Kadangi $m = 1$, $a = 0{,}0005$, o $n = 3000$, tai $Y \sim \mathcal{P}(1{,}5)$, kai H_0 teisinga. Todėl

$$P(Y \leqslant 1) = \mathrm{e}^{-1{,}5} + 1{,}5\mathrm{e}^{-1{,}5} = 0{,}5578 > 0{,}05.$$

Taigi hipotezės H_0 neatmetame. Neturime pagrindo teigti, kad skiepai statistiškai reikšmingai sumažino riziką susirgti, todėl jų efektyvumas abejotinas.

3.7. Hipotezė apie proporciją mažoms imtims

Ankstesnieji du skyreliai buvo skirti hipotezei apie proporcijos lygybę skaičiui, kai imtis didelė. Jeigu n nėra labai didelis, galima taikyti *tikslų* kriterijų. Didelėms imtims toks kriterijus netinkamas, nes skaičiavimų apimtys labai didelės.

Kaip ir anksčiau, tarsime, kad stebime binominį atsitiktinį dydį $X \sim \mathcal{B}(1, p)$ su nežinomu parametru p. Atsitiktinę imtį sudaro nepriklausomi atsitiktiniai dydžiai $(X_1, X_2, \ldots, X_n)$, turintys tą patį skirstinį kaip ir X. Imties elementų suma S_n turi binominį skirstinį su parametrais n ir p, t. y.

$$S_n = X_1 + X_2 + \cdots + X_n \sim \mathcal{B}(n, p). \tag{3.3.21}$$

Jeigu hipotezė apie parametro reikšmę H_0: $p = a$ teisinga, tai $S_n \sim \mathcal{B}(n, a)$ ir galime S_n konstruoti kritines sritis. Tačiau šiuo atveju patogiau kriterijų formuluoti p-reikšmėms.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Dvireikšmių duomenų aibę sudaro nuliai (matuotos savybės nerasta) ir vienetai (matuota savybė rasta).

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p = a, \\ H_1\colon\ p \neq a. \end{cases} \tag{3.3.22}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$P(S_n \geqslant m) \text{ ir } P(S_n \leqslant m), \tag{3.3.23}$$

čia $S_n \sim \mathcal{B}(n, a)$, o m yra imties vienetų skaičius.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi p statistiškai reikšmingai skiriasi nuo a), jeigu $P(S_n \geqslant m) < \alpha/2$ arba $P(S_n \leqslant m) < \alpha/2$. Kitais atvejais hipotezė H_0 neatmetama.

3.3.14 pavyzdys. Kauliuką metus 9 kartus, vieną kartą atsivertė 6 akutės. Ar galime teigti, kad 6 akučių atsivertimo tikimybė nelygi 1/6? ($\alpha = 0{,}05$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon\ p = 1/6, \\ H_1\colon\ p \neq 1/6. \end{cases}$$

Kadangi $n = 9$, o $a = 1/6$, tai $S_n \sim \mathcal{B}(9; 1/6)$, kai H_0 teisinga. Randame

$$P(S_n \geqslant 1) = 1 - P(S_n = 0) = 1 - (5/6)^9 = 0{,}806,$$

$$P(S_n \leqslant 1) = P(S_n = 0) + P(S_n = 1) = (5/6)^9 + 9(1/6)(5/6)^8 = 0{,}542.$$

Kadangi nė viena iš apskaičiuotųjų tikimybių nėra mažesnė už 0,025, tai H_0 *neatmetame.* Taigi negavome patvirtinimo, kad 6 akučių atsivertimo tikimybė nelygi 1/6. Jeigu vis dėlto įtariame kauliuko asimetriją, bandymą turėtume kartoti daugiau kartų.

Vienpusėms alternatyvoms naudojamos tos pačios tikimybės, tik jos lyginamos su α. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.7 lentelėje.

3.3.7 lentelė. H_0: $p = a$. Tikslus kriterijus

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p \neq a$	$P(S_n \geqslant m) < \alpha/2$ arba $P(S_n \leqslant m) < \alpha/2$	$P(S_n \geqslant m) \geqslant \alpha/2$ ir $P(S_n \leqslant m) \geqslant \alpha/2$
$p > a$	$P(S_n \geqslant m) < \alpha$	$P(S_n \geqslant m) \geqslant \alpha$
$p < a$	$P(S_n \leqslant m) < \alpha$	$P(S_n \leqslant m) \geqslant \alpha$

3.3.15 pavyzdys. Firmoje 30% operatorių sudarė moterys. Mažinant etatus, tarp dešimties atleistų operatorių buvo septynios moterys. Ar galima firmos vadovus įtarti moterų diskriminacija? ($\alpha = 0{,}05$.)

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p = 0{,}3, \\ H_1\colon\ p > 0{,}3. \end{cases}$$

Kadangi $m = 6$, $a = 0{,}3$, o $n = 10$, tai $S_n \sim \mathcal{B}(10; 0{,}3)$, kai H_0 teisinga. Todėl

$$P(S_n \geqslant 7) = 0{,}009 + 0{,}001 + 0{,}0001 + 0{,}000 = 0{,}001\ldots < 0{,}05.$$

Taigi hipotezė H_0 atmetama. Galime įtarti moterų diskriminaciją.

Pastaba. Mažas stebėjimų skaičius *visuomet palankus* H_0. Todėl nustačius statistiškai reikšmingą skirtumą dešimties elementų imčiai, jis iš tikro yra didelis.

Kartais net ir nedideliems n sunku apskaičiuoti binomines tikimybes. Todėl galima naudoti apytikslį kriterijų, kuris pagrįstas pasikliautinaisiais intervalais. Tegul m – imties vienetų skaičius,

$$p_1^*(u) = \frac{2\chi^2_{1-u}(2m)}{2(2n-m+1)+\chi^2_{1-u}(2m)}, \qquad p_2^*(u) = \frac{2\chi^2_u(2m+2)}{2(2n-m)+\chi^2_u(2m+2)}, \tag{3.3.24}$$

o $\chi^2_u(k)$ yra χ^2 su k laisvės laipsnių u lygmens kritinė reikšmė. Tuomet sprendimo taisyklės suformuluotos 3.3.8 lentelėje.

3.3.8 lentelė. H_0: $p = a$. Apytikslis kriterijus

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p \neq a$	$a < p_1^*(\alpha/2)$ arba $a > p_2^*(\alpha/2)$	$p_1^*(\alpha/2) \leqslant a \leqslant p_2^*(\alpha/2)$
$p > a$	$a < p_1^*(\alpha)$	$a \geqslant p_1^*(\alpha)$
$p < a$	$a > p_2^*(\alpha)$	$a \leqslant p_2^*(\alpha)$

3.3.16 pavyzdys. Užbaigsime šį skyrelį, atsakydami į 3.2.1 pavyzdžio klausimą, ar Murausko dėsniui (dėstyk kaip nori, – vis tiek kas penktas studentas nieko nesupras) prieštarauja tai, kad iš 30 studentų nieko nesuprato 5 studentai. Imsime $\alpha = 0{,}05$. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p = 0{,}2, \\ H_1\colon\ p < 0{,}2. \end{cases}$$

Kadangi $m = 5$, $a = 0{,}2$, o $n = 30$, tai $p_2^*(0{,}05) = 0{,}32 \geqslant 0{,}2$. Taigi gauti duomenys nepaneigia Murausko dėsnio (hipotezė H_0 neatmetama).

3.8. Hipotezė apie koreliacijos koeficiento lygybę nuliui

Tarkime, stebime intervalinių kintamųjų porą (X, Y), gautą matuojant dvimatį normalųjį atsitiktinį dydį. Atsitiktinę imtį sudaro poros $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Norime nustatyti, ar kintamieji X ir Y koreliuoja. Atsitiktinių dydžių tiesinę priklausomybę matuoja koreliacijos koeficientas ϱ, kurio įvertis R pateiktas 1.6 skyrelyje ((3.1.14) ir (3.1.15) formulės). Ten pat buvo nurodytos nusistovėjusios vartotojų normos, kokią koreliacijos koeficiento reikšmę laikyti didele. Tačiau tos normos sudarytos neatsižvelgiant į imties didumą, todėl lieka neaišku, ar koreliacija *statistiškai* reikšmingai skiriasi nuo nulio. Šiame skyrelyje šią problemą ir nagrinėsime.

Pastaba. Tvirtai nusistovėjusi tradicija hipotezes apie koreliacijos koeficientą nagrinėti kartu su *vienos* imties kriterijais, nors koreliacijos koeficientas yra dviejų imčių elgesį nusakantis dydis. Šio skyriaus kriterijų pagrindinis bruožas yra ne viena imtis (porinė ar ne), o tai, kad hipotezės formuluojamos *vienam* parametrui ir turime tik *vieną* empirinį to parametro įvertį.

Konstruojant kritines sritis remiamasi tuo, kad

$$\mathcal{T} = R\sqrt{\frac{n-2}{1-R^2}} \tag{3.3.25}$$

turi Stjudento skirstinį su $(n-2)$ laisvės laipsnių, jeigu $\varrho = 0$.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Intervalinių duomenų porinė imtis $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ gauta matuojant dvimatį normalųjį atsitiktinį dydį (X, Y).

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ \varrho = 0, \\ H_1\colon\ \varrho \neq 0. \end{cases} \tag{3.3.26}$$

3 *Kriterijaus statistika.* Randame

$$T = r\sqrt{\frac{n-2}{1-r^2}}. \tag{3.3.27}$$

Čia r yra koreliacijos koeficiento realizacija, skaičiuojama pagal formulę

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}. \tag{3.3.28}$$

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (X ir Y statistiškai reikšmingai koreliuoja), jeigu $|T| > t_{\alpha/2}(n-2)$. Čia $t_{\alpha/2}(n-2)$ yra Stjudento skirstinio su $(n-2)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|T| \leqslant t_{\alpha/2}(n-2)$.

Vienpusėms alternatyvoms naudojama ta pati statistika $\mathcal{T}$, apibrėžiama (3.3.25) formule. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.9 lentelėje.

3.3.9 lentelė. H_0: $\varrho = 0$

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\varrho \neq 0$	$\lvert T\rvert > t_{\alpha/2}(n-2)$	$\lvert T\rvert \leqslant t_{\alpha/2}(n-2)$
$\varrho > 0$	$T > t_\alpha(n-2)$	$T \leqslant t_\alpha(n-2)$
$\varrho < 0$	$T < -t_\alpha(n-2)$	$T \geqslant -t_\alpha(n-2)$

3.3.17 pavyzdys. Patikrinsime ar 3.1.5 pavyzdyje gauta koreliacija $\boxed{r = 0{,}915}$ statistiškai reikšmingai skiriasi nuo 0. Tegul $\alpha = 0{,}01$. Statistinė hipotezė:

n = 10

$$\begin{cases} H_0\colon \varrho = 0, \\ H_1\colon \varrho \neq 0. \end{cases}$$

Apskaičiuojame

$$T = 0{,}915\sqrt{\frac{10-2}{1-0{,}915^2}} = 6{,}4146.$$

Kadangi $|T| = 6{,}4146 > 3{,}355 = t_{0{,}005}(8)$, tai H_0 *atmetama*. Koreliacija tarp pardavėjų skaičiaus ir parduodamo produkcijos kiekio statistiškai reikšminga.

3.3.18 pavyzdys. Sociologas nori nustatyti, ar yra tiesioginė priklausomybė tarp ekonomisto studijų balų vidurkio ir pradinio atlyginimo. Reikšmingumo lygmuo $\alpha = 0{,}05$. Duomenys pateikti 3.3.10 lentelėje.

Sprendimas. Formuluojame *statistinę* hipotezę:

$$\begin{cases} H_0\colon \varrho = 0, \\ H_1\colon \varrho > 0. \end{cases} \tag{3.3.29}$$

Randame $r = 0{,}183$, $T = 0{,}6711$. Kadangi $T = 0{,}6711 \leqslant 1{,}77 = t_{0{,}05}(13)$, tai hipotezės H_0 neatmetame. Duomenys neleidžia teigti, kad pradinis atlyginimas tiesiogiai tiesiškai priklauso nuo studijų balo.

3.3.10 lentelė

Balai	Atlyginimas	Balai	Atlyginimas
5,58	1500	8,70	1800
6,27	2000	8,90	2900
6,85	2300	9,20	1200
6,50	1900	9,20	1600
6,33	1000	9,37	2000
5,89	2700	9,38	2700
7,23	3000	9,50	2800
8,43	2500		

Kriterijus, naudojant SPSS paketą. Tarkime, kad reikšmingumo lygmuo yra α, $H_0\colon \varrho = 0$, o taikant kriterijų gautoji p-reikšmė lygi p. Tuomet:

1 Jeigu $H_1\colon \varrho \neq a$, tai meniu pasirenkamas 'Test of significance: two-tailed'. H_0 *atmetama* (koreliacija nenulinė), jeigu $p < \alpha$. Hipotezė H_0 neatmetama, jeigu $p \geqslant \alpha$.

2 Jeigu tiriama vienpusė alternatyva, tai meniu pasirenkamas 'Test of significance: one-tailed'. Jeigu $p \geqslant \alpha$, tai H_0 neatmetama ir statistiškai reikšmingos koreliacijos nerasta. Jeigu $p < \alpha$ ir $r > 0$, tai H_0 *atmetama* ir lieka alternatyva $H_1\colon \varrho > 0$. Jeigu $p < \alpha$ ir $r < 0$, tai H_0 *atmetama* ir lieka alternatyva $H_1\colon \varrho < 0$.

SPSS paketu gautas rezultatas apie koreliaciją tarp vilkdalgių vainiklapių ilgių ir pločių pateikiamas 3.3.6 paveiksle. Matome, kad koreliacija $r = 0{,}956$ statistiškai reikšminga ($p < 0{,}01$). Be to, r teigiama – priklausomybė tiesioginė.

```
CORRELATIONS
  /VARIABLES=y u
  /PRINT= TWOTAIL NOSING
  /MISSING=PAIRWISE.
```

Correlations

CORRELATIONS

		Vainiklapių ilgis	Vainiklapių plotis
Pearson Correlation	Vainiklapių ilgis Vainiklapių plotis	1,000 ,956**	,956** 1,000
Sig. (2-tailed)	Vainiklapių ilgis Vainiklapių plotis	, ,000	,000 ,
N	Vainiklapių ilgis Vainiklapių plotis	150 150	150 150

** Corelation is significant at the 0.01 level (2-tailed).

3.3.6 pav. SPSS paketu gautas koreliacijos rezultatas

3.9. Hipotezė apie koreliacijos koeficiento lygybę skaičiui

Tarkime, stebime intervalinių kintamųjų porą (X, Y), gautą matuojant dvimatį normalųjį atsitiktinį dydį. Atsitiktinę imtį sudaro poros $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Norime nustatyti, ar koreliacija tarp X ir Y lygi skaičiui a (H_0: $\varrho = a$). Kadangi koreliacijos įvertis yra Pirsono koreliacijos koeficientas R, tai spręsdami turime lyginti jo realizaciją r su a. Situacija, palyginti su ankstesniu skyreliu, pasikeitė, kadangi (3.3.25) galioja tik tuo atveju, kai $a = 0$. Jeigu $a \neq 0$, tai statistika R turi labai asimetrišką skirstinį. Todėl jai netinka nei normalioji, nei Stjudento aproksimacija (abi jos simetriškos). Išeitį 1915 metais pasiūlė R. A. Fišeris. Asimetriją galima panaikinti *transformuojant* koreliacijos koeficientą.

Fišerio transformacija $z_r = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$.

Transformuotoji statistika Z_R apytiksliai turi normalųjį skirstinį, kurio dispersija yra $\sqrt{1/(n-3)}$. Analogiškai transformuojame a. Kai galioja H_0: $\varrho = a$,

$$Z = (Z_R - Z_a)\sqrt{n-3} \approx \mathcal{N}(0, 1). \tag{3.3.30}$$

Kritinės sritys konstruojamos remiantis šia formule.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Intervalinių duomenų porinė imtis $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ gauta matuojant dvimatį normalųjį atsitiktinį dydį (X, Y), $n > 3$.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \varrho = a, \\ H_1\colon \varrho \neq a. \end{cases} \tag{3.3.31}$$

3 *Kriterijaus statistika.* Randame

$$Z = (Z_r - Z_a)\sqrt{n-3}. \tag{3.3.32}$$

Čia r yra Pirsono koreliacijos koeficiento realizacija, skaičiuojama pagal (3.3.28) formulę.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (X ir Y koreliacija statistiškai reikšmingai skiriasi nuo a), jeigu $|Z| > z_{\alpha/2}$. Čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leqslant z_{\alpha/2}$.

Dažniausiai naudojamos z_α reikšmės pateiktos 3.1 skyrelyje (žr. p. 155).

3.3.19 pavyzdys. Psichologas mano, kad koreliacija tarp IQ ir alkoholio suvartojimo (mililitrais per savaitę) yra lygi $-0{,}5$. Ištyręs 30 žmonių, jis gavo $r = -0{,}45$. Ar tai neprieštarauja psichologo hipotezei? ($\alpha = 0{,}01$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon \varrho = -0{,}5, \\ H_1\colon \varrho \neq -0{,}5. \end{cases}$$

$Z = (Z_{0,45} - Z_{0,5})$

$-Z_{0,45} + Z_{0,5}$

Iš knygos pabaigoje esančios priedo 7 lentelės randame $z_r = -0{,}485$, $z_{-0,5} = -0{,}549$. Apskaičiuojame $Z = (-0{,}485 + 0{,}5)/\sqrt{27} = 0{,}0779$. Kadangi $|Z| = 0{,}0779 \leqslant 2{,}575 = z_{0,005}$, tai H_0 neatmetame. Psichologo hipotezei duomenys neprieštarauja.

Vienpusėms alternatyvoms naudojama ta pati Z, apibrėžiama (3.3.32) formule. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.3.11 lentelėje.

$-0{,}485 + 0{,}545$

$0{,}06$

3.3.11 lentelė. H_0: $\varrho = a$

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\varrho \neq a$	$\lvert Z\rvert > z_{\alpha/2}$	$\lvert Z\rvert \leqslant z_{\alpha/2}$
$\varrho > a$	$Z > z_\alpha$	$Z \leqslant z_\alpha$
$\varrho < a$	$Z < -z_\alpha$	$Z \geqslant -z_\alpha$

3.3.20 pavyzdys. Siuvykla kas mėnesį dalį lėšų išleidžia savo produkcijai reklamuoti. Jos direkcija nori sužinoti, kokia išleidžiamų reklamai pinigų ir parduodamos produkcijos kiekių priklausomybė. Priklausomybė laikoma pakankamai stipria, jeigu koreliacija ne mažesnė už 0,6. Ištyrus 12 mėnesių duomenis, gauta $r = 0{,}51$. ($\alpha = 0{,}05$.)

Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \varrho = 0{,}6, \\ H_1\colon \varrho < 0{,}6. \end{cases}$$

Randame $z_r = 0{,}563$, $z_{0,6} = 0{,}693$, $Z = (0{,}563 - 0{,}693)\sqrt{9} = -0{,}39$. Kadangi $Z = -0{,}39 \geqslant -1{,}64 = z_{0,05}$, tai hipotezės H_0 neatmetame. Duomenys neleidžia teigti, kad koreliacija yra statistiškai reikšmingai mažesnė už 0,6.

Fišerio transformacija — normalioji aproksimacija — Puasono aproksimacija

UŽDAVINIAI

1. Statistikos profesorius keletą metų egzaminams naudojo tą patį 100 balų testą. Daugelio metų rezultatų vidurkis yra 78,3 balo, o standartinis nuokrypis – 10 balų. Šių metų 49 studentų testo rezultatų vidurkis yra 85 balai, o standartinis nuokrypis – 10 balų. Ar duomenys patvirtina hipotezę, kad informacija apie užduotis „nutekėjo"? ($\alpha = 0{,}05$.)

2. Automatas pildo 0,5 ℓ talpos skardines. Automatas suderintas taip, kad pilstomo alaus standartinis kvadratinis nuokrypis yra 0,02 ℓ. Išmatavus 25 skardinių turinį, paaiškėjo, kad vidutiniškai skardinėje yra po 0,49 ℓ alaus. Ar tą galima paaiškinti atsitiktinumu? ($\alpha = 0{,}1$.)

3. Dr. K. Kiškis pasiūlė meninę dietą „graužiu ir lieséju" (pertraukose tarp pagraužimų dainuojamos liaudies dainos). Jis teigia, kad ši dieta leidžia numesti vidutiniškai po 5 kg svorio per pirmąjį mėnesį. Dešimt savanorių išbandė naująją dietą. Per mėnesį jie numetė atitinkamai 3; 2; 5; 6; 7; 4; 2; 3; 0 ir 6 kg svorio. Ar duomenys neprieštarauja Dr. Kiškio teiginiui? ($\alpha = 0{,}01$.)

4. Užkandžiais prekiaujanti firma nusprendė mėsainius su žuvimi pakeisti mėsainiais su bananais. Dvylikoje užkandinių per savaitę buvo parduota atitinkamai 530; 540; 510; 500; 520; 532; 540; 515; 517; 522; 530 ir 510 naujųjų užkandžių. Žinoma, kad kiekviena užkandinė parduodavo vidutiniškai po 520 senųjų užkandžių per savaitę. Ar naujoji produkcija blogiau perkama? ($\alpha = 0{,}05$.)

5. Ampulėje turi būti po 300 mg tam tikro preparato. Leistinas nukrypimas nuo normos toks: standartinis nuokrypis ne didesnis už 10 mg. Patikrinus 15 naujos siuntos ampulių, jose preparato atitinkamai rasta 310; 312; 298; 270; 280; 300; 305; 311; 290; 288; 302; 330; 320; 295 ir 289 mg. Ar ampulių siunta atitinka reikalavimus? ($\alpha = 0{,}01$.)

6. Juodojo šokolado „Vytautas juodjūrietis" gamintojai kokybės kontrolei teigia, kad 100 gramų produkto kalorijų kiekis nuo 1000 kcal skiriasi ne daugiau kaip 50 kcal. Kontrolė patikrino 20 šokolado plytelių ir nustatė, kad kalorijų standartinis nuokrypis $s = 15$ kcal. Ar tai neprieštarauja gamintojų teiginiui? (Paklaidą galima laikyti $\approx 3\sigma$, $\alpha = 0{,}05$.)

7. Naujo medikamento reklamoje teigiama, kad jis sukelia pašalines reakcijas ne daugiau kaip 1% pacientų. Ištyrus 1000 vaistą vartojusių ligonių, nustatyta, kad pašalinį poveikį pajuto 32 ligoniai. Ar duomenys neprieštarauja reklaminiam teiginiui? ($\alpha = 0{,}05$.)

8. Universiteto administracija sprendžia, ar reikia įrengti naują automobilių stovėjimo aikštelę. Jos manymu, per 50% studentų į paskaitas važinėja automobiliais. Apklausus 240 studentų, paaiškėjo, kad iš jų į paskaitas važinėja 140. Ar šie duomenys neprieštarauja administracijos manymui? ($\alpha = 0{,}05$.)

9. Parduotuvė garantuoja nemokamą metinį televizorių remontą. Televizorius gaminanti firma teigia, kad per pirmus jų eksploatavimo metus remontuoti tenka ne daugiau kaip 5% televizorių. Iš 250 parduotų televizorių parduotuvei teko remontuoti 15. Ar tai neprieštarauja firmos teiginiui? ($\alpha = 0{,}01$.)

10. Ekonomistas nori patikrinti, ar padaugėjo smulkių įmonių (procentais). Prieš 10 metų jos sudarė 20% visų įmonių. Šiuo metu iš 100 atsitiktinai parinktų įmonių 27 buvo smulkios. ($\alpha = 0{,}05$.)

11. Studentas Algirdas mano, kad tik 0,1% pirmakursių yra neragavę alkoholio. Tarp 3000 apklaustų studentų tokių atsirado 4. Ar duomenys neprieštarauja studento Algirdo teiginiui? ($\alpha = 0{,}05$.)

12. Pildant testą, kiekvienam klausimui reikia pasirinkti vieną atsakymą iš dviejų. Testą sudaro 100 klausimų. Į testo klausimus atsakinėjo 100 žmonių, suskaičiavome kiekvieno iš jų teisingai atsakytus klausimus. Kokia koreliacija tarp teisingai atsakytų ir neteisingai atsakytų klausimų skaičiaus?

13. Duomenys apie pardavėjo stažą (metais) ir jo pradinį atlyginimą (sutartiniais vienetais) pateikti 3.3.12 lentelėje. Ar atlyginimas tiesiškai priklauso nuo pardavėjo stažo? ($\alpha = 0{,}05$.)

3.3.12 lentelė

Stažas	Atlyginimas	Stažas	Atlyginimas
2	100	8	500
1,5	300	7	400
3	400	5	400
10	600	4	250
12	600	2	200
4	300	1	100
2	100	6	350

14. Leidykla spėja, kad koreliacija tarp knygos kainos ir parduodamo jų kiekio yra $-0{,}6$. Eksperimento metu buvo išbandytos 103 kainos. Gauta koreliacija $r = -0{,}7$. Ar šis rezultatas nepaneigia leidyklos spėjimo? ($\alpha = 0{,}01$.)

4. STATISTINĖS IŠVADOS DVIEM IMTIMS

Paklauskite žmogaus su „garsia" pavarde (tarkime, Erlicko) – „Tai jūs Erlicko sūnus?" Beveik visuomet išgirsite – „Ne." Pakartoję eksperimentą kelis kartus, gausite patikimą statistinį įrodymą, kad dauguma „garsių" pavardžių turėtojų laiko save nesantuokiniais vaikais.

Statistikams dažnai tenka lyginti dviejose populiacijose stebimus atsitiktinius dydžius. Atsitiktinių dydžių skirstinių skirtumai nustatomi remiantis atitinkamų imčių statistikų skirtumais. Žinoma, atsižvelgiama į atsitiktinę imčių prigimtį. Kai imčių statistikų skirtumai dideli, labai mažai tikėtina, kad tai atsitiktinumo pasekmė. Tuomet sakome, kad imčių statistikos *statistiškai* reikšmingai skiriasi ir didelė tikimybė, jog šia prasme skiriasi ir pačios populiacijos. Dvi imtis galima gauti ir pakartotinai matuojant tuos pačius objektus – prieš ir po dietos, mokslo metų pradžioje ir pabaigoje ir pan. Rezultatų skirtumai leidžia įvertinti dietos ar mokymo metodo efektyvumą.

Šiame skyriuje nagrinėsime parametrines hipotezes, t. y. tirsime dviejų populiacijų *parametrų* skirtumus. Tikrieji populiacijų parametrai paprastai esti nežinomi, todėl apie jų skirtumus spręsime pagal parametrų įverčių skirtumus. Sprendimo priėmimo principai lieka tokie pat kaip ir vienos imties atveju – konstruojama statistika, kuri remiasi parametrų įverčių skirtumu. Jeigu parametrai vienodi, ši statistika turi žinomą skirstinį. Hipotezę apie parametrų skirtumą galima suformuluoti dvejopai: H_0: $\mu_1 = \mu_2$ ir H_0: $\mu_1 - \mu_2 = 0$. Abi šios išraiškos ekvivalenčios. Dažniau naudosime pirmąjį užrašą.

4.1. Stjudento kriterijus, taikomas nepriklausomoms imtims

Tarkime, norime žinoti:

ar tam tikro amžiaus berniukų ir mergaičių vidutinis ūgis yra tas pats;

ar vyrai ir moterys vienodai greitai atlieka sudėtingas automobilio vairavimo užduotis;

ar vidutinė TV žiūrėjimo per parą trukmė mieste ir kaime ta pati;

ar vienodai ilgai tarnauja to paties modelio kompiuteris, surinktas Europoje ir Azijoje;

ar vidutinis žuvų mutacijų skaičius Ignalinos ežeruose skiriasi nuo mutacijų skaičiaus kituose Lietuvos ežeruose;

ar dvi tiriamos firmos tiekia vienodos kokybės žaliavas;

ar baigiamąjį matematikos egzaminą vienodai gerai išlaikė Vilniaus ir Balbieriškio moksleiviai ir pan.

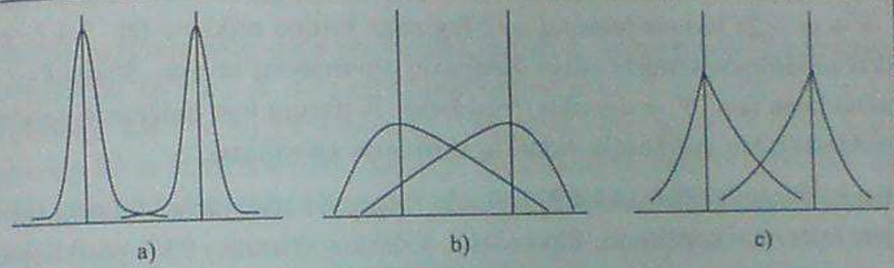
Šiame skyrelyje lyginsime kintamųjų, stebimų dviejose populiacijose, vidurkius. Visais šiais atvejais matuojami kintamieji laikomi normaliai pasiskirsčiusiais ir formuluojamas klausimas apie jų vidurkių lygybę.

Dažnai populiacijos parenkamos taip, kad galimi jų skirtumai atskleistų aplinkos poveikį. Pavyzdžiui, taip nustatoma: kuri mokymo programa efektyvesnė; ar naktinė ir dieninė pamainos vienodai našiai dirba; ar vienodai ilgai ligoninėse gydosi ta pačia liga sergantys skirtingų socialinių sluoksnių atstovai; ar po reklaminės kampanijos produkcijos vidutiniškai parduodama daugiau; ar išankstinis studentų gąsdinimas, kad egzaminas labai sunkus, turi įtakos egzamino rezultatams ir pan.

Vidurkių lygybei tikrinti naudojamos statistikos turi Stjudento skirstinius, todėl atitinkami kriterijai visuotinai vadinami Stjudento t kriterijais, arba t testais.

4.1.1. Stjudento kriterijus, kai populiacijų dispersijos lygios

Tarkime, atsitiktinės imtys $(X_1, X_2, \ldots, X_n)$ ir $(Y_1, Y_2, \ldots, Y_m)$ gautos stebint du *nepriklausomus* normaliuosius atsitiktinius dydžius $X \sim \mathcal{N}(\mu_X, \sigma^2)$ ir $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, kurių vidurkiai μ_X ir μ_Y nežinomi. Abiejų dydžių dispersija σ^2 ta pati ir nežinoma. Norime patikrinti hipotezę H_0: $\mu_X = \mu_Y$. Paprasčiausias populiacijų vidurkių skirtumo $\mu_X - \mu_Y$ įvertis yra $\overline{X} - \overline{Y}$. Tačiau vien imčių vidurkių skirtumas dar neatskleidžia pačių populiacijų skirtumų. Dviejų imčių histogramos (dėl vaizdumo jos nubraižytos suglodintos) pavaizduotos 3.4.1 paveiksle. Visais atvejais imčių vidurkių *skirtumas* toks pat. Atveju a), kai duomenų sklaida maža, tikėtina, kad skiriasi ir pačios populiacijos, tuo tarpu atveju b), kai sklaida didelė, populiacijų skirtumas labai abejotinas. Atveju c), kai sklaida vidutinė, populiacijos gali ir skirtis, ir nesiskirti.



3.4.1 pav.

Taigi vertinant vidurkių skirtumus, svarbi ir duomenų sklaida – stebimųjų dydžių dispersijos.

Nesunku įsitikinti, kad statistikos $\overline{X} - \overline{Y}$ vidurkis $\mathbf{E}(\overline{X} - \overline{Y}) = \mu_X - \mu_Y$, o dispersija

$$\begin{aligned}\mathbf{D}(\overline{X} - \overline{Y}) &= \frac{1}{n^2}(\mathbf{D}X_1 + \mathbf{D}X_2 + \cdots + \mathbf{D}X_n) \\ &\quad + \frac{1}{m^2}(\mathbf{D}Y_1 + \mathbf{D}Y_2 + \cdots + \mathbf{D}Y_m) = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).\end{aligned} \tag{3.4.1}$$

Tikrosios σ^2 reikšmės nežinome, todėl ją keičiame dispersijos įverčiu. Dispersijos σ^2 įverčiai yra ir S_X^2 ir S_Y^2. Iš šių įverčių sudarome naują jungtinį dispersijos įvertį S_p^2, skaičiuojamą pagal formulę

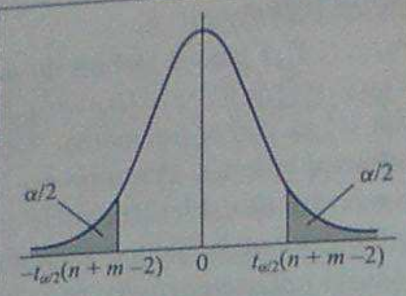
$$S_p^2 = \frac{\sum_1^n (X_i - \overline{X})^2 + \sum_1^m (Y_j - \overline{Y})^2}{n + m - 2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}. \tag{3.4.2}$$

Formulėje (3.4.1) σ^2 pakeitę S_p^2, gausime $\mathbf{D}(\overline{X} - \overline{Y})$ įvertį.

Kritinė sritis sudaroma remiantis tuo, kad statistika

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}} \tag{3.4.3}$$

turi Stjudento skirstinį su $(n + m - 2)$ laisvės laipsnių, kai $\mu_X = \mu_Y$. Stjudento skirstinys simetriškas nulio atžvilgiu, todėl dvipusės alternatyvos H_1: $\mu_X \neq \mu_Y$ kritinė sritis yra



3.4.2 pav. Vidurkių skirtumo kritinė sritis

aibė $W = (-\infty, -t_{\alpha/2}(n+m-2)) \cup (t_{\alpha/2}(n+m-2), \infty)$, čia $t_{\alpha/2}(n+m-2)$ yra Stjudento skirstinio su $(n + m - 2)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė (žr. 3.4.2. pav.).

Analogiškai sudaromos kritinės sritys vienpusių alternatyvų atveju. Statistika (3.4.3) dažnai interpretuojama taip: $T = signalas/triukšmas$. Iš tikrųjų kuo mažesnis standartinis nuokrypis (*triukšmas*), tuo svarbesnis vidurkių skirtumas (*signalas*).

Pastaba. Statistika T, apibrėžiama (3.4.3) formule, taip pat naudojama vidurkių skirtumo pasikliautinajam intervalui konstruoti. Pavyzdžiui, vidurkių skirtumo 95% pasikliautinasis intervalas

$$(\overline{X} - \overline{Y}) \pm t_{0,025}(n + m - 2)\sqrt{S_p^2(1/n + 1/m)}.$$

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Dvi intervalinių duomenų imtys $(x_1, x_2, \ldots, x_n)$ ir $(y_1, y_2, \ldots, y_m)$ gautos matuojant du nepriklausomus normaliuosius atsitiktinius dydžius $X \sim \mathcal{N}(\mu_X, \sigma^2)$ ir $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$. Vidurkiai μ_X, μ_Y ir dispersija σ^2 nežinomi.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X \neq \mu_Y. \end{cases} \tag{3.4.4}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{1/n + 1/m}} = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}; \tag{3.4.5}$$

čia $\overline{x}$, $\overline{y}$ yra imčių vidurkiai, s_x^2, s_y^2 – imčių dispersijos, o n, m – imčių didumai.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama*, jeigu $|t| > t_{\alpha/2}(n + m - 2)$. Čia $t_{\alpha/2}(n + m - 2)$ yra Stjudento skirstinio su $(n - 1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|t| \leqslant t_{\alpha/2}(n + m - 2)$.

Kritines reikšmes $t_{\alpha/2}(n + m - 2)$ galima rasti priedo 3 lentelėje.

3.4.1 pavyzdys. Lygių teisių komisija tikrina, ar didelėje draudimo firmoje nėra lyčių diskriminavimo. Atsitiktinai parinkus 20 draudimo agentų ir 25 draudimo agentes, paaiškėjo, kad agento vidutinės mėnesio pajamos yra 3050 Lt ($s_x = 200$), o agentės – 2900 Lt ($s_y = 300$). Lygių teisių komisija nori žinoti, ar nepažeidžiamos agenčių teisės ($\alpha = 0{,}05$).

Sprendimas. Formuluojame statistinę hipotezę:

X Y

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X \neq \mu_Y. \end{cases}$$

$n_x = 20$ $n_y = 25$

$\bar{X}_x = 3050$ $s_x = 200$
$\bar{X}_y = 2900$ $S_y = 300$

Randame statistiką

$$t = \frac{3050 - 2900}{\sqrt{19 \cdot 200^2 + 24 \cdot 300^2}} \sqrt{\frac{20 \cdot 25(20 + 25 - 2)}{20 + 25}} = 1{,}9187.$$

Kadangi $|t| = 1{,}9187 \leqslant 2{,}01 = t_{0{,}025}(43)$, tai H_0 *neatmetama*. Taigi draudimo agentų ir agenčių vidutinių atlyginimų skirtumas statistiškai nereikšmingas.

Vienpusėms alternatyvoms naudojama ta pati t, apibrėžiama (3.4.5) formule. Vienpusei alternatyvai $H_1\colon \mu_X < \mu_Y$ parenkama kritinė sritis $W = (-\infty, -t_\alpha(n + m - 2))$, t. y. H_0 atmetama, kai $t < -t_\alpha(n + m - 1)$. Vienpusei alternatyvai $H_1\colon \mu_X > \mu_Y$ parenkama kritinė sritis $W = (t_\alpha(n+m-2), \infty)$, t. y. H_0 atmetama, kai $t > t_\alpha(n+m-2)$. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.1 lentelėje.

3.4.1 lentelė. $H_0\colon \mu_X = \mu_Y$, kai dispersijos lygios

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\mu_X \neq \mu_Y$	$\lvert t\rvert > t_{\alpha/2}(n + m - 2)$	$\lvert t\rvert \leqslant t_{\alpha/2}(n + m - 2)$
$\mu_X > \mu_Y$	$t > t_\alpha(n + m - 2)$	$t \leqslant t_\alpha(n + m - 2)$
$\mu_X < \mu_Y$	$t < -t_\alpha(n + m - 2)$	$t \geqslant -t_\alpha(n + m - 2)$

3.4.2 pavyzdys. Firmos vadovai nori patikrinti, ar naktinės pamainos darbo našumas mažesnis nei dieninės. Per 20 darbo dienų dieninėje pamainoje pagaminta 20; 19; 22; 18; 22; 24; 25; 23; 22; 22; 19; 21; 26; 25; 23; 17; 18; 17; 22; 20 grężimo staklių. Per tą patį laikotarpį naktinėje pamainoje pagaminta 20; 19; 19; 17; 20; 23; 23; 20; 20; 21; 19; 21; 22; 20; 15; 16; 17; 17; 16; 17 grężimo staklių. Tarkime, kad reikšmingumo lygmuo $\alpha = 0{,}05$.

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X > \mu_Y. \end{cases}$$

Randame $\bar{x} = 21{,}25$; $s_x^2 = 7{,}247$; $n = 20$; $\bar{y} = 19{,}1$; $s_y^2 = 5{,}469$; $n = 20$; $t = 2{,}697$. Kadangi $t = 2{,}697 > 1{,}68 = t_{0{,}05}(38)$, tai hipotezę H_0 atmetame. Liko alternatyva $H_1\colon \mu_X > \mu_Y$. Taigi galime teigti, kad vidutinis naktinės pamainos darbo našumas mažesnis už dieninės pamainos darbo našumą.

4.1.2. Stjudento kriterijus, kai populiacijų dispersijos nelygios

Ankstesniame skyrelyje stebimų kintamųjų dispersijos buvo lygios. Kai taip nėra, susiduriame su vadinamąja Berenso–Fišerio problema. Žinomi keli apytiksliai šios problemos sprendimai. Pateiksime vieną iš jų:

1 *Duomenys.* Dvi intervalinių duomenų imtys $(x_1, x_2, \ldots, x_n)$ ir $(y_1, y_2, \ldots, y_m)$ gautos matuojant du nepriklausomus normaliuosius atsitiktinius dydžius $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ ir $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Vidurkiai μ_X, μ_Y ir dispersijos σ_X^2, σ_Y^2 nežinomi.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X \neq \mu_Y. \end{cases} \tag{3.4.6}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2/n + s_y^2/m}}; \tag{3.4.7}$$

čia $\overline{x}$, $\overline{y}$ yra imčių vidurkiai, s_x^2, s_y^2 – imčių dispersijos, o n, m – imčių didumai.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama*, jeigu $|t| > t_{\alpha/2}(k)$. Čia $t_{\alpha/2}(k)$ yra Stjudento skirstinio su k laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Laisvės laipsnių skaičius k yra mažiausias sveikasis skaičius, tenkinantis sąlygą

$$k \leqslant \frac{(s_x^2/n + s_y^2/m)^2}{s_x^4/n^3 + s_y^4/m^3}. \tag{3.4.8}$$

Hipotezė H_0 atmetama, jeigu $|t| > t_{\alpha/2}(k)$. Vienpusėms alternatyvoms naudojama ta pati statistika ir tas pats laisvės laipsnių skaičius k. Vienpusei alternatyvai $H_1\colon \mu_X < \mu_Y$ parenkama kritinė sritis $W = (-\infty, -t_\alpha(k))$, t. y. H_0 atmetama, kai $t < -t_\alpha(k)$. Vienpusei alternatyvai $H_1\colon \mu_X > \mu_Y$ parenkama kritinė sritis $W = (t_\alpha(k), \infty)$, t. y. H_0 atmetama, kai $t > t_\alpha(k)$.

3.4.3 pavyzdys. Rinkotyros specialistas nori nustatyti, ar naujoji pieno pakuotė padidino pirkėjų skaičių. Buvo pasirinkta 50 parduotuvių, kurių dienos apyvarta vienoda. Iš jų 30 atsitiktinai parinktų parduotuvių pardavinėjo produkciją senoje pakuotėje, o 20 likusių – naujojoje. Vidutinis per dieną parduotos produkcijos kiekis atitinkamai yra 130 ($s_x = 15$) ir 139 ($s_y = 3$) vienetų. Tarkime, kad reikšmingumo lygmuo $\alpha = 0{,}01$.

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X < \mu_Y. \end{cases}$$

Apskaičiuojame $n = 30$, $m = 20$, $\overline{x} = 130$, $\overline{y} = 139$, $s_x^2 = 225$, $s_y^2 = 9$, $t = -3{,}1919$. Randame laisvės laipsnių skaičių

$$k \leqslant \frac{(225/30 + 9/20)^2}{225^2/30^3 + 9^2/20^3} = 33{,}53.$$

Taigi $k = 33$. Kadangi $t = -3{,}19 < -2{,}45 = t_{0{,}01}(33)$, tai hipotezę H_0 atmetame. Liko alternatyva H_1: $\mu_X < \mu_Y$. Taigi galime teigti, kad vidutinis parduotos produkcijos naujojoje pakuotėje kiekis didesnis už vidutinį parduotą produkcijos senoje pakuotėje kiekį.

4.1.3. Stjudento kriterijaus, taikomo nepriklausomoms imtims, modifikacijos

Kartais dviejų populiacijų vidurkių skirtumą reikia palyginti su skaičiumi, nelygiu nuliui. Pavyzdžiui, norime žinoti: ar naujasis gydymo metodas, palyginti su senuoju, vidutiniškai 5 dienomis sutrumpina pooperacinės reabilitacijos trukmę; ar penkiolikmečių ir dvidešimtmečių IQ skirtumas ne mažesnis kaip 10 balų; ar naujoji Interneto adresų paieškos programa vidutiniškai 1 s greitesnė už senąją ir pan. Norint atsakyti į šiuos klausimus pakanka tik truputį pakeisti (3.4.5) ir (3.4.7) formules. Tarkime, kad H_0: $\mu_X - \mu_Y = C$. Tuomet (3.4.5) formulė virsta tokia:

$$t = \frac{\overline{x} - \overline{y} - C}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}. \qquad (3.4.9)$$

Visos sprendimo priėmimo taisyklės aprašytos 3.4.1 lentelėje, tik reikia atitinkamai pakeisti alternatyvas. Alternatyva H_1: $\mu_X \neq \mu_y$ keičiama H_1: $\mu_X - \mu_Y \neq C$; alternatyva H_1: $\mu_X < \mu_y$ keičiama H_1: $\mu_X - \mu_Y < C$; alternatyva H_1: $\mu_X > \mu_y$ keičiama H_1: $\mu_X - \mu_Y > C$.

3.4.4 pavyzdys. Apklausus 100 atsitiktinai parinktų aukštąjį išsilavinimą turinčių žmonių, paaiškėjo, kad vidutiniškai per pusmetį teatrams, koncertams, parodoms ir pan. jie vidutiniškai išleidžia po 300 Lt ($s_x = 18$ Lt). Apklausus 150 žmonių, neturinčių aukštojo išsilavinimo, nustatyta, kad analogiškoms reikmėms jie vidutiniškai išleidžia po 180 Lt ($s_y = 20$ Lt). Ar duomenys patvirtina hipotezę, kad žmonės su aukštuoju išsilavinimu kultūrinėms reikmėms išleidžia 100 Lt daugiau už žmones be aukštojo išsilavinimo? ($\alpha = 0{,}01$.)

Sprendimas. Akivaizdu, kad $\overline{x} - \overline{y} = 300 - 180 = 120 > 100$; taigi *imties* žmonių su aukštuoju išsilavinimu išlaidos viršija *imties* žmonių be aukštojo išsilavinimo išlaidas daugiau nei 100 Lt. Tačiau ar šis skirtumas toks didelis, kad galėtume kalbėti apie analogišką rezultatą *populiacijoms*? Formuluojame statistinę hipotezę:

$$\begin{cases} H_0: \mu_X - \mu_Y = 100, \\ H_1: \mu_X - \mu_Y > 100. \end{cases}$$

Apskaičiuojame statistiką

$$t = \frac{300 - 180 - 100}{\sqrt{99 \cdot 18^2 + 149 \cdot 20^2}} \sqrt{\frac{100 \cdot 150(100 + 150 - 2)}{100 + 150}} = 8{,}057.$$

Kadangi $t = 8{,}057 > 2{,}3 = t_{0{,}01}(248)$, tai H_0 *atmetama*. Taigi skirtingų visuomenės grupių išlaidų kultūrinėms reikmėms skirtumas viršija 100 Lt.

Analogiškai keičiama statistika ir nelygių dispersijų atveju.

4.1.4. Stjudento kriterijaus taikymas naudojantis SPSS paketu

Tarkime, kad reikšmingumo lygmuo yra α, H_0: $\mu_X = \mu_Y$.

SPSS pakete pateikiamos dvi statistikos realizacijos: viena lygių dispersijų, kita – nelygių. Todėl tikrinant hipotezę apie vidurkių lygybę, kartu patikrinama hipotezė apie imčių dispersijų lygybę (žr. taip pat 4.3). Priklausomai nuo to, ar dispersijas laikome statistiškai nereikšmingai (atitinkama p-reikšmė didesnė už α), ar statistiškai reikšmingai (atitinkama p-reikšmė mažesnė už α) besiskiriančiomis, hipotezei apie vidurkių lygybę tikrinti parenkama atitinkamai viršutinė arba apatinė stulpelio 'Sig. (2-tailed)' p-reikšmė. Tarkime, kad ji lygi p. Tuomet:

1. Jeigu H_1: $\mu_X \neq \mu_Y$, tai H_0 *atmetama*, kai $p < \alpha$. Hipotezė H_0 neatmetama, jeigu $p \geqslant \alpha$.
2. Jeigu H_1: $\mu_X > \mu_Y$ ir $\overline{x} > \overline{y}$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} > \overline{y}$ ir $p \geqslant 2\alpha$ arba $\overline{x} \leqslant \overline{y}$.
3. Jeigu H_1: $\mu_X < \mu_Y$, ir $\overline{x} < \overline{y}$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} < \overline{y}$ ir $p \geqslant 2\alpha$ arba $\overline{x} \geqslant \overline{y}$.

3.4.5 pavyzdys. Norima patikrinti dviejų migdomųjų preparatų efektyvumą. Matuojama (sekundėmis), per kiek laiko užmiega preparato gavusi jūrų kiaulytė. Pirmas preparatas buvo išbandytas 20 kiaulyčių. Gauta: 288; 253; 262; 279; 270; 262; 281; 247; 252; 292; 297; 293; 280; 295; 278; 282; 290; 281; 280; 279. Antras preparatas buvo išbandytas 15 kiaulyčių. Gauta: 263; 236; 250; 230; 279; 245; 290; 274; 235; 269; 262; 240; 260; 254; 266 ($\alpha = 0{,}05$).

Sprendimas. SPSS paketu gauti rezultatai pateikti 3.4.3 paveiksle. Matome, kad vidutinis užmigimo skiriant pirmąjį preparatą laikas yra 277,0500 s (standartinis nuokrypis 14,8305 s), o skiriant antrąjį preparatą – 256,8667 s (standartinis nuokrypis 17,5290 s).

Kaip rasti t kriterijaus p-reikšmę? Prieš tikrinant hipotezę apie vidurkių lygybę, reikia nuspręsti, ar populiacijų dispersijas galima laikyti lygiomis. Grafoje 'Levene's Test for Equality of Variances' pateikiama statistikos F reikšmė (0,821) ir p-reikšmė (0,372). Kai p-reikšmė ne mažesnė už pasirinktąjį reikšmingumo lygmenį α, dispersijos statistiškai reikšmingai nesiskiria. Jeigu p-reikšmė mažesnė už α, tai populiacijų dispersijos statistiškai reikšmingai skiriasi. Išvadoms apie pačių vidurkių lygybę skirta likusioji lentelės dalis ('t-test for Equality of Means'). Stulpelyje 't' pateiktos statistikos t reikšmės, stulpelyje 'df' – laisvės laipsnių skaičius, stulpelyje 'Sig. (2-tailed)' – p-reikšmės, kiti stulpeliai skirti vidurkių skirtumo įverčiams. Visų stulpelių pirmoji eilutė yra lygių dis-

GROUP STATISTICS

	Preparatas	N	Mean	Std. Deviation	Std. Error Mean
Užmigimo laikas	1.00	20	277.0500	14.8305	3.3162
	2.00	15	256.8667	17.5290	4.5260

INDEPENDED SAMPLES TEST

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Mean	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
Užmigimo laikas	Equal variances assumed	.821	.372	3.686	33	.001	20.1833	5.4756	9.0431	31.3235
	Equal variances not assumed			3.597	27.275	.001	20.1833	5.6109	8.6762	31.6904

3.4.3 pav. Nepriklausomų imčių t kriterijus. SPSS rezultatai persijų atvejui ('Equal variances assumed'), antroji – nelygių dispersijų atvejui ('Equal variances not assumed'). Taigi t kriterijaus p-reikšmė randama taip: pirmiausia stulpelio 'Sig.' reikšmę palyginame su reikšmingumo lygmeniu α. Jeigu ši reikšmė didesnė už α arba lygi jai, tai t kriterijaus p-reikšmė yra viršutinis stulpelio 'Sig. (2-tailed)' skaičius. Jeigu ši reikšmė mažesnė už α, tai t kriterijaus p-reikšmė yra apatinis stulpelio 'Sig. (2-tailed)' skaičius. Nagrinėjamojo pavyzdžio atveju $0{,}372 > 0{,}05$, todėl dispersijas galima laikyti nesiskiriančiomis ir t kriterijaus p-reikšmė yra viršutinis 'Sig. (2-tailed)' skaičius $p = 0{,}001$. Kadangi šis skaičius yra mažesnis už 0,05, tai darome išvadą, kad vidutinis užmigimo laikas statistiškai reikšmingai skiriasi – preparatų efektyvumas skirtingas.

4.2. Stjudento kriterijus, taikomas priklausomoms imtims

Tarkime, norime nustatyti dietos efektyvumą. Žinome tiriamosios žmonių grupės svorius prieš dietą ir po jos. Šiuo atveju svarbiausia yra svorių pokyčiai. Svorių pokyčių tyrimas pašalina potencialų paklaidų šaltinį – skirtingus pradinius žmonių svorius. Pradinių ir galutinių rezultatų skirtumas yra pagrindinis informacijos šaltinis tiriant: ar pardavėjų mokymas padidina parduodamos produkcijos kiekį; ar ramioje aplinkoje ir triukšmingoje išsprendžiamas vienodas uždavinių skaičius; ar knygoms skaityti praleidžiama daugiau laiko, nei žiūrėti TV. Visais šiais atvejais taikomas porinis Stjudento t kriterijus. Jis taikomas priklausomoms imtims. Nuo t kriterijaus nepriklausomoms imtims jis visų pirma skiriasi tuo, kad duomenys susiję – turime matavimų poras $(x_1, y_1), (x_2, y_2), \ldots$. Pavyzdžiui, tą pačią dieną matuodami pirmakursių ir antrakursių IQ, gauname dvi nepriklausomas imtis, o du kartus matuodami *tų pačių* pirmo ir antro kurso studentų IQ, gauname priklausomas imtis.

Porinis t kriterijus grindžiamas tuo, kad dviejų normaliųjų atsitiktinių dydžių skirtumas irgi turi normalųjį skirstinį. Radę kiekvienos poros duomenų skirtumus, gauname vieną duomenų aibę, kuriai tinka visi 3.2 skyrelio samprotavimai. Lygindami dviejų priklausomų imčių vidurkius pagal porinį t kriterijų, gauname tas pačias išvadas, kaip ir tikrindami hipotezę apie stebimų porų rezultatų skirtumų vidurkio lygybę nuliui. Iš tikrųjų

$$\begin{aligned}\overline{X} - \overline{Y} &= \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \\ &\quad - \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \frac{1}{n}((X_1 - Y_1) + \cdots + (X_n - Y_n)).\end{aligned}$$

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1. *Duomenys.* Intervalinių duomenų poros $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ gautos matuojant du priklausomus normaliuosius atsitiktinius dydžius $X \sim \mathcal{N}(\mu_X, \sigma_1^2)$ ir $Y \sim \mathcal{N}(\mu_Y, \sigma_2^2)$. Vidurkiai μ_X, μ_Y ir dispersijos σ_1^2, σ_2^2 nežinomi.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \mu_X = \mu_Y, \\ H_1\colon \mu_X \neq \mu_Y. \end{cases} \tag{3.4.10}$$

3 *Kriterijaus statistika.* Randame porų duomenų skirtumus:

$$d_1 = x_1 - y_1, \qquad d_2 = x_2 - y_2, \qquad \ldots, \qquad d_n = x_n - y_n.$$

Apskaičiuojame

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_d^2/n}} = \frac{\overline{d}}{\sqrt{s_d^2/n}}; \tag{3.4.11}$$

čia $\overline{x}$, $\overline{y}$ yra imčių vidurkiai, $\overline{d} = (d_1 + \cdots + d_n)/n$ – skirtumų vidurkis, o $s_d^2 = (d_1^2 + \cdots + d_n^2 - n\overline{d}^2)/(n-1)$ – skirtumų dispersija.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama*, jeigu $|t| > t_{\alpha/2}(n-1)$. Čia $t_{\alpha/2}(n-1)$ yra Stjudento skirstinio su $(n-1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|t| \leqslant t_{\alpha/2}(n-1)$.

3.4.6 pavyzdys. Savininkas nori palyginti dviejų savo kavinių pelną. Abiejų kavinių keturiolikos dienų pelnas litais pateiktas 3.4.2 lentelėje.

3.4.2 lentelė

Savaitės diena	I kavinė	II kavinė	Savaitės diena	I kavinė	II kavinė
Pirmadienis	2531	2262	Pirmadienis	2310	2131
Antradienis	3270	3110	Antradienis	2912	2750
Trečiadienis	3352	3063	Trečiadienis	3582	3334
Ketvirtadienis	4831	4342	Ketvirtadienis	4463	4278
Penktadienis	6351	5940	Penktadienis	6440	6091
Šeštadienis	6910	6570	Šeštadienis	7021	6831
Sekmadienis	6530	6030	Sekmadienis	6430	6020

Iš pirmo žvilgsnio atrodytų, kad kavinės nesusijusios, todėl ir jų duomenys sudaro nepriklausomas imtis. Pritaikę t kriterijų dviem nepriklausomoms imtims, gautume, kad vidutinio kavinių pelno skirtumai yra nereikšmingi. Tačiau net paviršutiniškai peržvelgę duomenis, pastebime, kad *kiekvieną* dieną pirmoji kavinė gauna didesnį pelną už antrąją. Vadinasi, pirmoji kavinė dirba pelningiau. Prieštara tarp statistinių išvadų ir šio fakto atsirado todėl, kad yra dvi priklausomos imtys, t. y. duomenų *poros*. Priklausomybė atsirado ne todėl, kad vienos kavinės gautas pelnas turi įtakos kitos kavinės pelnui. Priklausomybę lemia savaitės diena. Abiejų kavinių lankytojų skaičius skirtingomis savaitės dienomis yra nevienodas – mažiausias pirmadieniais ir didžiausias savaitgaliais. Todėl šio pavyzdžio atveju tikslingiau taikyti porinį t kriterijų. Randame $\overline{x} = 4780{,}9$, $\overline{y} = 4482{,}3$, $t = 9{,}33$, $t_{0{,}025}(13) = 2{,}16$. Kadangi $t > 2{,}16$, tai hipotezė apie vidurkių lygybę atmetama. Taigi nustatėme statistiškai reikšmingą vidutinio dviejų kavinių dienos pelno skirtumą.

Vienpusėms alternatyvoms bei kriterijaus modifikacijoms sprendimai priimami visiškai analogiškai. Tarkime, H_0: $\mu_X - \mu_Y = C$ (vidurkių lygybę atitinka atvejis $C = 0$). Tuomet vietoje (3.4.11) apskaičiuojama t':

$$t' = \frac{\overline{x} - \overline{y} - C}{\sqrt{s_d^2/n}} = \frac{\overline{d} - C}{\sqrt{s_d^2/n}}. \tag{3.4.12}$$

Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.3 lentelėje.

3.4.3 lentelė. H_0: $\mu_X - \mu_Y = C$

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\mu_X - \mu_Y \neq C$	$\lvert t' \rvert > t_{\alpha/2}(n-1)$	$\lvert t' \rvert \leqslant t_{\alpha/2}(n-1)$
$\mu_X - \mu_Y > C$	$t' > t_\alpha(n-1)$	$t' \leqslant t_\alpha(n-1)$
$\mu_X - \mu_Y < C$	$t' < -t_\alpha(n-1)$	$t' \geqslant -t_\alpha(n-1)$

3.4.7 pavyzdys. Psichologas nori patikrinti, ar nemiegota naktis stipriai sulėtina vairuotojų reakcijos greitį. Milisekundėmis buvo išmatuotas bandomųjų reakcijos greitis stabdant automobilį prieš ir po nemiegotos nakties. Psichologas mano, kad vidutiniškai greitis sulėtėja ne mažiau kaip 7 milisekundėmis. Duomenys pateikti 3.4.4 lentelėje.

3.4.4 lentelė

X Y

Žmogus	Ryte	Išvakarėse	Žmogus	Ryte	Išvakarėse
1	47	34	9	51	38
2	52	42	10	44	29
3	48	41	11	38	29
4	36	27	12	30	21
5	43	34	13	44	34
6	53	41	14	48	41
7	54	51	15	43	42
8	50	47	16	43	33

Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$.

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \mu_X - \mu_Y = 7, \\ H_1\colon \mu_X - \mu_Y > 7. \end{cases}$$

Apskaičiuojame $\overline{d} = 8{,}75$; $s_d^2 = 14{,}86$; $n = 16$; $t' = 1{,}81$. Kadangi $t' > 1{,}75 = t_{0{,}05}(15)$, tai hipotezę H_0 atmetame. Liko alternatyva H_1: $\mu_X - \mu_Y > 7$. Gavome statistiškai reikšmingą patvirtinimą, kad vidutinis stabdymo greitis po nemiegotos nakties sulėtėja ne mažiau kaip 7 milisekundėmis.

Kriterijus naudojant SPSS paketą. Tarkime, reikšmingumo lygmuo yra α, H_0: $\mu_X = \mu_Y$, o sprendžiant uždavinį gautoji p-reikšmė lygi p. Tuomet:

1. Jeigu H_1: $\mu_X \neq \mu_Y$, tai H_0 *atmetama*, kai $p < \alpha$. Hipotezė H_0 neatmetama, jeigu $p \geqslant \alpha$.
2. Jeigu H_1: $\mu_X > \mu_Y$, ir $\overline{x} > \overline{y}$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} > \overline{y}$ ir $p \geqslant 2\alpha$ arba $\overline{x} \leqslant \overline{y}$.
3. Jeigu H_1: $\mu_X < \mu_Y$, ir $\overline{x} < \overline{y}$, tai H_0 *atmetama*, kai $p < 2\alpha$. Hipotezė H_0 neatmetama, jeigu $\overline{x} < \overline{y}$ ir $p \geqslant 2\alpha$ arba $\overline{x} \geqslant \overline{y}$.

3.4.8 pavyzdys. Šešiolika pirkėjų balais įvertino siūlomos prekės patrauklumą (didesnis balas rodo didesnį norą prekę pirkti) prieš ir po reklamos. Norima nustatyti, ar reklama padidino prekės patrauklumą ($\alpha = 0{,}05$). Gauti duomenys (prieš, po): (11, 18), (9, 15), (11, 9), (14, 17), (15, 11), (11, 17), (12, 11), (11, 19), (14, 13), (8, 17), (10, 17), (6, 9), (9, 11), (12, 9), (15, 13), (14, 15). SPSS paketu gauti rezultatai pateikti 3.4.4 paveiksle. Pirmojoje lentelėje pateikti vertinimų vidurkiai (11,37 prieš ir 13,81 po reklamos) bei vertinimų standartiniai nuokrypiai (2,6 prieš ir 3,48 po reklamos). Antrojoje lentelėje įvertintas skirtumas $(x - y)$. Kriterijaus reikšmė yra stulpelyje 't' (−2,265), laisvės laipsniai – stulpelyje 'df' (15), o p-reikšmė – stulpelyje 'Sig. (2-tailed)' (0,039). Kadangi $\overline{x} = 11{,}37 < 13{,}81 = \overline{y}$ ir $p = 0{,}039 \leqslant 0{,}1 = 2 \cdot 0{,}05$, tai hipotezė H_0 atmetama. Taigi gavome statistiškai reikšmingą patvirtinimą, kad po reklamos vidutinis prekės patrauklumas padidėjo.

PAIRED SAMPLES STATISTICS

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Prieš	11.3750	16	2.6045	.6511
	Po	13.8125	16	3.4875	.8719

PAIRED SAMPLES TEST

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 Prieš – po	-2.4375	4.3046	1.0761	-4.7312	-.1438	-2.265	15	.039

3.4.4 pav. Porinis t kriterijus. SPSS rezultatai

Dažnai planuojant eksperimentą stengiamasi sudaryti specialias priklausomas imtis, parenkant vadinamąsias *suderintas poras*. Kartais tai nesunku. Pavyzdžiui, norint ištirti, kurios iš 2 trąšų efektyvesnės, užtenka pusę kiekvieno bandomojo laukelio (laukelių dirvų savybės skiriasi) tręšti vienomis trąšomis, o pusę – kitomis. Kartais suderintoms poroms sudaryti reikia itin sudėtingai planuoti eksperimentą. Pavyzdžiui, norint ištirti 2 mokymo metodų efektyvumą, negalima jų abiejų taikyti tiems patiems žmonėms. Todėl iš pradžių sudaromos bandymo dalyvių poros. Kiekvienos dalyvių poros atstovai parenkami taip, kad tiriamojo reiškinio atžvilgiu nesiskirtų – dažniausiai abu yra tos pačios lyties, turintys tą patį išsilavinimą ir pan. Idealiu atveju dalyvių porą sudaro identiški dvyniai. Pagrindinis reikalavimas parenkant dalyvių poras yra šis – tiriamojo reiškinio prasme vienos poros atstovų skirtumai turi būti mažesni nei skirtingų porų. Kai dalyvių poros sudarytos, vienas mokymo metodas taikomas dalyvių porų pirmųjų atstovų grupei, kitas – antrųjų atstovų grupei. Pasibaigus mokymams ir įvertinus žinių lygį (duomenys turi būti intervaliniai!), gaunami pirmosios dalyvių poros laimėjimai (x_1, y_1), antrosios dalyvių poros laimėjimai ir pan. Gautiems rezultatams jau galima taikyti porinį t kriterijų.

Suderintos poros

4.3. Hipotezė apie dviejų dispersijų lygybę

Nuo to, ar dispersijas galima laikyti lygiomis, priklauso Stjudento kriterijaus statistika (žr. 4.1). Dispersijų lygybė tampa svarbi ir tiriant: ar dviejų rūšių vertybinių popierių biržos kainos vienodai stabilios; ar moterys, vertindamos politikus, labiau linkusios į kraštutinumus nei vyrai; ar dviejų grupių testų rezultatai vienodai homogeniški ir pan. Šiame skyrelyje nagrinėsime dvi situacijas – kai lyginamos imtys yra nepriklausomos ir kai priklausomos.

4.3.1. Nepriklausomos imtys

Tarkime, atsitiktinės imtys $(X_1, X_2, \ldots, X_n)$ ir $(Y_1, Y_2, \ldots, Y_m)$ gautos stebint du *nepriklausomus* normaliuosius atsitiktinius dydžius, kurių dispersijos σ_X^2 ir σ_Y^2 nežinomos. Norime patikrinti hipotezę H_0: $\sigma_X^2 = \sigma_Y^2$. Dispersijų įverčiai yra S_X^2 ir S_Y^2. Dispersijų lygybės kriterijus grindžiamas tuo, kad jei H_0 teisinga, tai santykis

$$\mathcal{F} = \frac{S_X^2}{S_Y^2}$$

turi Fišerio skirstinį su $(n-1)$ ir $(m-1)$ laisvės laipsnių. Pažymėtina, kad čia nagrinėjamas ne parametrų įverčių skirtumas, o jų santykis.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Dvi intervalinių duomenų imtys $(x_1, x_2, \ldots, x_n)$ ir $(y_1, y_2, \ldots, y_m)$ gautos matuojant du nepriklausomus normaliuosius atsitiktinius dydžius, kurių dispersijos σ_X^2 ir σ_Y^2.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \sigma_X^2 = \sigma_Y^2, \\ H_1\colon \sigma_X^2 \neq \sigma_Y^2. \end{cases} \tag{3.4.13}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$F = \frac{s_x^2}{s_y^2}; \tag{3.4.14}$$

čia s_x^2, s_y^2 yra imčių dispersijos.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (dispersijos statistiškai reikšmingai skiriasi), jeigu $F > F_{\alpha/2}(n-1, m-1)$ arba $F < F_{1-\alpha/2}(n-1, m-1)$. Čia $F_{\alpha/2}(n-1, m-1)$ yra Fišerio skirstinio su $(n-1)$ ir $(m-1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $F_{1-\alpha/2}(n-1, m-1) \leqslant F \leqslant F_{\alpha/2}(n-1, m-1)$.

Kritines reikšmes $F_{\alpha/2}(n-1, m-1)$ galima rasti priedo 5 lentelėje. Reikia nepamiršti, kad

$$F_{\alpha/2}(n, m) = \frac{1}{F_{1-\alpha/2}(m, n)}, \tag{3.4.15}$$

todėl lentelėse užtenka nurodyti tik dalį visų kritinių reikšmių.

3.4.9 pavyzdys. Investuotojas nori palyginti dviejų vertybinių popierių kainų stabilumą. Trisdešimt vieną sesiją stebėjęs kainų kitimą, investuotojas gavo tokią informaciją: $\overline{x} = 35$, $s_x = 7{,}6$ Lt ir $\overline{y} = 50$, $s_y = 4{,}6$ Lt. Investuotoją domina ne vidutinės vertybinių popierių kainos, o jų stabilumas. Imkime $\alpha = 0{,}1$.

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \sigma_X^2 = \sigma_Y^2, \\ H_1\colon \sigma_X^2 \neq \sigma_Y^2. \end{cases}$$

Apskaičiuojame

$$F = \frac{(7{,}6)^2}{(4{,}6)^2} = 2{,}73.$$

Kadangi $F > 1{,}84 = F_{0{,}05}(30, 30)$, tai H_0 *atmetama.* Taigi vertybinių popierių kainų stabilumas statistiškai reikšmingai skiriasi.

Pastaba. Kadangi F yra arba didesnis už 1, arba mažesnis, tai *tą patį* kriterijų galima suformuluoti šitaip:

$$\mathcal{F} = \frac{\text{didesnioji dispersija}}{\text{mažesnioji dispersija}}.$$

Tuomet F yra ne mažesnis už 1, o hipotezė apie dispersijų lygybę atmetama, kai $F > F_{\alpha/2}(n_1 - 1, n_2 - 1)$. Čia n_1 yra imties su didesniąja dispersija elementų skaičius; n_2 – imties su mažesniąja dispersija elementų skaičius.

Suformuluosime kai kurias kriterijaus modifikacijas. Tegul H_0: $\sigma_X^2 = C\sigma_Y^2$, čia $C > 0$ konstanta. Tuomet skaičiuojama

$$F' = \frac{s_x^2}{Cs_y^2}. \tag{3.4.16}$$

Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.5 lentelėje.

3.4.5 lentelė. H_0: $\sigma_X^2 = C\sigma_Y^2$, kai imtys nepriklausomos

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\sigma_X^2 \neq C\sigma_Y^2$	$F < F_{1-\alpha/2}(n-1, m-1)$ arba $F > F_{\alpha/2}(n-1, m-1)$	$F_{1-\alpha/2}(n-1, m-1) \leqslant F \leqslant F_{\alpha/2}(n-1, m-1)$
$\sigma_X^2 > C\sigma_Y^2$	$F > F_\alpha(n-1, m-1)$	$F \leqslant F_\alpha(n-1, m-1)$
$\sigma_X^2 < C\sigma_Y^2$	$F < F_{1-\alpha}(n-1, m-1)$	$F \geqslant F_{1-\alpha}(n-1, m-1)$

3.4.10 pavyzdys. Grandines gaminanti firma svarsto, ar diegti naują grandžių gamybos technologiją. Grandinę sudaro grandys. Vidutinis grandies stiprumas tiek taikant senąją, tiek naująją technologiją toks pat. Tačiau teigiama, kad naująja technologija gamintų grandžių stiprumo įvairovė daugiau kaip dukart mažesnė, nei taikant senąją technologiją. Firma nori patikrinti, ar duomenys neprieštarauja šiam teiginiui. Išmatavus 41 senu metodu gamintą grandį, rasta, kad jų stiprumo dispersija $s_x^2 = 310$. Išmatavus 61 nauju metodu gamintą grandį, gauta $s_y^2 = 120$. Tarkime, kad reikšmingumo lygmuo $\alpha = 0{,}05$.

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \sigma_X^2 = 2\sigma_Y^2, \\ H_1\colon \sigma_X^2 > 2\sigma_Y^2. \end{cases}$$

Apskaičiuojame $F = 310/(2 \cdot 120) = 1{,}29$. Kadangi $F \leqslant 1{,}59 = F_{0,05}(40, 60)$, tai hipotezė H_0 neatmetama. Taigi duomenys neleidžia teigti, kad dispersija sumažėjo daugiau nei dvigubai. Beje, duomenys statistiškai reikšmingai patvirtina, kad grandžių stiprumo dispersija sumažėjo daugiau nei pusantro karto (įsitikinkite!).

Vidutinis grandies stiprumas tas pats,
vidutinis grandinės stiprumas – ne

Kriterijus dispersijų lygybei tikrinti netinka, jeigu duomenys yra gauti stebint kintamuosius, kurie nėra normaliai pasiskirstę.

Primename, kad SPSS paketu dviejų nepriklausomų dispersijų lygybė tikrinama Levenė kriterijumi, kuris automatiškai taikomas naudojant t kriterijų.

4.3.2. Priklausomos imtys

Tarkime, kad 62 studentai laikė tiksliųjų ir humanitarinių mokslų testus (kiekvieno testo maksimalus balų skaičius yra 100). Testų rezultatai koreliuoja ($r = 0{,}84$). Norima patikrinti, ar abiejų testų rezultatai vienodai homogeniški. Tiksliųjų mokslų testo rezultatų dispersija $s_x^2 = 100{,}2$, o humanitarinių mokslų $s_y^2 = 91{,}7$. Nagrinėjamu atveju turime dvi priklausomas imtis. Išvada grindžiama tuo, kad normuotas dispersijų įverčių skirtumas turi asimptotinį Stjudento skirstinį.

Suformuluosime nagrinėjamojo uždavinio sprendimo etapus:

1 *Duomenys.* Dvi intervalinių duomenų imtys $(x_1, x_2, \ldots, x_n)$ ir $(y_1, y_2, \ldots, y_m)$ gautos matuojant du priklausomus normaliuosius atsitiktinius dydžius, kurių dispersijos yra σ_X^2 ir σ_Y^2.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \sigma_X^2 = \sigma_Y^2, \\ H_1\colon \sigma_X^2 \neq \sigma_Y^2. \end{cases} \tag{3.4.17}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{s_x^2 - s_y^2}{\sqrt{4s_x^2 s_y^2(1-r^2)/(n-2)}}; \tag{3.4.18}$$

čia s_x^2, s_y^2 yra imčių dispersijos, r – imčių koreliacija.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (dispersijos statistiškai reikšmingai skiriasi), jeigu $|t| > t_{\alpha/2}(n-2)$. Hipotezė H_0 neatmetama, jeigu $|t| \leqslant t_{\alpha/2}(n-2)$. Čia $t_{\alpha/2}(n-2)$ yra Stjudento skirstinio su $(n-2)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė.

3.4.11 pavyzdys. Grįžkime prie skyrelio pradžioje nagrinėto pavyzdžio. Tegul $\alpha = 0{,}05$. Apskaičiuojame

$$t = \frac{100{,}2 - 91{,}7}{\sqrt{4 \cdot 100{,}2 \cdot 91{,}7(1-0{,}84^2)/60}} = 0{,}6329.$$

Kadangi $t < 2 = t_{0{,}025}(60)$, duomenys nepatvirtina, kad dispersijos skiriasi.

Vienpusėms alternatyvoms naudojama ta pati t, apibrėžiama (3.4.18) formule. Vienpusei alternatyvai H_1: $\sigma_X^2 < \sigma_Y^2$ parenkama kritinė sritis $W = (-\infty, -t_\alpha(n-2))$, t. y. H_0 atmetama, kai $t < -t_\alpha(n-2)$. Vienpusei alternatyvai H_1: $\sigma_X^2 > \sigma_Y^2$ parenkama kritinė sritis $W = (t_\alpha(n-2), \infty)$, t. y. H_0 atmetama, kai $t > t_\alpha(n-2)$.

4.4. Hipotezė apie dviejų proporcijų lygybę

Tarkime, kad mus domina:

ar vakcinacija sumažina labai reto susirgimo tikimybę;

ar dviejose populiacijose žmonių, turinčių tam tikrą genų anomaliją, skaičius skirtingas;

ar rizika gauti infarktą didesnė būnant A lygio valdininku, ar dirbant mokytoju ir pan.

Visais šiais atvejais reikia lyginti du nepriklausomus dvireikšmius kintamuosius.

Tarkime, X yra dvireikšmis kintamasis, stebimas pirmoje populiacijoje, Y – antrojoje. Tegul X ir Y reikšmės koduotos simboliais 0 ir 1. Pažymėkime $p_1 = P(X = 1)$, $p_2 = P(Y = 1)$. Tuomet $X \sim \mathcal{B}(1, p_1)$, $Y \sim \mathcal{B}(1, p_2)$. Pavyzdžiui, X kintamasis nusako kairiarankiškumą Lietuvoje: 1 – žmogus kairiarankis, 0 – nekairiarankis. Analogiškai Y – kairiarankiškumą kaimyninėje valstybėje. Tuomet p_1 yra kairiarankių Lietuvos gyventojų dalis, p_2 – kairiarankių kaimyninės valstybės gyventojų dalis.

Kintamuosius matuojant atitinkamai n ir m kartų, duomenys yra dvireikšmiai, juos sudaro tik vienetai ir nuliai. Imties vienetų skaičiai S_1 ir S_2 turi binominius skirstinius $S_1 \sim \mathcal{B}(n, p_1)$, $S_2 \sim \mathcal{B}(m, p_2)$. Taigi, tikrindami hipotezę apie p_1 lygybę p_2, lyginame du binominius skirstinius. Hipotezės tikrinimas grindžiamas tuo, kad

$$P(S_1 = j | S_1 + S_2 = s) = \frac{\binom{n}{j}\binom{m}{s-j}}{\binom{m+n}{s}}, \tag{3.4.19}$$

jei tik hipotezė H_0: $p_1 = p_2$ teisinga. Matome, kad sąlyginė tikimybė turi hipergeometrinį skirstinį. Tuo remiantis sudarome tikslų kriterijų proporcijų lygybei tikrinti.

Statistikas pabandė anginą gydyti kankorėžių uogiene, ir pacientas pasveiko. Statistikas išspausdino straipsnį „Kankorėžių uogienė išgydo anginą". Antrasis angina sirgęs pacientas nuo kankorėžių uogienės numirė. Statistikas išspausdino dar vieną straipsnį „Kruopštesni tyrimai atskleidė, kad kankorėžių uogienė išgydo 50% angina sergančių ligonių".

Hipotezės apie proporcijų lygybę tikrinimo etapai yra tokie:

1 *Duomenys.* Stebime du nepriklausomus binominius kintamuosius $X \sim \mathcal{B}(1, p_1)$ ir $Y \sim \mathcal{B}(1, p_2)$. Kintamąjį X stebime n kartų, kintamąjį Y stebime m kartų. Gauname dvi dvireikšmių duomenų aibes, kurias sudaro nuliai ir vienetai. Tarkime, kad pirmoje imtyje yra k_1, o antrojoje – k_2 vienetų.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p_1 = p_2, \\ H_1\colon\ p_1 \neq p_2. \end{cases} \tag{3.4.20}$$

3 *Kriterijaus statistika.* Pažymime $s = k_1 + k_2$ ir apskaičiuojame sumas:

$$Z_1 = \sum_{j=k_1}^{\min(s,n)} \binom{n}{j}\binom{m}{s-j} \Big/ \binom{m+n}{s}, \tag{3.4.21}$$

$$Z_2 = \sum_{\max(0,s-m)}^{k_1} \binom{n}{j}\binom{m}{s-j} \Big/ \binom{m+n}{s}. \tag{3.4.22}$$

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (taigi p_1 skiriasi nuo p_2), jeigu $Z_1 < \alpha/2$ arba $Z_2 < \alpha/2$. Kitais atvejais hipotezė H_0 neatmetama.

3.4.12 pavyzdys. Ar galima teigti, kad vienas skrandžio operavimo metodas geresnis už kitą, jei taikant pirmąjį iš 300 pacientų mirė 3, o taikant antrąjį iš 150 mirė 2? ($\alpha = 0{,}1$.)

Sprendimas. Šiuo atveju $k_1 = 3$, $k_2 = 2$, $n = 300$, $m = 150$, $s = 5$. Apskaičiuojame

$$Z_1 = \sum_{j=3}^{5} \binom{300}{j}\binom{150}{5-j} \Big/ \binom{450}{5} = 0{,}79.$$

$$Z_2 = \sum_{j=0}^{3} \binom{300}{j}\binom{150}{5-j} \Big/ \binom{450}{5} = 0{,}54.$$

Kadangi $Z_1 \geqslant 0{,}05$ ir $Z_2 \geqslant 0{,}05$, tai H_0 atmesti nėra pagrindo. Statistiškai reikšmingo mirčių skaičiaus skirtumo operuojant skirtingais metodais neradome.

Kriterijų galima modifikuoti ir vienpusių alternatyvų atvejais. Jeigu H_0: $p_1 = p_2$, tai sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.6 lentelėje. Reikšmingumo lygmuo yra α, o Z_1 ir Z_2 apibrėžtos (3.4.21) ir (3.4.22) lygybėmis.

3.4.6 lentelė. H_0: $p_1 = p_2$, kai kriterijus tikslus

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p_1 \neq p_2$	$Z_1 < \alpha/2$ arba $Z_2 < \alpha/2$	$Z_1 \geqslant \alpha/2$ ir $Z_2 \geqslant \alpha/2$
$p_1 > p_2$	$Z_1 < \alpha$	$Z_1 \geqslant \alpha$
$p_1 < p_2$	$Z_2 < \alpha$	$Z_2 \geqslant \alpha$

3.4.13 pavyzdys. Ištyrus 300 ešerių, sugautų Senupėje ties trąšų gamykla, pelekų mutacija nustatyta 5 ešeriams. Ištyrus 200 bandymų stoties tvenkinyje pagautų ešerių, mutacija nustatyta 1 ešeriui. Ar Senupėje statistiškai reikšmingai daugiau mutacijų? ($\alpha = 0{,}05$.)

Sprendimas. Turime $k_1 = 5$, $k_2 = 1$, $s = 6$, $n = 300$, $m = 200$. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p_1 = p_2, \\ H_1\colon\ p_1 > p_2. \end{cases}$$

Apskaičiuojame $Z_1 = 0{,}23$. Kadangi $Z_1 > 0{,}05$, tai hipotezės H_0 atmesti nėra pagrindo. Taigi statistiškai reikšmingo mutacijų skaičiaus skirtumo nerasta.

Tikslų kriterijų patogu taikyti, jeigu k_1, k_2 yra nedideli. Priešingu atveju dėl milžiniškos skaičiavimų apimties tikslus kriterijus netaikomas. Tačiau praktiškai k_1, k_2 maži būna retai. Tarkime:

rinkos ekspertas nori sužinoti, ar vyrų, mėgstančių Alytaus šampaną, procentas pirkėjų populiacijoje yra toks pat kaip ir moterų;

kandidatas į parlamento narius nori išsiaiškinti, ar jis vienodai populiarus tarp jaunimo ir tarp pensininkų;

medikas nori nustatyti, ar taikant naują metodą sėkmingų operacijų dalis didesnė, nei operuojant senuoju metodu;

sociologą domina, ar pritariančiųjų mirties bausmei Europos šalyse yra 2% mažiau nei Kanadoje.

Visais šiais atvejais k_1, k_2 yra nemaži skaičiai ir taikoma normalioji aproksimacija:

1 *Duomenys.* Tegul kaip ir anksčiau galioja tos pačios prielaidos ir žymenys yra tie patys, t.y. stebime du nepriklausomus binominius kintamuosius. Pirmoje n elementų imtyje yra k_1 vienetų (likę – nuliai), antrojoje m elementų imtyje yra k_2 vienetų (likę – nuliai).

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p_1 - p_2 = C, \\ H_1\colon\ p_1 - p_2 \neq C. \end{cases} \tag{3.4.23}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - C}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/n + \widehat{p}_2(1-\widehat{p}_2)/m}}; \tag{3.4.24}$$

čia $\widehat{p}_1 = k_1/n$, $\widehat{p}_2 = k_2/m$.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama*, jeigu $|Z| > z_{\alpha/2}$. Čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leqslant z_{\alpha/2}$.

Keletas dažnai naudojamų z_α reikšmių pateikta 3.1 skyrelyje (žr. p. 155). Sprendimo taisyklės, esant skirtingoms alternatyvoms, nurodytos 3.4.7 lentelėje.

3.4.7 lentelė. $H_0\colon\ p_1 - p_2 = C$, kai skirstinys normalusis

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$p_1 - p_2 \neq C$	$\lvert Z\rvert > z_{\alpha/2}$	$\lvert Z\rvert \leqslant z_{\alpha/2}$
$p_1 - p_2 > C$	$Z > z_\alpha$	$Z \leqslant z_\alpha$
$p_1 - p_2 < C$	$Z < -z_\alpha$	$Z \geqslant -z_\alpha$

$p_1 < p_2$

Jeigu H_0: $p_1 = p_2$ (t. y. $C = 0$), tai 3.4.7 lentelėje vietoje Z reikia imti

$$Z' = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\bar{p}(1-\bar{p})(1/n + 1/m)}}; \tag{3.4.25}$$

čia $\widehat{p}_1 = k_1/n$, $\widehat{p}_2 = k_2/m$, $\bar{p} = (k_1 + k_2)/(n+m)$.

3.4.14 pavyzdys. Sociologo klausimyną užpildyti sutiko 100 iš 200 vyrų ir 150 iš 270 moterų. Ar galima manyti, kad sutinkančių atsakinėti vyrų dalis statistiškai reikšmingai skiriasi nuo moterų dalies? Tegul $\alpha = 0{,}1$.

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon\ p_1 = p_2, \\ H_1\colon\ p_1 \neq p_2. \end{cases}$$

Apskaičiuojame:

$$\widehat{p}_1 = 100/200 = 0{,}5, \qquad \widehat{p}_2 = 150/270 = 0{,}555, \qquad \bar{p} = 250/470 = 0{,}532,$$

$$Z' = (0{,}5 - 0{,}555)/(\sqrt{0{,}532 \cdot 0{,}468(1/200 + 1/270)}) = -1{,}181.$$

Kadangi $|Z'| = 1{,}181 < 1{,}64 = z_{0{,}05}$, tai H_0 *neatmetama*. Sutinkančių atsakinėti vyrų ir moterų dalių skirtumo nerasta.

3.4.15 pavyzdys. Mėnesio pradžioje apklausus 100 atsitiktinai parinktų respondentų, paaiškėjo, kad 20 iš jų yra reguliarūs radijo stoties „Radiokavinė“ klausytojai. Radijo stoties savininkas visą mėnesį aktyviai reklamavo savo stotį. Mėnesio gale apklausus 150 atsitiktinai parinktų respondentų, paaiškėjo, kad iš jų šios stoties nuolat klausosi 50. Savininko nuomone, reklama atsiperka, jeigu klausytojų skaičius padidėja daugiau kaip 3 procentus. Ar duomenys leidžia teigti, kad reklama atsipirko? ($\alpha = 0{,}05$.)

Sprendimas. Tegul p_1 žymi reguliarių radijo stoties klausytojų dalį mėnesio pabaigoje, o p_1 – mėnesio pradžioje. Reklama atsipirks, jeigu $p_1 > p_2 + 0{,}03$. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p_1 - p_2 = 0{,}03, \\ H_1\colon\ p_1 - p_2 > 0{,}03. \end{cases}$$

Apskaičiuojame $\widehat{p_1} = 50/150 = 0{,}33$, $\widehat{p_2} = 20/100 = 0{,}20$,

$$Z = \frac{0{,}33 - 0{,}20 - 0{,}03}{\sqrt{0{,}33 \cdot 0{,}67/150 + 0{,}20 \cdot 0{,}80/100}} = 1{,}80.$$

Kadangi $Z > 1{,}64 = z_{0{,}05}$, tai hipotezę H_0 atmetame. Duomenys leidžia teigti, kad reguliarių klausytojų skaičius padidėjo daugiau kaip 3 procentus.

3.4.16 pavyzdys. Gamyklos vadovai nori sužinoti, ar dirbant naujomis staklėmis tikrai reikšmingai sumažėja broko. Iš 300 senomis staklėmis gamintų detalių buvo 30 brokuotų; iš 200 naujomis staklėmis gamintų detalių buvo 16 brokuotų. Kadangi naujųjų staklių diegimas susijęs su nemenkomis išlaidomis, hipotezę reikia patikrinti su $\alpha = 0{,}01$.

Sprendimas. Tegul p_1 žymi broko dalį, dirbant senomis staklėmis, o p_2 – dirbant naujosiomis. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon\ p_1 = p_2, \\ H_1\colon\ p_1 > p_2. \end{cases}$$

Apskaičiuojame $\widehat{p_1} = 30/300 = 0{,}10$, $\widehat{p_2} = 16/200 = 0{,}08$, $\bar{p} = (30 + 16)/(300 + 200) = 0{,}092$,

$$Z' = \frac{0{,}10 - 0{,}08}{\sqrt{0{,}092 \cdot 0{,}908(1/300 + 1/200)}} = 0{,}758.$$

Kadangi $Z' = 0{,}758 < 2{,}326 = z_{0{,}01}$, tai hipotezės H_0 neatmetame. Duomenys neleidžia teigti, kad broko procentas sumažėjo.

Hipotezę apie dviejų proporcijų lygybę galima tikrinti ir neparametriniais kriterijais (žr. 5.4).

4.5. Hipotezė apie dviejų koreliacijos koeficientų lygybę

Trečiajame skyriuje aptarėme hipotezę apie koreliacijos koeficiento lygybę skaičiui. Dažnai tyrėją domina ne tik koreliacijos koeficiento didumas, bet ir dviejų koreliacijos koeficientų skirtumas. Pavyzdžiui, norima nustatyti: ar studentų matematinio testo rezultatai su kalbos testo rezultatais koreliuoja stipriau nei studenčių; ar matematikos ir gimtosios kalbos testų rezultatų koreliacija yra stipresnė nei matematikos ir užsienio kalbos testų. Pirmu atveju lyginame du nepriklausomų imčių koreliacijos koeficientus, antruoju – priklausomų imčių. Kiekvieną modelį aptarsime atskirai.

4.5.1. Nepriklausomų imčių atvejis

Tarkime, stebime dvi nepriklausomas intervalinių kintamųjų poras (X_1, Y_1) ir (X_2, Y_2). Atsitiktinės imtis sudaro poros $(X_{11}, Y_{11}), (X_{12}, Y_{12}), \ldots, (X_{1n}, Y_{1n})$ ir $(X_{21}, Y_{21}), (X_{22}, Y_{22}), \ldots, (X_{2m}, Y_{2m})$.

Norime nustatyti, ar koreliacija X_1 su Y_1 (pažymime ją ϱ_1) skiriasi nuo koreliacijos X_2 su Y_2 (pažymime ją ϱ_2). Kadangi empirinių koreliacijos koeficientų R_1 ir R_2 skirtumo skirstinys yra asimetriškas, prieš taikydami normaliąją aproksimaciją, naudojamės Fišerio logaritmine transformacija:

$$\mathcal{Z}_1 = \frac{1}{2}\ln\frac{1+R_1}{1-R_1}, \qquad \mathcal{Z}_2 = \frac{1}{2}\ln\frac{1+R_2}{1-R_2}.$$

Kai teisinga hipotezė H_0: $\varrho_1 = \varrho_2$ ir $n > 3$, $m > 3$, statistika

$$(\mathcal{Z}_1 - \mathcal{Z}_2)/\sqrt{1/(n-3) + 1/(m-3)} \approx \mathcal{N}(0, 1). \tag{3.4.26}$$

Remiantis šia formule, sudaromos kritinės sritys.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Dvi porinės imtys (x_{11}, y_{11}), (x_{12}, y_{12}), $\ldots$, (x_{1n}, y_{1n}) ir (x_{21}, y_{21}), (x_{22}, y_{22}), $\ldots$, (x_{2m}, y_{2m}) gautos matuojant du nepriklausomus dvimačius normaliuosius atsitiktinius dydžius (X_1, Y_1) ir (X_2, Y_2). Koreliacija X_1 su Y_1 lygi ϱ_1, koreliacija X_2 su Y_2 lygi ϱ_2. Imčių didumai $n > 3$ ir $m > 3$.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon \varrho_1 = \varrho_2, \\ H_1\colon \varrho_1 \neq \varrho_2. \end{cases} \tag{3.4.27}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$Z = (z_1 - z_2)/\sqrt{1/(n-3) + 1/(m-3)}; \tag{3.4.28}$$

čia $z_i = \frac{1}{2}\ln\dfrac{1+r_i}{1-r_i}$, $i = 1, 2$, o r_1, r_2 yra Pirsono koreliacijos koeficientų R_1 ir R_2 realizacijos, skaičiuojamos pagal (3.3.28) formulę.

4 *Sprendimo priėmimo taisyklė.* Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 atmetama (X_1 ir Y_1 koreliacija skiriasi nuo X_2 ir Y_2 koreliacijos), jeigu $|Z| > z_{\alpha/2}$. Čia $z_{\alpha/2}$ yra standartinio normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|Z| \leqslant z_{\alpha/2}$.

3.4.17 pavyzdys. Psichologas mano, kad vyrų ir moterų matematikos ir gimtosios kalbos testų rezultatų koreliacija skirtinga. Ištyręs 43 vyrų rezultatus, psichologas gavo koreliacijos koeficiento įvertį $r_1 = 0{,}56$. Ištyręs 50 moterų rezultatus, psichologas gavo $r_2 = 0{,}62$. Ar gautieji duomenys patvirtina psichologo hipotezę? ($\alpha = 0{,}1$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon \varrho_1 = \varrho_2, \\ H_1\colon \varrho_1 \neq \varrho_2. \end{cases}$$

Pasinaudoję priedo 7 lentele, randame $z_1 = 0{,}633$, $z_2 = 0{,}725$. Apskaičiuojame

$$Z = (0{,}633 - 0{,}725)/\sqrt{1/40 + 1/47} = -0{,}427.$$

Kadangi $|Z| = 0{,}0779 \leqslant 1{,}64 = z_{0{,}05}$, tai H_0 neatmetame. Psichologo hipotezės duomenys nepatvirtino.

3.4.8 lentelė. H_0: $\varrho_1 = \varrho_2$, kai imtys nepriklausomos

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\varrho_1 \neq \varrho_2$	$\|Z\| > z_{\alpha/2}$	$\|Z\| \leqslant z_{\alpha/2}$
$\varrho_1 > \varrho_2$	$Z > z_\alpha$	$Z \leqslant z_\alpha$
$\varrho_1 < \varrho_2$	$Z < -z_\alpha$	$Z \geqslant -z_\alpha$

Vienpusėms alternatyvoms naudojama ta pati Z, apibrėžiama (3.4.28) formule. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.8 lentelėje.

3.4.18 pavyzdys. Sociologas nori nustatyti, ar IQ ir jo sukurto klausimyno rezultatų priklausomybė vyresniems žmonėms didesnė. Šimto iki 40 metų amžiaus respondentų IQ ir klausimyno rezultatų koreliacijos koeficientas $r_1 = 0{,}49$, o 130 vyresnių nei 40 metų respondentų $r_2 = 0{,}71$ ($\alpha = 0{,}05$).

Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \varrho_1 = \varrho_2, \\ H_1\colon \varrho_1 < \varrho_2. \end{cases}$$

Apskaičiuojame $z_1 = 0{,}536$, $z_2 = 0{,}887$, $Z = (0{,}536 - 0{,}887)/\sqrt{1/97 + 1/127} = -2{,}60$. Kadangi $Z = -2{,}60 < -1{,}64 = z_{0{,}05}$, tai hipotezė H_0 atmetama. Duomenys leidžia teigti, kad vyresnių respondentų rezultatų koreliacija yra statistiškai reikšmingai didesnė negu jaunesnių.

4.5.2. Priklausomų imčių atvejis

Tarkime, stebime tris normaliai pasiskirsčiusius kintamuosius X, Y ir Z. Koreliacija X su Y lygi ϱ_{XY}, X su Z yra ϱ_{XZ}, Y su Z yra ϱ_{YZ}. Norime patikrinti hipotezę apie koreliacijos koeficientų lygybę $\varrho_{XY} = \varrho_{XZ}$. Kriterijaus statistika grindžiama tuo, kad koreliacijos koeficientų skirtumas po specialaus normavimo turi asimptotiškai Stjudento skirstinį, jei tik teisinga hipotezė H_0.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Turime trijų priklausomų normaliai pasiskirsčiusių kintamųjų stebėjimus $(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)$. Koreliacija X su Y lygi ϱ_{XY}, X su Z – ϱ_{XZ}, Y su Z – ϱ_{YZ}. Imties didumas $n > 3$.

2 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon \varrho_{XY} = \varrho_{XZ}, \\ H_1\colon \varrho_{XY} \neq \varrho_{XZ}. \end{cases} \tag{3.4.29}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$t = \frac{(r_{xy} - r_{xz})\sqrt{(n-3)(1 + r_{yz})}}{\sqrt{2(1 - r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz})}}; \tag{3.4.30}$$

čia r_{xy}, r_{xz}, r_{yz} yra Pirsono koreliacijos koeficientų realizacijos, skaičiuojamos pagal (3.3.28) formulę.

Sprendimo priėmimo taisyklė. Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 *atmetama* (X ir Y koreliacija skiriasi nuo X ir Z koreliacijos), jeigu $|t| > t_{\alpha/2}(n-3)$. Čia $t_{\alpha/2}(n-3)$ yra Stjudento skirstinio su $(n-3)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė. Hipotezė H_0 neatmetama, jeigu $|t| \leqslant t_{\alpha/2}(n-3)$.

3.4.19 pavyzdys. Psichologas mano, kad matematikos (kintamasis X) ir gimtosios kalbos (kintamasis Y) testų rezultatų koreliacija skiriasi nuo matematikos ir užsienio kalbos (kintamasis Z) testų rezultatų koreliacijos. Ištyręs 123 atsitiktinai parinktų respondentų rezultatus, psichologas gavo koreliacijos koeficientų įverčius $r_{xy} = 0{,}63$, $r_{xz} = 0{,}79$, $r_{yz} = 0{,}52$. Ar gautieji duomenys patvirtina psichologo hipotezę? ($\alpha = 0{,}05$.)

Sprendimas. Statistinė hipotezė:

$$\begin{cases} H_0\colon \varrho_{XY} = \varrho_{XZ}, \\ H_1\colon \varrho_{XY} \neq \varrho_{XZ}. \end{cases}$$

Apskaičiuojame $t = -3{,}21$. Kadangi $|t| > 1{,}98 = t_{0{,}025}(120)$, tai H_0 atmetame. Psichologo hipotezę duomenys patvirtino.

Vienpusėms alternatyvoms naudojama ta pati t, apskaičiuojama pagal (3.4.30) formulę. Sprendimo taisyklės, esant skirtingoms alternatyvoms, pateikiamos 3.4.9 lentelėje.

3.4.9 lentelė. H_0: $\varrho_{XY} = \varrho_{XZ}$, kai imtys priklausomos

Alternatyva H_1	H_0 atmetama, jeigu	H_0 neatmetama, jeigu
$\varrho_{XY} \neq \varrho_{XZ}$	$\lvert t\rvert > t_{\alpha/2}(n-3)$	$\lvert t\rvert \leqslant t_{\alpha/2}(n-3)$
$\varrho_Y X > \varrho_{XZ}$	$t > t_\alpha(n-3)$	$t \leqslant t_\alpha(n-3)$
$\varrho_{XY} < \varrho_{XZ}$	$t < -t_\alpha(n-3)$	$t \geqslant -t_\alpha(n-3)$

3.4.20 pavyzdys. Polinkiui į šizofreniją nustatyti buvo pasiūlyti du klausimynai (kuo didesnis surinktų balų skaičius, tuo polinkis didesnis). Kiekvienas iš 63 pacientų, turinčių šizofrenijos požymių, užpildė abu klausimynus. Po to kiekvieno paciento šizofrenijos lygį balais (iki 20) įvertino grupė medikų ekspertų. Ekspertų parašytų balų vidurkio ir pirmojo klausimyno rezultatų koreliacijos koeficientas $r_{xy} = 0{,}75$. Ekspertų parašytų balų vidurkio ir antrojo klausimyno rezultatų koreliacijos koeficientas $r_{xz} = 0{,}64$. Ar duomenys leidžia teigti, kad pirmojo klausimyno rezultatai statistiškai reikšmingai stipriau koreliuoja su ekspertų parašytų balų vidurkiu nei antrojo klausimyno rezultatai, jeigu $r_{yz} = 0{,}45$? ($\alpha = 0{,}05$.)

Sprendimas. Formuluojame statistinę hipotezę:

$$\begin{cases} H_0\colon \varrho_{XY} = \varrho_{XZ}, \\ H_1\colon \varrho_{XY} > \varrho_{XZ}. \end{cases}$$

Apskaičiuojame $t = 1{,}43$. Kadangi $t \leqslant 1{,}67 = t_{0{,}05}(60)$, tai hipotezė H_0 neatmetama. Duomenys neleidžia teigti, kad pirmojo klausimyno rezultatų koreliacija statistiškai reikšmingai didesnė už antrojo klausimyno rezultatų koreliaciją.

dispersijų lygybė
koreliacijos koeficientų lygybė
porinis t kriterijus
proporcijų lygybė
Stjudento kriterijus

UŽDAVINIAI

1. Gamyklos vadovai nori nustatyti, ar dėl ligos daugiau darbadienių praleidžiama dieninėje pamainoje, ar naktinėje. Atsitiktinai parinkus 10 naktinės pamainos darbuotojų, paaiškėjo, kad per metus dėl ligos jie praleido 20; 10; 14; 32; 9; 2; 18; 6; 4; 13 dienų. Parinkus 10 dieninės pamainos darbuotojų, paaiškėjo, kad jie praleido 5; 12; 17; 0; 6; 17; 16; 3; 23; 2 dienų. Suformuluokite statistinę hipotezę ir ją patikrinkite ($\alpha = 0,05$).

2. Automobilius gaminanti firma nori žinoti, ar, naudodami dviejų skirtingų rūšių benziną, jos automobiliai nuvažiuoja tiek pat kilometrų. Ištyrus 200 automobilių, naudojusių pirmos rūšies benziną, paaiškėjo, kad 100 km vidutiniškai prireikė 8,2 ℓ benzino ($s_1 = 2\,\ell$). Ištyrus 150 automobilių, naudojusių antros rūšies benziną, paaiškėjo, kad 100 km vidutiniškai prireikė 7,8 ℓ benzino ($s_2 = 1,9\,\ell$). Suformuluokite statistinę hipotezę ir ją patikrinkite ($\alpha = 0,01$).

3. Nepriklausomas ekspertas nori palyginti du retorikos kursus. Trisdešimt du studentai ekspertų komisijai nurodyta tema pasakė po kalbą. Kiekvieno studento pasirodymą komisija įvertino balais (maksimalus balas 60). Po to 16 studentų išklausė pirmąjį retorikos kursą, o kiti 16 – antrąjį. Po kursų kievienas studentas tai pačiai komisijai vėl pasakė kalbą ir gavo įvertinimą. Kiekvieno studento balų, gautų po kurso išklausymo ir prieš kursą, skirtumas pateiktas 3.4.10 lentelėje. Ar galima teigti, kad antrasis retorikos kursas efektyvesnis? ($\alpha = 0,05$.)

3.4.10 lentelė. Retorikos kursai

Kursas							
I	II	I	II	I	II	I	II
15	10	7	11	5	0	−2	10
10	1	12	11	12	8	0	5
5	13	7	13	15	20	2	3
1	0	12	11	5	10	2	10

4. Dvylikai vyno ekspertų dukart buvo pateiktas tas pats šampanas. Pirmąkart šampanas buvo pateiktas butelyje su prancūziška etikete, antrąkart – butelyje su rusiška etikete. Ar galima teigti, kad etiketė turėjo įtakos vertinimams? Ekspertų vertinimai pateikti 3.4.11 lentelėje ($\alpha = 0,1$).

3.4.11 lentelė. Vyno ekspertų išvados

Ekspertas	Etiketė		Ekspertas	Etiketė	
	prancūziška	rusiška		prancūziška	rusiška
a	12	10	g	12	12
b	5	1	h	9	3
c	14	11	i	10	6
d	11	10	j	16	17
e	19	14	k	5	6
f	7	6	l	16	12

3.4.12 lentelė. Automobilių ekspertų išvados

Automobilis	I ekspertas	II ekspertas	Automobilis	I ekspertas	II ekspertas
1	200	200	11	300	350
2	1200	1500	12	3000	2500
3	750	1000	13	500	350
4	2300	3100	14	800	350
5	1300	2000	15	650	500
6	2900	2100	16	1000	1300
7	4000	3500	17	5000	5200
8	3300	3200	18	4300	4150
9	600	400	19	1300	950
10	2550	2350	20	3050	3150

5. Sociologas nori patikrinti, ar dviejų nepriklausomų ekspertų, vertinančių avarijos metu automobiliams padarytą žalą, išvados skiriasi. Abu ekspertai buvo paprašyti litais įvertinti 20 avarijose apdaužytų automobilių remonto kainas. Duomenys pateikti 3.4.12 lentelėje. Kokią išvadą padarys sociologas, jeigu reikšmingumo lygmuo $\alpha = 0{,}05$?

6. Tirdami dviejų policijos komisariatų darbą, sociologai matavo laiką nuo policijos iškvietimo iki jos atvykimo. Pirmojo komisariato policininkai buvo kviesti 30 kartų. Jie vidutiniškai sugaišo 10,72 min (standartinis nuokrypis 7,2 min). Antrojo komisariato policininkai buvo kviesti 32 kartus ir vidutiniškai sugaišo 10,02 min (standartinis nuokrypis 1,2 min). Kuris komisariatas dirba geriau? Suformuluokite statistinę hipotezę ir ją patikrinkite ($\alpha = 0{,}1$).

7. Taikant senąjį mokymo metodą 41 mokiniui, baigiamojo egzamino testo rezultatų sklaida $s_1^2 = 30{,}3$. Taikant naująjį mokymo metodą 31 mokiniui, rezultatų sklaida $s_2^2 = 15$. Ar galima sutikti su naujojo metodo kūrėjų teiginiu, kad naujasis metodas rezultatų sklaidą sumažina ne mažiau nei dvigubai? ($\alpha = 0{,}05$.)

8. Buvo dukart įvertintas 50 studentų konformizmas – tik įstojus į universitetą ir ji baigiant. Ar galima teigti, kad baigiantieji šiuo aspektu vienodesni, jei pradžioje rezultatų dispersija buvo $s_1^2 = 25$, o baigiant $s_2^2 = 12$? Atsakymų koreliacija $r = 0{,}8$ ($\alpha = 0{,}05$).

9. Nepriklausomas ekspertas tiria, kiek kartų garantinio TV taisymo prireikė televizoriams, surinktiems Pietryčių Azijoje, ir kiek – Rytų Europoje. Iš 150 azijinių televizorių garantinio remonto prireikė 4, iš 100 europinių – 2. Ar galima teigti, kad europiniams televizoriams garantinio remonto reikia rečiau? ($\alpha = 0{,}01$.)

10. Literatūros klasikas Juozas E. pareiškė, kad pagal dantų (savo) skausmą gali atpažinti, kurios politinės partijos atstovas šneka. Pasaulinės parapsichologų akademijos centras nusprendė patikrinti, ar Juozas E. iš tikrųjų toks paragabus. Išklausęs 30 politikų kalbų, Juozas teisingai klasifikavo 8 politikus. Tuos pačius 30 politikų išklausęs lietuviškai nesuprantantis Jagai Baba, atsitiktinai spėjo jų partinę priklausomybę ir pataikė 6 kartus. Ar atpažinimo gebėjimais Juozas skiriasi nuo Jagai Babos? ($\alpha = 0{,}05$.)

11. Dvi grupės po 50 žmonių lankė skirtingus parengiamuosius matematikos kursus. Prieš kursus buvo įvertintas kiekvieno lankytojo IQ. Pirmųjų kursų lankytojų baigiamojo

testo rezultatų ir IQ koreliacija yra 0,63; antrųjų kursų lankytojų yra 0,70. Ar galima teigti, kad šių koreliacijos koeficientų skirtumas statistiškai nereikšmingas? ($\alpha = 0{,}1$.)

12. Teisinių paslaugų firma surengė 43 naujai priimtų darbuotojų patikrinimą (kiekvienas iš jų turėjo išspręsti po 30 praktinių užduočių, o sprendimus balais vertino firmos ekspertai). Surinktų balų sumos ir universiteto diplomo pažymių vidurkio koreliacija yra 0,64. Surinktų balų sumos ir ankstesnės patirties (mėn.) koreliacija yra 0,71. Firmos vadovai nori žinoti, ar šie duomenys leidžia teigti, kad darbuotojo žinios labiau priklauso nuo turimos patirties nei nuo diplomo pažymių? Žinoma, kad diplomo pažymių vidurkio ir patirties koreliacija yra 0,4 ($\alpha = 0{,}05$).

5. DAŽNIŲ LENTELĖS

Gimtadienius švęsti sveika. Statistiniai duomenys liudija, kad žmonės, švenčiantys daugiausia gimtadienių, gyvena ilgiausiai.

Kodėl šio skyriaus pavadinimas „Dažnių lentelės“? Prisiminkime pirmąją dalį. Joje nagrinėjome, kaip duomenų aibė užrašoma dažnių lentelėmis, grupuotomis dažnių lentelėmis, porinėmis dažnių lentelėmis ir pan. siekiant koncentruotai pateikti duomenis bei išryškinti jų savybes. Šis statistinių išvadų dalies skyrelis taip pavadintas kitu tikslu – norima pabrėžti, kad čia pateikiami kriterijai taikomi *tiktai duomenų aibėms, užrašytoms dažnių lentelėmis*. Praktiškai šiame skyrelyje aprašomas tik vienas – χ^2 (chi kvadratu) kriterijus bei jo modifikacijos. Jis yra vienas iš populiariausių neparametrinių kriterijų (kiti neparametriniai kriterijai aprašomi II knygoje). Dėl χ^2 išskirtinio populiarumo ir plačios taikymo sferos šis kriterijus vertas atskiro skyrelio.

Ar stojančiųjų gebėjimo testo rezultatai aprašomi normaliuoju skirstiniu? Ar yra priklausomybė tarp žmonių akių ir plaukų spalvos? Ar tarp moterų daugiau religingų nei tarp vyrų? Pirmąjį uždavinį statistikai vadina skirstinių suderinamumo uždaviniu, antrąjį – kintamųjų (dažnai sakoma požymių) nepriklausomumo uždaviniu, trečiąjį – populiacijų homogeniškumo uždaviniu. Visiems šiems uždaviniams spręsti naudojamasi χ^2 kriterijumi.

Panagrinėkime tokią situaciją. Apklausta 100 atsitiktinai parinktų žmonių, ar jie buvo praeitą savaitę kino teatre. Gauti (stebimi) duomenys pateikiami 3.5.1 lentelėje.

Iš apklaustųjų 70% atsakė, kad jie buvo kino teatre praeitą savaitę. Jei nebūtų ryšio tarp lyties ir lankomumo, tai praeitą savaitę kino teatrus būtų aplankę 70% vyrų ir 70% moterų, todėl, apklausę 40 vyrų ir 60 moterų, galėtumėme tikėtis tokių rezultatų kaip 3.5.2 lentelėje.

Nurodyti 3.5.1 ir 3.5.2 lentelių duomenys skiriasi. Jų skirtumai pateikiami 3.5.3 lentelėje. Ar šie skirtumai yra statistiškai reikšmingi? Į šį klausimą atsako χ^2 kriterijus, t. y. juo nustatoma, ar yra priklausomybė tarp lyties ir kino teatrų lankomumo. Šioje konkrečioje situacijoje – skirtumai statistiškai reikšmingi, t. y. lytis ir kino teatrų lankomumas tarpusavyje susiję.

3.5.1 lentelė. Kino teatrų lankymas. Duomenys

	Taip	Ne	Iš viso
Vyrai	20	20	40
Moterys	50	10	60
Iš viso	70	30	100

3.5.2 lentelė. Kino teatrų lankymas. Prognozė

	Taip	Ne	Iš viso
Vyrai	28	12	40
Moterys	42	18	60
Iš viso	70	30	100

3.5.3 lentelė. Kino teatrų lankymas. Duomenų ir prognozės skirtumai

	Taip	Ne	Iš viso
Vyrai	−8	8	0
Moterys	8	−8	0
Iš viso	0	0	0

Iš pradžių aptarsime, kokia teorija grindžiamas χ^2 kriterijaus. Tai leis suvokti, kodėl šis kriterijus taip plačiai naudojamas įvairiems uždaviniams spręsti.

Pastaba. Toliau šiame skyrelyje simboliais O_{ij}, E_{ij} žymėsime atitinkamų gardelių (langelių) stebimus bei tikėtinus (prognozuojamus) dažnius, o o_{ij}, e_{ij} – jų realizacijas. Pavyzdžiui, $o_{12} = 20$, $e_{12} = 12$ ir pan.

5.1. Teoriniai modeliai

Aptarsime du teorinius modelius.

1 *Pirmasis modelis.* Atlikus eksperimentą, būtinai įvyksta vienas iš nesutaikomų įvykių $A_1, A_2, \ldots, A_k$. Kiekvieno eksperimento įvykio A_i tikimybė įvykti lygi $p_i > 0$ $(i = 1, \ldots, k;\ p_1 + p_2 + \cdots + p_k = 1)$. Atliekama n nepriklausomų eksperimentų. Pažymėkime ν_i $(i = 1, \ldots, k)$ įvykio A_i įvykimų skaičių, atlikus n eksperimentų. Apibrėžkime atsitiktinį dydį χ^2 taip:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}. \tag{3.5.1}$$

Kai n neaprėžtai didėja, (3.5.1) formule nurodytos statistikos χ^2 skirstinys artėja prie χ^2 skirstinio su $(k-1)$ laisvės laipsnių (t. y. dideliems n jį galima aproksimuoti χ^2 skirstiniu).

2 *Antrasis modelis.* Prielaidos tos pačios kaip ir pirmojo modelio, tik tikimybė p_i priklauso nuo s nežinomų parametrų $p_i = p_i(\alpha_1, \alpha_2, \ldots, \alpha_s)$. Įstatę į (3.5.1) formulę parametrų įverčius $\widehat{\alpha_1}, \widehat{\alpha_2}, \ldots, \widehat{\alpha_s}$ (jie gaunami χ^2 – minimumo metodu), gauname tokią išraišką:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i(\widehat{\alpha_1}, \widehat{\alpha_2}, \ldots, \widehat{\alpha_s})\right)^2}{np_i\left(\widehat{\alpha_1}, \widehat{\alpha_2}, \ldots, \widehat{\alpha_s}\right)}. \tag{3.5.2}$$

Šio atsitiktinio dydžio skirstinys, kai n neaprėžtai didėja, artėja prie χ^2 skirstinio su $(k - s - 1)$ laisvės laipsnių.

Kokia (3.5.1) ir (3.5.2) formulių fizikinė prasmė? Įvykio A_i *tikrasis* įvykimų dažnis yra ν_i, o *tikėtinasis* A_i įvykimų dažnis yra np_i. Aišku, kad skirtumai $\nu_i - np_i$ (kartu ir skirtumų kvadratų suma) neturėtų būti dideli. Priešingu atveju modelis netinkamas (matyt, p_i neteisingai pasirinktos!).

$E_i = np_i$ – tikėtinas dažnis.

Čia aprašyti rezultatai naudojami statistiniams kriterijams sudaryti. Juos taikysime šio skyriaus uždaviniams spręsti.

Pastaba. Norint, kad dydžio (3.5.1) aproksimavimas χ^2 skirstiniu nebūtų per grubus, kiekvienas iš įvykių A_i turi įvykti kuo daugiau kartų. Praktiškai aproksimavimas yra patenkinamas, jei kiekvienas A_i įvyksta daugiau nei 10 kartų. Plačiau χ^2 kriterijaus praktinio taikymo aspektai aprašyti 5.6 skyrelyje.

5.2. χ^2 suderinamumo kriterijus

55% žmonių reguliariai skaito horoskopus.
60% žmonių mano atrodą vidutiniškai.
20% moterų motinas laiko geriausiomis draugėmis.

χ^2 suderinamumo kriterijus naudojamas hipotezėms apie kintamojo skirstinį (binominį, Puasono, normalųjį ir pan.) populiacijoje tikrinti. χ^2 kriterijus parodo, ar empirinio ir teorinio skirstinių skirtumas yra reikšmingas, t. y. tikrinama, ar turimas empirinis skirstinys suderinamas su teoriniu modeliu.

Bendroji kriterijaus sudarymo schema yra tokia:

1. Jei stebimas diskretusis kintamasis, tai iš pradžių apskaičiuojami imties reikšmių (kategorijų) dažniai.
2. Jei stebimas tolydusis kintamasis, tai reikšmių sritis suskaidoma į nesikertančius intervalus ir apskaičiuojami intervaliniai dažniai.
3. Tarkime, kategorijų (intervaliniai) dažniai yra $O_1, O_2, \ldots, O_k$, čia k – kategorijų (intervalų) skaičius. Naudodamiesi teorinio skirstinio (nurodyto hipotezės H_0 formuluotėje) savybėmis, apskaičiuojame, kuri kintamojo reikšmių dalis turėtų būti priskirta kiekvienai kategorijai (patektų į kiekvieną intervalą), jei hipotezė apie kintamojo skirstinį būtų teisinga, t. y. randame tikėtinus dažnius $E_1, E_2, \ldots, E_k$.
4. Įvertiname tikėtinų ir stebimų dažnių skirtumus. Kuo šie skirtumai didesni, tuo labiau abejotina, kad hipotezė apie skirstinį teisinga. Sprendimo priėmimo taisyklės grindžiamos šiuo tikėtinų ir stebimų dažnių skirtumų didumu.

Diskretusis skirstinys. Išsamiau aprašysime, kaip taikomas suderinamumo kriterijus, kai stebimas diskretusis kintamasis. Tarkime, kad pagal stebimą kintamąjį populiaciją galima suskirstyti į k kategorijų. Populiacijos, priklausančios i-ajai kategorijai, dalį pažymime p_i, $i = 1, \ldots, k$ (ekvivalenti formuluotė: stebimas atsitiktinis dydis, su tikimybe p_i įgyjantis i-ąją reikšmę). Jei hipotezė apie skirstinį H_0 teisinga, tai stebimojo kintamojo skirstinys yra žinomas ir tikimybė priklausyti i-ajai kategorijai yra p_i^0. Todėl, kai H_0 teisinga, iš n stebėjimų imties $E_i = np_i^0$ stebėjimų turėtų priklausyti i-ajai kategorijai. Iš tikrųjų priklauso O_i. Skirtumas $O_i - E_i$ rodo, ar hipotezė H_0 tikėtina.

Diskrečiojo skirstinio atveju skirstinių suderinamumui tikrinti naudojame statistiką

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

Jei hipotezė H_0 yra teisinga, tai kriterijaus statistika aproksimuojama χ^2 skirstiniu su $k-1$ laisvės laipsnių.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Stebimas diskretusis atsitiktinis dydis. Iš n imties reikšmių i-ajai kategorijai priklauso o_i reikšmių ($i = 1, 2, \ldots, k$).

2 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ p_1 = p_1^0, \quad p_2 = p_2^0, \ \dots, \ p_n = p_n^0, \\ H_1\colon\ p_i \neq p_i^0, \quad \text{bent vienam iš } i = 1, \dots, n. \end{cases}$$

3 *Kriterijaus statistika*. Apskaičiuojame

$$\chi^2 = \sum_{i=1}^{n} \frac{(\sigma_i - e_i)^2}{e_i}, \tag{3.5.3}$$

čia $e_i = np_i^0$.

4 *Sprendimo priėmimo taisyklė*. Tarkime, reikšmingumo lygmuo lygus α. Jei $\chi^2 > \chi^2_\alpha(k-1)$, tai hipotezę H_0 atmetame (duomenys prielaidos apie kintamojo skirstinį populiacijoje nepatvirtino). Jei $\chi^2 \leqslant \chi^2_\alpha(k-1)$, tai hipotezės H_0 neatmetame (duomenys prielaidos apie kintamojo skirstinį populiacijoje nepaneigė). Čia $\chi^2_\alpha(k-1)$ yra χ^2 skirstinio su $k-1$ laisvės laipsnių α lygmens kritinė reikšmė.

3.5.1 pavyzdys. Atsitiktinai parinktų 60 žiūrovų atsakė į klausimą, kokia televizija geriausia. Rezultatai pateikti 3.5.4 lentelėje.

3.5.4 lentelė. Geriausias TV kanalas

TV kanalas	1	2	3	4	5	6
Dažnis	5	8	10	12	12	13

Ar remiantis šiais duomenimis galima sakyti, kad visų TV kanalų reitingai yra vienodi?

Sprendimas. Tarkime, atsitiktinis dydis (kintamasis) X yra numeris TV kanalo, kurį renkasi atsitiktinis žiūrovas. Tada teorinis X tikimybių skirstinys aprašomas 3.5.5 lentele. Joje p_i^0 žymi tikimybę, kad žiūrovas

3.5.5 lentelė. Geriausias TV kanalas. Teorinis modelis

X	1	2	3	4	5	6
P	p_1^0	p_2^0	p_3^0	p_4^0	p_5^0	p_6^0

geriausiu pripažins i-ąjį, $i = 1, \ldots, 6$, TV kanalą. Kitais žodžiais tariant, p_i^0 žymi populiacijos dalį, kuri geriausiu laiko i-ąjį kanalą.

Mus domina, ar tikrai i-ąjį kanalą geriausiu laiko p_i^0 populiacijos dalis. Formuluojant tikimybių teorijos terminais, mums reikia patikrinti hipotezę, kad turima duomenų aibė gauta iš populiacijos, kurios skirstinys apibrėžtas 3.5.5 lentele. Stebimieji dažniai o_i (O_i realizacijos) pateikti 3.5.4 lentelėje, tikėtinus dažnius e_i (E_i realizacijas) randame imdami p_i^0 iš 3.5.5 lentelės.

Žiūrovų populiacijoje, kurioje visi kanalai vertinami vienodai, kiekvienam iš kanalų pirmenybę atiduoda 1/6 dalis žiūrovų. Tada $p_i^0 = 1/6$, $i = 1, \ldots, 6$. Kadangi $n = 60$, o visi p_i^0 lygūs, tai ir visi $e_i = np_i^0 = 60 \cdot (1/6) = 10$. Pagal formulę (3.5.3) apskaičiuojame statistikos realizaciją

$$\chi^2 = \frac{(5-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(13-10)^2}{10} = 4{,}6.$$

Kadangi, $\chi^2 = 4{,}6 < 11{,}070 = \chi^2_{0{,}05}(5)$, hipotezės H_0 atmesti negalima.

Išvada. Duomenys neprieštarauja hipotezei, kad kanalų reitingai yra vienodi.

Klausimą apie populiacijos, patenkančios į skirtingas kategorijas, dalis galima suformuluoti įvairiai. Pavyzdžiui, hipotezę, kad, esant tam tikram radioaktyvumo lygiui, mutavusių pupų bus dukart daugiau nei nemutavusių, galima suformuluoti ir taip: a) mutavusių ir nemutavusių pupų santykis yra 2 : 1; b) 2/3 visų pupų mutuos, o 1/3 nemutuos. Taigi šiuo atveju $p_1^0 = 2/3$, $p_2^0 = 1/3$. Analogiškai teiginį, kad itin muzikalių, vidutiniškai muzikalių ir visai nemuzikalių vaikų santykis tam tikroje populiacijoje yra 1 : 10 : 4, performuluojame taip: itin muzikalūs vaikai sudaro 1/15 populiacijos dalį (p_1^0), muzikalūs vaikai sudaro 10/15 populiacijos (p_2^0), o nemuzikalūs sudaro 4/15 populiacijos (p_3^0).

χ^2 suderinamumo kriterijaus taikymas naudojantis SPSS paketu. Problemai spręsti galima taikyti SPSS paketą. Tuomet sprendimą priimti patogiausia naudojantis p-reikšme. Tarkime, kad ji lygi p. Tegul pasirinktas reikšmingumo lygmuo yra α. Tuomet, jeigu $p < \alpha$, hipotezė H_0 atmetama (duomenys prielaidos apie kintamojo skirstinį populiacijoje nepatvirtino). Jei $p \geqslant \alpha$, tai hipotezės H_0 neatmetame (duomenys patvirtino prielaidą apie kintamojo skirstinį populiacijoje).

SPSS rezultatai 3.5.1 pavyzdžio atveju pateikti 3.5.1 paveiksle. Pirmojoje lentelėje nurodytos stebimosios reikšmės o_i, tikėtinosios reikšmės e_i ir jų skirtumai $(o_i - e_i)$. Antrojoje lentelėje pateikiama statistikos realizacija *Chi-Square* = $(\chi^2) = 4{,}600$, laisvės laipsnių skaičius $df = (k-1) = 5$ ir p-reikšmė 0,467. Kadangi $0{,}467 \geqslant 0{,}05$, tai H_0 atmesti nėra pagrindo. Beje, SPSS pakete galima įvesti spėjamą tarpusavio santykį, o ne tik pačias tikimybes, t. y. santykį 2 : 1, nebūtinai 2/3 ir 1/3.

Normalusis skirstinys. Prielaida, kad stebimas kintamasis turi normalųjį skirstinį, remiamasi sprendžiant daugelį hipotezių tikrinimo uždavinių. χ^2 suderinamumo kriterijus yra įrankis šiai prielaidai tikrinti.

TV

	Observed N	Expected N	Residual
1	5	10,0	-5,0
2	8	10,0	-2,0
3	10	10,0	,0
4	12	10,0	2,0
5	12	10,0	2,0
6	13	10,0	3,0
Total	60		

TEST STATISTICS

	TV
Chi-Square[a]	4,600
df	5
Asymp. Sig.	,467

3.5.1 pav. SPSS rezultatai geriausiam TV kanalui nustatyti

Formaliai statistinę hipotezę apie stebimojo kintamojo normališkumą užrašome taip: H_0: $X \sim N(\mu, \sigma^2)$; H_1: $X \not\sim N(\mu, \sigma^2)$.

Stebimas kintamasis X yra tolydus, todėl, norėdami taikyti χ^2 kriterijų, turime jį sudiskretinti. Duomenų aibė pateikiama grupuotų dažnių lentele, kurioje yra k grupavimo intervalų (prisiminkime aprašomosios statistikos metodus!). Stebint X, įvyksta vienas iš k nesutaikomų įvykių A_i, $i = 1, \ldots, k$. Čia $A_i = \{X$ reikšmė patenka į i-ąjį intervalą$\}$. Apskaičiavę patekimo į i-ąjį intervalą tikimybę, gauname priklausymo i-ajai kategorijai tikimybę $p_i = P(A_i)$, kartu randame ir tikėtiną dažnį np_i.

Turime antrąjį matematinį modelį, aprašytą 5.1 skyrelyje, t. y. stebėjimo rezultatas – vienas iš k nesutaikomų įvykių A_i, o įvykių tikimybės $p_i = P(A_i)$ priklauso nuo dviejų nežinomų parametrų μ ir σ (juk tikrinama normališkumo prielaida, o normaliojo skirstinio tikimybės priklauso nuo μ ir σ!). Hipotezei tikrinti naudojama statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i))^2}{E_i} = \sum_{i=1}^{k} \frac{(O_i - np_i(\overline{X}, S))^2}{p_i(\overline{X}, S)}.$$

Čia O_i – stebimi intervaliniai dažniai, E_i – tikėtini intervaliniai dažniai, $\overline{X}$, S – normaliojo skirstinio vidurkio ir standartinio nuokrypio įverčiai[1], $p_i(\overline{X}, S)$ – normaliojo skirstinio su parametrais $\overline{X}$ ir S i-ojo intervalo tikimybės.

Jeigu duomenys gauti stebint normalųjį atsitiktinį dydį, tai statistika χ^2 turi χ^2 skirstinį su $k - 3$ laisvės laipsnių.

Nagrinėjamojo uždavinio sprendimo etapai yra tokie:

1 *Duomenys.* Stebimas tolydusis kintamasis. Galimų reikšmių aibė suskirstyta į nesikertančius intervalus $(-\infty; c_1)$; $[c_1, c_2), \ldots, [c_{k-1}, +\infty)$; i-ajam intervalui priklauso o_i stebėjimų.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ X \sim \mathcal{N}(\mu, \sigma^2), \\ H_1\colon\ X \not\sim \mathcal{N}(\mu, \sigma^2). \end{cases}$$

3 *Kriterijaus statistika.* Apskaičiuojame

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \qquad (3.5.4)$$

čia

$$e_i = n\left(\Phi\left(\frac{c_i - \overline{x}}{s}\right) - \Phi\left(\frac{c_{i-1} - \overline{x}}{s}\right)\right).$$

4 *Sprendimo priėmimo taisyklė.* Tarkime, reikšmingumo lygmuo yra α. Jei $\chi^2 > \chi^2_\alpha(k-3)$, hipotezę H_0 atmetame (duomenys neleidžia teigti, kad stebime normalųjį atsitiktinį dydį). Jei $\chi^2 \leqslant \chi^2_\alpha(k-3)$, hipotezės H_0 neatmetame (duomenys leidžia teigti, kad stebime normalųjį atsitiktinį dydį). Čia $\chi^2_\alpha(k-3)$ yra χ^2 skirstinio su $k-3$ laisvės laipsnių α lygmens kritinė reikšmė.

---

[1] Jeigu grupavimo intervalai parinkti iš anksto, tai μ ir σ^2 įverčiai yra $\overline{X}$ ir S^2. Priešingu atveju reikia naudoti įverčius, gautus χ^2 minimumo metodu, arba taikyti modifikuotą χ^2 kriterijų ([6]).

3.5.2 pavyzdys. Atsitiktinai parinktų 710 gamyklos darbininkų dalyvavo eksperimente, kurio tikslas – nustatyti, per kiek laiko pagaminama tam tikra detalė. Laiko (minutėmis), per kurį parinkti darbininkai pagamino detalę, grupuotų dažnių skirstinys pateikiamas 3.5.6 lentelėje. Ar galima tvirtinti, kad laikas, sugaištamas tam tikrai detalei pagaminti, turi normalųjį skirstinį?

3.5.6 lentelė. Laikas, reikalingas detalei pagaminti

Laiko intervalas	Dažnis	Laiko intervalas	Dažnis
$(c_0, c_1) = (0; 9{,}99)$	39	$[c_6, c_7) = [60; 69{,}99)$	83
$[c_1, c_2) = [10; 19{,}99)$	53	$[c_7, c_8) = [70; 79{,}99)$	73
$[c_2, c_3) = [20; 29{,}99)$	64	$[c_8, c_9) = [80; 89{,}99)$	62
$[c_3, c_4) = [30; 39{,}99)$	75	$[c_9, c_{10}) = [90; 99{,}99)$	75
$[c_4, c_5) = [40; 49{,}99)$	85	$[c_{10}, c_{11}) = [100; 109{,}99)$	75
$[c_5, c_6) = [50; 59{,}99)$	92	Iš viso	710

Sprendimas. Tarkime, kintamasis X yra laikas, sugaištamas detalei pagaminti. Iš pradžių apskaičiuojame parametrų μ ir σ įverčius pagal grupuotų duomenų charakteristikų skaičiavimo (1.3) ir (1.8) formules. Randame $\bar{x} = 54{,}71$; $s = 27{,}61$. Norint rasti konkrečią kriterijaus statistikos reikšmę, pirmiausia reikia apskaičiuoti *tikėtinus* dažnius. Kad būtų apimtos visos įmanomos situacijos, formaliai pirmąjį intervalą užrašome $(-\infty; c_1) = (-\infty; 9{,}99)$, o paskutinį – $[c_{10}; +\infty) = [100; +\infty)$, t. y. sąlygą keičiame taip: pirmasis intervalas – iki 10 minučių; paskutinis – ne mažiau kaip 100 minučių. Pasinaudoję normaliojo skirstinio savybėmis, galime užrašyti:

$$e_i = n \cdot P(c_{i-1} < X < c_i) = np_i(\bar{x}, s) = n\left(\Phi\left(\frac{c_i - \bar{x}}{s}\right) - \Phi\left(\frac{c_{i-1} - \bar{x}}{s}\right)\right)$$

$$= 710\left(\Phi\left(\frac{c_i - 54{,}71}{27{,}61}\right) - \Phi\left(\frac{c_{i-1} - 54{,}71}{27{,}61}\right)\right), \quad i = 1, \ldots, 11.$$

Čia $\Phi(x)$ – standartinio normaliojo atsitiktinio dydžio pasiskirstymo funkcija. Pavyzdžiui,

$$e_2 = 710\left(\Phi\left(\frac{19{,}9 - 54{,}71}{27{,}61}\right) - \Phi\left(\frac{10 - 54{,}71}{27{,}61}\right)\right) = 36{,}92.$$

Pastebėsime, kad

$$e_1 = n\left(\Phi\left(\frac{c_1 - \bar{x}}{s}\right) - \Phi\left(-\infty\right)\right) = n\Phi\left(\frac{c_1 - \bar{x}}{s}\right) = 710\Phi\left(\frac{9{,}99 - 54{,}71}{27{,}61}\right) = 37{,}63.$$

3.5.7 lentelė. Hipotezės apie normalųjį skirstinį tikrinimas

Intervalas	o_i	e_i	$(o_i - e_i)^2/e_i$
(0; 0,99)	39	37,63	0,049
[10; 19,9)	53	36,92	7,00
[20; 29,9)	64	58,93	0,436
[30; 39,9)	75	77,39	0,074
[40; 49,9)	85	96,56	1,384
[50; 59,9)	92	101,53	6,894
[60; 69,9)	83	93,72	1,226
[70; 79,9)	73	72,42	0,004
[80; 89,9)	62	62,48	0,004
[90; 99,9)	53	36,21	7,785
[100; 110)	36	36,21	0,001
Σ	710		$\chi^2 = 24{,}77$

$$e_{11} = n\left(\Phi\left(+\infty\right) - \Phi\left(\frac{c_{10} - \overline{x}}{s}\right)\right) = 710\left(1 - \Phi\left(\frac{100 - 54,71}{27,61}\right)\right) = 36,21.$$

Statistikos dėmenų skaičiavimo etapai pateikiami 3.5.7 lentelėje. Apskaičiuojame $\chi^2 = 24,77$. Kadangi $\chi^2 = 24,77 > 15,507 = \chi^2_{0,05}(8)$, hipotezę reikia atmesti.

Išvada. Negalime teigti, kad laikas, sugaištamas detalei pagaminti, turi normalųjį skirstinį.

5.3. Požymių nepriklausomumo tikrinimas

Dažnai reikia išsiaiškinti, ar stebimi kintamieji yra nepriklausomi, ar priklausomi. Pavyzdžiui, reikia nustatyti:

ar yra priklausomybė tarp sergamumo širdies ligomis ir rūkymo;

ar nusikalstamumo lygis priklauso nuo bedarbystės lygio;

ar užsienio politikos kursas priklauso nuo to, kokia partija yra valdžioje;

ar perkamo automobilio spalva priklauso nuo perkančiojo lyties.

Jei kintamieji yra intervaliniai, tai žinome, kad tiesinės priklausomybės matas yra Pirsono koreliacijos koeficientas. Tikrindami hipotezę apie Pirsono koreliacijos koeficiento lygybę nuliui, atsakome į klausimą apie tiesinę kintamųjų priklausomybę. Tačiau ką daryti, jei kintamieji yra kokybiniai? Tuo atveju galima taikyti χ^2 kriterijų. Beje, kategoriniai kintamieji dažnai vadinami *požymiais*. Iš čia šio skyrelio pavadinimas.

Tarkime, turime porinių stebėjimų imtį $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, gautą stebint kategorinių kintamųjų porą (X, Y). Kaip patikrinti hipotezę, kad kintamieji X ir Y yra nepriklausomi? Duomenų aibę užrašykime 3.5.8 porine dažnių lentele. Čia $n_{i.} = \sum_{j=1}^{c} o_{ij}$ yra imties narių, kurių požymio X reikšmė yra x_i, skaičius; $n_{.j} = \sum_{i=1}^{r} o_{ij}$ – imties narių, kurių požymio Y reikšmė yra y_j, skaičius; $n = \sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij}$ – imties didumas.

3.5.8 lentelė. Porinė dažnių lentelė

	y_1	y_2	$\ldots$	y_c	$\sum$
x_1	o_{11}	o_{12}	$\ldots$	o_{1c}	$n_{1.}$
x_2	o_{21}	o_{22}	$\ldots$	o_{2c}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_r	o_{r1}	o_{r2}	$\ldots$	o_{rc}	$n_{r.}$
$\sum$	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.c}$	n

Suformuluokime tikimybinį uždavinio modelį.

Tarkime, $p_{ij} = P(X = x_i, Y = y_j)$ – populiacijos dalis, kuriai matuojamų požymių pora (X, Y) įgyja reikšmę (x_i, y_j); $p_i = P(X = x_i)$ – populiacijos dalis, kuriai požymis X įgyja reikšmę x_i; $q_j = P(Y = y_i)$ – populiacijos dalis, kuriai požymis Y įgyja reikšmę y_j. Aišku, kad

$$\sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1.$$

Iš tikimybių teorijos žinoma, kad kintamieji yra nepriklausomi, jei $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ su visais $i = 1, \ldots, r$, $j = 1, \ldots, c$.

Taigi mums reikia patikrinti statistinę hipotezę

$$\begin{cases} H_0\colon\ p_{ij} = p_i q_j, & \text{su visais } i = 1, \ldots, r,\ j = 1, \ldots, c; \\ H_1\colon\ p_{ij} \neq p_i q_j, & \text{bent vienai porai } (i, j). \end{cases} \tag{3.5.5}$$

Jei hipotezė teisinga, tai p_{ij} yra nežinomų parametrų p_i, q_j, $i = 1, \ldots, r$, $j = 1, \ldots, c$, funkcijos. Prisiminkime, kaip teoriškai grindžiamas χ^2 kriterijus. Nagrinėjamu atveju tinka antroji iš 5.1 skyrelyje aprašytų situacijų. Iš tiesų, stebint kintamųjų porą (X, Y), gali įvykti $r{\cdot}c$ nesutaikomų įvykių ($r{\cdot}c$ yra (X, Y) skirtingų realizacijų skaičius). Taigi atsižvelgiant į tai, kad kiekvieną įvykį apibūdina du indeksai i ir j, kriterijaus statistiką galima užrašyti taip:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n\widehat{p_{ij}})^2}{\widehat{p_{ij}}}.$$

Nežinomų parametrų įverčiai, gauti χ^2 minimumo metodu, yra:

$$\widehat{p_i} = \frac{n_{i\cdot}}{n}, \quad i = 1, \ldots, r; \qquad \widehat{q_j} = \frac{n_{\cdot j}}{n}, \quad j = 1, \ldots, c.$$

Tikėtinieji dažniai E_{ij} apskaičiuojami pagal formulę

$$E_{ij} = n\widehat{p_{ij}} = \widehat{p_i} \cdot \widehat{q_j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}.$$

Iš čia išplaukia, kad esant teisingai hipotezei H_0 statistika

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{O_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right) \tag{3.5.6}$$

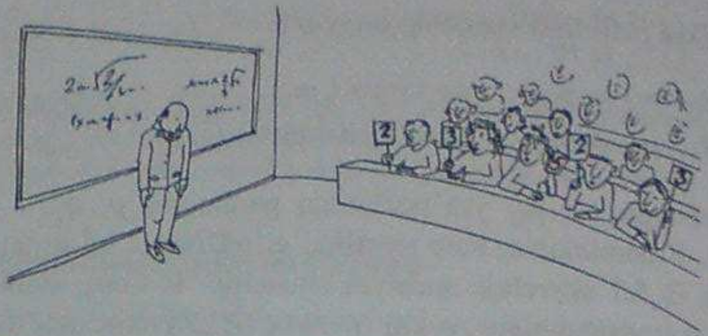
yra asimptotiškai pasiskirsčiusi pagal χ^2 dėsnį su $rc - [(r-1) + (c-1)] - 1 = rc - r - c + 1 = (r-1)(c-1)$ laisvės laipsnių. Hipotezė H_0 apie kintamųjų nepriklausomumą yra atmetama, kai apskaičiuotos statistikos χ^2 reikšmė yra didesnė už χ^2 skirstinio su $(r-1)(c-1)$ laisvės laipsnių α lygmens kritinę reikšmę; čia α – pasirinktas reikšmingumo lygmuo.

Panagrinėkime, kaip tai daroma praktiškai.

Dėstytojai teigia, kad studentai juos vertina pagal gautus pažymius. Norint šį teiginį patikrinti, buvo atsitiktinai parinkta 400 studentų. Kiekvienas studentas vertino tik po vieną dėstytoją. Gautų duomenų aibė pateikiama 3.5.9 lentelėje.

3.5.9 lentelė. Studentų vertinimai

Vertinimas	Studento gautas balas				
	0–4	5–6	7–8	9	10
Puikus	10	15	30	25	20
Kompetentingas	30	40	60	25	25
Netikęs	40	45	10	10	15



Ar remiantis šiais duomenimis galima sakyti, kad dėstytojų teiginys yra teisingas?

1 *Duomenys*. Tarkime, kintamasis X yra „vertinimas", kintamasis Y yra „pažymys". Kintamasis X turi 3 kategorijas. Pažymėkime jas taip: 1 – *puikus*, 2 – *kompetentingas*, 3 – *netikęs*. Kintamojo Y penkios kategorijos yra tokios: 1 atitinka 0–4, 2 – 5–6, 3 – 7–8, 4 – 9, 5 – 10. Kintamųjų X ir Y porinių stebėjimų duomenų aibė pateikta 3.5.9 lentele.

2 *Statistinė hipotezė*:

$$\begin{cases} H_0\colon\ p_{ij} = p_i q_j,\ i = 1, \ldots, 3,\ j = 1, \ldots, 5; \\ H_1\colon\ p_{ij} \neq p_i q_j,\ \text{bent vienai porai } (i, j). \end{cases}$$

Hipotezė H_0 teigia, kad vertinimas ir gautas pažymys nesusiję, H_1 – vertinimas ir gautas pažymys susiję. Čia p_{ij} yra ta studentų populiacijos dalis, kurios dėstytojo vertinimas atitinka i-ąją kategoriją, o atitinkamos disciplinos pažymys – j-ąją kategoriją; p_i yra ta studentų populiacijos dalis, kurios vertinimai atitinka i-ąją kategoriją, q_j – kurios gauti pažymiai atitinka j-ąją kategoriją.

3 *Kriterijaus statistika*. Apskaičiuojame

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{\left(o_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = 60{,}741.$$

Čia o_{ij} – stebimi dažniai, kurie pateikti 3.5.9 lentelėje, e_{ij} – tikėtini dažniai. Jie apskaičiuojami taip:

$$e_{11} = \frac{n_{1.}n_{.1}}{n} = \frac{80 \cdot 100}{400} = 20.$$

Analogiškai skaičiuojami ir kiti tikėtini dažniai. Stebimi ir tikėtini (skliausteliuose) dažniai pateikiami 3.5.10 lentelėje.

4 *Sprendimo priėmimo taisyklė*. Tarkime, reikšmingumo lygmuo yra $\alpha = 0{,}05$. Jei statistikos reikšmė didesnė už χ^2 skirstinio su $(3-1)(5-1) = 8$ laisvės laipsniais 0,05 lygmens kritinę reikšmę ($\chi^2_{0.05}(8) = 15{,}507$), hipotezę reiktų atmesti. Kadangi $60{,}741 > 15{,}507 = \chi^2_{0.05}(8)$, hipotezę (apie požymių nepriklausomumą) atmetame. Taigi vertinimai ir pažymiai susiję. Dėstytojų vertinimui turi įtakos gaunami pažymiai.

3.5.10 lentelė. Studentų vertinimai. Duomenys ir tikėtini dažniai

Vertinimas	Studento gautas balas				
	0–4	5–6	7–8	9	10
Puikus	10 (20)	15 (25)	30 (25)	25 (15)	20 (15)
Kompetentingas	30 (36)	40 (45)	60 (45)	25 (27)	25 (27)
Netikęs	40 (24)	45 (30)	10 (18)	10 (30)	15 (18)

χ^2 kriterijus atsako į klausimą, *ar požymiai priklausomi*. Pavyzdžio vertinimų ir pažymių priklausomybę nustatome analizuodami 3.5.10 lentelę. Požymių priklausomybę nagrinėti dar patogiau, kai duomenys išreikšti procentais.

SPSS paketu gautas pavyzdžio sprendimas pateikiamas 3.5.2 paveiksle. Viena lentelė skirta atsakyti į klausimą, ar pažymiai ir vertinimai statistiškai reikšmingai susiję. Grafoje 'Pearson Chi-Square' randame $\chi^2 = 60{,}741$ reikšmę, laisvės laipsnių skaičių $df = 8$ ir p-reikšmę 0,000. Kadangi p-reikšmė mažesnė už 0,05, hipotezę apie požymių nepriklausomumą atmetame – požymiai statistiškai priklausomi. Taigi gauname tą pačią išvadą kaip ir anksčiau. Beje, turint daug gardelių, kuriose mažai stebėjimų, χ^2 kriterijaus atsakymas būtų statistiškai nepatikimas (žr. 5.6).

Kitoje lentelėje pateikti ne tik duomenys apie stebimus tikruosius dažnius, tikėtinus dažnius, bet ir apie jų procentinę sudėtį. Taigi įsitikiname, kad 50% studentų, gavusių nuo 0 iki 4 balų, dėstytoją vertina kaip netikusį. Iš 7–8 balus gavusių studentų neigiamai dėstytoją vertinančių yra tik 10%. Analogiškai paanalizavę visą lentelę, matome, kad gavusieji blogesnius pažymius dėstytoją linkę vertinti blogiau nei gavusieji 7 ir daugiau balų.

5.4. Homogeniškumo tikrinimas

Norime išsiaiškinti:

ar rūkančių vyrų ir moterų procentas yra tas pats;

ar Lietuvoje gyvenančių įvairių tautybių žmonių išsilavinimas yra vienodas.

Statistikos terminais antrąjį uždavinį galima suformuluoti taip: turint keliolika populiacijų (lietuvių, rusų, lenkų ir pan.) reikia nustatyti, ar kintamojo „išsilavinimas" skirstinys šiose populiacijose yra vienodas, kitaip sakant, ar įvairių tautybių populiacijos nagrinėjamo kintamojo (požymio) „išsilavinimas" atžvilgiu yra homogeniškos.

Šiame skyrelyje nagrinėsime homogeniškumo tikrinimo uždavinius. Homogeniškumo hipotezėms tikrinti taip pat naudojamas χ^2 kriterijus. Kuo požymių nepriklausomumo kriterijus skiriasi nuo homogeniškumo kriterijaus? Visų pirma, skiriasi imties procedūra. Požymių nepriklausomumo kriterijus taikomas, kai vienoje populiacijoje stebima kintamųjų pora. Homogeniškumo kriterijus taikomas, kai keliose populiacijose stebimas vienas (ir tas pats) kintamasis. Antra, skirtinga tikėtinų dažnių fizikinė prasmė. Trečia, skiriasi rezultatų interpretavimas.

Teorinis uždavinio modelis atrodo taip.

Turime r nepriklausomų populiacijų. Jose yra stebimas kategorinis kintamasis X, kurio galimų reikšmių aibę sudaro c kategorijos. Tada duomenų aibė gali būti užrašoma 3.5.11 dažnių lentele.

VERTINIMAS * STUDENTŲ BALAS CROSSABULATION

			Studentų balas					
			0-4	5-6	7-8	9	10	Total
Vertinimas	Puikus	Count	10	15	30	25	20	100
		Expeced Count	20,0	25,0	25,0	15,0	15,0	100,0
		%within Vertinimas	10,0%	15,0%	30,0%	25,0%	20,0%	100,0%
		%within Studentų balas	12,5%	15,0%	30,0%	41,7%	33,3%	25,0%
	Kompetentingas	Count	30	40	60	25	25	180
		Expeced Count	36,0	45,0	45,0	27,0	27,0	180,0
		%within Vertinimas	16,7%	22,2%	33,3%	13,9%	13,9%	100,0%
		%within Studentų balas	37,5%	40,0%	60,0%	41,7%	41,7%	45,0%
	Netikęs	Count	40	45	10	10	15	120
		Expeced Count	24,0	30,0	30,0	18,0	18,0	120,0
		%within Vertinimas	33,3%	37,5%	8,3%	8,3%	12,5%	100,0%
		%within Studentų balas	50,0%	45,0%	10,0%	16,7%	25,0%	30,0%
Total		Count	80	100	100	60	60	400
		Expeced Count	80,0	100,0	100,0	60,0	60,0	400,0
		%within Vertinimas	20,0%	25,0%	25,0%	15,0%	15,0%	100,0%
		%within Studentų balas	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

CHI-SQUARE TESTS

	Value	df	Asymp. Sig. (2-tiled)
Pearson Chi-Square	60,741[a]	8	,000
likelihood Ratio	63,424	8	,000
Linear-by-Linear Association	31,886	1	,000
N of Valid Cases	400		

3.5.2 pav. χ^2 kriterijus pažymių ir vertinimų priklausomybei nustatyti

3.5.11 lentelė. Populiacijų dažnių skirstiniai

Populiacija	Kategorija				
	1	2	...	c	$\sum$
1 populiacija	o_{11}	o_{12}	...	o_{1c}	$n_{1\cdot}$
2 populiacija	o_{21}	o_{22}	...	o_{2c}	$n_{2\cdot}$
...	...	...	...	...	...
r populiacija	o_{r1}	o_{r2}	...	o_{rc}	n_{r}
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot c}$	n

Tarkime, p_{ij} yra i-osios populiacijos dalis, kuriai kintamojo X reikšmė patenka į j-ąją kategoriją. Tuomet formuluojama statistinė hipotezė:

$$\begin{cases} H_0\colon\ p_{11} = p_{21} = \cdots = p_{c1}, \\ \qquad p_{12} = p_{22} = \cdots = p_{c2}, \\ \qquad \cdots\cdots\cdots\cdots\cdots\cdots \\ \qquad p_{1r} = p_{2r} = \cdots = p_{cr}, \\ \\ H_1\colon\ H_0 \text{ neteisinga (nėra bent vienos lygybės).} \end{cases} \tag{3.5.7}$$

Nulinė hipotezė H_0 teigia, kad visos populiacijos tiriamo požymio atžvilgiu homogeninės. Kai populiacijų skirstinių skirtumai statistiškai reikšmingi, hipotezę H_0 atmetame.

Ir kriterijaus statistika χ^2 (apibrėžta (3.5.6) formule), ir sprendimo priėmimo taisyklės yra tokios pat, kaip ir požymių nepriklausomumo uždaviniuose, tačiau dar kartą norime pabrėžti, kad naudojamas kitoks statistinis modelis ir kitaip interpretuojami rezultatai. Panagrinėkime tokį pavyzdį:

Tiriama, ar verslininkai ir verslininkės ūkio perspektyvas ateinančiais metais vertina vienodai. Apklausus 200 atsitiktinai parinktų verslininkų ir 100 atsitiktinai parinktų verslininkių, gauti rezultatai, kurie pateikti 3.5.12 lentelėje.

3.5.12 lentelė. Verslininkų nuomonė apie ūkio perspektyvas

	Pagerės	Nepakis	Pablogės
Verslininkai	52	60	70
Verslininkės	47	58	70

Ar galima tvirtinti, kad verslininkių ir verslininkų ūkio perspektyvų vertinimai skiriasi?

1 *Duomenys*. Turime dvi populiacijas – verslininkų ir verslininkių. Kintamasis X, kurio skirstinius populiacijose norime palyginti, yra „ūkio perspektyvos“. Duomenų aibė – kintamojo X stebėjimai verslininkų ir verslininkių populiacijose. Duomenys pateikti 3.5.12 lentelėje.

2 *Statistinė hipotezė:*

$$\begin{cases} H_0\colon\ p_{11} = p_{21},\ p_{12} = p_{22},\ p_{13} = p_{23}; \\ H_1\colon\ H_0 \text{ neteisinga.} \end{cases} \tag{3.5.8}$$

Hipotezė H_0 teigia, kad verslininkai ir verslininkės ūkio perspektyvas vertina vienodai, H_1 – verslininkai ir verslininkės ūkio perspektyvas vertina nevienodai. Čia p_{11} yra dalis verslininkų, manančių, kad ūkio situacija kitais metais pagerės; p_{21} – dalis verslininkių, manančių, kad ūkio situacija kitais metais pablogės, ir pan.

3 *Kriterijaus statistika.* Apskaičiuojame

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{\left(o_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}} = 0{,}149. \tag{3.5.9}$$

4 *Sprendimo priėmimo taisyklė.* Tarkime, reikšmingumo lygmuo yra $\alpha = 0{,}05$. Jei statistikos reikšmė būtų didesnė už χ^2 skirstinio su $(2-1)(3-1) = 1 \cdot 2 = 2$ laisvės laipsnių 0,05 lygmens kritinę reikšmę ($\chi^2_{0,05}(2) = 5{,}99$), hipotezę reiktų atmesti. Kadangi $\chi^2 = 0{,}149 < 5{,}99 = \chi^2_{0,05}(2)$, hipotezės atmesti negalima. Duomenys suformuluotai hipotezei neprieštarauja. Verslininkių ir verslininkų požiūriai nesiskiria.

Kaip ir nepriklausomumo tikrinimo atveju, χ^2 homogeniškumo kriterijus tik atsako į klausimą, *ar* požymio skirtingose populiacijose skirstiniai skiriasi. Pačius skirtumus nustatome nagrinėdami porinę dažnių lentelę.

5.5. Dvireikšmių požymių dažnių lentelės

Tiek požymių nepriklausomumo, tiek homogeniškumo uždaviniuose duomenų aibė yra porinė dažnių lentelė, arba kitaip sakant, duomenų $r \times c$ eilės matrica. Praktiškai labai dažnai pasitaiko situacijų, kai duomenų matrica yra 2×2 eilės, t. y. stebima *dvireikšmių* kintamųjų pora arba tikrinamas dvireikšmio kintamojo skirstinio dviejose populiacijose homogeniškumas. Pavyzdžiui, norima žinoti, ar vyrai labiau remia mirties bausmės panaikinimą nei moterys; ar odos vėžiu dažniau serga vyresni nei 40 metų žmonės; ar žiūrinčių smurtines TV laidas ir agresyviai besielgiančių vaikų procentas didesnis nei nežiūrinčiųjų ir pan. Tokius uždavinius verta aptarti atskirai dėl dviejų priežasčių. Pirma, šiuo atveju χ^2 kriterijus bei statistikos išraiška įgyja daug paprastesnį pavidalą; kartu galima suprantamiau paaiškinti fizikinę taikomų formulių prasmę. Antra, vietoje χ^2 kriterijaus dažnai naudojamas Fišerio kriterijus (tikslus kriterijus!).

Fišerio kriterijus vadinamas tiksliuoju todėl, kad jį naudojant p-reikšmės apskaičiuojamos tiksliai, tuo tarpu χ^2 kriterijaus p-reikšmės yra apytikslės. Fišerio kriterijaus dažniai pateikti 3.5.13 lentelėje. Kadangi šiuo atveju yra tik keturios gardelės, tradiciškai jų reikšmės žymimos raidėmis ($o_{11} = a$, $o_{12} = b$, $o_{21} = c$, $o_{22} = d$).

3.5.13 lentelė

	x_1	x_2	Σ
y_1	a	b	$a+b$
y_2	c	d	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d$

Tarkime, kad 3.5.13 lentelės eilučių sumos $a+b$, $c+d$ ir stulpelių sumos $a+c$, $b+d$ yra fiksuoti dydžiai. Tada, žinant vienos iš gardelių dažnį (pavyzdžiui, a), žinomi ir visi likusieji. Tuomet tikimybė, kad tarp $n = a+b+c+d$ dvireikšmių kintamųjų poros (X, Y) stebėjimų reikšmė (x_1, y_1) pasirodys a kartų (kartu (x_1, y_2) – c kartų, (x_2, y_1) – b kartų, (x_2, y_2) – d kartų), užrašoma kaip hipergeometrinio skirstinio tikimybė:

$$P(a, b, c, d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}. \tag{3.5.10}$$

Kaip ir anksčiau, nulinė hipotezė H_0 teigia, kad požymiai nepriklausomi, alternatyva H_1 – kad priklausomi:

$$\begin{cases} H_0\colon\ p_{ij} = p_i q_j, & \text{su visais } i = 1, 2,\ j = 1, 2; \\ H_1\colon\ p_{ij} \neq p_i q_j, & \text{bent vienai porai } (i, j). \end{cases} \tag{3.5.11}$$

Hipotezei H_0 patikrinti reikia pagal (3.5.10) formulę apskaičiuoti tikimybes 3.5.13 dažnių lentelei ir visoms galimoms lentelėms, kurių pirmosios gardelės dažnis mažesnis už a. Jeigu sudėjus gautas tikimybes suma neviršija pusės pasirinkto reikšmingumo lygmens, t.y. $\alpha/2$, nulinę hipotezę H_0 reikia atmesti (požymiai statistiškai reikšmingai priklausomi). Jeigu suma viršija $\alpha/2$, tai reikia skaičiuoti (3.5.10) tikimybes 3.5.13 dažnių lentelei ir visoms galimoms, kurių pirmosios gardelės dažnis didesnis už a. Jeigu sudėjus gautas tikimybes suma neviršija $\alpha/2$, nulinę hipotezę H_0 reikia atmesti. Priešingu atveju hipotezės H_0 atmesti nėra pagrindo.

3.5.3 pavyzdys. Tarkime, atliktas tyrimas, kurio metu buvo apklausiami atsitiktinai parinkti 42 mokiniai, ar jiems patinka kūno kultūros pamokos. Be to, mokiniai buvo suskirstyti į turinčius antsvorio ir jo neturinčius. Gauti rezultatai pateikiami 3.5.14 lentelėje. Ar antsvorio neturintys mokiniai palankiau žiūri į kūno kultūros pamokas?

Fiksuojame suminius dažnius (19, 23, 5, 37). Galime sudaryti dvi lenteles, kuriose $a > 3$, t.y. 3.5.15 ($a = 4$) ir 3.5.16 ($a = 5$).

3.5.14 lentelė. Antsvoris ir požiūris į kūno kultūrą

	Neturi antsvorio	Turi antsvorio	Iš viso
Mėgsta	3	16	19
Nemėgsta	2	21	23
Iš viso	5	37	42

3.5.15 lentelė. Antsvoris ir požiūris į kūno kultūrą, $a = 4$

	Neturi antsvorio	Turi antsvorio	Iš viso
Mėgsta	4	15	19
Nemėgsta	1	22	23
Iš viso	5	37	42

3.5.16 lentelė. Antsvoris ir požiūris į kūno kultūrą, $a = 5$

	Neturi antsvorio	Turi antsvorio	Iš viso
Mėgsta	5	14	19
Nemėgsta	0	23	23
Iš viso	5	37	42

Šioms trims lentelėms pagal (3.5.10) formulę apskaičiuojame tikimybes:

$$P(3,16,2,21) = \frac{19!23!5!37!}{3!16!2!21!} = 0{,}288,$$

$$P(4,15,1,22) = 0{,}105, \qquad P(5,14,0,23) = 0{,}014.$$

Šių tikimybių suma $P(3,16,2,21) + P(4,15,1,22) + P(5,14,0,23) = 0{,}407$. Matome, kad $0{,}407 > 0{,}025$. Kadangi pasirinktos lentelės tikimybė $P(3,16,2,21) = 0{,}288 > 0{,}025$, tai aišku, kad ir lentelių su ne mažesniais už 3 pirmosios gardelės dažniais tikimybių suma viršys 0,025. Taigi atmesti hipotezę H_0 nėra pagrindo, t. y. negalime kalbėti apie statistiškai reikšmingą antsvorio ir kūno kultūros pamokų mėgimo priklausomybę.

Tikslusis Fišerio kriterijus reikalauja labai daug skaičiavimų ir yra konservatyvus – atmeta nulinę hipotezę tik esant dideliems skirtumams. Jeigu nesinaudojama kompiuteriais, tai Fišerio kriterijus patogus tik mažai stebėjimų turinčioms lentelėms. Jeigu stebėjimų daugiau, taikomas χ^2 kriterijus. Tada χ^2 statistiką su Jeitso[1] tolydumo pataisa galima užrašyti taip:

$$\chi^2 = \frac{(|ad - bc| - 0{,}5(a+b+c+d))^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}. \tag{3.5.12}$$

Pasirinkus reikšmingumo lygmenį α, gautąją statistikos χ^2 reikšmę reikia lyginti su $\chi^2_\alpha(1)$. Požymiai (X, Y) statistiškai priklausomi (populiacijos nehomogeniškos), jei $\chi^2 > \chi^2_\alpha(1)$.

3.5.4 pavyzdys. Tarkime, kad apklausti atsitiktinai parinkti 105 mokiniai, ar jiems patinka kūno kultūros pamokos. Kaip ir ankstesniajame pavyzdyje, mokiniai skirstomi į turinčius antsvorio ir jo neturinčius. Gauti rezultatai pateikiami 3.5.17 lentelėje. Ar antsvorio neturintys mokiniai palankiau žiūri į kūno kultūros pamokas?

Nagrinėjamuoju atveju

$$\chi^2 = \frac{(3625 - 3014)^2(105)}{(33)(39)(55)(50)} = 3{,}418.$$

Kol kas dar neparinkome reikšmingumo lygmens. Kadangi

$$\chi^2_{0,10}(1) = 2{,}704 < 3{,}418 < 3{,}841 = \chi^2_{0,05}(1),$$

tai pasirinkus reikšmingumo lygmenį $\alpha = 0{,}10$ nagrinėjamų grupių skirtumus reiktų pripažinti statistiškai reikšmingais, pasirinkus $\alpha = 0{,}05$ – statistiškai nereikšmingais. Taigi šiuo atveju atsakymas labai priklauso nuo pasirinkto reikšmingumo lygmens!

[1] Frank Yates (1902–1994) – anglų statistikas.

3.5.17 lentelė. Antsvoris ir požiūris į kūno kultūrą esant didesnei imčiai

	Neturi antsvorio	Turi antsvorio	Iš viso
Mėgsta	36 (a)	14 (b)	50 ($a+b$)
Nemėgsta	30 (c)	25 (d)	55 ($c+d$)
Iš viso	66 ($a+c$)	39 ($b+d$)	105 (n)

5.6. Pastabos apie χ^2 kriterijaus naudojimą

Skyriaus pradžioje rašėme, kad χ^2 kriterijaus statistika yra *aproksimuojama* χ^2 skirstiniu. Aproksimavimas laikomas pakankamai tiksliu (t. y. kriterijaus rezultatai patikimi), jei n ne mažesnis kaip 30 ir bent 75% dažnių lentelės gardelių tikėtini dažniai ne mažesni kaip 5. Be to, visi stebėjimai turi būti nepriklausomi ir kiekvienas stebėjimas turi patekti tik į vieną gardelę.

Svarbu nepulti taikyti χ^2 kriterijaus tik pamačius dažnių lentelę. Toliau pateikiame keletą humoristinių lentelių, kurios galbūt padės atidžiau ir kritiškiau vertinti pasirenkamus uždavinius.

	V	M
Laukiasi		
Nesilaukia		

	Laikė egzaminą	Nelaikė egzamino
Neišlaikė		
Išlaikė		

Ką daryti, jei tikėtini dažniai mažesni už 5? Jeigu turime 2×2 lentelę, galime taikyti tikslųjį Fišerio kriterijų. Stebint nedvireikšmius kintamuosius, problemą kartais galima išspręsti mažinant kategorijų skaičių.

3.5.5 pavyzdys. Sociologas gavo užduotį nustatyti, ar yra priklausomybė tarp išsilavinimo ir atsakymo į klausimą apie Lietuvos narystę ES. Tarkime, apklausus 120 atsitiktinai parinktų Lietuvos piliečių buvo gauti rezultatai, nurodyti 3.5.18 lentelėje. (Pateikiami hipotetiniai duomenys.)

3.5.18 lentelė. Išsilavinimas ir požiūris į ES

Išsilavinimas	Narystė ES		
	Taip	Ne	Nežino
Pradinis	6 (7,5)	12 (7,7)	10 (12,8)
Nebaigtas vidurinis	2 (3,5)	8 (3,6)	3 (6,0)
Vidurinis	3 (4,5)	5 (4,7)	9 (7,8)
Specialus vidurinis	3 (4,8)	3 (5,0)	12 (8,3)
Nebaigtas aukštasis	9 (5,1)	3 (5,2)	7 (8,7)
Aukštasis	9 (5,7)	2 (6,1)	14 (11,5)

3.5.19 lentelė. Išsilavinimas ir požiūris į ES

Išsilavinimas	Narystė ES		
	Taip	Ne	Nežino
Pradinis	6 (7,5)	12 (7,7)	10 (12,8)
Vidurinis, nebaigtas vidurinis, specialus vidurinis	8 (8,8)	13 (9,1)	12 (15,1)
Aukštasis ir nebaigtas aukštasis	18 (15,7)	8 (16,2)	33 (27,0)

Čia skliaustuose parašyti atitinkamų gardelių apskaičiuoti tikėtini dažniai. Matome, kad 6 gardelių (tai sudaro 33%) tikėtini dažniai mažesni už 5, t. y. nepatenkinta viena iš χ^2 kriterijaus taikymo sąlygų (tiksliau sakant, rekomendacijų). Pakeiskime 3.5.18 dažnių lentelę 3.5.19 dažnių lentele.

Šiai dažnių lentelei jau galima taikyti χ^2 kriterijų. Aišku, dalį informacijos praradome. Todėl tik tyrėjas gali nuspręsti, ar jungtiniams duomenims taikyti χ^2 kriterijų, ar pradinei dažnių lentelei parinkti kitą matematinį modelį.

5.7. Maknemaro kriterijus priklausomoms dvireikšmėms populiacijoms

Maknemaro[1] kriterijų galima laikyti vienu iš χ^2 kriterijaus variantų. Kada jis yra taikomas?

Tarkime, apklaustųjų nuomonė kokiu nors klausimu (pritaria, nepritaria) vertinama iki pokalbio su jais ir po pokalbio:

pacientų savijauta (patenkinama, nepatenkinama) iki terapijos ir po terapijos;

potencialių pirkėjų (prekę pirkti verta, prekės pirkti neverta) iki reklaminės kampanijos ir po ir pan.

Tiriamas *dvireikšmis* kintamasis (nuostata, gebėjimai, sveikata ir pan.) matuojamas *du kartus* (iki poveikio ir po jo). Dažnai tokio tipo uždaviniuose dvireikšmio kintamojo reikšmės koduojamos $+$, $-$ arba *taip*, *ne*. Nulinė hipotezė H_0 teigia, kad populiacijos dalis, kuriai matuojamo kintamojo reikšmė pasikeitė iš $+$ į $-$, lygi daliai, kuriai kintamojo reikšmė pasikeitė iš $-$ į $+$. Duomenų aibė užrašyta 3.5.20 lentele.

3.5.20 lentelė. Priklausomų požymių dažniai

Prieš	Po		
	$+$	$-$	Σ
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d$

Iš viso pakeitusių nuomonę respondentų yra $b+c$. Jeigu teisinga nulinė hipotezė, kad pakeitusių nuomonę iš $+$ į $-$ respondentų skaičius sutampa su pakeitusių nuomonę iš $-$ į $+$ respondentų skaičiumi, tai $e_{12} = e_{21} = (b+c)/2$. Kadangi mus domina

[1] O. McNemar – amerikiečių statistikas.

tik tie respondentai, kurie keitė nuomonę, tai skaičiuojant statistiką tik jų duomenys ir naudojami. Esant teisingai nulinei hipotezei dydis

$$\frac{(b-(b+c)/2)^2}{(b+c)/2}+\frac{(c-(b+c)/2)^2}{(b+c)/2}=\frac{(b-c)^2}{(b+c)}$$

aproksimuojamas χ^2 skirstiniu su vienu laisvės laipsniu. Atsižvelgus į tolydumo pataisą, užrašoma tokia statistika:

$$\chi^2=\frac{(|b-c|-1)^2}{(b+c)}. \tag{3.5.13}$$

Jei reikšmingumo lygmuo lygus α, tai hipotezę apie nuomonę pakeitusių respondentų vienodą skaičių atmetame. Hipotezei tikrinti galime naudoti ir statistiką

$$Z=\frac{|b-c|-1}{\sqrt{(b+c)}}.$$

Jei $|Z|>z_{\alpha/2}$, tai hipotezę atmetame (esant reikšmingumo lygmeniui α). Čia $z_{\alpha/2}$ – normaliojo skirstinio $\alpha/2$ lygmens kritinė reikšmė.

Prieš ir po dviejų pretendentų į prezidentus (tarkime, pono L. ir pono V.) TV debatų atsitiktinai buvo apklausta 500 TV žiūrovų, už ką jie ruošiasi balsuoti. Apklausos rezultatai pateikiami 3.5.21 lentelėje.

3.5.21 lentelė. TV debatai

Prieš debatus	Po debatų		
	Už L.	Už V.	Iš viso
Už L.	269	36	305
Už V.	21	174	195
Iš viso	290	210	500

Ar debatai pagausino kurio nors kandidato potencialių rinkėjų būrį?

1 *Duomenys*. Dvireikšmių porinių stebėjimų aibė pateikta 3.5.21 lentele.

2 *Statistinė hipotezė*. Nulinė hipotezė H_0 teigia, kad simpatizuojančių kandidatams L. ir V. žiūrovų dalys populiacijoje liko nepakitusios. Alternatyva H_1 teigia, kad dalis remiančių vieną iš kandidatų po debatų padidėjo (o remiančių kitą kandidatą sumažėjo). Taigi

$$\begin{cases} H_0:\ p_{12}=p_{21}, \\ H_1:\ p_{12}\neq p_{21}. \end{cases} \tag{3.5.14}$$

Čia p_{12} yra dalis žiūrovų, kurie prieš TV debatus ruošėsi balsuoti už kandidatą L., o po TV debatų – už V.; p_{21} – dalis žiūrovų, kurie prieš TV debatus ruošėsi balsuoti už kandidatą V., o po TV debatų – už L.

3 *Kriterijaus statistika.* Apskaičiuojame

$$\chi^2 = \frac{(|b-c|-1)^2}{(b+c)} = 3{,}439. \qquad (3.5.15)$$

4 *Sprendimo priėmimo taisyklė.* Tarkime, kad reikšmingumo lygmuo $\alpha = 0{,}01$. Jei statistikos reikšmė būtų didesnė už χ^2 skirstinio su $(2-1)(2-1) = 1$ laisvės laipsniu 0,01 lygmens kritinę reikšmę ($\chi^2_{0,01}(1) = 6{,}635$), hipotezę reiktų atmesti.

Kadangi $3{,}439 < 6{,}635$, nulinę hipotezę atmesti nėra pagrindo.

Išvada. TV debatai kandidatų rėmėjų skaičiaus pokyčiams įtakos neturėjo.

Atkreipiame dėmesį, kad nors kai kurie rinkėjai ir pakeitė savo nuomonę, tačiau dalis remiančių L. (taigi ir potencialių jo rinkėjų *skaičius*) nepakito. Tas pats pasakytina ir apie V. rinkėjus. Taigi nors per debatus kandidatai pritraukė šiek tiek naujų rinkėjų, tiek pat jų ir atgrasė.

Maknemaro kriterijus naudojamas ir proporcijų lygybei dviejose *priklausomose* imtyse tikrinti. Šiuo atveju užtenka 3.5.20 lentelėje *Po* ir *Prieš* pakeisti į *Pirmoji populiacija, Antroji populiacija*.

5.8. Kategorinių duomenų ryšio matai

Nuo imties didumo labai priklauso χ^2 kriterijaus taikymo rezultatai. Jei imties didumas n pakankamai didelis, χ^2 kriterijus fiksuoja mažiausius nuokrypius nuo nepriklausomumo. Praktiškai, jei n labai didelis ($n \gg 1000$), beveik visuomet hipotezė apie požymių nepriklausomumą atmetama. Bet ar kintamųjų (požymių) statistinė priklausomybė visada turi praktinės reikšmės? Kitaip sakant, ar žinodami parinkto individo vieno iš požymių reikšmę galime pakankamai tiksliai atspėti kito požymio reikšmę? Vien konstatavimas, kad požymiai yra priklausomi, nėra pakankamas atsakyti į čia pateiktus klausimus. Papildomai reikia įvertinti požymių ryšio stiprumą. Žinome, kad tolydžių kintamųjų atveju ryšio stiprumo matas yra Pirsono koreliacijos koeficientas. Šiame skyrelyje pateikiame keletą ryšio matų, kai stebimi kategoriniai kintamieji matuojami pagal nominalią skalę. Be abejo, šie ryšio matai tinka ir ranginiams kintamiesiems, tačiau tada geriau taikyti tikslesnius ranginius ryšio matus, kurie aprašyti II knygoje. Visiems čia aprašytiems ryšio matams galioja ta pati taisyklė – kuo jie absoliučiuoju didumu didesni, tuo požymių priklausomybė didesnė; kuo arčiau nulio – tuo priklausomybė silpnesnė. Daugelio iš jų maksimali reikšmė lygi vienetui.

5.8.1. *Ryšio matai 2×2 dažnių lentelėms*

Iš pradžių aptarsime ryšio matus 2×2 dažnių lentelėms, t. y. tarsime, kad duomenys yra tokie kaip 3.5.13 lentelėje.

Koeficientas ϕ dar vadinamas tarpusavio sutapimo rodikliu. Jis apibrėžiamas taip:

$$\phi = \sqrt{\frac{\chi^2}{n}}; \qquad (3.5.16)$$

čia χ^2 – statistikos, apskaičiuotos pagal (3.5.6) formulę, reikšmė. Įsistatę į (3.5.16) formulę χ^2 išraišką, gauname

$$\phi = \sqrt{\frac{(ad-bc)^2 n}{(a+b)(c+d)(b+d)(a+c)n}} = \frac{|ad-bc|}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}. \tag{3.5.17}$$

Koeficiento ϕ kitimo sritis yra $[0; 1]$.

Pastaba. Kartais (3.5.17) formulėje rašoma ne $|ad-bc|$, o $ad-bc$, t.y. absoliutusis didumas praleidžiamas. Tuomet koeficientas ϕ kinta nuo -1 iki 1.

Fiksuotoms eilučių ir stulpelių sumoms koeficiento ϕ minimali ir maksimali reikšmės gali labai skirtis nuo 0 ir 1. Norėdami išvengti šio trūkumo, dalis statistikų skaičiavimams naudoja koeficientą ϕ_{adj}:

$$\phi_{\mathrm{adj}} = \frac{\phi}{|\phi|_{\max}}. \tag{3.5.18}$$

Koeficientas $\phi_{\max}$ apskaičiuojamas taip:

1) randama mažiausia eilutės narių suma $eil = \min(a+b, c+d)$ ir mažiausia stulpelių suma $stulp = \min(a+c, b+d)$;
2) mažesnysis iš skaičių eil, $stulp$ padalijamas iš imties didumo $n = a+b+c+d$ ir pažymimas $p_{\min}$ (taigi $p_{\min} = \min(eil, stulp)/n$);
3) didesnysis iš skaičių eil, $stulp$ padalijamas iš imties didumo $n = a+b=c+d$ ir pažymimas $p_{\max}$ (taigi $p_{\max} = \max(eil, stulp)/n$);
4) gauti dydžiai įstatomi į formulę

$$\phi_{\max} = \frac{\sqrt{p_{\min}(1-p_{\max})}}{\sqrt{p_{\max}(1-p_{\min})}}. \tag{3.5.19}$$

3.5.6 pavyzdys. Apskaičiuokime ϕ ir ϕ_{adj} 3.5.1 pavyzdžiui. Duomenų aibė pateikta 3.5.17 lentelė. Taigi $eil = \min(50, 55) = 50$, $stulp = \min(66, 39) = 39$, $p_{\min} = \min(50, 39)/105 = 39/105 = 0{,}371\ldots$, $p_{\max} = \max(50, 39)/105 = 50/105 = 0{,}476\ldots$,

$$\phi_{\max} = \frac{\sqrt{0{,}371(1-0{,}476)}}{\sqrt{0{,}476(1-0{,}371)}} = 0{,}806.$$

Pasinaudoję (3.5.17) ir (3.5.18) formulėmis, randame:

$$\phi = \frac{|36\cdot 25 - 14\cdot 30|}{\sqrt{50\cdot 55\cdot 39\cdot 66}} = 0{,}180, \qquad \phi_{\mathrm{adj}} = \frac{0{,}180}{0{,}806} = 0{,}223. \tag{3.5.20}$$

Skaičiuojant ryšio matus rekomenduojama, rasti abu koeficientus ϕ, ϕ_{adj}.

Kaip keičiasi ϕ, kintant stebėjimų skaičiui gardelėse? Įsivaizduokime tokią 3.5.1 pavyzdžio situaciją. Tarkime, kad *visi* turintieji antsvorio pareiškė, kad kūno kultūros pamokų nemėgsta; *visi* neturintys – mėgsta. Duomenų aibė pateikta 3.5.22 lentelėje.

3.5.22 lentelė. Antsvoris ir požiūris į kūno kultūrą

	Neturi antsvorio	Turi antsvorio	Iš viso
Mėgsta	66 (a)	0 (b)	66 $(a+b)$
Nemėgsta	0 (c)	39 (d)	39 $(c+d)$
Iš viso	66 $(a+c)$	39 $(b+d)$	105 (n)

Aišku, kad „antsvoris" ir „vertinimas" yra visiškai priklausomi požymiai. Mums ši hipotetinė situacija reikalinga tam, kad galėtume įvertinti, kokia šiuo atveju ϕ reikšmė. Pagal (3.5.17) formulę apskaičiuojame

$$\phi = \frac{|66 \cdot 39 - 0|}{\sqrt{66 \cdot 39 \cdot 39 \cdot 66}} = 1. \tag{3.5.21}$$

Taigi, esant visiškai požymių priklausomybei, $\phi = 1$.

Pastaba. Iki šiol kalbėjome apie dvireikšmių kintamųjų (2×2 dažnių lentelių) koeficientą ϕ. Kartais koeficientas ϕ skaičiuojamas ir didesnėms, t. y. $r\times c$, dažnių lentelėms. Tuomet ϕ apibrėžiamas (3.5.16) formule, čia χ^2 – statistika, skaičiuojama pagal (3.5.6) formulę.

Julo*[1] *asociacijos koeficientas dar vadinamas *keturlaukės koreliacijos* koeficientu. Tarkime, kad dvireikšmių požymių duomenų aibė nusakyta 3.5.13 lentele. Tuomet Julo koeficientas apibrėžiamas taip:

$$Q = \frac{ad - bc}{ad + bc}. \tag{3.5.22}$$

Q kinta nuo -1 iki $+1$. Jei $Q = 0$, tai tarp stebimų dvireikšmių kintamųjų ryšio nėra. Apskaičiuokime Q 3.5.17 dažnių lentelei:

$$Q = \frac{36\times 25 - 30\times 14}{36\times 25 + 30\times 14} = \frac{480}{1320} = 0{,}36.$$

Galima padaryti išvadą, kad *tikėtina*, jog antsvorio turintys labiau nemėgsta kūno kultūros pamokų nei jo neturintys. Praktiškai dažnai naudojamas toks koeficiento Q absoliučiųjų reikšmių interpretavimas „iš akies":

$$\begin{aligned}0 \leqslant |Q| \leqslant 0{,}24 &- \text{ryšio nėra arba jis labai silpnas;}\\ 0{,}25 \leqslant |Q| \leqslant 0{,}49 &- \text{silpnas ryšys;}\\ 0{,}50 \leqslant |Q| \leqslant 0{,}74 &- \text{vidutinio stiprumo ryšys;}\\ 0{,}75 \leqslant |Q| \leqslant 1 &- \text{stiprus ryšys.}\end{aligned}$$

Pastaba. Q kintamųjų ryšio stiprumui vertinti nenaudojamas, kai vienoje iš gardelių dažnis lygus 0. Iš tikrųjų tuo atveju $|Q| = 1$, tačiau tai nereiškia visiškos stebimų dvireikšmių kintamųjų priklausomybės.

Kuo absoliučiosios koeficientų Q ir ϕ reikšmės didesnės, tuo kintamųjų ryšys stipresnis. Visada $|Q| \leqslant \phi$. Nagrinėto pavyzdžio kintamųjų ryšys yra vienpusis. Antsvoris sąlygoja požiūrį į kūno kultūros pamokas, o ne atvirkščiai. Tuo atveju Q geriau atskleidžia empirinį ryšį. Jei ieškoma lygiaverčių požymių ryšio (t. y. daroma prielaida, jog jie abu sąlygojami to paties faktoriaus arba abu vienas kitą veikia), svarbesnis yra dvipusis ryšys, o jį teisingiau nusako koeficientas ϕ. Pavyzdžiui, tiriant priklausomybę tarp žmonių plaukų ir akių spalvos, geriau skaičiuoti ϕ koeficientą!

[1] George Udney Yule (1871–1951) – britų statistikas.

5.8.2. Ryšio matai $r \times c$ dažnių lentelėms

Yra daug 3.5.8 lentelės duomenų požymių priklausomybę aprašančių koeficientų. Čia aptarsime tik kelis iš jų.

Kontingencijos koeficientas C dar vadinamas Pirsono kontingencijos matu. Jis apibrėžiamas taip:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{3.5.23}$$

Matome, kad C yra taip „pataisytas“ ϕ, kad niekada neviršytų 1. Naudojant šį koeficientą, reikia atsižvelgti į tai, kad didžiausia C reikšmė priklauso nuo eilučių ir stulpelių skaičiaus. C niekada neviršija

$$C_{\max} = \sqrt{\frac{k-1}{k}}; \tag{3.5.24}$$

čia $k = \min(r, c)$.

Kramero koeficientas V 3.5.8 dažnių lentelei apibrėžiamas taip:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}. \tag{3.5.25}$$

Kramero koeficientas $0 \leqslant V \leqslant 1$. Dažnių lentelei 2×2 Kramero koeficientas V sutampa su koeficientu ϕ.

Sąlyginis prognozės indeksas λ. Vienas iš svarbesnių ryšio matų yra sąlyginis prognozės indeksas, kurį įvedė Gudmenas[1] ir Kruskalas[2]. Šis koeficientas įvertina vieno požymio kategorijos nuspėjamumo santykinį klaidos sumažėjimą, kai žinoma kito požymio kategorija.

Tarkime, stebima dviejų tam tikros populiacijos požymių pora (X, Y). Požymių (X, Y) poros *skirstinys* pateikiamas 3.5.23 lentele.

Pavyzdžiui, tikimybė, kad atsitiktinai parinktam individui vektoriaus (X, Y) reikšmė yra (x_2, y_1), lygi 0,06. Kitaip sakant, populiacijos dalis, kuriai požymio X reikšmė yra x_2 ir požymio Y reikšmė yra y_1, lygi 0,06.

3.5.23 lentelė. Požymių poros (X, Y) skirstinys

	x_1	x_1	x_3	Σ
y_1	0,09	0,06	0,15	0,30
y_2	0,06	0,04	0,05	0,15
y_3	0,05	0,40	0,10	0,55
Σ	0,20	0,50	0,30	1,00

[1] L. A. Goodman – amerikiečių statistikas.
[2] W. H. Kruskal – amerikiečių statistikas.

3.5.24 lentelė. Požymio X skirstinys

X	x_1	x_1	x_3
P	0,20	0,50	0,30

Tarkime, norime atspėti, kokia bus požymio X reikšmė atsitiktinai parinktam individui, jei požymio Y reikšmė yra nežinoma. Požymio X ribinis skirstinys, kai Y bet koks, pateiktas 3.5.24 lentelėje. Jis gaunamas sumuojant 3.5.23 lentelės elementus.

Aišku, spėtume, kad X įgis reikšmę x_2, kadangi $P(X = x_2) = \max_{j=1,2,3} P(X = x_j) = 0{,}50$. Kokia tikimybė apsirikti teigiant, kad įgyjamoji reikšmė yra x_2? Ši tikimybė yra lygi tikimybei, kad X įgis *ne* x_2, t. y.

$$P(\text{klaida, kai į } Y \text{ neatsižvelgiama}) = 1 - P(X = x_2) = 1 - 0{,}5 = 0{,}5. \tag{3.5.26}$$

Kokia klaidos tikimybė naudojantis informacija apie požymį Y? Pasirėmę pilnosios tikimybės formule, užrašome:

$$\begin{aligned} P(\text{klaida, kai į } Y \text{ atsižvelgiama}) &= P(\text{klaida } | Y = y_1)P(Y = y_1) \\ &+ P(\text{klaida } | Y = y_2)P(Y = y_2) + P(\text{klaida } | Y = y_3)P(Y = y_3). \end{aligned} \tag{3.5.27}$$

Tarkime, žinome, kad parinkto individo požymio Y reikšmė yra y_1. Tuomet geriausia tokia spėjamoji X reikšmė x_j, $j = 1, 2, 3$, kuri maksimizuoja tikimybę $P((X = x_j) \cap (Y = y_1))$. Tokia reikšmė yra x_3. Iš tiesų $P((X = x_3) \cap (Y = y_1)) = 0{,}15 = \max(0{,}09; 0{,}06; 0{,}15)$,

$$\begin{aligned} P(\text{klaida } |Y = y_1) &= P(\overline{X} = x_3 | Y = y_1) = 1 - P((X = x_3)|(Y = y_1)) \\ &= 1 - \frac{P((X = x_3) \cap (Y = y_1))}{P(Y = y_1)} = 1 - \frac{0{,}15}{0{,}30} = 0{,}5. \end{aligned}$$

Analogiškai apskaičiavę kitas sąlygines tikimybes ir įrašę į (3.5.27) formulę, gauname

$$\begin{aligned} &P(\text{klaida, kai į } Y \text{atsižvelgiama}) \\ &= \left(1 - \frac{0{,}15}{0{,}30}\right)0{,}30 + \left(1 - \frac{0{,}06}{0{,}15}\right)0{,}15 + \left(1 - \frac{0{,}40}{0{,}55}\right)0{,}55 = 0{,}39. \end{aligned}$$

Klaidos sumažėjimas, kai X reikšmės spėjimui naudojama informacija apie požymį Y, yra lygus

$$\begin{aligned} &P(\text{klaida, kai į } Y \text{ neatsižvelgiama}) - P(\text{klaida, kai į } Y \text{ atsižvelgiama}) \\ &= 0{,}50 - 0{,}39 = 0{,}11. \end{aligned}$$

Taigi jei spėdami naudojamės informacija apie Y, klaida sumažėja 11%. *Santykinis* klaidos tikimybės sumažėjimas vadinamas *sąlyginiu prognozės indeksu* λ. Jis skaičiuojamas pagal formulę

$$\lambda_X = \frac{P(\text{klaida, kai į } Y \text{ neatsižvelgiama}) - P(\text{klaida, kai į } Y \text{ atsižvelgiama})}{P(\text{klaida, kai į } Y \text{ neatsižvelgiama})}.$$

Mūsų atveju

$$\lambda = \frac{0,11}{0,50} = 0,22.$$

Vadinasi, galima sakyti, kad spėjimo klaida sumažėja 22%, jei spėjant naudojamasi informacija apie Y.

Aišku, kad praktiškai požymių poros (X, Y) skirstinio nežinome. Tuo atveju galima apskaičiuoti tik indekso λ_X įvertį l_X. Tarkime, požymių (X, Y) stebėjimų duomenų aibė užrašoma 3.5.25 dažnių lentele.

3.5.25 lentelė. Porinė (X, Y) dažnių lentelė

	x_1	x_2	$\dots$	x_c	$\sum$
y_1	o_{11}	o_{12}	$\dots$	o_{1c}	R_1
y_2	o_{21}	o_{22}	$\dots$	o_{2c}	R_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_r	o_{r1}	o_{r2}	$\dots$	o_{rc}	R_r
$\sum$	C_1	C_2	$\dots$	C_c	n

Tada

$$l_X = \frac{\max(o_{11}, \dots, o_{1c}) + \dots + \max(o_{r1}, \dots, o_{rc}) - \max(C_1, \dots, C_c)}{n - \max(C_1, \dots, C_c)}. \qquad (3.5.28)$$

Pastaba. $0 \leqslant \lambda_X \leqslant 1$. Jei požymiai X ir Y yra „visiškai" nepriklausomi, tai $\lambda_X = 0$. Jei požymio X kategorijos yra tiksliai atspėjamos, kai žinomos požymio Y kategorijos, tai $\lambda_X = 1$. Atkreipiame dėmesį, kad lygybė $\lambda_X = 0$ dar nereiškia, jog X ir Y yra nepriklausomi. Šiuo atveju galima tik sakyti, kad informacija apie Y nepagerina X įgyjamos reikšmės spėjimo rezultatų. Analogiškai apskaičiuojamas ir λ_Y. Tuo atveju (3.5.38) formulėje eilutės ir stulpeliai susikeičiami vietomis. Bendruoju atveju $\lambda_X \neq \lambda_Y$.

3.5.5 pavyzdys. Apskaičiuokime 3.5.2 pavyzdžiui $l_{\text{antsvoris}}$ ir $l_{\text{mėgimas}}$. Šiuo atveju $C_1 = 66$, $C_2 = 39$, $n = 105$, $R_1 = 50$, $R_2 = 55$,

$$l_{\text{antsvoris}} = \frac{36 + 30 - \max(66, 39)}{105 - \max(66, 39)} = 0,$$

$$l_{\text{mėgimas}} = \frac{36 + 25 - \max(50, 55)}{105 - \max(50, 55)} = 0,22.$$

Gavome, kad informacija apie kūno kultūros pamokų mėgimą nesumažina klaidos tikimybės spėjant, ar mokinys turi antsvorio. Tačiau informacija apie mokinio antsvorį 22% sumažina klaidos tikimybę spėjant, ar mokiniui patinka kūno kultūros pamokos.

---

keturlaukės koreliacijos koeficientas
kontingencijos koeficientas
Kramero koeficientas V
Maknemaro kriterijus
sąlyginis prognozės indeksas
tikslusis Fišerio kriterijus
ϕ koeficientas
χ^2 homogeniškumo kriterijus
χ^2 nepriklausomumo kriterijus
χ^2 suderinamumo kriterijus

UŽDAVINIAI

1. Atliktas tyrimas, kurio tikslas – nustatyti, kokios spalvos automobiliai populiariausi. Atsitiktinai apklausus 200 potencialių pirkėjų, gauti rezultatai, kurie pateikiami 3.5.26 lentelėje.

3.5.26 lentelė. Automobilio spalva ir perkamumas

Spalva	Raudona	Geltona	Mėlyna	Žalia	Ruda
Dažnis	39	65	46	37	13

Patikrinkite hipotezę, kad pirkėjai nevienodai vertina automobilių spalvas ($\alpha = 0{,}05$).

2. Rinkos analitikas mano, kad A, B, C ir D rūšies dantų pastos vartotojų dalis yra atitinkamai 0,30; 0,60; 0,08 ir 0,02. Atsitiktinai apklausus 600 žmonių, kokią pastą jie vartoja, gauti rezultatai, kurie pateikiami 3.5.27 lentelėje.

3.5.27 lentelė. Dantų pastų vartotojai

Rūšis	A	B	C	D
Dažnis	192	342	44	22

Ar šie duomenys leidžia suabejoti rinkos analitiko teiginiu ($\alpha = 0{,}01$)?

3. Tarkime, turime 100 pacientų vyrų nuo 29 iki 59 metų amžiaus sistolinio kraujo spaudimo duomenų aibę, kuri pateikiama 3.5.28 lentele.

3.5.28 lentelė. Sistolinis kraujo spaudimas

Intervalas	Dažnis	Intervalas	Dažnis
90–99	5	140–149	9
100–109	8	150–159	5
110–119	22	160–169	5
120–129	27	170–179	2
130–139	17		

Ar galima teigti, kad pacientų sistolinis kraujo spaudimas pasiskirstęs pagal normalųjį dėsnį?

4. Buvo tirta, ar užimamos pareigos ir pasitenkinimas darbu yra tarpusavyje susiję dalykai. Atsitiktinai apklausus 800 aukštųjų mokyklų dėstytojų, buvo gauti tokie rezultatai:

3.5.29 lentelė. Pasitenkinimas darbu

	Asistentas	Vyr. asistentas	Docentas	Profesorius
Patenkintas	40	60	52	63
Neturi nuomonės	78	87	82	88
Nepatenkintas	57	63	66	64

Pasirinkę reikšmingumo lygmenį $\alpha = 0{,}05$, patikrinkite hipotezę apie pareigų ir pasitenkinimo darbu priklausomybę.

5. Atliktas tyrimas, ar vitamino C kasdienis vartojimas padeda išvengti peršalimo. Atsitiktinai apklausus 100 žmonių, ar jie buvo peršalę praėjusiais metais, buvo gauti tokie rezultatai:

3.5.30 lentelė. Peršalimas ir vitaminas C

	Sirgo	Nesirgo
Vartojo	10	20
Nevartojo	21	14

Ar kasdienis vitamino C vartojimas apsaugo nuo peršalimo?

6. Sveikatos apsaugos ministerija tyrė, ar įvairių profesijų žmonių alkoholio vartojimo įpročiai yra tokie pat. Atsitiktinai apklausus 200 mokytojų, 300 teisininkų ir 400 gydytojų, buvo gauti tokie rezultatai:

3.5.31 lentelė. Profesijos ir alkoholis

	Mokytojai	Teisininkai	Gydytojai
Mažai	100	50	100
Vidutiniškai	50	150	200
Daug	50	100	200
Σ	200	300	400

Ar galima teigti, kad šių trijų profesijų atstovų alkoholio vartojimo įpročiai tokie pat? Tarkime, reikšmingumo lygmuo $\alpha = 0{,}05$.

7. Reklamos agentūra atliko tyrimą, norėdama nustatyti, ar skalbimo miltelių nemokamas dalijimas bandymams namie turi įtakos potencialių pirkėjų pasirinkimui. Atsitiktinai apklausti 100 pirkėjų turėjo pasakyti, kurie skalbimo milteliai – A ar B geresni. Pakartotinai savo nuomonę pirkėjai pareiškė abiejų rūšių miltelius išbandę namuose. Buvo gauti tokie rezultatai:

Prieš	Po		
	A	B	Σ
A	41	3	44
B	9	47	56
Σ	50	50	100

Ar skalbimo miltelių bandymas namuose turėjo įtakos pirkėjų nuomonei? ($\alpha = 0{,}05$.)

8. Įrodykite, kad $|Q| \leqslant \phi$.

9. Šeštajam uždaviniui apskaičiuokite ϕ ir $\lambda_{\text{profesija}}$.

Žymenys[1]

n	imties didumas
$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$	variacinė eilutė
f_i	dažnis
$\overline{X}$	imties vidurkis (statistika)
$\overline{x}$	imties vidurkis (realizacija)
S^2, S_X^2	imties dispersija (statistika)
s^2, s_x^2	imties dispersija (realizacija)
S, S_x	standartinis nuokrypis (statistika)
s, s_x	standartinis nuokrypis (realizacija)
Mo	moda
Md	mediana
Q_1, Q_3	pirmasis ir trečiasis kvartiliai
IQR	kvartilių skirtumas
IQV	kokybinės įvairovės indeksas
CV	populiacijos kitimo (variacijos) koeficientas
cv	imties kitimo (variacijos) koeficientas
CVP	procentinis populiacijos kitimo koeficientas
cvp	procentinis imties kitimo koeficientas
Ω	būtinasis įvykis
ω_i	elementarusis įvykis
$\varnothing$	negalimasis įvykis
$A \cup B$	įvykių sąjunga
$A \cap B$	įvykių sankirta
$\bar{A}$	įvykis, priešingas įvykiui A
$P(A)$	įvykio A tikimybė
$P(A \mid B)$	įvykio A sąlyginė tikimybė
$X, Y, Z, \ldots$	atsitiktiniai dydžiai
$\mathbf{E}X$	atsitiktinio dydžio X vidurkis
$\mathbf{D}X$	atsitiktinio dydžio X dispersija
$S_{\overline{X}}$	standartinė vidurkio paklaida
$X \sim \mathcal{P}(\lambda)$	atsitiktinis dydis X turi Puasono skirstinį
$X \sim \mathcal{B}(n, p)$	atsitiktinis dydis X turi binominį skirstinį
$X \sim \mathcal{N}(\mu, \sigma^2)$	atsitiktinis dydis X turi normalųjį skirstinį
$\Phi(x)$	standartinio normaliojo skirstinio pasiskirstymo funkcija
$p(x)$	tankio funkcija
$\varphi_{\mu,\sigma^2}(x)$	normaliojo skirstinio (su parametrais μ ir σ^2) tankio funkcija
z_α	standartinio normaliojo skirstinio α lygmens kritinė reikšmė
$\chi_\alpha^2(n)$	χ^2 skirstinio su n laisvės laipsnių α lygmens kritinė reikšmė
$t_\alpha(n)$	Stjudento skirstinio su n laisvės laipsnių α lygmens kritinė reikšmė
ϱ, ϱ_{XY}	koreliacijos koeficientas
r, r_{xy}	koreliacijos koeficientas (realizacija)
α	reikšmingumo lygmuo
W	kritinė sritis
CRT	centrinė ribinė teorema

[1] Rečiau pasitaikančioms charakteristikoms vartojami tarptautiniai žymenys.

ATSAKYMAI

Įvadas

5. a) diskretusis, b) tolydusis, c) tolydusis, d) diskretusis, e) tolydusis.
6. a), b), d), e), g) kiekybiniai; c), f) kokybiniai.
7. a), b), d) pavadinimų, c) rangų, e) santykių, f) intervalų.
8. a) rangų, b) pavadinimų, c) santykių, d) rangų, e) pavadinimų, f) pavadinimų, g) rangų, h) rangų, i) rangų.

I dalis

4. a) modą, b) modą, c) $\overline{x}$, d) modą. **5.** $\overline{x} = 23{,}76$, $s = 2{,}57$, sąlyginė išskirtis 30.
6. *Nurodymas.* $f(M) = \sum x_i^2 - n\overline{x}^2 + n(\overline{x} - M)^2$. **9.** $-2, -1, 0, 100, 100$.
10. $\overline{x} = 329{,}4$, $Md = 332{,}5$, $Mo = 305$ ir $Mo = 335$, $Q_1 = 259$, $Q_3 = 385$.
14. $(-1, 3)$. **15.** $Mo = 55$, $Md = 50$. **16.** Visi duomenys lygūs. **17.** Sumažės.
18. Homogeniškesnė pirmoji grupė ($IQV_1 = 0{,}88$, $IQV_2 = 0{,}95$). **20.** -7.
21. 20/3. **22.** Ne. **24.** $\overline{x} = 4{,}82$, $s = 0{,}44$, 6,39 – išskirtis, duomenys pasiskirstę nesimetriškai.

II dalis

1. Į $A_1 \cap B_4$ patenka 0 darbuotojų; $A_2 \cap B_3$ – 6; $A_3 \cup B_2$ – 63; $A_1 \cup A_3$ – 61; $B_1 \cup B_2$ – 81; $A_1 \cap (B_1 \cup B_4)$ – 20; $(A_1 \cup A_3) \cap B_2$ – 20.
3. a) $1 - \binom{2}{2}/\binom{7}{2}$, b) $\binom{5}{1}\binom{2}{1}/\binom{7}{2}$. **4.** $1/7!$, $4!2!2!/8!$. **5.** $1/10^4$, $1/A_{10}^4$. **7.** 2/5.
8. 2/21. **9.** 0,56. **10.** $1 - (1-p)^{150}$, $(1 - (1-p)^{15})^{10}$. **11.** $\binom{12}{6}\binom{12}{6}/\binom{24}{12}$.
12. $P(A) = 160/500$, $P(B) = 240/500$, $P(A \cap B) = 100/500$, A ir B priklausomi.
13. $1 \cdot 0{,}60 + 0 \cdot 0{,}30 + 0{,}5 \cdot 0{,}10 = 0{,}65$.
14. $1 - (0{,}80 \cdot 0{,}99 \cdot 4/24 + 0{,}70 \cdot 0{,}99 \cdot 8/24 + 0{,}90 \cdot 0{,}99 \cdot 2/24 + 0 \cdot 10/24) = 0{,}56275$.
15. $(1/6 \cdot 1/20)/(0{,}8 \cdot 3/6 + 0{,}6 \cdot 2/6 + 1/20 \cdot 1/6) = 0{,}0137$.
16.

X	1	2	3	4	5
P	$\frac{4}{8}$	$\frac{4}{8} \cdot \frac{4}{7}$	$\frac{4}{8} \cdot \frac{3}{7} \cdot \frac{4}{6}$	$\frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{4}{7}$	$\frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6} \cdot \frac{1}{5} \cdot \frac{4}{4}$

17. 6/7. **18.** $\mathbf{D}X$, 0, $\mathbf{E}X$, 0. **19.** 20. **22.** $-0{,}2182$.
23. $P(X = k) = 0{,}5(k-1)0{,}5(10{,}5)^{k-2} = (k-1)0{,}5^k$, $k = 2, 3, \ldots$.
25. Razinų bandelėje skaičius $X \sim \mathcal{P}(\lambda)$, čia $\lambda = \mathbf{E}X$ yra vidutinis razinų bandelėje skaičius. Ats. $\lambda \geqslant \ln 10 = 3{,}175$.
26. $2(1 - \Phi(0{,}5)) = 0{,}62$, $\Phi(0{,}75) + \Phi(0{,}25) - 1 = 0{,}48$.

III *dalis*

3.1. *Imties skirstiniai. Įverčiai*

1. a) b) 42/90; c)

(X_1, X_2)	(1,1)	(1,0)	(0,1)	(0,0)
P	42/90	21/90	21/90	6/90

X	0	1	2
P	6/90	42/90	42/90

2. a) $X \sim \mathcal{N}(60; 25)$, $\overline{X} \sim \mathcal{N}(60; 0{,}25)$. b) $P(\overline{X} \leqslant 55) = 0$.
3. $X \sim \mathcal{N}(8; 4)$, $\overline{X} \sim \mathcal{N}(8; 1/9)$. $P(\overline{X} \leqslant 8{,}3) = 0{,}8159$. **5.** Abiem atvejais $\widehat{\lambda} = 1/\overline{X}$.
6. Dešimt kartų stebime $X \sim \mathcal{B}(100, p)$. Todėl $n = 10$. Didžiausio tikėtinumo įvertis $\widehat{p} = \overline{X}/100$. Įverčio realizacija 0,678.
8. (4,7; 9,5).
9. Vidurkio pasikliautinasis intervalas (−2,20; 1,47), dispersijos pasikliautinasis intervalas (3,64; 22,96).
10. $z_\alpha = 2$, $\alpha = 0{,}023$, $Q = 0{,}954$.
11. $n \geqslant (20 \cdot 1{,}96)^2$. Taigi imtyje turi būti ne mažiau kaip 1537 elementų.
12. $\widehat{\sigma}_1^2 = 1{,}44$, $\widehat{\sigma}_2^2 = 4{,}02$. Stakles galima laikyti išsiderinusiomis.
13. Pirmosios populiacijos vidurkio pasikliautinasis intervalas yra (75,71; 84,29). Antrosios populiacijos vidurkio pasikliautinasis intervalas yra (55,33; 64,67). Intervalai nesikerta, todėl galima teigti, kad populiacijų vidutiniai raštingumai statistiškai reikšmingai skiriasi. Jungtinis pasikliautinasis vidurkio intervalas yra (61,83; 71,51).
14. Stebimasis atsitiktinis dydis X – užsisegęs keleivis diržą ($X = 1$) ar ne ($X = 0$). Taigi $X \sim \mathcal{B}(1, p)$, čia p – nežinomas parametras, atitinkantis diržų nenaudojančių keleivių dalį visoje populiacijoje. Pasikliautinasis intervalas yra (0,32; 0,48).

3.3. *Statistinės išvados vienai imčiai*

1. „Nutekėjo". **2.** Ne. **3.** Neprieštarauja. **4.** Neblogiau. **5.** Neatitinka. **6.** Neprieštarauja. **7.** Prieštarauja. **8.** Prieštarauja. **9.** Neprieštarauja. **10.** Padvigubėjo (alternatyva vienpusė). **11.** Ne. **12.** 1. **13.** Priklauso.

3.4. *Statistinės išvados dviem imtims*

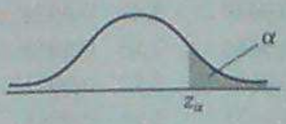
3. Negalima. **4.** Turėjo. **5.** Nesiskiria. **6.** Pirmasis. **7.** Galima sutikti. **8.** Taip. **9.** Negalima. **10.** Nesiskiria. **11.** Nereikšmingas. **12.** Neleidžia.

3.5. *Dažnių lentelės*

1. Nevienodai. **2.** Neleidžia, $p = 0{,}016$. **4.** Nepriklauso. **5.** Taip. **6.** Skirtingus. **7.** Neturėjo. **9.** $\phi = 0{,}30$, $l_{\text{profesija}} = 0{,}2$.

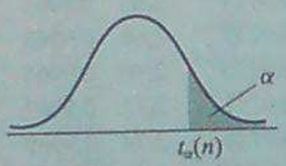
PRIEDAS

Lentelių aprašymas

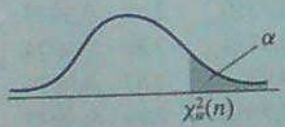


1 lentelėje pateiktos standartinio normaliojo atsitiktinio dydžio pasiskirstymo funkcijos $\Phi(x)$ reikšmės. $\Phi(-x) = 1 - \Phi(x)$.

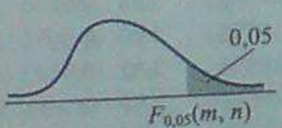
2 lentelėje pateiktos standartinio normaliojo atsitiktinio dydžio α lygmens kritinės reikšmės z_α, t. y. lygties $\Phi(z_\alpha) = 1 - \alpha$ sprendiniai (z_α lygus $1 - \alpha$ lygmens kvantiliui).



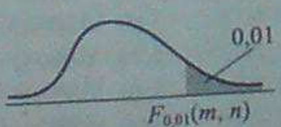
3 lentelėje įvairiems m ir n pateiktos Stjudento atsitiktinio dydžio T su n laisvės laipsnių α lygmens kritinės reikšmės $t_\alpha(n)$, t. y. lygties $P(T > t_\alpha(n)) = \alpha$ sprendiniai ($t_\alpha(n)$ lygus $1 - \alpha$ lygmens kvantiliui).



4 lentelėje įvairiems n ir α pateiktos χ^2 atsitiktinio dydžio su n laisvės laipsnių α lygmens kritinės reikšmės $\chi^2_\alpha(n)$, t. y. lygties $P(\chi^2 > \chi^2_\alpha(n)) = \alpha$ sprendiniai ($\chi^2_\alpha(n)$ lygus $1 - \alpha$ lygmens kvantiliui).



5 lentelėje įvairiems m ir n pateiktos Fišerio atsitiktinio dydžio F su m ir n laisvės laipsnių 0,05 lygmens kritinės reikšmės $F_{0.05}(m, n)$, t. y. $P(F > F_{0,05}(m, n)) = 0{,}05$ sprendiniai ($F_{0.05}(m, n)$ lygus 0,95 lygmens kvantiliui).



6 lentelėje įvairiems m ir n pateiktos Fišerio atsitiktinio dydžio F su m ir n laisvės laipsnių 0,01 lygmens kritinės reikšmės $F_{0.01}(m, n)$, t. y. $P(F > F_{0,01}(m, n)) = 0{,}01$ sprendiniai ($F_{0,01}(m, n)$ lygus 0,99 lygmens kvantiliui).

7 lentelėje pateiktos funkcijos $z_r = \operatorname{arth} r = \frac{1}{2} \ln \frac{1+r}{1-r}$ reikšmės.

1 lentelė. Funkcijos $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ reikšmės

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,5000	0,43	0,6664	0,86	0,8051	1,29	0,9015	1,72	0,9573	2,30	0,9893
0,01	0,5040	0,44	0,6700	0,87	0,8078	1,30	0,9032	1,73	0,9582	2,32	0,9898
0,02	0,5080	0,45	0,6736	0,88	0,8106	1,31	0,9049	1,74	0,9591	2,34	0,9904
0,03	0,5120	0,46	0,6772	0,89	0,8133	1,32	0,9066	1,75	0,9599	2,36	0,9909
0,04	0,5160	0,47	0,6808	0,90	0,8159	1,33	0,9082	1,76	0,9608	2,38	0,9913
0,05	0,5199	0,48	0,6844	0,91	0,8186	1,34	0,9099	1,77	0,9616	2,40	0,9918
0,06	0,5239	0,49	0,6879	0,92	0,8212	1,35	0,9115	1,78	0,9625	2,42	0,9922
0,07	0,5279	0,50	0,6915	0,93	0,8238	1,36	0,9131	1,79	0,9633	2,44	0,9927
0,08	0,5319	0,51	0,6950	0,94	0,8264	1,37	0,9147	1,80	0,9641	2,46	0,9931
0,09	0,5359	0,52	0,6985	0,95	0,8289	1,38	0,9162	1,81	0,9649	2,48	0,9934
0,10	0,5398	0,53	0,7019	0,96	0,8315	1,39	0,9177	1,82	0,9656	2,50	0,9938
0,11	0,5438	0,54	0,7054	0,97	0,8340	1,40	0,9192	1,83	0,9664	2,52	0,9941
0,12	0,5478	0,55	0,7088	0,98	0,8365	1,41	0,9207	1,84	0,9671	2,54	0,9945
0,13	0,5517	0,56	0,7123	0,99	0,8389	1,42	0,9222	1,85	0,9678	2,56	0,9948
0,14	0,5557	0,57	0,7157	1,00	0,8413	1,43	0,9236	1,86	0,9686	2,58	0,9951
0,15	0,5596	0,58	0,7190	1,01	0,8438	1,44	0,9251	1,87	0,9693	2,60	0,9953
0,16	0,5636	0,59	0,7224	1,02	0,8461	1,45	0,9265	1,88	0,9699	2,62	0,9956
0,17	0,5675	0,60	0,7257	1,03	0,8485	1,46	0,9279	1,89	0,9706	2,64	0,9959
0,18	0,5714	0,61	0,7291	1,04	0,8508	1,47	0,9292	1,90	0,9713	2,66	0,9961
0,19	0,5753	0,62	0,7324	1,05	0,8531	1,48	0,9306	1,91	0,9719	2,68	0,9963
0,20	0,5793	0,63	0,7357	1,06	0,8554	1,49	0,9319	1,92	0,9726	2,70	0,9965
0,21	0,5832	0,64	0,7389	1,07	0,8577	1,50	0,9332	1,93	0,9732	2,72	0,9967
0,22	0,5871	0,65	0,7422	1,08	0,8599	1,51	0,9345	1,94	0,9738	2,74	0,9969
0,23	0,5910	0,66	0,7454	1,09	0,8621	1,52	0,9357	1,95	0,9744	2,76	0,9971
0,24	0,5948	0,67	0,7486	1,10	0,8643	1,53	0,9370	1,96	0,9750	2,78	0,9973
0,25	0,5987	0,68	0,7517	1,11	0,8665	1,54	0,9382	1,97	0,9756	2,80	0,9974
0,26	0,6026	0,69	0,7549	1,12	0,8686	1,55	0,9394	1,98	0,9761	2,82	0,9976
0,27	0,6064	0,70	0,7580	1,13	0,8708	1,56	0,9406	1,99	0,9767	2,84	0,9977
0,28	0,6103	0,71	0,7611	1,14	0,8729	1,57	0,9418	2,00	0,9772	2,86	0,9979
0,29	0,6141	0,72	0,7642	1,15	0,8749	1,58	0,9429	2,02	0,9783	2,88	0,9980
0,30	0,6179	0,73	0,7673	1,16	0,8770	1,59	0,9441	2,04	0,9793	2,90	0,9981
0,31	0,6217	0,74	0,7704	1,17	0,8790	1,60	0,9452	2,06	0,9803	2,92	0,9982
0,32	0,6255	0,75	0,7734	1,18	0,8810	1,61	0,9463	2,08	0,9812	2,94	0,9984
0,33	0,6293	0,76	0,7764	1,19	0,8830	1,62	0,9474	2,10	0,9821	2,96	0,9985
0,34	0,6331	0,77	0,7794	1,20	0,8849	1,63	0,9484	2,12	0,9830	2,98	0,9986
0,35	0,6368	0,78	0,7823	1,21	0,8869	1,64	0,9495	2,14	0,9838	3,00	0,99865
0,36	0,6406	0,79	0,7852	1,22	0,8888	1,65	0,9505	2,16	0,9846	3,20	0,99931
0,37	0,6443	0,80	0,7881	1,23	0,8907	1,66	0,9515	2,18	0,9854	3,40	0,99966
0,38	0,6480	0,81	0,7910	1,24	0,8925	1,67	0,9525	2,20	0,9861	3,60	0,999841
0,39	0,6517	0,82	0,7939	1,25	0,8944	1,68	0,9535	2,22	0,9868	3,80	0,999928
0,40	0,6554	0,83	0,7967	1,26	0,8962	1,69	0,9545	2,24	0,9875	4,00	0,999968
0,41	0,6591	0,84	0,7995	1,27	0,8980	1,70	0,9554	2,26	0,9881	4,50	0,999997
0,42	0,6628	0,85	0,8023	1,28	0,8997	1,71	0,9564	2,28	0,9887	5,00	0,9999997

2 lentelė. Normaliojo skirstinio $N(0,1)$ kritinės reikšmės z_α

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
0,50	0,000	0,37	0,331	0,24	0,706	0,11	1,226	0,025	1,959	0,012	2,257
0,49	0,025	0,36	0,358	0,23	0,738	0,10	1,281	0,024	1,977	0,011	2,290
0,48	0,050	0,35	0,385	0,22	0,772	0,09	1,340	0,023	1,995	0,010	2,326
0,47	0,075	0,34	0,412	0,21	0,806	0,08	1,405	0,022	2,014	0,009	2,365
0,46	0,100	0,33	0,439	0,20	0,841	0,07	1,475	0,021	2,033	0,008	2,408
0,45	0,125	0,32	0,467	0,19	0,877	0,06	1,554	0,020	2,053	0,007	2,457
0,44	0,150	0,31	0,495	0,18	0,915	0,05	1,644	0,019	2,074	0,006	2,512
0,43	0,176	0,30	0,524	0,17	0,954	0,04	1,750	0,018	2,096	0,005	2,575
0,42	0,201	0,29	0,553	0,16	0,994	0,03	1,880	0,017	2,120	0,004	2,652
0,41	0,227	0,28	0,582	0,15	1,036	0,029	1,895	0,016	2,144	0,003	2,747
0,40	0,253	0,27	0,612	0,14	1,080	0,028	1,911	0,015	2,170	0,002	2,878
0,39	0,279	0,26	0,643	0,13	1,126	0,027	1,926	0,014	2,197	0,001	3,090
0,38	0,305	0,25	0,674	0,12	1,174	0,026	1,943	0,013	2,226		

3 lentelė. Stjudento skirstinio α lygmens kritinės reikšmės $t_\alpha(n)$

$n\backslash\alpha$	0,40	0,10	0,05	0,025	0,01	0,005
1	0,325	3,078	6,314	12,706	31,821	63,657
2	0,289	1,886	2,920	4,303	6,965	9,925
3	0,277	1,638	2,353	3,182	4,541	5,841
4	0,271	1,533	2,132	2,776	3,747	4,604
5	0,267	1,476	2,015	2,571	3,365	4,032
6	0,265	1,440	1,943	2,447	3,143	3,707
7	0,263	1,415	1,895	2,365	2,998	3,499
8	0,262	1,397	1,860	2,306	2,896	3,355
9	0,261	1,383	1,833	2,262	2,821	3,250
10	0,260	1,372	1,812	2,228	2,764	3,169
11	0,260	1,363	1,796	2,201	2,718	3,106
12	0,259	1,356	1,782	2,179	2,681	3,055
13	0,259	1,350	1,771	2,160	2,650	3,012
14	0,258	1,345	1,761	2,145	2,624	2,977
15	0,258	1,341	1,753	2,131	2,602	2,947
16	0,258	1,337	1,746	2,120	2,583	2,921
17	0,257	1,333	1,740	2,110	2,567	2,898
18	0,257	1,330	1,734	2,101	2,552	2,878
19	0,257	1,328	1,729	2,093	2,539	2,861
20	0,257	1,325	1,725	2,086	2,528	2,845
21	0,257	1,323	1,721	2,080	2,518	2,831
22	0,256	1,321	1,717	2,074	2,508	2,819
23	0,256	1,319	1,714	2,069	2,500	2,807
24	0,256	1,318	1,711	2,064	2,492	2,797
25	0,256	1,316	1,708	2,060	2,485	2,787
30	0,256	1,310	1,697	2,042	2,457	2,750
40	0,255	1,303	1,684	2,021	2,423	2,704
60	0,254	1,296	1,671	2,000	2,390	2,660
120	0,254	1,289	1,658	1,980	2,358	2,617
∞	0,253	1,282	1,645	1,960	2,326	2,576

4 lentelė. χ^2 skirstinio α lygmens kritinės reikšmės $\chi^2_\alpha(n)$

$n\backslash\alpha$	0,9995	0,999	0,995	0,99	0,975	0,95	0,90
1	0	0	0,00003	0,00015	0,00098	0,00393	0,0158
2	0,00100	0,00200	0,0100	0,0201	0,0506	0,103	0,211
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,064
5	0,158	0,210	0,412	0,554	0,831	1,145	1,610
6	0,299	0,381	0,676	0,872	1,237	1,635	2,204
7	0,485	0,598	0,989	1,239	1,690	2,167	2,833
8	0,710	0,857	1,344	1,646	2,180	2,733	3,490
9	0,972	1,153	1,735	2,088	2,700	3,325	4,168
10	1,265	1,479	2,156	2,558	3,247	3,940	4,865
11	1,587	1,834	2,603	3,053	3,816	4,575	5,578
12	1,934	2,214	3,074	3,571	4,404	5,226	6,304
13	2,305	2,617	3,565	4,107	5,009	5,892	7,042
14	2,697	3,041	4,075	4,660	5,629	6,571	7,790
15	3,108	3,483	4,601	5,229	6,262	7,261	8,547
16	3,536	3,942	5,142	5,812	6,908	7,962	9,312
17	3,980	4,416	5,697	6,408	7,564	8,672	10,085
18	4,439	4,905	6,265	7,015	8,231	9,390	10,865
19	4,912	5,407	6,844	7,633	8,907	10,117	11,651
20	5,398	5,921	7,434	8,260	9,591	10,851	12,443
21	5,896	6,447	8,034	8,897	10,283	11,591	13,240
22	6,404	6,983	8,643	9,542	10,982	12,338	14,041
23	6,924	7,529	9,160	10,196	11,688	13,091	14,848
24	7,453	8,085	9,886	10,856	12,401	13,848	15,659
25	7,991	8,649	10,520	11,524	13,120	14,611	16,473
26	8,538	9,222	11,160	12,198	13,844	15,379	17,292
27	9,093	9,803	11,808	12,879	14,573	16,151	18,114
28	9,656	10,391	12,461	13,565	15,308	16,928	18,939
29	10,227	10,986	13,121	14,256	16,047	17,708	19,768
30	10,804	11,588	13,787	14,953	16,791	18,493	20,599
31	11,389	12,196	14,458	15,655	17,539	19,281	21,434
32	11,979	12,811	15,134	16,362	18,291	20,072	22,271
33	12,576	13,431	15,815	17,073	19,047	20,867	23,110
34	13,179	14,057	16,501	17,789	19,806	21,664	23,952
35	13,788	14,688	17,192	18,509	20,569	22,465	24,797
36	14,401	15,324	17,887	19,233	21,336	23,269	25,643
37	15,020	15,965	18,586	19,960	22,106	24,075	26,492
38	15,644	16,611	19,289	20,691	22,878	24,884	27,343
39	16,273	17,262	19,996	21,426	23,654	25,695	28,196
40	16,906	17,916	20,707	22,164	24,433	26,509	29,051
50	23,461	24,674	27,991	29,707	32,357	34,764	37,689
60	30,340	31,738	35,535	37,485	40,482	43,188	46,459
80	44,791	46,520	51,172	53,540	57,153	60,391	64,278
100	59,896	61,918	67,328	70,065	74,222	77,929	82,358

4 lentelės tęsinys

0,10	0,05	0,025	0,01	0,005	0,001	0,0005	$\alpha \backslash n$
2,706	3,841	5,024	6,635	7,879	10,828	12,116	1
4,605	5,991	7,378	9,210	10,597	13,816	15,202	2
6,251	7,815	9,348	11,345	12,838	16,266	17,730	3
7,779	9,488	11,143	13,277	14,860	18,467	19,997	4
9,236	11,070	12,832	15,086	16,750	20,515	22,105	5
10,345	12,592	14,449	16,812	18,548	22,458	24,103	6
12,017	14,067	16,013	18,475	20,278	24,322	26,018	7
13,362	15,507	17,535	20,090	21,955	26,125	27,868	8
14,684	16,919	19,023	21,666	23,589	27,877	29,666	9
15,987	18,307	20,483	23,209	25,188	29,588	31,420	10
17,275	19,675	21,920	24,725	26,757	31,264	33,136	11
18,549	21,026	23,336	26,217	28,300	32,909	34,821	12
19,812	22,362	24,736	27,688	29,819	34,528	36,478	13
21,064	23,685	26,119	29,141	31,319	36,123	38,109	14
22,307	24,996	27,488	30,578	32,801	37,697	39,719	15
23,542	26,296	28,845	32,000	34,267	39,252	41,308	16
24,769	27,587	30,191	33,409	35,718	40,790	42,879	17
25,989	28,869	31,526	34,805	37,156	42,312	44,434	18
27,204	30,144	32,852	36,191	38,582	43,820	45,973	19
28,412	31,410	34,170	37,566	39,997	45,315	47,498	20
29,615	32,671	35,479	38,932	41,401	46,797	49,010	21
30,813	33,924	36,781	40,289	42,796	48,268	50,511	22
32,007	35,172	38,076	41,638	44,181	49,728	52,000	23
33,196	36,415	39,364	42,980	45,558	51,179	53,479	24
34,382	37,652	40,646	44,314	46,928	52,620	54,947	25
35,563	38,885	41,923	45,642	48,290	54,052	56,407	26
36,741	40,113	43,194	46,963	49,645	55,476	57,858	27
37,916	41,337	44,461	48,278	50,993	56,892	59,300	28
39,087	42,557	45,722	49,588	52,336	58,301	60,735	29
40,256	43,773	46,979	50,892	53,672	59,703	62,162	30
41,422	44,985	48,232	52,191	55,003	61,098	63,582	31
42,585	46,194	49,480	53,486	56,328	62,487	64,995	32
43,745	47,400	50,725	54,776	57,648	63,870	66,402	33
44,903	48,602	51,966	56,061	58,964	65,247	67,803	34
46,059	49,802	53,203	57,342	60,275	66,619	69,199	35
47,212	50,998	54,437	58,619	61,581	67,985	70,588	36
48,363	52,192	55,668	59,892	62,882	69,346	71,972	37
49,513	53,384	56,895	61,162	64,181	70,703	73,351	38
50,660	54,572	58,120	62,428	65,476	72,055	74,725	39
51,805	55,758	59,342	63,691	66,766	73,402	76,095	40
63,167	67,505	71,420	76,154	79,490	86,661	89,561	50
74,397	79,082	83,298	88,379	91,952	99,607	102,695	60
96,578	101,879	106,629	112,329	116,321	124,839	128,261	80
118,498	124,342	129,561	135,807	140,169	149,449	153,167	100

5 lentelė. Fišerio skirstinio $\alpha = 0{,}05$ lygmens kritinės reikšmės $F_{\alpha}(m, n)$

$n \backslash m$	1	2	3	4	5	6	8	10	12	15	20	40
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	241,88	243,91	245,95	248,01	251,14
2	18,513	19,000	19,164	19,247	19,296	19,330	19,371	19,396	19,413	19,429	19,446	19,471
3	10,128	9,5521	9,2766	9,1172	9,0135	8,9406	8,8452	8,7855	8,7446	8,7029	8,6602	8,5944
4	7,7086	6,9443	6,5914	6,3883	6,2560	6,1631	6,0410	5,9644	5,9117	5,8578	5,8025	5,7170
5	6,6079	5,7861	5,4095	5,1922	5,0503	4,9503	4,8183	4,7351	4,6777	4,6188	4,5581	4,4638
6	5,9874	5,1433	4,7571	4,5337	4,3874	4,2839	4,1468	4,0600	3,9999	3,9381	3,8742	3,7743
7	5,5914	4,7374	4,3468	4,1203	3,9715	3,8660	3,7257	3,6365	3,5747	3,5108	3,4445	3,3404
8	5,3177	4,4590	4,0662	3,8378	3,6875	3,5806	3,4381	3,3472	3,2840	3,2184	3,1503	3,0428
9	5,1174	4,2565	3,8626	3,6331	3,4817	3,3738	3,2296	3,1373	3,0729	3,0061	2,9365	2,8259
10	4,9646	4,1028	3,7083	3,4780	3,3258	3,2172	3,0717	2,9782	2,9130	2,8450	2,7740	2,6609
11	4,8443	3,9823	3,5874	3,3567	3,2039	3,0946	2,9480	2,8536	2,7876	2,7186	2,6464	2,5309
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,8486	2,7534	2,6866	2,6169	2,5436	2,4259
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,7669	2,6710	2,6037	2,5331	2,4589	2,3392
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,8477	2,6987	2,6021	2,5342	2,4630	2,3879	2,2664
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,6408	2,5437	2,4753	2,4035	2,3275	2,2043
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,7413	2,5911	2,4935	2,4247	2,3522	2,2756	2,1507
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,6987	2,5480	2,4499	2,3807	2,3077	2,2304	2,1040
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,6613	2,5102	2,4117	2,3421	2,2686	2,1906	2,0629
19	4,3808	3,5219	3,1274	2,8951	2,7401	2,6283	2,4768	2,3779	2,3080	2,2341	2,1555	2,0264
20	4,3513	3,4928	3,0984	2,8661	2,7109	2,5990	2,4471	2,3479	2,2776	2,2033	2,1242	1,9938
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4205	2,3210	2,2504	2,1757	2,0960	1,9645
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,5491	2,3965	2,2967	2,2258	2,1508	2,0707	1,9380
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,5277	2,3748	2,2747	2,2036	2,1282	2,0476	1,9139
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,3551	2,2547	2,1834	2,1077	2,0267	1,8920
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,4904	2,3371	2,2365	2,1649	2,0889	2,0075	1,8718
26	4,2252	3,3690	2,9751	2,7426	2,5868	2,4741	2,3205	2,2197	2,1479	2,0716	1,9898	1,8533
27	4,2100	3,3541	2,9604	2,7278	2,5719	2,4591	2,3053	2,2043	2,1323	2,0558	1,9736	1,8361
28	4,1960	3,3404	2,9467	2,7141	2,5581	2,4453	2,2782	2,1900	2,1179	2,0411	1,9586	1,8203
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,4324	2,2782	2,1768	2,1045	2,0275	1,9446	1,8055
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,4205	2,2662	2,1646	2,0921	2,0148	1,9317	1,7918
40	4,0848	3,2317	2,8387	2,6060	2,4495	2,3359	2,1802	2,0772	2,0035	1,9245	1,8389	1,6928
60	4,0012	3,1504	2,7581	2,5252	2,3683	2,2540	2,0970	1,9926	1,9174	1,8364	1,7480	1,5943
120	3,9201	3,0718	2,6802	2,4472	2,2900	2,1750	2,0164	1,9105	1,8337	1,7505	1,6587	1,4952
∞	3,8415	2,9957	2,6049	2,3719	2,2141	2,0986	1,9384	1,8307	1,7522	1,6664	1,5705	1,3940

6 lentelė. Fišerio skirstinio $\alpha = 0{,}01$ lygmens kritinės reikšmės $F_{\alpha}(m, n)$

$n\backslash m$	1	2	3	4	5	6	8	10	12	15	20	40
1	4052,2	4999,5	5403,3	5624,6	5463,7	5859,0	5981,1	6055,8	6106,3	6157,3	6208,7	6268,6
2	98,503	99,000	99,166	99,249	99,299	99,332	99,374	99,399	99,416	99,432	99,449	99,474
3	34,116	30,817	29,457	28,710	28,237	27,911	27,489	27,229	27,052	26,872	26,690	26,411
4	21,198	18,000	16,694	15,977	15,522	15,207	14,799	14,546	14,374	14,198	14,020	13,745
5	16,258	13,274	12,060	11,392	10,967	10,672	10,289	10,051	9,8883	9,7222	9,5527	9,2912
6	13,745	10,925	9,7795	9,1483	8,7459	8,4661	8,1016	7,8741	7,7183	7,5590	7,3958	7,1432
7	12,246	9,5466	8,4513	7,8467	7,4604	7,1914	6,8401	6,6201	6,4691	6,3143	6,1554	5,9084
8	11,259	8,6491	7,5910	7,0060	6,6318	6,3707	6,0289	5,8143	5,6668	5,5151	5,3591	5,1156
9	10,561	8,0215	6,9919	6,4221	6,0569	5,8018	5,4671	5,2565	5,1114	4,9621	4,8080	4,5667
10	10,044	7,5594	6,5523	5,9943	5,6363	5,3858	5,0567	4,8492	4,7059	4,5582	4,4054	4,1653
11	9,6460	7,2057	6,2167	5,6683	5,3160	5,0692	4,7445	4,5393	4,3974	4,2509	4,0990	3,8596
12	9,3302	6,9266	5,9526	5,4119	5,0643	4,8206	4,4994	4,2961	4,1553	4,0096	3,8584	3,6192
13	9,0738	6,7010	5,7394	5,2053	4,8616	4,6204	4,3021	4,1003	3,9603	3,8154	3,6646	3,4253
14	8,8616	6,5149	5,5639	5,0354	4,6950	4,4558	4,1399	3,9394	3,8001	3,6557	3,5052	3,2656
15	8,6831	6,3589	5,4170	4,8932	4,5556	4,3183	4,0045	3,8049	3,6662	3,5222	3,3719	3,1319
16	8,5310	6,2262	5,2922	4,7726	4,4374	4,2016	3,8896	3,6909	3,5527	3,4089	3,2588	3,0182
17	8,3997	6,1121	5,1850	4,6690	4,3359	4,1015	3,7910	3,5931	3,4552	3,3117	3,1615	2,9205
18	8,2854	6,0129	5,0919	4,5790	4,2479	4,0146	3,7054	3,5082	3,3706	3,2273	3,0771	2,8354
19	8,1850	5,9259	5,0103	4,5003	4,1708	3,9386	3,6305	3,4338	3,2965	3,1533	3,0031	2,7608
20	8,0960	5,8489	4,9382	4,4307	4,1027	3,8714	3,5644	3,3682	3,2311	3,0880	2,9377	2,6947
21	8,0166	5,7804	4,8740	4,3688	4,0421	3,8117	3,5056	3,3098	3,1729	3,0299	2,8796	2,6359
22	7,9454	5,7190	4,8166	4,3134	3,9880	3,7583	3,4530	3,2576	3,1209	2,9780	2,8274	2,5831
23	7,8811	5,6637	4,7649	4,2635	3,9392	3,7102	3,4057	3,2106	3,0740	2,9311	2,7805	2,5355
24	7,8229	5,6136	4,7181	4,2184	3,8951	3,6667	3,3629	3,1681	3,0316	2,8887	2,7380	2,4923
25	7,7698	5,5680	4,6755	4,1774	3,8550	3,6272	6,3239	3,1294	2,9931	2,8502	2,6993	2,4530
26	7,7213	5,5263	4,6366	4,1400	3,8183	3,5911	3,2884	3,0941	2,9579	2,8150	2,6640	2,4170
27	7,6767	5,4881	4,6009	4,1056	3,7848	3,5580	3,2558	3,0618	2,9256	2,7827	2,6316	2,3840
28	7,6356	5,4529	4,5681	4,0740	3,7539	3,5276	3,2259	3,0320	2,8959	2,7530	2,6017	2,3535
29	7,5976	5,4205	4,5378	4,0449	3,7254	3,4995	3,1982	3,0045	2,8685	2,7256	2,5742	2,3253
30	7,5625	5,3903	4,5097	4,0179	3,6990	3,4735	3,1726	2,9791	2,8431	2,7002	2,5487	2,2992
40	7,3141	5,1785	4,3126	3,8283	3,5138	3,2910	2,9930	2,8005	2,6648	2,5216	2,3689	2,1142
60	7,0771	4,9774	4,1259	3,6491	3,3389	3,1187	2,8233	2,6318	2,4961	2,3523	2,1978	1,9360
120	6,8510	4,7865	3,9491	3,4796	3,1735	2,9559	2,6629	2,4721	2,3363	2,1915	2,0346	1,7628
∞	6,6349	4,6052	3,7816	3,3192	3,0173	2,8020	2,5113	2,3209	2,1848	2,0385	1,8783	1,5923

7 lentelė. Fišerio transformacija $z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$; $z_{-r} = -z_r$

r	z_r	r	z_r	r	z_r	r	z_r	r	z_r
.000	.000	.200	.203	.400	.424	.600	.693	.800	1.099
.005	.005	.205	.208	.405	.430	.605	.701	.805	1.113
.010	.010	.210	.213	.410	.436	.610	.709	.310	1.127
.015	.015	.215	.218	.415	.442	.615	.717	.815	1.142
.020	.020	.220	.224	.420	.448	.620	.725	.820	1.157
.025	.025	.225	.229	.425	.454	.625	.733	.825	1.172
.030	.030	.230	.234	.430	.460	.630	.741	.830	1.188
.035	.035	.235	.239	.435	.466	.635	.750	.835	1.204
.040	.040	.240	.245	.440	.472	.640	.758	.840	1.221
.045	.045	.245	.250	.445	.478	.645	.767	.845	1.238
.050	.050	.250	.255	.450	.485	.650	.775	.850	1.256
.055	.055	.255	.261	.455	.491	.655	.784	.855	1.274
.060	.060	.260	.266	.460	.497	.660	.793	.860	1.293
.065	.065	.265	.271	.465	.504	.665	.802	.865	1.313
.070	.070	.270	.277	.470	.510	.670	.811	.870	1.333
.075	.075	.275	.282	.475	.517	.675	.820	.875	1.354
.080	.080	.280	.288	.480	.523	.680	.829	.880	1.376
.085	.085	.285	.293	.485	.530	.685	.838	.885	1.398
.090	.090	.290	.299	.490	.536	.690	.848	.890	1.422
.095	.095	.295	.304	.495	.543	.695	.858	.895	1.447
.100	.100	.300	.310	.500	.549	.700	.867	.900	1.472
.105	.105	.305	.315	.505	.556	.705	.877	.905	1.499
.110	.110	.310	.321	.510	.563	.710	.887	.910	1.528
.115	.116	.315	.326	.515	.570	.715	.897	.915	1.557
.120	.121	.320	.332	.520	.576	.720	.908	.920	1.589
.125	.126	.325	.337	.525	.583	.725	.918	.925	1.623
.130	.131	.330	.343	.530	.590	.730	.929	.930	1.658
.135	.136	.335	.348	.535	.597	.735	.940	.935	1.697
.140	.141	.340	.354	.540	.604	.740	.950	.940	1.738
.145	.146	.345	.360	.545	.611	.745	.962	.945	1.783
.150	.151	.350	.365	.550	.618	.750	.973	.950	1.832
.155	.156	.355	.371	.555	.626	.755	.984	.955	1.886
.160	.161	.360	.377	.560	.633	.760	.996	.960	1.946
.165	.167	.365	.383	.565	.640	.765	1.008	.965	2.014
.170	.172	.370	.388	.570	.648	.770	1.020	.970	2.092
.175	.177	.375	.394	.575	.655	.775	1.033	.975	2.185
.180	.182	.380	.400	.580	.662	.780	1.045	.980	2.298
.185	.187	.385	.406	.585	.670	.785	1.058	.985	2.443
.190	.192	.390	.412	.590	.678	.790	1.071	.990	2.647
.195	.198	.395	.418	.595	.685	.795	1.085	.995	2.994

Vartojamų terminų anglų-lietuvių kalbų žodynėlis

Statistiniai, kaip ir kiti okultiniai pranašysčių metodai turi savą žargoną, tyčia išgalvotą tam, kad juos padarytų sunkiai suprantamus žmonėms, tų metodų netaikantiems.

G. O. Ellis

a posteriori (posterior) probability – aposteriorinė tikimybė
a priori (prior) probability – apriorinė tikimybė
absolute continuous random variable – absoliučiai tolydus atsitiktinis dydis
alternative hypothesis – alternatyva
bar graph – stulpelių diagrama
box-and-whiskers plot – stačiakampė diagrama
categorical variable – kategorinis kintamasis
central limit theorem – centrinė ribinė teorema
certain event, Ω – būtinasis įvykis
Chebyshev rule – Čebyšovo taisyklė
chi-square goodness-of-fit test – χ^2 suderinamumo kriterijus
chi-square test of homogeneity – χ^2 homogeniškumo kriterijus
chi-square test of independence – χ^2 nepriklausomumo kriterijus
cluster sample – lizdinė imtis
coefficient of variation (CV) – kitimo (variacijos) koeficientas
complementary (contrary) event – priešingasis įvykis
conditional probability – sąlyginė tikimybė
confidence interval – pasikliautinasis intervalas
confidence level – pasikliovimo lygmuo
consistent estimator – suderintasis įvertis
contingency coefficient – kontingencijos koeficientas
continuous variable – tolydusis kintamasis
correlation coefficient – koreliacijos koeficientas
covariance – kovariacija
cumulative frequency – sukauptasis dažnis
data set – duomenų aibė
degrees of freedom – laisvės laipsniai
density – tankis
descriptive statistics – aprašomoji statistika
dichotomious variable – dvireikšmis kintamasis
discrete random variable – diskretusis atsitiktinis dydis
discrete variable – diskretusis kintamasis
distribution – skirstinys
distribution function – pasiskirstymo funkcija
effective estimator – efektyvusis įvertis
empirical rule – empirinė taisyklė
entropy – entropija
error margin – paklaidos rėžis
estimate – įverčio realizacija
estimator – įvertis, įvertinys
expected frequency – tikėtinasis dažnis
expected value of random variable – atsitiktinio dydžio vidurkis, matematinė viltis
first quartile (Q_1) – pirmasis kvartilis
Fischer transformation – Fišerio transformacija
frequency – dažnis
frequency distribution – dažnių skirstinys
frequency distribution function – dažnių pasiskirstymo funkcija
frequency function – empirinė dažnio (tankio) funkcija
frequency polygon – dažnių daugiakampis
frequency table – dažnių lentelė
grouped data – grupuotieji duomenys
histogram – histograma
impossible event, $\varnothing$ – negalimasis įvykis
independent events – nepriklausomieji įvykiai
index of predictive association (λ) – sąlyginės prognozės indeksas
index of qualitative variation (IQV) – kokybinės įvairovės indeksas
interquartile range (IQR) – kvartilių skirtumas
intersection of events – įvykių sankirta
interval scale – intervalų skalė
judgment sample – ekspertinė imtis
kurtosis – eksceso koeficientas
law of large numbers – didžiųjų skaičių dėsnis
level of significance (α) – reikšmingumo lygmuo
maximum likelihood method – didžiausiojo tikėtinumo metodas
measures of dispersion – sklaidos charakteristikos
measures of location – padėties charakteristikos
median – mediana

method of moments – momentų metodas
mode – moda
moment – momentas
mutually exclusive events – nesutaikomieji įvykiai
nominal scale – pavadinimų skalė
non-probability sample – netikimybinė imtis
normal approximation – normalioji aproksimacija
normal curve – normalioji kreivė
null hypothesis – nulinė hipotezė
ogive – sukauptųjų santykinių dažnių laužtė, ogivė
one-sided test – vienpusis kriterijus
opportunity sample – proginė imtis
ordered array – variacinė eilutė
ordinal scale – rangų skalė
outlier – išskirtis
paired sample – porinė imtis
parallel sample – lygiagrečioji imtis
Pareto diagram – Pareto diagrama
percentiles – procentiliai
pie chart – skritulinė diagrama
Poisson approximation – puasoninė (Puasono) aproksimacija
population – populiacija
population parameter – populiacijos parametras
prediction interval – prognozės intervalas
probability – tikimybė
probability sample – tikimybinė imtis
qualitative variable – kokybinis kintamasis
quantiles – kvantiliai
quantitative variable – kiekybinis kintamasis
quartiles – kvartiliai
quota sample – kvotinė imtis
random event – atsitiktinis įvykis
random sampling error – atsitiktinė imties paklaida
random variable – atsitiktinis dydis
range – duomenų aibės plotis, amplitudė
ratio scale – santykių skalė
rejection region (critical region), W – kritinė sritis, atmetimo sritis
relative frequency – santykinis dažnis
representative sample – reprezentatyvi imtis
response rate – atsakymo lygis
sample – imtis
sample mean – imties vidurkis, empirinis vidurkis
sample size – imties didumas
sample statistics – imties statistika
sample variance – imties dispersija, empirinė dispersija
sampling with replacement – grąžintinis ėmimas
sector – išpjova, sektorius
simple (elementary) event – elementarusis įvykis
simple random sample – paprastoji atsitiktinė imtis
skewness – asimetrijos koeficientas
standard deviation (Std) – standartinis nuokrypis
standard normal distribution – standartinis normalusis skirstinys
standard score – standartizuotoji reikšmė
statistical inference – statistinės išvados
stem-and-leaf plot – diagrama medis
strata – sluoksniai
stratified sample – sluoksninė imtis
systematic sample – sistemingoji imtis
systematic sampling error – sistemingoji imties paklaida
tail probability – uodegos tikimybė
test statistic – kriterijaus statistika
third quartile (Q_3) – trečiasis kvartilis
trimmed mean – nupjautasis vidurkis
two-sided test – dvipusis kriterijus
type I and type II errors – I ir II rūšies klaidos
unbiased estimator – nepaslinktasis įvertis
union of events – įvykių sąjunga
variable – kintamasis
variance – dispersija
p-value – p-reikšmė
t-test – t kriterijus
z-score – z reikšmė

Dalykinė rodyklė

Literatūra

Išsami matematinės statistikos literatūros iki 1993 metų bibliografija pateikta J. Kruopio [6] knygoje. Todėl čia nurodomi tik didesnės apimties lietuviški mokomieji statistikos ir tikimybių teorijos leidiniai, išleisti po 1993 metų.

1. Aprašomoji statistika: mokomoji priemonė / Parengė S. Martišius ir kt.; red. J. Markelevičius. Vilnius: VU l-kla, 1994, 138 p.
2. Aprašomoji statistika: mokomoji priemonė. Vilnius: VU l-kla, 1998, 136 p.
3. Bikelienė V. Taikomosios matematinės statistikos elementai. Vilnius: VU l-kla, 1993, 101 p.
4. Bučys K. Tikimybių teorija ir matematinė statistika: mokomoji priemonė. Klaipėda: KU, 1994, 159 p.
5. Eidukevičius R., Juknevičienė D., Kosareva N., Pamerneckis S. Matematinė statistika istorijoje. Vilnius: VU l-kla, 1998, 280 p.
6. Kanišauskas V. Tikimybių teorijos ir matematinės statistikos pagrindai: mokomoji knyga. Šiauliai: ŠU l-kla, 2000, 147 p.
7. Kruopis J. Matematinė statistika. Vilnius: Mokslas, 1993, 416 p.
8. Kubilius J. Tikimybių teorija ir matematinė statistika. Vilnius: Mokslas, 1996, 439 p.
9. Martišius S. Statistinių išvadų teorijos pradmenys: mokomoji priemonė. Vilnius: VU l-kla, 1997, 119 p.
10. Mišeikis F. Statistika ir ekonometrija: vadovėlis studijuojantiems pagal vadybos studijų programas. Vilnius: Technika, 1997, 275 p.
11. Sakalauskas V. Statistika su „Statistica": aukštųjų mokyklų studentams. Vilnius: Margi raštai, 1998, 227 p.

Anglų kalba yra išleista gausybė statistikos vadovėlių, iš kurių pateikiami tik keli populiariausi.

1. Afifi A., Clark V. A. Computer-aided Multivariate Analysis. 3rd edn. London: Chapman and Hall, 1996.
2. Berenson M. L., Levine D. M. Basic Business Statistics: Concepts and Applications. 7th edn. Prentice Hall, 1999, 1114 p.
3. Healey J. F. Statistics: A Tool for Social Research. 5th edn. Wadsworth Inc., 1999.
4. Hinkley D. E., Wiersma W., Jurs S. G. Applied Statistics for the Behavioral Sciences. 4th edn. Boston, MA: Haughton Mifflin, 1997.
5. Howell D. C. Statistical Methods for Psychology. 4th edn. Boston, 1997.

1

Vydas Čekanavičius, Gediminas Murauskas

STATISTIKA IR JOS TAIKYMAI I

2001 11 22. 15 sp. l. Tiražas 2000 egz. Užs. Nr. 1211
Leidykla TEV, Akademijos g. 4, LT-2600 Vilnius
Spausdino AB „Vilspa" spaustuvė
Viršuliškių skg. 80, LT-2056 Vilnius

5

Prieš pradėdama masinę
dietinių „mėsainių su lašinių kvapu"
gamybą, užkandinė „Mak-kauskas"
paprašė 100 lankytojų įvertinti naująjį produktą.
Teigiamai naująjį produktą įvertino 63 lankytojai.
Ar šie duomenys neprieštarauja naujojo
mėsainio kūrėjo reklaminiam teiginiui,
kad pagamintas produktas patiks
bent dviem iš trijų lankytojų?

ISBN 9986-546-93-1

Graikiškos raidės

Ar galima fizikoje ar astronomijoje išsiversti be graikiškų raidžių? Sunkiai :)

Raidė	Kaip užrašyti	Raidė	Kaip užrašyti
α	\alpha	σ	\sigma
β	\beta	ς	\varsigma
γ	\gamma	τ	\tau
δ	\delta	υ	\upsilon
ϵ	\epsilon	ϕ	\phi
ε	\varepsilon	φ	\varphi
ζ	\zeta	χ	\chi
η	\eta	ψ	\psi
θ	\theta	ω	\omega
ϑ	\vartheta	Γ	\Gamma
ι	\iota	Δ	\Delta
κ	\kappa	Θ	\Theta
λ	\lambda	Λ	\Lambda
μ	\mu	Ξ	\Xi
ν	\nu	Π	\Pi
ξ	\xi	Σ	\Sigma
π	\pi	Υ	\Upsilon
ϖ	\varpi	Φ	\Phi
ρ	\rho	Ψ	\Psi
ϱ	\varrho	Ω	\Omega